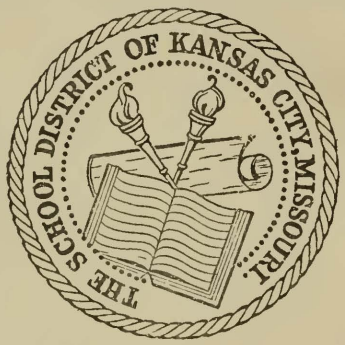
Bound Periodical

649013

Kansas City Public Library



This Volume is for

REFERENCE USE ONLY

10 29—6m—P

From the collection of the



San Francisco, California
2008

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
H. P. Charlesworth — E. H. Colpitts
L. F. Morehouse — H. D. Arnold — O. B. Blackwell
W. H. Harrison — Philander Norton, *Editor*

VOLUME VIII
1929

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

PUBLIC LIBRARY
KANSAS CITY
MO

Bound
Periodical
649013
Je 2 '30

The Bell System Technical Journal

January, 1929

Decibel—The Name for the Transmission Unit

By W. H. MARTIN

IN 1923 the "mile of standard cable" was replaced in the Bell System by a new unit for expressing telephone transmission efficiencies and levels. At that time the generic term "transmission unit" was taken to designate this new unit, since it was considered desirable to defer the adoption of a more distinctive name until this unit had been given further consideration by others who would have use for a unit of this type. This new unit is defined by the statement that two amounts of power differ by one transmission unit when they are in the ratio of $10^{.1}$, and any two amounts of power differ by N transmission units when they are in the ratio of $10^{N(.1)}$. In accordance with this, the number of transmission units corresponding to the ratio of any two powers is ten times the common logarithm of that ratio.

For a unit of this kind, it is evidently desirable to have universal use. Accordingly, the Bell System, prior to its adoption of the transmission unit, discussed this matter with various foreign telephone administrations, and suggested their consideration of the use of the proposed "transmission unit." A number of these administrations expressed a favorable attitude towards this unit.

In 1924 there was organized the International Advisory Committee on Long Distance Telephony in Europe. The purpose of this committee, which is composed of representatives of the various telephone administrations of Europe, is to recommend standards and practices for the development of telephone service between the European countries. One of the early considerations of this committee was this proposal of the universal standardization of a unit for telephone transmission work. This brought forth a difference of view, since some of the countries represented on this committee wished to continue their use of a unit based on naperian or natural logarithms, for which the basic power ratio is e^2. The characteristics of the unit based on decimal logarithms and that based on natural logarithms and their relative merits were discussed in a number of papers which were published at that time * and so need not be rehearsed here.

At the request of the International Advisory Committee, representatives of the Bell System attended some of their meetings at which

this matter was discussed. In this joint consideration there arose the suggestion that the fundamental unit on the decimal basis be defined to be equal in magnitude to that of ten transmission units, so that the basic power ratio would be 10^1. The units proposed thus came to one based on the power ratio of 10^1 and the other on the power ratio of e^2, with the provision that decimal submultiples of either unit could be employed, using the customary prefixes to give the proper indication. On this basis, the numbers of the two kinds of units corresponding to a given power ratio, differ by about 15 per cent. It was further suggested that the naperian unit be called the "neper" and that the fundamental decimal unit be called the "bel," these names being derived respectively from the names of Napier, the inventor of natural logarithms, and Alexander Graham Bell.

These joint considerations have had the following results. The European International Advisory Committee has recommended to the various European telephone administrations that they adopt either the decimal or naperian unit and designate them the "bel" and "neper" respectively. The Bell System has adopted the name "decibel" for the "transmission unit," based on a power ratio of $10^{.1}$. This is in accordance with the terminology for the decimal unit, the prefix "deci" being the usual one for indicating a one-tenth relation. For convenience, the symbol "db" will be employed to indicate the name "decibel."

* *References:*

"The Transmission Unit and Telephone Transmission Reference Systems," W. H. Martin, *A.I.E.E. Trans.*, Vol. 43, pp. 797–801, 1924, and *B.S.T.J.*, Vol. 3, pp. 400–408, July, 1924.

"The Transmission Unit," R. V. L. Hartley, *Electrical Communications*, Vol. 3, pp. 34–42, June, 1924.

"Telephone Transmission Standards," Dr. F. Breisig, *The Electrician*, Vol. 94, pp. 114–115, Jan. 30, 1925.

"Telephone Transmission Units," Colonel T. F. Purves, *The Electrician*, Vol. 94, pp. 144–145, Feb. 6, 1925.

"Telephone Transmission," G. D. Hood, *The Electrician*, Vol. 94, p. 369, March 27, 1925.

"Telephone Transmission Standards," Dr. F. Breisig, *The Electrician*, Vol. 94, pp. 454–455, April 17, 1925.

"The Telephone Transmission Unit," Colonel T. F. Purves, *The Electrician*, Vol. 94, pp. 535–542, May 8, 1925.

"Telephone Transmission Unit," F. B. Jewett, *The Electrician*, Vol. 94, pp. 562–563, May 15, 1925.

"Telephone Transmission Units," Dr. F. Breisig, *The Electrician*, Vol. 94, p. 630, May 29, 1925.

"The Telephone Transmission Unit," Dr. F. Luschen, *The Electrician*, Vol. 95, pp. 94–95, July 24, 1925.

The Principles of Electric Circuits Applied to Communication [1]

By F. S. OSBORNE

SYNOPSIS: This paper discusses the method of presenting in the curricula of engineering schools the fundamental electrical principles, emphasizing the desirability of presenting them as far as practical in a general way and of making clear the relations of specific applications, such as the relation between circuit theory equations as applied to power systems and to telephone systems, and the relation between ordinary circuit theory and the generalized electromagnetic equations. An outline is given of some interesting problems arising and results obtained in the application of electric principles to telephone systems.

ENGINEERING education shares with other forms of education the general movement toward greater emphasis on the unity of subject matter which plays such an important part in the development of modern educational methods. By the unity of subject matter I refer to the aim to reduce as far as practicable the number of separate compartments in which the educational subject matter is kept, and to present this subject matter under a smaller number of broader headings. Whereas the curriculum must be divided into a certain number of different courses for administrative and practical reasons, I take it the modern educational method is opposed to the presentation of these courses as individual entities, separated from other subjects, but insists rather that the curricular partitions be kept as low as possible so that the student may appreciate as fully as possible the continuity of each subject with its neighbors, and may obtain a good perspective of the close mutual relations of the different parts of the educational material and realize their mutual dependence and the large areas in which they are jointly applicable.

The wisdom of this move is evident to men in the industries as well as to educators. The tremendously rapid growth of fundamental electrical science and of the electrical industries have both worked rapidly in this direction. The time has long passed when it was at all possible to cover both fundamental electrical science and its applications in a four year engineering course. I judge that the old question "Is it more important to teach the fundamentals of electrical science

[1] Presented at Pittsburgh July 18, 1928, at the Summer School for Electrical Engineering Teachers under the auspices of the Society for the Promotion of Engineering Education.

The author expresses his appreciation of the assistance given him by Mr. C. O. Gibbon, American Telephone and Telegraph Company, in the preparation of this paper.

or their application?" is no longer even discussible, as far as the engineering school is concerned. Applications must indeed be learned by the student, but this can best be done after graduation, with the help of the industry employing him, and under the stimulus of the necessity of learning his job and preparing himself for greater usefulness in the organization. The question before the colleges is not "Shall we teach fundamentals?" but "What fundamentals shall we teach, and how can these most effectively be presented? How much of the fundamentals of our present far flung electrical science can we convey to the students in a four year course?"

The same question naturally presents itself in this discussion. The communication field, from the nature of the problems which it presents, has a great wealth of material relating to the principles of electric circuits, both as regards the practical application of these principles and research extending our fundamental knowledge of these principles. What phases of this subject matter should be presented here? In this Mr. Hammond's circular to the members of the summer school staffs is a guide. He points out that among the various purposes of the summer schools the principal aim is to produce tangible results for improving methods of teaching. I do not understand that I am expected, in talking with a group of educators, to discuss teaching methods directly, and will not presume to do so. However, in response to your invitation to discuss "The Principles of Electric Circuits Applied to Communication" I shall refer to the use in the communication field of those principles within the scope of student work which appear to have the broadest general application, including all fields of electrical engineering. Also, it will no doubt be of interest for us to give some consideration to the relative conditions, including similarities and differences, of application of these principles to communication problems and to other branches of the electrical industries.

Limiting the discussion in this way necessarily results in leaving untouched many phases of the application of electrical principles to communication which are of great interest but of less application in other fields, and so I have left out such important matters as, for example, modulation and demodulation, the balancing of line impedance characteristics by artificial lines, inductive effects between different circuits, and the performance characteristics of various types of apparatus.

Also, of course, this will in no sense be a general discussion of the work of the transmission engineers of the telephone companies. While their work is based on the application of electrical principles, and requires that they understand those principles, the theoretical

work which they do on electric circuits is, with the exception of a few men, relatively small, and their work is also very largely based on the use of economic principles, knowledge of the telephone system, general business principles and common sense.

General Principles

Communication circuits in general are very complicated networks and the application of electric principles to these networks involves the solution of numerous new problems and the development of a great many practical approximations. To be effective in this work it is of prime importance that the young engineer have a good grounding in the general fundamental principles of direct and alternating current theory. By a good grounding we mean an appreciation of the generality of these principles so that they can be applied by the student to the problems of his particular work, no matter in what branch of electrical engineering that work may be. He needs also to have a facility in their application to new problems. The relations of resistance, reactance, conductance and susceptance, the use of Kirchhoff's laws, the relations of resonance and conditions for maximum transfer of energy between two branches of a network—all these we would list as a matter of course. We should include also in the list of basic fundamentals some simple but extremely useful practical theorems, including the reciprocity theorem and Pollard's or Thevenin's theorem.

Along with these fundamental principles we believe it helpful to a man in any branch of electrical work to have absorbed the idea of the equivalence of networks of different types; for example, the expression, in terms of equivalent T or Pi circuits or other convenient form of the characteristics of any 4 pole network (that is, a network with two input and two output terminals) from the measurements which can be made at its terminals. Fundamental training of this sort gives the man a mastery in the solution of electrical problems, not only through ability to place a given problem in its most convenient form, but by assisting the engineer in the formation of correct and simple physical ideas regarding the processes which are taking place.

A very good illustration of this point is given by the transformer. Most young men graduating from engineering schools, I believe, think of the operation of the transformer in terms of its vector diagram. This is a valuable way of getting a physical picture of the effect of a transformer which is useful in certain types of problems but less convenient in others. If in addition to the vector diagram the student is taught the equivalent network of a transformer, as is done in some

texts, he can more easily apply it to other types of problems, and the equivalent T network of the transformer is so simple that it is very helpful in showing the variation in the performance of the circuit with changes in the constants of any part of the circuit.

Familiarity with the fundamental principles of networks and equivalent circuits is particularly helpful in case of a man who comes in contact with problems associated with networks made up of a number of similar sections, as for example, electric filters, which already play such an important part in certain branches of the communication art.

In the study of fundamental principles it would appear to be very helpful in enabling the student to get an appreciation of their generality if the specific problems and illustrations used in his work are drawn from the various fields of electric work rather than from a single field. The communication field is replete with specific problems illustrating these principles which are very suitable for the use of the student.

Transmission Line Theory

In discussing the application of the principles of electric circuits to communication it is natural to give particular attention to transmission line theory because of its importance in connection with the transmission of electrical energy for any purpose whatever, including both power and communication services, and because of the interest of the problems it involves. Transmission line theory in one sense dates back to Lord Kelvin who in 1855 applied laws of diffusion of heat to the determination of the flow of electricity through long submarine cables. This solution ignored the effect of line inductance which was unimportant in the particular problem to which Lord Kelvin applied this solution, but which is very important in any general transmission line theory. Through the work of Heaviside and others the general transmission line theory was at an early date applied to telephony. It is, of course, in relatively recent years that the great development of long distance power transmission lines has made the general theory of value in power transmission work, the performance of early alternating current systems being adequately represented by approximate formulas, entirely neglecting the effect of the capacity of the line. It is indeed not long ago that the effect of line capacity, assumed lumped at one or two points, became important and still more recently that it became necessary in power problems to take more accurate account of the distribution of the capacity.

It is no doubt partly as a result of the historical development of the application of transmission line theory to telephony and to power transmission problems, and partly the result of differences in conditions

which the solution must meet in the two fields of application that it seems in the past not always to have been made clear to students that the power line equations and the telephone line equations are simply special solutions of the same general line formula.

To illustrate this point it is desirable to refer to a few well-known equations. The differential equations for what may be called the classical transmission line theory are given in equation 1. Equation 2 is a solution of these differential equations for the steady state, for the circuit indicated in Fig. 1.

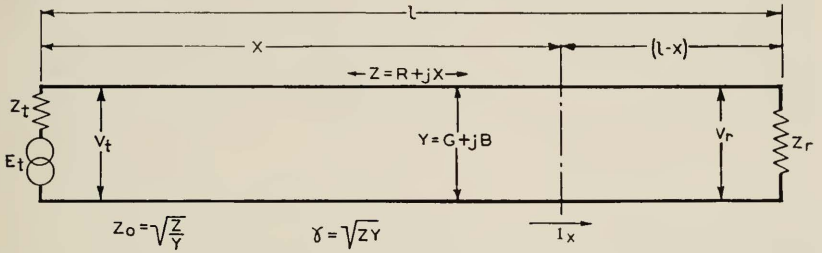


Fig. 1

$$\left.\begin{aligned} \left(L\frac{d}{dt}+R\right) I &= -\frac{\partial}{\partial x} V, \\ \left(C\frac{d}{dt}+G\right) V &= -\frac{\partial}{\partial x} I, \end{aligned}\right\} \tag{1}$$

$$\begin{aligned} I_x = \frac{E_t}{Z_0+Z_t}\Bigg\{ & \epsilon^{-\gamma x}\left[1+\left(\frac{Z_0-Z_t}{Z_0+Z_t}\cdot\frac{Z_0-Z_r}{Z_0+Z_r}\right)\epsilon^{-2\gamma l}\right. \\ & \left.+\left(\frac{Z_0-Z_t}{Z_0+Z_t}\cdot\frac{Z_0-Z_r}{Z_0+Z_r}\right)^2\epsilon^{-4\gamma l}+\cdots\right] \\ & +\epsilon^{-\gamma(l-x)}\left[\frac{Z_0-Z_r}{Z_0+Z_r}\epsilon^{-\gamma l}+\frac{Z_0-Z_t}{Z_0+Z_t}\left(\frac{Z_0-Z_r}{Z_0+Z_r}\right)^2\epsilon^{-3\gamma l}\right. \\ & \left.\left.+\left(\frac{Z_0-Z_t}{Z_0+Z_t}\right)^2\left(\frac{Z_0-Z_r}{Z_0+Z_r}\right)^3\epsilon^{-5\gamma l}+\cdots\right]\right\}. \end{aligned} \tag{2}$$

This equation indicates the current flowing at any point in the circuit for a given impressed voltage. The solution in this form seems to have special educational value because it gives a very clear physical picture of what is taking place in the transmission line. As was early pointed out by Heaviside, the current flowing in any simple circuit such as indicated can be considered to be built up from a directly transmitted wave which at the transmitting end has the magnitude of $\frac{E_t}{Z_t+Z_0}$. This direct wave is attenuated as it is propagated along the

line, and at the receiving end is reflected in the ratio of the difference between the receiving impedance and characteristic line impedance to the sum of these impedances; propagated back toward the transmitting end with continued attenuation; reflected there if the transmitting end impedance is not equal to the characteristic line impedance, the doubly reflected wave propagated toward the receiving end, and so on in an infinite series of propagations and reflections.

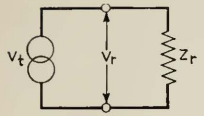
In both power and communication work one quantity which is of great importance in considering the characteristics of the transmission line is the ratio of the voltage at the transmitting end to that at the receiving end. This is, of course, readily derived from equation 2, and is presented in equation 3 in the beautifully compact form offered by the use of hyperbolic functions of the propagation constant of the line.

$$\frac{V_t}{V_r} = \cosh \gamma l + \frac{Z_0}{Z_r} \sinh \gamma l. \tag{3}$$

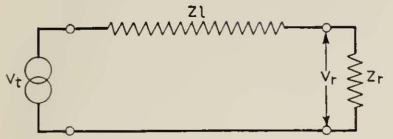
Considering first the application of this general transmission line theory to power transmission circuits, it would appear to be of great value for the student to appreciate the relationship between the general formula and the approximations used for short lines. This is brought out clearly by equation 4 and the diagrams and equations presented under 5.

$$\frac{V_t}{V_r} = 1 + \frac{Zl}{Z_r} + \frac{YZl^2}{2!} + \frac{YZ^2l^3}{3!Z_r} + \frac{Y^2Z^2l^4}{4!} + \frac{Y^2Z^3l^5}{5!Z_r} + \cdots \tag{4}$$

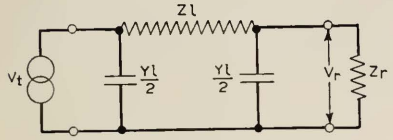
Equation 4 is simply the development of the general formula of equation 3 into a series of terms of ascending powers of Zl, the total resistance and reactance of the line, and of Yl the total shunt admittance of the line. The comparison of this series expansion with the results indicated by various approximate methods is of considerable interest. The first term, unity, is naturally the ratio of transmitting and receiving voltages with no transmission line, as indicated in 5*a*. With the addition of the second term $\left(\frac{Zl}{Z_r}\right)$ one has the result obtained by entirely ignoring the capacity of the line as indicated in 5*b*. The first three terms of the series give the result obtained by assuming that one half of the capacity of the line is concentrated at each end of the line as indicated in 5*c*, namely, a simple Pi network. The simple T network, assuming the capacity all concentrated at the middle as indicated in 5*d*, gives 4 terms, but you will note that the fourth term



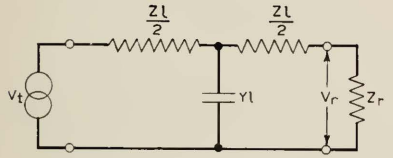
$$\frac{V_t}{V_r} = 1 \qquad (5a)$$



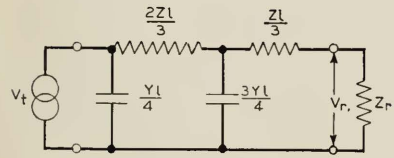
$$\frac{V_t}{V_r} = 1 + \frac{Zl}{Z_r} \qquad (5b)$$



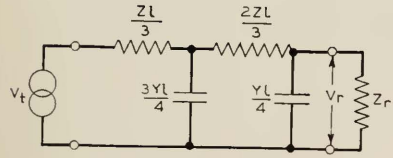
$$\frac{V_t}{V_r} = 1 + \frac{Zl}{Z_r} + \frac{YZl^2}{2!} \qquad (5c)$$



$$\frac{V_t}{V_r} = 1 + \frac{Zl}{Z_r} + \frac{YZl^2}{2!} + \frac{YZ^2l^3}{4Z_r} \qquad (5d)$$



$$\frac{V_t}{V_r} = 1 + \frac{Zl}{Z_r} + \frac{YZl^2}{2!} + \frac{YZ^2l^3}{3!Z_r} \qquad (5e)$$



$$\frac{V_t}{V_r} = 1 + \frac{Zl}{Z_r} + \frac{YZl^2}{2!} + \frac{YZ^2l^3}{3!Z_r} + \frac{Y^2Z^2l^4}{4!} \qquad (5f)$$

is 50 per cent. greater in magnitude than the fourth term of the series, and this approximation, therefore, does not commend itself for determining the voltage ratio, since if one wishes a precision requiring four terms of the series it is naturally better to use the correct fourth term. In order to correctly represent four terms of the series by an equivalent network it is necessary, as indicated in 5*e*, to assume one fourth of the capacity concentrated at the sending end of the line and the other three fourths concentrated at a point two thirds distant from the sending end. Finally, if this unsymmetrical network be reversed in direc-

tion, as indicated in 5*f*, the first five terms of the series are accurately reproduced.

The degree of approximation represented by dropping off various terms of this series is indicated for three typical cases in Figures 2, 3, and 4. Fig. 2 represents a typical 11,000 volt distribution line. It is to be noted that even neglecting the third term, which is the first in which the capacity of the line appears, results in an error of only 0.4 per cent. as indicated by consulting the values of ratios between successive terms of the series which are given with the diagram. The first ratio shown is that between the second and first term, and the



Fig. 2

20 Miles, 11 Kv., Three Phase Distribution Line
200 Kw. at 90% Power Factor, 60 ~
Ratio of Terms:

$$\frac{Zl}{Z_r} = .0540/2°[1\ \tfrac{1}{3}\ \cdots]$$

$$YlZ_r = .0668/116°[\tfrac{1}{2}\ \tfrac{1}{4}\ \cdots]$$

second ratio when divided by two is that between the third and second term. The ratios between successive terms in the series are equal alternately to these two values divided by coefficients increasing in simple arithmetical progression as indicated. This fact would seem to make the series very convenient for computations.

To take the opposite extreme of power transmission, Fig. 3 has been prepared showing the degree of approximation resulting from the use of different numbers of terms of the series for a 220 kv. line, 250 miles

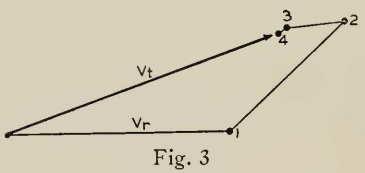


Fig. 3

250 Mile, 220 Kv., Three Phase Power Line
100,000 Kw. at 90% Power Factor, 60 ~
Ratio of Terms:

$$\frac{Zl}{Z_r} = .464/60°[1\ \tfrac{1}{3}\ \cdots]$$

$$YlZ_r = .588/116°[\tfrac{1}{2}\ \tfrac{1}{4}\ \cdots]$$

long, transmitting 100,000 kilowatts. Here it is seen that in order to get a precision of a fraction of one per cent. four terms are necessary. The results obtained by a simple Pi network are indicated by the results of the first three terms. It is seen that even for this somewhat extreme case the computation by the series expansion is not at all laborious. This is, of course, due to the fact that even for 250 miles at 60 cycles the power line represents only about one-fifteenth of a wave-length.

Series expansions of the type discussed are, of course, not novel. I have dwelt on these somewhat, however, because of the value which

they appear to have in clarifying the students' ideas about electric transmission, and because few students appear to be familiar with this method of treatment.

It is no doubt true that in many cases the student can best start transmission line theory with a simple approximation. It would seem, however, that before he gets through his study of the principles of electric circuits he should have a clear picture of the physical processes involved in the propagation of electric power over transmission lines, such as is given by equation 2, and of the assumptions involved in the various approximations which may be presented to him. If the scope of the course is not such as to permit the derivation of the general equations from the differential equations, it is possible to get equation

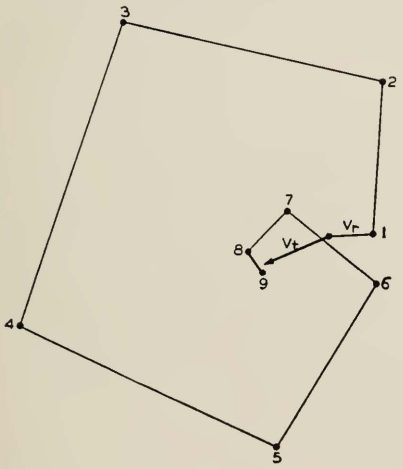


Fig. 4

100 Mile, 8BWG Copper Telephone Circuit, 1000 ~
Terminating Impedance Equals Characteristic Impedance

Ratio of Terms:

$$\frac{Zl}{Z_r} = YlZ_r$$
$$= 3.5/83°[1\ \tfrac{1}{2}\ \tfrac{1}{3}\ \tfrac{1}{4}\ \tfrac{1}{5}\ \tfrac{1}{6}\ \tfrac{1}{7}\ \tfrac{1}{8}\ \cdots]$$

4 by a method of successive approximations as is shown in at least one textbook, and from 4 to derive the general equation 2.

In contrast with the power line cases, Fig. 4 indicates the results obtained by the series computation of a relatively short telephone toll line, an open wire circuit 100 miles long, and using 1,000 cycles as one typical telephone frequency. Although this line is only a little more than half a wave-length, the solution by the series for this case is quite laborious and indeed impracticable, and of course, would be even more so with the longer lengths or the higher attenuations ordinarily involved in telephone circuits.

The form of the general equation under discussion which is found most convenient for telephone use is given in equation 6.

$$\frac{E_t}{V_r} = \frac{Z_t + Z_r}{Z_r} \cdot \epsilon^{\gamma l} \cdot \frac{\dfrac{Z_0 + Z_t}{2\sqrt{Z_0 Z_t}} \cdot \dfrac{Z_0 + Z_r}{2\sqrt{Z_0 Z_r}}}{\dfrac{Z_t + Z_r}{2\sqrt{Z_t Z_r}}} \cdot \left[1 - \frac{Z_0 - Z_t}{Z_0 + Z_t} \cdot \frac{Z_0 - Z_r}{Z_0 + Z_r} \cdot \epsilon^{-2\gamma l} \right]. \quad (6)$$

This applies to the same circuit as equations 3 and 4, the only difference being that for telephone purposes the ratio E_t/V_r is used rather than V_t/V_r since in the telephone case the internal voltage of the generating apparatus rather than the terminal voltage is the more convenient reference voltage. Otherwise the two equations are identical consisting simply of a rearrangement of terms. The outstanding feature of the arrangement shown in equation 6 is that it consists of the product of a number of terms rather than the sum of a series of terms as in equation 4. The first term represents the ratio of voltages which would be obtained if there were no line present, namely $\frac{Z_t + Z_r}{Z_r}$. The second term illustrates the effect of propagation over the line itself without allowance for reflection at the terminals. The next three terms, two in the numerator and one in the denominator are the factors which make allowance for this reflection. Each is dependent upon simply the magnitude of two impedances, and their inclusion in the equation represents the fact that inserting the line between the two impedances inserts the reflection factors between the line and the transmitting impedance at one end, and the line and receiving impedance at the other end, and takes out the reflection factor directly between the transmitting and receiving impedances. The reflection factor becomes unity in any case in which the two impedances are equal. The last term of the equation is called the interaction factor because it represents the effect of multiple reflection back and forth between the two terminals of the line. This factor necessarily is complicated as it depends upon the characteristics of the line and on both transmitting and receiving impedances. It will be noted that this factor becomes substantially equal to unity in case either the transmitting impedance and the line impedance are approximately equal, the receiving impedance and the line impedance are approximately equal, or the attenuation of the circuit is considerable. In most practical telephone circuits these conditions are approximated sufficiently closely so that the departure of this factor from unity can usually be neglected.

In general, practical telephone circuits are, of course, a good deal

more complicated in form than the simple circuit indicated above. It is, however, necessary to have for practical telephone purposes, means for readily determining to a good degree of approximation the overall efficiency of these complicated circuits as a part of the every-day work of certain departments of the telephone companies. The process of expressing the efficiency in terms of the product of a number of factors provides a convenient means for doing this. Under these conditions the large factors such as line attenuation and certain other factors are determined for the individual circuits, whereas factors which are close to unity can be treated approximately by tables representing various types of cases rather than individual circuits.

The convenience of treatment of circuit equations in this way for telephone use has led to the use of a logarithmic measure for expressing the efficiencies of telephone circuits. Although the computations can often be made in terms of currents or voltages, where changes of impedance are involved, such as inequality ratio transformers, we are, of course, concerned with variations in power rather than variations in either the current or voltage. The losses in a telephone circuit are therefore expressed in terms of the logarithm of the ratio of input power to output power. The unit is so defined that 10 transmission units correspond to a ratio of 10, 20 transmission units correspond to a ratio of 100, etc. The overall efficiency of practical telephone circuits from transmitter to receiver is in many cases in the order of 20 transmission units, that is, the circuit delivers at the receiving end one per cent. of the power delivered to it at the transmitting end.

Line Efficiency

A problem of great importance in electric transmission in all fields is that of obtaining the maximum transmission line efficiency practicable within economic limits. Certain comparisons between power transmission and telephone transmission would seem to be of interest.

The losses in a unit length of transmission circuit include both resistance losses and leakage losses, and may be represented by the formula $I^2r + V^2g$. If both r and g are constant with variations in voltage and current, it is easy to show that the maximum efficiency of transmission takes place when the voltage and current are so adjusted that these two parts of the line losses are equal in magnitude.

Actually in a well designed power transmission circuit without Corona losses V^2g is very small. Hence the solution of the problem of increasing line efficiency is to raise the transmission voltage, thus decreasing the transmission current. This is accomplished in the shorter power transmission lines by using step-up and step-down

transformers at the ends of the line, the line itself having a relatively small effect on the impedances thus obtained. For the longer lines, however, the line characteristics play a very important part in this process, and this is best illustrated by consideration of the telephone case.

In the ordinary telephone line the leakage losses are also small compared with the resistance losses, and the voltage can be raised by a factor of 2 or 3 before these losses become equal. The solution of the problem here as in the case of the power line is therefore to raise the transmission voltage and decrease the current, that is, to increase the impedance of the line. In a telephone case, however, this cannot be done by changing the impedance of the terminal apparatus since, as has been pointed out, practically all of the power is absorbed in line losses, and therefore the impedance at the transmitting end of the line is not appreciably affected by the impedance of the receiving apparatus. In order to raise the ratio of voltage to current on the telephone line, it is therefore necessary to operate on the line itself.

The impedance of an electrically long transmission circuit is very approximately equal to $\sqrt{L/C}$. In telephone lines the most practical way to increase this impedance is to increase the inductance. This may be done, as you know, either by uniformly distributed inductance or by lumped inductance, providing certain essential conditions are met, and the result is what is called a "loaded" telephone circuit. It is particularly to be noted that this is not a resonance phenomenon. On the contrary, loading in this way tends to decrease the variations with frequency of the efficiency of transmission of the circuit, and when so proportioned as to give maximum power efficiency, results in distortionless transmission.

In the very long power lines on the other hand, it is desired to transmit efficiently only one frequency, the fundamental. The use of methods depending upon resonance is therefore permissible, and in fact the method which will undoubtedly have increasing use in the future as with increasing length of power transmission lines the effects of line capacity become more important, will be the partial neutralization of the effects of this capacity by shunt inductances distributed along the line, that is, induction machines, or synchronous machines underexcited. This reduces the equivalent capacity of the lines by supplying at least a part of the charging current at the intermediate points. Thus is a similar end obtained in the two cases by different means.

Long Line Phenomena

The notable success of the vacuum tube amplifier has made a great change in the character of the problems encountered in the design of very long telephone circuits, and our discussion would not be complete without a brief consideration of the nature of these problems. A detailed discussion of these problems is given in numerous papers in recent technical literature.

With the amplifier it is possible in a very large measure to overcome at a relatively small cost the effects of power losses in the telephone circuit. Whereas before the general use of the amplifier it was necessary to make every possible effort to further improve the efficiency of the long circuits in order to increase the distance over which satisfactory telephone transmission could be given, with the amplifier of today there is a limit to the amount of money which can properly be spent merely for the improvement of the volume efficiency of the line as the line losses can always be made up if desirable by the use of amplifiers. As a result, in general, very long telephone circuits have become electrically so long that factors other than power efficiency determine the limits of their effectiveness. While a quantitative theoretical discussion of these problems is necessarily in large measure beyond the scope of undergraduate work at the present time, this may not long be the case, and in any event a general appreciation of these phenomena is of a good deal of interest.

Although these effects are common to all long circuits in principle, they are most prominent in very long telephone cable circuits as these are electrically the longest circuits in use. For example, the propagation constant of a toll circuit in cable between New York and Chicago is at 1,000 cycles approximately $50 + j300$. This is approximately the same as the propagation constant of a high voltage power line transmitting power at 60 cycles of 25,000 to 50,000 miles in length. If there were no intermediate amplification in the circuit the ratio of input to output power would be ten to the 45th power so that with our usual telephone input of about one milli-watt the circuit would deliver only one electron in each two months, and even if all the power available in New York City or Chicago could be used at the input without burning up the circuit, the received current would be utterly inappreciable. However, there are far more practical reasons than this for frequent intermediate amplification. The lower limit to which the power level can be permitted to fall in the circuit is limited by the disturbances picked up from other telephone circuits in the same cable or from other electric circuits outside the cable, and the maximum power level is, of course, limited by considerations of economy in the

design of the amplifiers. These limits result in the use of amplification in these circuits at approximately 50 mile intervals, there being 17 intermediate points of amplification between New York and Chicago on the shortest route.

With such a circuit the variation of efficiency with temperature is very rapid and the emergence of the sun from under the clouds could make as much as 1,000 fold-difference in the amount of power delivered. It is therefore necessary for practical operation to control the power gain introduced by the amplifiers by means of pilot wires in the cable subject to the same temperature variations as the talking circuit, and in this way to compensate automatically for the effect on the propagation constant of varying temperature.

In order to obtain for these cable circuits as far as possible the benefits of high voltage transmission, the circuits are loaded. This loading also in large measure equalizes the efficiency of transmission over the frequency range required for the transmission of speech.

The velocity of wave propagation over conductors loaded with inductance is, of course, relatively slow compared with the velocity of light. In the case of loaded telephone circuits in cable the velocity for the two types of circuit in general use is respectively 10,000 miles a second and 20,000 miles a second. The low velocity circuits are loaded with more inductance and are of higher efficiency and therefore preferable from the standpoint of volume. It is necessary, however, for the long circuits to use facilities of higher velocity and lower efficiency because of several very interesting phenomena.

For one thing, with circuits of high efficiency conforming to present day standards, the currents reflected from the distant end because of irregularity of impedance between the line and the terminal apparatus are by no means inappreciable. When the time for the propagation of these currents over the line and back to the transmitting end is very short, the reflected currents can be large without interference with service as they are indistinguishable from the sound directly heard by the speaker. If a circuit from New York to Chicago were used on the lower speed of the two types of circuit mentioned above, however, the reflected currents would arrive about one fifth of a second late. An interval of this magnitude would result in serious confusion to the speaker due to hearing his words twice, by direct transmission and after reflection from the distant end of the circuit. With the high speed facilities the time interval is reduced to one tenth of a second, and the interfering effect is very much smaller. Even with the high speed facilities, however, the effect is sufficient so that on circuits of over a few hundred miles in length special devices known as echo

suppressors are used to intercept the echo currents and prevent their interference with the speaker.

Another important effect is the imperfect equalization of the transmission of different frequencies within the range of important telephone frequencies. This imperfection can be offset by the use at intervals of correcting networks which introduce a distortion opposite to that produced by the line, and the design of such networks is one of the interesting problems which has been worked out in connection with these very long circuits. The distortion is not only one of magnitude but also one of phase due to the difference in the velocity of wave propagation of component currents of different frequencies. This may be extremely important on long circuits of the heavier type of loading in which any minor disturbance at the transmitting end of the line is transmitted in such a way that the low frequency components appear first at the receiving end, followed by progressively higher frequency components and causing disturbing transient noises somewhat similar to the chirrupping of a bird. Phase distortion is less on the more lightly loaded circuits but still remains of enough importance to require the use in some cases of networks to equalize the distortion of phase.

In this discussion of very long circuits I have talked of telephone circuits. In the case of both telephone and telegraph circuits the fundamental requirements are the same, namely, the propagation of currents within a certain range of frequencies without excessive distortion and without interference from other electric circuits. The principal difference in the problems is in the range of frequencies which is important in the two cases, that for telegraph being much lower and more limited in extent than that for telephone. Another difference is that in the case of telephony phase distortion is important only in producing different time of arrival of different components, whereas in telegraphy the effect of phase distortion in distorting wave shape is also of importance. The telegraph problem can, like the problem of transients in telephone lines, be approached theoretically from the performance of a circuit when a potential is applied suddenly at one end. It has been shown, however, that the treatment of the circuit in terms of its steady state characteristics for the propagation of alternating currents over a range of frequencies leads to results identical with those reached by the transient treatment, and for most cases the steady state method of treatment rather than the transient method of treatment is found to be more convenient to handle for purposes of circuit and apparatus design.

Electromagnetic Theory

There is a further gap which it is to be hoped can in the future be bridged for those students who are sufficiently advanced to become familiar with the general electromagnetic theory. To what extent it is practical to teach general electromagnetic theory to undergraduates is a question which is perhaps beyond the scope of this paper. However, those five differential equations as formulated by Maxwell and Lorentz which form the general mathematical statement of the fundamental discoveries of Ampere, Faraday and the other pioneers of electric science are the Magna Charta of electric science of today, and with the rapid development of the electrical arts in various directions constant recourse must be had to these fundamental equations for the establishment of correct electrical principles.

The simplified electric circuit theory which we have just discussed, which may be called the classic theory, serves very well for the great bulk of problems of electrical transmission of today. However, already there are situations both in the power transmission art and in the communication art in which the approximations which these equations involve are not valid, and for the solution of practical cases recourse must be had to the more general equations. These practical cases include the distribution of current in the earth when one side of a circuit is grounded, the inductive effects produced in other electrical circuits from a grounded circuit, and the transmission characteristics of submarine cables in which the sea water forms a part of the return path.

The classic circuit theory expresses the electrical quantities in terms of the total currents flowing in conductors and the voltages between these conductors, and expresses the aggregate energies in terms of inductances and capacities, and the dissipation in terms of resistances. The general equations of the electromagnetic theory express the electric quantities in terms of elementary current and charge densities, and the electric and magnetic fields are expressed in terms of field strengths. These, then, are the rigorous equations in differential form. In the classic theory the current and charge densities and field strengths are integrated into more easily manipulated totals. In other words, the electric circuit theory deals with macroscopic or large scale phenomena and the electromagnetic theory deals with microscopic or small scale phenomena.

What are the approximations involved in the classic theory and what conditions must be met for these to be good approximations? This matter has been treated in a very interesting way in some recent papers by Mr. John R. Carson. In brief, Mr. Carson's papers point

out that the classic circuit theory applied to transmission lines involves the following assumptions:

1. That the solution is concerned only with conditions at some distance from the terminals of the circuit and can therefore ignore the changes in distribution of electric and magnetic fields near the terminals known as end effects. That means that the electric and magnetic fields are propagated along the line in the same way as the currents and voltages.
2. That the propagation constant (per centimeter) is very small compared with unity and that the real part of the propagation constant is not large compared with the imaginary part, that is, the attenuation is not large compared with the phase change.
3. That in the conductors the loss due to the transmission current (that is, the axially flowing current) is large compared with the loss due to the charging currents.
4. That in the dielectric the propagation of energy is nearly parallel to the axis of the conductors and the dissipation in the dielectric is negligible.
5. That the fields of the currents and charges are propagated at an infinite velocity, that is, that radiation is neglected.

These assumptions, it can be shown, are very good for the ordinary case of an efficient transmission system. The effect of modification of the field at the terminals influences only a few feet of the line and is negligible in amount. Assumptions 2, 3 and 4 can very readily be shown to be true from the characteristics of the conductors and dielectrics involved in transmission, and as regards neglecting of radiation Mr. Carson has shown that for the ordinary transmission system the losses by radiation are in the order of one ten-thousandth of those in the conductors.

It is to be noted, however, that these assumptions place certain limitations on the application of the classic theory which are important in certain cases. The limitations are as follows:

1. The electric and magnetic fields are accurately expressed only for points relatively near the conductors. At great distances from the conductors the radiation field becomes important in comparison with the inductive field because of the much more rapid rate of decrease in intensity of the inductive field with distance.
2. The electric and magnetic fields are not accurately expressed near the terminals of the circuits.

3. The approximations do not apply if the conductors are quite imperfect or if the dielectric is highly dissipative. This affects the application in certain practical cases as indicated above.
4. The classic theory does not apply to circuits of the usual dimensions for extremely high frequencies in the order of millions of cycles.

From the standpoint of the best appreciation of the fundamentals of the electrical arts it might be said that the ideal course for students in electrical theory would start with the fundamental discoveries of Ohm, Oersted, Ampere, Faraday and Henry, pass through the generalized mathematical statement of the laws which they discovered to the general electromagnetic equations of Maxwell and Lorentz, and then from this focal point derive the various approximations of electric theory as applied to the various electrical arts; electric light and power, electric transportation, telephone, telegraph and radio. Unfortunately, teachers and students as well as the rest of us are hampered by questions of time and this approach is not proposed as a practical undergraduate course at the present time.

Conclusion

Electrical science and its practical applications are undergoing a very rapid development and expansion. The science of today is the engineering of tomorrow. These facts result in increasing importance in the mastery by the engineering student during his time at college of the electrical principles of broadest general usefulness, rather than learning specific applications of these principles. By a mastery is meant such an appreciation of the scope and limitations of the principles that he is able to apply them correctly to new conditions as they come up.

It is not intended to express a judgment on the extent to which the curricula of engineering schools should go in presenting electrical theory, and it is, of course, recognized that different schools have different conditions to meet which will naturally result in somewhat different courses. It is not proposed that the specific forms of electrical theory applying directly to telephone problems should be taught.

It is proposed for consideration, however, that whatever the scope of general electrical principles which is taught, these be so presented that the student have a clear picture of what they mean and of how and where they apply, and that he also should appreciate the relation to the general principles of any specific cases presented and the approximations which they involve. As far as practicable, all principles should be related to the general electromagnetic theory, the fundamental basis of all our electrical science.

Magnetic Properties of Perminvar [1]

By G. W. ELMEN

Synopsis: This paper describes the magnetic properties of a group of iron-nickel-cobalt alloys, named "perminvar." With certain heat treatments these alloys have unusual constancy of permeability and extremely small hysteresis losses at low flux densities, and peculiarly shaped hysteresis loops constricted in the middle as the maximum flux densities of the loops are increased. Methods of preparing and heat treating the alloys are described, limits of composition, and changes in the magnetic properties with composition and with different heat treatments are illustrated. A theory of constitutional changes effected by heat treatment and responsible for the unusual magnetic properties is suggested.

IN 1921 the writer was investigating the magnetic properties of a series of permalloys [2] to which a few per cent of a third metal was added to the nickel and iron. One of these alloys contained cobalt. Magnetic measurements indicated that up to moderate field strengths the permeability of this nickel-cobalt-iron alloy was remarkably constant. The constancy was materially better than for soft iron, notwithstanding the fact that the initial permeability was several times higher. This was unusual, as small permeability variation ordinarily is found only in materials with low permeability. Measurements of other magnetic properties were equally surprising. When the hysteresis loop was traced for a cycle which carried the flux density up to a few thousand gauss, it was found to have an extraordinary form in that it was sharply constricted in the middle. These and other differences which were observed indicated that this alloy was a new type of magnetic material in which the magnetic properties were fundamentally different from those of previously known materials.

This discovery aroused a great deal of interest for it was recognized that magnetic materials possessing these properties were of great scientific and technical importance. In order to develop the possibilities which this alloy suggested, an exploration of the whole field of the iron-nickel-cobalt series was undertaken. For it was, of course, apparent that the alloy which had aroused our interest must be one of a group of compositions which possessed similar properties in a greater or less degree. In this survey, alloys varying in 10 per cent steps in composition and including the whole range of the ternary

[1] Reprinted from *The Journal of the Franklin Institute*, Vol. 206, No. 3, September, 1928.

[2] Arnold & Elmen, *Jour. of Frank. Inst.*, May, 1923, p. 621.

series of these metals were made up and their magnetic properties measured.

These measurements showed the range of compositions which shared in such unusual magnetic properties, and indicated that heat treatment was an important factor in the development of these properties. A large number of alloys have been made up in this range, for which the variations in composition were evenly distributed but much smaller than for the initial survey. From these alloys a few were selected which appeared to be specially suited for magnetic uses in electrical communication circuits. Our experience with these alloys has been that when good grades of commercial materials are used, the castings are readily reduced mechanically to the desired dimensions, and the magnetic properties from different castings of the same composition are quite uniform.

We felt that these alloys were so unique as regards magnetic quality that they should be grouped in a class under a common name which should readily distinguish them from other materials. We have chosen "perminvar" as the name for alloys in the iron-cobalt-nickel series, which are characterized, when properly heat treated, by constancy of permeability for a considerable range of the lower part of the magnetization curve, by small hysteresis loss throughout the same range of flux densities, and by a hysteresis loop constricted at the origin for medium flux densities.

This paper describes the magnetic properties of the perminvar group of alloys. Results are given for several alloys selected to show the variation in magnetic properties when the proportions of the constituent metals are varied over a wide range. Detailed measurements under a variety of magnetic conditions and heat treatments are recorded for the composition 45 per cent nickel, 25 per cent cobalt and 30 per cent iron. This composition is a typical one and was chosen early in our experimental work as specially suitable for commercial uses, for it had, in addition to the unusual properties in which we were most interested, a fairly high initial permeability.

Preparation of Alloys

The alloys were cast from the best available commercial materials. Armco iron, electrolytic nickel and commercial cobalt were melted together in the desired proportions in a silica crucible in a high frequency induction furnace. Before pouring, one half of one per cent of metallic manganese was added to the molten metal. Part of this manganese deoxidized the metal and went into the slag, and the remainder, usually about one half of the added amount, remained in

the alloy. The alloys also contained small amounts of carbon (less than .03 per cent), silicon (less than .1 per cent) and traces of sulphur and phosphorus. The alloys were cast into bars 18 in. long and 3/4 in. in diameter. The bars were rolled or swedged into 1/4 in. rods and drawn from that size to .062 in. diameter wire. This wire was flattened and trimmed into tape 1/8 in. × .006 in. The material was annealed several times in the reduction process, for the cold working hardened the alloys rapidly and made them difficult to work.

To prepare the tape for heat treatment and subsequent magnetic measurements, about 30 ft. of it was wound spirally into a ring of 3 in. inside diameter, the ends being spot welded to the adjacent turns. Care was taken to wind the rings loosely to prevent the turns of tape from sticking during annealing.

A number of such rings were packed in a nichrome pot. Some iron dust was usually placed in the pot to take up the oxygen and thus prevent the oxidation of the rings. Further protection was secured by luting the joint between the pot and its cover. The pot was placed in an electrical resistance furnace, the temperature of the furnace raised to 1000° C. and held at that temperature for one hour. The current was then turned off and the pot cooled with the furnace. Ten hours were required for the furnace to cool to the temperature of the room. Between 700° C. and 400° C. the rate of cooling was approximately 1.5° per minute.

Three rings of each composition were always annealed together. One of these rings received no further heat treatment. The second ring was placed for 15 minutes in a furnace held at 600° C., then removed and cooled rapidly on a copper plate. In some cases, the third ring was heated 24 hours at 425° C.

In the discussions and in the figures and tables, the rings which received the first heat treatment only are referred to as "annealed," those reheated to 600° C. and rapidly cooled as "air quenched," and those held for a long time at 425° C. as "baked."

Magnetic Measurements

Permeabilities at low magnetizing forces were measured on unwound rings with an inductance bridge, and an a.c. permeameter.[3] From these measurements initial permeabilities were computed. For elevated temperature, measurements were made with a similar permeameter provided with a furnace compartment.[4] The bridge was also used for measuring permeabilities to small a.c. magnetizing forces when d.c. forces are superposed on the magnetic circuit. For these

[3] G. A. Kelsall, *J. O. S. A.* and *R. S. I.*, **8**, pp. 329–338, 1924.
[4] G. A. Kelsall, *J. O. S. A.* and *R. S. I.*, **8**, pp. 669–674, 1924.

measurements the rings were wound with insulated copper wire after being placed in thin annular wooden boxes to protect them from strain. Hysteresis losses were computed for a few alloys from effective resistance measurements at low flux densities, made on wound rings.

Magnetization and permeability curves and hysteresis loops were plotted from ballistic galvanometer measurements. Galvanometer measurements also were made on a few alloy rods 11 in. long and 1/8 in. diameter, at a magnetizing force of 1500 gauss. For these measurements the rods were placed in a long solenoid and the induction measured by means of an exploring coil at the center of the rod.

Properties of the 45 Per Cent Ni, 25 Per Cent Co, 30 Per Cent Fe, Composition

Measurements for the composition 45 per cent nickel, 25 per cent cobalt and 30 per cent iron in the annealed condition are plotted in Figs. 1–7 and tabulated in Table I. The curves in Fig. 1 illustrate the permeability characteristics for this composition (No. 858–1) and for a sample of annealed Armco iron. For magnetizing forces below 1.7 gauss, the permeability is substantially constant, the variation being less than 1 per cent. This constancy is remarkable for a magnetic material having an initial permeability nearly double that of iron. Within the same range of field strengths the permeability of the iron sample rises from an initial value of 250 through a maximum of 7,000 at a magnetizing force of 1.3 gauss and decreases to 6,300.

TABLE I.

Hysteresis Loss with Different Heat Treatments for [45% Ni—25% Co—30% Fe] Composition Perminvar

Heat Treatment	B	Ergs per Cm.^{3} per Cycle
Air Quenched	568	18.7
	722	32
	993	57
	1,503	119
	5,010	850
	14,810	2,500
Baked at 425° C. for 24 Hours	600	0
	795	0
	1,003	15.27
	1,604	163
	4,950	1,736
	13,810	4,430
Annealed	570	0
	820	9.54
	974	15.65
	1,508	93.20
	5,050	1,185
	8,480	2,500
	14,900	3,375

Another property of the perminvar alloy closely related to the constancy of permeability is the extremely small hysteresis loss in the range of magnetizing forces and flux densities in which the permeability is constant. This is illustrated in Fig. 2 where line *a* represents the plot of the upper half of the hysteresis loop for a maximum flux

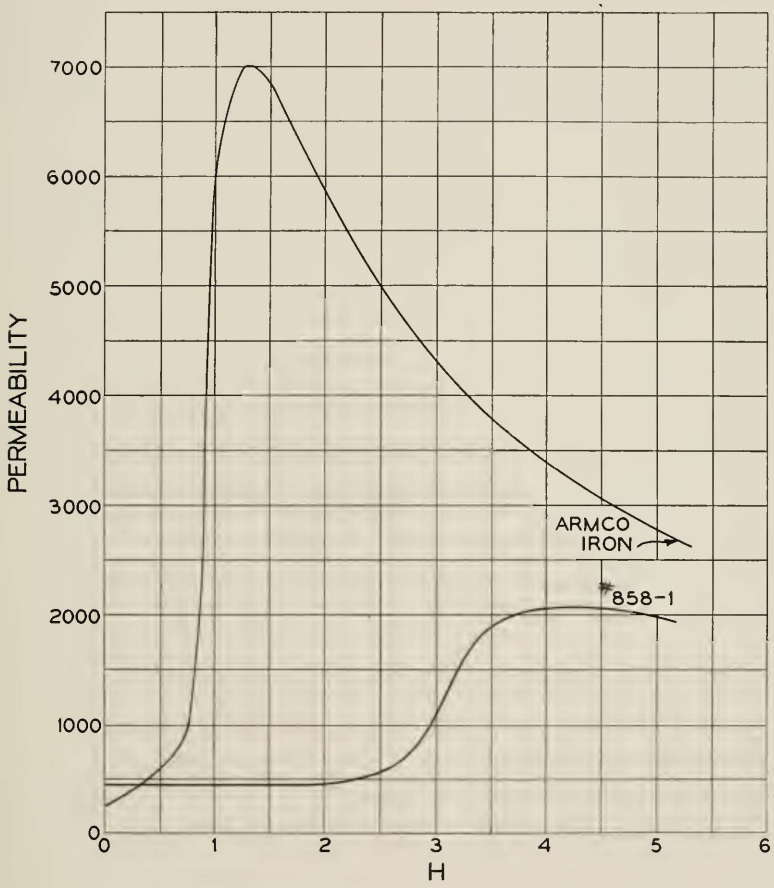


FIG. 1—Permeability curves for Armco iron and Perminvar (45% Ni—25% Co—30% Fe)

density of approximately 600 gauss. Curves *b* and *c* are similar plots for silicon steel ($3\frac{1}{2}$ per cent silicon) and Armco iron respectively. While the hysteresis loops for Armco iron and silicon steel have considerable areas amounting to 33 and 14 ergs per cubic centimeter per cycle respectively, there is no measurable area for the perminvar alloy. Although the ballistic method of measurements which was

used in obtaining these curves, does not indicate very small losses readily it is evident that the losses in the perminvar alloy are of a different order of magnitude from those of the other two materials. In order to obtain additional information in regard to the hysteresis loss of this alloy at low flux densities, the sample was measured by the inductance bridge method. It was found that the hysteresis loss at a flux density of 100 gauss was .024 $\times$ 10^{-3} ergs per cubic centimeter per cycle. The best material in this regard previously known was permalloy, for which a sample containing approximately $78\frac{1}{2}$ per cent nickel, measured under similar conditions, had a hysteresis loss of 33 $\times$ 10^{-3} ergs per cubic centimeter per cycle.

The growth of the hysteresis loss and the appearance of measurable areas, and the peculiar shapes of the loops for this composition as the flux densities increase are illustrated in Fig. 3. The curve for a

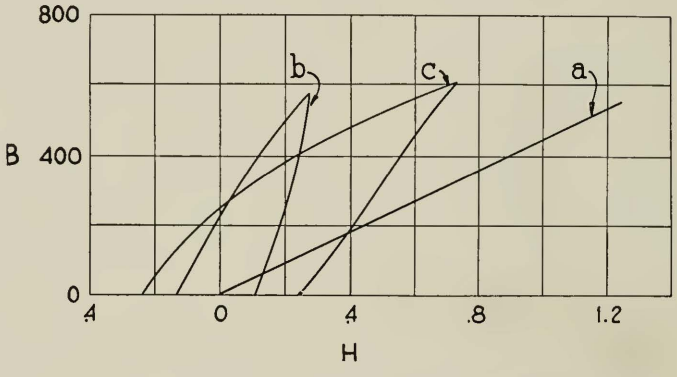


FIG. 2—Hysteresis characteristics: *a*—Perminvar (45% Ni—25% Co—30% Fe); *b*—silicon steel; *c*—Armco iron

maximum flux density of 580 gauss in Fig. 3, is from the same data as Curve *a* in Fig. 2. The circles in this plot indicate points on the ascending branch, and the dots, points on the descending branch. The hysteresis loop broadens out so that it has a measurable area when the maximum flux density is increased to 800 gauss. The existence of a close relation between the hysteresis losses and the constancy of permeability is quite apparent from the permeability curve in Fig. 1 and the curves in Fig. 3. While the permeability remains constant there is practically no hysteresis loss but as it begins to change this loss appears and increases quite rapidly with increase in permeability. The increase in the energy loss and the changes in the shapes of the loops as the flux density increases also are illustrated by these curves. The most striking hysteresis characteristic of these

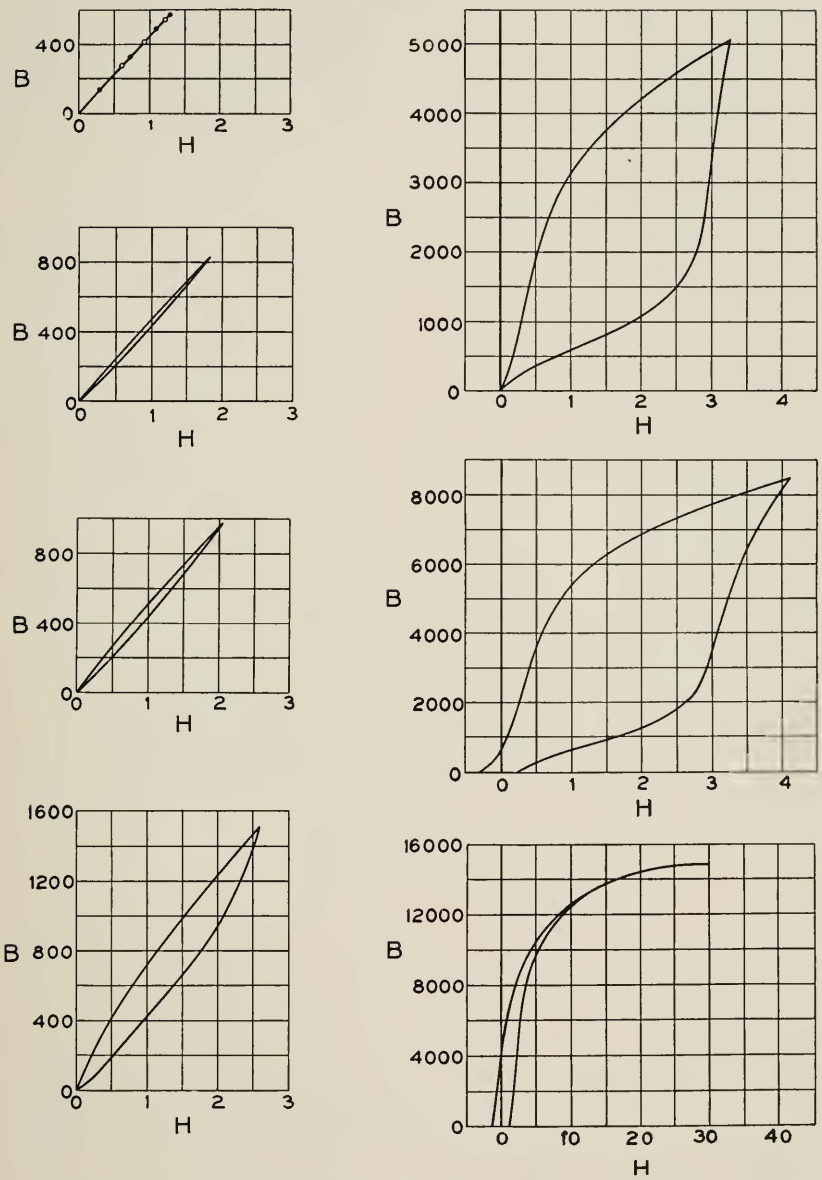


FIG. 3—Upper halves of hysteresis loops of Perminvar (45% Ni—25% Co—30% Fe) annealed

loops is the absence of coercivity. For loops having maximum flux densities of 5,000 or less the ascending and descending branches pass through the origin. For greater flux densities the coercivity begins to be measurable, but there is still a considerable constriction of the loop for a maximum flux density of 8,000 gauss. It is only in the loop for 15,000 gauss, that the constriction at the origin has disappeared and the loop resembles those for ordinary magnetic materials. In Table I the hysteresis losses for the complete loops are tabulated.

Fig. 4 illustrates graphically how the permeability measured with a constant alternating current magnetizing force of about .0021 gauss and 200 cycles per second is affected when a steady magnetizing force

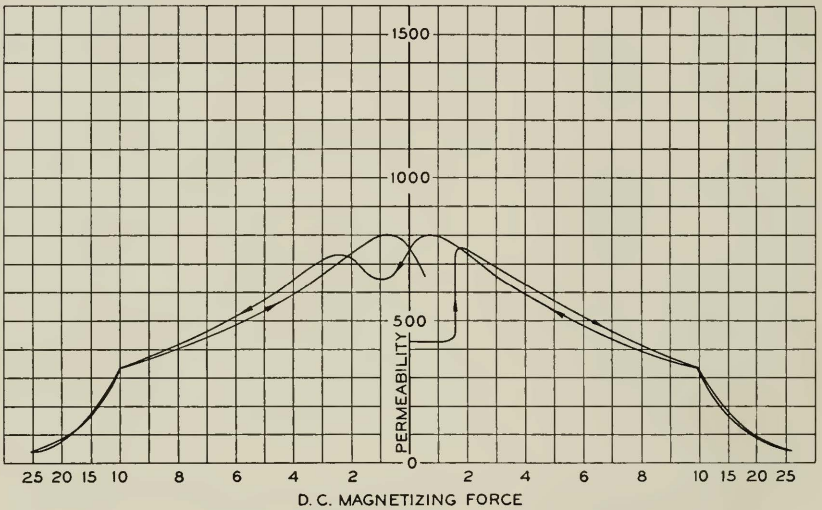


FIG. 4—The effect of superposed d.c. fields on the a.c. permeability of Perminvars (45% Ni—25% Co—30% Fe)

is superposed on the magnetic circuit, the steady force being produced by a direct current. The arrows in the figure indicate the direction in the progress of the permeability as the direct current magnetizing force is varied. The permeability is substantially constant as the direct current magnetizing force increases up to approximately 1.7 gauss and it then suddenly rises as the force is increased beyond that value. This is the same field strength at which the permeability begins to increase as shown in Fig. 1.

Another characteristic of this material not found in ordinary magnetic substances also is shown in Fig. 4. After an applied d.c. magnetizing force of 25 gauss is removed, the permeability has risen from 460 to 750. With ordinary materials, after such magnetization,

the permeability is reduced, in some cases from 40 to 60 per cent. With the increase in permeability its constancy disappears both for the superposed condition as shown in Fig. 4, and for ordinary magnetizations at low field strengths. The hysteresis losses are also increased for corresponding flux densities. These changes in the

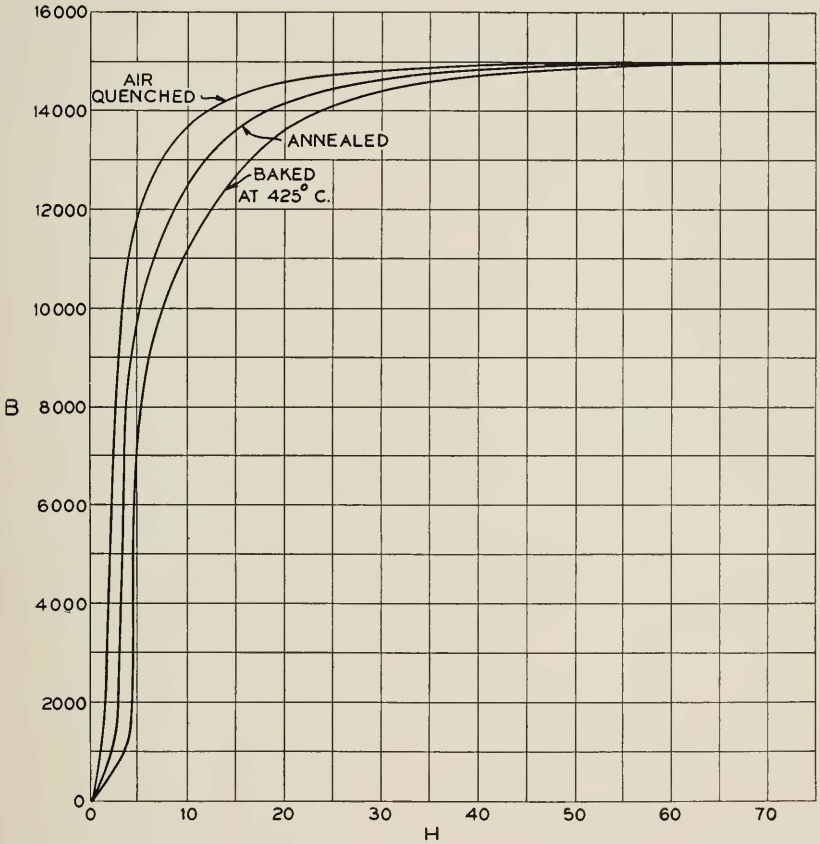


FIG. 5—Magnetization curves for Perminvar (45% Ni—25% Co—30% Fe)

magnetic properties are largely removed by demagnetization. The ordinary method of demagnetization by reversals of a slowly decreasing magnetizing force has been less successful than it is with iron in returning the material to its initial state. Addition of an a.c. force superposed on the d.c. helps materially in restoring the original magnetic properties.

Effects of Heat Treatment

The manner in which the magnetic properties of this composition are affected by the rate of cooling is illustrated in Figs. 5–7. The measurements plotted in these figures are from three rings, air quenched, annealed and baked, respectively. For weak fields there are large differences in the magnetic properties for these rings. The initial permeability for the quenched ring is more than twice that of the baked one. With increased field strength this difference decreases and disappears for fields over 50 gauss. The permeability variations as the strength of the field increases also show the remarkable change

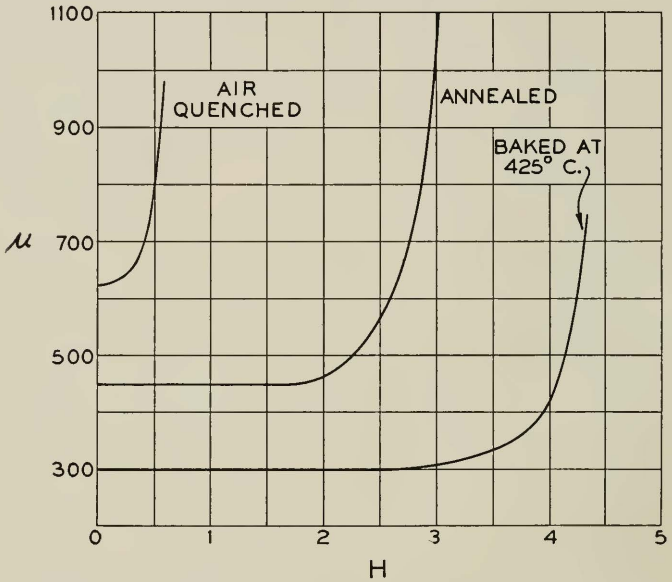


Fig. 6—Permeability curves for Perminvar (45% Ni—25% Co—30% Fe)

which heating for a long time in the critical temperature range produced in these alloys. For the quenched ring the permeability increased from 620 to 800 for an increase of field strength from 0 to .5 gauss. For the baked ring the permeability remains constant for fields up to 2.5 gauss.

The hysteresis loss and the shapes of the loops also are affected greatly by the heat treatment. This is illustrated in Fig. 7, where loops for a number of flux densities are plotted for two sample rings, one baked at 425° C. and the other air quenched, and in Fig. 3 where loops for the same maximum flux densities are plotted for an annealed

ring. The energy losses integrated for complete loops are tabulated in Table I.

These curves show that the rate of cooling determines the magni-

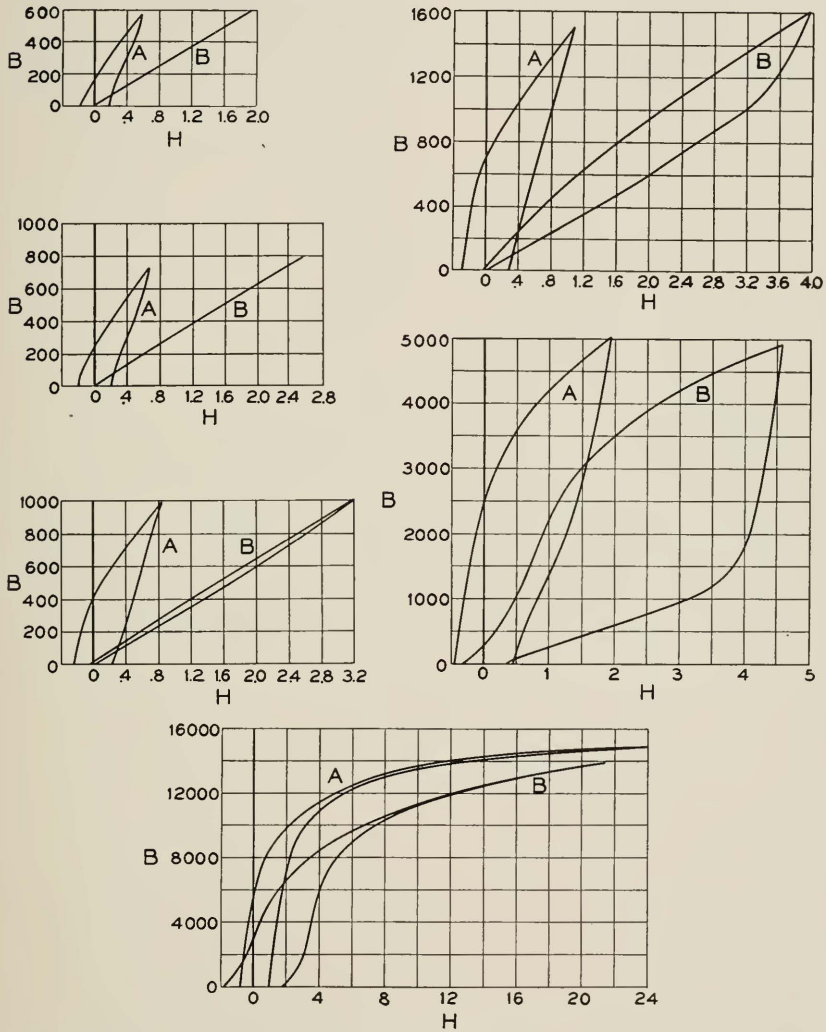


FIG. 7—Upper halves of hysteresis loops for Perminvar (45% Ni—25% Co—30% Fe) *A*—air quenched, *B*—baked at 425° C.

tudes of the hysteresis losses and the shapes of the hysteresis loops. For the air quenched rings the shapes of the loops for the different flux densities resemble those for ordinary magnetic materials although a few of them show traces of the perminvar characteristics. If a

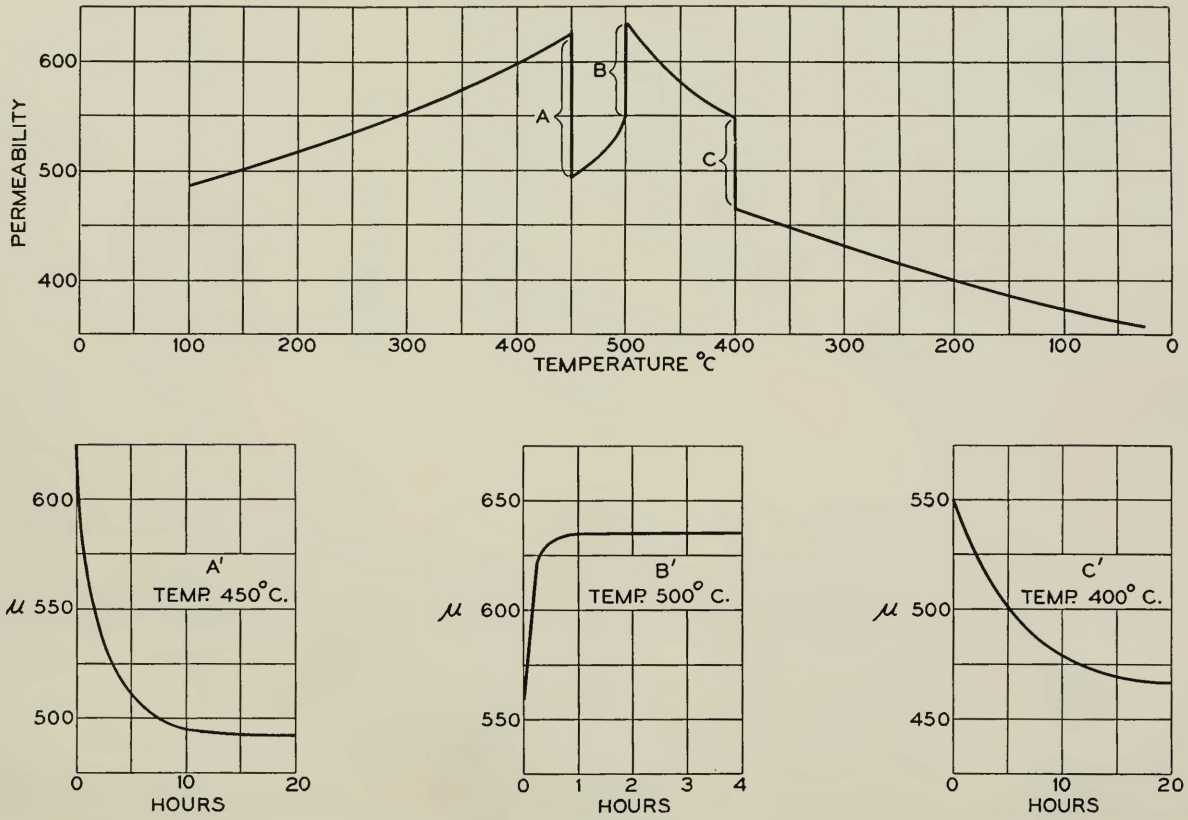


FIG. 8—Permeability—Temperature curve for Perminvar (45% Ni—25% Co—30% Fe)

more rapid cooling rate had been used these characteristics no doubt would have disappeared completely. The hysteresis loops all have considerable areas. The one for 568 gauss, the lowest flux density measured, represents an energy loss of 18.7 ergs per centimeter cube. For the same flux density, the hysteresis loops for the annealed and the baked rings have no measurable areas, the ascending and descending branches of the loops falling on the straight lines shown in

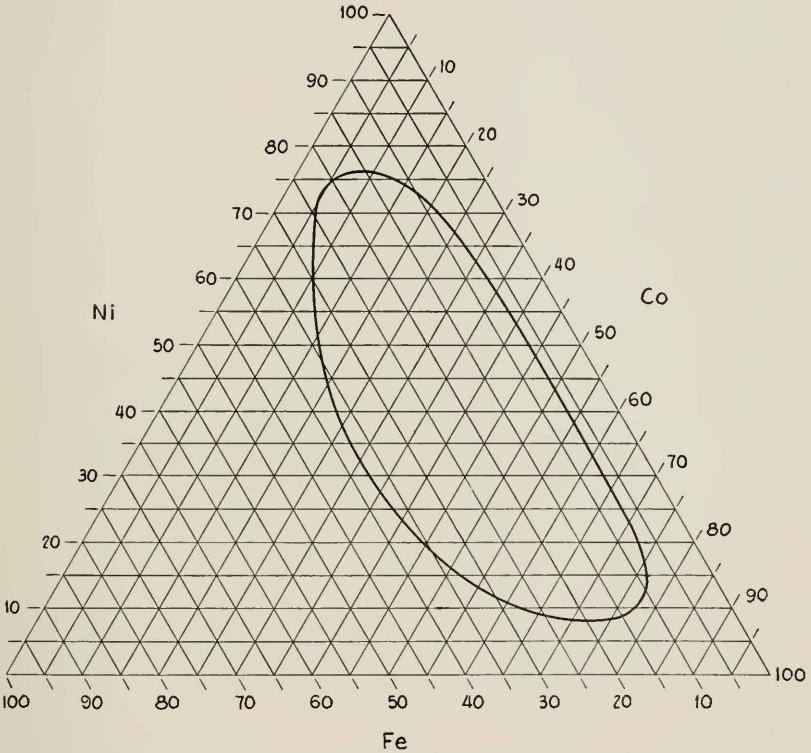


FIG. 9—Composition diagram Ni—Fe—Co Series. Area enclosed by the curve shows compositions with marked Perminvar characteristics

the figure. This absence of measurable area extends up to a flux density of nearly 1,000 gauss for the baked ring. The increase of energy loss as the flux density increases above this value, however, is considerably more rapid than for the quenched ring. At 1,500 gauss, the hysteresis loss is a little greater than for the quenched ring and when the flux density is increased to 5,000 gauss, the loss is more than double.

It was shown above that the degree to which perminvar charac-

teristics are developed depends on the rate of cooling through the critical temperature range, and that baking at 425° C. gave the most characteristic results. The manner in which the permeability of the 45 per cent nickel, 25 per cent cobalt and 30 per cent iron composition changes in this temperature range is illustrated in Fig. 8. The temperature of an annealed ring was increased from that of the room to 450° C. where it was held constant for twenty hours. It was then raised to 500° C. and held for four hours, then lowered to 400° C. where it was held for twenty hours, and finally cooled to room temperature. Permeability measurements were made at these temperatures with an a.c. magnetizing force of .02 gauss.

Inspection of these curves shows that in the range 400°–500° C., the permeability lags behind the temperature, and that the time required for the permeability to reach a constant value increases very rapidly below 450° C. The changes in final permeabilities with temperature decrease also rapidly below 450° C. In fact, when the difference in permeability caused by the temperature coefficient is corrected for, the permeability of the alloy after heating at 400° C. is not very different from what it is after heating at 450° C. Other experiments show that the critical temperature range extends below 400° C., but, as would be expected, the decrease in permeability is very small. The range also extends above 500° C. for this alloy and some experiments indicate that the upper limit is the magnetic transformation temperature which for this alloy is 725° C.

Effects of Variation of Compositions

The composition range within which the magnetic properties characteristic of perminvar are developed pronouncedly by annealing, is represented by the area enclosed by the curve in the triangular composition diagram Fig. 9. Magnetic properties for a few of the compositions in this area are plotted in Figs. 10 and 11. Table 2 gives their chemical analyses, initial permeability (μ_0), the maximum permeability ($\mu_{\max.}$), the magnetizing forces (H) and the flux densities (B) in gauss to which the alloys may be brought with a permeability variation not over 1 per cent, also the ($B - H$) values for a magnetizing force of 1,500 gauss for some of the alloys, and the resistivity in microhms-cm. The hysteresis losses for a number of flux densities are given in Table 3.

The area enclosed in Fig. 9 shows that approximately one third of the alloys in the Ni-Fe-Co series show some of the characteristic perminvar properties in the annealed condition. The proportions of nickel and cobalt may be varied through a wide range. A great deal

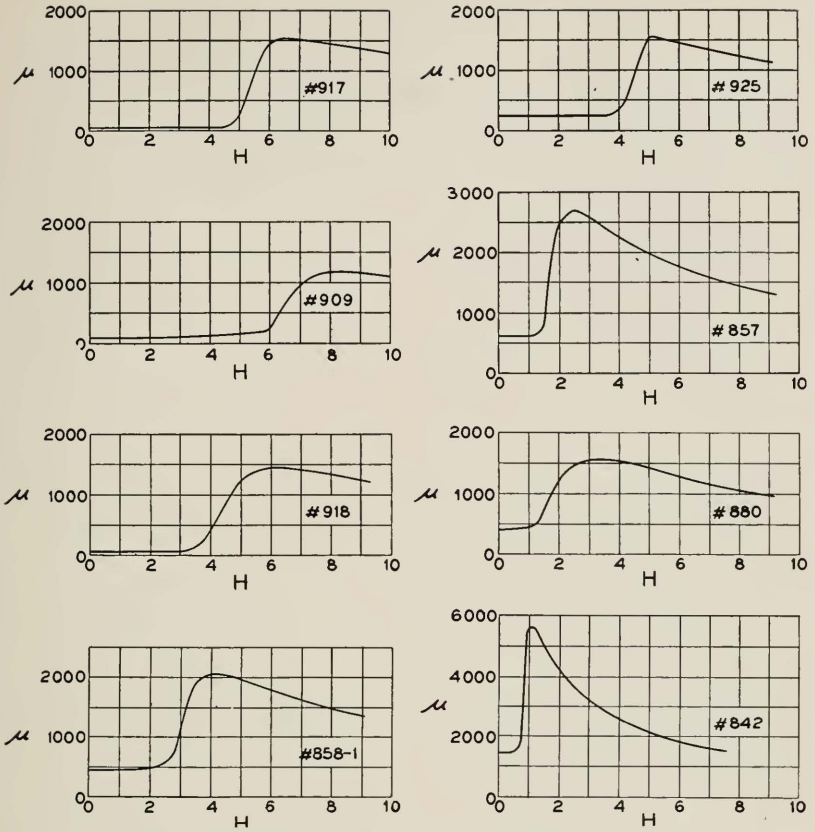


FIG. 10—Permeability curves for several compositions of Perminvar. Chemical analysis given in Table II

TABLE II

CHEMICAL COMPOSITION AND MAGNETIC PROPERTIES OF PERMINVARS

Casting Number	Chem. Analysis				Magnetic Properties					Resistivity
	Ni	Co	Fe	Mn	μ_0	μ_m	H for 1% Ch. in μ	B for 1% Ch. in μ	$B - H$ for $H = 1{,}500$	Microhm-Cm.
917	11.35	68.10	20.36	.35	57	1,545	4.2	242	18,400	15.38
909	20.85	49.18	29.74	.31	98	1,180	4.0	396	18,200	16.59
918	20.73	68.35	10.58	.39	51	1,447	2.5	129	17,400	12.35
858–1	45.12	23.83	30.69	.46	449	2,075	1.75	793	15,600	18.63
925	50.47	29.28	20.15	.33	231	1,555	3.2	746	14,600	14.55
857	59.66	14.76	24.97	.60	631	2,680	1.15	733		17.5
880	70.29	15.23	14.57	Tr.	390	1,570	.55	216		14.13
842	73.29	5.97	20.7	.20	1,430	5,600	.55	795		15.56

less variation in the iron content is permissible, being less than one half of the amounts of the other two constituents. The manner in which each of the metals affects the magnetic properties is not very clearly indicated by the numerical values in the table. Iron and

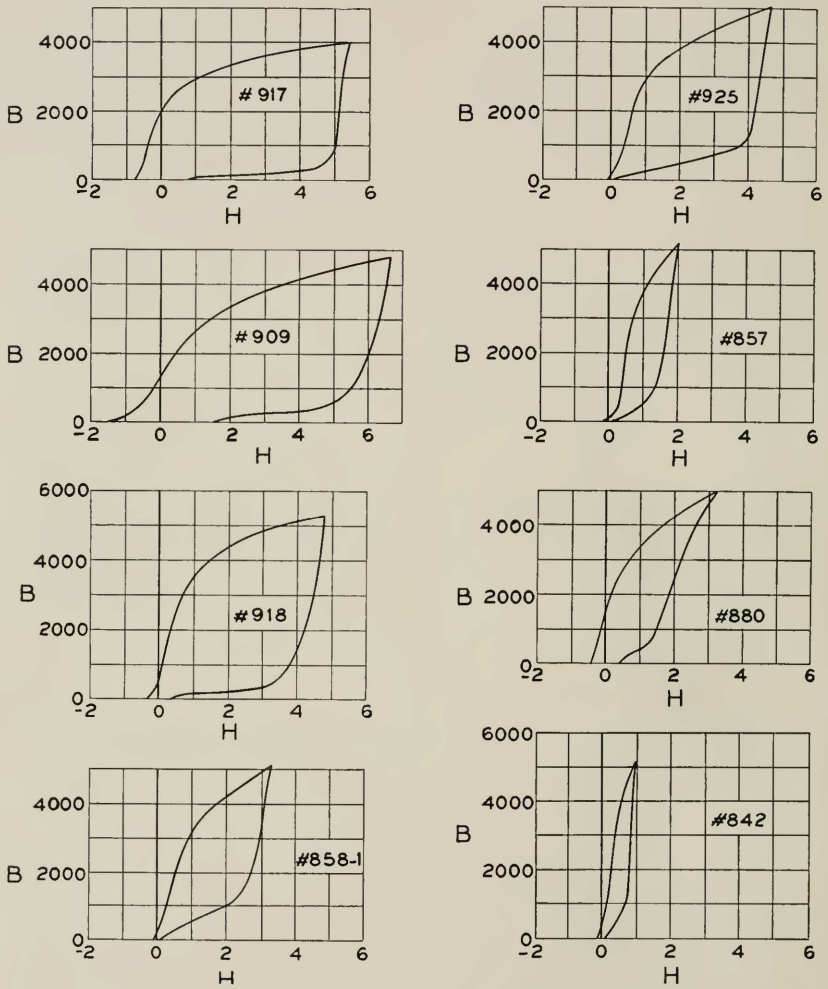


FIG. 11—Upper halves of hysteresis loops for several Perminvar compositions. Chemical analysis is given in Table II

cobalt appear to increase the constancy of the permeability but decrease the initial permeability values. Nickel increases the initial permeability but large percentages decrease the constancy. On the whole, in the alloys with high nickel content, the combination of high

TABLE III

HYSTERESIS LOSS

Casting Number	B	Ergs per Cm.^{3} per Cycle	B	Ergs per Cm.^{3} per Cycle	B	Ergs per Cm.^{3} per Cycle	B	Ergs per Cm.^{3} per Cycle	B	Ergs per Cm.^{3} per Cycle
917	280	8	625	150	1,700	1,040	3,950	2,740	15,500	14,160
909	—	—	620	36	1,030	299	4,800	3,330	15,200	12,460
918	—	—	750	247	—	—	5,270	2,605	—	—
858–1	566	0	827	9.5	1,508	93	5,050 8,480	1,185 2,500	14,900	3,375
925	560	0	—	—	2,000	468	5,000	2,020	13,200	4,965
857	145	0	840	8.4	1,520	88	5,200 8,250	632 1,240	13,250	1,508
880	320	0	—	—	1,470	116	5,000	1,012	10,300	2,435
842	500	0	—	—	1,530	23	5,100	348	11,500	783

permeability and fair constancy makes for a larger range of flux densities in which the permeability is constant and consequently also increases the range of flux densities with low hysteresis loss.

Experiments on several alloys of this series indicated that by baking the alloys at 425° C. the area enclosed by the curve in Fig. 8 would be increased considerably, and possibly would include some of the binaries of these metals.

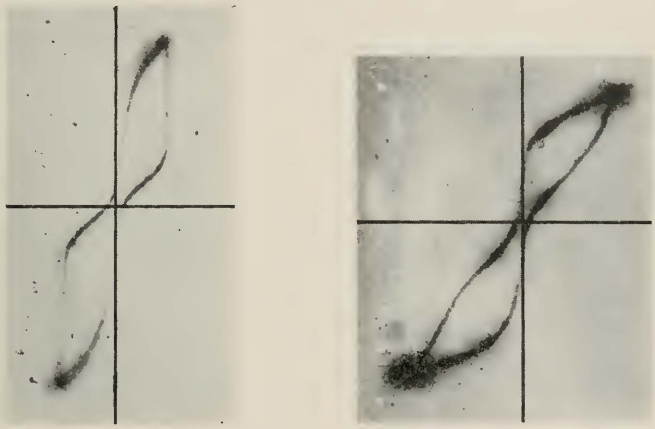
Discussion

While this paper is concerned primarily with the study of the magnetic properties of these alloys and the dependence of these properties on composition and on heat treatment, some of the results are of considerable theoretical interest, as they suggest the manner in which the unusual magnetic properties are acquired by the alloys. It was shown in the heat treating experiments that the unusual magnetic properties resulted from suitable heat treatment of certain compositions. Slow cooling through a rather narrow temperature range, or continuous heating for a long time at the lower end of this range resulted in alloys which had marked perminvar characteristics. Rapid cooling through this temperature range usually did not develop these characteristics. From the measurements at elevated temperatures, Fig. 8, it was shown that in the temperature range from 400° C. to 500° C., the change in the alloys is quite rapid at the higher temperature, but that the rate of stabilization slows up as the temperature decreases. When the alloy is heated and cooled through a temperature cycle in this manner, the permeability changes progressively and at each temperature in the cycle the alloy reaches a stable condition if the rate of cooling or heating is slow enough. There is a striking similarity in the manner in which these changes in permeability are developed, and in the progress of the constitutional changes in an alloy which at high temperatures is a homogeneous solid solution, but as the temperature falls becomes saturated and segregates into a mixture of two solid solutions of different concentration.

That such a segregation takes place in the slowly cooled alloys is also supported by a study of the differences in the shapes of the hysteresis loops of the quenched and the slowly cooled alloys. Ordinarily, the widest part of a hysteresis loop of a homogeneous material is the intercept on the H axis. All the loops of the air-quenched alloys have these characteristics. Gumlich [5] has shown that if a magnetic circuit is made of two materials of different magnetic properties, the loops may assume a variety of shapes, ranging from

[5] E. Gumlich, *Arch. f. Elektrotechnik*, Vol. 9, p. 153, 1920.

that of the homogeneous material to one in which two branches converge at the origin into a single line. This constriction of the hysteresis loop is also illustrated for a parallel bimetallic magnetic circuit in Fig. 12 where loops *a* and *b* are traced for a perminvar core and a bi-metallic rod, respectively. The rod was 15 in. long and consisted of a core of .04 in. diameter unannealed piano wire and a .006 in. wall permalloy tube, heat treated to give high permeability, and fitting closely to the wire. Though the magnetic circuit condition for the perminvar core is not the same as for the bi-metallic rod, the similarity of the two loops is marked and supports the theory that the constricted loop of the perminvar core is caused by segregation in the alloy.



a *b*

FIG. 12—Hysteresis loops: *a*, Perminvar; *b*, Bi-metallic rod. Loops traced with a cathode ray oscillograph

The electrical resistance also is affected by the slow cooling. An air-quenched alloy of the 45 per cent nickel, 25 per cent cobalt and 30 per cent iron composition was 10 per cent lower in resistivity after it had been baked at 425° C. This change is also in line with the idea that segregation takes place when the perminvar alloys are cooled slowly.

While these considerations point to a satisfactory explanation for the constriction of the hysteresis loops they do not explain the extremely low hysteresis losses of the alloys at low flux densities. This characteristic of perminvar suggests that one of the constituents which is segregated by the heat treatment is itself a material of much lower hysteresis loss than any previously known material and that

the other constituent suffers relatively little change of magnetization at low magnetizing forces.

To the engineer these alloys are of unusual interest. They may be used to advantage for magnetic structures where the magnetizing forces do not exceed the limits of constancy of permeability for the various compositions. Interesting results have been obtained with the 45 per cent nickel, 25 per cent cobalt and 30 per cent iron composition for continuous loading of telephone conductors and for cores of loading and filter coils used in high quality transmission and in carrier current circuits. For such purposes high resistivity is also desired, and it has been found that the addition of a few per cent of other metals such as molybdenum serves for this purpose. For circuits requiring greater constancy or higher permeability other compositions are more suitable. The best alloy for any specific circumstance may be selected from a study of the magnetic properties of the various compositions.

The Aluminum Electrolytic Condenser [1]

By H. O. SIEGMUND

SYNOPSIS: In this paper the anodic film-forming properties of aluminum are discussed and the unique electrical qualities of film-coated aluminum anodes are described. Special reference is made to an aluminum electrolytic condenser of the type used in low pass electric wave-filters of direct-current telephone power plant equipment. Electrical characteristics of condensers are given and the manner is described in which the operation and life of the units are influenced by variations in composition of the electrodes and the electrolyte.

INTRODUCTION

SINCE the discovery about 75 years ago [2] of the unusual polarizing effect of aluminum it has become well known that certain metals, notably aluminum and tantalum, as anodes in a suitable electrolyte become coated with a film having remarkable electrical properties. Films formed in this manner are characterized by the influence of impressed potential on their electrical resistance.

A representative relationship between applied voltage and resistance per 1,000 sq. cm. of film on an aluminum anode in an ammonium borate electrolyte [3] is shown in Fig. 1. This resistance characteristic imparts to the film the capability of conducting current more freely in one direction than in the other; of breaking down as an insulation between the metallic electrode and the solution when voltages above a critical value are applied; and in combination with the thinness of the film of holding a substantial charge of electricity at potentials below the breakdown voltage.

Each of these characteristics provides the principle around which a distinctive class of electrical apparatus has been developed. The electrolytic rectifier, widely used in small direct-current supply sets for battery charging and radio purposes, employs the unidirectional conducting characteristic. The aluminum electrolytic lightning arrester, used extensively for protection of direct-current railway equipment, depends for its operation upon the breakdown characteristic of the film. And finally the aluminum electrolytic condenser, now being

[1] Presented before the American Electrochemical Society, at Bridgeport, Conn., April 26, 1928.

[2] Wheatstone, *Phil. Mag.*, **10**, 143 (1855).

[3] The exact values of resistance are somewhat unstable, and depend on the time between readings and whether successively increasing or decreasing values of potential are applied. However, the general shape of the curve and the magnitude of the values are representative.

used in direct-current telephone power plant equipment [4] utilizes the dielectric property of the film to provide electrostatic capacity.

Electrical Qualities of Films on Aluminum

There are a number of electrolytes, including various concentrations of phosphates, borates, tartrates, carbonates and others in which films can be formed on aluminum to withstand maximum potentials

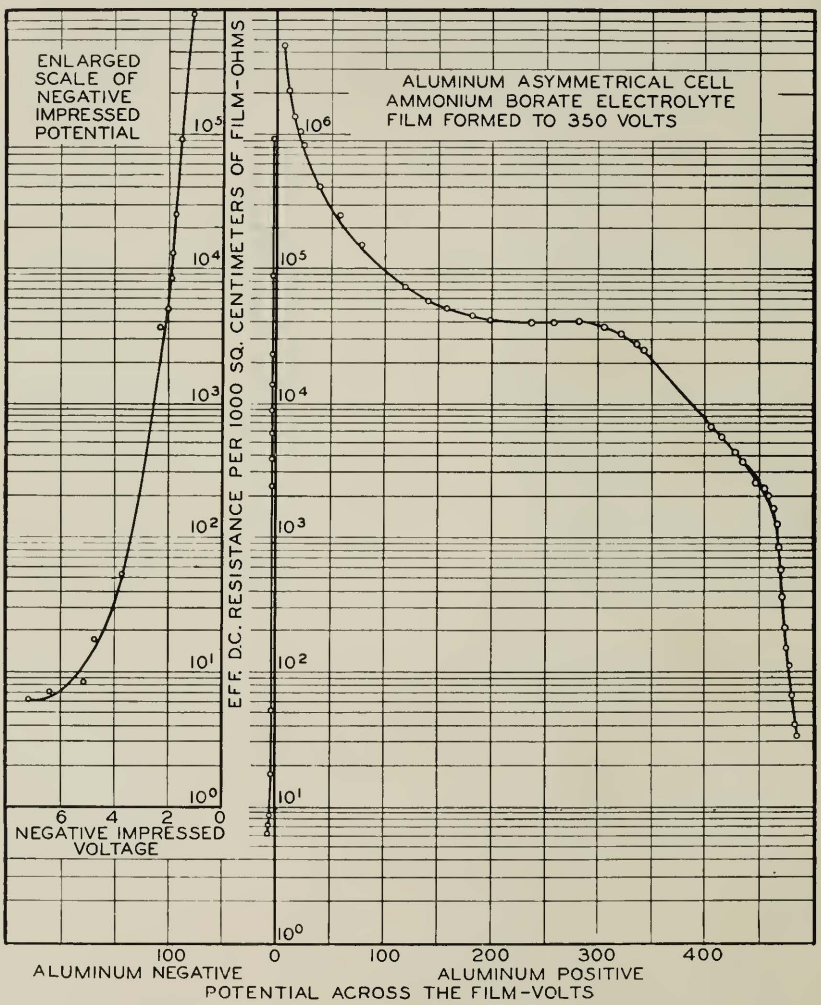


Fig. 1—Influence of impressed potential upon the electrical resistance of the film on an aluminum electrode

[4] Young, R. L., *Bell System Techn. J.*, **6**, 708 (1927).

upwards of 300 volts, at least for limited periods. If a film is formed on a piece of aluminum to this maximum voltage and the metal is then made the anode in an electrolytic cell across which variable potential can be applied, a current corresponding to a density of less than a microamp. per sq. cm. of filmed surface will flow when a potential of one tenth the maximum voltage is impressed.

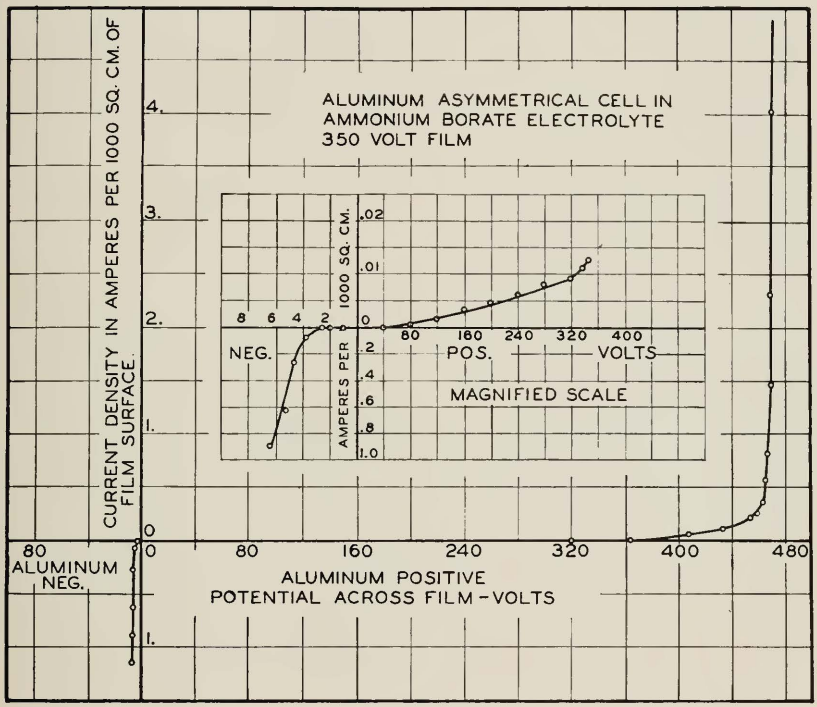


FIG. 2—Influence of potential on the current through an electrolytic cell with a film-coated aluminum electrode

As the potential is increased this "leakage" current will increase at a rate somewhat greater than proportionate to the voltage. As the maximum or breakdown potential is approached it will be noticed, if the room is darkened, that the anode begins to glow uniformly over the surface with a pale light and with further increases in voltage sparks begin to scintillate over the entire electrode, being noticed first at the surface of the electrolyte. The current through the cell becomes appreciable under this condition and increases more rapidly until at voltages slightly above the sparking potential the cell acts virtually as a short circuit.

Upon reduction of the voltage, however, the insulating properties of the film are restored and the current decreases with decreasing potential in substantially the same relation to voltage as before. The sparking over the surface will be observed to cease at about the same potential at which it began, the glow will disappear and the low leakage-current values will be obtained when the voltage is reduced sufficiently.

Upon reversal of potential on the aluminum electrode, however, there is a much larger flow of current, the value of which is limited by a counter voltage of several volts, and by the low internal resistance of the cell with negative potential applied. Typical current-voltage relations for an aluminum cell are shown graphically in Fig. 2. These relations correspond to the curve in Fig. 1, showing the variation of resistance with potential of a "filmed" aluminum electrode in ammonium borate electrolyte.

Capacity of Aluminum Films

Like the ordinary paper or mica static condenser, the electrolytic condenser consists of two conducting surfaces separated by an insulator. The high-resistance film constitutes the insulator in the electrolytic cell, and the electrolyte on one side of the film and the metal of the film-bearing electrode on the other provide the two conducting surfaces. The cathode in this type of cell merely provides a means for making electrical contact with the electrolyte.

When a film is formed upon a smooth polished aluminum surface the coating is transparent. If observed under favorable illuminating conditions the "filmed" surface is seen to be colored and may be either green, yellow, red or blue, depending upon the thickness of the film. This is attributed to light interference and indicates that the thickness of the film is in the order of the length of light waves. The actual thicknesses of films on aluminum have been determined to be from 0.001 to 0.00001 mm.,[5] depending upon conditions of formation.

Because of this extreme thinness of the dielectric and its high insulation resistance when positive potential is applied, unusually large capacities per unit area of surface can be obtained. The capacity of a film formed to 30 volts on aluminum is about 0.18 microfarad per sq. cm. of dielectric surface, or about 1,000 times that of paper condensers. The capacity per unit area is approximately inversely

[5] Zimmerman, *Trans. Am. Electrochem. Soc.*, **7**, 309 (1905); Sutton and Willstrop, *Engineering*, **124**, 442 (1927); Slepian, *Trans. Am. Electrochem. Soc.*, September, 1927, to be printed in Vol. 54 of the *Transactions*.

proportional to the potential at which the film is formed, indicating that the thickness of the dielectric is directly proportional to the voltage of formation.

Effect of Impressed Voltage on Capacity

When an electrolytic cell with a "formed" anode has impressed on its terminals a voltage greater than the formation voltage, the film must build up to the new potential before the electrical characteristics of the cell become stable. At this higher voltage the capacity of the cell will be reduced to correspond to the increased potential. Where large plate areas are involved the direct application of a potential above the formation voltage results in a heavy flow of current, which may overheat and damage the cell if not properly limited.

If a voltage is impressed on a condenser lower than the potential applied during the formation of the film, the cell will operate satisfactorily, but the capacity will not be immediately affected and will correspond to the potential at which the film was originally formed. This is illustrated in Table I, which shows the capacities of an electrolytic cell measured with applied voltages of different values below the voltage of formation.

However, if a condenser operates for a long time at a reduced voltage the excess film will be removed slowly by the chemical action of the electrolyte, and the capacity will increase gradually to a value depending upon the operating voltage. The rate of change of capacity under these conditions is affected by the temperature at which the cell operates, and by the conductivity of the electrolyte. As is illustrated by the curves shown in Figs. 3 and 4 the change becomes more rapid when these factors are increased.

TABLE I

Applied Potential Volts—D.C.	Capacity Readings Microfarads
49.2	1008
43.0	1002
36.7	1008
30.4	1013
24.3	1013
18.3	1013

Series Connected Condensers for Direct and Alternating Current Service

In the discussion of the current-voltage relations in Fig. 1, it was noted that an aluminum cell with one electrode of non-film-forming metal conducts current freely when the aluminum is cathode. Accordingly this type of cell is capable of holding a charge of electricity and

serving as a condenser only while the aluminum is at a higher positive potential than the electrolyte.

A cell with a non-film-forming cathode makes a suitable condenser to operate on direct-current or pulsating-current circuits, in

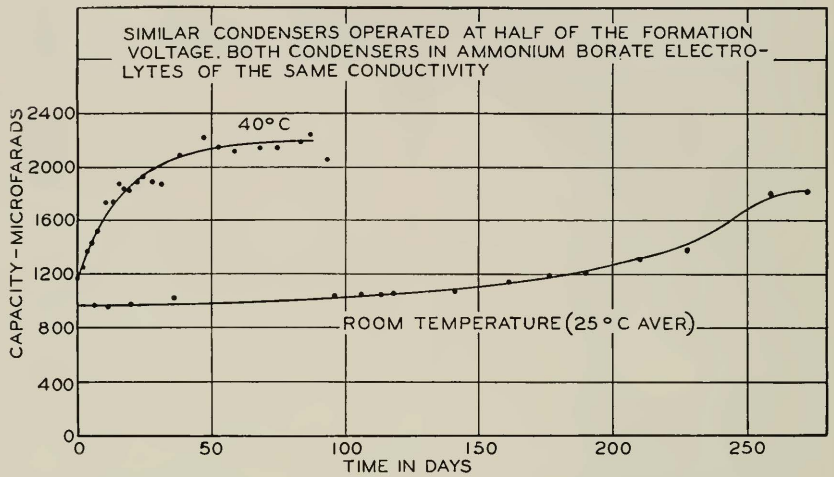


FIG. 3—Effect of electrolyte temperature on the rate of capacity change in condensers operating at voltages below the formation voltage

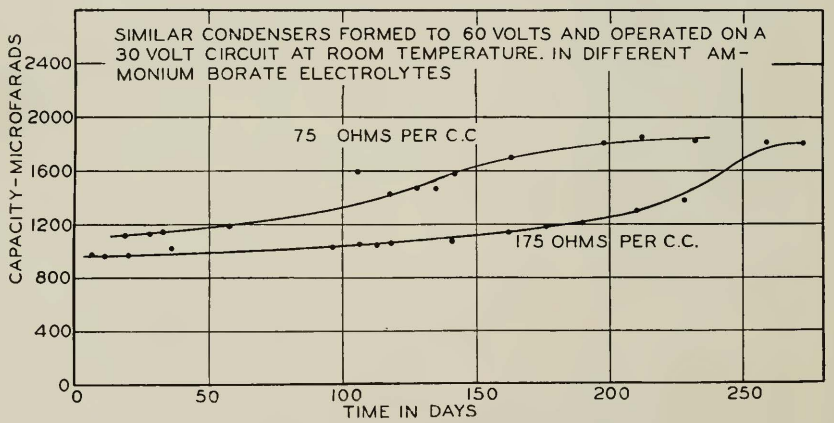


FIG. 4—Effect of electrolyte conductivity on the capacity change due to reduced operating voltage

which the aluminum always remains positively charged. On alternating-current circuits, however, such a cell will operate as a rectifier rather than as a condenser, unless two similar units are connected in a series-opposed relationship.

A suitable condenser for operation on alternating current can also

be made by having two electrodes of film-forming metal in the same solution, the electrical relations between the "filmed" electrodes being the same in this case as in the series-opposed arrangement of two asymmetrical cells. In either case one or the other of the film-forming electrodes opposes the flow of current during each half cycle.

If we consider the conditions that exist at the instant of maximum potential in one direction, the electrode that is then anode acting as a condenser, receives its maximum charge. As soon as the potential begins to decrease from this maximum the accumulated electricity begins to flow from the charged plate through the circuit, but in so doing the opposing film-forming electrode becomes an anode, enabling it to hold the charge given up by the discharging electrode.

In this way, as the alternating potential varies between maximum values in each direction, the charge is transferred from the capacity provided by one film-forming electrode to the other, the sum of the charges on these two electrodes at every instant remaining constant. It can be shown that two series-opposed asymmetrical cells of capacities C_1 and C_2, or two "formed" electrodes of these capacities in the same electrolyte, have a resulting capacity equal to

$$\frac{C_1 C_2}{C_1 + C_2}.$$ [6]

As illustrated in Table II, which gives the results of measurements on two asymmetrical cells formed to different voltages and connected in various series combinations, it will be noted that this relationship

TABLE II

Measurements at 60 Cycles

	Biasing Potential Volts D.C.	Capacity Microfarads	Equivalent Series Res. Ohms
Condenser "A"	24.6	1529	0.165
Condenser "B"	44	962	0.23
"A" and "B" series—aiding	50	587	0.385
	43.8	582	0.38
	25	588	0.38
"A" and "B" series—opposing			
"B" positive	37.3	588	0.395
	18.3	588	0.38
No bias	0	594	0.385
"A" positive	12	588	0.39
	24	588	0.39
Calculated from measurements of "A" and "B"		589	0.395

[6] Zimmerman, *Trans. Am. Electrochem. Soc.*, **7**, 323 (1905).

is true both on alternating and direct-current circuits. If unidirectional potential is applied this relation holds without regard to the polarity or the magnitude of the impressed voltage so long as this voltage does not make the potential across the film opposing the flow of direct current greater than the voltage to which this film was formed.

In the case of asymmetrical cells connected in series-opposed relation the expression is correct either when the positive plates

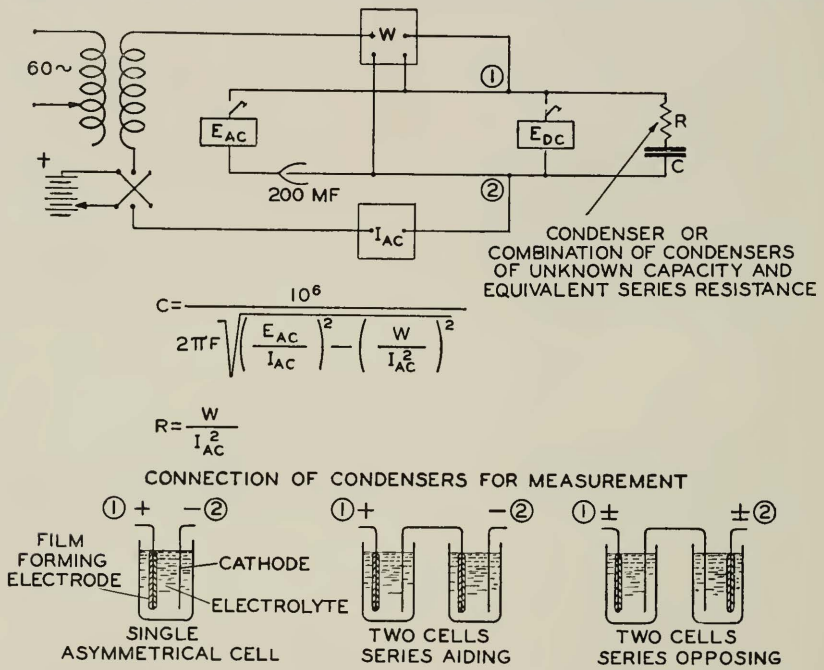


FIG. 5—Determination of capacity and equivalent series resistance of electrolytic condensers with 60-cycle alternating current superimposed upon a variable unidirectional potential.

of the two cells are connected together at the mid-point or when the negatives are so connected. On direct current it applies for series-aiding as well as for series-opposed connections.

It will be recognized that the resultant capacity $\frac{C_1C_2}{C_1+C_2}$ is the same as is obtained when two ordinary static condensers of values C_1 and C_2 are connected in series. However, the internal distribution of electrostatic charges in opposed electrolytic condensers is quite different from that in series-connected static condensers, due to the rectifying characteristic of the films.

These determinations were made with a voltmeter, ammeter and wattmeter at 60 cycles, and were in conformity with measurements over a range of frequencies made on an impedance bridge. In the cases of measurements made with unidirectional potential applied, suitable bias was provided by superimposing the alternating current by means of a transformer in a battery circuit, the d.c. potential of which could be varied by changing the number of cells in series.

The apparatus for these measurements is shown in Fig. 5. It will be noted that a high-capacity blocking condenser is required in the a.c. voltmeter circuit, but is omitted from the potential circuit of the wattmeter. This condenser is inserted to block the unidirectional potential, which otherwise would be read by the voltmeter, but the power that the wattmeter indicates due to this potential is of no importance, and except for direct-current leakage would actually be zero.

Losses in Aluminum Cells and their Effect on Electrical Impedance

In the matter of electrical impedance characteristics, the electrolytic condenser does not approach a perfect capacitance as nearly as the more familiar forms of static condensers. Three sources of energy loss in the electrolytic condenser impart to it an equivalent series resistance, as a result of which the condenser current leads the impressed voltage by a phase angle somewhat less than 90°.

The first of these losses is the dielectric hysteresis loss, which, as in the case of the paper condenser, is approximately proportional to the frequency. The second loss is due to the resistance of the electrolyte and, in the case of aluminum condensers, may be of appreciable magnitude because of the low electrical conductivity of suitable electrolytes. This electrolyte resistance remains practically constant over a wide range of frequencies.

The third possible loss is due to the leakage-resistance of the film, which in its effect is similar to a high resistance in parallel with the condenser. Ordinarily this loss is negligible because the leakage current is less than a microampere per sq. cm. of film surface.

Conditions Affecting the Life of Condensers

To be successful from a commercial point of view an electrolytic condenser must have long life and must not require frequent attention. Otherwise the advantage in the matter of mounting space and the cost per unit capacity is offset by the depreciation and maintenance costs involved. There are two common conditions affecting the life

of aluminum condensers that must be controlled if the cells are to operate satisfactorily.

The first concerns the chemical action of the electrolyte on the electrodes and the film. This action, which is merely a matter of the film dissolving and forming aluminum hydroxide in the solution, takes place when the cell is off circuit as well as when potential is impressed. With impressed potential, new film forms under the influence of the leakage current to replace that which is dissolved, but in time the fluid becomes saturated with aluminum hydroxide, which may precipitate as a white jelly and adversely affect the life of the condenser.

The second consideration involves corrosion of the positive electrodes. The susceptibility of aluminum to corrosion is well known, and in the use of electrolytic condensers anodic corrosion is the most damaging irregularity that can occur.

Obviously then an electrolyte must be chosen that does not rapidly dissolve the film, and the material for the electrodes as well as for the electrolyte must be selected and prepared to prevent serious corrosion of the "formed" aluminum plates.

Commercial Applications and Designs of Electrolytic Condensers

Reference has already been made to the use in telephone systems of electrolytic condensers. The principal applications of this device involve its use in low-pass electric wave-filters. These filters are placed in the supply circuits associated with central office storage batteries to eliminate noise-producing ripples and pulsations, introduced by battery charging-apparatus and signaling equipment, from the direct current furnished to telephone instruments. That is, the filters are used to exclude hum and other disturbing noises from the subscribers' circuits.

In Fig. 6 is shown an electrolytic condenser of the type designed for direct-current filter service. When prepared for operation on 24-volt d.c. circuits, the capacity of this cell is nominally 1,000 mf. at 1,000 cycles, and for 48 volts is about 600 mf. at the same frequency. The cell is 8 inches wide, 10.25 inches long and 14.25 inches high (20 x 26 x 36 cm.). Completely assembled it weighs about 42 pounds (19 kg.), including 22 pounds (10 kg.) of electrolyte.

The container for the condenser is made of heat-resisting glass which reduces possible breakage due to temperature variations. The electrodes, both of aluminum, are rigid and are bolted to a porcelain cover to keep them in proper space relation. Two supporting bolts,

one from each electrode properly marked with respect to polarity, extend through the cover to provide the terminals for the condenser.

A thin layer of high grade paraffine oil is used on top of the con-

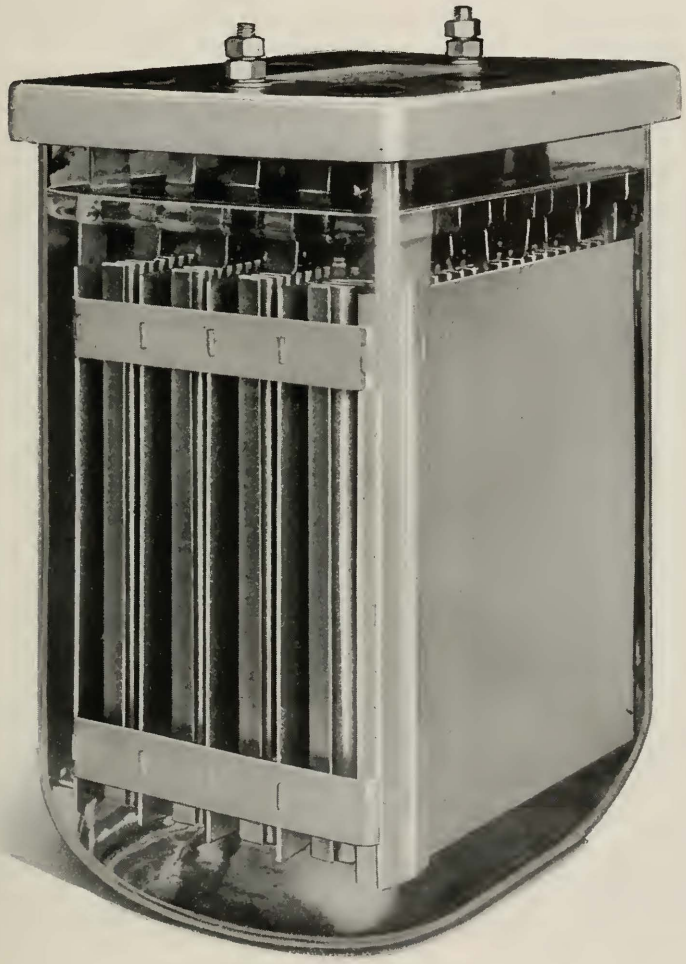


FIG. 6—Aluminum electrolytic condenser, designed for direct-current filter service

denser fluid to prevent evaporation and to keep the inside of the cell from sweating under varying room temperature conditions. The cover is sealed to the glass jar with paraffine to provide additional protection against evaporation and to prevent dirt from getting into the cell.

The Anode Construction and Material

The construction of the electrodes is shown in Fig. 7. The positive electrode on which the dielectric film is formed is made of four corrugated aluminum plates, each supported by four integral ears. In an assembled condenser the positive plate surfaces are entirely immersed in electrolyte, the ears extending up through the oil and providing contact with the positive terminal.



Fig. 7—Aluminum electrolytic condenser, showing construction of the electrodes

The material for the positive plates is aluminum of special composition, selected on a basis of properties which influence the formation of the film, the leakage current and the life of the metal. In general, the higher the purity of the aluminum the more rapid is the formation of the film and the lower is the resultant leakage current.

It has been noticed that the unit-area capacity for high-purity

metals is slightly lower than for metals containing small quantities of alloying materials, but this difference is of inconsequential importance. The difference in the rate of film formation under similar conditions between 99.1 per cent aluminum and 99.6 per cent aluminum is shown in Fig. 8. After 24 hours in these particular cases the

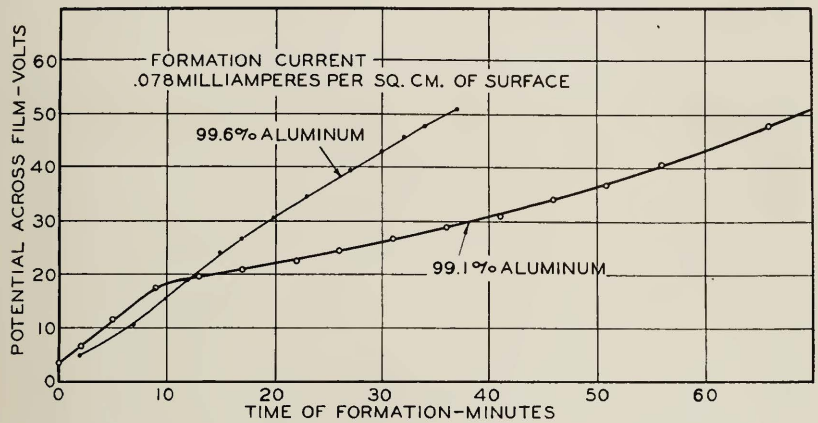


FIG. 8—Rate of film formation as influenced by composition of metal

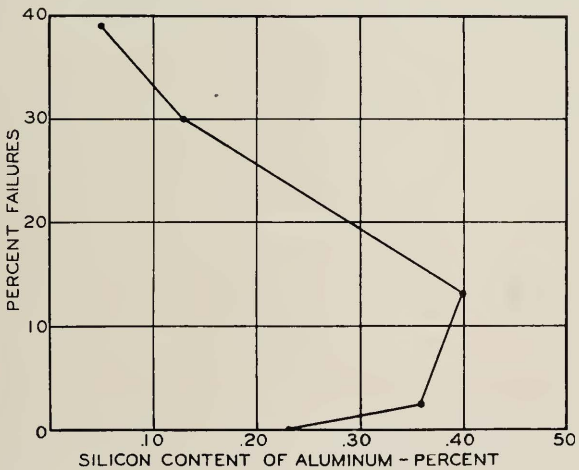


FIG. 9—The relation between silicon contents of anodes and failures due to corrosion in groups of aluminum condensers classified with respect to anode composition

leakage current of the 99.1 per cent aluminum was about 3 microamperes per sq. cm., or nearly six times that of the 99.6 per cent material at the formation potential of 60 volts.

In the matter of life, however, the purer metals seem to be more readily attacked by agencies capable of causing electrolytic corrosion.

This is illustrated by the performance of groups of condensers having anodes of different compositions as represented in Fig. 9. This curve shows the percentage of the cells in each group affected by corrosion in relation to the amount of silicon in the anode aluminum. It will be noted that the group of condensers with anodes of the purest metal (that is, least silicon) gave the least satisfactory results.

The Negative Electrode

The negative electrode consists of five rectangular flat plates, having a combined useful surface area about 35 per cent of the total positive surface. The negative plates or cathodes are of aluminum, but they do not have a film formed on them because their sole function is to provide contact with the condenser fluid. In an ammonium borate electrolyte, such as is used in these condensers, there are a number of other materials, including tin and carbon which can be used for the negative electrodes.

However, aluminum was chosen because it is light, relatively strong, easily worked and mounted, and can be cleaned by the same process used to clean the positive plate metal. A question might properly be raised as to the use of an aluminum negative electrode, particularly one having less area than the positive, because the formation of a film on the negative electrode will result in a reduction of the electrostatic capacity of the cell.

In normal operation with aluminum negatives there is a tendency for a film to form, even though the condenser is operated on direct-current circuits, because the negative electrode is an anode during the interval that the condenser discharges. However, this disadvantage with respect to the use of aluminum cathodes is overcome by making these plates of metal, which is really a rich aluminum alloy, containing enough other substances such as silicon to impede the formation of a film on its surface.

With material containing less than 99 per cent aluminum it is possible to have as much as 3.5 amperes a.c. in the condenser circuit or an alternating-current density of about 1 milliamp. per sq. cm. of negative plate surface without forming sufficient film on the negative plates to affect the capacity of the cell.

Prevention of Contamination in Manufacture and Installation

The initial formation of the film on the positive electrodes is carried out by chemical cleaning and electrochemical formation processes, in which the purity of the materials used as well as the composition of

the electrodes is of importance. Because of the delicate nature of the film and the desirability of keeping it clean, the electrodes are designed so that in the manufacture, packing and installation of the condenser it is unnecessary to touch either the positive or the negative plate surfaces after the film is formed.

The fluid in which the condensers operate is shipped in sealed glass containers to prevent contamination, and the routine to be followed in the installation of the condensers emphasizes the need for cleanliness on the part of the installer. However, the design

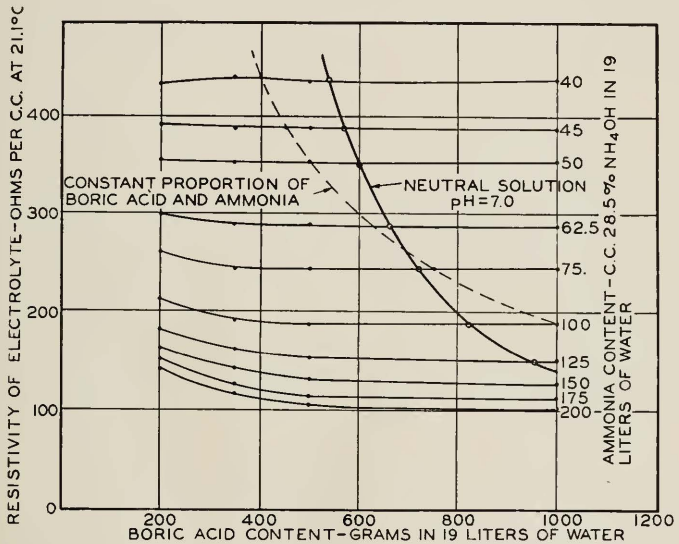


FIG. 10—Specific resistance chart of ammonium borate electrolyte

of the condenser and the method of supplying the condenser fluid have been arranged so that satisfactory installations can be made with ordinary skill and simple precautions.

THE COMPOSITION OF THE ELECTROLYTE

The electrolyte or condenser fluid as the solution is called is a mixture of ammonia, boric acid and water. A chart showing the specific resistance of different ammonium borate electrolytes is shown in Fig. 10. Two curves on this chart illustrate an interesting characteristic of these solutions. The heavy solid curve shows the neutral ammonium borate solutions, that is, those having a hydrogen ion concentration of 10^{-7} mols per liter.

This curve was determined colorimetrically with suitable indicators against standard solutions, and was checked by measurements with

the hydrogen electrode. The dotted curve shows different dilutions for the same proportion of ammonia and boric acid. Since these curves intersect it is evident that the acidity of an ammonium borate solution decreases with dilution, and acid solutions may become alkaline by the addition of sufficient water.

Within the ranges of compositions and concentrations shown, it will also be noted from the set of approximately horizontal full line curves that the specific resistance of the electrolyte is practically independent of the boric acid content. That is, within these limits, the conductivity of the solution is substantially determined by the ammonia content, the amount of boric acid in the electrolyte affecting principally the degree of acidity or alkalinity of the solution.

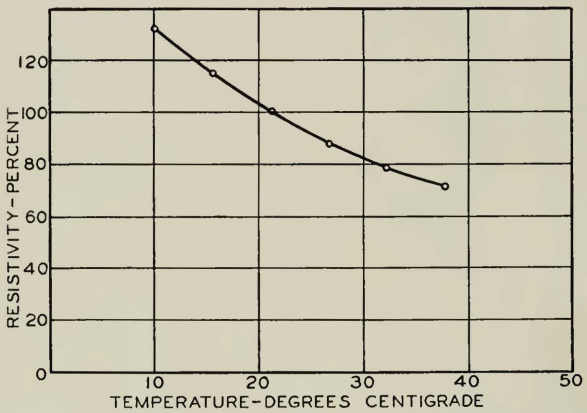


FIG. 11—Effect of temperature on the resistivity of ammonium borate electrolyte

The specific resistance of the solution is, of course, affected by heat and decreases with increasing temperature, as shown graphically in Fig. 11. In this figure the specific resistance in per cent of the specific resistance at 21.1° C. is plotted against the temperature of the electrolyte.

In the selection of a suitable electrolyte for a condenser the choice is influenced by the life of the solution, the effect of the specific resistance of the electrolyte on the electrical characteristics of the cell, and the susceptibility of the electrodes to corrosion in the solution.

The life of the solution, as has been explained, is determined by the rate at which it becomes saturated with aluminum hydroxide. In general, the lower the specific resistance of the solution, the more quickly does aluminum hydroxide form. On this basis the advantage of a high-resistance electrolyte is obvious.

But a limit is reached beyond which increases in specific resistance produce objectionable additions to the electrical impedance of the condensers, because of the high internal resistance set up in the cells.

In a condenser, as shown in Fig. 6 at room temperatures of about 25° C., a fluid having a specific resistance of 75 ohms per cc. will last from six months to a year without need for renewal because of the precipitation of aluminum hydroxide.

With fluid of 150 ohms per cc. the period of useful life is from one to three years, and with a 300-ohm solution is upward of five years, possibly never requiring renewal within the useful life of the cell. The rate at which the precipitate forms in a given solution is greatly accelerated at elevated temperatures and at 40° C., for example, the solution remains free from a white precipitate only about one third as long as at 25° C.

With respect to the effect of the acidity or alkalinity of the solution on the operation of aluminum condensers the difference is not readily distinguishable. Films can be formed and cells can be operated both in acid and alkaline electrolytes, and the electrical characteristics, except for resistance effects due to different solution conductivities, are essentially the same in both kinds of electrolyte. Somewhat better results, with respect to corrosion of electrodes have been obtained, however, with alkaline solutions, particularly under unfavorable operating conditions.

The Relation between Voltage of Film Formation and Operating Voltage

Electrolytic condensers, used on circuits associated with storage batteries, must be capable of operating at potentials throughout the range of voltage variations due to charging and discharging the batteries. Both the 24-volt and the 48-volt type condensers described can be used in circuits up to 140 per cent of their normal voltage, provision for this variation being made by the initial formation of a film to a potential somewhat above the maximum operating value.

After a condenser is connected in service the excess thickness of film is removed slowly by the chemical action of the electrolyte, because the film on the anode is maintained only at a thickness corresponding to the operating voltage. Ordinarily, therefore, the capacities of these condensers increase from their initial values, and stabilize at new values depending upon the maximum potential in the cycle of operating-voltage variations.

Capacity and Resistance Characteristics

The capacity and resistance characteristics plotted against frequency for a condenser of the type illustrated in Fig. 6, with a film formed to 46 volts d.c. are shown in Fig. 12. Curves No. 1 show the values

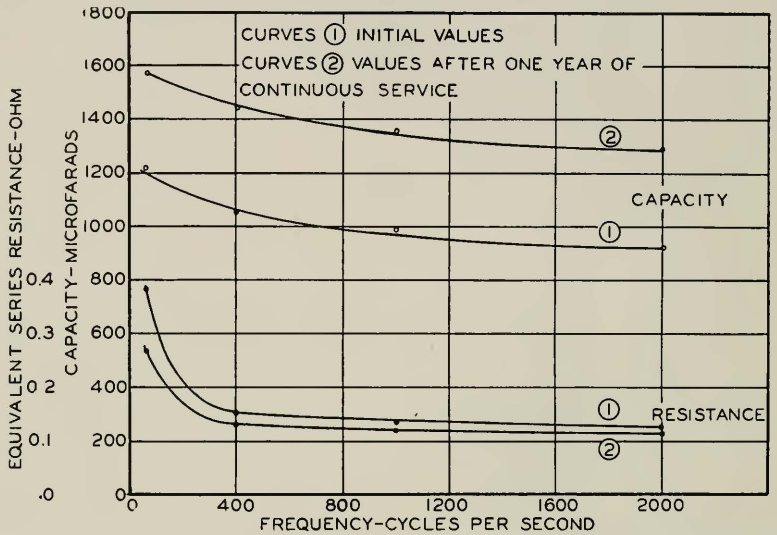


Fig. 12—Effect of frequency on capacity and resistance of an aluminum condenser formed to 46 volts and operated at 28 volts

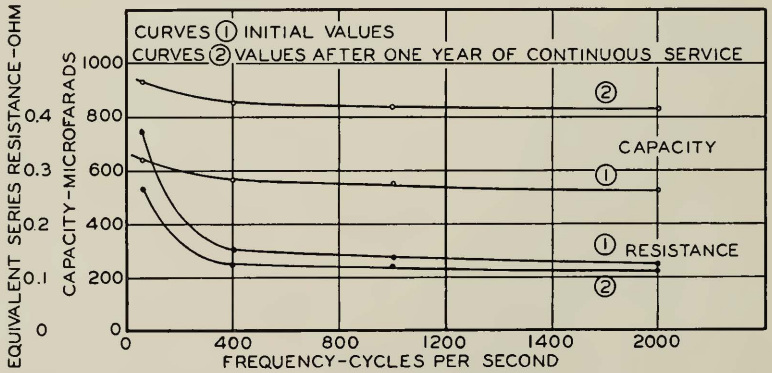


Fig. 13—Effect of frequency on capacity and resistance of an aluminum condenser formed to 100 volts and operated at 66 volts

measured at the time the condenser was put in service, and curves No. 2 show the values after one year of continuous service on a 28-volt battery with maximum voltage of 32. A second set of curves for a similar condenser formed to 100 volts and operated on a 66-volt battery, maximum 75, are shown in Fig. 13.

It will be noted that the capacities of the condensers decrease with increasing frequency. While there is a slight decrease in the unit-area capacity of films that accompanies a rise of frequency [7] this drooping characteristic is due principally to the corrugated shape of the plates on which the film is formed. The resistance through the electrolyte from the negative electrode to the film at the mouth of a "U" shaped corrugation, is less than that to the portion of the film at the bottom of the "U."

Thus, as the frequency increases, the alternating-current density at the mouth of the corrugations increases, while that in the trough decreases, resulting in a decrease of the effective capacity of the unit. The change in capacity due to frequency is greater in con-

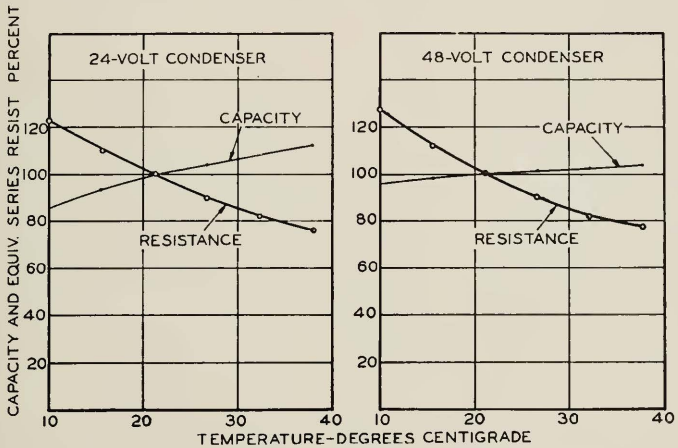


FIG. 14—Effect of temperature upon capacity and equivalent series resistance of 24-volt and 48-volt condensers at 1,000 cycles

densers with high-resistance electrolytes, and is more pronounced in the case of low-voltage films. The series resistance of the condensers is also observed to fall quite rapidly as the frequency is increased. This is due to the inverse proportionality between frequency and the component of the total condenser resistance, representing dielectric loss.

Since the change in capacity with frequency is caused by the difference at the crest and in the trough of the corrugations of the ratios of unit-area capacity to associated electrolyte resistance, changes in electrolyte resistance, caused by variations in temperatures, also influence the effective capacity of the cells.

The temperature effect is particularly noticeable at high frequencies,

[7] De Bruyne and Sanderson, *Trans. Faraday Soc.*, **23**, 42 (1927).

and, as in the case of the change of capacity with frequency, is more pronounced in condensers with low-voltage films. In Fig. 14 the effect of temperature on the electrical characteristics, at 1,000 cycles for 24-volt and 48-volt condensers, is shown. Here the change in resistance is due to the negative temperature coefficient of the electrolyte which, with increasing temperature, causes a reduction in that component of the total condenser resistance representing the resistance of the electrolyte.

In the normal adaptations of these condensers in low-pass electric wave-filters, the inherent changes in capacity and resistance with frequency and temperature cause no serious engineering difficulties. The limitations imposed by these variations are more than offset by the advantages of the corrugated structure in the matter of compactness and simplicity of design.

Corrosion of the Condenser Anodes

Notwithstanding the care taken in the manufacture and installation of condensers, there remains some possibility that the positive electrode will be attacked by corrosion. This corrosion may make its appearance as a gray growth or pitting on the surface of the anodes, or on the anode supports where they extend through the electrolyte. It is usually accompanied by the deposit of a granular or finely divided gray substance, probably aluminum oxide, which collects in the bottom of the condenser jar. Corrosion may occur shortly after a condenser is put in service or months can elapse before it appears.

In cases where corrosion has occurred, provided proper materials were used in the manufacture of the condenser, it has been noticed that the electrical properties of the cells were not seriously impaired. Such units have often continued to perform satisfactorily from a circuit standpoint for a number of years after the electrodes were attacked, even though the mechanical structure of the electrodes was damaged and weakened.

The long life that can be obtained from a corroded condenser under these circumstances is due in part to the tendency for areas affected to heal and restore the condenser to normal conditions. This indicates that the influence responsible for corrosion dissipates itself or it may be carried away from the corroded area by some of the products of the action to lie inert in the bottom of the cell. A number of spots on aluminum anodes have been observed, where the attack on the metal has ceased and a film has formed over the affected surface.

The leakage current of a condenser increases substantially when

corrosion occurs, going up in some cases 20 or 30 times, but this current again decreases and approaches the original value. The capacity likewise increases when corrosion occurs, but this merely lowers the impedance of the condenser which, in most cases, is not objectionable. The performance of a condenser that has corroded and continued to operate satisfactorily is shown in Fig. 15. The curves on the chart show how the leakage current, capacity and resistance varied during continuous operation on a 65-volt d.c. circuit over a period of two years, both before and after the anodes were attacked by corrosion.

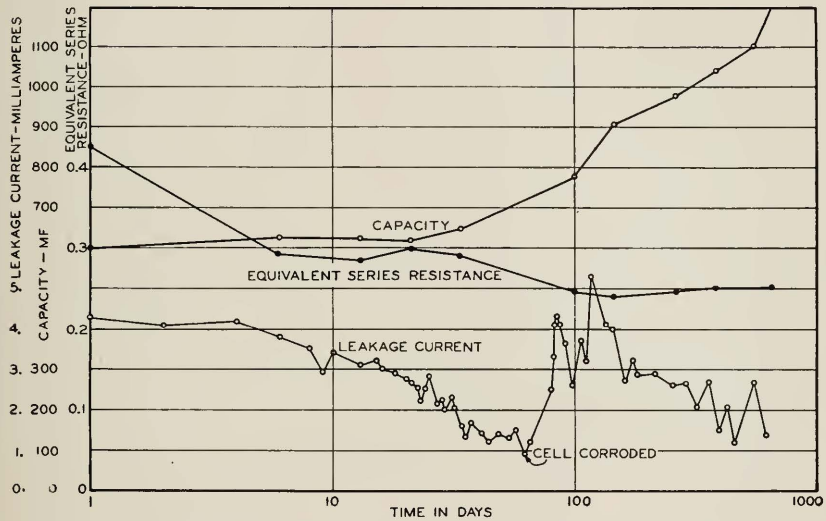


FIG. 15—The performance of a condenser, the anodes of which were attacked by corrosion after 62 days of operation at 65 volts direct current

Under certain conditions it is possible for the products of corrosive action to cause short circuits within a condenser, not by bridging from the positive to the negative plates because the corrosion product, aluminum oxide, is a non-conductor, but by accumulating between the plates and forcing the positive plates out of position into contact with the negative.

Also if corrosion occurs on the positive terminals eating through the supports one of the plates may drop and cause a short circuit. These possibilities of trouble are minimized by suitable design, and can be cleared up when they occur by removing the electrodes from the solution and repositioning or removing the deranged anode plates.

Because of the tendency for the vigor of the corrosive attack on aluminum anodes to decrease as the action continues and because the cell will continue to operate satisfactorily from an electrical standpoint, the best practice for condensers affected by corrosion is to leave them alone unless it is necessary to clear a short circuit due to a buckled plate or a severed anode support. Several cases have been reviewed where condensers are continuing to operate without maintenance, though corrosion was experienced nearly four years ago.

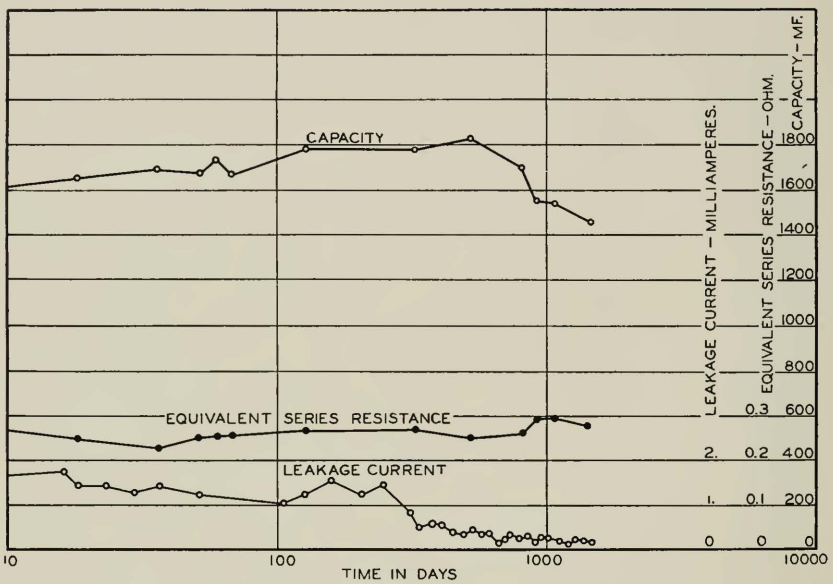


FIG. 16—Performance of a normal condenser operating continuously at 28 volts, direct current

THE LIFE OF ALUMINUM CONDENSERS

With respect to the life of normal aluminum condensers, a number of cells have been on trial in actual service installations for more than five years, and several have been operating under test conditions for about eight years. In general, after the capacity of a condenser has stabilized at the value corresponding to the operating voltage, the cell continues to operate at a comparatively constant but slowly decreasing capacity due, perhaps, to a gradual thickening of the film with age.

The leakage current also decreases with length of service. The capacity, resistance and leakage current of a typical condenser operated for several years is shown in Fig. 16. In this case the fluid had

never been renewed since the condenser was installed and after the period of service shown the electrolyte was still free from the white precipitate of aluminum hydroxide.

What the ultimate life of a condenser of this type will be remains for future determination. Judging from the appearance of cells that have operated for six and seven years there would seem to be years more of useful service to be rendered by these cells.

Contemporary Advances in Physics, XVII

The Scattering of Light with Change of Frequency

By KARL K. DARROW

SCATTERING of light is one of the commonest of all phenomena, which does not in the least imply that it is one of the most commonplace. Even its practical importance entitles it to high respect. We are often told that were it not for scattering, the sky would not be blue; the sun and the stars would stand out amazingly brilliant against a background black as coal. It is probable, however, that if scattering were suddenly to be suspended, the disappearance of the sky would be one of the least of our worries. Everything else would disappear, except what was self-luminous. The visible world would consist of the sun, the other stars, and flames, some electrical discharges, the filaments of incandescent lamps, and some substances glowing feebly with fluorescence or phosphorescence. Nothing else could be seen except as a silhouette, apart from objects so translucent that they could be viewed as a stereopticon slide against a flame. Happily no such calamity impends; and we may unconcernedly consider the theoretical importance of the process, which is great. As some might say, the scattering of light is one of the battlegrounds between the undulatory and the corpuscular theories. Metaphors of combat are however not appropriate; it is necessary to reconcile the theories, not to smash one or the other. Now it happens that some of the phenomena of scattering may be interpreted by the one theory, and some by the other; and some can be explained by either, which is most auspicious; for if this can some day be said of all the phenomena of light, the goal of our desires will have been attained.

Also scattering of light has just sprung into prominence as the most inviting and the most ardently invaded field of physics, because of a discovery such as, it had been supposed, could never happen again. It seemed that experimental physics had been so thoroughly developed —or, to change over to an ancient metaphor, that the field had been so thoroughly harvested and then so exhaustively gleaned—that nothing important could possibly remain to be discovered unless by measurements of great precision, or by radically new apparatus, or by applying voltages or other agencies on a scale as yet untried. Yet in the spring of 1928 a mode of scattering of visible light was discovered,

by a physicist working with a quite ordinary spectroscope, a quite ordinary source of light and some very familiar chemicals, all of which had been available to everyone for at least fifty years. The physicist who thus saw what for half a century the whole world had overlooked was C. V. Raman of Calcutta.

Within a few months it was observed that what Raman had discovered was one of the special cases of a very general principle, which had already been stated almost as clearly as one can state it even today, but had failed of due recognition, evidently for want of some conspicuous example. Other special cases had indeed already been discovered among the phenomena of X-rays, but for some reason or other all except one had failed to make the impression which they should have made. Raman's discovery brought the principle sharply into relief. I shall not state it fully at this point; but this is its practical consequence: *when light falls upon matter, the scattered rays need not all have the same frequency as the infalling.* Most of the scattered light may preserve the initial frequency unchanged as formerly it was supposed that all did; but some part of it may be modified or shifted.

The present paper is devoted to this principle almost entirely; but to give it the proper background, I must at least mention some of the other aspects of scattering. Light may be scattered by particles of any size, from atoms up to grains of dust or droplets of mist; even the reflection from larger bodies is to be considered as the resultant of light-scattering from the particles of which these are made. Scattering by granules or droplets may be analyzed by the electromagnetic theory, the substance of the particle being considered as a medium with a different index of refraction and a different coefficient of absorption from the surrounding air. This however is distinct from the case with which we have now to deal, the scattering of light by molecules or atoms—although the transition from one type to the other should be very interesting and instructive.

The effect produced by molecules or atoms—it is known as the Tyndall effect, from its earliest thorough student—is fairly conspicuous in liquids or crystalline solids, though a very small trace of dust or finely-dispersed precipitate produces a brilliant scattering which completely overwhelms it. In gases it is difficult to see, but not impossible. The blue of the sky is a specific instance which is easy to observe, because the air is so deep.

The features of the scattered light which are commonly studied—I am speaking now of the time when it was assumed that there is no change in frequency—are its intensity, its polarization, and a property which is rather vaguely known as "coherence." The intensity of the

light scattered from an incident beam, of which the wave-length is varied while the intensity is kept constant, increases very rapidly indeed with decreasing wave-length—in some cases, as the inverse fourth power of the wave-length. (This is the reason why the sky is blue; the molecules of oxygen and nitrogen are not especially tuned to blue light, but the waves from the sun are more powerfully scattered the nearer they lie to the violet end of the visible spectrum.) The scattered light is more or less polarized, even when the primary light is not polarized at all. In some cases the light deflected through 90° is perfectly plane-polarized; by the undulatory theory this signifies a very simple sort of vibrator in the atom, a vibrator which is attracted by an equal restoring-force whichever way it is displaced from its centre of vibration. In other cases the polarization is different, and other inferences about the atom-model may be drawn from it. As for the "coherence," it is a very important property—important for the theorist. If the scattering atoms contain vibrators which the infalling waves maintain in forced oscillations, and which themselves send out the scattered light, then these scattered or "secondary" wave-trains should interfere with one another. If in particular the vibrators, or let us say the atoms, form a regular lattice in space—a cubic lattice, for example—there should be destructive interference; the secondary wavetrains should completely destroy one another in all directions save that of the ongoing primary beam,[1] and there should be no perceptible scattering at all. The perceptible scattering is then a measure, to speak rather vaguely, of the *irregularity* in the arrangement of the atoms. This theory seems to be confirmed, for the light scattered without change of wave-length. Whether it is true also for the shifted light will probably soon be known.

Scattering of light with unchanged frequency is easy to explain by either wave-theory or corpuscle-theory. To those who think of light as waves and of atoms as systems of vibrators, it is a consequence of forced vibrations. The fluctuating or alternating electric field which—coupled with an alternating magnetic field—constitutes the beam of light, seizes upon an electrified portion of the atom and swings it to and fro in synchronous vibration. From this swinging electric charge, the scattered waves originate. It is evident that the forced vibrations and the scattering should be especially intense, when the frequency of the light coincides with a natural frequency of vibration of the atom, for then there is resonance. Now it is a fact that scattering is especially intense, when the infalling light agrees in frequency with any of

[1] More precisely, in all directions save those of the Laue diffraction-beams, which however for crystals and visible light do not occur.

certain spectrum lines which the atom may emit spontaneously. To those who think of light as a hail of corpuscles—"quanta"—scattering is rebounding of the quanta from atoms which they strike "elastically"; that is to say, as one elastic sphere striking another. This is easy to picture; but then we are left without any obvious explanation of the fact which was just mentioned—the fact that this sort of rebounding takes place especially often, when the quanta agree in frequency with those which the atom can naturally emit. Finally, for the single case in which the incident frequency agrees with that of a spectrum line and the scattering is very abundant, one can employ a compromise-theory; the atom is struck as by a bullet which sets it to vibrating freely with one of its own natural frequencies, as a bell which is struck by its clapper.

Scattering of light with change of frequency is certainly more complicated. The advocates of waves and oscillators must conceive that in the atom there goes on a process similar to what, in the art of electrical communication, is known as *modulation*. The frequency of the infalling light is modulated with some frequency characteristic of the atom. If the compromise-theory is valid, there are several cases in which one easily sees how this happens. Thus if a straight spring is alternately contracting and expanding with a frequency n_0, and at the same time is revolving around an axis perpendicular to its length with a constant angular velocity $2\pi n_1$, its ends will seem to a stationary observer to be moving with a motion compounded of two frequencies—$(n_0 + n_1)$ and $(n_0 - n_1)$; and if waves are sent out, they will have these frequencies jointly.[2] If the "spring" is an electrical doublet lying perpendicular to a magnetic field, it revolves automatically as it vibrates, and sends forth electromagnetic waves which are discriminated by the spectroscope into two lines of these two frequencies. Such is the explanation by wave-theory of the "normal Zeeman effect"; and while the actual effect of magnetic fields upon the light emitted by the atoms which they influence is not often exactly thus, it is sufficiently nearly so to prove that this interpretation is a step on the right path. If however one wishes to maintain the uncompromising wave-theory, and suppose that the vibrators in the atoms are kept going in forced vibration by the continually-acting waves of light, then modulation does not necessarily occur—not at least with the con-

[2] This idea was introduced by the elder Lord Rayleigh in the course of some speculations on the emission of light by rotating atoms, and was later turned to account in explaining the fine-structure of the bands which constitute the spectra of molecules.

ventional atom-models.[3] Models however can be devised which account for modulation, and perhaps they will become more popular.[4] In one simple case related to the Compton effect, forced vibrations result in waves which do not coincide in frequency with the primary waves; this is the case of a free electron, which the magnetic force in the light-stream pushes more and more rapidly forward as the electric force makes it swing more and more rapidly crosswise; and as the electron gains speed, the frequency of the waves which it sends to a stationary observer steadily sinks.

To the thoroughgoing advocate of the corpuscle-theory, however, the problem of the scattered light of shifted frequency seems simple; or, at all events, the first step in explaining it seems obvious. Frequency of light, when multiplied by the universal constant h, is the measure of the energy of the corpuscles of the light. Change of frequency therefore means transfer of energy. If a quantum of frequency n_0 flies onto an atom and a quantum of frequency n_1 flies away, energy in the amount $h(n_0 - n_1)$ stays behind with the atom. If n_1 is greater than n_0, as sometimes happens, the departing quantum takes with it some energy which belonged to the atom as well as all that was brought by the oncoming quantum. As yet there is no picture of the process by which the energy is passed between the matter and the light. But we are not supposed to ask the quantum-theory for such pictures. Perhaps one reason why it seems so much stronger than the wave-theory is, that of the latter we have expected so much more.

However, visualizable or not, the corpuscle-theory implies that if the scattered light differs in frequency from the infalling light, individual molecules or atoms are receiving or giving energy in quantities equal to the frequency-difference multiplied by h. Not any and every amount of energy may be annexed or ceded by a molecule or an atom—only certain sharply definite, distinct and separate amounts, equal to the energy-differences between the state in which the particle happens initially to be, and one or another of its various other "permitted" stationary states. There are exceptions to the rule, as I will state immediately; but in experiments performed with visible light they are not apparent. Correspondingly the frequency-shifts of the scattered light are limited to certain distinct and separate values;

[3] Thus an electron which is subject at once to a quasi-elastic restoring-force, a sinusoidal electric field, and a constant magnetic field perpendicular to the electric field, describes a fixed orbit with a single frequency equal to that of the electric force, and there is no modulation. Mistakes in this respect have been made by various people who theorized about the Wood-Ellett effect.

[4] Such models have been devised by Hartley and by Kennard.

the spectrum of the scattered light due to a single primary spectrum line consists of that line accompanied by a number of others, separate from it and from one another. Intermediate frequencies do not occur, for they would correspond to transfers of energy in quantities which the atoms are not able to offer or accept. The shifted lines which do occur, the "Raman lines," reveal the energy-values of the stationary states of the scattering particles.

We consider next the exceptions to the rule—the cases in which a scattering particle may accept or surrender any quantity of energy whatever within an appreciable continuous range, instead of merely certain separate discrete amounts. This may be possible if the energy conceded by the quantum is employed in altering the speed of the particle, or in breaking the particle into pieces and imparting speed to these—if it becomes kinetic energy of translatory motion of the molecule or atom, or of the fragments thereof. Translatory motion is non-quantized, which is a way of saying that it is not under the dominion of quantum-conditions which allow to it some values and deny it others. Any amount of kinetic energy of translation is permitted to a molecule or an atom, so far as we know. This suggests that any amount of energy may be transferred when such a particle meets a corpuscle of light, provided that so long as the energy is held by the molecule or the atom it is held in this form. But there is another limitation to be remembered—that imposed by Newton's principle of the conservation of momentum. If a swiftly-moving corpuscle of relatively small mass m strikes a slowly-moving body of much larger mass M, the latter cannot gain much speed in the encounter; for it cannot acquire speed without acquiring momentum, and if it were to accept for that purpose more than a very small fraction of the energy of m, it would have to take more momentum than all that m possesses.

Now relatively to an atom, a corpuscle of light is a body of very small mass and very swift flight indeed; and a quantum of frequency n cannot transfer to an atom of mass M, for use as kinetic energy of translation, more than the fraction $2hn/MC^2$ of its own initial energy—more than the quantity $2h^2n^2/MC^2$ altogether.[5] For a quantum belonging to the visible spectrum the fraction $2hn/MC^2$ is of the order of 10^{-8} even for an impact with the lightest of all atoms. The utmost possible shift in frequency of the scattered light would bear only this proportion to the primary frequency, and would be indistinguishable. But the higher the frequency of the quantum, and the lower the mass of the

[5] The formula is approximate, but the approximation is very close in all practical cases. For the derivation of this and the accurate formula, see A. H. Compton, *Bull. Nat. Res. Council*, 20 (1922) or my *Introduction to Contemporary Physics*, pp. 148–149.

scattering particle, the greater this maximum possible transfer of energy and this maximum possible frequency-shift become; and for a collision between an X-ray quantum and a free electron, it attains the order 10^{-2} of the primary frequency, and is very appreciable. In fact, the frequency-shift occurring when X-ray quanta transfer energy to free electrons and these employ it as kinetic energy was the first of all to be observed. It is simply the Compton effect. It was noticed first towards 1904 and was described as "softening of the scattered X-rays," and in 1922 was for the first time properly measured and properly interpreted by Arthur Compton. The scattered rays include every frequency from that of the primary rays, n_0 let us call it, downward to the lower limit $(1 - 2hn/MC^2)n_0$, as they should.[6]

If the primary quantum breaks the particle which it strikes into two or more fragments—as for instance when an atom is ionized or a molecule dissociated—the requirement of conservation of momentum no longer limits the amount of energy which it may pass to these. It must give at least enough energy to ionize or to dissociate the particle; beyond this, so far as we know *a priori*, any extent of transfer is permitted. Hence we should expect to observe in the spectrum of the scattered light a continuous band, commencing at the frequency which is less than that of the primary light by the quotient of h into the ionizing-potential or the dissociation-potential, and extending towards lower frequencies indefinitely far. More precisely, we should expect to observe as many of these bands as there are modes of dissociation or modes of ionization feasible by light.

No one, so far as I know, has yet observed any bands corresponding to dissociation of molecules or to the detachment of loosely-bound electrons by visible or ultra-violet light. In the X-ray region, however, it is different. In the spectra of scattered X-ray bands answering to this description, and suggesting that X-ray quanta have extracted deep-lying tightly-bound electrons from atoms and have conferred kinetic energy upon them, have in fact been reported. Several such spectra were depicted in 1923 and 1924 by G. L. Clark and W. Duane.

If it should turn out in any special case that quanta could extract electrons from atoms, but could not confer extra kinetic energy of translatory motion on them—a restriction which there is no evident

[6] This is disguised by the fact that the rays scattered in any one direction (relatively to the primary beam) are of a single frequency. If we observed simultaneously rays scattered in all directions, we should see a continuous band of light extending between n_0 and the stated lower limit. This condition was approached, though not purposely, in some of the earlier researches on the Compton effect.

The shift which should occur if quanta of the visible spectrum are scattered by free electrons is very small, but sufficiently large to be appreciable; however, this type of scattering does not seem to occur to a perceptible extent, for it has been sought in vain (P. A. Ross).

reason to foresee—then the aforesaid bands would be reduced to lines, shifted from the primary line through frequency-intervals equal to the ionizing-potentials divided by h. Such lines were observed in the early spring of 1928 by B. Davis and D. P. Mitchell, in the spectrum of X-rays scattered by graphite.

There is one more way in which corpuscles of light can dispose of part of their energy—in setting into vibration atoms which are built into the structure of crystal lattices. Many crystals behave as if they contained oscillators having natural frequencies of the order 10^{12}–10^{14}, and able to emit light of the corresponding wavelengths, which are in the infra-red region of the spectrum. Some of the quanta which strike such a crystal lose energy in being scattered, and the energy which they lose is equal to h times one or another of these oscillation-frequencies. This effect was discovered independently in the late winter or early spring of 1928 by G. Landsberg and L. Mandelstam, and by C. V. Raman and K. S. Krishnan.

Presently I will quote in more detail the data which establish all these facts; but first it is urgent to point out that there is another phenomenon in nature, a phenomenon long known and well known, with which the scattering of light with altered frequency can readily be confused; indeed it is often difficult, and I suspect that it may sometimes be impossible, to tell whether in an actual case we have the one or the other before us. I refer, of course, to fluorescence. The description of fluorescence, indeed, reads exactly like the description of scattering of light with change of frequency. Light of one frequency falls upon a substance, and light of another frequency emerges from it. How then shall we discriminate between the two?

According to the ordinary conception of fluorescence—a conception which has attained to the rank of a definition—the molecule or the atom absorbs a quantum of the incident light, and is put thereby into an excited state; and after a longer or a shorter time, it passes spontaneously into a state different both from the excited and from its original state, and emits a quantum which is not of the same frequency as the one which it absorbed. (I am considering fluorescence of gases or of dilute solutions, where one can suppose that the quanta are absorbed or emitted by individual molecules; more complex cases are too complex, for the time being.) Let the original state be symbolized by N, and the final state by A, and the temporary state by B; denote by $E_{AN} = hn_{AN}$ the energy-difference between A and N, positive if the energy in state A is the greater; use the letters n_{BN} and n_{BA} correspondingly, and denote by n_0 the primary frequency. Then there will be no fluorescence at all unless $n_0 = n_{BN}$; unless, that is to

say, the primary quanta have exactly the right energy to transfer a molecule from its original state to some other (excited) state. Suppose however that this condition is fulfilled; then the quantum of fluorescence-light is emitted when the molecule passes from state B to state A, and therefore has the frequency n_{BA}. But because of the energy-relations, we have

$$n_{BA} = n_{BN} - n_{AN} = n_0 - n_{AN},$$

which means that the quantum of fluorescence-light has exactly the same energy as a quantum of the primary light would retain, if it had been scattered from the molecule after communicating to this latter the energy requisite to transfer it from its original state N to the state A. The fluorescence-light is shifted in frequency from the primary line by exactly the same interval as the Raman line corresponding to the transfer of the molecule from N to A.

It follows then that one can never decide by measurement of wave-length whether a line in the scattered light is due to fluorescence or to scattering with change of frequency.[7] This is inevitable; for in either case the molecule involved in the process starts from the same initial and ends in the same final state, and the frequency of the departing quantum depends on nothing but the difference between these two. The only question at issue is, whether the molecule has gone from the initial to the final state directly, or *via* the temporary state B.

One way of solving the question seems obvious: to vary the frequency of the light with which the molecules are irradiated, and notice whether the shifted line which is under observation—a line is identified by the amount of its shift, not by its actual frequency, so that a given line travels along the spectrum *pari passu* with the primary light—makes its appearance when and only when the quantum-energy hn_0 of the primary beam coincides with the energy-difference between the initial state of the molecule and any one of its other states. In other words: does the shifted line appear only when the primary quanta can be absorbed by the molecules—when the infalling light coincides with a line of the absorption-spectrum of the substance? or does it appear always? If the latter is the case, it is the Raman effect which we have before us, at least when the wave-length of the infalling light is such that its quanta are not absorbed.

If however the shifted line is most intense when the primary fre-

[7] This may be too strongly stated; one might observe a fluorescence-line emitted—to use the foregoing symbols—by reason of the transition from A to N (not that from B to A) which could not be a Raman line correlated with a transfer of the molecule out of the state N. On the other hand it could be a Raman line associated with a transfer of the molecule out of state A, so that one would have to assess the relative likelihoods of the states A and N among the molecules.

quency coincides with a frequency in the spectrum of the scattering substance, it does not necessarily follow that we are dealing with a case of fluorescence. Scattering without change of frequency is very much intensified, when such coincidence is brought about. Experience teaches this, and the wave-theory also; for a vibrator scatters waves most powerfully, when they and it are in resonance together. Scattering with change of frequency may follow the same rule. The proof of fluorescence, then, turns finally on this: can it be shown that there is an interval of time between the moment when the primary quantum impinges on the molecule, and the moment when the secondary quantum leaves it?

Often with solid substances one can actually see that the secondary rays continue to emerge for an appreciable time after the primary rays are discontinued; but with gases no such great delay has so far been observed.[8] However, there are sometimes indications that between the arrival of the primary quantum and the departure of the secondary quantum, there is an interval of time during which something can happen to the molecule or atom—something which changes the nature of the departing quantum, and may even prevent it from ever being born. Pure rarefied gaseous mercury and sodium and iodine, to take three instances, emit light vividly when they are illuminated; but if they are made very dense, the intensity of the emitted light is much reduced, or its spectrum is entirely changed, or both of these things happen; so also when they are mixed with gases such as hydrogen or argon.

Such results, it is clear, are difficult or impossible to explain if the emitted light consists of primary quanta which have rebounded *instantaneously* from collisions with (say) mercury atoms; for such collisions would be more numerous when the gas became denser, and not much less numerous when the gas was diluted with argon; and the rebounding quanta would vary proportionately in number, while the frequency-shifts which they display would not be changed (unless one were to hit two or more atoms in succession). However, if the mercury atoms absorb the primary quanta, and hold on to their energy for a while, and subsequently by some independent process release it, all these effects are quite easy to interpret. The atom which has accepted the energy of a quantum, and has not yet decided to disgorge it in the form of a secondary quantum, may meet another atom and unload the energy in part or altogether. It is certain that this can happen; for when mercury vapor is mixed for instance with thallium

[8] Methods whereby it might be possible to measure the time-interval have been suggested by Ruark.

vapor and the mixture is bombarded with quanta which mercury atoms can absorb and thallium atoms cannot, we nevertheless presently find the thallium atoms possessed of some of the energy which was sent in. Collisions of atoms occur more frequently, the denser and the warmer the gas or the mixture of gases becomes; the opportunities for diversion of energy, which otherwise would be re-radiated as fluorescent light, become correspondingly more numerous. When therefore the light emitted by an illuminated gas changes its spectrum or fades away as the gas is densified or contaminated, the probabilities are that it is true fluorescence-light.[9] The influence of a magnetic field upon the character of the emitted light may also furnish evidence.

One sees therefore that the distinction between scattering and fluorescence is by no means immediate. Even the case which seems most explicit of all—where the primary light agrees with one of the spectrum-frequencies of the atom and the secondary light is unshifted, as when rarefied sodium vapor is illuminated by one of the *D*-lines and re-radiates it—is not exempt from doubt. Very likely part of the re-radiated rays is fluorescence-light and part consists of scattered quanta. It is obvious why Raman thought at first that he was observing fluorescence. Others very likely had already noticed the Raman effect, and classified it merely as another instance of the already well-known phenomenon.

I will now relate some of the details of the recent experiments which have suggested that light may actually be scattered with change of frequency.

The Raman Effect

The scattering of visible[10] light with change of frequency was first discovered by a man who was working with molecular liquids. It is interesting and instructive to consider why the effect, so obvious under these circumstances, had eluded the numerous and notable physicists who had studied—exhaustively, it was thought,—the influences of light on gases and of gases on light.

In the first place, a liquid contains many more molecules per unit volume than a gas, and therefore offers many more opportunities for collisions of quanta with molecules. This is essential, for collisions which result in excitation of the molecule and in scattering of the quan-

[9] I suspect that Saha, in concluding that the resonance-spectra of vapors discovered by Wood are actually examples of scattering with change of frequency, did not take sufficient account of some of these phenomena. Consider for instance those observed by Wood and Loomis (*Jour. Franklin Inst., 205*, pp. 489–495).

[10] I will reserve the name "Raman Effect" for the scattering with shift-of-frequency of light of the visible and adjacent ranges of the spectrum, as in the X-ray region the effect was earlier discovered.

tum with altered frequency are evidently relatively rare; otherwise they could not have escaped the notice of those who have studied gases.[11] Even scattering without change of frequency is unusual, unless the primary light coincides exactly with a spectrum-line of the molecule; the blue of the sky is conspicuous only because the air is so thick; in the laboratory, light scattered with unshifted wave-length by a gas can be seen only if the gas is dense, the primary light blindingly brilliant, and the eye thoroughly rested.

But if it had occurred to any physicist to seek for the effect with (say) mercury atoms, by crowding the atoms together into the liquid form, he would certainly have rejected the idea the moment after it flashed across his mind; indeed it would probably never have flashed; for as

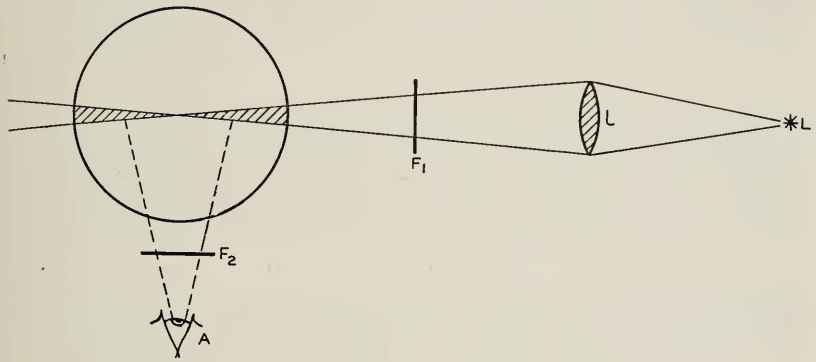


Fig. 1—Sketch of scheme for observing fluorescence and scattering. (After Pringsheim.)

soon as atoms are forced into such close proximity, their excited states, or those at least with which we are now concerned, simply disappear. The *free* mercury atom, for instance, has a "4.9-volt" excited state—that is to say, a stationary state into which it will pass over, if being initially in its normal state it receives an acceptable offer of 4.9 equivalent volts of energy. Were we to bombard mercury vapor with 6-volt quanta—i.e. with corpuscles of light, each possessing 6 equivalent volts of energy—we might expect some of these to transfer 4.9 equivalent volts to atoms which they strike, and rebound as 1.1-volt quanta. But there is no reason to expect anything of the sort with *liquid* mercury; there is no reason to suppose that the atoms are avid to grasp this particular amount of energy, and plenty of reason to suppose that they are not. The same holds for every other excited state of a free atom, of which the energy-excess over the normal state is smaller

[11] Also one would expect to find, in the light scattered from liquids, quanta which have suffered two or more collisions; such have not yet been reported, so far as I know.

than the energy at the disposal of corpuscles of visible or ultra-violet light. Modern atomic theory makes this vanishing of the excited states seem very plausible; for the said excited states correspond to particular arrangements of the electrons at the surface of the atom, which are completely disorganized when atoms are crowded close together. But, apart from theory, it is an experimental fact.

One can therefore scarcely hope to find scattered light with a spectrum composed of discrete separate lines, unless one can find atoms or complexes of atoms of which the low-energy excited states remain discrete, separate and accessible when the substance is liquefied or solidified. With atoms, as I have said, this appears to be impossible. The high-energy excited states, in which one or another of the deep-lying electrons is absent from the atomic system, are indeed the same whether the atoms are free or are crowded together into a solid or a liquid; but corpuscles of light of the visible or the ultra-violet spectrum have not energy enough to excite them, and therefore for the time being they fall out of our purview. Molecules, however, do possess excited states, into which they may be transferred from the normal state by offering them one or two or three, or even a fraction of one equivalent volt of energy; and these they possess, even when jammed together in the liquid state.

These excited states of molecules correspond, according to modern theory, to various amplitudes of vibration of the atom-nuclei relatively to one another within the molecule. The simplest case, of course, is that of the diatomic molecule, of which there are so many examples—oxygen and hydrogen and nitrogen, for instance. The nuclei of the two atoms, being positively charged, repel each other; but the electrons, aided possibly by additional magnetic fields, exert upon each nucleus a force which tends to push them together; and there is a certain internuclear distance of equilibrium, for which the two opposing forces balance. If the nuclei are displaced slightly from their points of equilibrium, they vibrate. Vibration is a quantized form of motion; only certain amplitudes and certain energy-values are permitted. The low-energy excited states of the molecules correspond to the permitted amplitudes and the permitted vibration-energies; when a quantum excites one of these and rebounds with the remainder of its energy, the energy which it gives up is spent in augmenting the vibrations of the nuclei.

These low-energy excited states are responsible for Raman's discovery; partly because, as I have stated, they survive when the molecules are jammed together into a liquid—a fact which evidently means that the electrons, which produce the force upon the nuclei

countervailing their reciprocal repulsion, are shielded from the outer world, presumably by other electrons lying still farther outward from the nuclei; and partly because their energy-values are so conveniently low. This latter point can best be illustrated with an example. To perceive a relatively feeble optical effect, or one which is expected to be feeble, it is best to produce it in the visible spectrum—not merely in order to observe it with the eye; the major reasons are rather, that in

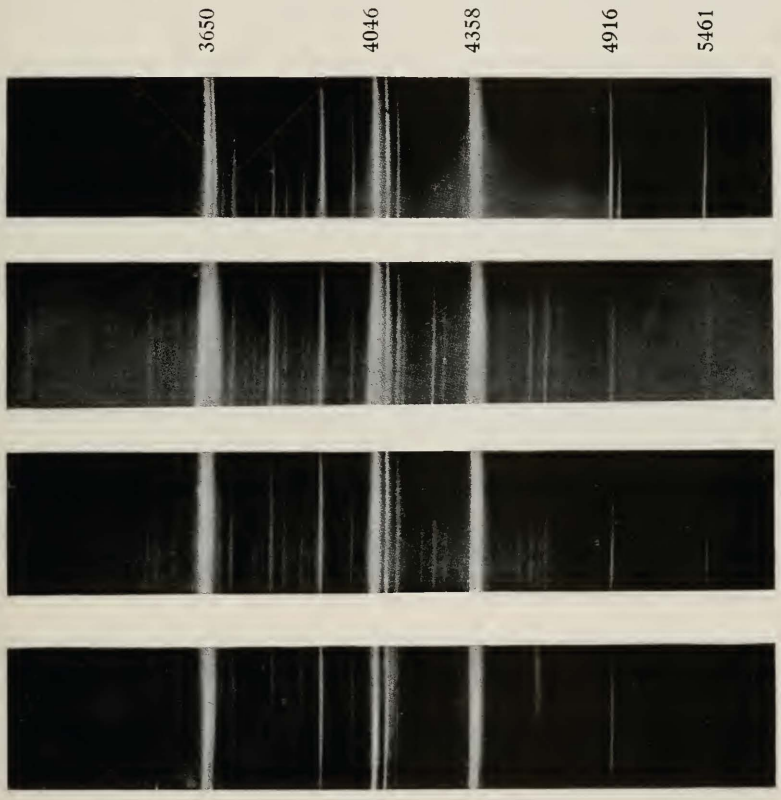


Fig. 2—At top, spectrum of primary light (mercury arc); below, spectra of light scattered by benzene, toluene, pentane respectively. (C. V. Raman; *Indian Journal of Physics.*)

going away from the visible spectrum-range in one direction we find the photographic plates becoming rapidly less sensitive, while in the other direction the transmission of the rays through matter grows steadily worse. Suppose then that one tries to produce the Raman effect by light near the high-frequency limit of the visible—say about 4000 Å., where the quantum-energy is about 3 equivalent volts. If

the scattered quantum in its turn is to be in the visible spectrum, its wave-length must be less than some 8000 Å., its energy more than roughly 1.5 equivalent volts. The primary corpuscle of light must therefore not cede to the molecule or atom more than (3 − 1.5) or 1.5 equivalent volts of energy—the material particle must therefore be able to receive energy in quantities less than this, quantities preferably which are small fractions of an equivalent volt; its excited states should differ from the normal state by energy-differences of this order; and molecules satisfy this condition.[12]

En somme, then, the scattering of light with change of frequency was never discovered in all the abundant work on the common monatomic



Fig. 3—At top, spectrum of primary light, reduced by a filter almost to a single strong line; below, that of light scattered by benzene, showing shifted lines due to that single line (4358). (C. V. Raman; *Indian Journal of Physics.*)

gases, because: first, the atoms were too few in unit volume to cause much scattering; second, they could not be squeezed together without destroying their power of entering distinct excited states; third, the excited states had energy-values so high, that a quantum of the visible or the near ultra-violet spectrum striking a normal atom would simply not have had energy enough to transfer it into one of them. It was discovered by Raman in the way in which he discovered it, because he was working with molecular liquids where: first, the molecules were numerous in unit volume and scattering was frequent; second, in spite of being squeezed together the molecules retained their power of enter-

[12] Atoms however do not always satisfy it; in particular, those of the noble gases and of mercury, the substances most often used in optical experiments on monatomic gases, possess no excited states differing from the normal state by less than four equivalent volts—another reason why the Raman effect was not sooner discovered. The more massive of the alkali metals have excited states superior to the normal by about 1.5 equivalent volts, while metals of the third column of the periodic table would be very favorable.

ing distinct excited states; third, the excited states had energy-values so low that a quantum of the violet, blue or green regions of the spectrum striking a molecule had plenty of energy to excite it and yet have some left over. Subsequently one of Raman's associates (Ramdas), photographing with very long exposure, detected the effect with ether vapor. Carrelli and his colleagues obtained it with molecular salts in aqueous solution;[13] and perhaps in the course of time somebody may overcome the obstacles, and demonstrate the scattering of light with change of frequency by free atoms of a monatomic gas.

We will now consider some of the photographs which were made by Raman, and later with improvements of technique by Wood, Langer and Meggers, Brickwedde and Peters.

Some of Raman's earliest published pictures are reproduced in Fig. 2. At the top we see the spectrum of the primary light—that of the mercury arc, the source of light employed, I think, in all the researches thus far published. The strong lines near 4046 and 4358 are responsible for most of the Raman lines thus far observed by anyone; other strong lines are those near 5461, and the pair at 5770 and 5790. (These last look much fainter in Fig. 2 than in some of the others, but such variations are due to the photographic plates employed, and should not be heeded.) The three spectra below are, in order, those of the light scattered by benzene, toluene and pentane. The new lines are extremely numerous—so much so, that some care is required to determine for each new line which is the primary line whence it is shifted. This may be done by filtering out from the primary light all but one of its strong lines. In making the photographs in Fig. 3, Raman and Krishnan used a filter which removed from the infalling light almost all the quanta but those of the wave-length 4358 (though 3650 and 4046 are still seen dimly in the spectrum, the topmost one in the figure). The spectrum of the scattered light, below, now shows additional lines which are certainly made of quanta which originally had the wave-length 4358.

Figs. 4 and 5 show the spectra scattered by benzene and carbon tetrachloride, as photographed by Wood.[14] The "fat" lines from left to right are the unshifted lines 4046, 4358, 5461 and the aforesaid doublet 5770–5790. (The rich adjacent spectrum is a "comparison" spectrum of iron.) Most of the lines companioning 4358 and 5461 on both sides are shifted lines. Notice, in the spectrum scattered by

[13] With salts which are completely dissociated in solution they failed, as they expected, to obtain it. Possibly the effect may some day be used as a measure of percentage of dissociation!

[14] I am much indebted to Professor Wood for furnishing me with prints of these, and to Dr. Langer for plates from which the next two figures were made.



Fig. 4—Light scattered by benzene (with comparison spectrum). The four fat lines from left to right are 4046, 4358, 5461 and the doublet 5770–5790 of the unshifted scattered light. (R. W. Wood.)



Fig. 5—Light scattered by carbon tetrachloride; note the array of lines shifted each way from 4358. (R. W. Wood.)

carbon tetrachloride, the triad of lines to the right of 4358, and the equally-spaced triad to its left. These latter are "anti-Stokesian" lines—I will presently explain the name—and consist of quanta which have received as much energy from molecules as the quanta of lowered frequencies have given up. With carbon tetrachloride they are extraordinarily bright. Notice again how these lines and the hazy doublet still further out are repeated to the right of 4046, where they are interspersed with other lines which are primary lines unshifted; and again on both sides of 5461.



Fig. 6—Ultraviolet light scattered by sulphuric acid; at the right, 2536 accompanied by numerous shifted lines. (R. M. Langer.)

In Fig. 6 we pass to another region of the spectrum, the ultra violet. On the extreme right is the strong line 2536 of the mercury spectrum, scattered unshifted by sulphuric acid; the numerous lines beside it are Raman lines.

In Fig. 7 the scattering substance is water; the novel feature of the scattered light is a set of diffuse bands, each shifted from a certain line of the primary spectrum. It is frequently observed that the shifted lines, and the unshifted lines as well, are broader and hazier than those of the primary light, an effect attributed to conversion of the energy of the quanta into energy of rotation of the molecules. Water however shows a quite remarkable broadening, if this be the



Fig. 7—Light scattered by water; note the diffuse bands shifted from the sharp lines. (R. M. Langer.)

way to describe it; it has been attributed to association of molecules.

A few special words must be said about the shifted lines which are of higher frequency than the primary light responsible for them—not because they are really more remarkable than the others, but because somehow they seem less to be expected. They are called by the monstrous name "anti-Stokesian" because in fluorescence such lines run counter to a principle laid down by Stokes; it seems cruel to perpetuate a mistake in this way. They consist of quanta which have

received energy from the molecules which they have struck. They must therefore have struck molecules which were not originally in the state of lowest energy, or "lowest state" for short. Such a molecule might have been in the lowest state, until some antecedent quantum came along and gave it energy, lifting it into a higher state; the second quantum then undid the work of the first, taking away the energy which the first had given and being shifted as far towards higher as the former towards lower frequencies. It does not seem likely that this double process occurs often, although pairs of lines with equal and opposite shifts are reported in several cases. Much more commonly, in all probability, the molecules which at any moment are in other states than the lowest are there because of the interchanges of energy which are always taking place between the particles of substances in thermal equilibrium. The laws of thermal equilibrium are such, that if in a substance at room-temperature the molecules have one or more excited states differing from the lowest state only by fractions of a volt, quite an appreciable fraction among them are at any moment in one or another of those states. The higher the temperature, the greater this fraction; in consonance with which fact it is observed, that the warmer the scattering liquid the more prominent are these "anti-Stokesian" lines.

The shifted lines, in the light scattered at right angles to the primary beam (the only direction which has been utilized to any extent), are partially polarized. The electric vector is stronger in the direction perpendicular to the primary beam than in the direction parallel to it, as one would expect. The degree of polarization varies enormously from one shifted line to another, and may be either greater or less than that of the unshifted lines. Cabannes thought it to be constant for lines shifted by the same amount from different primary lines, but the ampler data of Carrelli do not seem to bear him out.

The test for "coherence" of the shifted light, which is made by bringing the scattering substance near to the state of "critical opalescence" where the irregularity of the arrangement of the atoms is greatest—the shifted lines should brighten *pari passu* with the unshifted, if their light is coherent—has been made by at least four people (Raman, Bogros and Rocard, Martin); but the results are oddly discordant.

One more most valuable service of the shifted lines remains to be mentioned. In the foregoing pages I have stressed the fact that some of them are known to agree, that is to say their frequency shifts are known to coincide, with the frequencies of lines of the infra-red spectrum of the scattering substance. But there are cases in which the

infra-red spectrum of the scatterer has not yet been explored; and there we may deduce its lines from the frequency-shifts of the Raman lines in the visible spectrum. Moreover, there are regions of the infra-red spectrum which are very difficult to explore, because of such technical reasons as the insensitiveness of photographic plates; and the Raman lines make it possible to discover some at least of the features of these, by observations made in the most convenient region of the spectrum. Perhaps this will turn out to be the most fruitful of the consequences of Raman's discovery.

Scattering of Light with Transfer of Energy to Vibrations in Solids

The scattering of light with shift of frequency from solids was discovered, independently and almost simultaneously, by C. V. Raman and K. S. Krishnan in India and by G. Landsberg and L. Mandelstam in Russia. As seen on the photographs of the spectra of the scattered light, the effect is altogether like the Raman effect of liquids and vapors. The lines of the primary spectrum, scattered without change of frequency, are accompanied by companions shifted mostly towards lower, but in occasional cases towards higher frequencies. The infalling quanta therefore sometimes cede energy to quantized motions within the solid substance, and sometimes—but much less frequently—receive energy from these.

The first and obvious question is: do the frequency-shifts agree with lines of the infra-red spectra of the solid substance? For studying the infra-red spectra of solids there are, be it remembered, two classical methods. One is the familiar way of dispersing a beam of light which has traversed the solid, and looking for absorption lines or bands in its spectrum. To find a good dispersing-agent in the far infra-red is however not easy; and there is an alternative method, in which the beam of light is reflected several times over from samples of the solid. At each incidence of the beam upon the crystal, the waves of frequencies which do not coincide with natural frequencies of the substance go on through, while the waves of frequencies which do coincide are mostly reflected. Thus, after several reflections, the "residual" beam is composed of one or a few wave-lengths, those of the principal absorption-lines of the crystal; and these are measured by operating on the beam with a special interferometer, or in some other way. This is the method of "residual rays," or "Reststrahlen," which was developed and much exploited during the nineties of the last century and the opening years of this. The spectrum-lines of the crystalline substance are its "Reststrahlen"; and these are to be compared with the fre-

quency shifts observed in and near the visible spectrum. Here is an instance: Landsberg and Mandelstam working with quartz observed frequency-shifts corresponding to infra-red lines of wave-lengths 9μ, 13.5μ, 21.5μ, 48μ, and 81μ respectively; there are Reststrahlen of wave-lengths 8.7, 12.8 and 20.7, while the other two wave-lengths cited lie in gaps of the infra-red spectrum unexplored as yet.

A photograph by F. G. Brickwedde [15] which I reproduce as Figure 8 illustrates this effect. The very broad black band is due to primary light of wave-length about 2536, scattered without change of fre-

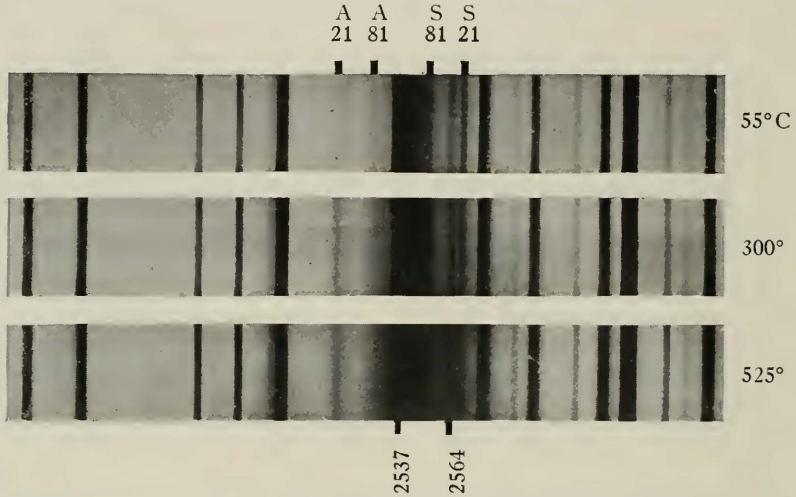


Fig. 8—Light scattered by a quartz crystal; unshifted line is 2536, lines denoted by S and A are shifted towards lower and higher frequencies respectively. (F. G. Brickwedde.)

quency; its excessive width is due to the great intensity of this light. The lines marked S21 and S81 consist of quanta which have spent part of their energy in exciting vibrations, those corresponding to the Reststrahlen of wave-lengths 21μ and 81μ respectively. The lines marked A21 and A81 consist of quanta which have received energy from these vibrations. These "Anti-Stokesian" lines evidently grow more intense as the temperature of the crystal is raised, as they should, for the reason which I have already stated: as the substance grows warmer, the percentage of the molecules spontaneously in vibration is increased. In addition, the shifted lines all move inward toward the position of the unshifted light as the temperature rises—one may see this by comparing those on the low-frequency side with

[15] I am much obliged to Dr. Brickwedde for furnishing me with a print of this photograph.

the line marked 2564, which is a faint line of the primary light scattered without change in frequency. This diminution in the shift signifies that the natural frequencies of the lattice-vibrations are declining, which is to be expected from the expansion and relaxation of the crystal which the rise of temperature brings about.

There arises now an interesting question. X-ray analysis reveals that there are certain chemical compounds, organic chiefly, of which the molecules retain their identity when crystallization occurs; they set themselves side by side in a regular lattice, but the arrangement of the atoms in each of them is not greatly altered. At the other extreme, there are compounds of which the molecules disintegrate completely when solidification takes place, and the atoms arrange themselves without any reminiscence of their earlier relations; a familiar instance is sodium chloride, in the crystal of which every atom of either kind, Na or Cl, is surrounded by six of the other kind all equally distant from it. Intermediate cases occur, as for instance that of $CaCO_3$, where each atom-group or "radical" CO_3 retains its identity but not its coupling to one single Ca atom. Now when molecules or radicals survive within the crystal, oscillations of atoms inside these atom-groups are probably not different in character from the oscillations which occur within the same molecules when they are wandering freely in a liquid or a gas. But in the case of a crystal like sodium chloride, the oscillations must be controlled by the forces which hold the atoms of the crystal together; they are truly lattice-vibrations. Perhaps the difference between the two is not really profound; but it will be interesting to find out whether quanta may or may not transfer energy with equal ease to vibrations of either type, *i.e.* whether in the two cases the shifted lines are comparably bright. According to Carrelli, Pringsheim and Rosen, all of the shifted lines thus far observed with solids, except probably those obtained with quartz, correspond to vibrations within molecules or radicals which remain intact in the crystal.

The Compton Effect

The Compton effect, in the restricted sense—the sense in which I shall use the term—is simply the scattering ensuing on collisions of corpuscles of light with free electrons. So much has been written about the effect [16] that it is scarcely necessary for me to do more in this place than mention the laws of these collisions. The energy which the quantum loses is converted into kinetic energy of translatory

[16] Cf. for instance this Journal, April, 1925; "Introduction to Contemporary Physics," pp. 146–160; H. Kallmann, H. Mark, "Ergebnisse der exakten Naturwissenschaften," *5* (1926).

motion of the electron. Momentum also is conserved in the encounter; the momentum of the quantum is equal in magnitude to hn/c before and to hn'/c after the collision, n and n' standing for the frequency before and after; these momenta are of course vectors parallel to the directions along which the corpuscle of light approaches and recedes, respectively; and their difference is the momentum which the electron acquires. These two conditions limit very severely the transfer of energy from quantum to electron. All of the quanta deflected through a given angle from their original line of flight suffer the same loss of energy and the same shift in frequency. The relation between shift of frequency and angle of deflection θ takes its simplest form when we write it as a relation between shift of wave-length, $\Delta\lambda$, and angle θ:

$$\Delta\lambda = \frac{h}{mc}(1 - \cos\theta).$$

The maximum shift of wave-length occurs when the quantum is reflected straight backward along its original line of approach; it is evidently $2h/mc$, and the corresponding maximum frequency shift is $2hn^2/mc^2$, as I stated earlier.

The predicted relation of wave-length shift and angle of scattering has been verified to the most thoroughgoing extent;[17] and the recoiling electrons have been observed as they dash off with the energy which the corpuscle of light has lost. Research on the Compton effect is now confined almost entirely to the problem of its likelihood of occurrence—*i.e.*, given a substance with P atoms per unit volume irradiated by a stream of X-rays composed of N quanta per unit area per second, what are the relations between the number of these quanta which are scattered in the fashion just described, and the frequency of the X-rays and the nature of the atoms? The corresponding problem for the Raman effect will undoubtedly soon come into the foreground. It is, of course, slightly annoying that we do not know *a priori* how many of the electrons of (say) the carbon atom, or how many of the electrons in a piece of graphite, are to be regarded as "free" electrons. This seems to be one of the facts which we shall have to deduce as best we may from these researches on the likelihood of the Compton effect. The data thus far acquired may be summarized in this way: the higher the frequency of the quanta, and the lower the mass of the atoms, the more abundant these collisions are. High-frequency X-rays poured

[17] Even to the point where with general assent it is taken for granted and the measurements are used as data for evaluating the constant m, the mass of the electron!

upon lithium or paraffin give the best "yield"; visible light, none perceptible.

The shifted X-rays scattered at 90° are polarized, very nearly completely (Lukirsky, Kallmann and Mark).

The Scattering of Quanta Attended by Extraction of Bound Electrons

Almost immediately after Arthur Compton had measured and interpreted the scattering which is due to collisions of quanta with free electrons, he and others realized that corpuscles of light might possibly encounter atoms in such a way that they extracted bound electrons, and thereupon were scattered with a corresponding abatement of their energy. In fact it seemed most probable that the electrons responsible for the Compton effect were themselves not quite free, but very lightly bound; and that a careful study of the scattered quanta, the "shifted" or "modified" line, would reveal that they had spent energy in dissolving the bonds as well as in imparting kinetic energy to the electrons. Researches on this topic were numerous, and are still continuing. At that time, the Raman effect had not yet been discovered; and perhaps it did not seem natural to accept the transfer of energy from corpuscles of light to atoms as a *general* phenomenon, apart from special cases so easy and so beautiful to visualize as the elastic impacts of quanta against free electrons. At all events, when in 1923 and 1924 data were published by G. L. Clark and W. Duane which to the present-day onlooker seem to declare the effect in the most forthright fashion, they made no such impression.

The experiments of Clark and Duane were involved in a long controversy, in which the reality even of the Compton effect was called into question. Data were obtained by some experimenters, which others could not or at least did not reproduce. The questions at issue speedily reached the point, where no outsider could risk a judgment unless he was himself a great expert in the study of X-ray scattering. Unfortunately the experiments were terminated, when the reality of the Compton effect was established. I say "unfortunately," for it now seems as if in the nature of things both sides must have been right. Compton had discovered the transfer of energy from quanta to free or nearly free electrons; Clark and Duane must have discovered the transfer of energy from quanta to bound electrons, or the process in which a corpuscle of light uses part of its energy in ionizing an atom, part in giving speed to the liberated electron, and retains the remainder. It would be a very desirable result of the present-day revival of interest in scattering, if somebody should reinvestigate this entire field.

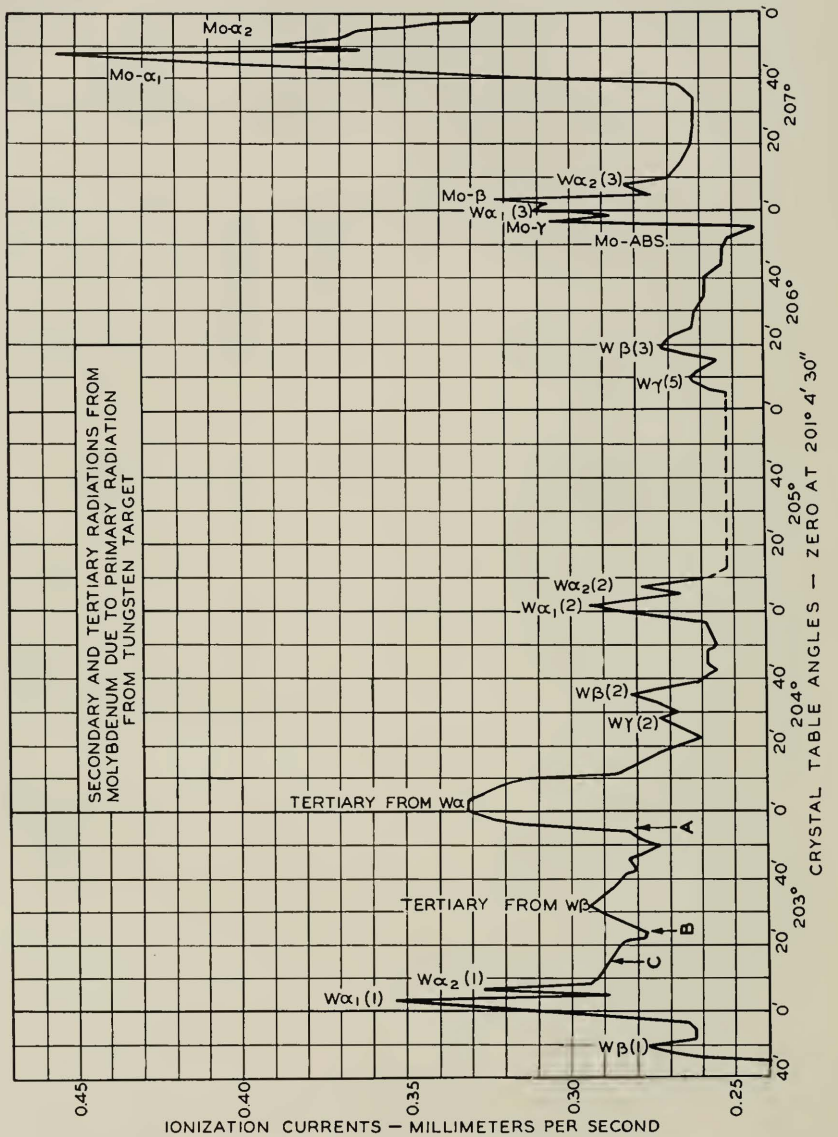


Fig. 9—X-rays scattered by molybdenum, the sharply-pointed peaks being unshifted lines. (G. L. Clark, W. Duane; *Proc. Nat. Acad. Sciences.*)

I will pass over the theory first propounded for these data, which I have quoted elsewhere, and over the paper in which Compton advanced practically the present theory; and reproduce one of the spectrum-curves of scattered radiation from the work of Duane and Clark. This curve represents the scattering of high-frequency X-rays by molybdenum. The primary light was composed of several lines, the various lines of the K-spectrum of tungsten. In the scattered spectrum, these lines appear unshifted as narrow sharp peaks. We consider the two at the extreme left, and then the two humps which rise from the points marked B and A. These points correspond to the wave-lengths of quanta which originally belonged to the spectrum-lines at the extreme left, and have lost exactly the energy necessary to extract a K-electron from the molybdenum atom. The humps in all probability are composed of quanta, which have extracted such electrons and in addition have given them greater or smaller amounts of extra kinetic energy.

Years after the work of Clark and Duane had been discontinued, Bergen Davis and D. P. Mitchell undertook to study what they designated as the "fine structure" of the lines in the spectra of scattered X-rays. They had many improvements of technique at their disposition, improvements many of which were due to Davis himself; for instance they had a spectrometer of Davis' design, by which it was possible to appreciate the true narrowness of a very narrow X-ray line, instead of having it spread out by defects of the apparatus into a simulation of a wide band. They irradiated graphite with the spectrum-line known as $K\alpha_1$ of molybdenum, of which the wave-length is 0.721 Å.; and in the spectrum of the rays scattered at 90° they found not only this line, but four others of slightly greater wave-lengths of which three are shown in Fig. 10. The outermost, beyond the right-hand limit of the picture, comprises quanta which collided with free electrons —it is the Compton shifted line. The outermost of the other three consists of incident quanta which have given up just the amount of energy required to extract one of the K electrons—one of the pair which are by far the most tightly bound of all the six which belong to the carbon atom. More precisely, the extraction energy of these electrons is evaluated from the X-ray spectrum of carbon at 287 equivalent volts, while the loss of energy suffered by the quanta is estimated by Davis and Mitchell at 279; the difference of less than 4 per cent is within the uncertainty of experiment. The two innermost of the shifted lines, composed as they are of quanta which have yielded up 29 and 50 equivalent volts respectively, are presumably tokens of collisions in which superficial electrons were torn away from carbon

atoms. That their shifts do not agree very well with the values of extraction-energies suggested by certain spectroscopic data is not in the least surprising. None of the spectroscopic data was obtained with solid carbon; and the superficial electrons of the atom are most sensitive of all to such changes of environment as occur when it is incorporated into a lattice. In all probability, the frequency-shifts of these lines, when divided by h, are the best values yet available for the amounts of energy required to extract superficial electrons from carbon atoms in the graphite lattice.[18]

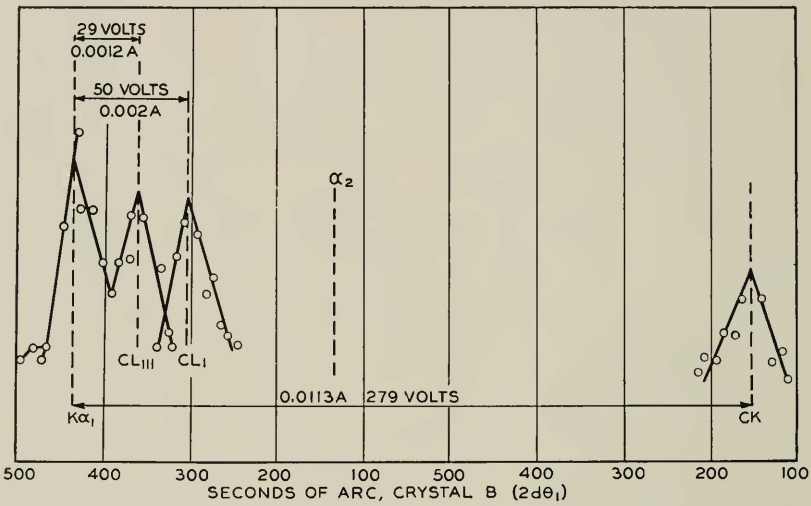


Fig. 10—X-rays scattered by graphite, the peak on the left being the unshifted line. (B. Davis, D. P. Mitchell; *Physical Review.*)

Davis and Mitchell point out a feature of these scattered rays which should be stressed: the shifted lines are sharp, implying that the quanta simply extract electrons, without endowing them with divers quantities of kinetic energy to boot. They later observed such lines in the rays scattered by aluminium.

The General Principle

The general principle of which all these scattered phenomena of scattering are special illustrations now stands forth very clearly, and

[18] It seems paradoxical that the quanta which collide with free electrons should confer on them several hundred of equivalent volts of energy, while those which strike these loosely-bound superficial electrons should lose but half-a-dozen. The contrast is due to the requirement of conservation of momentum; where only two particles are involved the relatively large energy-transfer is entailed, but where there are three among which the momentum may be divided, the limitation ceases.

may be expressed in any of several ways, laying greater emphasis now on one aspect and now on another.

From the viewpoint of the atom, and using the notions of the undulatory theory, one may say that *the atom (or the molecule) modulates the incident light with frequencies of its own.*

Again from the viewpoint of the atom, but using now the notions of the corpuscle-theory of light, one may say that *the atom or molecule may take part but not necessarily all of the energy of an incident quantum, converting this energy in any of numerous ways.*

From the viewpoint of the quantum, however, the essential feature of the principle is this: *a quantum may lose part of its energy or receive energy in an encounter with a molecule or atom, retaining its identity even though its frequency is changed.*

The first who stated the principle with anything like its proper generality was probably Smekal; in the following year (1924) it was developed by Kramers and Heisenberg. They knew of no examples but the Compton effect, and curiously enough no one was tempted to search for other instances, though Foote and Ruark considered whether any of the phenomena already known in optical spectra could be related to it.[19] Partial anticipations crop out here and there, especially in the work of Compton, Jauncey and their associates; for it was early suspected that the electrons responsible for the Compton shift are not altogether free, but very loosely bound to atoms; it was assumed that the incident quanta must spend energy enough to break the bonds as well to set the electrons into motion, and efforts were made to disclose this breaking of the bonds.[20]

It is the third of the foregoing formulations of the principle which I wish to stress in closing—the principle from the viewpoint of the quantum, the authorization of the quantum to give up part of its energy and retain the rest. To the unprejudiced mind this must seem very natural. We have accepted for years the principle that an electron may give up part of its energy and keep the rest—that the life-history of an electron is an endless sequence of gains and losses of kinetic energy, of speedings-up and slowings-down, during which the identity of the electron is never lost. Why should we not have thought likewise about the quantum? Yet it has been almost an article

[19] I am told that Kramers tried vainly to persuade a number of experimental physicists to look for the effect. At present they must be feeling like the astronomers whom Adams vainly pressed to make haste in looking for the planet Neptune, until finally someone else discovered it.

[20] Jauncey and Compton anticipated in 1927 the idea that atoms in a lattice may acquire energy of vibration from incident quanta, and discovered an important restriction which should be noted; apparently the lattice or some third particle must be involved in the impact.

of faith that a quantum must give all of its energy, or none—either vanish altogether, or retain its frequency unchanged.

Of course, till 1922 there was no compelling evidence that a corpuscle of light may suffer a change of frequency in rebounding from a particle of electricity or matter. However, it does not seem to have occurred to anyone that the want of evidence was in any way surprising, or that it should be possible to find quanta scattered with change of energy. The reason for this satisfaction, I suspect, was perfectly simple. It did not seem possible that a quantum should give up part of its energy, for its energy was inseparably linked with its frequency, and its frequency seemed to be its one indissoluble and characteristic feature. As well say that an electron might lose part of its charge and still be the same electron, or that an atom might lose part of its mass and yet remain the same atom, as that a quantum might give up part of its frequency without ceasing to be itself!

Now this contention—if one may call it a contention—lost its force through the discovery that electrons also are endowed with frequency and wave-length, or in other words that negative electricity like light possesses both qualities of corpuscles and qualities of waves. Whenever a corpuscle of electricity parts with kinetic energy, whenever a corpuscle of light parts with energy, the associated wave-length is augmented. If we suppose that an electron retains its identity when its wave-length changes, how can we deny like continuity of existence to a quantum? If we admit that an electron may suffer change of wave-length in rebounding from an atom, how may we be surprised when a quantum does the like? It is true that the corpuscle of electricity has other features than wave-length: a charge which apparently never changes, a mass which apparently never falls below a certain minimum. The quantum does not have an immutable quality corresponding to charge, and we do not know of any lower limit to its mass short of complete disappearance. But for either sort of corpuscle the wave-length is in principle variable. We say that all electrons are of one kind, but may have any speed. Should we not also say that all quanta are of a single kind, though they may have any frequency?

BIBLIOGRAPHY

(The year is 1928, unless otherwise stated)

S. K. Allison, W. Duane: Proc. Nat. Acad. Sci., *10*, pp. 196–199 (1924).
J. A. Becker: Proc. Nat. Acad. Sci., *10*, pp. 342–345 (1924).
C. E. Bleecker: Zs. f. Phys., *50*, pp. 781–786.
A. Bogros, Y. Rocard: C. R., *186*, pp. 1712–1713.
M. Born: Naturwiss., *16*, p. 673.
F. G. Brickwedde, M. F. Peters: Bull. Amer. Phys. Soc., Nov. 15.
M. de Broglie: C. R., *187*, p. 697.
J. Cabannes: C. R., *186*, pp. 1201–1202, 1714–1715; *187*, pp. 654–656.

J. Cabannes, P. Daure: C. R., *186*, pp. 1533–1534.
A. Carrelli, P. Pringsheim, B. Rosen: Zs. f. Phys., *51*, pp. 511–519.
G. L. Clark, W. Duane: Proc. Nat. Acad. Sci., *9*, pp. 413–424 (1923); *10*, pp. 41–47, 92–96, 148–152, 191–196 (1924).
A. H. Compton: Phys. Rev., (2) *24*, pp. 168–177 (1924).
A. Cotton: C. R., *186*, pp. 1475–1476.
B. Davis, D. P. Mitchell: Phys. Rev., (2) *31*, p. 1119; *32*, pp. 331–335; Bull. Am. Phys. Soc., Dec. 12.
P. Daure: C. R., *186*, pp. 1833–1835; *187*, pp. 940–941.
K. K. Darrow: Science, *68*, pp. 488–490.
P. D. Foote, A. E. Ruark: Science, *61*, pp. 263–264 (1925).
R. V. L. Hartley: Bull. Am. Phys. Soc., Dec. 12.
G. E. M. Jauncey, A. H. Compton: Nature, *120*, p. 549 (1927).
H. Kallmann, H. Mark: Zs. f. Phys., *36*, pp. 120–142 (1926).
E. H. Kennard: Bull. Amer. Phys. Soc., Dec. 12.
H. Kornfeld: Naturwiss., *16*, p. 653.
H. A. Kramers, W. Heisenberg: Zs. f. Phys., *31*, pp. 681–706 (1925).
H. S. Krishnan: Nature, *122*, pp. 477–478, p. 650.
G. Landsberg, L. Mandelstam: Naturwiss., *16*, pp. 557–558; C. R., *187*, pp. 109–111.
R. M. Langer, W. F. Meggers: Bull. Amer. Phys. Soc., Nov. 15.
F. A. Lindemann, T. C. Keeley, N. R. Hall: Nature, *122*, p. 921.
P. Lukirsky: Nature, *122*, pp. 275–276.
W. H. Martin: Nature, *122*, pp. 506–507.
P. Pringsheim: Naturwiss., *16*, pp. 597–605; Zs. f. Phys., *50*, pp. 741–755.
I. Ramakrishna Rao: Indian Jour. of Phys., *3*, pp. 123–130.
C. V. Raman: Indian Jour. of Phys., *2*, pp. 387–398.
C. V. Raman, K. S. Krishnan: Nature, *121*, pp. 501–502, 619, 711; *122*, pp. 12–13, 169, 278; Indian Jour. of Phys., *2*, pp. 398–420.
L. A. Ramdas: Nature, *122*, p. 57; Indian Jour. of Phys., *3*, pp. 131–136.
Y. Rocard: C. R., *186*, pp. 1107–1109; Annales de Physique, *10*, pp. 116–232.
A. E. Ruark: Nature, *122*, pp. 312–313.
M. N. Saha, D. S. Kothari, G. R. Toshniwai: Nature, *122*, p. 398.
A. Smekal: Naturwiss., *11*, pp. 873–875 (1923); *16*, pp. 612–613 (1928).
S. Venkateswaran: Indian Jour. of Phys., *3*, pp. 105–122.
D. L. Webster: Proc. Nat. Acad. Sci., *10*, pp. 191–196 (1924).
R. W. Wood: Nature, *122*, p. 349; Phil. Mag., (7) *6*, pp. 729–742.

Ground Return Impedance: Underground Wire with Earth Return

By JOHN R. CARSON

SYNOPSIS: In certain transmission problems principally those relating to induction and interference phenomena, it is necessary to know the transmission characteristics of a circuit composed of an underground wire with earth return. These can be evaluated by well known engineering formulas provided the ground return impedance is known. The present paper gives the mathematical solution of this problem and shows that the ground return impedance is substantially independent of the depth of the wire below the surface.

THE object of this note is to give the solution of a problem of considerable interest and practical importance which does not appear to have been solved heretofore; this is the "ground return" impedance, per unit length, of a circuit composed of an underground wire or cable with earth return.

The physical system and the problem may be more explicitly described and explained as follows: An underground wire or cable parallel to and at depth h below the surface of the ground is surrounded by a concentric dielectric cylinder of external radius a. The earth then forms the return path for currents flowing in the wire. The ground return impedance Z_g is then defined as the ratio of the mean axial electric intensity at the external surface of the dielectric sheath to the current flowing in the wire.

When the earth extends indefinitely in all directions about the wire so that circular symmetry obtains, the problem is quite simple, and the formula for the ground return impedance, denoted in this case by Z_g^0, is well known. In practice, however, we are interested principally in the case where the wire is close to the surface of the earth, so that the distribution of return current in the ground is anything but symmetrical. For this case the formula for the ground return impedance, which it is the object of this note to state and discuss, is

$$Z_g = (1 + c)Z_g^0. \tag{1}$$

Here the correction term c, which takes care of the departure from circular symmetry, is given by

$$c = \frac{1}{K_0(a'i\sqrt{i})} \int_0^\infty \frac{\sqrt{\mu^2 + i} - \mu}{\sqrt{\mu^2 + i} + \mu} \cdot \frac{e^{-2h'\sqrt{\mu^2+i}}}{\sqrt{\mu^2 + i}} d\mu. \tag{2}$$

In formula (2),

$a' = a\sqrt{\alpha}$,

$h' = h\sqrt{\alpha}$,

$\alpha = 4\pi\lambda\omega$ where λ is the conductivity of the ground in elm. c.g.s. units, ω is 2π times the frequency,

$i = \sqrt{-1}$.

$K_0(a'i\sqrt{i})$ is the Bessel Function of the second kind; it is equal to $\frac{i\pi}{2} H_0^{(1)}(a'i\sqrt{i})$ where $H_0^{(1)}$ is the Hankel function as defined by Jahnke u. Emde in their Funktionentafeln. Denoted by ker $a' + i$ kei a' the function $K_0(a'i\sqrt{i})$ has been computed and tabulated by the British Association. The only restriction on formula (2) is that the radius a is supposed small compared with the depth h.

Now the ground conductivity λ lies between 10^{-14} and 10^{-12}, while the depth h will not in practice exceed a few meters ($h \leq 10^3$). Under such circumstances, at ordinary frequencies, h' will be exceedingly small compared with unity, and a' still smaller. Consequently in evaluating the infinite integral in (2), it is permissible to take $e^{-2h'\sqrt{\mu^2+i}}$ as unity, since for $\mu > 2$, the rest of the integrand converges as $i/4\mu^3$.

Now we have

$$\int_0^\infty \frac{\sqrt{\mu^2+i} - \mu}{\sqrt{\mu^2+i} + \mu} \cdot \frac{d\mu}{\sqrt{\mu^2+i}} = \frac{1}{i}\int_0^\infty \left\{ \sqrt{\mu^2+i} - 2\mu + \frac{\mu^2}{\sqrt{\mu^2+i}} \right\} d\mu = \frac{1}{2}$$

and hence c of formula (2) becomes

$$c = \frac{1}{2} \frac{1}{K_0(a'i\sqrt{i})}.$$

Furthermore since a' by hypothesis is very small compared with unity, we can replace K_0 by its limiting form for vanishingly small arguments which is approximately

$$\log (1/a').$$

We thus get, finally, the approximate formula, valid for most practical applications,

$$Z_g = \left\{ 1 + \frac{1}{2 \log (1/a')} \right\} Z_g^0. \tag{3}$$

The interesting and somewhat surprising feature of this formula is that the value of the correction term $1/2 \log (1/a')$ likely to occur in

practical applications amounts at most to 0.05 to 0.10. On the other hand, with the wire close to the surface of the ground, the conducting area of the ground return path is just one half the area available when the ground extends indefinitely in all directions and the return impedance is Z_g^0. In other words, the departure from circular symmetry means only a very small increase in the ground return impedance. In fact this increase is so small and the ground conductivity actually so variable, the correction is hardly justified by the precision of the data, so that, in most engineering applications, we may take Z_g as equal to Z_g^0 with an error probably less than that involved in other factors, and lack of precision in data.

Derivation of Preceding Formulas

The derivation of the preceding formulas is not without interest. Since, however, this derivation is, in general, an adaptation of the methods employed in my paper 'Wave Propagation in Overhead Wires with Ground Return' (*B. S. T. J.*, Oct., 1926) it will be outlined rather than given in detail.

Take the axis of the wire as the origin and Y as the vertical axis; then the surface of the ground is the plane $y = h$. Let a unit current flow in the wire and take the axis of the wire as the Z axis. In the ground ($\rho = \sqrt{x^2 + y^2} \geqq a$) the axial electric intensity will be written

$$E = \frac{Z_g^0}{K_0(a'i\sqrt{i})} K_0(\rho'i\sqrt{i}) + E' = E^0 + E', \tag{4}$$

where $\rho' = \sqrt{\alpha}\sqrt{x^2 + y^2}$ and K_0 is the Bessel function of the second kind, related to the Hankel function by the equation

$$K_0(a'i\sqrt{i}) = \frac{\pi i}{2} H_0^{(1)}(a'i\sqrt{i}).$$

The first term on the right hand side of (4) represents the circularly symmetrical distribution which would alone exist if the surface of the ground were removed to an infinite distance, while E' is a secondary distribution due to reflection at the surface of the earth ($y = h$). Inspection of equation (4) shows that when $\rho = a$, E is the required return impedance Z.

Strictly speaking E^0 should be written as

$$E^0 = \frac{Z_g^0}{K_0(a'i\sqrt{i})} \{K_0(\rho'i\sqrt{i}) + h_1 K_1(\rho'i\sqrt{i})\cos\theta + h_2 K_2(\rho'i\sqrt{i})\cos 2\theta + \cdots\}, \tag{5}$$

the harmonic terms representing secondary reflection at the surface of the dielectric cylinder ($\rho = a$). If a is made sufficiently small, however, the harmonic terms become negligible. In view of this fact, the large amount of tedious additional analysis required, if the harmonic terms are retained, is not believed to be justified by the practical applications contemplated. E^0 will therefore be taken as in formula (4).

The secondary electric intensity E' can always be written as the Fourier integral

$$E' = \int_0^\infty F(\mu) e^{y' \sqrt{\mu^2+i}} \cos x'\mu \, d\mu, \quad 0 \leqq y \leqq h, \quad (6)$$

where $x' = x\sqrt{\alpha}$, $y' = y\sqrt{\alpha}$, and the Fourier function F is to be determined. For the formulation of the boundary conditions at $y = h$ we also require the expansion of $K_0(\rho' i\sqrt{i})$ as a Fourier integral; the required expansion is *

$$K_0(\rho' i\sqrt{i}) = \int_0^\infty \frac{1}{\sqrt{\mu^2 + i}} e^{-y' \sqrt{\mu^2+i}} \cos x'\mu \, d\mu, \quad \rho > 0. \quad (7)$$

In the dielectric, the magnetic forces H_x, H_y are taken as

$$\begin{aligned} H_x &= \int_0^\infty \phi(\mu) e^{-y\mu} \cos x\mu \, d\mu \\ H_y &= -\int_0^\infty \phi(\mu) e^{-y\mu} \sin x\mu \, d\mu \end{aligned} \qquad y \geqq h. \quad (8)$$

In the ground, on the other hand, we have

$$\begin{aligned} i\omega H_x &= -\frac{\partial}{\partial y} E, \\ i\omega H_y &= \frac{\partial}{\partial x} E. \end{aligned} \quad (9)$$

In order to satisfy the boundary conditions at the plane $y = h$, we equate H_x as given by (8) and (9), and H_y as given by (8) and (9). The explicit formulas for H_x and H_y are derived from (9) by substituting the Fourier integral for K_0, as given by (7), in (4) and differentiating as indicated.

The two equations resulting from equating the two expressions for H_x and the two expressions for H_y can be solved for the Fourier function

* See note at end of this paper.

$F(\mu)$. With this determined the required impedance Z_g is simply by (4)

$$Z_g = Z_g^0 + E_{\rho'}' = 0 \tag{10}$$

on the assumption that $a' = a\sqrt{\alpha}$ is quite small compared to unity. This gives formula (1).

Note: The expansion (7), which is believed to be novel, was derived by a limiting process rather too long and unsuitable for inclusion in this paper. It and the following additional expansions are quite useful in certain problems on wave propagation.

$$\cos\theta \cdot K_1(\rho' i\sqrt{i}) = -\sqrt{i}\int_0^\infty e^{-y'\sqrt{\mu^2+i}} \cos x'\mu \, d\mu,$$

$$\sin\theta \cdot K_1(\rho' i\sqrt{i}) = -\sqrt{i}\int_0^\infty \frac{\mu}{\sqrt{\mu^2+i}} e^{-y'\sqrt{\mu^2+i}} \sin x'\mu \, d\mu,$$

where

$$\cos\theta = y/\rho = y/\sqrt{x^2+y^2},$$

$$\sin\theta = x/\rho = x/\sqrt{x^2+y^2}.$$

Application to the Binomial Summation of a Laplacian [1] Method for the Evaluation of Definite Integrals

By E. C. MOLINA

Introduction

THE numerical evaluation of the incomplete Binomial Summation, a problem of major importance for many statistical and engineering applications of the Theory of Probability, is a question for which a satisfactory solution has not as yet been obtained. Several approximation formulas have been presented,[2] each of which gives good results for some limited range of values of the variables involved; but a formula of wide applicability is still a desideratum.

The purpose of this paper is to submit for consideration an approximation formula which seems to meet the situation to a measurable extent. The writer derived it by applying to the equation

$$(1) \qquad \sum_{x=c}^{x=n} \binom{n}{x} p^x(1-p)^{n-x} = \frac{\int_0^p x^{c-1}(1-x)^{n-c}dx}{\int_0^1 x^{c-1}(1-x)^{n-c}dx},$$

a method which is peculiarly efficacious for approximately evaluating definite integrals when the integrands contain factors raised to high powers.

The method used constitutes the subject matter of Chapter I, Part II, Book I of Laplace's "Théorie Analytique des Probabilités." Poisson applied the method to the integrals in the equation

$$(2) \qquad \sum_{x=c}^{x=n} \binom{n}{x} p^x(1-p)^{n-x} = \frac{\int_{(1-p)/p}^{\infty} x^{n-c}/(1+x)^{n+1}dx}{\int_0^{\infty} x^{n-c}/(1+x)^{n+1}dx}$$

and published a first approximation, together with its derivation, in his "Recherches sur la Probabilité des Jugements." Poisson's approximation seems never to have been used and was less fortunate than his famous limit to the binomial expansion which also was lost sight of until it reappeared under the caption "law of small numbers."

[1] Presented before International Congress of Mathematicians at Bologna, Italy in September, 1928.

[2] For an excellent resumé of some well-known formulas, together with a discussion of their limitations, reference may be had to C. Jordan, "Statistique Mathématique," articles 37 and 38.

While the integrals in equations (1) and (2) are well known equivalent forms for the complete and incomplete Beta functions, the equations themselves are not so familiar although one or the other will be found in Laplace, Poisson, Boole (Differential Equations) and at least two other places.

Approximate Formula

The approximate formula derived from equation (1) and submitted herewith for consideration is

$$(3) \qquad \sum_{x=c}^{x=n} \binom{n}{x} p^x(1-p)^{n-x} = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{T} e^{-t^2} dt - \frac{S_i e^{-T^2}}{2\sqrt{\pi}},$$

where S_i is the ith approximation to the infinite series

$$(4) \quad S = \frac{\sum_{s=1} R_s T^{s-1}[1 + (s-1)T_1^{-2} + (s-1)(s-3)T_1^{-4} \cdots]}{1 + \sum_{s=1} R_{2s}[1 \cdot 3 \cdot 5 \cdots (2s-1)]2^{-s}},$$

$$T_1 = T\sqrt{2},$$

$$(5) \quad T^2 = (n-1)\log\frac{n}{n-1} + (c-1)\log\frac{c-1}{a} + (n-c)\log\frac{n-c}{n-a},$$

and $a = np$; T to be taken negative when $a < (c-1)n/(n-1)$.

The first, second and third approximations to the infinite series S are

$$S_1 = R_1, \qquad S_2 = \frac{R_1 + R_2 T}{1 + R_2/2}, \qquad S_3 = \frac{R_1 + R_2 T + R_3(1 + T^2)}{1 + R_2/2},$$

where

$$R_1 = 4[(n-c) - (c-1)]/3\sqrt{2(n-1)(n-c)(c-1)},$$
$$R_2 = (1/6)[1/(n-c) + 1/(c-1) - 13/(n-1)],$$
$$R_3 = -(4/15)R_1[R_2 + 6/(n-1)].$$

It will be noted that R_2, $|R_1|$ and $|R_3|$ are symmetric functions of $(n-c)$ and $(c-1)$.

For the limiting case (Poisson's Exponential Binomial Limit) where $n = \infty$, $p = 0$ but $np = a$, we have

$$T^2 = 1 + (c-1)\log(c-1)/a + (a-c),$$
$$R_1 = 4/3\sqrt{2(c-1)},$$
$$R_2 = 1/6(c-1),$$
$$R_3 = -(4/15)R_1R_2.$$

Numerical Results

Since it is easy to compute the binomial summation directly when either c or $n - c$ is small, the practical value of an approximate formula depends on its efficiency for large values of these quantities.

The analysis given below under the heading "Derivation of the Approximate Formula" indicates that the successive R_s's in the series for S decrease when $\sqrt{c-1}$ and $\sqrt{n-c}$ increase. Therefore, when these two quantities are large, a few terms of the approximate formula (3) may be expected to give satisfactory results. As a matter of fact, the formula gives good results when $\sqrt{c-1}$ and $\sqrt{n-c}$ are not large. To confirm this statement the Tables [3] given at the end of this paper are submitted. In the 4th column of each table are given 10^6 times the true values of

$$P = \sum_{x=c}^{x=n} \binom{n}{x} p^x (1-p)^{n-x}.$$

In the columns headed Δ_1, Δ_2 and Δ_3 are given 10^6 times the differences between the true values and those obtained by applying formula (3) with the first, second and third approximations to S respectively. Table I in Czuber's "Wahrscheinlichkeitsrechnung" was used for evaluating the probability integral in equation (3).

The range of values of P covered by the tables is such that at the lower end of each section $P \not> .0005$ while at the upper end $P \not< .9995$, except where this latter condition would call for a value of $c < 2$. Of course, a larger or smaller range might have been given. The decision as to this question was based on the fact that several writers on the theory of statistics, when dealing with the normal law of errors, speak of an error exceeding 3 or 4 times the standard deviation as being a very improbable event. In order to keep the number of pages required for the tables within reasonable bounds computations were made only for even values of c.

The values of $a = np$ used are such that each of the values $p = 1/2$, $p = 1/10$ and $p = 1/20$ occurs twice; likewise each of the values $n = 100$, $n = 50$ and $n = 30$ occurs twice.

A greater degree of accuracy than that indicated by the tables can, of course, be obtained by working out and using R_4, $R_5 \cdots$; for this purpose, recourse should be had to equation (12) below and the details immediately following it. The only practical limitation to the use of formula (3) would appear to be the number of places given

[3] I am greatly indebted to Miss Nelliemae Z. Pearson of the Department of Development and Research both for supervising the work of my computers and contributing personally several sections of the tables.

by the existing tables for the probability integral. However, this difficulty is encountered only when P, or $(1 - P)$, is small, in which case T is large and the integral

$$\int_{-\infty}^{T} e^{-t^2} dt$$

may be readily evaluated by computing the first few terms of the series

$$[e^{-T^2}/2T\sqrt{\pi}][1 - T_1^{-2} + (1\cdot 3)T_1^{-4} - (1\cdot 3\cdot 5)T_1^{-6} \cdots],$$

where, as above, $T_1 = T\sqrt{2}$.

When P is very small, the difference $c - a = c - np$ is relatively large compared to a, and for this latter case recourse may be had to the approximate formula published by the writer in the *American Mathematical Monthly* for June, 1913.

Derivation of the Approximate Formula

Following Laplace closely, let us set

$$y(x) = Ye^{-t^2}, \tag{6}$$

where $Y = y(x_0)$ is the maximum value of $y(x)$. Then

$$\int_0^p y dx = Y \int_{-\infty}^{T} e^{-t^2} \left(\frac{dx}{dt}\right) dt, \tag{7}$$

the upper limit T being given by the equation

$$y(p) = y(x_0)e^{-T^2}. \tag{8}$$

Assuming dx/dt expanded in powers of t so that

$$dx/dt = \sum_{s=0} D_{s+1} t^s \tag{9}$$

and setting $R_s = D_{s+1}/D_1$, equation (7) reduces to

$$\int_0^p y dx = YD_1 \sum_{s=0} R_s \int_{-\infty}^{T} t^s e^{-t^2} dt.$$

Our fundamental equation (1) may now be written

$$\sum_{x=c}^{x=n} \binom{n}{x} p^x (1 - p)^{n-x} = \frac{\sum_{s=0} R_s \int_{-\infty}^{T} t^s e^{-t^2} dt}{\sum_{s=0} R_s \int_{-\infty}^{\infty} t^s e^{-t^2} dt}. \tag{10}$$

Integrating by parts and separating the terms involving $\int e^{-t^2}dt$ from the terms containing e^{-t^2}, we obtain equations (3) and (4).

To determine $R_s = D_{s+1}/D_1$, note that equation (6) gives $t = (\log Y - \log y)^{1/2}$ and set $v(x) = (x - x_0)/(\log Y - \log y)^{1/2}$ so that x may be written in the form

$$x = x_0 + v(x)t.$$

This form for x gives the expansion (Lagrange's Theorem for the simple case where $f(x) = x$; see "Modern Analysis" by Whittaker and Watson)

$$x = \sum_{s=0} \frac{t^s}{s!} \left(\frac{d^{s-1}v^s}{dx^{s-1}} \right)_{x=x_0}.$$

Comparing this expansion for x with the previous expansion (9) for dx/dt, we obtain

$$D_1 = v(x_0)$$

and

$$\frac{D_{s+1}}{D_1} = R_s = \left(\frac{1}{s!v(x)} \cdot \frac{d^s v^{s+1}}{dx^s} \right)_{x=x_0}.$$

Up to this point no particular form has been attributed to the function $y(x)$. From now on we deal with the function which constitutes the integrand of the integrals in equation (1).

The function $y(x) = x^{c-1}(1 - x)^{n-c}$ gives the expansion $(\log Y - \log y) = (x - x_0)^2[A_0 + A_1(x - x_0) + A_2(x - x_0)^2 \cdots]$, where $x_0 = (c - 1)/(n - 1)$ is the value of x for which $y(x)$ is a maximum and

$$A_s = \frac{1}{(s + 2)!} \left[\frac{d^{s+2}(\log Y - \log y)}{dx^{s+2}} \right]_{x=x_0}$$

or

$$A_s = \frac{(n - 1)^{s+2}}{s + 2} \left[\left(\frac{1}{n - c} \right)^{s+1} + (-1)^s \left(\frac{1}{c - 1} \right)^{s+1} \right]. \tag{11}$$

We are now prepared to evaluate R_s. Set

$$g = A_0 + A_1(x - x_0) + A_2(x - x_0)^2 \cdots$$

and

$$g_s = d^s g/dx^s.$$

Then

$$v = g^{-1/2},$$

$$\frac{dv^2}{dx} = -g^{-2}g_1,$$

$$\frac{d^2v^3}{dx^2} = (3/2)g^{-7/2}[(5/2)g_1^2 - g_2 g],$$

$$\frac{d^3v^4}{dx^3} = -2g^{-5}[g_3 g^2 - 9g_2 g_1 g + 12g_1^3].$$

Therefore, since $g_s = s!A_s$ when $x = x_0$,

$$R_1 = -A_0^{-3/2}A_1,$$
$$R_2 = (3/2)A_0^{-3}[(5/4)A_1^2 - A_0A_2],$$
$$R_3 = -2A_0^{-9/2}[A_3A_0^2 - 3A_2A_1A_0 + 2A_1^3].$$

Substituting for A_0, A_1, A_2 and A_3 the expressions derived by giving s the values 0, 1, 2 and 3 respectively in equation (11), we obtain for R_1, R_2 and R_3 the functions of n and c given on page 2.

For values of s greater than 3 the direct evaluation of d^sv^{s+1}/dx^s by successive differentiation becomes very tedious. It will be found much more practical to use the following procedure,[4] where D is a symbol of operation, $A = A_0$ and $b = A_1$.

$$A_0^{-1/2}R_s = (1/s!)\left(\frac{d^sg^{-(s+1)/2}}{dx^s}\right)$$
$$= \left[\frac{dA^{-(s+1)/2}}{1!dA}\right]D^{s-1}b + \left[\frac{d^2A^{-(s+1)/2}}{2!dA^2}\right]D^{s-2}b^2 + \cdots$$
$$+ \left[\frac{d^{s-1}A^{-(s+1)/2}}{(s-1)!dA^{s-1}}\right]Db^{s-1} + \left[\frac{d^sA^{-(s+1)/2}}{s!dA^s}\right]b^s$$

or

$$R_s = A_0^{1/2}\sum_{m=1}^{m=s}\left[\frac{d^mA^{-(s+1)/2}}{m!dA^m}\right](D^{s-m}b^m). \tag{12}$$

The following equations give the details requisite for the formation of R_s to R_8 inclusive; A_s can be computed from equation (11).

$Db = A_2$, $D^2b = A_3$, $D^3B = A_4$, $D^4b = A_5$,

$D^5b = A_6$, $D^6b = A_7$, $D^7b = A_8$,

$Db^2 = 2A_1A_2$,

$D^2b^2 = 2A_1A_3 + A_2^2$,

$D^3b^2 = 2A_1A_4 + 2A_2A_3$,

$D^4b^2 = 2A_1A_5 + 2A_2A_4 + A_3^2$,

$D^5b^2 = 2A_1A_6 + 2A_2A_5 + 2A_3A_4$,

$D^6b^2 = 2A_1A_7 + 2A_2A_6 + 2A_3A_5 + A_4^2$,

$Db^3 = 3A_1^2A_2$,

$D^2b^3 = 3A_1^2A_3 + 3A_1A_2^2$,

$D^3b^3 = 3A_1^2A_4 + 6A_1A_2A_3 + A_2^3$,

[4] See DeMorgan's "Differential and Integral Calculus," 1842, page 328, art. 214.

$$D^4b^3 = 3A_1^2A_5 + 6A_1A_2A_4 + 3A_1A_3^2 + 3A_2^2A_3,$$

$$D^5b^3 = 3A_1^2A_6 + 6A_1A_2A_5 + 6A_1A_3A_4 + 3A_2^2A_4 + 3A_2A_3^2,$$

$$Db^4 = 4A_1^3A_2,$$

$$D^2b^4 = 4A_1^3A_3 + 6A_1^2A_2^2,$$

$$D^3b^4 = 4A_1^3A_4 + 12A_1^2A_2A_3 + 4A_1A_2^3,$$

$$D^4b^4 = 4A_1^3A_5 + 12A_1^2A_2A_4 + 6A_1^2A_3^2 + 12A_1A_2^2A_3 + A_2^4,$$

$$Db^5 = 5A_1^4A_2, \qquad Db^6 = 6A_1^5A_2,$$

$$D^2b^5 = 5A_1^4A_3 + 10A_1^3A_2^2, \qquad D^2b^6 = 6A_1^5A_3 + 15A_1^4A_2^2,$$

$$D^3b^5 = 5A_1^4A_4 + 20A_1^3A_2A_3 + 10A_1^2A_2^3, \quad Db^7 = 7A_1^6A_2.$$

To illustrate the use of the procedure given above, let us evaluate R_4. We have

$$A_0^{-1/2}R_4 = \left(\frac{dA^{-5/2}}{1!dA}\right)D^3b + \left(\frac{d^2A^{-5/2}}{2!dA^2}\right)D^2b^2 + \left(\frac{d^3A^{-5/2}}{3!dA^3}\right)Db^3 + \left(\frac{d^4A^{-5/2}}{4!dA^4}\right)b^4$$

$$= -(5/2)A_0^{-7/2}(A_4) + (1/2)(5/2)(7/2)A_0^{-9/2}(2A_1A_3 + A_2^2)$$

$$-(1/6)(5/2)(7/2)(9/2)A_0^{-11/2}(3A_1^2A_2)$$

$$+(1/24)(5/2)(7/2)(9/2)(11/2)A_0^{-13/2}A_1^4$$

or

$$R_4 = (5/2)A_0^{-6}[-A_0^3A_4 + (7/2)A_0^2(A_1A_3 + A_2^2/2)$$

$$-(1/2)(7/2)(9/2)A_0A_1^2A_2$$

$$+(1/24)(7/2)(9/2)(11/2)A_1^4].$$

Tables Indicating Degree of Accuracy Obtainable by Use of Formula (3) for Evaluating

$$P = \sum_{x=c}^{x=n} \binom{n}{x} p^x(1-p)^{n-x}.$$

P_1 = 1st approximation, $\Delta_1 = (P - P_1)10^6$,

P_2 = 2d approximation, $\Delta_2 = (P - P_2)10^6$,

P_3 = 3d approximation, $\Delta_3 = (P - P_3)10^6$,

$a = np$,

$$T^2 = (n-1)\log\frac{n}{n-1} + (c-1)\log\frac{c-1}{a} + (n-c)\log\frac{n-c}{n-a},$$

$$I = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{T} e^{-t^2}dt.$$

TABLE I

c	T	$I(10^6)$	$P(10^6)$	Δ_1	Δ_2	Δ_3
			$a = 1.5, n = \infty, p = 0$			
2	+ .3074653	668154	442174	15991	9518	−1348
4	− .7612106	140849	65643	10816	1989	30
6	−1.5874105	12386	4456	1641	303	8
8	−2.2985028	577	170	103	20	0
			$a = 1.5, n = 30, p = .05$			
2	+ .2865166	657333	446458	21266	17382	−2083
4	− .8219430	122536	60772	3684	5617	− 97
6	−1.6966449	8211	3282	− 113	914	40
8	−2.4640017	246	85	− 24	50	7

TABLE II

c	T	$I(10^6)$	$P(10^6)$	Δ_1	Δ_2	Δ_3
			$a = 5, n = \infty, p = 0$			
2	1.5461442	985613	959576	−1681	2590	−798
4	.6837566	833222	734978	−2036	1897	−138
6	.0000000	500000	384044	2986	1036	− 4
8	− .5960752	199621	133376	4219	616	17
10	−1.1358169	54105	31832	2128	286	13
12	−1.6349406	10385	5452	604	123	2
14	−2.1027717	1471	692	107	11	− 5
16	−2.5454242	159	68	14	1	− 1
			$a = 5, n = 100, p = .05$			
2	1.5596227	986295	962920	− 373	3331	−732*
4	.6839234	833281	742162	639	3248	−141
6	− .0162780	490817	384001	2889	2108	− 38
8	− .6310024	186097	127961	1989	1431	− 8
10	−1.1912234	46029	28188	566	703	1
12	−1.7124507	7723	4274	75	208	2
14	−2.2138799	914	463	2	38	0
16	−2.6715388	79	37	− 1	5	0
			$a = 5, n = 50, p = .1$			
2	1.5742756	987005	966214	830	3962	−724
4	.6844600	833471	749706	3289	4437	−212
6	− .0335708	481067	383877	2599	3022	− 62
8	− .6689826	172053	122145	− 334	2076	31
10	−1.2524740	38258	24538	− 789	954	51
12	−1.7994619	5467	3220	− 278	244	24
14	−2.3191412	520	285	− 48	36	6
16	−2.8175965	34	17	− 5	3	0

* $|P - P_3| > |P - P_1|$.

TABLE III

c	T	$I(10^6)$	$P(10^6)$	Δ_1	Δ_2	Δ_3
			$a = 10, n = \infty, p = 0$			
2	2.5879363	999874	999499	− 47	.67	−37
4	1.8406742	995381	989662	− 533	275	−53
6	1.2386541	960089	932912	−1532	518	−50
8	.7094191	842135	779778	−1585	547	−27
10	.2274981	626172	542069	79	425	− 8
12	− .2200272	377838	303223	1786	322	1
14	− .6408864	182375	135535	2077	239	4
16	−1.0401811	70640	48740	1374	146	4
18	−1.4215063	22199	14277	628	68	0
20	−1.7875189	5737	3454	216	25	1
22	−2.1402533	1236	699	58	7	0
24	−2.4813121	225	119	12	1	− 1
			$a = 10, n = 100, p = .1$			
2	2.6528972	999912	999679	− 3	71	−24*
4	1.8917619	996268	992164	− 15	432	−39*
6	1.2715533	963931	942424	278	1070	−45
8	.7213308	846163	793949	997	1349	−36
10	.2161911	620099	548710	1213	1179	−19
12	− .2564838	358406	296967	503	982	− 2
14	− .7042404	159638	123877	− 222	771	10
16	−1.1320595	54691	39891	− 376	470	13
18	−1.5434535	14526	10007	− 222	206	9
20	−1.9410214	3025	1979	− 80	66	4
22	−2.3267578	500	312	− 20	15	1
24	−2.7022383	66	40	− 4	2	0

* $|P - P_3| > |P - P_1|$.

TABLE IV

c	T	$I(10^6)$	$P(10^6)$	Δ_1	Δ_2	Δ_3
			$a = 15, n = \infty, p = 0$			
4	2.6780004	999924	999788	− 18	10	− 4
6	2.1229551	998660	997207	− 141	54	− 9
8	1.6324888	989520	981998	− 525	146	− 14
10	1.1843012	953019	930147	−1066	242	− 15
12	.7670044	860975	815249	−1198	274	− 10
14	.3737500	701445	636783	− 515	245	− 4
16	.0000000	500000	431911	582	203	0
18	− .3574541	306598	251141	1311	168	2
20	− .7009899	160758	124781	1351	132	2
22	−1.0324325	72134	53106	961	89	2
24	−1.3532229	27826	19464	523	49	1
26	−1.6645241	9287	6184	228	22	0
28	−1.9672925	2700	1715	82	8	− 1
30	−2.2623270	689	418	25	2	0

TABLE IV—*Continued*

$a = 15, n = 30, p = .5$

6	2.6019552	999883	999837	52	24	− 12
8	2.0184138	997845	997388	559	152	− 77
10	1.4626537	980704	978613	2676	411	−239
12	.9237039	904277	899756	5946	488	−343
14	.3942720	711436	707667	5025	227	−205
16	− .1313195	426335	427768	−1916	− 73	72
18	− .6581761	175978	180797	−6392	−382	301
20	−1.1915875	45979	49369	−4405	−497	316
22	−1.7378702	6991	8062	−1346	−278	149
24	−2.3057782	555	715	− 191	− 69	32
26	−2.9097701	19	30	− 11	− 7	3

TABLE V

c	T	$I(10^6)$	$P(10^5)$	Δ_1	Δ_2	Δ_3
			$a = 25, n = \infty, p = 0$			
10	2.6086661	999888	999778	− 11	3	− 1
12	2.2291734	999191	998583	− 51	11	− 2
14	1.8705496	995920	993531	− 159	30	− 4
16	1.5289263	984700	977705	− 364	59	− 6
18	1.2015564	955365	939522	− 617	90	− 7
20	.8863972	894998	866422	− 765	109	− 7
22	.5818753	794717	752697	− 651	110	− 7
24	.2867455	657452	606120	− 253	101	− 6
26	.0000000	500000	447076	268	92	− 3
28	− .2791919	346482	299814	678	84	1
30	− .5515253	217703	182105	837	75	2
32	− .8175896	123790	100070	761	62	3
34	−1.0778902	63708	49782	561	45	3
36	−1.3328647	29718	22460	350	29	2
38	−1.5828952	12593	9212	189	17	2
40	−1.8283181	4860	3445	90	9	2
42	−2.0694313	1713	1178	38	4	1
44	−2.3065005	553	370	15	2	1
			$a = 25, n = 50, p = .5$			
14	2.3698187	999598	999531	80	16	− 9
16	1.9447371	997023	996699	404	57	− 32
18	1.5274793	984621	983580	1329	126	− 74
20	1.1159208	942735	940539	2844	170	−113
22	.7083182	841759	838881	3763	140	−109
24	.3031406	665931	664094	2415	60	− 56
26	− .1010188	443200	443862	− 873	− 20	19
28	− .5055162	237333	239944	−3426	−102	87
30	− .9117246	98634	101319	−3499	−166	117
32	−1.3211006	30859	32454	−2056	−157	95
34	−1.7352770	7063	7673	− 774	− 91	50
36	−2.1561545	1147	1301	− 191	− 33	17
38	−2.5860897	128	153	− 31	− 8	4

A New Method for Obtaining Transient Solutions of Electrical Networks

By W. P. MASON

Synopsis: A new method for obtaining transient solutions of electrical networks is developed in this paper which depends upon the fact that a distortionless line can be made to approach as a limit all three of the circuit elements, resistance, inductance and capacity. The process of solution consists in solving for the current in a distortionless line—which is ordinarily a simple process—and then proceeding to the limiting case of the distortionless line which approaches the element or elements of interest. Some examples are worked out and a derivation of the Laplacian integral solution is given. It is interesting to note that this method gives a formal solution of the Laplacian integral equation.

THE following paper sets forth a new method for obtaining the transient solutions of electrical networks, which it is believed has some advantages over other methods of solution, in that the operations required for solution are quite simple, and also because this method presents a more definite physical picture of the processes involved. By means of this method, the current at any time t can be obtained, due to an applied voltage which is zero when t is less than zero, and is $E_0 \cos(\omega t + \theta)$ when t is greater than zero. This type of voltage includes as a special case the applied voltage, which is zero when t is less than zero, and is unity when t is greater than zero, and hence the solutions obtained by this method reduce to the cases discussed by Heaviside,[1] when ω and θ are taken equal to zero.

This method gives directly the more compact Laplacian integral equation solution, first obtained by Carson, and in addition gives a method for solving this integral equation, if its solution is not already known.

I. Method of Solution

All practical schemes for solving the transient type of circuit problem, including the Laplacian integral equation, and the generalized Fourier integral solution, are made to depend on the known and easily determined steady state solution. This implies that all circuits which have the same steady state solution, have also the same transient or time solution. The method described in the present paper rests on the same basis.

The method of solution used here depends upon the fact that the

[1] Heaviside, "Electromagnetic Theory," Volume II.

distortionless line can be made to approach as a limit, all three of the electrical elements, resistance, inductance, and capacity, and that the complete solution for the current in a distortionless line can be obtained by adding the incident current and the sum of the reflected currents which can occur up to the time of interest. That is the distortionless line has a true velocity of propagation, and hence the current at any time will be the initial current and the sum of the reflections which can occur up to the time of interest. All of the three electrical elements, resistance, inductance, and capacity, can be considered as limiting cases of the distortionless line. Hence the process of solution consists in solving for the current in the distortionless line, and then proceeding to the limiting value of the line which coincides with the element of interest.

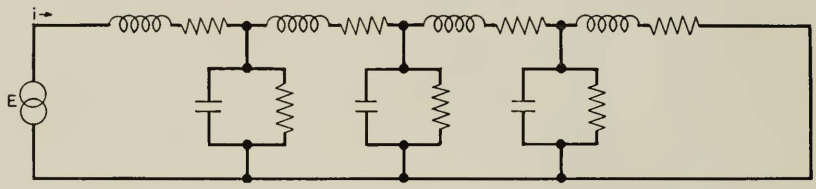


Fig. 1-A.

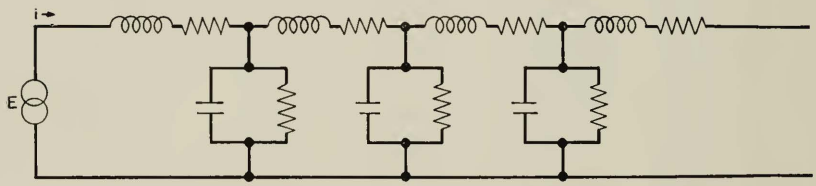


Fig. 1-B.
Diagrammatic representations of lines.

A. The Distortionless Line

Since the distortionless line is the tool by means of which problems are solved by this method, a brief discussion of lines [2] is given here. If a voltage is suddenly applied to a transmission line, the current at any point in the line is zero for a certain time and then begins to build up to its final or steady state value. If there is no distributed inductance in the line, the current begins to build up immediately.

For a distortionless line, however, the current is zero for the time required to propagate the wave to the point of interest and then instantaneously reaches its steady state value. To show this let us consider the equations for a transmission line. A line has distributed

[2] For a more complete discussion of lines see "Transmission Circuits for Telephone Communication," K. S. Johnson, page 144.

series resistance R_u and inductance L_u, and distributed shunt capacity C_u and leakage conductance G_u, as shown on Fig. 1-A, where the letter u indicates the values per unit length. If i denotes the current and v the voltage at a distance x from one end of the line, the well known differential equations are

$$\left.\begin{aligned}\left(L_u\frac{d}{dt}+R_u\right)i &= -\frac{\partial}{\partial x}v,\\ \left(C_u\frac{d}{dt}+G_u\right)v &= -\frac{\partial}{\partial x}i.\end{aligned}\right\} \tag{1}$$

If we eliminate v from these two equations we have

$$L_uC_u\frac{d^2i}{dt^2}+(R_uC_u+G_uL_u)\frac{di}{dt}+R_uG_ui=\frac{\partial^2 i}{\partial x^2}. \tag{2}$$

Similarly, if i is eliminated, the resulting equation is the same as (2) with i replaced by v. Since we are dealing with simple harmonic forces, the current i varies as $\cos(\omega t+\theta)$ where ω is 2π times the frequency of vibration and θ is an arbitrary phase angle. It is usually more convenient to consider i as varying according to the time factor

$$i=\dot{i}e^{j(\omega t+\theta)}=\dot{i}[\cos(\omega t+\theta)+j\sin(\omega t+\theta)],$$

where $\dot{i}$ is the maximum amplitude of i. The solution obtained on this assumption is called the symbolic solution, and the real solution is obtained from the symbolic solution by taking the real part. Substituting the symbolic form of i in equation (2), this equation reduces to

$$[(j\omega)^2L_uC_u+j\omega[R_uC_u+G_uL_u]+R_uG_u]i=\frac{\partial^2 i}{dx^2}. \tag{3}$$

The solution for a line can be specified in terms of two parameters, the characteristic impedance and the propagation constant of the line. To show this we note that the solution of (3) is

$$i=Ae^{-\Gamma x}+Be^{\Gamma x}, \tag{4}$$

where $\Gamma^2=[R_u+j\omega L_u][G_u+j\omega C_u]$ and A and B are constants. From the last of equations (1) we have

$$v=-\frac{\frac{\partial}{\partial x}i}{G_u+j\omega C_u}=\frac{\Gamma}{G_u+j\omega C_u}[Ae^{-\Gamma x}-Be^{\Gamma x}]. \tag{5}$$

When $x = 0$, $i = i_0$, and $v = v_0$. From (4) and (5) solving for A and B we have

$$A = \frac{i_0}{2} + \frac{v_0/2}{\sqrt{\frac{R_u + j\omega L_u}{G_u + j\omega C_u}}}; \qquad B = \frac{i_0}{2} - \frac{v_0/2}{\sqrt{\frac{R_u + j\omega L_u}{G_u + j\omega C_u}}}.$$

Substituting these values in (4) and (5), we have the equations

$$\begin{aligned} i &= i_0 \cosh \Gamma x - \frac{v_0 \sinh \Gamma x}{\sqrt{\frac{R_u + j\omega L_u}{G_u + j\omega C_u}}}, \\ v &= v_0 \cosh \Gamma x - i_0 \sqrt{\frac{R_u + j\omega L_u}{G_u + j\omega C_u}} \sinh \Gamma x. \end{aligned} \tag{6}$$

In this equation $\Gamma x = P$, the propagation constant of the line, and

$$\sqrt{\frac{R_u + j\omega L_u}{G_u + j\omega C_u}} = Z_0,$$

the characteristic impedance of the line. If we are interested in a given length of line l, the parameters are

$$P = \sqrt{(R + j\omega L)(G + j\omega C)}; \qquad Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}}, \tag{7}$$

where R, L, G, and C are the total distributed constants for the length of line considered. The characteristic impedance is the impedance looking into a line of infinite length as can readily be seen from either of equations (6) by letting x, the length, approach infinity. In this case $\cosh \Gamma x = \cosh P$ approaches $\sinh P$ and both approach infinity. Then from (6)

$$\frac{v_0}{i_0} = \left(\frac{\cosh P - i/i_0}{\sinh P}\right) Z_0 \to Z_0 \quad \text{when} \quad x \to \infty,$$

since i/i_0 can never be larger than 1.

The physical significance of the propagation constant is that e^{-P} represents the ratio of currents or voltages at the two ends of the line when the line is connected to an infinite line of the same characteristic impedance. To show this, suppose we terminate the section considered in an infinite line, which as we have seen above has an impedance Z_0. Then in equations (6), $v = v_1$ the output voltage and

$i = i_1$ the output current. We let $v_1 = i_1 Z_0$. Eliminating either v_0 or i_0, we have

$$v_1 = v_0 e^{-P} \qquad \text{or} \qquad i_1 = i_0 e^{-P}. \tag{8}$$

In the following work it is necessary to know the impedance of a short circuited line and that of an open circuited line. For the short circuited line, the voltage at the far end is zero, so putting $v = 0$ in the last of equation (6), we have

$$v_0/i_0 = Z_0 \tanh P. \tag{9}$$

For the open circuited line we put $i = 0$, obtaining

$$v_0/i_0 = Z_0 \coth P. \tag{10}$$

So far we have discussed the general transmission line. For the distortionless line there exists the relation

$$\frac{R}{L} = \frac{G}{C}. \tag{11}$$

Substituting this relation in equation (7), these parameters reduce to

$$Z_0 = R_0 = \sqrt{\frac{L}{C}} = \sqrt{\frac{R}{G}} \quad \text{and} \quad P = \sqrt{RG} + j\omega\sqrt{LC} = A + j\omega D. \tag{12}$$

This equation shows that the characteristic impedance becomes a resistance R_0, while the propagation constant equals a real constant A plus $j\omega$ times the constant D. To show how wave transmission takes place in an infinite line, these values are substituted in equation (8), giving

$$v_1 = v_0 e^{-(A+j\omega D)}; \qquad i_1 = i_0 e^{-(A+j\omega D)}.$$

To find the real solution, we take the real part of this symbolic solution, obtaining

$$v_1 = \bar{v}_0 e^{j(\omega t+\theta)} e^{-(A+j\omega D)} = \bar{v}_0 e^{-A} e^{j[\omega(t-D)+\theta]},$$

or, taking the real part,

$$\begin{aligned} v_1 &= \bar{v}_0 e^{-A} \cos [\omega(t - D) + \theta] \\ \text{and} \qquad i_1 &= \bar{i}_0 e^{-A} \cos [\omega(t - D) + \theta], \end{aligned} \tag{13}$$

where the dash over v_0 and i_0 indicate the maximum amplitude of these quantities. Since $v_0 = \bar{v}_0 \cos(\omega t + \theta)$, these equations show that either v_1 or i_1 has the same form as v_0 or i_0 respectively, attenuated by a factor e^{-A} and delayed in time by an amount D.

B. Condition for Obtaining Lumped Constants from a Distortionless Line

In a distortionless line there is only one necessary relation between the constants, equation (11). Hence, we are at liberty to vary the constants subject only to this relation. In the following work we wish to make the distortionless line degenerate into resistances, inductances, and capacities, or combinations of these.

For example, suppose that we wish to obtain a resistance from a distortionless line. To obtain this we take a short circuited line as shown on Fig. 1-A and let R be finite, $L \to 0$, $G \to 0$, and $C \to 0^2$ in order to satisfy equation (11). The shunt elements will all vanish and the series inductance disappears, leaving only the series resistance in the line. Since the line is short circuited, the line degenerates into a resistance. The line parameters for this case become

$$R_0 = \sqrt{\frac{R}{G}} = \sqrt{\frac{L}{C}} \to \infty^{1/2}; \quad P = \sqrt{RG} + j\omega\sqrt{LC} \to (0^{1/2} + j\omega 0^{3/2})$$

and

$$R_0 P = \sqrt{\frac{R}{G}}(\sqrt{RG}) + \sqrt{\frac{L}{C}}(j\omega\sqrt{LC}) \to R. \tag{14}$$

There are three combinations of lumped elements which can be obtained from the short circuited line and three combinations which can be obtained from the open circuited line. These are listed in the following table, together with the resulting line parameters.

LIMITING CASES OF THE DISTORTIONLESS LINE

A. Limiting Values with Short Circuited Line

Equation	Assumed Line Constants				Resulting Line Parameters			Resulting Lumped Element
	R	L	G	C	R_0	P	R_0P	
(14)	Finite	0	0	0^2	$\infty^{1/2}$	$0^{1/2} + j\omega 0^{3/2}$	R	Resistance
(15)	0	Finite	0^2	0	$\infty^{1/2}$	$0^{3/2} + j\omega 0^{1/2}$	$j\omega L$	Inductance
(16)	Finite	Finite	0	0	$\infty^{1/2}$	$0^{1/2} + j\omega 0^{1/2}$	$R+j\omega L$	Resistance and Inductance

The first three cases result from the suppression of the shunt elements in the short circuited line, and the line parameters are characterized by $R_0 \to \infty$; $P \to 0$. The last three cases result from the suppression of the series elements in the open circuited line, and are characterized by $R_0 \to 0$; $P \to 0$.

B. Limiting Values with Open Circuited Line

Equation	Assumed Line Constants				Resulting Line Parameters			Resulting Lumped Element
	R	L	G	C	R_0	P	R_0/P	
(17)	0	0^2	Finite	0	$0^{1/2}$	$0^{1/2} + j\omega 0^{3/2}$	$1/G$	Resistance
(18)	0^2	0	0	Finite	$0^{1/2}$	$0^{3/2} + j\omega 0^{1/2}$	$1/j\omega C$	Capacity
(19)	0	0	Finite	Finite	$0^{1/2}$	$0^{1/2} + j\omega 0^{1/2}$	$1/(G+j\omega C)$	Resistance and Capacity

C. Solution for a Resistance and Inductance in Series

As a first example of a transient solution obtained by this method let us consider the case of an inductance and resistance in series with a source of alternating voltage. To solve this problem, consider the case of a voltage in series with a distortionless line, short circuited, as shown on Fig. 1-*A*. The current immediately flowing on application of the voltage will be i_0 where

$$i_0 = \frac{E}{R_0}. \tag{20}$$

This current is transmitted down the line and completely reflected at the far end, returning to the near end. The first reflected current entering the generator is then

$$i_1 = i_0 e^{-2P}, \tag{21}$$

where P is the propagation constant of the line. Upon reaching the near end, the current is completely reflected in the same phase and again enters the line. At the end of the first reflection, the current entering the line is

$$i = i_0(1 + 2e^{-2P}). \tag{22}$$

After $(n - 1)$ reflections and passages through the line, the current is

$$\begin{aligned} i &= i_0[1 + 2e^{-2P} + 2e^{-4P} + \cdots + 2e^{-2(n-1)P}] \\ &= i_0\left[2\left(\frac{1 - e^{-2nP}}{1 - e^{-2P}}\right) - 1\right]. \end{aligned} \tag{23}$$

Now the time at which the nth reflection occurs will be

$$t = n(2D),$$

where D is the time of delay in passing the network once. For a distortionless line

$$D = \sqrt{LC}. \tag{24}$$

Hence, we can replace n by

$$n = \frac{t}{2D} = \frac{t}{2\sqrt{LC}}. \tag{25}$$

Since $D \to 0$ and $n \to \infty$, the time scale becomes continuous. In equation (23) we insert the values given in (16) and (25), appropriate to the limiting case considered here, namely

$$P = \frac{R + j\omega L}{R_0}; \qquad R_0 \to \infty; \qquad n = \frac{t}{2\sqrt{LC}} = \frac{tR_0}{2L}$$

and note that $2P \to 0$ so that $e^{-2P} \to 1 - 2P$; then

$$i \to \frac{E}{R_0}\left[\frac{2(1 - e^{-tR_0/L[(R+j\omega L)/R_0]})}{1 - 1 + 2(R + j\omega L)/R_0} - 1\right]$$

$$= E\left[\frac{1 - e^{-t(R/L+j\omega)}}{R + j\omega L} - \frac{1}{R_0}\right].$$

But $R_0 \to \infty$ and hence the solution is

$$i = E\left[\frac{1 - e^{-t(R/L+j\omega)}}{R + j\omega L}\right]. \tag{26}$$

This is the symbolic or complex algebra solution of the equation

$$L\frac{di}{dt} + Ri = E. \tag{27}$$

In general it is desirable to obtain the current due to an applied voltage of the form

$$E = E_0 \cos(\omega t + \theta).$$

This solution can be obtained directly from the symbolic solution given in (26) by taking the real part. The result is

$$i = E_0\left[\frac{\cos(\omega t + \theta - \varphi) - \cos(\theta - \varphi)e^{-tR/L}}{\sqrt{R^2 + \omega^2 L^2}}\right]. \tag{28}$$

It will be found that (28) is a solution of (27) for an applied voltage $E_0 \cos(\omega t + \theta)$.

D. General Method for Determining Reflections

The method for obtaining the successive current reflections of the line given in the preceding section, is laborious to carry out in complicated cases and hence it is desirable to obtain a simple method for determining the reflections. Such a method is the expansion of the expression for the impedance by the binomial theorem in order to get the successive reflections. In the above example the current i is

$$i = E/(R + j\omega L).$$

We note that the expression for the impedance is approached by that of a short circuited line when the R and L of the line are finite and the capacity and leakance approach zero. Hence

$$i = \frac{E}{R + j\omega L} \rightarrow \frac{E}{R_0 \tanh P} = \frac{E(1 + e^{-2P})}{R_0(1 - e^{-2P})}.$$

Now the expansion of

$$\frac{1}{1 - e^{-2P}} = 1 + e^{-2P} + e^{-4P} + \cdots.$$

Hence

$$i \rightarrow \frac{E}{R_0}(1 + e^{-2P})(1 + e^{-2P} + e^{-4P} + \cdots)$$

$$= \frac{E}{R_0}[1 + 2e^{-2P} + 2e^{-4P} + \cdots + 2e^{-2(n-1)P} + \cdots],$$

which is the expression for the reflections given by equation (23).

In all the following problems it will be found that a similar process for obtaining the reflections can be followed. It is evident that any method which gives an expression for the current in the form

$$i = E[a_0 + a_1 e^{-2j\omega D} + a_2 e^{-4j\omega D} + \cdots + a_n e^{-2nj\omega D} + \cdots] \quad (29)$$

will give the reflections, for if we take the real part of this expression we have

$$i = E_0[a_0 \cos(\omega t + \theta) + a_1 \cos[\omega(t - 2D) + \theta] + \cdots + a_n \cos[\omega(t - 2nD) + \theta] + \cdots].$$

Each term represents a current which adds to the original current after a time of delay $2D$, $4D$, $\cdots$ $2nD$, and hence the nth terms represents the nth reflection. Therefore any method, such as the above, which gives the current in the form of equation (29), will give the reflections.

E. Simpler Form for Replacing an Impedance

In the preceding section, the transient solution of an inductance and resistance in series was obtained by replacing the impedance $R + j\omega L$ by the expression

$$R_0 \tanh P, \text{ where } R_0 P = R + j\omega L \text{ and } R_0 \to \infty\ ;\ P \to 0.$$

Tanh P has a numerator and a denominator both of which must be expanded in order to obtain the reflections. If a single term can be used, the expansion becomes simpler. In order to effect such a simplification, it is necessary to find a physical structure, which has only one term in its impedance expression and which approaches a resistance and inductance as a limit.

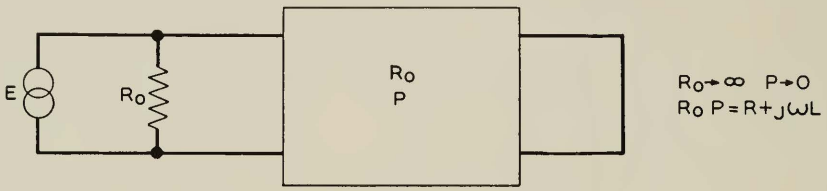


Fig. 2—Short circuited line and shunt resistance.

Such a structure is shown on Fig. 2. It consists of a short circuited line shunted by a resistance R_0. The current into the combination is

$$i = \frac{E}{\dfrac{R_0 \times R_0 \tanh P}{R_0 + R_0 \tanh P}} = \frac{E}{\dfrac{R_0(1 - e^{-2P})}{2}}. \tag{30}$$

If now in the short circuited line, we let R and L be finite and $G \to 0$, $C \to 0$, the combination obviously reduces to a resistance and inductance in series, since the infinite shunt will not affect the result. Hence the replacement of a resistance and inductance in the equation

$$i = \frac{E}{R + j\omega L}$$

by the expression in (30), is justified.

The solution of (30) is gotten by expanding the expression and is

$$i = \frac{2E}{R_0}\left[1 + e^{-2P} + e^{-4P} + \cdots + e^{-2(n-1)P} + \cdots\right] = \frac{2E[1 - e^{-2nP}]}{R_0[1 - e^{-2P}]}.$$

Upon substituting in the values $R_0 P = R + j\omega L$, $n = t/2D$, and letting $R_0 \to \infty\ ;\ P \to 0$, we have

$$i = E\left[\frac{1 - e^{-t[R/L + j\omega]}}{R + j\omega L}\right]$$

in agreement with equation (26).

Similarly, when we have the expression

$$\frac{1}{G + j\omega C}$$

we can replace it by

$$\frac{2R_0}{1 - e^{-2P}} \quad \text{where} \quad \frac{R_0}{P} = \frac{1}{G + j\omega C} \quad \text{and} \quad R_0 \to 0, \; P \to 0. \qquad (31)$$

The structure which gives the impedance in (31) is an open circuited line in series with a resistance R_0. The impedance of the combination is

$$R_0 + R_0 \coth P = R_0 \left[1 + \frac{1 + e^{-2P}}{1 - e^{-2P}} \right] = \frac{2R_0}{1 - e^{-2P}} .$$

If then the series impedances of the line approach zero, $R_0 \to 0$ and the impedance of the combination approaches

$$\frac{1}{G + j\omega C} .$$

F. Solution for a Resistance and Capacity in Series

As a second example let us obtain the solution of a resistance and capacity in series. To obtain the solution we solve the case of a

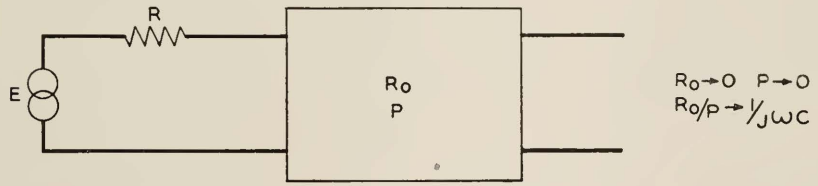


Fig. 3—Open circuited line and series resistance.

resistance in series with an open circuited line as shown on Fig. 3. The steady state solution for the current in this circuit is

$$i = \frac{E}{R + \dfrac{1}{j\omega C}} .$$

Replacing $\dfrac{1}{j\omega C}$ by $\dfrac{2R_0}{1 - e^{-2P}}$, and substituting in the above equation there results

$$i = \frac{E}{R + \dfrac{2R_0}{1 - e^{-2P}}} \quad \text{when} \quad R_0 \to 0 \quad \text{and} \quad P \to 0, \quad \text{and} \quad \frac{R_0}{P} \to \frac{1}{j\omega C}$$

in accordance with equation (18).

After some rearrangements this can be put into the form

$$i = \frac{E}{R + R_0}\left[1 - \frac{R_0}{R}\,\frac{e^{-2F}(1 + e^{-2P})}{1 - e^{-2(F+P)}}\right], \tag{32}$$

where

$$e^{-2F} = \frac{R}{R + 2R_0}.$$

Expanding equation (32) in the form of a series, there results

$$i = \frac{E}{R+R_0}\left[\frac{R+3R_0}{R+2R_0} - \frac{R_0}{R}\left(\frac{2R+2R_0}{R+2R_0}\right)(1+e^{-2(F+P)}+e^{-4(F+P)} + \cdots)\right].$$

Summing up n terms of this series, we have

$$i = \frac{E}{R + R_0}\left[\frac{R + 3R_0}{R + 2R_0} - \frac{R_0}{R}\left(\frac{2R + 2R_0}{R + 2R_0}\right)\left(\frac{1 - e^{-2n(F+P)}}{1 - e^{-2(F+P)}}\right)\right]. \tag{33}$$

Since in the above expression $R_0 \to 0$, we can obtain the value of F by writing the first terms of the expansion for the exponential

$$e^{-2F} = 1 - 2F + \frac{(2F)^2}{2!} + \cdots = \frac{R}{R + 2R_0} = 1 - \frac{2R_0}{R} + \cdots.$$

Hence

$$F \to \frac{R_0}{R}.$$

If now in equation (33) we proceed to the limit, letting

$$R_0 \to 0; \quad P \to 0; \quad \frac{R_0}{P} = \frac{1}{j\omega C}; \quad n = \frac{t}{2D}$$

there results the equation

$$i = \frac{E}{R}\left[1 - \frac{1 - e^{-t(1/RC+j\omega)}}{1 + j\omega CR}\right]. \tag{34}$$

This equation is the symbolic solution of the integral equation

$$Ri + \frac{1}{C}\int idt = E_0 e^{j(\omega t+\theta)}.$$

If we wish the solution corresponding to the impressed voltage, $E_0 \cos(\omega t + \theta)$, we take the real part of (34), obtaining

$$i = E_0\,\frac{\cos(\omega t + \theta + \delta) - [\sin(\theta - \delta)\tan\delta]e^{-t/RC}}{\sqrt{R^2 + 1/\omega^2C^2}},$$

where

$$\tan\delta = \frac{1}{\omega RC}.$$

II. Solution for M Sections of All-Pass Lattice Network

The process for solving any type of problem is to replace any resistance and inductance in series by $R_{0_1}(1 - e^{-2P_1})/2$, and any capacity and leakance in parallel by $(1 - e^{-2P_2})/2R_{0_2}$, where $R_{0_1} \to \infty$; $R_{0_2} \to 0$; $R_{0_1}P_1 = R + j\omega L$; $R_{0_2}/P_2 = 1/(G + j\omega C)$.

The next problem considered here is the solution for any number of sections of all-pass lattice type network [3] as shown on Figure 4.

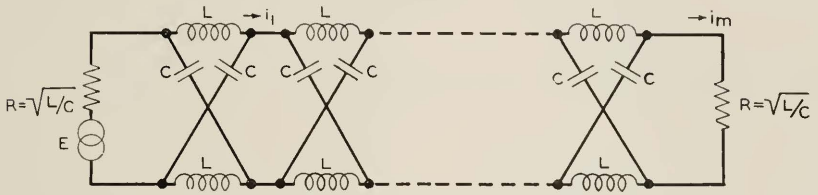


Fig. 4—Sections of all pass lattice network.

These networks have the property of passing all frequencies without attenuation, and they are much used as phase equalizers.

The steady state equation for the current at the end of the first section, when this section is terminated at each end by a resistance $R = \sqrt{Z_1 Z_2}$, is

$$i_1 = \frac{E}{2R}\left[\frac{\sqrt{Z_2} - \sqrt{Z_1}}{\sqrt{Z_2} + \sqrt{Z_1}}\right].$$

The current flowing out of the mth section of such a structure takes the form

$$i_m = \frac{E}{2R}\left[\frac{\sqrt{Z_2} - \sqrt{Z_1}}{\sqrt{Z_2} + \sqrt{Z_1}}\right]^m. \tag{35}$$

In the structure considered $Z_1 = j\omega L$; $Z_2 = 1/j\omega C$, and $\sqrt{L/C} = R$. In accordance with section I-B, we replace the inductance by a short circuited line, and the capacity by an open circuited line. For the first line, in the limiting case, we have by equation (15),

$$R_0 \to \infty; \quad R_0 P = j\omega L.$$

For the second line, by equation (18), we have

$$\bar{R}_0 \to 0; \quad \frac{\bar{R}_0}{\bar{P}} = \frac{1}{j\omega C}.$$

There is no loss of generality if we take the propagation constants for

[3] See for example *B. S. T. J.*, July 1928, page 510.

the two lines equal so that

$$\frac{\bar{R}_0}{\bar{P}} \times R_0 P = j\omega L \times \frac{1}{j\omega C} = \frac{L}{C} = R^2.$$

Hence $\bar{R}_0 = R^2/R_0$. Substituting these values in equation (35), we have

$$i_m = \frac{E}{2R}\left[\frac{1 - [R_0/2R](1 - e^{-2P})}{1 + [R_0/2R](1 - e^{-2P})}\right]^m. \tag{36}$$

After some simple rearrangements, equation (36) takes the form

$$i_m = \frac{E}{2R}\left[\frac{4R}{R_0}\left(\frac{1}{1 - e^{-2[(R/R_0)+P]}}\right) - 1\right]^m. \tag{37}$$

If m equals 1, the solution can be obtained exactly as discussed in the first example in section (1), and it is

$$i_1 = \frac{E}{2R}\left[\frac{1 - j\omega\sqrt{LC} - 2e^{-t[(1/\sqrt{LC})+j\omega]}}{1 + j\omega\sqrt{LC}}\right].$$

The solution for m sections of lattice network is discussed in the Appendix, and it is there shown that the solution can be written in the form

$$\begin{aligned} i_m = \frac{E}{2R}\Bigg[& \left(\frac{1 - j\omega\sqrt{LC}}{1 + j\omega\sqrt{LC}}\right)^m - e^{-t[(1/\sqrt{LC})+j\omega]}\Bigg[\frac{\left(\frac{t}{\sqrt{LC}}\right)^{m-1}}{(m-1)!} \times \frac{2^m}{1 + j\omega\sqrt{LC}} \\ & + \frac{\left(\frac{t}{\sqrt{LC}}\right)^{m-2}}{(m-2)!}\left(\frac{2^m}{(1 + j\omega\sqrt{LC})^2} - \frac{m2^{m-1}}{1 + j\omega\sqrt{LC}}\right) \\ & + \cdots + \frac{t}{\sqrt{LC}}\left[\frac{(1 - j\omega\sqrt{LC})^m}{(1 + j\omega\sqrt{LC})^{m-1}} - (-1)^m\left(m + 1 - \frac{m}{1 + j\omega\sqrt{LC}}\right)\right] \\ & + \left(\frac{1 - j\omega\sqrt{LC}}{1 + j\omega\sqrt{LC}}\right)^m - (-1)^m\Bigg]\Bigg]. \end{aligned} \tag{38}$$

Equation (38) represents the symbolic or complex algebra solution for the current at the end of the mth section of a lattice network as shown on Fig. 4. It is usually desirable to obtain the current due

to an applied voltage $E_0 \cos(\omega t + \theta)$. This can be obtained from equation (38) by taking the real part of the equation and rejecting the imaginary part. The process of doing this is simple and the result obtained is

$$i_m = \frac{E}{2R}\Bigg[\cos(\omega t + \theta - 2m\varphi) - 2e^{-(t/\sqrt{LC})}\Bigg[\frac{\left(\frac{2t}{\sqrt{LC}}\right)^{m-1}}{(m-1)!}\cos(\theta-\varphi)\cos\varphi + \frac{\left(\frac{2t}{\sqrt{LC}}\right)^{m-2}}{(m-2)!}[2\cos^2\varphi\cos(\theta-2\varphi) - m\cos\varphi\cos(\theta-\varphi)] + \frac{\left(\frac{2t}{\sqrt{LC}}\right)^{m-3}}{(m-3)!}\Bigg[4\cos^3\varphi\cos(\theta-3\varphi) - 2m\cos^2\varphi\cos(\theta-2\varphi) + \frac{m(m-1)}{2!}\cos\varphi\cos(\theta-\varphi)\Bigg] + \cdots\Bigg]\Bigg], \tag{39}$$

where $\tan\varphi = \omega\sqrt{LC}$.

For example, the solutions for one and two section networks take the form

$$i_1 = \frac{E}{2R}[\cos(\omega t + \theta - 2\varphi) - 2e^{-(t/\sqrt{LC})}\cos\varphi\cos(\theta-\varphi)]$$

$$i_2 = \frac{E}{2R}\Bigg[\cos(\omega t + \theta - 4\varphi) - 2e^{-(t/\sqrt{LC})}\Bigg[2\cos^2\varphi\cos(\theta-2\varphi) - 2\cos\varphi\cos(\theta-\varphi) + \frac{2t}{\sqrt{LC}}\cos(\theta-\varphi)\cos\varphi\Bigg]\Bigg]. \tag{40}$$

It appeared desirable to obtain some numerical calculations on the building up of current in this type of network. This calculation has been carried through for two section, four section, and six section networks. The current has been calculated for a constant voltage, for an alternating voltage whose frequency is the resonant frequency of the network, and for one whose frequency is twice the resonant frequency of the network. The current building up for a constant applied voltage is shown for the three networks on Fig. 5. The current building up in a two section network, in which an alternating voltage whose frequency equals the resonant frequency of the network, is shown on Fig. 6. The steady state and the transient terms

are shown separately in the dotted lines, and the complete solution in the full line. The applied voltage is of the form $E_0 \cos \omega t$ and hence θ is taken as zero in equation (39). Similarly, curves for two, four,

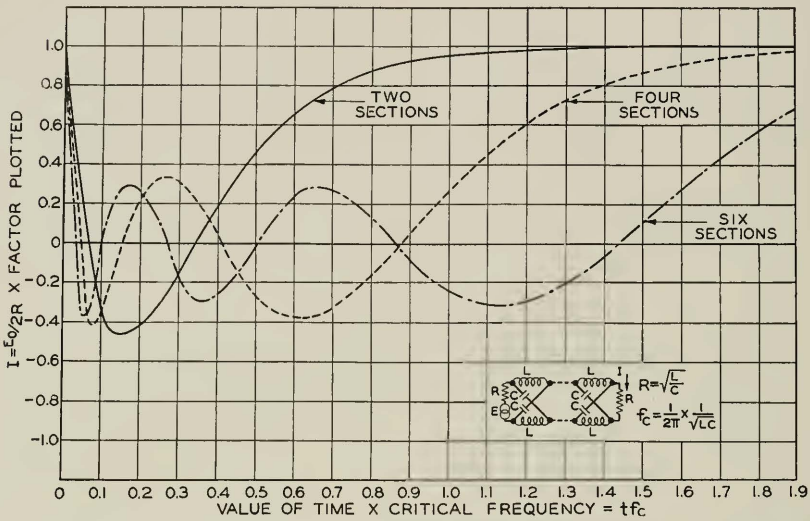


Fig. 5—Current resulting from the application of a constant voltage, E, on several sections of lattice network. The current plotted is the current in the termination of the network.

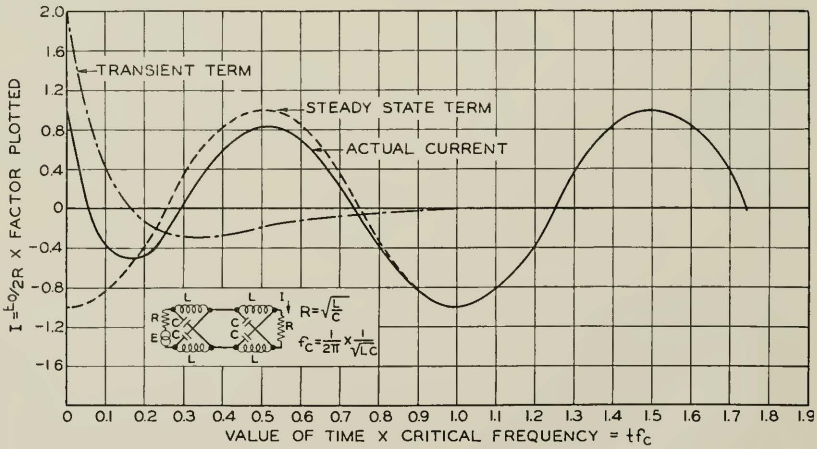


Fig. 6—Current resulting from the application of an alternating voltage, $E = E_0 \cos \omega_c t$, on a two section lattice network. The current plotted is the current in the termination of the network. The frequency of the applied voltage is the resonant frequency, f_c, of the network.

and six sections are shown on Fig. 7 and Fig. 8. Fig. 7 shows the transient terms and Fig. 8 the complete solution. Fig. 9 and Fig. 10 show similar curves for a frequency twice the resonant frequency of

the network. In addition, the solution for infinite frequency is readily obtained from (39) since for this frequency $\varphi = 90°$. Then

$$i = \frac{E}{2R}[\cos(\omega t + \theta - m\pi)].$$

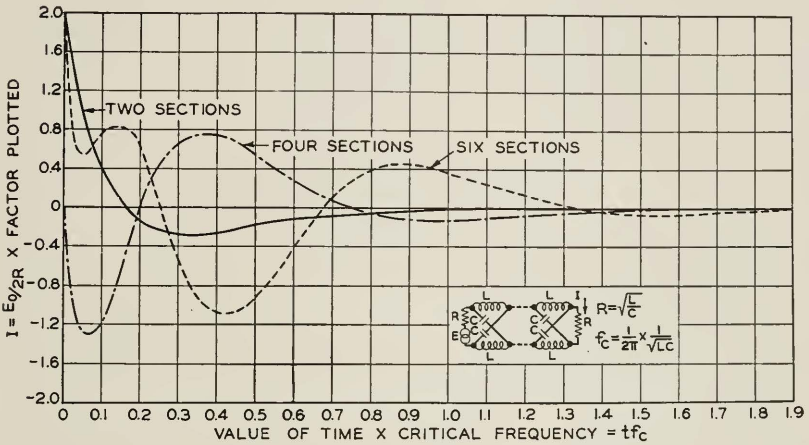


Fig. 7—Transient current resulting from the application of an alternating voltage, $E = E_0 \cos \omega_c t$, for several sections of lattice network. The current plotted is the current in the termination of the network. The frequency of the applied voltage is the resonant frequency, f_c, of the network.

III. Laplacian Infinite Integral Equation and Its Formal Solution

The solution of circuit problems by means of the Laplacian integral equation has been used by Carson [4] to a large extent. It is interesting to note that this integral form can be derived in a simple manner by means of this expansion method, and that this method provides a means for solving the Laplacian integral equation.

Any impedance Z is made up of resistances, inductances and capacities, and hence the current i can be represented by a series

$$i = \frac{E}{Z} = E[a_0 + a_1 e^{-j2\omega D} + a_2 e^{-j4\omega D} + \cdots + a_n e^{-j2n\omega D} + \cdots]. \quad (41)$$

The interpretation of this expansion from a physical standpoint is that the current is Ea_0, for the first interval of time $2D$, $E[a_0 + a_1 e^{-j2\omega D}]$ for the next interval of time $2D$, etc. Hence at the time $t = n(2D)$, the current i will be given by the sum of n terms

[4] See "Electric Circuit Theory and the Operational Calculus," *B. S. T. J.*, October 1925, and following.

of this series. We can, therefore, express the current i at the time t by the integral

$$i = E\left[\int_0^t a(t)e^{-j\omega t}dt + a_0\right], \tag{42}$$

where the value of $a(t)$ for any interval of time $(n-1)2D$ to $n(2D)$ is the constant of the above series a_{n-1} divided by $2D$. The value of this integral for an infinite time must reduce to the steady state value of $i = E/Z$, hence

$$\frac{E}{Z} = E\left[\int_0^\infty a(t)e^{-j\omega t}dt + a_0\right].$$

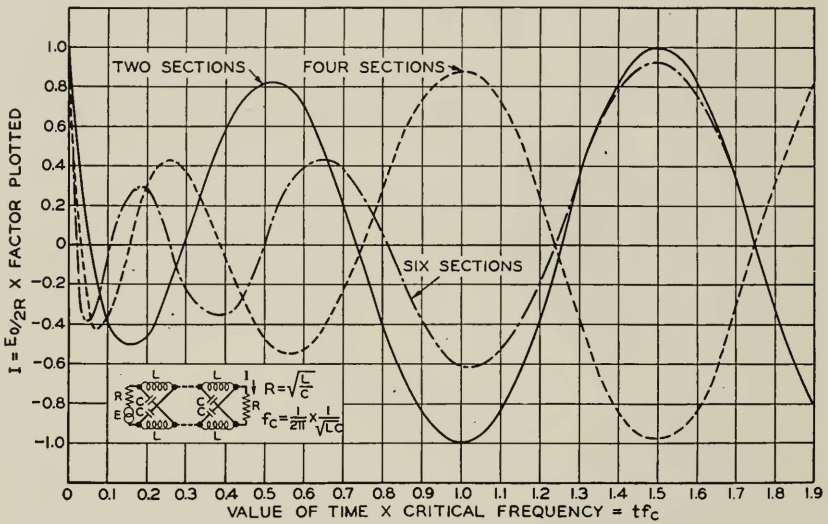


Fig. 8—Current resulting from the application of an alternating voltage, $E = E_0 \cos \omega_c t$, on several sections of lattice network. The current plotted is the current in the termination of the network. The frequency of the applied voltage is the resonant frequency, f_c, of the network.

Cancelling out the common factor E, we have the infinite integral equation

$$\frac{1}{Z(j\omega)} = \left[\int_0^\infty a(t)e^{-j\omega t}dt + a_0\right]. \tag{43}$$

The physical interpretation of the quantity $a(t)$ is readily obtained by reference to equation (42). If we set $\omega = 0$ and $E = 1$ in this equation we have

$$i = \int_0^t a(t)dt + a_0 = \int_0^t a(t)dt + h(0),$$

where $h(0)$ is a constant denoting the current when t is zero. Now i at any time t is the indicial or direct current admittance, designated by $h(t)$, hence $a(t)$ is

$$a(t) = \frac{d}{dt}(h(t)).$$

The infinite integral equation (43) takes the form

$$\frac{1}{Z(j\omega)} = \left[\int_0^\infty \frac{d}{dt}(h(t))e^{-j\omega t}dt + h(0)\right]. \tag{44}$$

This integral equation does not have quite the same form as Carson's integral equation but can be readily put into that form by means of

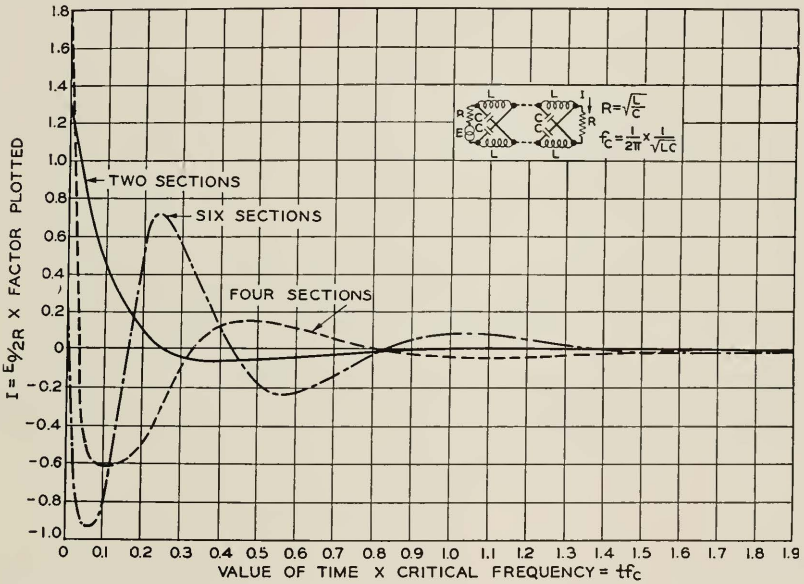


Fig. 9—Transient current resulting from the application of an alternating voltage, $E = E_0 \cos 2\omega_c t$, on several sections of lattice network. The current plotted is the current in the termination of the network. The frequency of the applied voltage is twice the resonant frequency, f_c, of the network.

Borel's theorem [5] which is given below. Suppose that $1/Z'$ and $1/Z''$ are two admittances, which when multiplied together give the admittance $1/H$. The admittances $1/Z'$ and $1/Z''$ have the expansions

$$\frac{1}{Z'} = [a_0' + a_1'e^{-2j\omega D} + a_2'e^{-4j\omega D} + \cdots];$$

$$\frac{1}{Z''} = [a_0'' + a_1''e^{-2j\omega D} + a_2''e^{-4j\omega D} + \cdots].$$

[5] See *B. S. T. J.*, October 1925, page 722.

The expansion $1/H$, then takes the form

$$\frac{1}{H} = \frac{1}{Z'Z''} = [a_0 + a_1 e^{-2j\omega D} + \cdots + a_n e^{-2nj\omega D} + \cdots]$$
$$= [a_0 a_0'' + [a_0 a_1'' + a_0'' a_1']e^{-2j\omega D} + \cdots$$
$$+ [a_0 a_n'' + \cdots + a_k a_{n-k}'' + \cdots + a_n a_0'']e^{-2nj\omega D} + \cdots].$$

We have then the relation

$$a_n = [a_0 a_n'' + \cdots + a_k a_{n-k}'' + \cdots + a_n a_0''].$$

Now a_n'' is the value of a'' when $t = n(2D)$, hence the above relation can be put into the form of an integral

$$a(t) = \int_0^t a''(\tau)a'(t - \tau)d\tau,$$

where $\tau = k/2D$, and the complete relation is

$$\frac{1}{H} = \frac{1}{Z'Z''} = \int_0^\infty a(t)e^{-j\omega t}dt + a_0$$
$$= \int_0^\infty \left[\int_0^t a''(\tau)a'(t - \tau)d\tau\right] e^{-j\omega t}dt + a_0 a_0''.$$

Suppose now that we let $Z' = j\omega$; $Z'' = Z(j\omega)$. We know from Heaviside's rule and from Section I that $1/j\omega$ has the direct current or indicial admittance solution

$$\frac{1}{j\omega} = t = h(t).$$

Hence

$$a'(t) = \frac{d}{dt}h'(t) = \frac{d}{dt}t = 1 \qquad \text{and} \qquad a_0' = 0,$$

and the infinite integral equation

$$\frac{1}{H} = \frac{1}{j\omega Z''(j\omega)} = \int_0^\infty \left[\int_0^t a''(\tau)a'(t - \tau)d\tau\right] e^{-j\omega t}dt + a_0 a_0''$$

takes the form

$$\frac{1}{j\omega Z''(j\omega)} = \int_0^\infty \left[\int_0^t a''(\tau)d\tau\right] e^{-j\omega t}dt = \int_0^\infty h''(t)e^{-j\omega t}dt.$$

Dropping the primes, we have the Laplacian integral equation

$$\frac{1}{j\omega Z(j\omega)} = \int_0^\infty h(t)e^{-j\omega t}dt. \tag{45}$$

Hence (43) is equivalent to Carson's integral equation, if $(j\omega)$ is replaced by p.

It will be noted that in deriving this equation use is made only of the general form of the expansion of admittances. For particular admittances, the values of the a's in equation (43) or the h's in equation (45) can be derived directly from an expansion of the admittance function, as shown in the foregoing work. Hence, if the solution of

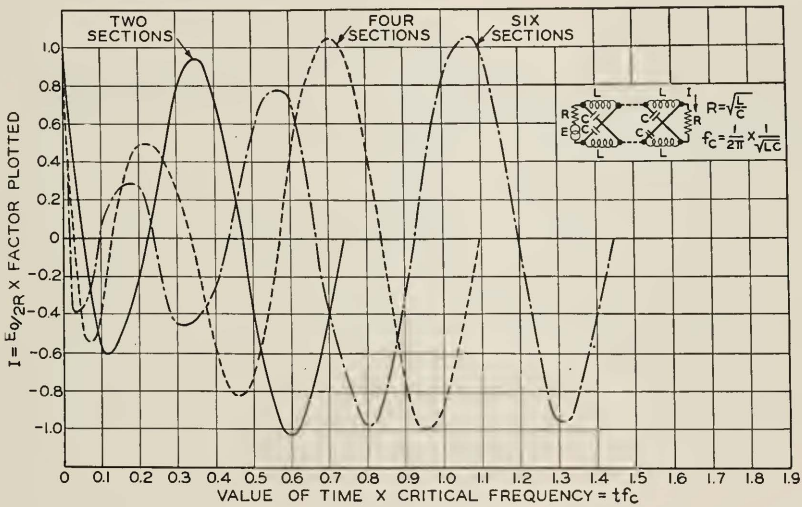


Fig. 10—Current resulting from the application of an alternating voltage, $E = E_0 \cos 2\omega_c t$, on several sections of lattice network. The current plotted is the current in the termination of the network. The frequency of the applied voltage is twice the resonant frequency, f_c, of the network.

the integral equation is not known from a table of integrals, one method for obtaining its solution is the expansion method developed above. This method may then have some application as a method for solving integral equations.

A. Illustrative Example

As an illustration of the use of this method in solving integral equations we will consider the equation

$$\frac{1}{\sqrt{(j\omega)^2 + 2\lambda j\omega}} = \int_0^\infty a(t)e^{-j\omega t}dt. \tag{46}$$

The expression on the left can be written

$$\frac{1}{\sqrt{j\omega}\sqrt{2\lambda + j\omega}}. \tag{47}$$

Noting that the square of the first factor has the form of an inductance and the second the form of a resistance and inductance in series we replace

$$j\omega \to R_{0_1}\left(\frac{1 - e^{-2P_1}}{2}\right) \qquad \text{where} \qquad R_{0_1} \to \infty\,;\ P_1 \to 0$$

$$R_{0_1}P_1 = j\omega$$

and $2\lambda + j\omega \to R_{0_2}\left(\dfrac{1 - e^{-2P_2}}{2}\right)$ where $R_{0_2} \to \infty\,;\quad P_2 \to 0\,;$ and $R_{0_2}P_2 = 2\lambda + j\omega$. We note that P_1 has the form $j\omega D$ where $R_{0_1}D = 1$, while P_2 has the form $A + j\omega D$ where $R_{0_2}A = 2\lambda$ and $R_{0_2}D = 1$. Substituting these values in (47), we have

$$\frac{1}{\sqrt{\dfrac{R_{0_1}R_{0_2}}{4}(1 - e^{-2j\omega D})(1 - e^{-2(A+j\omega D)})}}.$$

Expanding these two factors by the binomial theorem, we have

$$\frac{2}{\sqrt{R_{0_1}R_{0_2}}}\left[1 + \frac{1}{2}e^{-j(2\omega D)} + \frac{1/2 \times 3/2}{2!}e^{-j(4\omega D)} + \cdots + \frac{(2n)!e^{-j(2n\omega D)}}{2^{2n}(n!)^2} + \cdots\right]$$
$$\times\left[1 + \frac{1}{2}e^{-2(A+j\omega D)} + \cdots + \frac{(2n)!e^{-2n(A+j\omega D)}}{2^{2n}(n!)^2} + \cdots\right].$$

At this point we make use of Stirling's theorems on factorials which states that when K is large

$$K! = \left(\frac{K}{e}\right)^K\sqrt{2\pi K}.$$

The typical terms of the above expression become

$$\frac{(2n)!e^{-j(2n\omega D)}}{2^{2n}(n!)^2} = \frac{\left(\dfrac{2n}{e}\right)^{2n}\sqrt{4\pi n}\ e^{-j(2n\omega D)}}{2^{2n}\left(\dfrac{n}{e}\right)^{2n}2\pi n} = \frac{e^{-j(2n\omega D)}}{\sqrt{\pi n}}.$$

Inserting this value in the above expression and multiplying, there results

$$\sqrt{\frac{4}{R_{0_1}R_{0_2}}}\left[1 + e^{-j(2\omega D)}[1/2e^{-2A} + 1/2] + \cdots + e^{-j(2n\omega D)}\right.$$
$$\left(\frac{e^{-2nA}}{\sqrt{\pi n}} + \frac{e^{-2(n-1)A}}{2\sqrt{\pi(n-1)}} + \cdots + \frac{e^{-2(n-k)A}}{\sqrt{\pi(n-k)}\sqrt{\pi k}} + \cdots \right. \tag{48}$$
$$\left.\left. + \frac{1}{\sqrt{\pi n}}\right) + \cdots\right].$$

Since the value of $a(t)$ is given by the factor multiplying the term $e^{-j2n\omega D}$ divided by $2D$ we can write

$$a(t) = \frac{2}{\sqrt{R_{0_1}R_{0_2}}2D}\left[\frac{e^{-2nA}}{\sqrt{\pi n}} + \frac{e^{-2(n-1)A}}{2\sqrt{\pi(n-1)}} + \cdots \right.$$
$$\left. + \frac{e^{-2(n-k)A}}{\sqrt{\pi(n-k)}\sqrt{\pi k}} + \cdots + \frac{1}{\sqrt{\pi n}}\right]. \tag{49}$$

This expression can be written as the sum

$$a(t) = \sum_{k=0}^{k=n} \frac{e^{-2(n-k)A}}{\pi\sqrt{(n-k)k}}.$$

We introduce now the value $n = t/2D$ and define a new variable τ by $k = \tau/2D$. Inserting these values in the above summation and noting that $A/D = 2\lambda/R_{0_2}/1/R_{0_2} = 2\lambda$, we have

$$a(t) = \sum_{k=0}^{k=n} \frac{e^{(t-\tau)2\lambda}2D}{\pi\sqrt{(t-\tau)\tau}}. \tag{50}$$

But $2D = d\tau$, the element of time, so that the summation can be written as the integral

$$a(t) = \frac{e^{-2\lambda t}}{\pi}\int_0^t \frac{e^{2\lambda\tau}d\tau}{\sqrt{(t-\tau)\tau}}. \tag{51}$$

The value of the integral is $\pi e^{\lambda t}I_0(\lambda t)$, where $I_0(\lambda t)$ is the Bessel's functions $J_0(i\lambda t)$, when $i = \sqrt{-1}$. To show this we can expand the exponential and integrate the series giving

$$\pi\left[1 + \lambda t + \frac{1.3(\lambda t)^2}{(2!)^2} + \frac{1.3\cdot 5(\lambda t)^3}{(3!)^2} + \cdots\right].$$

This can be recognized as the series expansion of $e^{\lambda t}I_0(\lambda t)$. Hence

the value of (51) is

$$a(t) = e^{-\lambda t} I_0(\lambda t), \tag{52}$$

which is the solution of the integral equation.

IV. Other Types of Boundary Conditions

The solutions obtained before are all on the assumption that no energy exists in the network before the voltage is applied. Other types of boundary conditions are sometimes desirable, but these can all be derived from the above solutions.

The next most important case is the case where the network has come to its equilibrium value and the voltage is suddenly taken off. This condition is the same as would result if a negative voltage E were suddenly applied to the circuit, and hence the solution is the steady state value of the current minus the current which flows on application of the voltage E.

Another type of boundary condition sometimes occurring is the one where energy exists in the network when t is zero. This may be taken account of by assuming that the voltage is applied before t equals zero. To take account of this condition analytically, examine the expansion

$$i = \frac{E}{Z} = E[a_0 + a_1 e^{-2j\omega D} + a_2 e^{-4j\omega D} + \cdots + a_n e^{-2nj\omega D} + \cdots].$$

According to the above assumption, the voltage is applied when $t = -t_0$, hence for n we substitute

$$n = \frac{t}{2D} + \frac{t_0}{2D}.$$

The above series is then replaced by the integral

$$i = E \int_{-t_0}^{t} a(t + t_0) e^{-j\omega(t+t_0)} dt. \tag{53}$$

Another boundary condition of interest occurs when the voltage is taken off before an equilibrium value has been reached. If we count time as starting when the voltage is taken off, or what amounts to the same thing, when a negative voltage is applied, the symbolic solution takes the form

$$i = E\left[\int_{-t_0}^{t} a(t + t_0) e^{-j\omega(t+t_0)} dt - \int_{0}^{t} a(t) e^{-j\omega t} dt\right]. \tag{54}$$

Appendix

The expression

$$\left[\frac{4R}{R_0}\left(\frac{1}{1-e^{-2(R/R_0+P)}}\right)-1\right]^m$$

can be expanded into the form

$$\left[\frac{4R}{R_0}\left(\frac{1}{1-e^{-2(R/R_0+P)}}\right)\right]^m - m\left[\frac{4R}{R_0}\left(\frac{1}{1-e^{-2(R/R_0+P)}}\right)\right]^{m-1} + \cdots + (-1)^m. \quad (55)$$

and hence the general solution depends only on the solution of the general form

$$\left[\frac{4R}{R_0}\left(\frac{1}{1-e^{-2(R/R_0+P)}}\right)\right]^m. \quad (56)$$

If equation (56) is expanded by the binomial theorem, there results the expression

$$\left(\frac{4R}{R_0}\right)^m\left[1+me^{-2(R/R_0+P)}+\frac{m(m+1)}{2!}e^{-4(R/R_0+P)}+\cdots + \frac{(m+n-1)!e^{-2n(R/R_0+P)}}{n!(m-1)!}+\cdots\right]. \quad (57)$$

For any value of m, we can write the typical term of (57) as

$$\frac{(m+n-1)!}{n!(m-1)!} = \frac{\left(\frac{m+n-1}{e}\right)^{m+n-1}\sqrt{2\pi(m+n-1)}}{\left(\frac{n}{e}\right)^n\sqrt{2\pi n}(m-1)!}$$

$$= \frac{e^{-(m-1)}\left(\frac{m+n-1}{n}\right)^n(m+n-1)^{m-1/2}}{(m-1)!\sqrt{n}}. \quad (58)$$

by Stirling's theorem on factorials. Now n for any finite value of time approaches infinity, while m for any finite term in the series is finite. Hence (58) can be written as

$$\frac{e^{-(m-1)}\left(1+\frac{m-1}{n}\right)^n n^{m-1}}{(m-1)!}.$$

The limit of $\left(1+\frac{m-1}{n}\right)^n$ as $n\to\infty$ is $e^{(m-1)}$.[6] Hence

$$\frac{(m+n-1)!}{n!(m-1)!} = \frac{n^{m-1}}{(m-1)!}.$$

[6] See "Probability and Its Engineering Uses," T. C. Fry, page 107.

The value of (57), then reduces to

$$\left(\frac{4R}{R_0}\right)^m \sum_{t=0}^{t=n} \frac{n^{m-1}}{(m-1)!} e^{-2n(R/R_0+P)}. \tag{59}$$

If now we substitute $n = t/2D$, $P = j\omega D$; $R_0P = j\omega L$ (59) reduces to the integral

$$\left(\frac{4R}{R_0}\right)^m \times \frac{1}{(2D)^m} \int_0^t \frac{t^{m-1} e^{-[(1/\sqrt{LC})+j\omega]t}}{(m-1)!} dt$$
$$= \left(\frac{2}{\sqrt{LC}}\right)^m \times \frac{1}{(m-1)!} \int_0^t t^{m-1} e^{-t[(1/\sqrt{LC})+j\omega]} dt. \tag{60}$$

If we integrate (60) by parts, successively, there results the series

$$2^m \left\{ \left(\frac{1}{1+j\omega\sqrt{LC}} \right)^m - e^{-t(1/\sqrt{LC}+j\omega)} \left[\frac{1}{(1+j\omega\sqrt{LC})^m} + \frac{\left(\frac{t}{\sqrt{LC}}\right)}{(1+j\omega\sqrt{LC})^{m-1}} + \cdots + \frac{\left(\frac{t}{\sqrt{LC}}\right)^{m-1}}{(m-1)!\,(1+j\omega\sqrt{LC})} \right] \right\}. \tag{61}$$

The complete solution of (55) is then obtained by adding terms of the kind given in (61). For example the steady state term is given by the series

$$\frac{2^m}{(1+j\omega\sqrt{LC})^m} - \frac{m2^{m-1}}{(1+j\omega\sqrt{LC})^{m-1}} + \frac{\frac{m(m-1)}{2!}2^{m-2}}{(1+\omega\sqrt{LC})^{m-2}} + \cdots + (-1)^m$$
$$= \frac{2^m - m2^{m-1}(1+j\omega\sqrt{LC}) + \frac{m(m-1)}{2!}2^{m-2}(1+j\omega\sqrt{LC})^2 + \cdots + (-1)^m(1+j\omega\sqrt{LC})^m}{(1+j\omega\sqrt{LC})^m}.$$

This is readily seen to be the binomial expansion of

$$\left[\frac{2-(1+j\omega\sqrt{LC})}{1+j\omega\sqrt{LC}}\right]^m = \left[\frac{1-j\omega\sqrt{LC}}{1+j\omega\sqrt{LC}}\right]^m. \tag{62}$$

The other terms given in equation (38) follow directly by addition and reference to equations (55) and (61).

Acoustic Considerations Involved in Steady State Loud Speaker Measurements

By L. G. BOSTWICK

Synopsis: Certain difficulties encountered in acoustic measurements of the performance of loud speakers are described. Because of the nature of these difficulties it has not yet been possible to specify a complete and simple set of measurements or conditions which will completely express the performance of a loud speaker. Data are given showing the performance of two representative types of loud speakers both when measured in outdoor space free from reflections and when measured under varying conditions in a specially treated acoustic laboratory. The differences serve to emphasize the importance of certain precautions in the making of indoor acoustic measurements.

Scope

In view of the general misconception of the meaning of many claims which are made regarding the operation of loud speakers, it appears desirable to discuss in some detail the requirements which should be taken into account in making measurements for setting up ratings of loud speaker performance. For example, claims to "uniform response at all frequencies in the audible range" or "flat characteristic" can not be accurately applied to loud speakers which have thus far been made available. In many cases the claims for a loud speaker are based upon carefully made electrical measurements but these are often obtained in such a manner and under such conditions that they do not represent the performance of the loud speaker as it would be normally observed, and therefore are misleading. The main consideration in making loud speaker measurements is not the electrical circuit arrangement or apparatus of the measuring system but rather the acoustic conditions under which the magnitude of the sound output of the loud speaker is determined. It is the purpose of this paper to discuss steady state loud speaker measurements particularly from the standpoint of the more important acoustic factors which are involved and which must be properly considered in order to be able to interpret the significance of any measurements obtained.

Loud Speaker Performance Indices and Factors Involved in Their Determination

Efficiency.—In power engineering and other branches of engineering, efficiency (a measure of the degree to which a device performs the functions for which it is designed) is defined as the ratio of the power delivered to a load to the power absorbed from a source of

supply. Since in power transmission systems the purpose of a machine is to draw a limited amount of power from a relatively unlimited source and to deliver this power to a load with a minimum loss in the machine itself, this ratio constitutes a useful measure of the performance. If it were of interest a similar quantity could likewise be used as a measure of the performance of a loud speaker. In this latter case however, the function is not in general to draw a limited amount but as much power as possible from a supply source and to radiate maximum power to the air or load. A measure of the efficiency would therefore have to involve the ability of the loud speaker to take maximum power from the supply and might be defined as the ratio of acoustic power P_A radiated to the maximum electrical power P_E which the supply circuit is capable of delivering under optimum impedance conditions. Thus the efficiency η at a specified frequency would be defined by the ratio

$$\eta = \frac{P_A}{P_E}. \tag{1}$$

Assuming the impedance of the electrical supply source for a loud speaker essentially non-reactive (as is almost invariably the case) and of a constant magnitude r suitable to the requirements of the loud speaker, then the maximum power which the supply circuit is capable of delivering under optimum impedance conditions with an open circuit supply voltage e would be

$$P_E = \frac{e^2}{4r}. \tag{2}$$

These quantities are all readily measureable. The determination of the quantity P_A however is more difficult.

For measuring the acoustic energy or power stored in or transmitted through the medium adjacent to a loud speaker, the condenser transmitter is probably the most suitable free space acoustic measuring device. The ruggedness of this transmitter for an instrument of this type and the straightforward manner in which it can be used recommend it for practical loud speaker measurements. The condenser transmitter is not, however, an acoustic power indicating device but is a device having a high impedance compared to the impedance of the acoustic system in which it is used. It is therefore an acoustic measuring device which is analogous to an electrical voltmeter and can be calibrated by the thermophone [1] or other means to measure the excess pressure in a medium resulting from a sound wave. Other acoustic measuring devices such as the Rayleigh disc, thermal devices

[1] "The Thermophone," E. C. Wente, *Physical Review*, Vol. XIX, No. 4, April, 1922.

of different kinds, etc. can of course be used but these in general are considerably more difficult to use than the condenser transmitter, especially for free space acoustic measurements and the measured quantities bear no more simple relation to the acoustic power or energy.

Assuming the condenser transmitter then to be the acoustic measuring or indicating device, the problem becomes one of how and where in the medium to measure the pressure so that the measurement will bear some readily deducible relation to the acoustic power delivered by the loud speaker. The answer depends upon the nature of the acoustic medium in which the measurements are made. The simplest relations between excess r.m.s. pressure in the medium and the acoustic power exist when the pressure measurements are made in an infinite medium or in a room in which the incident energy at the walls is completely absorbed. Under such conditions the acoustic power from a loud speaker could be obtained by measuring the pressure at all points on the surface of a sphere having a radius several times that of the largest dimension of the sound radiating surface and with the sound radiating surface at the center of the sphere. The acoustic power would then be

$$P_A = \frac{1}{\rho c} \int\int p^2 ds,^2 \tag{3}$$

where ρ is the density of the air; c is the velocity of sound propagation; p is the excess r.m.s. pressure; and ds is the surface of the sphere. This relation, however, is generally true only when the radius of the spherical surface is sufficiently large so that the sound radiating surface appears as a point source. From equations (1), (2) and (3) the efficiency of a loud speaker could then be expressed in terms of excess pressure measurements over the surface of the sphere in an infinite medium as follows:

$$\eta = \frac{P_A}{P_E} = \frac{1}{\rho c} \frac{\int\int p^2 ds}{\frac{e^2}{4r}} = \frac{K \int\int p^2 ds}{\frac{e^2}{r}} . \tag{4}$$

Within an enclosure where there are sound reflections from the bounding surfaces, the determination of the acoustic power delivered by a loud speaker would involve the measurement of the pressure at

[2] "Theory of Vibrating Systems and Sound," Crandall, pp. 92 and 120.

all points within the enclosure in order to obtain the average energy density and then making use of known relations between energy density and the rate of energy flow into the room. Under steady state conditions, the total potential energy stored within the room would be

$$E_p = \frac{1}{2\rho c^2}\int\int\int p^2 dv.^3 \tag{5}$$

Assuming the room to be large so that the region within say a wave-length of the loud speaker is a small portion of the total volume of the room, it can be said with a reasonable approximation that the potential and kinetic energies stored within the room as the sound is transmitted are equal. The total energy would therefore be twice the potential energy or

$$E = 2E_p = \frac{1}{\rho c^2}\int\int\int p^2 dv. \tag{6}$$

This latter statement may be roughly justified in a simple manner by considering the sound radiated by the loud speaker as consisting of two components, one of which is completely absorbed at the walls and the other completely reflected. In considering separately that component which is absorbed, the loud speaker can be thought of as in an infinite medium and under these conditions (excluding the region within say a wave-length of the loud speaker) the acoustic impedance of the medium is essentially non-reactive. The potential and kinetic energies of the sound transmitted would therefore be equal. That component which is transmitted to the medium and completely reflected at the walls produces an ideal standing wave system. In such a system along the direction of the standing wave the total energy is alternately all kinetic and all potential and since this transition takes place the potential and kinetic energies must be equal.

Considering the room volume V, the average energy density would be

$$\bar{E} = \frac{1}{\rho c^2 V}\int\int\int p^2 dv. \tag{7}$$

If the loud speaker emits power into the room at a rate P_A, the average energy density in the room after a steady state has been reached is

$$\bar{E} = \frac{4P_A}{ac},^4 \tag{8}$$

[3] "The Dynamical Theory of Sound," Lamb, p. 208, Second Edition.
[4] "Theory of Vibrating Systems and Sound," Crandall, p. 210.

where a is the absorbing power of the room obtained from the sum of the products of the areas of the absorbing surfaces in the room and their respective absorption coefficients. From (7) and (8) the acoustic power delivered by the loud speaker could then be expressed in terms of the excess r.m.s. pressure throughout a large room as follows:

$$P_A = \frac{a}{4V\rho c}\int\int\int p^2 dv \tag{9}$$

and the efficiency of the loud speaker would therefore be

$$\eta = \frac{P_A}{P_E} = \frac{\dfrac{a}{4V\rho c}\displaystyle\int\int\int p^2 dv}{\dfrac{e^2}{4r}} = \frac{K'\displaystyle\int\int\int p^2 dv}{\dfrac{e^2}{r}}. \tag{10}$$

Response.—By determining the mean square pressure at all points in the measuring room or at all points on the surface of a large sphere in an infinite medium as discussed above, it is therefore possible to measure the "efficiency" of a loud speaker with a pressure indicating device. Such a method for determining the merits of a loud speaker at all frequencies of usual interest, however, would obviously be quite impracticable. Furthermore, unless the radiation from the loud speaker is uniform over a spherical surface it is not of particular interest to know the magnitude of the total acoustic power or the value of the quantity η since the configuration of the sound field about a loud speaker may change decidedly with frequency, with the result that variations in sound loudness at different frequencies in a particular region may be large even though the total power output from the loud speaker may be constant. In order then that the measured characteristic shall convey a true idea of the performance as it might be observed by the ear, the square of the pressure at one representative listening position or the average of the squares of the pressures in a small region wherein an observer might normally be located may be considered instead of the average throughout the room. In this manner a sort of specific efficiency measure would be obtained in that it is a measure of the efficiency with respect to the acoustic power transmitted through the specified position or region. Throughout the remainder of this paper, this specific measure of the efficiency is called

the "response" of the loud speaker as measured at a specified position or positions and is expressed in transmission units (TU) relative to a fixed arbitrary reference condition of 1 volt, 1 ohm, and 1 bar. In other words the acoustic power density at a specified position or the average acoustic power density at specified positions in the medium produced by the loud speaker under test per unit of available electrical supply power, is expressed relative to the acoustic power density produced by a fictitious standard loud speaker which when placed in the location of the loud speaker under test will produce a mean square pressure of one unit at the specified position or positions in the room when the ratio

$$\frac{e^2}{r} = 1.$$

The response in TU is thus expressed by the relation:

$$\mathrm{TU} = 10 \log_{10} \frac{\dfrac{K'\overline{p^2}}{\dfrac{e^2}{r}}}{\dfrac{K'\overline{1^2}}{1}} = 10 \log_{10} \frac{\overline{p^2}}{\dfrac{e^2}{r}}$$

$$= 20 \log_{10} \frac{\bar{p}}{\dfrac{e}{\sqrt{r}}},$$

where $\bar{p}$ is the r.m.s. pressure at a specified position or the square root of the mean square r.m.s. pressures at specified positions in the medium; and e and r are as defined above.

Measuring Considerations in a Reflectionless Medium.—In a medium where there are no sound reflections from the bounding surfaces, two factors are most likely to cause the measured response of a loud speaker to vary with frequency. These factors are independent and their effects of about equal importance. The first and most apparent is the inherent dynamical characteristics of the loud speaker which involves its ability to transfer maximum power from the electrical supply to the acoustic medium. Any variation with frequency in the acoustic power output of a loud speaker when supplied by constant, available electrical power will, of course, cause corresponding variations in the response provided the square of the pressure at the measuring position is indicative of the power transmitted through this position. In order that this latter condition be strictly true, the measuring position in general should be, at a distance from the loud speaker,

large compared to the dimensions of the radiating surface. Otherwise, the wave front at the measuring position would not be spherical and the indicated pressure might result largely from the cyclic storage and absorption of energy by the loud speaker in the immediate vicinity of the radiator. With a proper location of the condenser transmitter, however, a response-frequency characteristic provides a useful measure of the dynamical perfection of a loud speaker.

The second factor which may cause large variations in the response of a loud speaker in a reflectionless medium is the change in the space distribution of the radiated sound with frequency. Although the total acoustic power delivered by a loud speaker may be constant, the power density at certain positions in the medium may change greatly with frequency due to the interference of sound originating at different parts of the radiating surface. Unless the radiation from the loud speaker is spherical, this interference phenomenon will result in a concentration of sound power in certain regions in the medium and a diminution in others. The locations of these regions change with frequency, radiator dimensions and the mode of vibration of the radiating surface.

For the case of a piston diaphragm radiator in a large rigid wall, it is possible to calculate the variations with frequency in the excess pressure at points in the sound field. Such calculations [5] and confirming experimental data show that in the sound field along the center perpendicular (a line normal to the surface of the piston at the center) to a piston radiator there is a succession of sound pressure maxima and minima out to a distance equal to approximately $\frac{D^2 f_1}{4500}$ feet (where D is the piston diameter and f_1 is the energizing frequency). Beyond this distance these maxima and minima points disappear and the pressure varies inversely as the distance. If then, the response of a loud speaker with a piston diaphragm is measured with the condenser transmitter at a distance less than $\frac{D^2 f}{4500}$ feet (where f is the highest measuring frequency), the response-frequency characteristic will have a succession of peaks and depressions which are independent of but which may be difficult to distinguish from those caused by poor dynamical characteristics of the loud speaker itself. On the other hand, if the condenser transmitter is located at any distance greater than $\frac{D^2 f}{4500}$ feet, the response-frequency characteristic obtained will not

[5] "The Directional Effect of Piston Diaphragms," Backhaus and Trendelenburg, *Zeitschrift f. Techn. Physik.*, Vol. 7, pp. 630–635, 1926. Also "Theory of Vibrating Systems and Sound," Crandall, pp. 137–149.

have abrupt irregularities due to interference and any two curves so obtained will only differ in magnitude. A curve obtained under this latter condition would therefore show the response-frequency variations as these would be observed at any distance greater than $\frac{D^2f}{4500}$ feet in this same direction.

While the above facts relate to a piston diaphragm radiator because more definite statements can be made regarding its sound field, similar effects due to the irregular distribution of the sound field are involved in the case of any other loud speaker which does not radiate as a pulsating sphere. Using such piston diaphragm considerations as a basis it has been found possible to predict suitable measuring conditions for any particular loud speaker. The fundamental requisite is that the pressure indicator be located at a distance from the loud speaker commensurate with the typical listening distance in order that the response-frequency characteristic shows variations which would normally be observed and which are therefore of interest. If, however, the typical listening distance is greater than $\frac{D^2f}{4500}$ feet (where D is roughly the diameter of the radiating surface and f is the highest measuring frequency) response-frequency measurements at this distance will show the response-frequency variations at any greater distance in the same direction so that measurements at the greater distances would not be necessary. If the most likely position of a listener is at a distance less than $\frac{D^2f}{4500}$ feet, the response-frequency characteristic obtained with the condenser transmitter at such a position will have irregularities due to interference but since these irregularities would be heard they should be charged against the loud speaker and such a curve would be indicative of the performance. In this latter case response measurements at a distance greater than $\frac{D^2f}{4500}$ feet are sometimes valuable for loud speaker design work in order to distinguish those variations due to poor dynamical characteristics of the loud speaker itself and those due to poor sound field distribution characteristics.

Measuring Considerations in a Medium with Reflections.—If the sound energy reflected to the condenser transmitter position from the bounding surfaces of the medium is comparable in magnitude with the energy reaching this position directly from the loud speaker, standing waves will exist and the sound pressure may vary greatly with frequency at any fixed transmitter position although the acoustic power

transmitted through this position may be constant. Response measurements with the condenser transmitter at any one position can therefore mean very little under such conditions.

To our knowledge it is practically possible only by working outdoors under very particular conditions to obtain a medium sufficiently free from reflections to make suitable response measurements at all frequencies with the condenser transmitter located at any one position. By using a room with all dimensions very large compared to the distance between the condenser transmitter and the loud speaker (which distance is determined by the size of the loud speaker and the highest measuring frequency as discussed above) and covering the walls with sound absorbing material, it is possible to reduce the reflected energy at the transmitter position to a small value over a considerable frequency range but any practical method of reducing the reflected sound to a negligible value at all frequencies of interest in loud speaker measurements is as yet not available.

In a plane standing wave system the energy density at points of maximum pressure or minimum velocity is equal to $\frac{p^2}{\rho c^2}$, where p is the r.m.s. pressure at these points. The locations of these maximum pressure points change with frequency but if the position of the condenser transmitter in loud speaker response measurements is changed at each measuring frequency to a maximum pressure point within a suitable region the indicated pressure will be a measure of the energy transmitted through this region. The measured response would then be approximately the same as would be measured in a reflectionless medium except for a magnitude difference due to the addition of the reflected energy. Such a procedure for loud speaker response measurements indoors would thus be suitable if it were not for the fact that the standing wave system in the room is usually of a very complicated configuration in three dimensions instead of being simple. The probability of being able to locate a position in any desired region of the sound field of a loud speaker where the pressures of each of the standing waves which may traverse this position are a maximum, is obviously remote.

A method of response measurement making use of the mean square pressure instead of the maximum value is more practicable. With a single frequency sinusoidal sound source, the pressure-space distribution of each of the standing waves in a room is likewise sinusoidal. The maximum r.m.s. pressure squared of each standing wave would then be twice the mean r.m.s. pressure squared over a half wave-length or any multiple of a half wave-length; also approximately twice the mean over

any distance large compared to a wave-length since this latter average approaches the half wave-length mean. The mean square pressure is therefore just as suitable as the maximum value as a measure of the energy density and further, it lends itself more readily to the determination of the energy density in the case where there are standing waves in several directions. For this latter case the energy density within a specified region is proportional to the mean square pressure in all directions or at all points within the volume of a sphere having a diameter large compared to the wave-length. The response of the loud speaker at any frequency can accordingly be measured in a room with reflections by averaging the squares of the pressures throughout a suitable volume.

The above method of response measuring indoors however, is not entirely independent of the measuring room. If the reflected energy is large the response measurements will be affected by the variation in the absorption power of the room with frequency so that a large room with absorbing material having as uniform and large absorption characteristics as possible over the measuring frequency range is still desirable. Some sound absorbing materials have uniform absorption characteristics but when this is the case the absorbing power is apt to be very low. The use of such materials results in extremely large pressure variations within the room so that a measuring device having a sufficient amplitude range to average the squares of the pressures is difficult to obtain. For this reason and because of the fact that the region through which it is necessary to average the squares of the pressures at low frequencies becomes prohibitive, a large room is most desirable so that the difference between the direct and reflected energies at the transmitter position will be as large as possible. Indoor measurements under these conditions approach infinite medium measurements.

The use of a large room also results in less reaction of the room inclosure on the loud speaker itself. While under most conditions such reactions have little effect on the acoustic output power of the loud speaker, in small measuring rooms at very low frequencies where the absorption is low and the radiator of the loud speaker (perhaps designed for a large auditorium) is large, the phase and magnitude of the reflected energy at the radiating surface of the loud speaker may be such as to cause large variations in the acoustic impedance of the medium on the area adjacent to this surface. This variation in the acoustic impedance of the loud speaker load will cause variations in the acoustic power density at the transmitter position. Response measurements on loud speakers of large dimensions and particularly measure-

ments on such devices at very low frequencies are consequently more indicative of the capabilities of the loud speaker when obtained outdoors or in a very large room.

Experimental Data

General.—To determine the extent to which the above discussed acoustic effects may influence the result of a loud speaker performance measurement and to show the measuring conditions under which such acoustic factors are encountered, response-frequency characteristics of loud speakers were measured under various acoustic conditions. These are described in the following paragraphs.

In all these measurements a calibrated condenser transmitter approximately $2\frac{7}{8}''$ in diameter was used as the acoustic detector. The thermophone calibration on this transmitter showed that with constant pressure on the diaphragm the voltage produced between the grid and filament of the associated vacuum tube was very nearly independent of frequency. Corrections for such small variations as did exist however have been made in all the following curves. In addition a tapered correction of .6 TU per 100 cycles increased in frequency between 1,100 and 2,100 cycles and a constant correction of 6 TU for frequencies above 2,100 cycles have been subtracted from the response measurements to correct for the fact that the indicated pressure approaches twice the free space pressure at frequencies at which there is reflection from the diaphragm. This latter correction was obtained by making response-frequency measurements on a loud speaker under a fixed set of conditions; first, with the condenser transmitter freely suspended in the sound field as in all the following curves and then with the transmitter located at the center of a round baffle 12″ in diameter. When in the baffle complete sound reflection from the transmitter occurred at a frequency lower than that at which reflection began to take place from the transmitter alone. From the difference between the two response-frequency curves so obtained, it was therefore possible to definitely locate the transition frequency range between 1,100 and 2,100 cycles and to evaluate the transmitter reflection correction.

The number of measurements made in order to define any response-frequency characteristic depended upon the nature of the curve. If no abrupt changes in the response were observed in making the measurements, approximately 10 measurements per octave were obtained. Otherwise the frequency of the oscillator was changed by small steps and a sufficient number of measurements made to clearly define the curve.

Measuring System.—The circuit arrangement of the measuring system used in making the measurements is shown in schematic form on Fig. 1. An oscillator having a suitable frequency range and power output was alternately connected by means of a two-position switch through a transformer to the loud speaker under test and to the input terminals of an attenuator calibrated in TU. The transformer had a ratio such that the loud speaker being measured was always connected to an impedance equal to that for which it was designed to be connected. The condenser transmitter in series with the output terminals of the attenuator and located in the medium as will be discussed later, was connected to a voltage indicating system consisting of an amplifier, a thermocouple and a meter. A low-pass filter was included in the

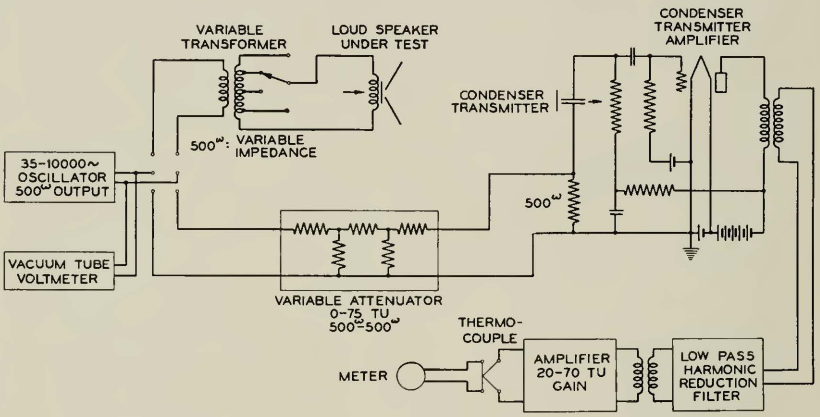


Fig. 1—Schematic circuit of loud speaker response measuring system.

indicating system as shown to insure that only the fundamental frequency of the output from the condenser transmitter or attenuator was indicated. The measuring procedure was as follows: The output or terminal voltage of the oscillator when open-circuited or connected to the attenuator was kept constant at all frequencies by means of a vacuum tube voltmeter. With the loud speakers considered the sound output over a wide magnitude range was proportional to the oscillator voltage so that the magnitude of this voltage was governed entirely by the sound pressure in the medium most suitable for making measurements. With the oscillator connected to the loud speaker (through the transformer) the sensitivity of the voltage indicating system was adjusted at each frequency until a mid-scale deflection of the meter was obtained as a result of the voltage generated by the condenser transmitter. After each adjustment the oscillator was then switched from

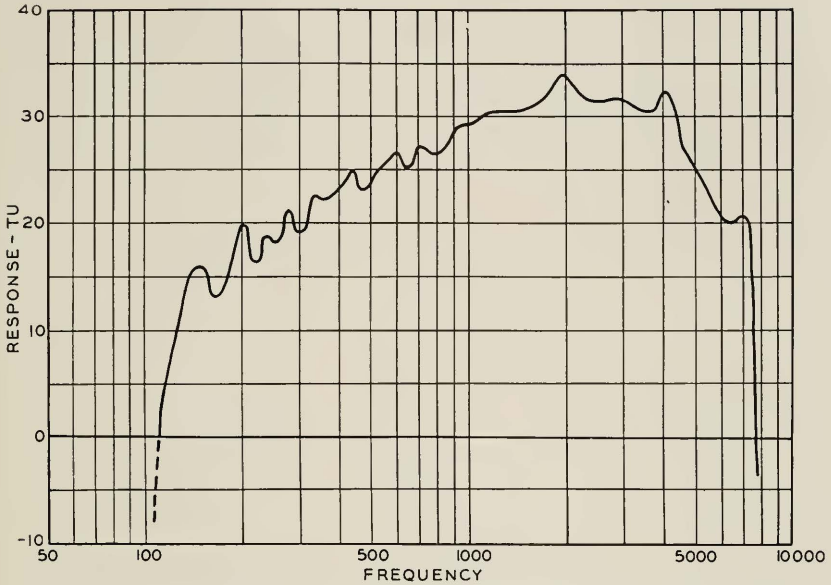


Fig. 2—Response-frequency characteristic of 115 cycle cut-off exponential horn with moving coil type receiver. Measured outdoors at a distance of 12′ from horn mouth on the axis.

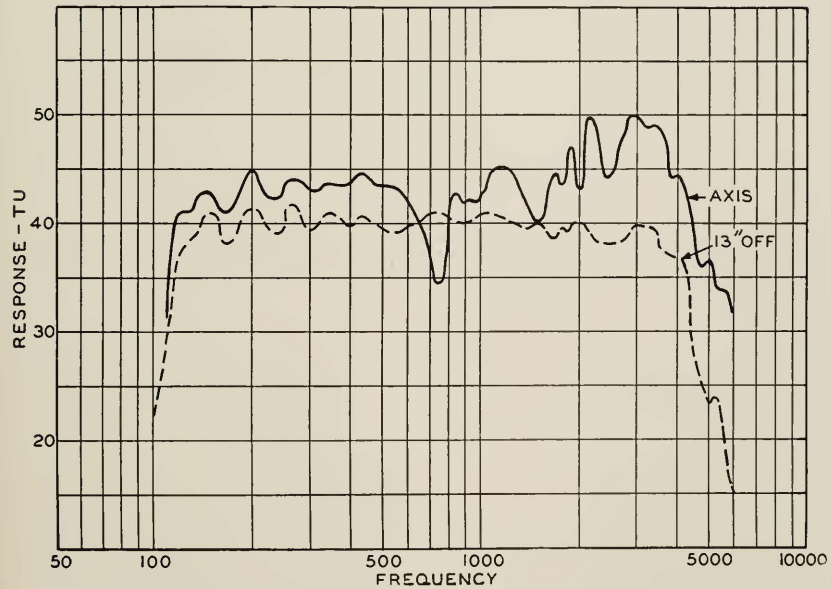


Fig. 3—Response-frequency characteristics of 115 cycle cut-off exponential horn with moving coil type receiver. Measured outdoors 2″ from plane of horn mouth with center of condenser transmitter diaphragm on horn axis and 13″ from axis.

the loud speaker to the input terminals of the attenuator and the attenuator adjusted to give the same meter deflection. The variations in the attenuator settings with frequency showed the variations in the performance of the loud speaker in TU. When the ratio of the open-circuit voltage of the oscillator to the square root of its output impedance equalled 1, the setting of the attenuator which gave a voltage between the attenuator output terminals equal to the voltage across the condenser transmitter terminals with a pressure of one bar on the diaphragm, gave the reference zero of one volt, one ohm, and one bar

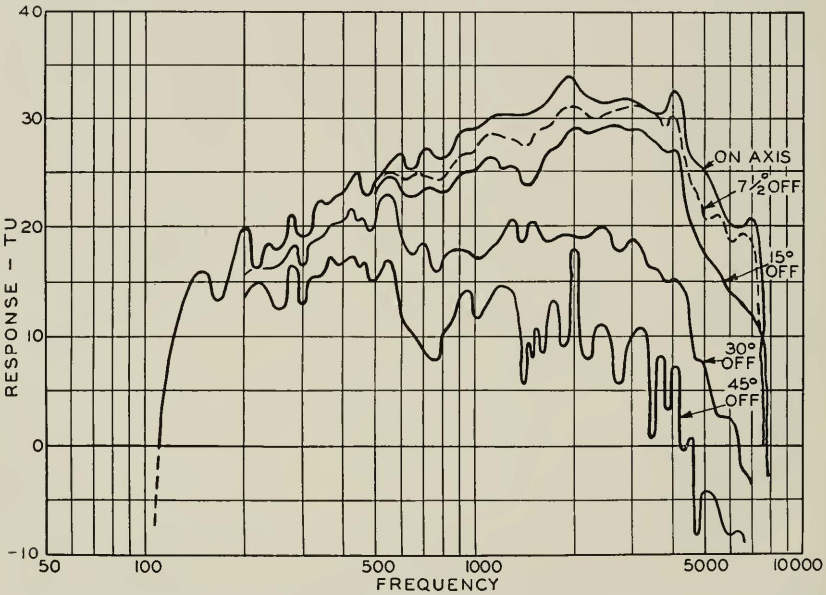


Fig. 4—Response-frequency characteristics of 115 cycle cut-off exponential horn with moving coil type receiver. Measured outdoors 12 feet from horn mouth at the specified angles with the axis.

described above. The response of the loud speaker was then read directly from the attenuator setting.

This measuring method has the commendable feature that the results obtained are independent of any variations from day to day in the sensitivity in the voltage indicating system. The voltage indicating system simply serves to compare the voltage produced by the condenser transmitter with the voltage across the output terminals of the attenuator and any variations in battery voltage, tubes, etc., or any variations in the amplification with frequency can in no way affect the accuracy of the measurements. Aside from the condenser transmitter, the only element in the measuring system which must be calibrated and

maintain its calibration closely, is the attenuator which involves only a group of resistances.

Outdoor Measurements.—As an illustration of those acoustic effects involved in loud speaker measurements in a reflectionless medium, the following response data obtained outdoors in an open field will be of

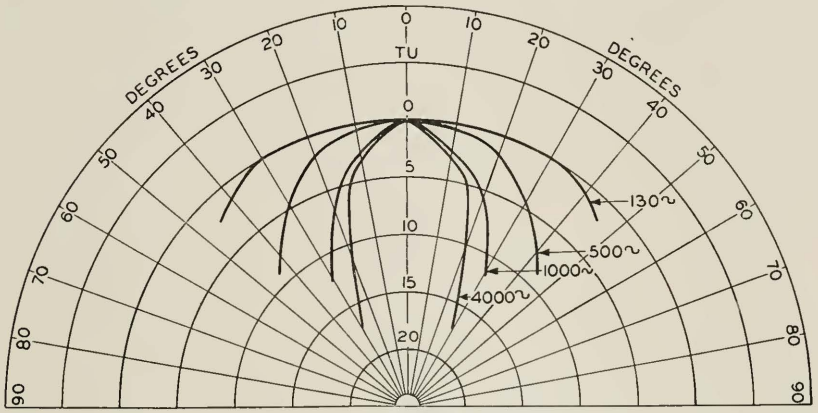


Fig. 5—Polar curves showing response (expressed relative to the axis response) of 115 cycle cut-off exponential horn with moving coil type receiver at various angles from horn axis and 12 feet from the mouth.

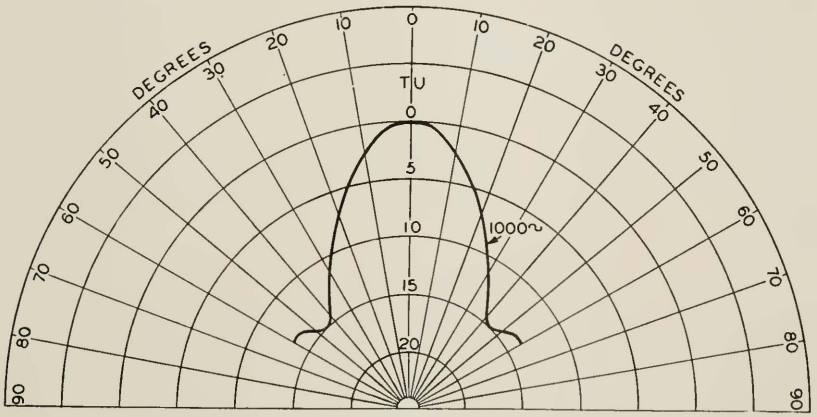


Fig. 6—Polar curve showing response (expressed relative to the axis response) of 115 cycle cut-off exponential horn with moving coil type receiver at different angles from axis and at a distance of 12 feet from center of horn mouth.

interest. Two loud speakers having uniform sound power output over a wide frequency range so that the dynamic characteristics would not obscure the acoustic effects were used. The loud speakers were placed at the edge of a skeleton platform approximately 15′ above the ground and the condenser transmitter suspended at various positions

in the sound field. Care was taken to suspend the transmitter and its small associated amplifier between small poles in such a way that any possible reflections from such objects in the sound field would not reach the transmitter position. As for reflections from the ground, the distance of the loud speaker from the ground with the consequent sound divergence from the radiating surface, and also the absorption and diffraction at the ground, caused by the magnitude of the sound reflected to the transmitter position to be quite undetectable.

One of the loud speakers was a 115 cycle cut-off exponential horn with a moving coil type receiver.[6] The mouth of the horn was located at the platform edge with the axis making an angle upward from



Fig. 7—Response-frequency characteristic of 3½″ piston diaphragm loud speaker. Measured outdoors 12 feet from and on a line perpendicular to the center of the diaphragm.

the horizontal of approximately 15°. Fig. 2 shows the response frequency characteristic with the condenser transmitter on the axis at a distance of 12′ from the mouth. Except for variations near the horn cut-off frequency due to the horn itself, note the absence of any large irregularities and also the rising trend of the curve with frequency. The diameter of the horn mouth was 30″ so that 12′ is greater than the distance $\frac{D^2f}{4500}$ feet discussed above.

Fig. 3 shows response-frequency characteristics of the same loud speaker measured under exactly the same conditions except that the condenser transmitter was located on the axis only 2″ from the horn

[6] This type of receiver was described by Wente and Thuras in *The Bell System Technical Journal*, for January, 1928.

mouth. This curve differs considerably from the one obtained at a distance of 12′. The marked depression in the curve at 750 cycles checks very closely the first interference frequency as calculated for a piston radiator approximately 30″ in diameter and allowing for a slight contraction of the radiating surface as the frequency is increased (which assumption would be quite reasonable for the horn), the irregularities at the higher frequencies are also explained in the same manner. For the piston, however, the minimum pressure point would be zero, which fact indicates that the wave-front at the horn mouth either is not plane or is not of uniform intensity over the radiating surface. Below 1,000 cycles the average trend of this curve is very nearly parallel to the axis of abscissæ while as noted for the curve

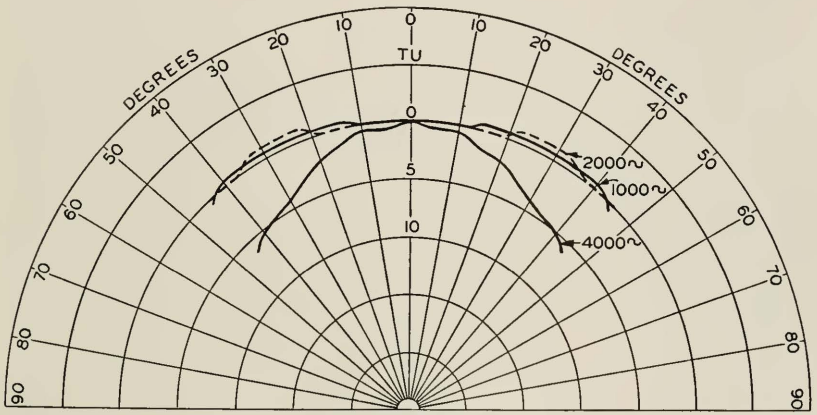


Fig. 8—Polar curves showing response (expressed relative to axis response) of 3½″ piston diaphragm loud speaker at various angles from perpendicular to center of diaphragm and 12 feet away.

obtained at a distance of 12′, there is a very definite downward slope. This is as would be expected if there were an increasing concentration of the sound field about the axis as the frequency increased. The fact that there is such a varying concentration is shown by the data on Fig. 4. These curves were obtained with the condenser transmitter at a distance of 12′ in each case, but with a line from the center of the horn mouth to the center of the transmitter making various angles with the horn axis as specified. In making these measurements, the transmitter remained fixed and the horn was rotated upward in a vertical plane about the center of the mouth. It is apparent from these curves that as the angle is increased the response at the higher frequencies becomes lower, while at lower frequencies the change is slight. The irregularities in the 45° curve are probably due to inter-

ference. Note that a curve of almost any desired trend may be obtained by locating the condenser transmitter at the proper position.

On Fig. 5 the data of Fig. 4 are plotted on polar coordinate paper to show more clearly the approximate manner in which the sound field varies. On these curves the magnitude of the response is expressed relative to that on the axis and the approximate distribution at each of four frequencies is shown. At the larger angles if a sufficiently large number of measurements are made an irregular interference pattern is obtained like that shown on Fig. 6 for 1,000 cycles.

The second loud speaker measured consisted of a $3\frac{1}{2}''$ piston diaphragm (inertia control) mounted in one side of a cubical box approxi-

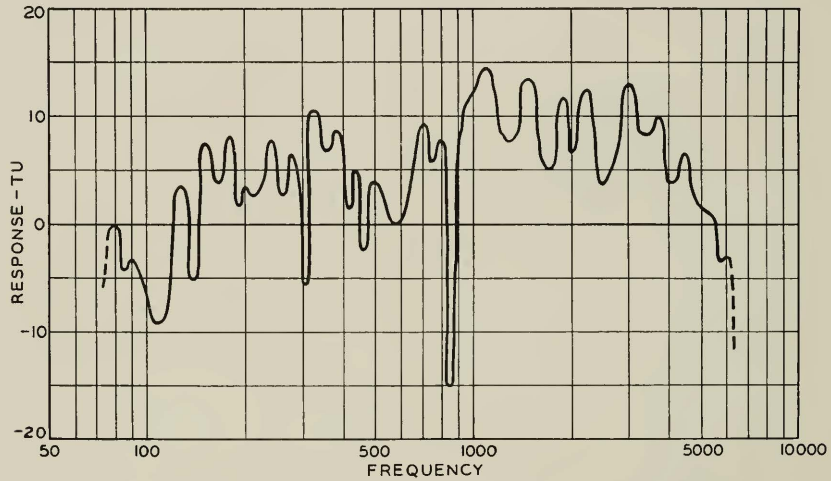


Fig. 9—Response-frequency characteristic of $3\frac{1}{2}''$ piston diaphragm loud speaker. Measured in highly absorbing room 12 feet from and on a line perpendicular to the center of the diaphragm.

mately 12″ on a side and filled with wool. This loud speaker was thus of a radically different type from the first and was chosen because the size and nature of the diaphragm was such that any reaction of the medium on its vibration amplitude would be unlikely. A direct comparison of its performance under one medium condition with that under another would therefore be justifiable. The response-frequency characteristic of this loud speaker as measured outdoors with the condenser transmitter at a distance of 12′ on the diaphragm center perpendicular is shown on Fig. 7. The irregularity in this curve at 600 cycles has been shown by other tests to be due to poor dynamical characteristics of the loud speaker itself.

On Fig. 8 are polar coordinate curves for the piston diaphragm loud

speaker similar to those shown on Fig. 5 for the horn type loud speaker. Note for the same frequency the greater concentration of the sound field in the case of the horn. This is due to its larger radiating surface. From these data it might be inferred that if the sound field of a loud speaker of either of these types is to be the same at all frequencies, the size of its radiating surface must decrease as the frequency increases.

Indoor Measurements.—Making use of the above described outdoor measurements as standards of comparison, response measurements were made indoors on the same loud speakers and at the same relative positions in the sound field in order to determine the magnitude of the effect of reflections on such measurements. The room available for

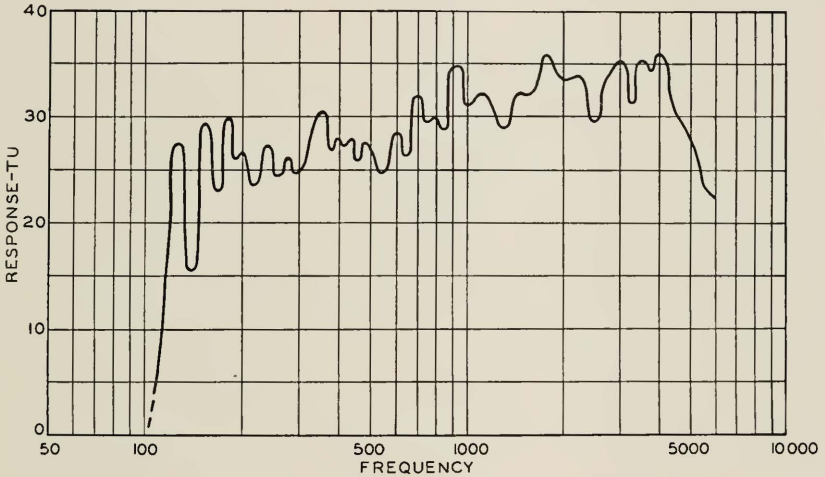


Fig. 10—Response-frequency characteristic of 115 cycle cut-off exponential horn with moving coil type receiver. Measured in highly absorbing room 12 feet from horn mouth on axis.

this work was approximately $34' \times 18' \times 9'$. The six bounding surfaces of the room were covered with $\frac{1}{2}''$ asbestos hair felt with heavy monks cloth curtains loosely and irregularly draped over the walls. The condenser transmitter and the loud speaker being measured were located along the major axis equi-distant from and on opposite sides of the center of the room and suspended about mid-way between the ceiling and the floor. A moderately large room with the usual precautions to eliminate reflections was thus used.

Fig. 9 shows the response frequency characteristic of the $3\frac{1}{2}''$ piston diaphragm loud speaker as obtained in this room with the condenser transmitter located on the diaphragm center perpendicular at a distance of $12'$. A comparison of this curve with that of the same loud

speaker obtained outdoors with the condenser transmitter at the same relative position in the medium gives an idea of the magnitude of the effect of room reflections even under comparatively favorable indoor measuring conditions. The same frequencies were measured in both the indoor and outdoor curves and no attempt was made to locate all the irregularities in the indoor curve.

Fig. 10 shows an indoor curve of the 115 cycle cut-off exponential horn measured in the same room and under the same conditions at a distance of 12′. The variations in this latter curve as compared to the outdoor curve shown on Fig. 2 appear to be less than in the case of the piston type loud speaker, probably because the horn is more directive,

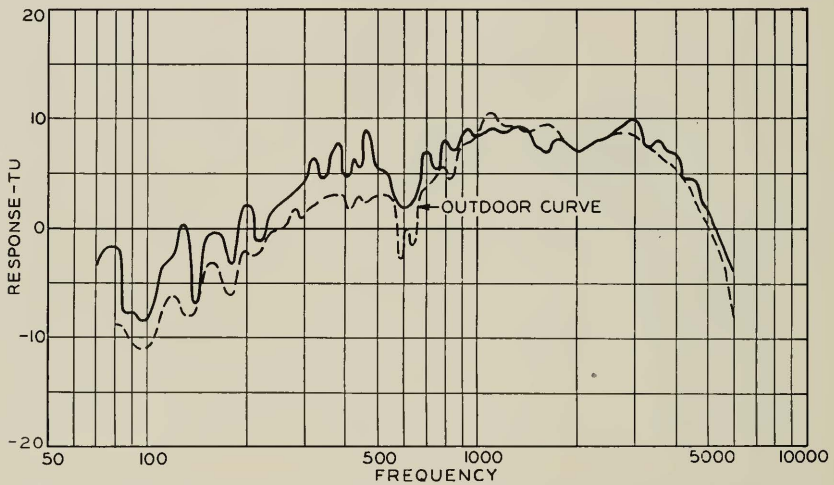


Fig. 11—Response-frequency characteristic of 3½″ piston diaphragm loud speaker. Measured in highly absorbing room 12 feet from diaphragm with rotating condenser transmitter.

resulting in a larger difference between the direct and the reflected sound energy at the transmitter position.

The variations in these two indoor response-frequency characteristics resulting from reflections could have been reduced by making the measurements in a much larger room treated with sound absorbing material in the same manner. Such a room for loud speaker measuring purposes is not usually available, but if sufficiently large and of a suitable shape does afford the most satisfactory indoor measuring conditions especially at very low frequencies.

Another method of obviating the effects of reflections is to measure the mean square pressure throughout a suitable volume as discussed previously. A practicable means of making such a mean square

pressure measurement and one which has been found to be quite satisfactory, consists in approximating the volume measurement by rotating the condenser transmitter in a circle 69″ in diameter, inclined 45° to the horizontal. A mechanism so arranged that the plane of the condenser transmitter diaphragm always remains perpendicular to the loud speaker axis is used and the condenser transmitter is connected to the same voltage indicating system described above. As noted, this indicating system involves a thermocouple and as a result, the meter deflection is very nearly proportional to the square of the input

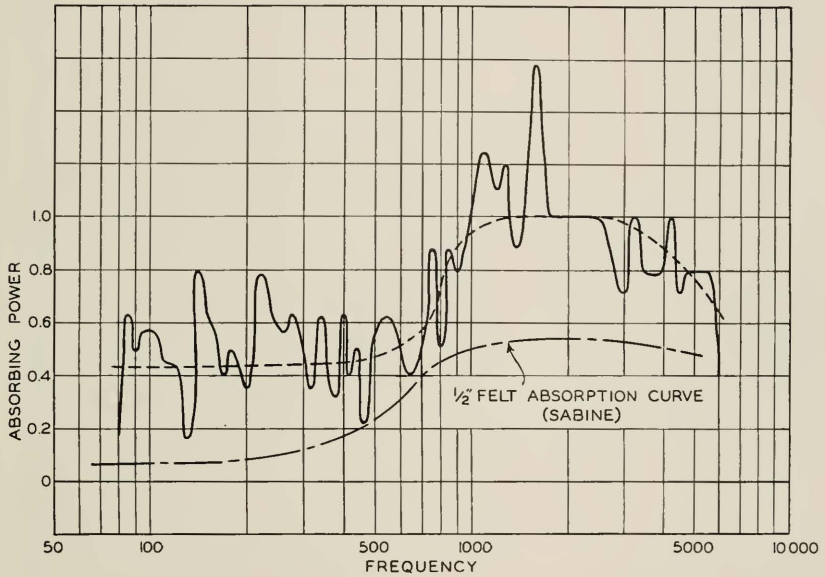


Fig. 12—Curve showing variation with frequency in the effective absorbing power of a felt lined room with respect to a region near the center and relatively near the sound source. Determined from loud speaker measurements in this room and outdoors as shown on Fig. 11.

voltage. As the condenser transmitter is rotated about the periphery of the circle, therefore, the average meter deflection is proportional to the average of the squares of the transmitter terminal voltages or the average of the squares of the pressures throughout the revolution.

Fig. 11 is a response-frequency characteristic of the piston diaphragm loud speaker measured in the same room and under the same conditions as the curve in Fig. 9 except with the rotating condenser transmitter. The center of the circle was located at the same point as the condenser transmitter for Fig. 9.

While rotating the transmitter in this manner does not average the

squares of the pressures throughout a volume, it does give a measure which is closely proportional to the power density. This is apparent from a comparison of Fig. 11 with a curve on the same loud speaker measured outdoors shown on Fig. 7 and replotted on Fig. 11 for comparison. Note that these two curves very closely coincide between 1,000 and 3,000 cycles. Below 1,000 cycles and above 3,000 cycles the uniformly greater response indoors can be explained in the following manner.

As given by equation (8) above, the average energy density in a room resulting from a loud speaker emitting sound power at a constant

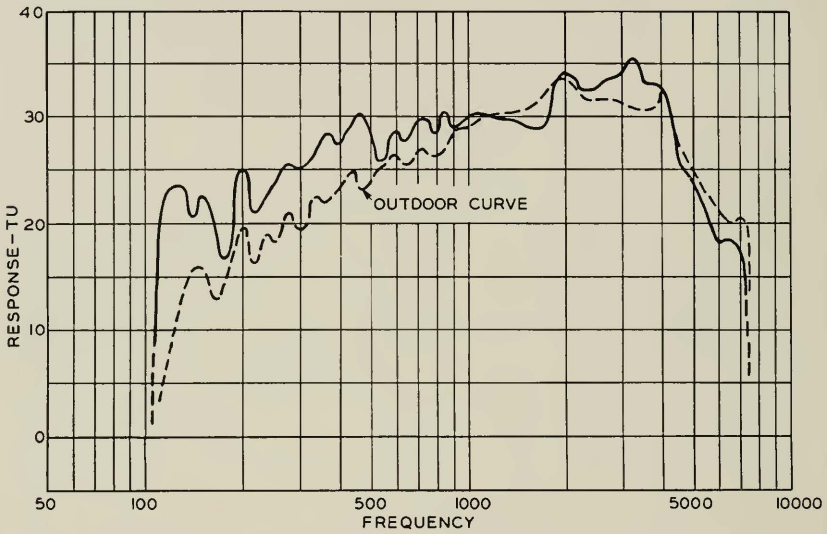


Fig. 13—Response-frequency characteristic of 115 cycle cut-off exponential horn with moving coil type receiver. Measured in highly absorbing room 12 feet from horn mouth with rotating condenser transmitter.

rate is inversely proportional to the absorbing power of the room. Inasmuch as the two curves in Fig. 11 were obtained with the same loud speaker with the condenser transmitter located at the same distance and at approximately the same relative position in the medium, the difference in the two curves would appear to be due only to the difference between the indoor and outdoor absorption. Assuming for the outdoor case a fictitious bounding surface making an enclosure of the same shape and size as the indoor room and that the energy striking this fictitious surface is completely absorbed (as would be the case) the area element of the factor "a" in equation (8) becomes the same for the indoor and outdoor tests, and the ratio of the indoor energy to the outdoor energy would therefore bear some simple

relation to the average absorption coefficient of the sound absorbing material of the indoor measuring room. From the ratios of the outdoor to indoor energy densities at different frequencies as determined from Fig. 11, the solid curve on Fig. 12 is obtained. The irregular character of this curve is probably due to the fact that the rotating indoor transmitter did not give an exact measurement of the energy density at each frequency. The trend of this curve, however, is quite definite as is indicated by the dotted curve which is an average curve obtained by inspection. A comparison of this mean curve with the dot-dash curve showing the approximate absorption power of a $\frac{1}{2}''$ layer of asbestos hair felt [7] indicates an interesting correlation in the trend of the two

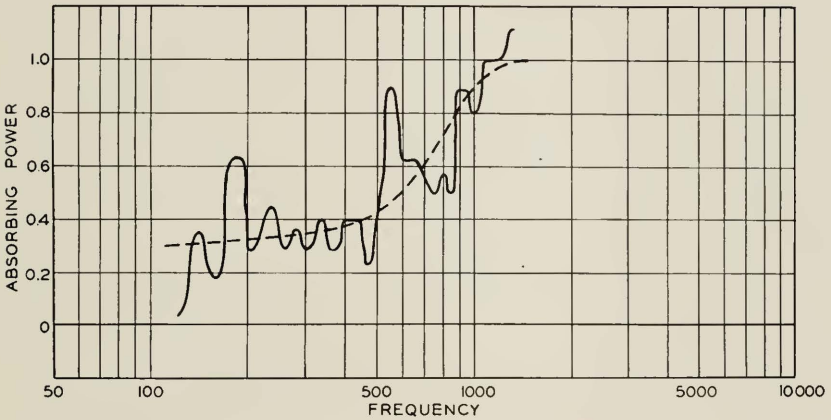


Fig. 14—Curve showing variation with frequency in the effective absorbing power of a felt lined room with respect to a region near the center and relatively near the sound source. Determined from loud speaker measurements in this room and outdoors, as shown on Fig. 13.

curves. The difference in magnitude is probably due to sound diffusion at the walls resulting in the ratio of the energy reflected to the energy direct from the source being smaller than would be the case were the measuring region located close to the absorbing surface.

The solid curve in Fig. 13 is a response-frequency characteristic of the 115 cycle cut-off exponential horn obtained with the rotating condenser transmitter under the same conditions as the curve in Fig. 11. The center of the horn mouth was placed at the same position as was the center of the diaphragm of the $3\frac{1}{2}''$ diameter piston radiator. The dotted curve in Fig. 13 is the outdoor characteristic obtained with the condenser transmitter on the horn axis at a distance of 12′. It will be noted that these indoor and outdoor curves diverge at the low

[7] See "Collected Papers on Acoustics," W. C. Sabine, p. 213, Fig. 4.

frequencies. From this divergence the absorption curve shown in Fig. 14 was obtained. Due to the relatively small angle subtended by the sound field of the horn at the higher frequencies, the indoor data where the condenser transmitter was rotated throughout a relatively large region in front of the horn, are not comparable with the outdoor data where the transmitter was located in one position on the horn axis.

Conclusion

From the above considerations it is obviously quite impossible to interpret the significance of response measurements on loud speakers in general unless such measurements are qualified by statements regarding the acoustic measuring conditions. Especially must information be given as to the position of the condenser transmitter relative to the loud speaker when the measurements were made, the method of measurement (pressure measured at one position or averaged within a region), and the size and nature of the medium. In general response measurements to be most indicative of the capabilities of the loud speaker should be made with the condenser transmitter at a distance from the loud speaker commensurate with or equivalent to the most likely listening distance of an observer.

To determine which of two loud speaker response-frequency characteristics is the better involves in addition to the above discussed acoustic considerations, an interpretation of the physiological significance of the magnitude and position in the frequency spectrum of departures in the curves from a straight horizontal line. Such an interpretation involves many physiological factors, the discussion of which is not within the scope of this paper. It should also be borne in mind that the response-frequency characteristics described in this paper are determined from steady state amplitude measurements and that they therefore give little information as to transient or phase distortion. However the cause of transient or phase distortion (the storage and release of energy in the reactive elements of the loud speaker) is also a cause of poor dynamical characteristics so that the peaks and depressions in a response-frequency characteristic may also be an indication of the phase and transient distortion. On the whole the response-frequency characteristic even though complicated by such a wide variety of factors has been found to be the most significant single criterion upon which to base a judgement of the merits of a loud speaker.

Recent Advances in Wax Recording [1]

By HALSEY A. FREDERICK

SYNOPSIS: This paper considers chiefly the frequency-response characteristics and limitations of the lateral cut "wax" record. It shows that the frequency range from 30 to 8,000 cycles can be recorded and reproduced from the record with practically negligible deviation from a flat frequency-response characteristic. The paper brings out the ease with which the record can immediately be replayed from the "wax" as an aid in assisting the artist to obtain the best results. A brief description is given of commercial processing methods including both plating and pressing. These methods give essentially a perfect copy of the original "wax." The time required for this work has been considerably reduced of late so that a test pressing can be obtained within three hours of the cutting of the original "wax."

IN the recording and reproducing of sound by the so-called "electric" method with the "wax" disc, the process may be considered as consisting of eleven steps. In order, these are: (1) studio, with its acoustic conditions, (2) microphone, (3) amplifier, (4) electro-mechanical recorder, (5) "wax" record, (6) copying or reproducing apparatus, (7) hard record or "pressing," (8) electric pickup, (9) amplifier, (10) loud speaker, (11) auditorium.

With this chain of apparatus the chief problem is that of making the reproduced sound in the auditorium a perfect copy of that in the studio. This is a matter of quality or fidelity of reproduction. There are other problems of cost, reliability, time required, etc., which are important but secondary to that of fidelity. While it may be necessary or convenient to introduce distortion in one of these links to compensate for such unavoidable distortion as may occur in other links, experience shows that it is desirable for the sake of simplicity, reliability and flexibility to reduce such corrective warping to a minimum and to make each step in the process as nearly perfect as possible. Perfection of a complete system may be judged by the practical method of listening to the overall result. It is necessary, however, to analyze each element of the complete system. To do this, other more analytical methods of test and standards of performance must be used. One of the most useful of these is the response-frequency curve. In order that all frequencies be reproduced equally and that the ordinary faults of resonance be avoided, this must be flat and free from sharp peaks. Good reproduction requires that frequencies from 50 to 5,000 cycles be included without discrimination. If, however,

[1] Presented before Society of Motion Picture Engineers at Lake Placid, New York, September 26, 1928.

the low frequency range be lowered to 25 or 30 cycles, a noticeable improvement will be obtained with some classes of music, whereas if the upper limit be increased to 8,000 or even 10,000 cycles, the naturalness and smoothness of practically all classes of reproduction will be noticeably improved.

A second important requirement in the judgment or analysis of any such system is that the ratio of output to input shall not vary over the range of currents or loudnesses (as well as frequencies) from the minimum up to the maximum used. If this requirement is not met, sounds or frequencies not present in the original reproduction will be introduced. This type of distortion has probably been heard by all of us in listening to an overloaded vacuum tube amplifier and is often referred to as "non-linear" distortion.

A third requirement not entirely disassociated from the first two is that any shifts in the phase relations shall be proportional to frequency.

Our judgment of the degree of perfection needed in sound reproduction systems is changing and growing more critical, so that what seemed excellent yesterday may be only fair today and tomorrow may seem intolerable. It is therefore necessary that our consideration and analysis be continually more searching and fundamental.

Of the eleven links in the chain of apparatus required for electric "wax" recording and reproduction, only five are peculiar to the "wax" method. These are the electromechanical recorder, "wax" record, the copying apparatus, the "pressing" and the pickup or reproducer. The extent to which the "wax" method is capable of the highest quality of reproduction will be disclosed by an examination of these five links. Any consideration of the practical advantages or disadvantages of the method can logically follow this examination into the quality possibilities.

The consideration which follows refers to the so-called "lateral" cut record; that is, a record in which the groove is of constant depth and oscillates or undulates laterally about a smooth spiral. This is the type used in the Western Electric Company disc record type of synchronized motion picture system. Some, but not all of the considerations and conclusions might apply to the "hill and dale" type record. It is not the purpose of this paper to consider the relative characteristics of "hill and dale" and "lateral" "wax" records.

Electromechanical Recorder

It is the task of the electromechanical recorder to take power from the amplifier and drive a mechanical recording stylus. The present-

day recorder is a highly developed apparatus based on extensive experimental as well as theoretical studies. A diagrammatic view is given in Fig. 1.[2] Recorders which have been supplied by the Western Electric Company have been designed to operate over a range of frequencies from 30 to 5,500 cycles. A typical frequency characteristic

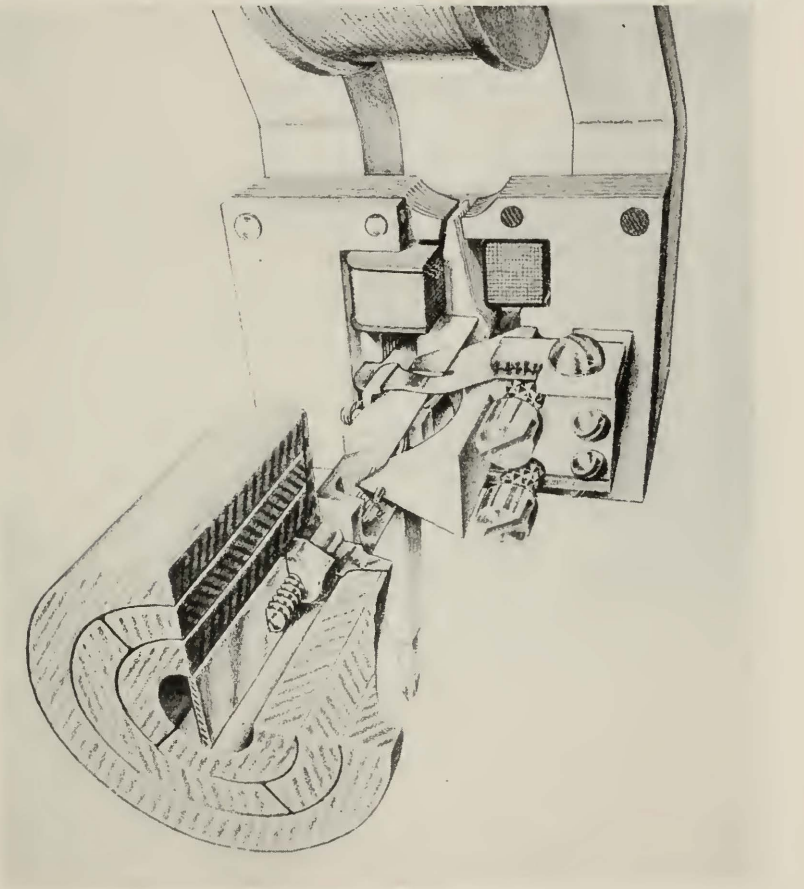


Fig. 1—Diagrammatic view of the electromechanical recorder.

is shown in Fig. 2. The device operates in linear fashion over the range of amplitudes involved in speech and music. As is seen, the response falls off below about 250 cycles. This falling characteristic is necessary in order that the maximum loudness be obtained from a record for a given spacing between grooves without cutting over

[2] "High Quality Recording and Reproducing of Music and Speech," by J. P. Maxfield and H. C. Harrison, presented at 14th Midwinter Convention of the American Institute of Electrical Engineers, New York, N. Y., Feb., 1926.

from groove to groove. In order that a lateral oscillation in a groove may represent constant intensity of sound or a constant energy over a range of frequencies, not the amplitude of the oscillation but the velocity, which is proportional to the product of the amplitude and the frequency, must be maintained constant. With the characteristic shown with these recorders, constant velocity is obtained from about

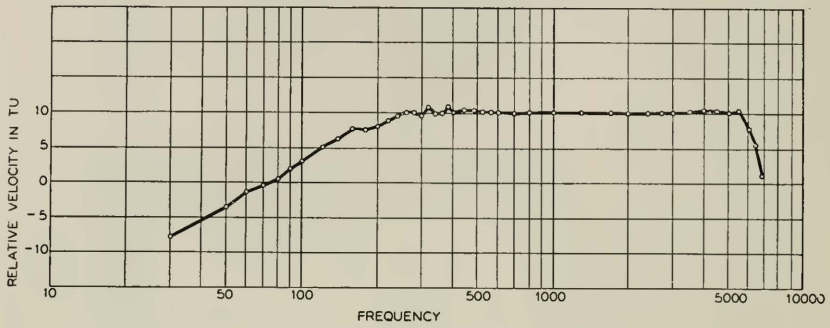


Fig. 2—Typical frequency characteristic of a commercial recorder.

250 cycles to 5,500 cycles. Below 250 cycles an approximately constant amplitude is obtained. If, therefore, sounds of constant absolute intensity are to be recorded over this range of 30 to 250 cycles, there is equal tendency for sounds of the different frequencies in this range to over-cut the record groove. It may be corrected in reproduction by a suitable electric network. Such a network will increase the subsequent amplification required but, as this additional

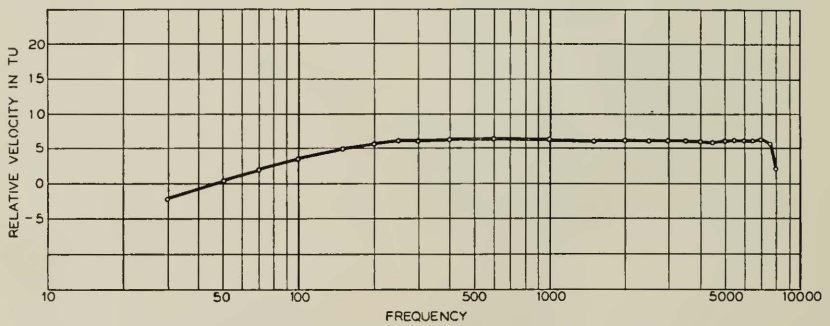


Fig. 3—Frequency characteristic of a laboratory model recorder.

amplification occurs in the first stages, it is not expensive. Practically it has not been found necessary or desirable to introduce such a corrective network since the correction has been largely cared for by the characteristics of the pickups used.

Recent development studies have established the possibility of flattening the response at the low frequency end and of raising the

high frequency cut-off of the recorder. Fig. 3 shows a characteristic obtained with such a laboratory model. This shows uniform performance within ± 1 TU from 250 to 7,500 cycles and within ± 4 TU from 30 to 8,000 cycles. Although its immediate practical value might be limited by other portions of the system, this device is of great interest in that it establishes beyond question the fact that an extremely broad range of frequencies can be successfully recorded in the "wax."

The broad, flat characteristic obtained with electric recorders has been made possible by so designing their elements that they constitute correctly designed transmission systems. In such a transmission system, whether it be an electrical recorder or a long telephone line, a correct terminating impedance is required. The load imposed by the "wax" is somewhat variable but fortunately is rather small. It has been found desirable to make the other impedances in the recorder relatively large so as to dominate the system and thus minimize the effects of any changes in the impedance imposed on the stylus by the "wax." The mechanical load used as a terminating impedance and to control the device has consisted of a rod of gum rubber 25 cm. long. Torsional vibrations are transmitted along this rod. The rate of propagation is about 3,000 centimeters per second so that its length is equivalent to an ideal electrical line of about 1,500 miles. The dissipation along this rubber rod is such that a vibration is substantially dissipated by the time it has travelled down the line and back. It thus constitutes substantially a pure mechanical resistance, the magnitude of this resistance being approximately 2,500 mechanical abohms, referred to the stylus point as its point of application.

"Wax" Record [3]

In recording the usual procedure is to use a disc from 1 in. to 2 in. thick and from 13 in. to 17 in. in diameter, composed of a metallic soap with small amounts of various addition agents to improve the texture. This is shaved to a highly polished surface on a lathe. This polished disc or so-called "wax" is placed in a recording machine. In Fig. 4 is shown what is essentially a high grade lathe arranged to rotate the "wax" in a horizontal plane at a very uniform speed in a definite relation to the film with which it is synchronized. The recorder with its cutting tool or stylus is moved relative to the surface of the disc, common phonograph procedure being to record from the outer edge of the disc towards the center, whereas in the Western

[3] "Some Technical Aspects of the Vitaphone," by P. M. Rainey, presented at the meeting of the Society of Motion Picture Engineers at Norfolk, Va., April, 1927.

Electric Company theater system the direction of cutting is reversed. After a record has been cut into the "wax," the "wax" may be handled and with proper precautions readily shipped from place to place.

The shape of the groove varies somewhat in commercial practice. The groove and stylus most commonly used with Western Electric apparatus are shown in Fig. 5. The groove is approximately .006 in. wide and .0025 in. deep. The pitch of the groove is between .010 in. and .011 in. so that the space between grooves is about .004 in. Thus

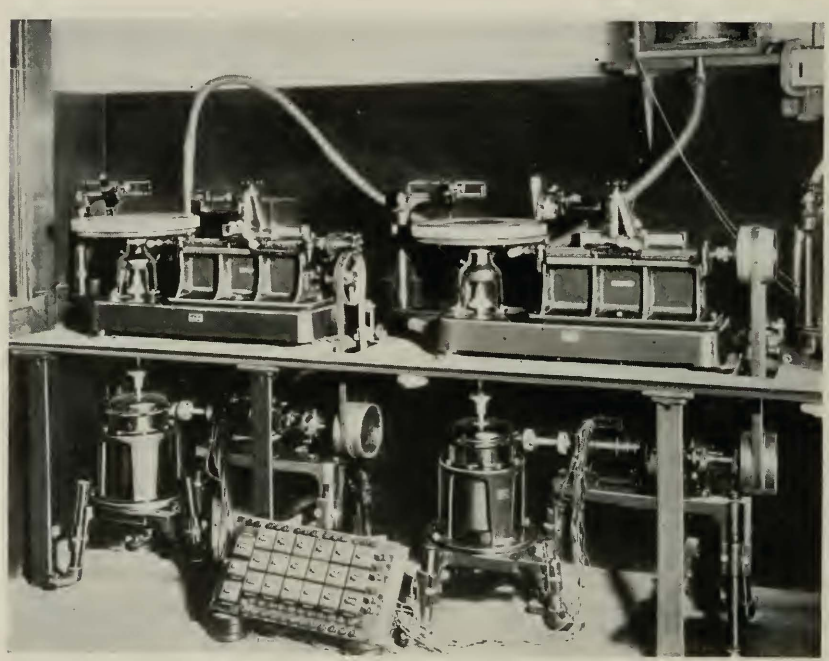


Fig. 4—The recording machine.

the maximum safe amplitude is about .002 in. If this occurs at 250 cycles the corresponding amplitude at 5,000 cycles, assuming constant absolute intensity of sound over this range, would be .0001 in.

It is important that a smooth groove be cut as any roughness in the walls introduces extraneous noise in the reproduced sound. To insure a truly smooth groove the surface of the "wax" must be shaved to a high polish. The texture of the "wax" must be fine and homogeneous which requires not only that the "wax" composition be correct, but that it be operated at the proper temperature. "Waxes" may be obtained commercially which will operate satisfactorily over

the ordinary range of room temperature. The "wax" must be levelled in the recording machine with reasonable care. The stylus must be sharp and so ground that the cut will be very clean. The "wax" shaving is removed as cut by air suction. The operator is aided in maintaining the correct depth of cut by the use of a so-called "advance" ball which rides lightly on the "wax" and serves to maintain uniform depth of cut in spite of small inaccuracies of leveling of the "wax" or deviations from planeness. The "advance" ball is adjusted relative to the stylus by observing the cut with a calibrated microscope. A satisfactory operation of the recording machine requires an ordinarily skilled mechanic with reasonable experience.

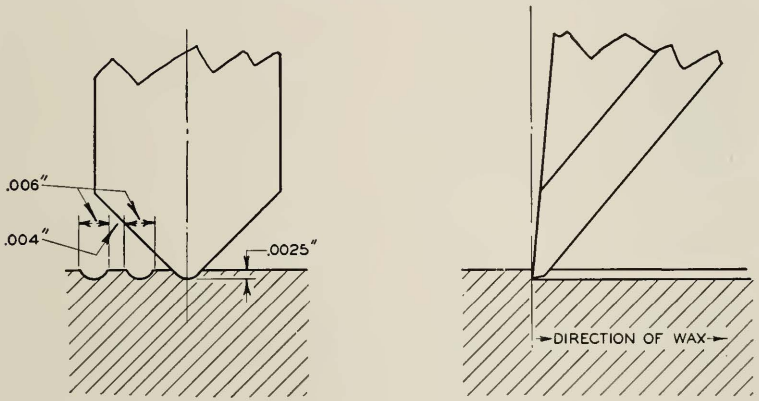
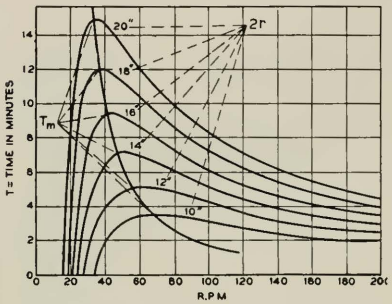


Fig. 5—Recorder stylus.

The rate of rotation is dependent upon the diameter of the record groove which is determined primarily by the length of time which it is desired to have covered by a single disc. The controlling element is the linear speed of the groove past the recorder or reproducer. In the Western Electric system the speed varies from 70 ft. to 140 ft. per minute, in other words, of the same order of magnitude as with the film record. The wave-lengths also are about the same as for a sound record on a film. At the minimum linear speed the half wave-length for a 5,000 cycle wave is .0014 in. If the minimum linear speed is fixed at 70 ft. per minute and the groove spacing is fixed, there is an optimum relation between the size of the record, the rate of rotation and the playing time. This is illustrated graphically in Fig. 6.

After a record has been cut, the sound may be reproduced directly from the "wax" by using a suitable pickup or reproducer. Ordinary reproducers or pickups rest much too heavily on the records to be

used on ordinary "wax." That this would be so is obvious from the fact that the vertical pressures between the point of the needle and the record in an ordinary phonograph are of the order of 50,000 pounds per square inch. Obviously any such pressures would destroy a groove cut in soft "wax." These high pressures have been necessary in order that the groove might properly drive the needle point of the reproducer. Reduction of this pressure requires reduction of the impedance offered by the needle point to transverse vibration.



$$T_m = \frac{\pi N R^2}{24 \bar{V}_o}$$

T_m = MAX. PLAYING TIME IN MINUTES
N = GROOVES PER INCH
R = OUTSIDE GROOVE RADIUS-INCHES
$\bar{V}_o$ = MIN. LINEAR SPEED-FT. PER MINUTE
r = INSIDE GROOVE RADIUS-INCHES

Fig. 6—Relation between playing time and rate of rotation of disc for various values of $R(R = 2r)$.

The design of a suitable "wax" "playback" requires reduction of both the mass and the stiffness of the reproducing system to a minimum. In the past such "playbacks" have failed to reproduce the higher and lower frequencies with much satisfaction. The device shown in Fig. 7 represents a large advance toward ideal reproduction.

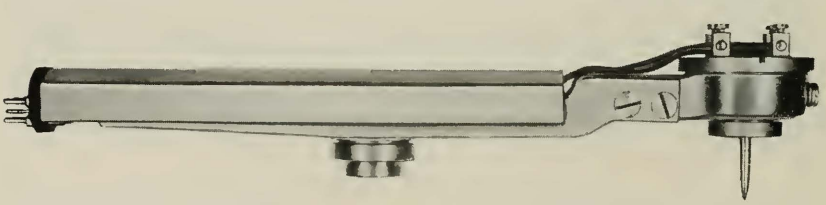


Fig. 7—Playback pickup for "wax" records.

The response of such a device when driven by a "wax" record recorded at constant velocity over the frequency range is shown in Fig. 8. This reproduction is not widely dissimilar from that obtained from finished records with the best electric pickups now commercially available and is sufficiently good to serve as a very valuable criterion in judging the quality of the record. The record may be played a number of times without great injury. The extent of the injury is indicated by the frequency characteristics obtained on successive

playings shown in Fig. 9. They show little change in the low frequency response and a loss of about 2 TU per playing at high frequencies. The practical value in studio work of being able to let an artist immediately hear and criticize the results of his own efforts can hardly be overestimated.

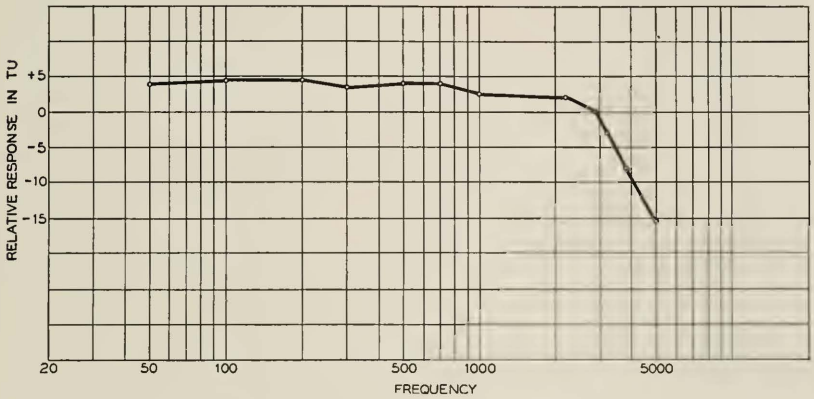


Fig. 8—Response of a "wax playback" driven by constant velocity wax record.

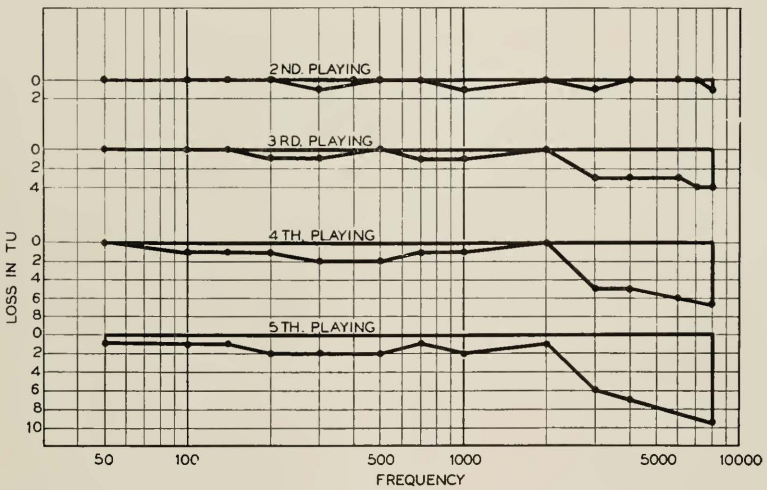


Fig. 9 Loss in response on successive playings of a wax playback driven by constant velocity wax record.

Copying (Processing) of the "Wax" Record

After a groove has been satisfactorily cut into the "wax" record, the usual procedure is to render the surface of the "wax" conducting by brushing into it an extremely find conducting powder. It is then electroplated. The technique in this step varies somewhat with the

different companies doing such work, although not in any fundamental manner. The negative electroplate thus made may be used to hot-press a molding compound such as shellac containing a finely ground filler. This first electroplate is called a "master." From it two test pressings are usually made. If satisfactory the matter is then electroplated with a positive, being first treated so that this positive plate may be easily removed. This positive is sometimes called an "original." From this in turn is plated a metal mold or "stamper." From these, duplicate "originals" may be plated and from them, duplicate "molds" or "stampers." These processes involve no measurable injury to the quality of the record and are comparatively simple and extremely safe in practice. By this practice of making a number of duplicates it is possible to safeguard the "master" and insure against any accident which might destroy a valuable record. From a single "stamper" it is not unusual to make a thousand finished pressings. The time required for these operations is such that test pressings are commonly obtained from the "wax" in 12 hours. Recent refinements in the art have reduced the time required so that finished records may, if necessary, be obtained in 3 hours after delivery of the "wax."

Hard Record or "Pressing"

Various materials have been used in making the hard record or "pressing." In some cases the material has been made homogeneous and in others the surface is of a different material from that used in the body of the record. Some have used a laminated structure. There has not, however, been much latitude allowed the experimenter concerned with materials for the hard record. The material has had to be quite hard and, in order to show a reasonable life, it has had to contain sufficient abrasive to grind the needle quickly to a good fit. At the beginning of the run of a new needle due to the small bearing surface, the pressures are very high. They rapidly decrease so that with an ordinary loud steel needle after one minute's wear in the ordinary phonograph, the bearing area is increased to such an extent that the pressure is only about 50,000 pounds per square inch. As the needle continues to wear to a larger bearing surface, the pressure obviously continues to decrease. These high pressures and necessary abrasive characteristics of the record have introduced irregularities which are responsible for most of the extraneous noise commonly known as "surface" or "needle scratch."

The "pressing" copies the "wax" record with a very high degree of accuracy so that if our attention be confined to frequency characteristics alone, the "pressing" shows almost complete perfection.

Moreover, it is cheap and durable, and reproduction of the sounds from this record calls for no fine adjustments or intricate apparatus as has been long evidenced by the broad use of the ordinary phonograph.

The major part of the extraneous or "surface" noise found with this method of reproduction comes from the material of the finished record. Recent progress has been made in reducing this noise. As a result of this, together with refinement in the plating processes, records used with Western Electric Company theater equipment during the last two years have shown a reduction of 3 to 6 TU in "surface" noise. This corresponds to eliminating 50 per cent to 75 per cent of that previously present. It is not necessary to reduce the level of "surface" noise to the zero point but merely to the threshold of audibility under the conditions of minimum auditorium noise which are of interest. This noise masks the surface. Moreover, it is not the absolute amplitude of the imperfection giving rise to "surface" noise but the relative magnitude in comparison with the useful sound amplitudes which counts. Thus, an effective reduction in "surface" could be made if we were willing to use larger records or if we were willing to reduce the playing time of the present records by increasing the spacing of the grooves and the amplitude at which the grooves are cut. Any large reduction in "surface" noise made by a reduction in the irregularities in the record material would open the door to increasing the playing time of a record of given size. There is no known absolute or fundamental reason why further improvements in record materials may not be expected to reduce further the amount of "surface" noise. Moreover, large advances in pickup design open distinctly new possibilities as to reductions in "surface."

It has sometimes been thought that in order to reproduce high frequencies properly, the linear record speed would have to be increased or the size of the needle point reduced. At present the diameter of the bearing portion of a representative needle is about .003 in. whereas, as mentioned before, the half wave-length for a 5,000 cycle wave is .0014 in. The factor determining whether a needle will follow the undulation of the groove is not any consideration of the relative diameter of the needle point and the undulation of the groove but rather the radius of curvature of the needle and the bend of the groove. As indicated before, the amplitude at 5,000 cycles would be only about .0001 in. if sounds of that frequency were as intense as those of lower frequencies (.002 in. at 250 cycles). As a matter of fact, sounds of 5,000 cycles or more in speech or music are characterized by lower

intensity than those of lower frequency. If, however, we assume an amplitude of .0001 in. at 5,000 cycles and assume a linear record speed of 70 ft. per minute, then the minimum radius of curvature of the undulation of the groove is .00193 in.[4] With the foregoing assumption, the radius of curvature of the undulation of the groove becomes equal to that of the needle point at about 7,000 cycles. Taking into account the lower intensities of sounds encountered at these high frequencies, it is obvious that present commercial needle points are quite capable of following the high frequency undulations of the groove up to frequencies of at least 10,000 cycles. The limitations of high frequency reproduction commonly found in the past are associated with limitations in the design of the pickup or reproducer and relate either to inability of the record groove to drive the needle point, with resultant chatter, or inability of the pickup structure to transmit high frequency motions from the needle point to the armature.

Electric Pickup

Large advances have been made within the last two or three years in designing electric reproducing structures. The mechanical im-

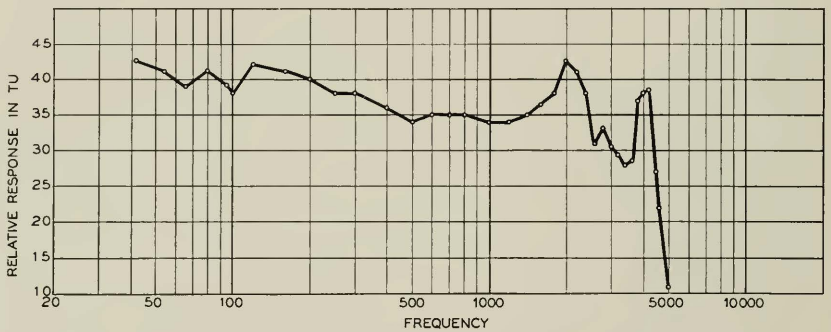


Fig. 10—Response of a 2-a pickup driven by constant velocity pressings.

pedance at the needle point has been reduced so that the needle point truthfully follows the undulations in the groove without necessitating excessive and somewhat destructive bearing pressures. At the

[4] The minimum radius of curvature is computed by the formula

$$R_c = \frac{V^2}{100\pi^2 A f^2},$$

where:

R_c = minimum radius of curvature in inches.

V = linear speed in feet per minute.

A = amplitude of vibration in inches.

f = frequency in c.p.s.

same time the transmitting structure has been so designed that a very broad range of frequencies is properly transmitted from the needle point to the armature. Moreover, proper mechanical loads have been provided so that the motions after transmission are absorbed and hence not reflected back. This is another way of stating the fact

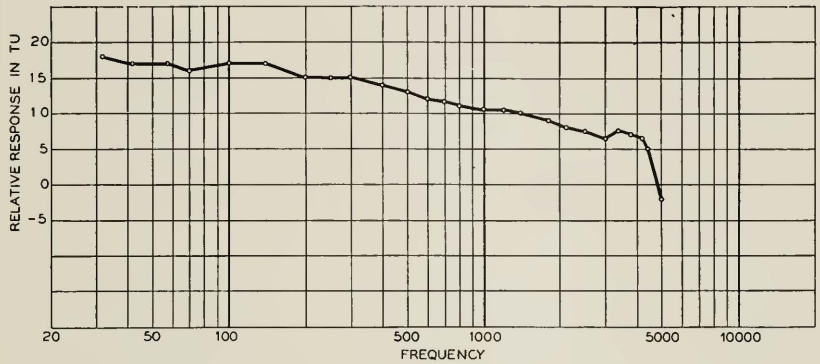


Fig. 11—Response of a 4-a pickup driven by constant velocity pressings.

that resonance as ordinarily considered has been eliminated from these structures.

The curves shown in Figs. 10 to 12 illustrate the steps which have been taken. The pickup shown in Fig. 11 is free from the resonances shown in that of Fig. 10. The resonances present in the earlier

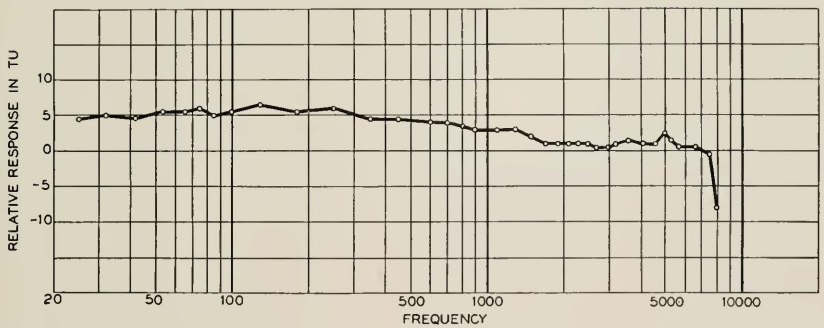


Fig. 12—Response of experimental pickup driven by constant velocity pressings.

pickup involved high needle point impedances in the region of these resonances. These high impedances involved large driving forces destructive both to needle and record. Certain records were injured after only a few playings with this reproducer. The later reproducer is characterized not only by considerably reduced *average* needle point impedance, but, as shown by the curves, the resonance is

practically eliminated and hence there is an even greater reduction from the maximum impedance which occurred in the earlier reproducer at resonance. Both needles and records have a relatively long life with the later type pickup, which has been in commercial use for some months. As is seen, the higher frequencies are reproduced in considerably better fashion. A third curve is given in Fig. 12. This was obtained with a more recent experimental model in which a further large reduction in the needle point impedance had been effected and in which, in addition, a very much more rigid, though a lighter structure served to connect the needle point with the armature. This model shows further reduction in wear and tear on the record and very greatly improved reproduction at the high frequency end of the scale.

The application of the processes of sound recording on "wax" to the synchronized film has involved meeting a number of conditions not previously encountered in the phonograph field. One of the most important of these relates to editing, cutting and rearranging of the picture. Various methods have long been used to copy or "dub" a disc record. The prime requirement is that there be no sacrifice in quality. To attain this end records have sometimes been copied at very low speed. This method appears unnecessarily laborious and slow and the results obtained are not altogether satisfactory in the light of possibilities presented by pickups and recorders of the characteristics shown above. Rearrangement of material on records is entirely practicable, portions may be deleted or new portions added either as a whole or the new sounds added to those already on a record—in fact any changes of this type may be made which can be made in the picture.

The detailed technique of "dubbing" appears to offer no serious technical difficulties. The refinement reached and the extent of its future use may be expected to be governed by the demand in the synchronized motion picture field.

Sound Recording with the Light Valve [1]

By DONALD MacKENZIE

Synopsis: The light valve developed by Bell Telephone Laboratories is an electromagnetic shutter consisting of a loop of duralumin tape formed into a slit at right angles to a magnetic field. Sound currents from the microphone and amplifier flow in this loop causing it to open and close in accordance with the current variations.

The slit is focussed by a lens on the sound negative film. An incandescent ribbon filament is focussed on the light valve, and the light passed by the undisturbed slit appears on the film as a line at right angles to the direction of the film travel. As the valve aperture is modulated by sound currents, the film receives a varying exposure and a sound record of the variable density type is obtained.

For talking pictures such a sound film is made on a separate recording machine synchronized with the camera and is printed alongside the picture on the finished positive. The prints are displaced so that the sound is advanced over the corresponding picture. This is in order that the sound may be projected at a point of continuous film motion below the picture gate.

THE sound records I am about to describe are of the variable density type, and the method of making them is that developed by Bell Telephone Laboratories.

It is not difficult to specify the requirements of this type of sound film. So far as possible the exposure of the negative must be kept within the straight line portion of the Hurter and Driffield curve for the emulsion chosen, and the print must be timed with the same restriction. The development of the negative and of the print must result in a positive where the transmission of each element of length is proportional to the exposure of the corresponding element of the negative. The light modulator must be supplied with undistorted power from the recording microphone and amplifier. When the positive is projected, the striations of the sound track must be enabled to modulate the illumination of a photo-sensitive cell to retranslate the photographic effect into electrical current which shall be a fair copy of the microphone current generated by the original sound. From this point on the problem is the familiar one of sound reenforcement, the film and cell having taken the places of the sound source and microphone.

Fig. 1 shows a photograph of the light valve, invented in 1922 by E. C. Wente of the Bell Telephone Laboratories. Essentially, it consists of a loop of duralumin tape suspended in a plane at right

[1] Presented before Society of Motion Picture Engineers at Lake Placid, New York, September 25, 1928.

angles to a magnetic field. The tape, 6 mils wide and 0.3 mil thick, is secured to windlasses A and A' and stretched tight by the spring held pulley B. At points C and C' insulated pincers confine the central portions of the tape between windlasses and pulley to form a slit 2 mils wide. Supporting this loop and adjusting devices is a slab of metal with central elevation D, which constitutes the armature of an electromagnet. The central portions of the loop are supported on insulating bridges to lie 3 mils above the face of D; here the sides of the loop are centered over a tapered slot, 8 mils wide by 256 mils long in this plane, opening to 204 mils by 256 mils at the outside face of the armature. Viewed against the light, the valve appears as a slit 2 mils by 256 mils.

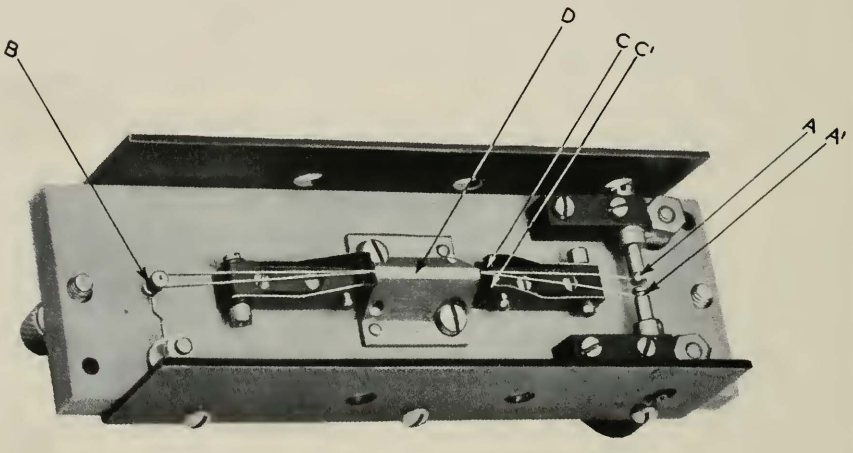


Fig. 1—The light valve.

The electromagnet core has a similar elevation opposing D across an air gap of 8 mils which closes to 7 mils when the magnet is energized from a 12 volt battery. A tapered slot in the magnet core begins 8 mils wide by 256 mils long and opens with the same taper as the slot in the armature. When the assembly of magnet and armature is complete, the valve constitutes a slit 2 mils by 256 mils, its sides lying in a plane at right angles to the lines of force and approximately centered in the air gap. The windlasses A and A', one of which is grounded, are connected to the output terminals of the recording amplifier. If the magnet is energized and the amplifier supplies a sine wave current from an oscillator, the duralumin loop opens and closes in accordance with the current alternations.

When one side of the wave opens the valve to 4 mils and the other side closes it completely, full modulation of the aperture is accom-

plished. The natural frequency of the valve is set by adjusting the tension applied by the pulley B; for reasons which involve many considerations the valve is tuned to 7,000 cycles per second. Under these circumstances about 10 milliwatts of A.C. power are required for full modulation at a frequency remote from resonance; about one one-hundredth of this power at the resonant frequency. The impedance of the valve with protecting fuse is about 12 ohms.

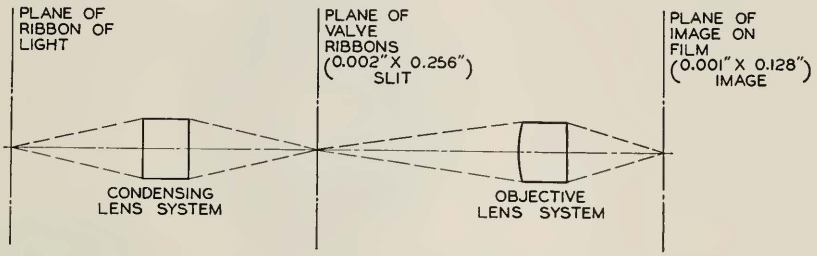


Fig. 2—Diagram of the optical system for studio recording.

If this appliance is interposed between a light source and a photographic film we have a camera shutter of unconventional design. Fig. 2 shows a diagram of the optical system for studio recording. At the left is a light source, a ribbon filament 18 ampere projection lamp, which is focussed on the plane of the valve. The light passed by the valve is then focussed with a 2 to 1 reduction on the photographic film at the right. A simple achromat is used to form the image of the filament at the valve plane, but a more complicated lens, designed to exacting specifications by Bausch and Lomb, is required for focussing the valve on the film. The undisturbed valve opening appears on the film as a line 1 mil by 128 mils, its length at right angles to the direction of film travel. The width of this line varies with the sound currents supplied to the valve, so that the film receives a varying exposure: light of fixed specific intensity through a varying slit.

Fig. 3 shows a studio recording machine with the door of the exposure chamber open. In this machine the film travels at 90 feet per minute, and the sound track is made at the edge away from the observer. The line of light, the image of the valve, overruns the perforations by 6 mils, extending toward the center of the film 122 mils inside the perforation line. The right-hand sprocket serves to draw film from the feed magazine above and to feed it to the take-up magazine below; this sprocket is driven from the motor shaft through a worm and worm-wheel. The left-hand sprocket engages 20 perforations and is driven through a mechanical filter from a worm and worm-wheel

similar to that driving the feed sprocket. The mechanical filter enforces uniform angular velocity of the left-hand sprocket which carries the film past the line of exposure: the focussed image of the valve; balancing of the flywheel which forms part of this mechanical filter holds the angular velocity constant to one-tenth of one per cent, despite the imperfections of the driving gears.

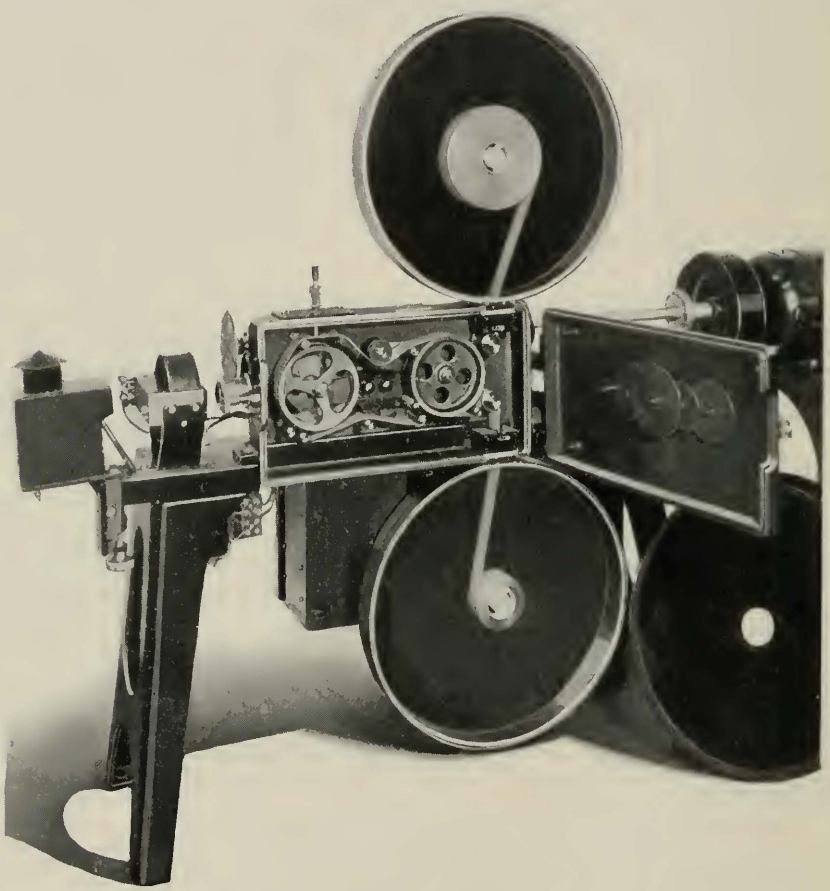


Fig. 3—Studio recording machine.

So far we have provided a means for driving the film and a means for modulating the light thereon, but we have not chosen the average illumination about which the modulation is to take place. The maximum exposure corresponds to the maximum opening of the valve and is therefore double the average.

Choose now the contrast to which the negative sound record is to

be developed and draw the Hurter and Driffield curve for this contrast for the emulsion chosen for the negative sound record. The maximum exposure should correspond to the beginning of over-exposure, the average should be half this. The Hurter and Driffield curve will give the density of the over-exposure point for the chosen contrast and the density for half this exposure. Let the machine be run to expose film to light through the unmodulated valve for several values of the lamp current. Develop the film and measure the densities due to the various values of lamp current. Select, by interpolation if necessary, the lamp current which corresponds to half over-exposure. With this current in the lamp the machine is ready to make a sound record, since the focussing of the valve has already been done and manufacturing specifications insure that the line of illumination shall lie, within 3 minutes of arc, at right angles to the direction of film travel.

Consider at this point the procedure in the recording studio. Adding sound to the picture introduces no complication of technique other than to require sufficient rehearsing to make sure of satisfactory pick-up of the sound: microphone placement must be established and amplifiers adjusted to feed the light valve currents which just drive it to the edge of overload in the fortissimo passages of music or the loudest utterances of speakers.

In Fig. 3 the photograph shows a photoelectric cell mounted inside the left-hand sprocket, which carries the film past the line of exposure. Fresh film transmits some 4 per cent of the light falling on it, and modulation of this light during the record is appreciated by the cell inside the sprocket. This cell is connected to a preliminary amplifier mounted below the exposure chamber, and with suitable further amplification the operator may hear from the loud speaker the record as it is actually being shot on the film. Full modulation of the valve implies complete closing of the slit by one side of the wave of current; this modulation should not be exceeded or photographic overload will abound.

One or more cameras and one or more sound recording machines are driven by motors electrically synchronized from a common distributor. Speed control and synchronization of these motors are described in Mr. Stoller's paper. At the beginning of the day's work a check is made of the operation of the driving motors, and the tuning-and-spacing of the valves is verified.

Fig. 4 is a schematic diagram of the studio equipment for sound recording. Provision is made for combining if desired the contributions of several microphones on the set. This combination is

under the control of the mixer operator in the monitoring room, viewing the set through a double window in the studio wall. The mixer controls also the gain of the amplifiers for the recording machines.

The diagram shows relays which permit the mixer to connect the horn circuit either directly to the recording amplifier or to one or the other of the monitoring photoelectric cells in the film recorders. The direct connection is used in preparing the sound pick-up in the studio: the program is rehearsed until satisfactory arrangement of microphones and of amplifier gain is effected. The electrical charac-

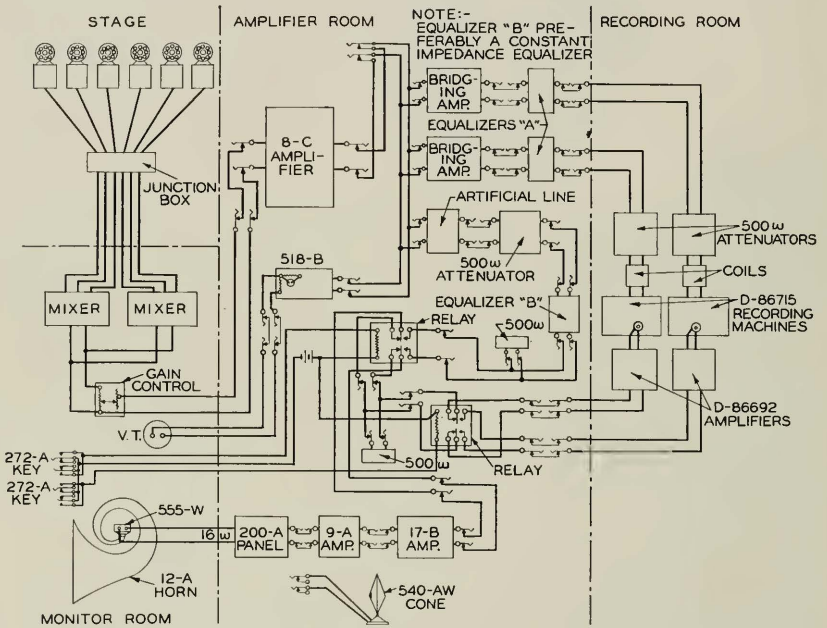


Fig. 4—Schematic diagram of the studio equipment for sound recording.

teristic of this direct monitoring circuit is so designed that the sound quality heard in the horns shall be the same as the quality to be expected in the reproduction of the positive print in the theater. Acoustic treatment of the walls of the monitoring room secures the reverberation characteristic of the theater, and the monitoring level is so adjusted that the mixer operator hears the same loudness that he would wish to hear from the theater horns. It is capitally important that the operator judge his pick-up on the basis of sound closely identical in loudness and quality with that to be heard later in theater reproduction.

After the pick-up has been established on the direct monitoring

circuit, the output of the recording amplifier is applied to the light valves and the monitoring horns are connected to the photo cell amplifiers on the recording machines. With no film in the machine and at a convenient lamp current a complete rehearsal is made to verify the operation of the valves at the proper level. Film is then loaded, cameras and sound recorders are interlocked and starting marks made on all films by punches or light flashes.

A light signal from the recording room warns the studio, which after lighting up signals back its readiness to start. The machine operator starts the cameras and sound recorders, brings up the lamp current to the proper value, and when the machines are up to speed signals the studio to start. During the recording, the mixer operator monitors the record through the light valves, thereby assuring himself that no record is lost.

In the choice of emulsion for the sound negative, the usual designation of speed may be disregarded, because it is desired to make the exposure of the unmodulated track many times the under-exposure of the emulsion used. The advantages of positive emulsion for the sound negative have come to be generally recognized; positive has been used by Bell Telephone Laboratories since 1924. The scale of Eastman positive film is about 20 to 1; we adjust the recording lamp current to give an illumination on the film for the unmodulated track of 10 times the under-exposure. After one lamp has been calibrated as described before it may be replaced when necessary by another in which the wattage in the ribbon filament is the same; the light emission is very closely correlated with the wattage. Where the unmodulated or average exposure is ten times the under-exposure minimum, 90 per cent modulation of the light can be permitted without running into under-exposure on the faint side of the wave. For sound currents reaching 100 per cent modulation of the light, 90 per cent of the wave is free from distortion; if the average light were halved, still 80 per cent would be free from distortion. There is therefore considerable latitude in the average exposure, and the negative is satisfactory if the transmission of the unmodulated track lies between fairly wide limits.

The choice of the negative sound gamma is determined by the practice of the laboratory in regard to picture development. It is usual to see on the screen pictures whose overall gamma considerably exceeds unity. On the sound track the overall gamma should equal unity, and the development of sound negatives should be uniform, though that of picture negatives is left to the judgment of the finisher.

Theoretically, it should be immaterial what combination of reciprocal values is chosen for the negative and positive sound gammas. Prac-

tically, we have to recognize the existence of ground noise in all records and take precaution to minimize it. No matter how excellently we reproduce the fortissimo passages, our record is unsatisfactory unless the ground noise is low enough for a wide volume range, that is, a wide range in level between fortissimo and pianissimo. Whether our negative sound record is made on negative or positive emulsion, there is always the danger that in reproduction we shall encounter variations in transmission from point to point due to local variations in the celluloid base, to local action of the developing agent, or to a developer excessively granular in action. The photoelectric cell is able to recognize variations of 1/10 of 1 per cent, whereas the eye ignores contrasts under 2 per cent. These local variations in transmission, continued to the positive print, constitute the ground noise.

The remedy is, in part, to choose a developer as little granular in its effect as possible. In part, to insist on machine development of the sound film with thoroughly agitated developer. Further, to carry the sound development to a high gamma; this obviates to a large extent flow marks of the developer, and goes a long way to escape local variations in the base by developing the negative striations to be conspicuous in comparison.

In 1924 we concluded that the optimum choice was positive emulsion developed to unit gamma for both sound negative and sound print. This is feasible for sound records separate from pictures, but a compromise must be made for the combination of sound and picture in a single positive print. Here the positive development required for a satisfactory picture is always to a gamma far above unity.

It is customary to develop picture negatives by inspection, having in mind the uniform positive development to be undergone by the prints from these negatives. The gamma of these positives need never exceed 1.8; the sound negative then should be developed to 0.55. In order not to disturb the practice of the film laboratory, we ask that the positive development be standardized and its gamma ascertained, the reciprocal of this gamma then arranged for in the standardized negative development. A negative gamma above 0.5, together with the precautions of careful handling, permits the realization of an adequate volume range.

It is beyond the scope of this paper to discuss the details of manipulation and of choice of developer, but I wish to acknowledge the cooperation of Mr. J. W. Coffman in the solution of such problems. The problem is the reduction of ground noise, and its seriousness is not to be diminished by choosing a different recording method.

In printing the sound negative, a uniform density for the print of the unmodulated track is desired. The volume of reproduced sound

for a given reproducing light source, varies directly with the average transmission and the per cent modulation of this average. This average density should be on the straight line portion of the positive Hurter and Driffield curve, far enough to keep the denser negative portions from reaching the under-exposure region. For Eastman positive film a suitable transmission of the unmodulated portion of the sound print is 35 per cent, referred to air, for the usual values of positive gamma: 1.4 to 1.8. At this average transmission only the peaks of the recorded sound will encroach on the region of under-exposure. For the reciprocally developed negative track the region of under-exposure will have been reached by occasional peaks on the other side of the wave, and such photographic distortion as exists will be balanced between positive and negative.

Here we appropriately consider the photographic distortion as it occurs in variable density records. If the entire negative exposure has been confined to the under-exposure region of the emulsion chosen, a huskiness will result in the reproduction which can not be corrected by any known technique. But if the unmodulated negative transmission, for a gamma of 0.55, is about 16 per cent referred to air, 90 per cent of the wave will be clear of under-exposure, and experience shows that the ear detects no distortion. In telephonic terms, everything at a level 1 TU below full modulation will be free from distortion, and the peaks will be substantially perfect. The same may be said of the positive printed to an average transmission of 35 per cent, provided the overall gamma approximates unity.

It has been calculated that if the overall gamma departs from unity by 0.2 in either direction, a harmonic of 5 per cent amplitude of the fundamental will be introduced. Experimentation has shown that a 5 per cent harmonic is the least detectible. We state then the tolerance on the overall gamma for the sound track as 0.8 to 1.2. Variation of corresponding amount in the contrast of a picture print is intolerable; therefore greater latitude in contrast is permissible in the sound record than could be tolerated in the accompanying picture.

In printing these sound negatives in combination with pictures for projection in the theater, it is customary at the present time to print one negative, masking the space needed for the other, then run the positive again through the printer with the other negative, masking now the space already printed. In printing the picture negative, light changes are made as usual; for the sound negative the light is regulated to result in 35 per cent transmission of the unmodulated track after positive development. Provision of suitable masks in the camera has been made to show in the finder and expose on the film only the portion which will be available for picture projection.

In the theater projector, the sound gate is located 14.5 inches below the picture gate, in order to project the sound record at a point where the film is in continuous motion. Therefore in the printing it is arranged to print the sound negative displaced along the length of the positive enough to bring the sound 14.5 inches ahead

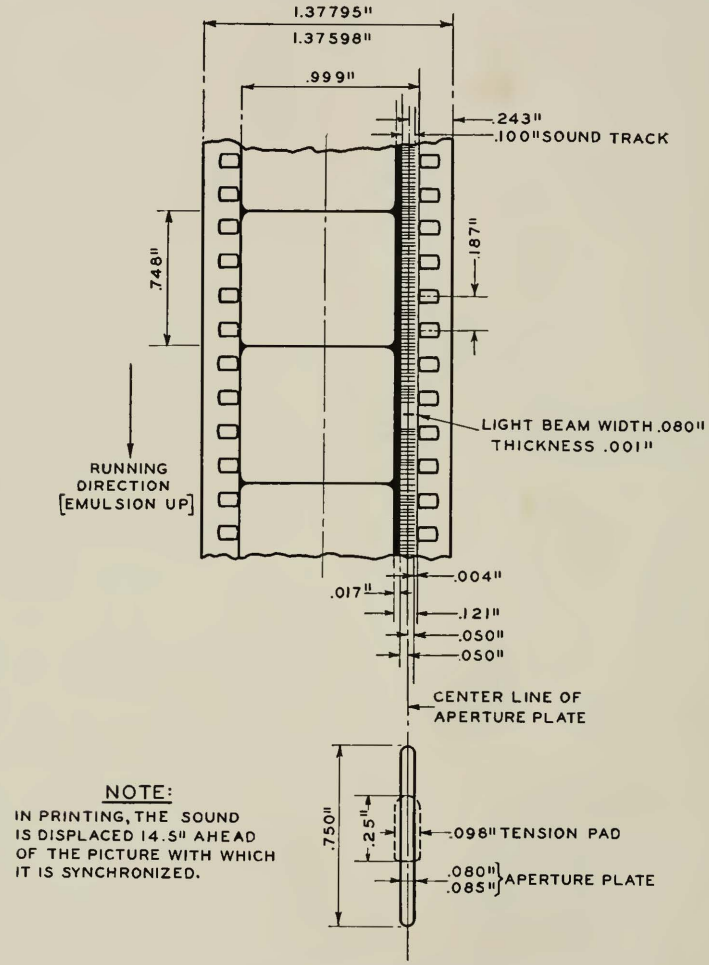


Fig. 5—Picture and sound track dimensions of synchronized sound film for standard 35 mm. positive stock.

of the corresponding frame. The printer apertures are chosen to give a dark no man's land 17 mils wide between picture and sound track; the latter at the outside is separated 4 mils from the inner perforation edge.

Fig. 5 exhibits the present practice for the finished positive. It will be seen that the sound track covers 100 mils clear, and is illuminated

in the projector by a line of light 80 mils long, 1 mil wide, centered on the striations. This gives a margin of 10 mils at each end of the reproducing line, an allowance for lateral shifting of the film on the sprocket teeth.

In conclusion, let me estimate the quality of the sound record to be expected. Assume that the recording lamp current has been set to within 5 per cent of the theoretical optimum, the overall gamma held between 0.8 and 1.2, and the final average positive transmission is between 32 per cent and 38 per cent. Then the distortion of wave form due to photographic handling is so small that the ear can not distinguish the record from a theoretically perfect one. The frequency-amplitude characteristic of the reproduced sound remains to be stated.

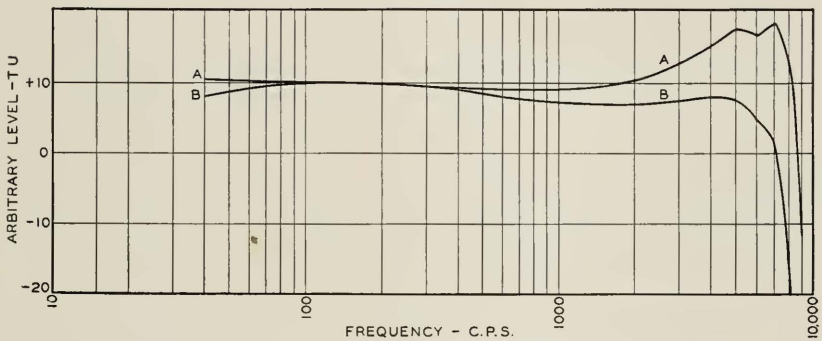


Fig. 6—Characteristic curves of sound recording.

Due to the fact that the element of illumination, both in recording and in reproducing, is 1 mil wide instead of infinitely narrow, the final print will not reproduce the higher frequencies as efficiently as the lower. For example, at the standard speed of 90 feet per minute, the line of illumination covers on the film an entire cycle length of the frequency of 18,000 cycles. This frequency is therefore extinguished completely. The drooping characteristic resulting from this effect, called the film transfer loss, may be largely offset by judicious choice of electrical characteristics and by taking advantage of the mechanical tuning of the light valve.

In Fig. 6 is shown in curve *A* the light modulation by the valve in recording for constant sound pressure of various frequencies at the transmitter; in curve *B* the overall characteristic of the reproduction in terms of electrical power delivered to the loud speaker for constant sound pressure at the transmitter in the studio. Experience shows that curve *B* is close enough to flat; the success of the record, as of the picture, depends on the director.

Synchronization and Speed Control of Synchronized Sound Pictures [1]

By H. M. STOLLER

SYNOPSIS: The reproduction of the synchronized sound picture of today presents no serious problem of synchronization, for this factor has been practically eliminated by the perfection of electrical means for reproducing sound with equipment which may be coupled mechanically to the picture projector.

The important problem of the present day, in connection with the reproduction of synchronized sound pictures, is the provision of suitable means for maintaining a constant speed of the sound reproducing mechanism in order that the pitch of the sound being reproduced may not suffer any sudden change which would be sensed by a good musical ear. Control circuits using vacuum tubes with a frequency bridge as a speed standard with provision for manual variable speed control are described and explained for use with both A.C. and D.C. motors. Remote synchronization permitting the recording of pictures and sound simultaneously on equipment located some distance apart is obtained by a modification of the Michalke electric gear system.

WHEN Thomas A. Edison gave a demonstration of his talking motion pictures nearly sixteen years ago one of his chief problems was proper synchronization between his acoustic phonograph and the motion picture projector. It was then necessary to locate the phonograph behind the screen in order to make the sound appear to come from the picture. A system of belts and pulleys running from one end of the theater to the other was used to secure synchronization with the projector in the booth.

The development of the electrical reproducer has made it possible to locate the turntable and reproducing mechanism in the projection booth permitting a direct mechanical coupling between it and the projector. The horns are located behind the screen and electrically connected by wires with the electrical reproducer.

Thus there is no problem of synchronization in reproducing except to set the needle on the disc at the proper point before starting. However, such mechanical coupling between the projector and sound recorder (either of the disc or film type) makes it necessary to provide very close speed regulation on the projector motor, since variations in speed produce proportional changes in the pitch of the sound.

This paper will describe the speed regulating system employed in reproducing and the synchronization system used in recording.

[1] Presented before Society of Motion Picture Engineers at Lake Placid, New York, September 24, 1928.

Speed Regulation Requirements

A good musical ear while having a sense of absolute pitch of only about 3 per cent is extremely sensitive to sudden changes in pitch. It has been found that a sudden change in pitch as small as one half of 1 per cent may be noticed if made abruptly. In order to properly take care of this requirement, therefore, the speed regulation or change in speed of the motor drive over normal variations in line voltage and load should be held within 2/10 of 1 per cent.

The absolute speed must also be held near these limits since at the end of a film it is necessary to switch from one projector to another with minimum change in the pitch of the sound reproduction.

Voltage, Frequency and Load Variations

A study of the voltage variations in power supply systems indicated a range from 100 to 125 volts. At a particular location the normal variation of voltage was found to be 5 per cent above or below the mean value with occasional momentary variations of as much as 10 per cent above and below mean value.

An investigation of variations in frequency of the supply voltage showed that in the large cities the frequency was held very accurately at 60 cycles. In New York City for example the frequency stays within one quarter of 1 cycle and does not change rapidly. However, in some small power systems the frequency varied as much as 5 cycles and in some cases was subject to rapid changes in frequency.

The load of the motor is due mainly to mechanical friction in the projector and take-up mechanism. This load was found to be on the average 1/10 of a horse power but subject to wide variations. In the case of a new machine with a stiff adjustment of the take-up mechanism, the load was found to be as high as one-fifth of a horse power.

Speed Control Circuit

A consideration of the variables just discussed imposes rather severe requirements of speed control, the two extremes being (1) the combination of low line voltage, low frequency and heavy load, (2) the combination of high line voltage, high frequency and light load. Ordinarily it might be possible to compromise and not provide for such an extremely unfavorable combination of requirements. However, it must be borne in mind that in the case of a musical program the failure of the speed regulating system for even as short a time as a fraction of a second would be a very serious matter causing the music to sound off pitch similar to a phonograph which has run down while in operation.

An examination of the standard commercial types of speed control indicated that there was nothing exactly suitable. The nearest approach to a suitable governor is the standard type of phonograph governor but this friction brake type of governor has serious objections if applied to a motor of considerable power output. In order, therefore, to have a control system which would be free from maintenance, it was necessary to develop a special form of control system for the purpose. Fig. 1 shows a photograph of the A.C. motor and its control cabinet.

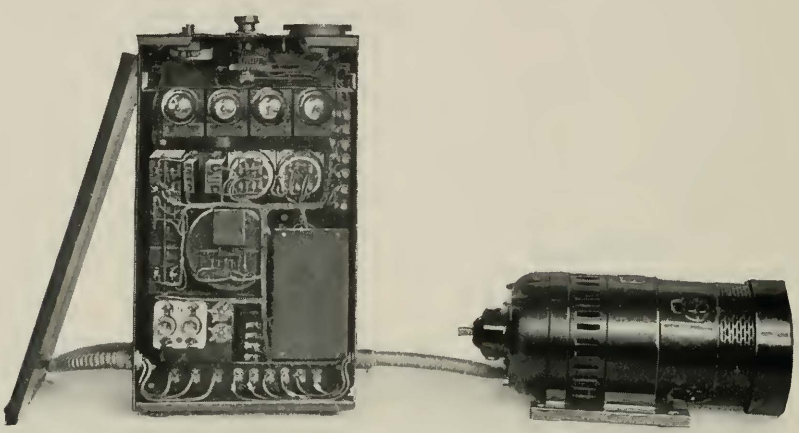


Fig. 1—A. C. Motor with control cabinet.

A.C. Motor Control Circuit

Fig. 2 shows the A.C. motor circuit which consists of a repulsion type of motor coupled to a small auxiliary alternator providing a frequency of 720 cycles which through a control circuit is made to operate a variable reactor across the armature terminals of the motor. If the speed of the motor is too high, the control circuit produces a maximum impedance in the reactor L_1 thereby reducing the armature current of the motor and causing it to slow down. While if the speed is too low the control circuit causes the reactor L_1 to have a minimum of impedance increasing the armature current and causing the motor to speed up.

This reactor L_1 is of the D.C. saturating type having two outer legs with A.C. windings and the middle leg with a D.C. winding. The A.C. flux circulates around the two outer legs. The D.C. flux flows from the middle leg and returns through the outer legs in parallel. When D.C. flux is sent through the middle leg it saturates

the outer legs thereby reducing their impedance to A.C. This type of reactance is old and was employed by Alexanderson as a magnetic modulator in his early radio work.

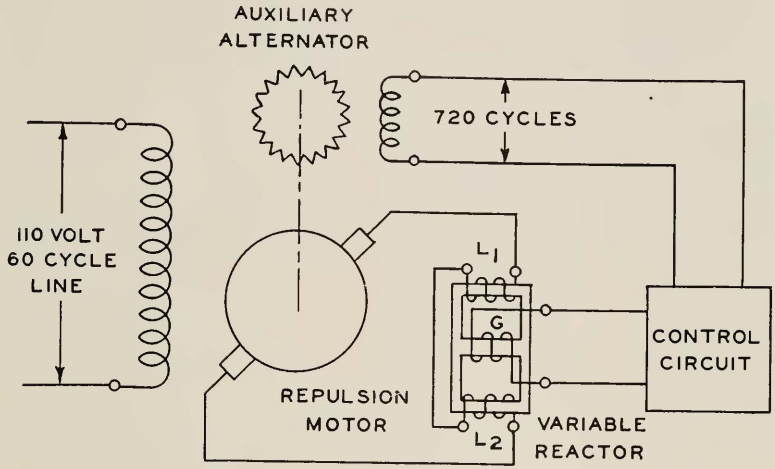


Fig. 2—A. C. Motor circuit diagram.

Fig. 3 shows one element of the speed control circuit. This consists of a bridge circuit having one variable arm and three fixed arms. The variable arm comprises a tuned circuit consisting of the in-

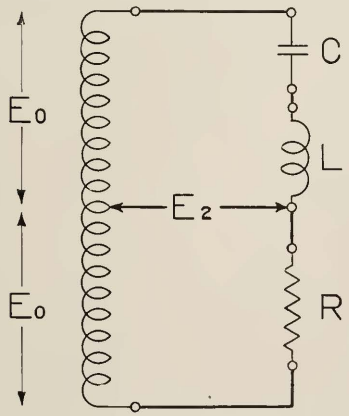


Fig. 3—Bridge circuit in speed control system.

ductance L and the capacity C which are designed to tune at exactly the frequency corresponding to the desired motor speed. When this circuit is in tune it has a resistive impedance which is balanced by the fixed arm R of the bridge. The other two fixed arms on the left

are windings of a transformer with a mid tap. If a voltage $2E_0$ having a frequency of 720 cycles (which is the frequency corresponding to the desired speed of 1,200 R.P.M.) is supplied to the bridge it will be apparent that the output voltage E_2 will be zero. If, however, the speed is low the tuned circuit will have a condensive reactance while if the speed is high it will have an inductive reactance. The output voltage E_2 will, therefore, change abruptly 180 electrical degrees from a speed below 1,200 to a speed above 1,200. This characteristic is shown in Fig. 4.

The use of the above described bridge circuit gives a very sharp characteristic due to the fact that the effective resistance component of the tuned circuit is balanced out by the adjacent resistance arm of

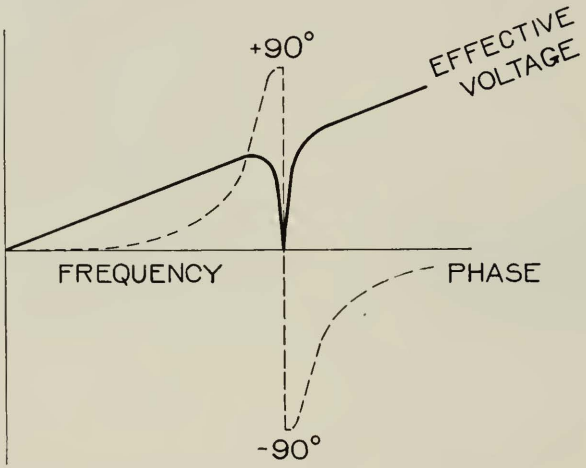


Fig. 4—Output voltage characteristic of bridge circuit.

the bridge. In this way an overall characteristic of the desired sharpness is secured using a comparatively small and inexpensive coil and condenser.

Fig. 5 shows the complete control circuit. The output from the bridge circuit is supplied to the grid of tube V_4 which is called the detector tube. The plate voltage of this tube comes from the 720-cycle generator through the step-up transformer T_4. The phase of this voltage, therefore, remains constant. The phase angle of the grid voltage, however, comes from the bridge output circuit through the step-up transformer T_3 and as previously explained suffers a sudden reversal of phase as the speed passes through 1,200 R.P.M. Fig. 6 shows the resulting current characteristic through tube V_4. This current flows through coupling resistance R_1 which drives the grids

of tubes V_1 and V_2 negative. This in turn reduces the plate current through tubes V_1 and V_2 and hence through the D.C. winding of the inductance L_1 controlling the armature current of the motor.

Tube V_3 is a rectifier tube supplying excitation to the field of the

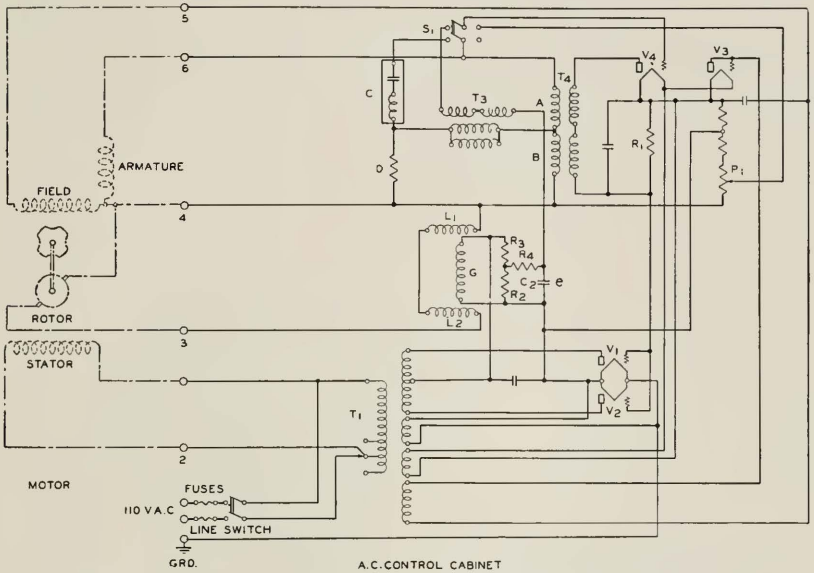


Fig. 5—A. C. control circuit diagram.

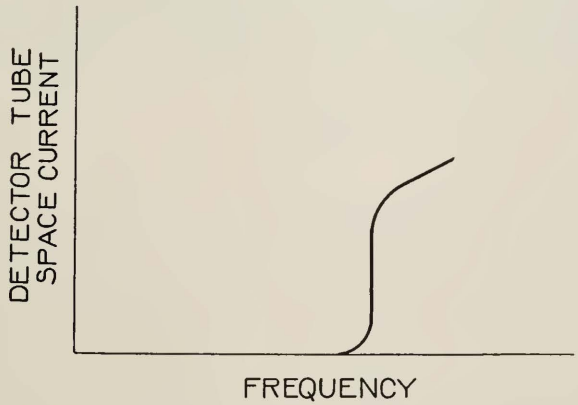


Fig. 6—Characteristic of space current through detector tube V_4.

720-cycle alternator and the negative "C" voltage along potentiometer P_1.

Fig. 7 shows the performance characteristics of the motor. It will be noted that the actual speed characteristic is practically flat. This

flat characteristic is secured by a compensating network consisting of the resistances R_2, R_3 and R_4 and the condenser C_2. This compensating network feeds back on the grid of tube V_4 a portion of the voltage drop across the D.C. winding of inductance L_1 thereby correcting for the "static fluctuation" of the control circuit. By a suitable adjustment of this compensating resistance the control circuit may be arranged to give flat regulation, under regulation or even over regulation if desired. Fig. 7 has been drawn with line voltage as the variable. A similar characteristic is also obtained with load as the variable instead of voltage.

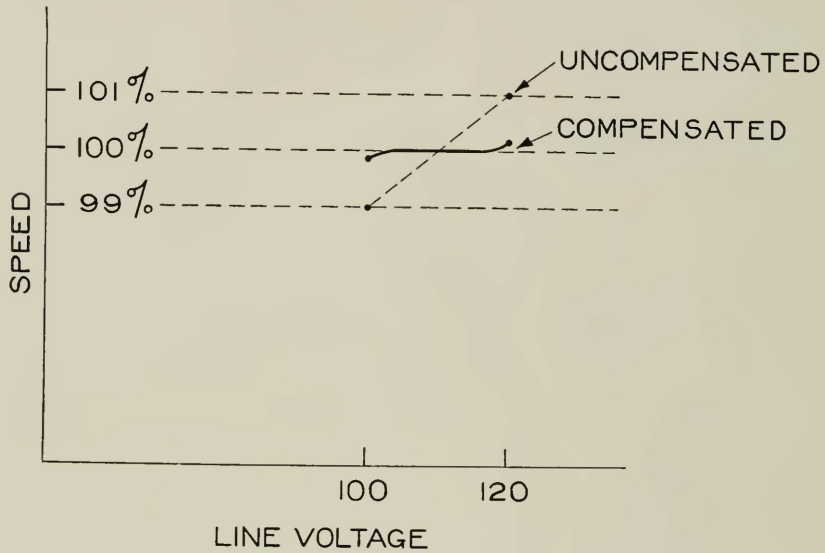


Fig. 7—Performance characteristics of motor.

An interesting point in connection with this compensation circuit is the necessity for avoiding hunting or surging of the speed. It is a well-known property of all forms of governors that if they are adjusted to too great a sensitivity the speed instead of remaining constant will fluctuate up and down about a mean value. The simplest method of preventing such speed fluctuations is to decrease the sensitivity of the governor allowing a bigger change in speed with load (or voltage) and then compensating for this change of speed or "static fluctuation" by means of a delayed action compensator. This phenomenon is well-known in the mechanical governor art and is described by Trinks in his book "Governors and the Governing of Prime Movers." The electrical equivalent of this mechanical system is obtained by intro-

ducing the condenser C_2 in series with the high resistance R_4. When a change in current through the regulating reactance L_1 occurs the corresponding change in voltage drop is not transmitted to the condenser C_2 immediately, but C_2 changes its voltage after a certain time lag (approximately 1 second), required to charge the condenser through the resistance R_4. The introduction of this time lag restores the precision of the circuit to the flat characteristic desired without introducing hunting.

Variable Speed Operation

By throwing the switch S_1 to the right the operator can disconnect the tuned circuit control and substitute a potentiometer P_1 as a

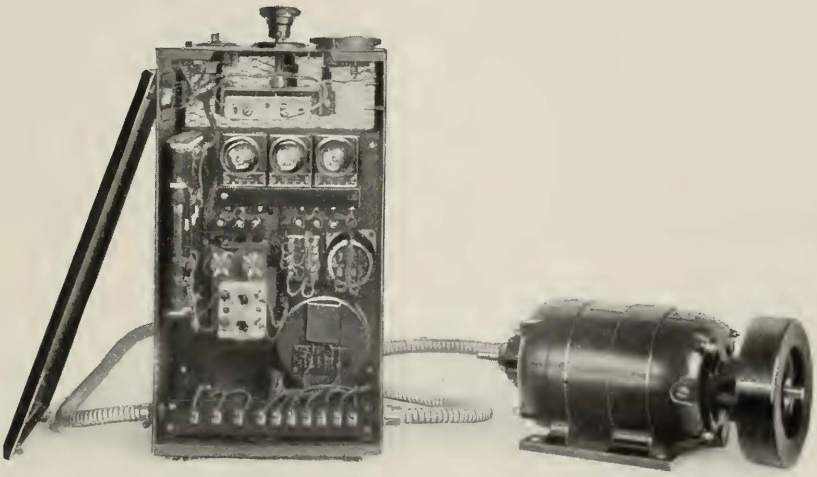


Fig. 8—D. C. motor with control cabinet.

source of grid voltage for tube V_4. By means of this potentiometer the operator can adjust the speed of the motor at any speed from 900 to 1,500 R.P.M. corresponding to 68 to 112 feet of film per minute. This feature is employed for ordinary motion picture work where it is unnecessary to synchronize the picture with the sound. The regulation of the circuit under these conditions is sufficiently good for ordinary motion pictures.

An interesting feature in this connection is that theaters in many cases have preferred to use the regulated speed position for ordinary motion pictures as well as synchronized pictures. The reason for this being that with the speed of the projector precisely controlled the orchestra leader is better able to keep his orchestra in step with the picture indicating apparently that closer speed regulation than is

at present provided would be desirable for ordinary motion pictures as well as synchronized pictures.

D.C. Control Circuit

A circuit very similar to the one just described is employed in the case of the D.C. motor. Fig. 8 shows a photograph of this motor and its control cabinet. The circuit is shown in Fig. 9. It differs from the A.C. circuit in that an auxiliary regulating field winding is employed on the motor instead of a variable reactor. The source of power for the plates of the vacuum tubes is obtained from the auxiliary 720-cycle generator instead of from a 60-cycle transformer as in the

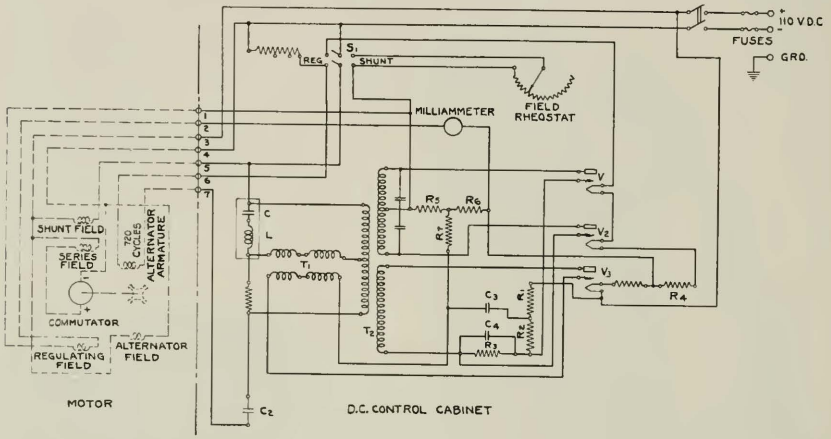


Fig. 9—D. C. control circuit diagram.

A.C. circuit. Since a strengthening of the field of the D.C. motor is required in order to reduce the speed it is necessary to reverse the phase relationship of the transformer T_1, so that the current in the detector tube decreases at speeds above 1,200 instead of increasing as in the case of the A.C. circuit.

The operation of the circuit is as follows: When the line switch is first thrown the motor acts as an ordinary D.C. shunt motor and accelerates. At low speeds the output from the 720-cycle generator is low and consequently there is no plate voltage supplied to the tubes and no current through the auxiliary field winding. The field is, therefore, weak and the motor speeds up. This condition is maintained until the equilibrium speed of 1,200 R.P.M. is approached. The phase angle of the voltage supplied to the grid of the tube V_3 is then in phase with the voltage supplied to the plate so that the grid of the tube goes positive at the same time that the plate is positive.

This causes a current to flow through the coupling resistances R_1 and R_2 which drives the grids of tubes V_1 and V_2 negative, thereby keeping down the current through these tubes and hence maintaining a weak motor field. The motor, therefore, continues to accelerate until a speed of 1,200 R.P.M. is reached. At this point as previously explained under the description of the bridge circuit, the phase of the output suddenly reverses whereupon the grid of the detector tube goes negative at the same time that the plate goes positive, thereby cutting off the current through the detector tube V_3 and reducing the negative C voltage on the grids of tubes V_1 and V_2. This increases the plate current through the regulating field thereby stiffening the field of the motor and checking its rise in speed. In practice the current through the detector tube is neither at one extreme nor the other but reaches an equilibrium at the speed of 1,200 R.P.M. A feedback network having the delay feature for prevention of hunting is included in the same manner as previously described for the A.C. circuit. The characteristic curves for the D.C. motor are similar to those shown in Fig. 7 for the A.C. motor.

For the operation of ordinary motion pictures the motor is changed to a simple shunt D.C. motor by the switch S_1 and the speed varied by means of the field rheostat.

Motor Drive of Recording System

It might appear that the simplest method of securing synchronization in recording work would also be mechanical connection between the recording machine and the camera. It has been found desirable, however, from a practical standpoint to have the camera movable with respect to the recording machine as the recorder has to be accurately lined up and adjusted and is not essentially a portable machine whereas the camera in ordinary motion picture work must be a portable piece of equipment. It has been necessary, therefore, to develop a motor drive equipment which will satisfactorily interlock the camera and the recording machine but leave the camera unit portable. It is essential that the interlock should hold not only during normal conditions but during acceleration and deceleration. In other words, the system must be the full equivalent of a mechanically geared system. The principle employed is old being disclosed in a patent issued to Michalke in 1901. In Fig. 10, A and B are two units which it is desired to interlock. Each unit has a three phase stator and a three phase rotor, the latter provided with slip rings. Magnetizing current for the system is supplied from an independent three phase, 60-cycle source. If the rotors of A and B are in exactly

the same positions with respect to the stators it is evident that the e.m.f.'s produced in them by transformer action will be identical as to voltage and phase. Consequently there will be no flow of current over the rotor leads and hence no torque developed. If, however, unit A is turned through a small angle then the phase of the e.m.f.'s produced in the rotor circuits will differ from that in B and a current will flow in the rotor circuits producing a torque which will tend to make unit B assume the same position as A. If A is rotated continuously B will follow it up to synchronous speed of the stator field at which point the torque will drop to zero since no e.m.f. is induced in the rotor of either machine.

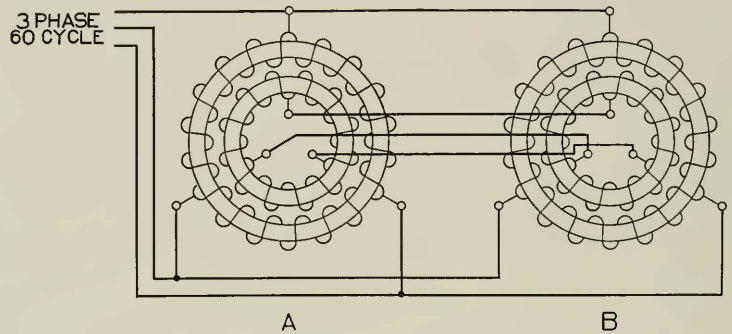


Fig. 10—Electrical driving gear.

Complete Recording Circuit

The portion of the circuit shown in Fig. 10, is merely the equivalent of a mechanical gear, neither unit tending to rotate as a motor by itself. In order to produce such rotation, therefore, a distributor set is added as shown in Fig. 11, the distributor acting, to use a mechanical analogy, as the driving gear of the system and each of the individual units of the system as driven gears. The distributor is itself driven by a D.C. motor provided with the speed control circuit previously described and shown in Fig. 9. The speed of the system is thus solely dependent on the D.C. driving motor and independent of the 60-cycle excitation frequency.

In practice the system is controlled by an operator at the distributor set and by means of switches any desired number of cameras, recording machines, or projectors may be employed. The projecting machines are used in case it is desired to make up a sound record to accompany an ordinary motion picture film which has previously been recorded without sound accompaniment. It has been found that the system operates very satisfactorily and requires very little maintenance.

When starting up for the first time it is necessary that the various units should line up properly as to phase otherwise there will be a local flow of current in the rotor circuits, which will cause the motors to operate as induction motors and run away. Under running conditions the system is very stable showing no tendency to hunt or surge between units, for the reason that being polyphase each phase as it becomes inactive (when the induced e.m.f. passes through zero) acts as a damping winding for the other two active phases.

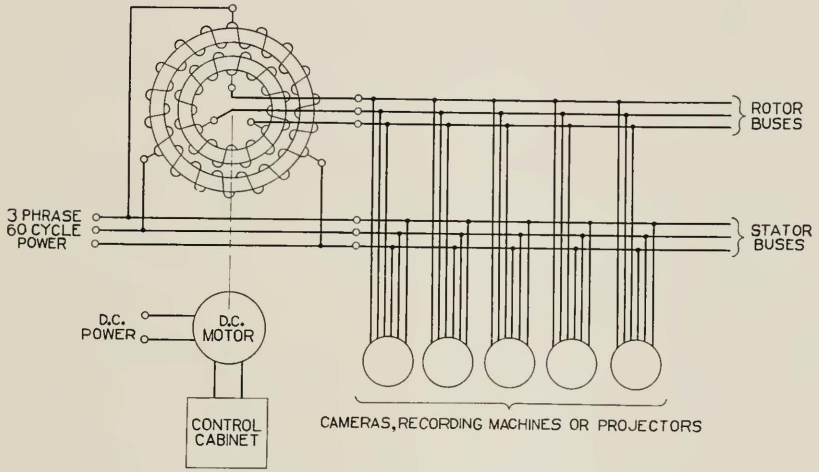


Fig. 11—Diagram of synchronizing system for recording.

If the load on a particular unit of this system is varied there will be a variation in the phase angle between this unit and other parts of the system in the same manner as in the case of a synchronous motor of the ordinary type. The magnitude of this phase angle, however, does not vary more than 30 electrical degrees or 15 mechanical degrees and is sufficiently small so that it produces an inappreciable effect on the synchronization.

Discussion

The above described motor equipment with its associated control circuits has been in practical use for over a year both in recording and reproducing work and there have been practically no troubles in service.

In the design of this equipment, first consideration has been given to its precision and reliability in operation and the provision of adequate margins to care for all variations in service conditions. As a result it has been possible to maintain a high standard of quality in music and speech reproduction.

A Sound Projector System for Use in Motion Picture Theaters[1]

By E. O. SCRIVEN

SYNOPSIS: The general problem involved in the design of a system suitable to be used to record and reproduce sounds such as are required for "talking" motion pictures is outlined. The general method of attack is indicated. There follows a description of the several pieces of apparatus which comprise the theatre equipment, including a discussion of some of their salient features and of the part each plays in the sound projector system.

IN order to reproduce in a theater the pictorial record of events accompanied by the sound associated with those events, it is, of course, necessary to add equipment to that installed to produce only the silent motion picture. It is the purpose of this paper to outline and discuss briefly the major items of such equipment as developed by Bell System engineers.

In the design of sound equipment a primary requisite is that there shall be freedom from distortion. Distortion may be of the sort which is independent of load and is evident in that the intensity of some portion or portions of the sound spectrum is increased or decreased in comparison with the rest; or there may be the distortion which is a function of the level at which the device is operated and is characterized by the reduction of a pure tone into fundamental and one or more harmonics. This latter condition is most often the consequence of operating a vacuum tube amplifier above its proper energy handling capacity.

It is the resonances of vibrating strings or reeds or air columns or vocal cords that give us the music we record but we are careful that a minimum of the resonances of the recording system itself shall go into the record, and that any resonances of the reproducing system shall not appear in the output of the sound projectors. Aside from the effects of overloading, the prevention of distortion is largely a matter of getting away from resonance phenomena since it is the characteristic of the resonant system to respond with disproportionate amplitude to stimuli in the region of its own natural period. The whole story of the passage from sound energy through the various recording and reproducing devices back to sound energy again is one of contest with this fundamental physical phenomenon.

[1] Presented before Society of Motion Picture Engineers at Lake Placid, New York, September, 1928.

There are in general two things one can do to avoid the harmful effects of resonance in vibration transmitting or transforming apparatus: (1) the period of resonance of each piece of equipment can be moved outside the range of frequencies one wishes to transmit, at the same time providing damping means to minimize free vibrations; (2) the distortion produced by resonance in one piece of apparatus can be compensated for or equalized by similar and opposite distortion

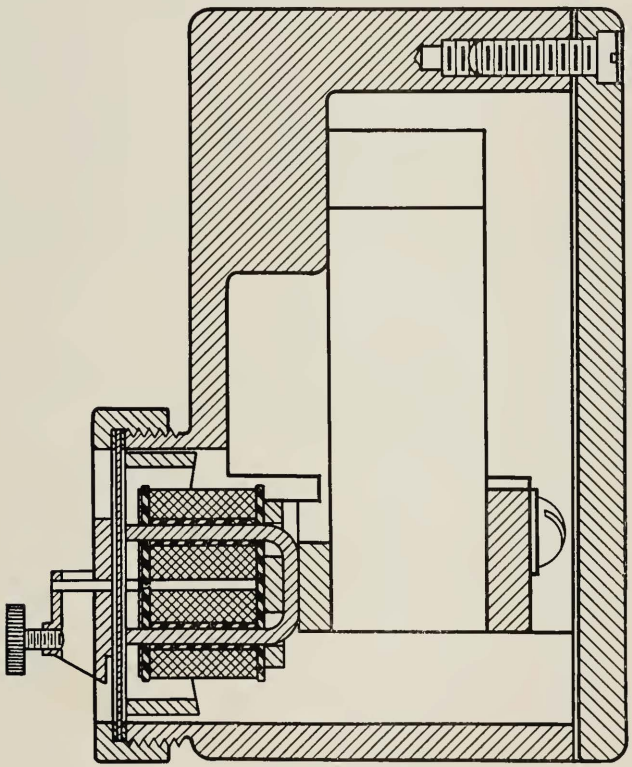


Fig. 1—Reproducer for disc record.

in some associated apparatus. The first is not always easy of practical accomplishment in any particular device and generally results in an instrument of very low inherent efficiency; the second usually involves loss of energy. In both cases increased amplification is required.

The sound record comes to the theater either as a wavy groove in a composition disc or as a striated track of varying density at one side of the picture film. It is the function of the apparatus being considered to derive from these records an electric current in which

all the variations in pitch and loudness are accurately represented, to suitably amplify this current, to effect its conversion into sounds approximating those from which the records were made, and to so direct those sounds as to reasonably create the illusion that sound and picture are cognate.

The disc records do not differ essentially from those used in the ordinary phonograph except that they are considerably larger and run at a much slower speed so that a single record will play throughout an entire reel. The reproducer used is in some ways similar to that

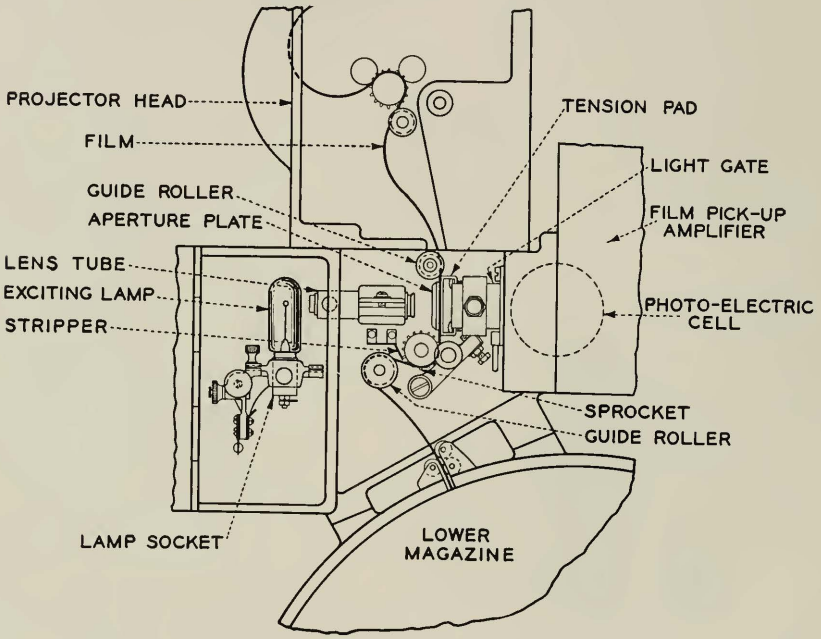


Fig. 2—Diagram of motion picture projector equipped for reproducing sound from film.

used on the acoustic phonograph, the needle holder being connected to a clamped diaphragm. This diaphragm is of highly tempered spring steel and to it there is fastened an armature made of a special high permeability alloy and so arranged that as the diaphragm vibrates the flux in the air-gap of a permanent magnet varies correspondingly, thereby inducing in appropriately placed coils currents which are the electric representation of the wavy groove which the needle travels. This reproducer is shown in Fig. 1. Although the energy delivered by this instrument is comparatively low it has a very uniform response over a wide frequency range. This result is largely brought about

by moving all resonances out of the working range and by filling the magnet chamber back of the diaphragm with a heavy damping oil. The film used with the disc record, called a synchronized film, differs from ordinary film only in that one frame at the beginning is specially marked to give the starting point.

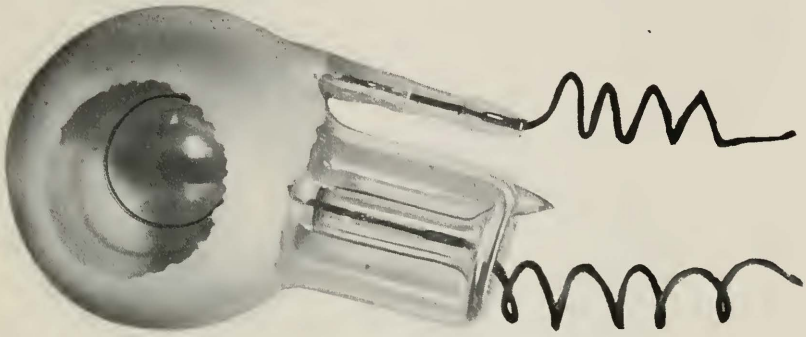


Fig. 3—Photoelectric cell.

The film sound record, as has been said, consists of a track of varying density running along one side of the picture. This sound track is 1/10″ wide. Differences or changes in intensity of sound are represented by differences in the density of the record, while pitch

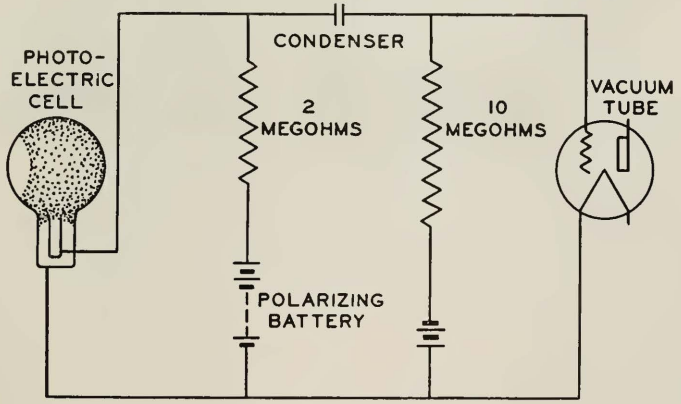


Fig. 4—Photoelectric cell circuit.

is represented by the number of changes from dark to light and back again in a given length of track. This sound record is converted into a corresponding electric current by arranging that a narrow high intensity beam of light shall pass through it and fall upon a photo-electric cell. The arrangement is shown in Fig. 2. The light from

the bright filament of the exciting lamp is focused as a very narrow line upon the film by passing through a system of lenses and an aperture plate. The lamp filament is focused upon a slit of dimensions .0015″ x 3/16″. The image of this slit is then brought to focus upon the film as a .001″ line whose length has been reduced in passing through the aperture plate to .080″. This reduction in length allows .010″ on either side for variations in position of the .100″ sound track. The position and focus of the lens tube are fixed, but the carriage of

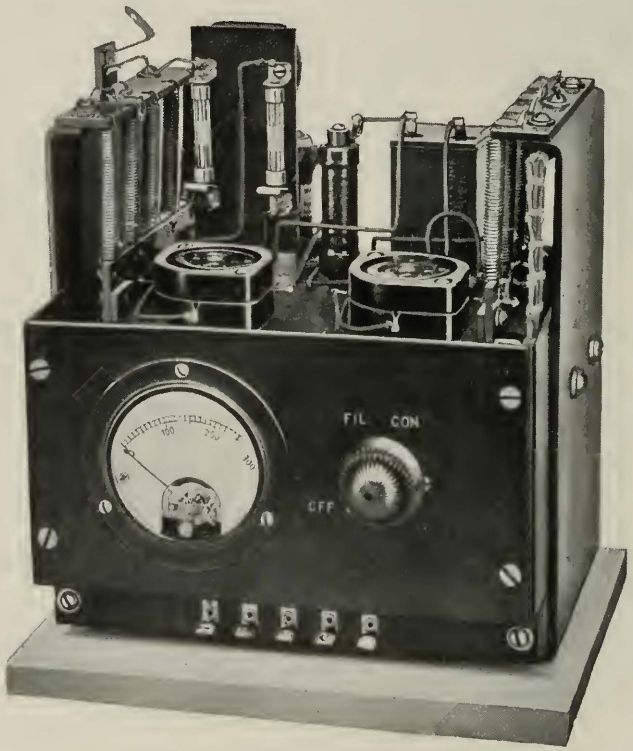


Fig. 5—Amplifier used at projector.

the exciting lamp is movable so that when replacing lamps the filament may be properly brought on focus.

A photoelectric cell of the type used is shown in Fig. 3. The characteristic of this device is that when it is polarized by a proper voltage and is used within proper limits the current through it is proportional to the incident light. The circuit is shown in Fig. 4. It is to be noted that the polarizing voltage is supplied to the photo cell through a very high resistance and there is, therefore, obtained across

this resistance a voltage which is proportional to the light falling upon the cell and accordingly bears a direct relation to the varying density of the sound track interposed between the exciting lamp and the cell.

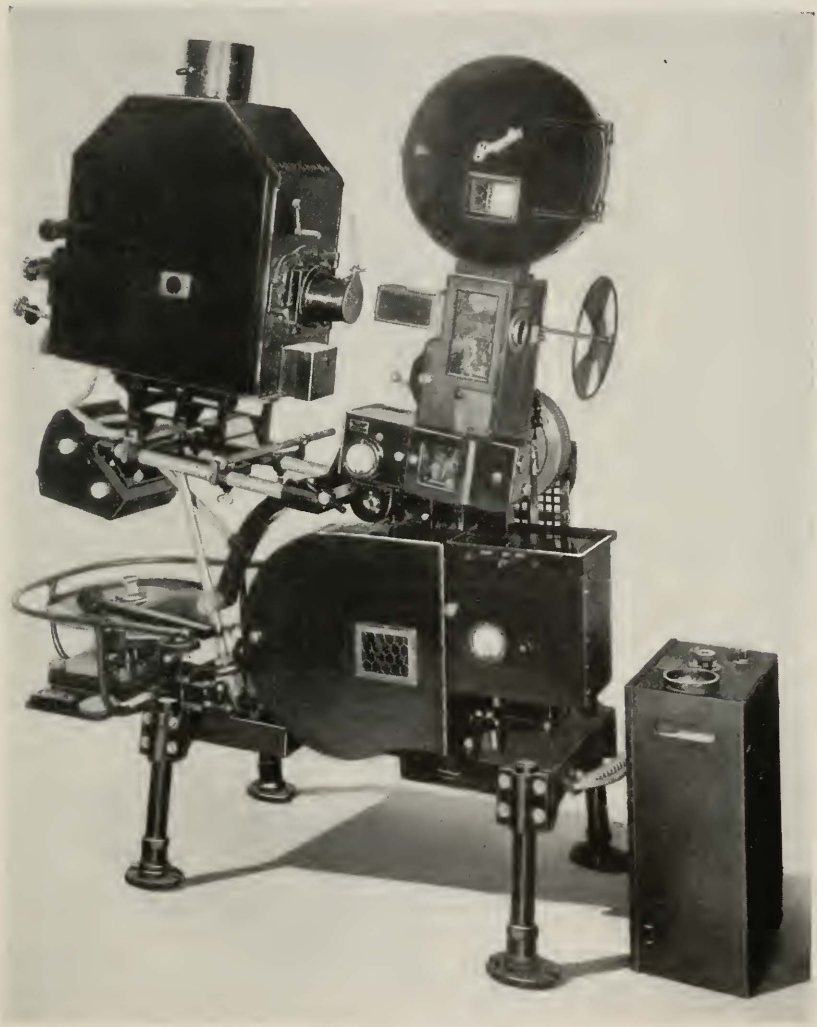


Fig. 6—Sound projector equipped with Simplex head.

The photo cell circuit is inherently one of high impedance. In such a circuit there are two matters which require attention. (1) Local interference—"static," to use the radio expression, is most readily picked up, and at this point where the energy level is low,

may be appreciable in comparison with the sound currents themselves; (2) also the shunting effect of capacity between the electrical conductors becomes noticeable, particularly at the higher frequencies.

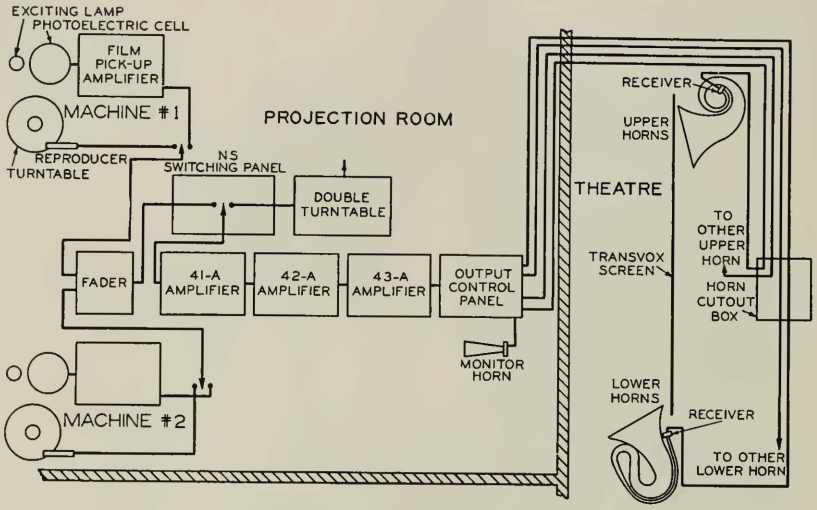


Fig. 7—General layout of equipment.

Hence a vacuum tube amplifier, which serves both to increase the energy and to make that energy available across a low impedance circuit, is closely associated with the cell upon the projector itself.

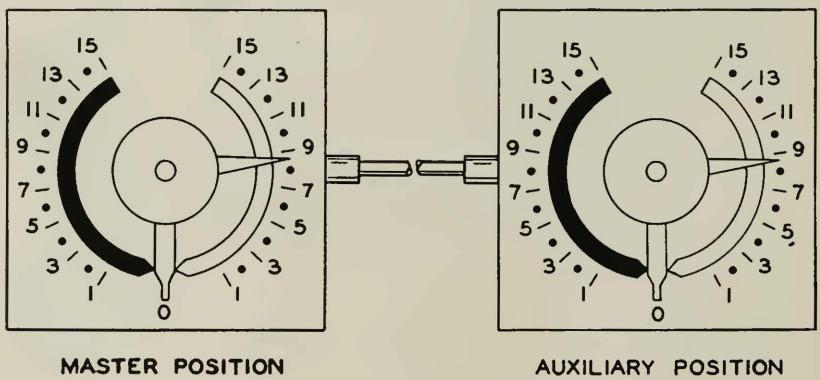


Fig. 8.

The cell and amplifier are enclosed in a heavy metal box or shield which is made fast to the frame of the projector and the projector itself is carefully grounded. This amplifier is shown in Fig. 5. It is designed to bring the level of the electric counterpart of the film sound

record up substantially to the same energy value as that obtained from the magnet coils of the disc reproducer. The filaments are heated from a 12-volt storage battery. Small dry batteries supply its plate

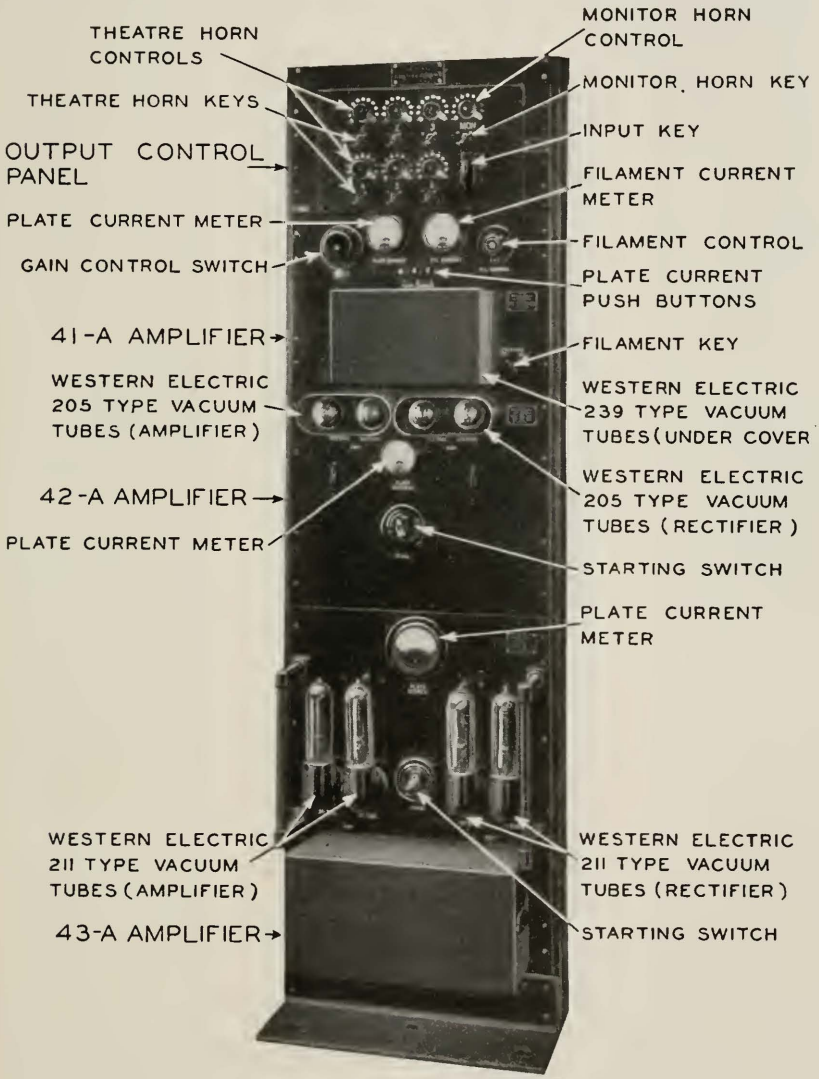


Fig. 9—Amplifier Panel.

current and also the polarizing potential for the photo cell. These batteries and the battery leads are shielded.

Vibration of a vacuum tube often produces sufficient motion of its elements with respect to each other to effect changes in the stream

of electrons which appear when sufficiently amplified as noise from a loud speaker. In spite of all precautions there is a certain amount of vibration of the projector when in operation and it has therefore been necessary to design a rather elaborate shock-proof mounting for the photo cell amplifier.

It is evident from the relative location of apparatus as shown in Fig. 2 that it is not feasible to print the film sound record directly beside the picture to which it applies. As a matter of fact, there is a spacing of $14\frac{1}{2}''$ between picture and corresponding sound record and a certain amount of slack is allowed between the sprocket which carries the picture with an intermittent motion before the picture projection lens and the sprocket which must carry the sound record

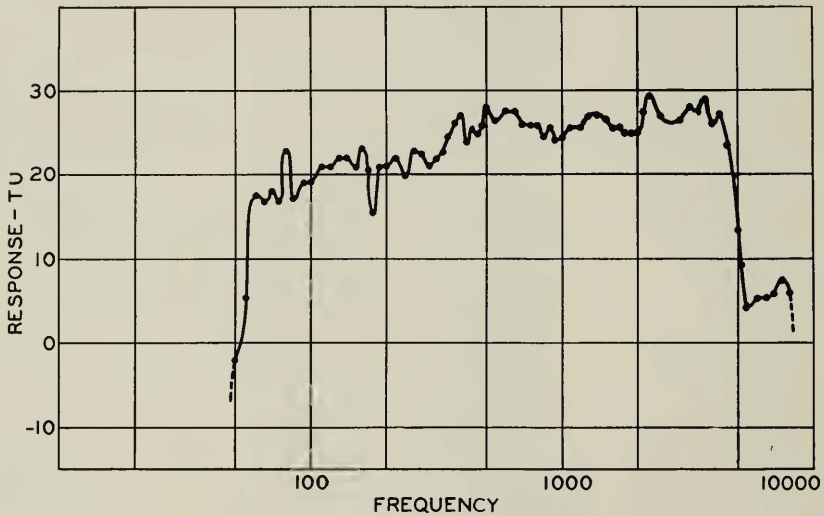


Fig. 10.

with a uniform motion in front of the photoelectric cell. In this connection it is noteworthy that special precautions are necessary in order to prevent vibrations and speed fluctuations due to either varying supply voltage or varying load from affecting the uniformity of rotation of this sound sprocket. This is taken care of by the very effective means of automatically controlling the speed of the driving motor and by means of a mechanical device interposed between the sound sprocket and the rest of the moving equipment of the projector which effectively opposes the transmission of any abrupt change of speed to this sprocket.

The control box which contains the apparatus for governing the speed of the driving motor is arranged to hold the record speed the

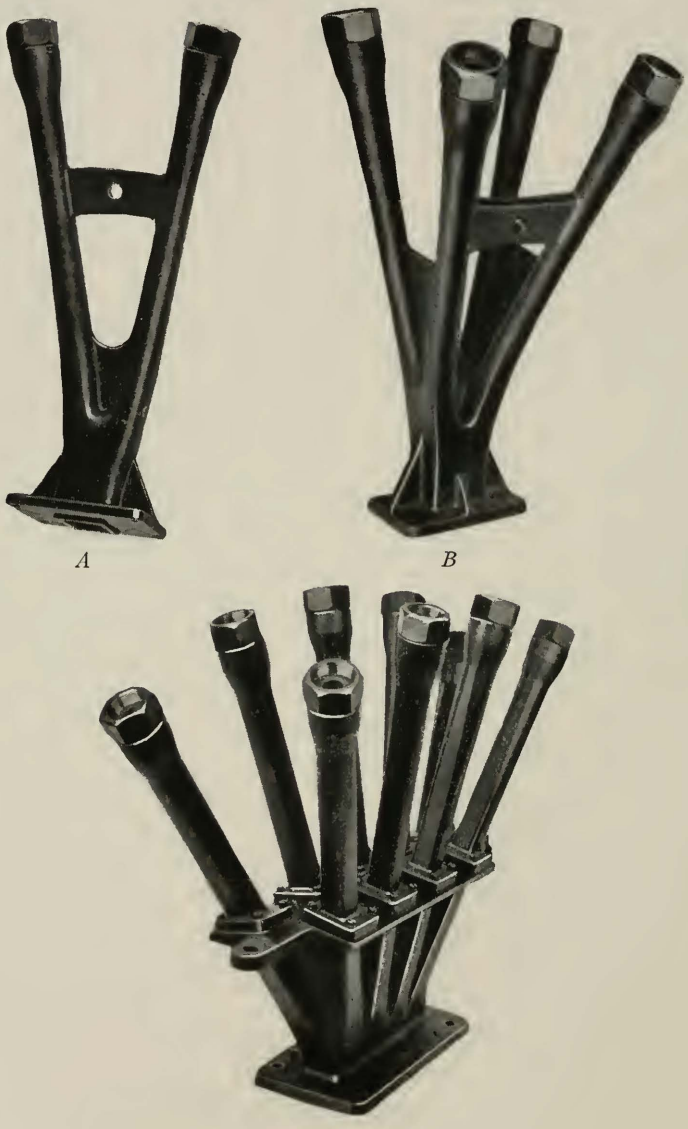
same as that at which the records are made, i.e. 90 feet per minute in the case of synchronized sound and picture productions. By throwing a switch the automatic feature may be cut out and the speed of the machine may then be manually controlled by the operator.

This completes the apparatus associated directly with the projector. The general arrangement of the latest type of projection machine, equipped with a Simplex head, is shown on Fig. 6. Incidentally this projector is also arranged to be fitted with the Powers or the Motiograph head. Fig. 7 shows a typical layout of a sound projector system as installed for use with talking motion pictures.

As in ordinary pictures, in order to run a continuous program, it is necessary to use two projectors alternately. As the picture from one machine is faded imperceptibly into that on the other so the sound record may be faded from one machine to the other without the audience being aware that a change has been made. At the end of each record or sound film the music overlaps the beginning of the next and a device called a fader is employed in making the transition. All that is necessary is to turn the fader knob when the incoming machine is started. This fader is in fact a double potentiometer. In the upper or normal operating range the change in volume in moving from one step to the next is hardly more than perceptible whereas in the lower range used only in fading the steps are large and the volume decreases to zero on one machine and builds up on the other very rapidly. By choosing the proper step in the upper range one can obtain any volume of sound desired within reasonable limits and thereby equalize the level obtained from different sound records. The fader is ordinarily installed with one or more auxiliary dials and handles interconnected so that it may be operated from any projector position. In connection with the fader there is provided a switch for changing from the film to the disc input system and also a key for switching a spare projector in place of either of the regular machines. Fig. 8 shows a fader with one auxiliary position.

Following the fader, we come to the main amplifier which raises the energy of the feeble electric currents to a level adequate to supply the loud speakers with sufficient volume to serve the particular theater. Fig. 9 shows a typical amplifier panel. This combination is capable of an energy amplification of about 100,000,000 times and is so designed that all frequencies in the range from 40 to 10,000 cycles are amplified practically equally. A potentiometer is provided on the amplifier but while its handle is readily accessible it is ordinarily not used after having once been set at the time of installation to give proper results in the particular theater. Necessary adjustments are

made on the fader. The amplifier shown consists of three units. The first consists of three low power tubes in tandem, resistance coupled, and requiring a 12-volt battery delivering 1/4 ampere to heat their filaments. The second consists of a single stage of two medium power tubes, connected in push-pull arrangement with filaments



A *B*

C

Fig. 11.

heated by low voltage alternating current. Two similar tubes in this unit operate as a full wave rectifier and supply rectified alternating current for the plate circuits of the amplifier tubes of both the first and second units. The third unit has a single stage of high power push-pull amplifier tubes and push-pull rectifier tubes and also operates entirely on alternating current.

These three types are capable of arrangement into combinations to meet the particular need. For small theaters only No. 1 and No. 2 are required. In the larger houses the high power unit No. 3 is added, while to meet exceptional conditions two or more of the high power amplifiers may be operated in parallel from the output of No. 2.

Following the amplifier there is an output control panel. This consists of an auto-transformer having a large number of taps, the taps being multipled to a number of dial switches, to which the sound projectors or loud speakers are connected. By means of this panel, it is possible to match the impedance of the amplifier output to the desired number of horns in order to obtain the most efficient use of the power available and also to adjust the relative volume of the individual horns.

The ordinary theater installation employs four horns, two mounted at the line of the stage and pointed upward toward the balconies and two mounted at the upper edge or above the screen and pointed downward. This combination has been found to give good distribution throughout the house.

The loud speaker unit used with the horns in theater equipments is essentially that recently described by Messrs. Wente and Thuras.[2] As brought out in this article, this unit shows extremely high efficiency; about 30 per cent of the electrical power supplied is radiated in the form of sound. This is important since the higher the loud speaker efficiency, the smaller the power capacity of the amplifier needed in the system. The frequency-response characteristic of a typical receiver and horn is given in Fig. 10. An individual horn may be equipped with two, four or nine loud speakers by using the throats shown in Fig. 11. The power capacity for continued safe operation of the horn with one, four and nine throats is approximately 5, 20 and 45 watts, respectively (electrical input). The number of horns used is dependent upon the particular installation and is related to the directive characteristic of the horn. If it is necessary to disperse the sound over a large angle, more horns are needed than when it is desired to concentrate over a comparatively small angle. This directive characteristic of the horn is important in talking motion pictures as it is responsible for the illusion of the sound coming directly

from the mouth of the horn; that is, from the screen. If the horn is replaced by a loud speaker of otherwise identical characteristics but which radiates its sound over a very wide angle, there is a tendency for the sound to appear to come from a point some distance back of the screen, thus tending to destroy the illusion.

The power supply equipment has been fairly completely covered in discussing various parts of the system. Under ordinary conditions the requisite power is obtained from the electric mains in the theater except for the 12-volt battery required for some of the vacuum tube filaments and for the electromagnets in the loud speakers and the dry cells used with the photo cell and photo cell amplifier. Where 110-volt D.C. only is available there is a projector-driving equipment which operates on this voltage, but a D.C. motor driving a 60-cycle generator is required for supplying the amplifiers. Where 110-volt A.C. is available, it is only necessary to connect the projector motor and the amplifiers to this supply.

[2] *Bell System Technical Journal*, January, 1928—"A High Efficiency Receiver of Large Power Capacity for Horn-type Loud Speakers," by E. C. Wente and A. L. Thuras.

Abstracts of Bell System Technical Papers Not Appearing in this Journal

The Communication System of the Conowingo Development.[1] W. B. Beals and E. B. Tuttle. This paper describes the communication system which has been installed to serve the power plant at Conowingo, Maryland, and its associated transmission line.

The important features to be considered in designing a telephone system for a power plant are pointed out. The types of telephone switchboard and telephone instruments chosen in this case to meet the special requirements of the generating station, together with the layout and cabling arrangement, are outlined.

The paper also discusses the possible ways of providing for the needs of the load dispatcher and the plan adopted at Conowingo; the facilities provided the patrolmen for calling from points along the transmission line; the connection from the private branch exchange to the general telephone system; and the special electrical protection installed on the long lines leaving the power house.

Reflection and Refraction of Electrons by a Crystal of Nickel.[2] C. J. Davisson and L. H. Germer. This is a report of further observations on the regular reflection of electrons from the surface of a nickel crystal; an earlier report was published in the same journal.[3] In the present report data are given of the selectivity of reflection for angles of incidence from 10 to 50 degrees, and for electrons of wave-lengths 0.6 to 1.5 Å. The previously found result is confirmed that to explain the occurrence of the intensity maxima of the reflected beam it is necessary to assume that electron waves are refracted on passing into the crystal. The data are used for calculating indices of refraction for nickel for electrons of various speeds or wave-lengths, and a dispersion curve is constructed. This curve displays a feature suggestive of the optical phenomenon of anomalous dispersion.

Optical Experiments with Electrons.[4] L. H. Germer. A semi-popular account of a series of experiments performed by C. J. Davisson and the author upon the scattering of electrons by single crystals of

[1] *Journal of the A. I. E. E.*, October 1928, pp. 737–741.

[2] *Proceedings of the National Academy of Sciences*, August 1928, pp. 619–627.

[3] *Proceedings of the National Academy of Sciences*, April 1928, pp. 317–322.

[4] *Journal of Chemical Education*, Part I, Sept. 1928, pp. 1041–1055. Part II, Oct. 1928, pp. 1255–1271.

nickel. These experiments establish the fact that under certain conditions moving electrons behave like trains of waves. In the interaction of these waves with a single crystal the optical phenomena of diffraction, reflection and refraction have been observed. Scientific accounts of these experiments are contained in the following papers: *Nature*, 119, 558 (1927); *Phys. Rev.*, 30, 705 (1927); *Proc. Nat. Acad. Sci.*, 14, 317 (1928); *Proc. Nat. Acad. Sci.*, 14, 619 (1928). Although the present paper is of a popular nature it aims to be quite comprehensive. It attempts to represent the status of this series of experiments in August 1928.

Rubber Compression Testing Machine.[5] C. L. Hippensteel. This paper gives a brief account of a new compression test developed at the Bell Telephone Laboratories for more reliably judging the ability of rubber insulation on metallic conductors to withstand certain service conditions to which it is subjected. A recording compression testing machine, which has been built for applying the test, and typical results are illustrated. Other possible test uses for the machine are suggested.

New Languages from Old—How Secrecy is Gained by the Inversion of Speech Sounds.[6] C. R. Keith. The inversion of speech sounds may be accomplished with the aid of methods used in radio broadcasting and in carrier telephony. Among the possible applications, it is illustrative of methods used to achieve secrecy in electrical communications.

The character of speech sounds is determined by the frequencies and amplitudes of the component waves into which the sound may be resolved. The process of inversion consists effectively in altering the frequency distribution of these components so that low tones appear as high tones, while high tones appear as low tones. To the untrained observer, inverted speech is unintelligible, although the characteristic cadence is preserved. Inversion of the frequency scale is produced by modulating speech with a carrier wave which lies just above the highest speech frequency which is to be transmitted, and selecting the lower sideband. For practical reasons connected with undesired distortion, it is more desirable to break up the modulating process into two distinct steps. The original speech sounds may then be regained by repeating the process which led to its inversion.

[5] *India Rubber World*, Sept. 1928, pp. 55–56.

[6] *Scientific American*, October 1928, pp. 310–311.

Joint Pole Use with Power Companies.[7] D. E. LOWELL. The relations between the telephone company and the other wire using companies, especially the power companies operating in the same area, are discussed in this paper. It recognizes the responsibility of the telephone company as well as that of the power company for good operating conditions in areas where both types of line are involved and also points out the necessity of close cooperation between Connecting and Bell Telephone Companies. The considerations involved in the joint use of poles by telephone and power companies are given with particular mention of the general joint use agreement. The importance of mutual advance notice of plans is developed. The reports of the Joint General Committee of the N. E. L. A. and Bell System form the background of the talk and are recommended to those who have not already read them.

Adsorption of Gases by Graphitic Carbon. II—X-ray Investigation of the Adsorbents.[8] H. H. LOWRY and R. M. BOZORTH. This paper is supplementary to one by Lowry and Morgan appearing in the *Journal of Physical Chemistry* in 1925 [9] and gives direct evidence that the adsorbents studied were graphitic carbon. The X-ray data show that carbon prepared by the explosion of graphitic acid is graphitic in structure and that the individual particles are flakes averaging approximately 50 atom diameters in breadth and 10 atom layers in thickness. The significance of this finding is discussed in relation to current views of the nature of active carbon adsorbents.

Recent Toll Cable Construction and its Problems.[10] H. S. PERCIVAL. One of the outstanding developments in the Bell System has been the rapid extension of toll cables. This has required the development of new methods and apparatus. Material is carried into rough right of way and installed through the use of tractors, with equipped trucks and various types of automotive equipment. The development of permalloy now allows the complete loading of a full-sized cable in two pots where six were required before. Crossings over rivers are made in submarine cable or by long span construction with catenary suspension. Cables are tested before completion for sheath damage, defective splices, etc., which might cause service failures, by means of dry gas under pressure.

[7] *Telephony*, September 8, 1928, pp. 22–24.
[8] *Journal of Physical Chemistry*, October 1928, pp. 1524–1527.
[9] *Journal of Physical Chemistry*, Vol. 29 (1925), p. 1105.
[10] *Telephone Engineer*, September 1928, pp. 31–33.

Quality Control by Sampling.[11] W. L. ROBERTSON. A discussion of the application of the mathematical theory of sampling to commercial shop inspection. Also gives tables illustrating numerically the results obtained from the various sampling plans in use.

Problems in Power Line Carrier Telephony and Recent Developments to Meet Them.[12] J. D. SARROS and W. V. WOLFE. Power transmission lines as commonly encountered present relatively complex networks having irregular and unstable attenuation-frequency characteristics within the 50–150 K.C. band employed for power line carrier telephony. The high frequency noise may be very high.

A single side band carrier suppressed system operating on a single frequency duplex basis has been developed to overcome these transmission difficulties.

A comparison of this system with other types shows its superiority.

The initial installation of this equipment was made on the 220 K.V. lines of the Pacific Gas and Electric Company.

The Planning of Telephone Exchange Plants.[13] W. B. STEPHENSON. This paper discusses procedures followed in planning future extensions to telephone exchange plants to care for increased demand for telephone service. An outline is given of the methods employed in forecasting future demand for telephone service and in determining the most efficient design of the plant to meet the service requirements. The uses made of engineering comparisons in solving the economic phases of various kinds of telephone engineering problems are discussed, with particular reference to location and size or extent of major items of plant as well as the time when they should be ready to give service. Emphasis is placed upon the importance of those factors less readily evaluated, such as service factors, practicability from a construction and operating standpoint, flexibility, etc.

The Effect of the Acoustics of an Auditorium on the Interpretation of Speech.[14] E. C. WENTE. Studies of speech sounds in the Bell Telephone Laboratories have shown that 60 per cent of the acoustic energy in speech lies below 500 c.p.s., although the intelligibility of individual speech sounds is reduced by only 2 per cent if all the

[11] *Factory and Industrial Management*, pp. 503–505, Sept. 1928; pp. 724–726, Oct. 1928.

[12] *Journal of the A. I. E. E.*, October 1928, pp. 727–731 (abridgment).

[13] *Journal of the A. I. E. E.*, July 1928, pp. 500–503 (abridgment).

[14] *The American Architect*, August 20, 1928, pp. 259–261.

energy below this frequency is completely suppressed. These results indicate that the sound absorption coefficient of materials placed in an auditorium for reducing the reverberation time should be high for tones of low frequency and low for those of high frequency. Most porous materials commonly used for this purpose have absorption characteristics quite the reverse. Rooms that have been treated with a rather large amount of such materials are therefore often unsatisfactory for speaking purposes, although the adjustment for reverberation time may have been carried out according to accepted standards.

Contributors to this Issue

W. H. Martin, A.B., Johns Hopkins University, 1909; S.B., Massachusetts Institute of Technology, 1911. American Telephone and Telegraph Company, Engineering Department, 1911–19; Department of Development and Research, 1919–. Mr. Martin's work has related particularly to transmission of telephone sets and local exchange circuits, transmission quality and loading.

H. S. Osborne, B.S., Massachusetts Institute of Technology, 1908; Austin Research Fellow in Engineering, 1908–10; Eng.D., 1910; American Telephone and Telegraph Company, Engineering Department, 1910–19; Department of Development and Research, 1919–20; Department of Operation and Engineering, 1920–. Mr. Osborne is Transmission Engineer and as such is responsible for assisting the Associated Companies in connection with telephone and telegraph transmission and protection matters.

G. W. Elmen, B.S., University of Nebraska, 1902; M.A., 1904; Research Laboratories of the General Electric Company, 1904–06; Engineering Department of the Western Electric Company, 1906–25; Bell Telephone Laboratories, 1925–. Mr. Elmen's principal line of work has been magnetic investigations. He is the inventor of the permalloy alloys. For this he was awarded the John Scott Medal in 1926 and the Elliott Cresson Medal in 1928.

H. O. Siegmund, B.S., University of Illinois, 1917; E.E., University of Illinois 1926; Instructor, U. S. Army School of Military Aeronautics, 1917–1918; Assistant Professor of Electrical Engineering, Drexel Institute, 1918–1919; Engineering Department, Western Electric Company, 1919–1925; Bell Telephone Laboratories, 1925–. Mr. Siegmund has been engaged in apparatus development work and has made contributions relating to telephone power plant apparatus and circuits, to provide more quiet transmission.

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, 1917–25; Bell Telephone Laboratories, Inc., 1925–. Mr. Darrow has been engaged largely in writing studies and analyses of various fields of physics and the allied sciences. Some of his earlier articles on Contemporary Physics form the nucleus of a recently published book entitled "Introduction to Contemporary Physics" (D. Van Nostrand Company).

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 is well known through his theoretical transmission studies and has published extensively on electric circuit theory and electric wave propagation.

EDWARD C. MOLINA, Engineering Department of the American Telephone and Telegraph Company, 1901–19, as engineering assistant; transferred to the Circuits Design Department to work on machine switching systems, 1905; Department of Development and Research, 1919–. Mr. Molina has been closely associated with the application of the mathematical theory of probabilities to trunking problems and has taken out several important patents relating to machine switching.

W. P. MASON, B.S., University of Kansas, 1921; M.A., Columbia, 1924; Ph.D., Columbia, 1928. Engineering Department, Western Electric Company, 1921–25; Bell Telephone Laboratories, 1925–. Mr. Mason's work has been largely in transmission studies.

L. G. BOSTWICK, B.S. in E.E., University of Vermont, 1922; American Telephone and Telegraph Company, Development and Research Department, 1922–1926; Bell Telephone Laboratories, Inc., Research Department, 1926–. While with the Development and Research Department, Mr. Bostwick's work involved general problems on systems for the high quality transmission of speech and music; since then his work has been largely on loud speakers and loud speaker measuring methods.

H. A. FREDERICK, B.S., Princeton, 1910, E.E., Princeton, 1912; Engineering Department, Western Electric Company, 1912–1925; Bell Telephone Laboratories, 1925–. Mr. Frederick is in charge of researches and engineering on telephone transmission instruments.

DONALD MACKENZIE, Ph.D., Johns Hopkins University, 1914; Assistant in Astronomy, Johns Hopkins, 1914–17; Ensign, National Naval Volunteers, 1917–1918; Bureau of Standards, 1918–1920; Engineering Department, Western Electric Company, 1920–1925; Bell Telephone Laboratories, 1925–. Engaged since 1922 in the development of a system of sound recording by photographic means.

H. M. STOLLER, E.E., Union College, 1913; M.S. in electrical engineering, 1915; Engineering Department of Western Electric Company, 1914 and 1916–1925; Bell Telephone Laboratories, 1925–. Most of Mr. Stoller's work has dealt with special problems connected

with electrical power machinery, particularly voltage and speed regulators. He designed a multi-frequency generator which is now employed in the voice frequency carrier telegraph system.

E. O. SCRIVEN, B.S., Beloit College, 1906; instructor, Fort Worth University, 1906–08; S.M., Massachusetts Institute of Technology, 1911; Engineering Department, Western Electric Company, 1911–25; Bell Telephone Laboratories, 1925–. Mr. Scriven has been identified with the development of apparatus employing vacuum tubes, e.g., amplifiers, oscillators, carrier current equipment, etc. At present his work is largely confined to public address and related audio frequency systems.

The Bell System Technical Journal

April, 1929

Electrons and Quanta [1]

By C. J. DAVISSON

The experiments by the author and L. H. Germer, by G. P. Thomson and by others from which the wave properties of electrons are adduced are briefly described. The agreement between the results of these experiments and the prediction of L. de Broglie is pointed out. The wave and corpuscular properties of electrons are compared with the similar properties of light quanta.

WHEN I discovered on looking over the announcement of this meeting that Professor Compton is to speak on "X-rays as a Branch of Optics," I realized that I had not made the most of my opportunities. I should have made a similar appeal to the attention of the Society by choosing as my subject, "Electrons as a Branch of Optics." And a very good case can be made out that electrons should be so regarded. During the last few years we have come to recognize that there are circumstances in which it is convenient, if not indeed necessary, to regard electrons as waves rather than as particles, and we are making more and more frequent use of such terms as diffraction, reflection, refraction and dispersion in describing their behavior. If this in itself is not enough to mark electrons as a branch of optics, it is sufficient at least to establish a certain community of ideas between the subjects of optics and electronics which cannot but be of interest to the members of this Society.

The evidence that electrons are waves is similar to the evidence that light and X-rays are waves. A beam of electrons is scattered by a grating—either the lattice grating of a crystal or an ordinary optical grating—and the intensity of scattering, as measured by the current density of electrons proceeding in different directions, is such as can be explained by assuming as is done in optics that what we are dealing with is the superposition of trains of scattered waves proceeding from the grating elements. In other words, current density of scattered electrons displays in these experiments the same type of spacial distribution as flux density in the analogous experiments in optics, and the observations are given a similar interpretation—an interpretation, that is, in terms of the interference of coherent wave-trains. The

[1] Presented at the Michelson Meeting of the Optical Society, Washington, D. C., November 1–3, 1928.

standard methods of optics are at once available for calculating the wave-lengths of electrons of various speeds. We do not hesitate to make these calculations, nor do we hesitate to attach physical significance to the results.

The experiments by which these phenomena are revealed have been made by Dr. Germer and myself in New York, by Thomson and Reid in Scotland, by Rupp in Germany, by Rose in England, by Nishikawa and Kikuchi in Japan, and by Szczeniewski in France. The subject is being actively cultivated at present, and there may be still other experiments of which I have not yet heard. I do not propose to give a detailed description of any of the investigations. The type of result and the methods of treating the data are so exactly those of optics—including, of course, X-rays as one of its branches—that the details would hardly interest you. I shall have something to say later on about the general nature of the results, but to begin with I shall attempt a brief account of certain theoretical speculations in which these results were more or less definitely anticipated.

It is a remarkable circumstance, and one that attests to the exceptional insight and daring of Louis de Broglie, that these newly discovered properties of the electron were suspected, and a definite hypothesis concerning them was formulated, two or more years before any of the experiments I have mentioned had been performed; even the exact relation between the speed and wave-length of the electron was accurately predicted. It is true that Einstein had at an even earlier date made use of the idea that an assemblage of gas molecules may for certain theoretical purposes be regarded as equivalent to a system of standing waves, but de Broglie seems first to have seen clearly that the duality of wave and corpuscular properties to which we are becoming reconciled in the phenomena of light might be characteristic also of electrons and material particles in general.

If light and X-rays behave in certain circumstances as if they are particles, why should there not be circumstances in which particles behave as waves? This question was suggested to de Broglie, not by any idea of the general fitness of things nor by any sense of symmetry in the universe, but by the realization, which he shared with others, that the laws of classical mechanics had been so amended in the Bohr atom model as to have become all but non-existent. It was generally felt that the Bohr atom had become too artificial to be acceptable, and that the real trouble arose from an unwarranted extrapolation of classical mechanics to systems of atomic size.

To understand the hypothesis on which de Broglie hoped to build a new model of the atom it will be necessary to have clearly in mind

the evidence that light is in some sense corpuscular. This idea had its inception in Einstein's speculations in regard to Planck's theory of the distribution of energy in the spectrum of a black body radiator. It was conceived that the energy radiated by one atom remained in some way localized in space, and could be delivered in toto to another atom or resonator suitably constituted to receive it. From this hypothesis Einstein predicted the relation which was later found to obtain between frequency of radiation and maximum electron energy in photoelectric emission, and the idea has become more and more essential to the understanding of the photoelectric phenomenon as the facts concerning it have been more and more fully revealed to us by experiment. Energy in amounts $h\nu$ is absorbed from radiation of frequency ν, not by slow accretion from waves, but instantaneously as from particle impact. The picture seemed clear enough; the energy of a beam of light is carried by corpuscles, each corpuscle transporting the amount $h\nu$. If we were willing, for the sake of pursuing this idea further, to disregard the whole gallery of interference phenomena with which it apparently conflicted, we could show that if the corpuscles possess energy $h\nu$ they must possess also momentum $h\nu/c$— this relation being required to explain the observations on light pressure in terms of impinging particles.

Thus the corpuscular theory seemed to be required to explain the photoelectric phenomena, and it might be made to explain also the phenomenon of radiation pressure. On the other hand the light corpuscle seemed a strange and ephemeral sort of particle, lacking that continuity in time which we attribute to electrons and atoms. Apparently it was manufactured within an atom for the express purpose of carrying away a part of its energy and was later destroyed in another part of the material universe to which this quantum of energy was delivered. It was difficult to regard an apparently transient entity of this sort as a particle in good standing to be classed with electrons and alpha rays.

If a certain suspicion still attaches to the light quantum in respect to its continuity, this suspicion has at any rate been considerably allayed, and the reputation of the quantum as an authentic particle correspondingly enhanced, by the discovery of the Compton effect, and again quite recently by the discovery of the Raman effect. In the first of these phenomena we see the quantum surviving an encounter with an essentially free electron, with which it exchanges energy and momentum in accordance with the ordinary laws of elastic collision; in the latter we see the quantum preserving its identity through an encounter with a molecule to which it imparts a part only of its energy.

If we are required by these recent developments to accept quanta as actual particles which carry the energy and momentum of light, how, in terms of such particles, are we to explain interference? How are we possibly to get on without waves? We even depend upon the waves to supply us with information concerning the energy and momentum of the quanta. One way in which it has been proposed to resolve the difficulty is to relegate the waves to the comparatively unimportant rôle of supplying the laws of motion of the quanta. Let us assume, for example, that when a stream of quanta passes through a narrow slit the particles do not continue in straight lines as Newton supposed, but that they spread out in such a fashion that the current density of quanta proceeding in different directions is proportional to the intensity of the light proceeding in these directions as calculated on the wave theory. In making this assumption we have given over classical mechanics and explained diffraction—or at least described it—by setting up a form of wave mechanics in its place.

With this rather crude and incomplete picture before us of light quanta being guided in their motion by waves, it is not difficult to imagine the general trend of de Broglie's speculations. de Broglie sensed that electrons like quanta might have waves to guide them—to supply the laws of their motion. That the ordinary laws of mechanics are adequate to describe the motions of electrons in discharge tubes is not inconsistent with this view, for it is well known that these laws are adequate also for a corpuscular theory of light to within the accuracy with which the phenomena are described by geometrical optics. It is only when one tries to explain diffraction that the simple corpuscular theory fails him. de Broglie envisaged a similar situation in regard to electrons—a range of small scale phenomena requiring a wave theory for their proper description. Assuming the frequency of these hypothetical waves to be given by the total energy of the electron divided by h, de Broglie was able to show that the length of the waves would be given by h divided by the momentum of the electron—and this as it happens is just the relation which obtains between the wave-length and momentum of quanta.

The goal toward which de Broglie was striving, as I have mentioned, was a new theory of the atom, and he was able to point at once to a suggestive relationship which exists between the lengths of these hypothetical electron-waves and the lengths of the circular orbits in the Bohr atom. The permitted orbits are just those which contain an integral number of these electron wave-lengths. But it was Schroedinger, as we all know, who elaborated these ideas into a comprehensive wave theory of mechanics, and showed the tremendous

possibilities of this theory in explaining the properties of the atom as revealed to us by the data of spectroscopy. At last we have an atom endowed with a constitution rather than a set of by-laws.

The success of the Schroedinger theory in explaining in a natural way the stationary states of the atom and the various rules governing transitions between these states, not to mention numerous others of its successes, must be taken as very strong evidence in favor of the fundamental idea upon which the theory is based—namely, that the duality of wave and corpuscular properties which characterizes light is characteristic also of electrons. If the evidence supplied by these data lacks something in the matter of directness, this deficiency is made good by the experiments on the scattering of electrons by crystals about which I am supposed to be speaking.

If I have been a long time in coming to the point, the time has not been wasted, for with the picture before us of the energy and momentum of a beam of light being carried by a stream of quanta for which the waves serve only to supply the laws of motion, a workable theory of the scattering of electrons is at once at hand—to a first approximation we merely read "electrons" for "quanta," and there we are. The observations on electron scattering are consistent with the view that the electrons are being guided by waves in just the way we have imagined quanta to be guided in the phenomena of optical diffraction. The only real difficulty seems to be that in the light phenomena it is not easy to believe in the particles, while in the electron phenomena it is hard to have faith in the waves.

Before going further I should like to point out that we now have two wave-lengths associated with an electron of given speed: one is the length of the X-ray waves which will be generated if the whole of the kinetic energy of the electron is converted into radiation and the other is the length of this new de Broglie wave, the so-called phase wave. The first of these wave-lengths is inversely proportional to the energy of the electron while the second is inversely proportional to its momentum. In terms of the equivalent voltage V of the kinetic energy of the electron, the lengths of the two waves are given in Ångstrom units by the formulæ

$$\lambda_x = \frac{12{,}350}{V} \qquad \text{and} \qquad \lambda_p = \left(\frac{150}{V}\right)^{1/2}$$

and their ratio is given approximately by

$$\frac{\lambda_x}{\lambda_p} = \frac{1000}{V^{1/2}}.$$

For values of V below, say, 10,000 volts the X-ray wave-length is much greater than the corresponding de Broglie wave-length.

The lengths of de Broglie waves of electrons which have been accelerated through potential differences comparable with 100 volts are the same as the lengths of moderately hard X-rays. For this reason crystal diffraction of de Broglie waves is observed with electrons of relatively low speeds—speeds corresponding to 100 volts or less—whereas, to observe the same phenomenon with X-rays, the tube producing the radiation must be operated at potential differences comparable with 10,000 volts.

The first clear evidence of the diffraction of X-rays was obtained when Laue and his collaborators investigated the scattering of X-rays—of X-ray quanta, shall we say—by a single crystal of zincblende. The analysis of this phenomenon led to the prediction and discovery of the Bragg reflection as a special case of crystal diffraction, and later on to the prediction and discovery of the special case of diffraction by aggregates of small crystals of random orientation. All three of these types of diffraction have now been observed with electrons. The Laue type of diffraction, and also the Bragg type, have been observed and investigated by Dr. Germer and myself. Diffraction by the crystal aggregates has been studied by Thomson and Reid, by Ironside and by Rupp. And observations by the Bragg method have been made also by Szczeniewski and by Rose.

I must now modify to a certain extent the picture of electron diffraction which I suggested to you a while ago. It is not quite true, as I suggested, that the only difference between the diffraction of light waves and the diffraction of electron waves is that in one case the pattern is formed by light quanta and in the other by electrons. In our investigation of the Laue type of diffraction we find, for example, that the streams of electrons which issue from the crystal do not coincide exactly in direction with the streams of quanta which would issue from the same crystal if the experiment were made with X-rays. In the case of X-ray diffraction the streams of quanta proceed from the crystal in the directions of regular reflection from important sets of atom planes, or nearly so. It is recognized that the Laue beams do not, in general, lie precisely in these directions because of a very slight refraction of the rays by the crystal. The situation in regard to electrons seems to be that electrons also are refracted and much more strongly than X-rays. The refractive indices of a metal such as nickel for electrons of low speed depart from unity much more widely than do the indices for X-rays of equal wave-length. It is a consequence of this difference that the departure from the simple law

governing the directions of beams, which in the case of X-rays is negligible, is in experiments with low-speed electrons marked and important.

Fortunately, we are not prevented by this complication from arriving at perfectly definite values of wave-lengths from observations on the Laue type of electron diffraction, and these wave-lengths turn out to be in acceptable agreement with the values of h/mv, as predicted by de Broglie.

Further evidence of electron refraction is contained in the observations we have made on the electron analogue of the Bragg X-ray reflection beam. And from the data of these experiments we have constructed a dispersion curve for nickel which displays some of the features to be expected from certain theoretical considerations. In conjunction with these measurements we have made additional determinations of electron wave-lengths, and these agree within one per cent or less with the values calculated from de Broglie's formula.

In the similar experiments made by Szczeniewski and by Rose, no certain evidence of electron refraction has been found. This may be due to some important difference in regard to refraction between bismuth and aluminium, the crystals employed in their experiments, and nickel, the crystal upon which our measurements were made. On the other hand Rupp has found evidence of refraction for a number of metals in measurements which he has made on the diffraction of low-speed electrons by crystal aggregates.

Electron diffraction differs from X-ray diffraction also in the matter of resolution. The X-ray beams are ordinarily extremely sharp because of the very great number of elements comprised in the diffracting lattice. Much broader beams are met with in electron diffraction—particularly in the diffraction of low-speed electrons—and occurrences of the beams are much less critical in wave-length. These characteristics are explained by the slight penetration of the electrons —and therefore of the electron waves—into the crystal; the effective number of scattering centers is small and the resolving power of the grating is correspondingly low.

The diffraction of electrons by crystal aggregates has been studied in Aberdeen by G. P. Thomson, who first observed this phenomenon, and by Rupp in Göttingen. Thomson has worked with thin polycrystalline foils of various metals and with high-speed electrons for which the refractive indices are practically unity. The results which he has obtained are in perfect agreement with those obtained in the corresponding experiments with X-rays—electrons of a given wave-length form exactly the same series of diffraction rings as would be formed by X-rays of the same wave-length.

Those of us who are studying electron diffraction are most fortunate in having before us a perfect model for our experiments and a fund of valuable data in the vast amount of work that has been done in the last fifteen years on the diffraction of X-rays. It is for this reason that, in spite of a rather difficult technique, so many and such varied results have been obtained in less than two years. Already we have passed on from crystal diffraction to diffraction by optical gratings. The first results of this sort were reported a month or so ago by Rupp and are in agreement with our expectations. Electrons are diffracted by an optical grating as if they were waves of length h/mv.

I have still to mention the beautiful but puzzling results which have been obtained in Japan by Nishikawa and Kikuchi. It is too bad to have to conclude my remarks with mention of the only results so far obtained which are distinctly puzzling. Nishikawa and Kikuchi have been studying the scattering of high-speed electrons by thin sheets of mica and calcite. The method of their experiment is identical with that of the original Laue experiment except that the heterogeneous beam of X-rays is replaced by a homogeneous beam of electrons. The results, as I have mentioned, are puzzling. If the incident beam is homogeneous, as stated, it is equivalent to a beam of monochromatic waves, and no diffraction pattern—or at most a very simple one—should be observed; and yet, when extremely thin sheets of mica are employed, elaborate and beautiful patterns of sharply defined spots are obtained—and patterns which cannot be readily explained even on the assumption that the incident beam contains a large range of wave-lengths, instead of a single wave-length only. When the speed of the incident beam is changed, the form of the pattern remains the same but its scale factor is altered. This also is unlike anything observed with X-rays. The results are such as might be expected if the diffracting system were a two-dimensional mesh rather than a three-dimensional lattice.

When somewhat thicker sheets of mica are used, the pattern of sharply defined spots is replaced by an array of rather fuzzy rings and lines. Again the observations are contrary to our expectations, and their explanation is far from obvious.

It may be significant that these are the only experiments, so far reported, in which the diffracting material is an insulator. But whether the clue lies here or elsewhere, it is highly unlikely, I think, that the explanation of these results will conflict with the conception we now have of electrons which are sometimes particles and sometimes waves.

The Predominating Influence of Moisture and Electrolytic Material Upon Textiles as Insulators [1]

By R. R. WILLIAMS and E. J. MURPHY

The insulating qualities of textiles vary with the amount of moisture present in them from hour to hour and are also strongly influenced by the amount of electrolytic material (salts, etc.) which the textiles contain. Electrolytic material may be washed out producing a commercially realizable increase in insulation resistance of the order of 50 times the original value.

The resistance of the animal fibers, silk and wool, is far greater for a given moisture content than that of cotton or of cellulose acetate, a derivative of cotton. It appears probable that the distribution of water as well as the quantity is important and that the two classes of fibers are characterized by different space patterns according to which the water is distributed. It is suggested that the space distribution patterns are associated with the colloidal structures of the materials and in turn with their chemical classification as proteins and celluloses respectively. Cellulose acetate absorbs little water as compared with cotton and is correspondingly superior electrically. However its resistance varies with moisture content in the same way as that of cotton.

A GREAT diversity of materials is used for insulating purposes. No simple descriptive term includes them all as the term "metals" includes commercial conducting materials. Yet in spite of this diversity it is to a great extent the quantity and mode of distribution of water in all insulators that determines their relative excellence. Were it not for the accumulation of moisture in it or on it the cheapest and mechanically most convenient material could, with rare exceptions, be used for the most exacting service.

At first glance it might seem possible to select insulating materials very simply according to moisture content, but a few illustrations will serve to show that wide contrasts exist in the response of insulations to a given amount of moisture. At one extreme is gutta percha, the classical insulation of submarine cables. If dry at the outset, it very gradually absorbs one or two per cent of moisture from the sea, but undergoes only a slight change in electrical characteristics in the process. Thereafter its water content and electrical properties are extremely stable in use. Rubber insulations used in air partake of these properties to some degree. Fluctuations in their electrical behavior never are large or sudden so long as they are mechanically intact. At the opposite extreme are the textile insulations which, especially if unimpregnated, are subject to every whim of the weather. Their water contents rise suddenly with corresponding changes in the relative humidity of the atmosphere, and the dielectric qualities

[1] Presented at the Winter Convention of the A. I. E. E., Jan. 28–Feb. 1, 1929.

faithfully reflect the moisture supplied from the air. A one or two per cent increment of moisture affects gutta percha scarcely at all but an equal amount has a most profound effect on the textiles.

The phenolized fibers, the impregnated papers, the cellulose esters, insulating varnishes and enamels, as well as glass and porcelain, are intermediate between the "waterproof" insulations and the textiles in their sensitivity to atmospheric moisture. We refer to these insulations, of course, in the forms in which they are ordinarily used, for brevity neglecting distinctions which might properly be made as to relative importance of surface and volume characteristics in the several cases.

Diverse as are the insulators in use, they have another common property of importance. They often contain, or have deposited on their surfaces, electrolytic material which dissolves in the absorbed water to form conducting solutions which are injurious to the insulating qualities of the material. This fact seems to be second in importance only to the prevalence of water in insulating materials. These electrolytic substances may be present as part of the natural constituents of the insulating material or as accidental contaminants; they may consist of the saline or organic constituents of the vegetable tissues which furnished the raw material, of by-products of the processes of manufacture, of degradation products of the insulating substances resulting from atmospheric oxidation or hydrolysis, or of atmospheric dust. Illustrative of the diversity of electrolytic material in commercial insulating materials are the natural ash constituents of textiles, pulp woods and other materials of vegetable origin; saline diluents of dyes used in fibrous materials; the quebrachitol of the latex of the rubber tree; acid resins produced by the atmospheric oxidation of rubber and gutta percha; and the free phenol present in phenol condensation products.

While the importance of moisture and of electrolytic contaminants in practical insulations has long had some recognition by electrical engineering opinion, especially in the telephone field, the foregoing general philosophy has been emphasized in the minds of the authors and their associates by the results of extended experimental studies of submarine insulation [2] and of textiles.[3]

[2] Williams, R. R. and Kemp, A. R., *Jour. Frank. Inst.*, 35 (1927). Lowry, H. H. and Kohman, G. T., *Jour. Phys. Chem. 31*, 23 (1927).

[3] *a*. Murphy, E. J. and Walker, A. C., "Electrical Conduction in Textiles. I. Dependence of the Resistivity of Cotton, Silk, and Wool upon Relative Humidity and Moisture Content," *Jour. Phys. Chem. 32*, 1761 (1928).

b. Murphy, E. J., "Electrical Conduction in Textiles. II. Alternating Current Conduction in Cotton and Silk," *Jour. Phys. Chem. 33*, 200 (1929).

c. Murphy, E. J., "Electrical Conduction in Textiles. III. Anomalous Properties of Conduction in Textiles," *Jour. Phys. Chem. 33* (1929).

Important contributions to the knowledge of the quantitative relations between the electrical properties of insulating materials and the moisture which they take up from the air have been made by Evershed,[4] Curtis,[5] Kujirai and Akahari,[6] Setoh and collaborators,[7] and other investigators. But in no published work, so far as we are aware, have data been given showing the quantitative relationships between the electrical properties of textile insulations and the electrolytic material which they contain. Data of this kind were obtained in the investigation of textiles mentioned above. Part of these data have been reported elsewhere;[3] but the investigation is being continued and a further report will be made when it is completed. It will require much further work to establish in detail the importance of contamination with aqueous solutions of electrolytes for every commercial insulating material. However, the presentation of the main thesis as a general one is abundantly justified by our constantly growing experience with cases in which such contamination of a variety of insulating materials has actually been found responsible for poor insulating qualities and for corrosion of metallic conducting or supporting elements in contact with them in electrical systems. This paper is intended to emphasize the importance of moisture and electrolytic material on the behavior of textiles as insulators and to discuss briefly the relation of electrical characteristics to physical structure and chemical constitution, so far as possible with the available facts.

General Characteristics of Textiles

It is obvious that the rapidity of response of textiles to atmospheric moisture is due first of all to their fibrousness which permits ready access to the interior of the mass through the large surfaces exposed. By contrast, the relative stability of rubber insulations, for example, is clearly due in part to the smaller ratio of surface to volume.

Since textiles are composed of fibers, it might seem that the resistance of a thread or the serving on a wire should depend largely on the resistances of the contacts between fibers. Further, the fibers themselves have superficial irregularities which would suggest that their resistance might vary widely from fiber to fiber of the same material. Table I*A* shows that single fibers of cotton and silk have a resistance [8]

[8] The experimental procedure is described elsewhere.[3a]

[4] Evershed, *Inst. of Elec. Eng. Jl.* (London) *52*, pp. 51–83, 1914.
[5] Curtis, Bur. of Standards, *Sci. Paper No. 234* (1915).
[6] Kujirai and Akahari, *Sci. Papers, Inst. Phys. & Chem. Res.* (Tokyo), *1*, pp. 94–124, 1923.
[7] Setoh and Toriyama, *Sci. Papers Inst. Phys. & Chem. Res.* (Tokyo), *3*, pp. 285–323, 1926.

TABLE I*A*

RESISTANCE OF COTTON AND SILK FIBERS

R is the resistance in megohms of a single fiber ½ in. long. Time allowed for equilibrium, 20 hrs. or more. Room temperature

Humidity, 99 Per Cent (About)			Humidity, 77 Per Cent (About)		
Fiber No.	*R*		*R* *		
	Cotton	Silk (Tussah)	Group No.*	Cotton	Silk †
1	2600	3300	1	563,000	
2	4700	4500	2	736,000	
3	3800	5140	3	680,000	
4	5600	4740	4	822,000	
5	3000	6650	5	625,000	
6	4600		6	577,000	
7	6000		7	822,000	
8	4500		8	733,000	
9	5500		9	736,000	
10	6000		10	830,000	
11	4000		11	653,000	
12	4000		12	824,000	
			13	760,000	
			14	867,000	
			15	725,000	
			16	938,000	
			17	820,000	
			18	682,000	

* Each group consisted of 60 single fibers attached to the electrodes so that they were in parallel. *R** is the average resistance per single fiber calculated from that of 60 in parallel. The experimental technique is described elsewhere.[3a]

† The values of *R** for silk at this humidity were found to be of the order of several million megohms, i.e., beyond the limit of accurate measurement with equipment available at the time the study of single fibers was under way.

which, considering the nature of the material, is surprisingly uniform for different fibers taken from the same material. Similarly, Table I*B* shows that threads [9] of cotton and silk also have a uniform resistance; this suggests that the interfiber contacts do not have a large effect on the resistance of the thread as a whole. Table I*C* shows the resistance of the servings on wires; the resistances are for short twisted pairs (2 in. long). This shows also that even where the voltage is applied transversely to the long axis of the fibers—which would tend to make contact resistances more important than when the voltage is applied parallel to the long axis—the resistance of different samples of the same material is fairly uniform. These facts suggest that inter-

[9] Because of their uniformity, small samples of thread (½ in. lengths) have been used in this laboratory as a convenient means of comparing the insulating quality of cottons and other textiles.

TABLE 1*B*

Resistance of Cotton and Silk Threads

Length ½ in. Temperature 25 deg. cent.

Sample	Resistance Megohms Cotton Humidity 77%	Silk (Spun) Humidity 90%
1	4160	21,100
2	4220	22,700
3	4100	28,000
4	3730	14,500
5	4020	30,600
6	3820	
7	3900	
8	3715	
9	4050	
10	4100	
	Aver. 3982	23,380

TABLE 1*C*

Resistance between Twisted Pairs of Cotton and Silk Insulated Wires

Humidity 77 per cent Temperature 25 deg. cent.

Sample	Resistance Megohms Cotton	Silk (Tussah)
1	5.45	2200
2	3.96	1890
3	5.20	1685
4	3.96	1685
5	3.50	2250
6	4.05	1970
7	5.60	
8	5.15	
9	4.37	
10	4.76	
11	4.00	
12	4.37	
	Aver. 4.53	1942

fiber contact resistances are only secondary or negligible in determining the resistance of a thread or other mass of fibers. Further evidence of this is given by the data in Table II, which show that even when the length of a thread considerably exceeds the length of a single cotton fiber, the resistance is approximately proportional to the length; if interfiber resistances were large, the resistance per unit length would increase considerably with the length of the thread measured.

The above results also suggest that electrical conduction takes place primarily through moisture in the interior of the fibers rather

than through moisture condensed on their surfaces. Other evidences that this is the case may be found in the relationships of conductivity to humidity, moisture content and electrolyte content, as well as the absence of any obvious relationship between the physical dimensions of different classes of fibers and their electrical behavior.

TABLE II

RESISTANCE OF DIFFERENT LENGTHS OF COTTON THREAD

Humidity about 77 per cent Room temperature

Length Inches	Resistance Megohms Total	Per Inch
0.5	21,700	43,400
1	41,750	41,750
3	153,000	51,000

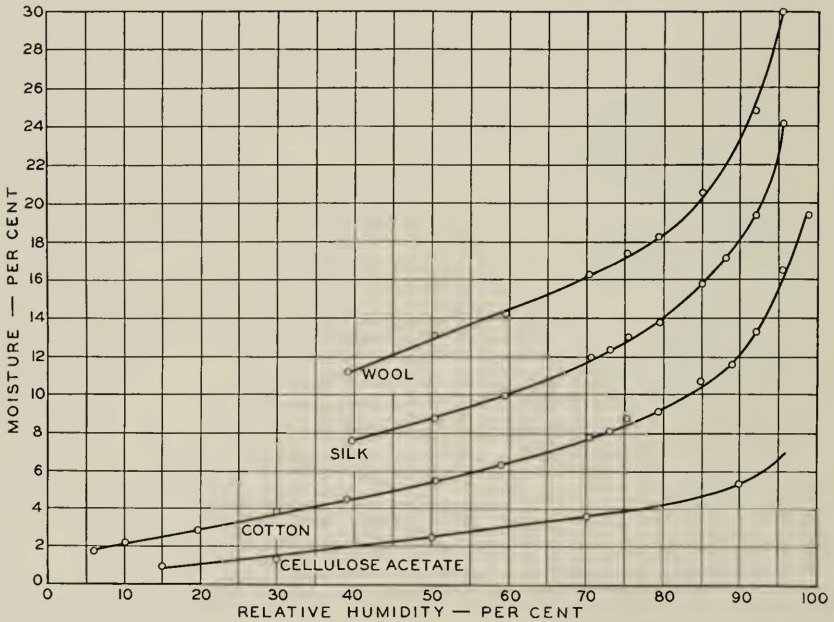


Fig. 1—Dependence of moisture content of textiles upon relative humidity of atmosphere with which they are equilibrated

While the form of the sample is not of predominating importance with reference to the insulation resistance of either cotton or silk, the marked contrast except at very high humidity between cotton and silk in all forms of samples should be noted. Both these facts and other available data justify the inference that the dielectric properties of textiles are determined primarily by the composition or

internal structure of the fibers, not by the twist of threads or the lay of servings.

The moisture content of each sort of textile depends directly on the humidity of the atmosphere. Fig. 1 shows the best data available for the moisture content of silk, wool, cotton, and cellulose acetate in equilibrium with air over considerable ranges of relative humidity.

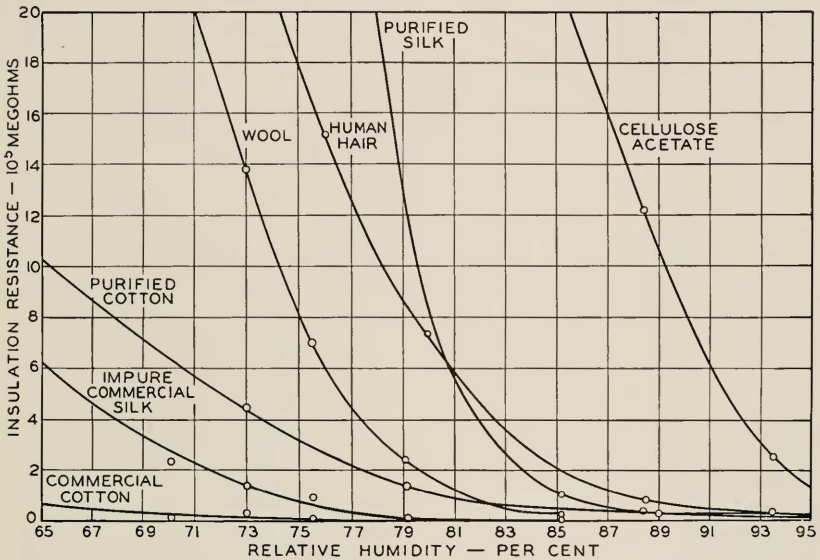


Fig. 2—Insulation resistance of ½ inch lengths of textile threads as affected by relative humidity of atmosphere. The purified cotton and purified silk had been submitted to a washing procedure to remove electrolytes. The impure silk was a commercial specimen representing somewhat more than usual contamination with electrolytes, while the commercial cotton is representative of its class.

The data for silk and wool were taken from a paper by Schloessing;[10] those for cotton are due to Urquhart and Williams;[11] while those for cellulose acetate represent the figures of Wilson and Fuwa,[12] who also give corresponding data for several textiles and many other substances. It is sufficient for our present purpose to emphasize the orderly dependence of moisture content upon the relative humidity of the atmosphere without discussing secondary phenomena or the full significance of the curves.

The relation of electrical behavior of each textile to relative humidity is also very close. Fig. 2 shows the insulation resistance of each of

[10] Schloessing, Th., *Bul. Soc. Encour. Indust. Nat. 8*, 717 (1893); C. R. *116*, 808, 1893. Text. World Record, Boston, Nov. 1908, p. 219.

[11] Urquhart and Williams, *J. Textile Inst. 15*, 143 (1924).

[12] Wilson, R. E. and Fuwa, Tyler, *Ind. & Eng. Chem. 14*, 913 (1922).

the above fibers plotted against relative humidity over the upper part of the range of atmospheric humidities. It is not practicable to plot the resistance over the entire range of humidity directly in this way, on account of the wide range of insulation resistance values which are obtained. In order to depict the fact that there is a consistent relationship throughout the range, we have plotted in Fig. 3

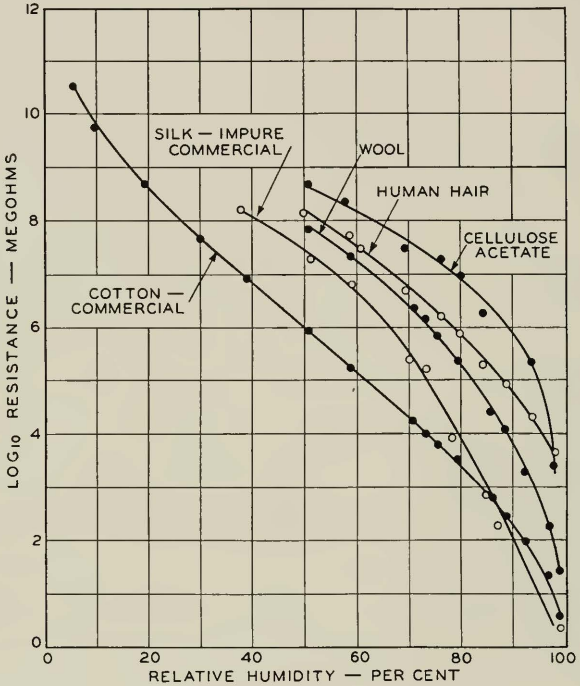


Fig. 3—Log insulation resistance of ½ inch lengths of textile threads *vs.* relative humidity of the atmosphere. For description of samples see Fig. 2. For tabulated data see reference 3*a*

Log Insulation Resistance *vs.* Relative Humidity as far as values are at present available. When considered together, these three charts show that the insulation resistance of a textile depends on its moisture content, which in turn is a function of relative humidity. In the series of papers previously mentioned,[3] the electrical behavior of textiles in relation to relative humidity and moisture content is discussed more fully.

Aside from the common property of dependence of electrical characteristics of textiles upon their moisture contents, several other phenomena of electrical behavior have been encountered which are

common to all. If any textile within our experience, including cellulose acetate (or even glass), be brought in contact with two electrodes of opposite polarity in the presence of atmospheric moisture, the electrical properties of the material undergo a change with a rapidity dependent on the current and in turn upon the voltage, the length of path, and the humidity. Such a change in the properties of insulating materials with continued application of voltage has been discussed recently by Granier, who advanced the explanation [13] that it is due to the presence of electrolytic impurities in the materials. However, the great magnitude of this change, which may occur in textiles when freely exposed to ordinary atmospheres, seems to have been very little appreciated.

TABLE III

RATE OF CHANGE OF RESISTANCE WITH TIME OF APPLICATION OF VOLTAGE FOR SOME FIBROUS MATERIALS

Separation of Electrodes ½ in. Humidity 97 per cent (approx.) Voltage 275

Cotton		Silk		Cellulose Acetate Silk		Paper	
Time Secs.	Resist. Megohms	Time Secs.	Resist. Megohms	Time Secs.	Resist. Megohms	Time Secs.	Resist. Megohms
70	19.8	30	8.1	80	9.3×10^3	55	4.54
150	33.6	90	12.8	1250	1.01×10^4	100	8.30
200	40.5	120	19.8	2300	1.06×10^4	200	11.2
300	45.1	153	26.7	3200	1.11×10^4	260	19.8
400	50.0	220	40.5			520	32.2
500	57.5	300	47.8			615	38.5
900	68.2	640	54.8			1000	42.7
1900	82.0	1250	58.4				
3000	91.0	2100	79.5				
4000	99.0	2320	107.0				
		3000	153.0				

In our experiments the insulation resistance rises to a value perhaps 10 to 100 times the original value, depending on the nature and condition of the fiber. A few typical cases are given in Table III. This phenomenon will be referred to as polarization. This rise in resistance appears to be largely due to substantial denudation of some intermediate portion of the fiber of electrolytic impurities, which in general tend to accumulate in the vicinity of the electrodes. The phenomenon involves the possibility of chemical reactions between the products of the electrolysis and the material of the electrodes, as well as the evolution of gases and the conversion of soluble salts into insoluble products. As regards the mineral constituents of

[13] Granier, J., *Soc. français Elec. Bull.*, *3*, 480 (1923).

cotton, it can be observed by ashing the polarized fiber that the ash lies largely in those portions which were adjacent to the electrodes and especially to the cathode. The electrical resistance is very unequally distributed along such a polarized fiber or thread, the positions of maximum resistance depending upon the conditions of polarization.[3c] Interruption of the current after polarization leads to a gradual restoration of the original electrical properties and a redistribution of the ash constituents with a speed depending largely on the humidity. Reversal of polarity is accompanied by a rapid drop in insulation resistance, followed upon continued application of voltage in the reverse direction by polarization in the opposite sense. Interruption of the circuit after polarization leaves large potential differences on the opposite ends of the fiber, which persist for several minutes. The cathode region is found to be alkaline in reaction, the anode region acidic. The polarized fiber is therefore a concentration cell.

The electrolysis of cotton may be carried out experimentally in another way. If cotton yarn is immersed in water in each of a series of cells separated from one another by a parchment paper membrane and a direct current is passed through the cells for some hours, the impurities tend to accumulate near the electrodes, with the development of acidity at the anode and alkalinity at the cathode, as is usual in the electrolysis of a saline solution. If the samples of cotton yarn be now removed and brought into equilibrium with an atmosphere of standard humidity, the insulation resistance is found to vary fairly regularly with the original position of the sample in the series of cells, being greatest in some intermediate cell and diminishing toward either electrode. The precise position of the maximum varies with the nature of impurities present in the system. The highest insulation resistance may be many times that of the original cotton.

Perhaps the most significant evidence of the importance of electrolytic impurities in silk, wool, cotton, and to some extent other textiles, is the fact that their electrical characteristics can be greatly improved by thorough washing with water though without altering qualitatively the general nature of the electroconducting phenomena which characterize them. Fig. 2 illustrates the result of washing upon the insulation resistance of cotton and silk threads. The improvement in insulation resistance of cotton and silk upon washing ranges commonly from fifty to one hundred fold, under any of the commonly prevailing conditions of atmospheric temperature and humidity. This improvement is accompanied by diminution of the ash content, in the case of cotton from about 1.0 per cent to 0.15 or 0.25 per cent. It produces only a slight reduction in the equilibrium moisture content

of the cotton over the ordinary ranges of atmospheric humidity. The sensitivity of the washed cotton to continued application of voltage is much less than that of the original, but polarization still occurs. Commercial silks are similarly affected by washing.

If the mineral contents of cottons which have undergone washing are compared quantitatively with the original contents a decrease is observable, particularly as to potash, but the calcium and magnesium contents are much less altered. Fairly complete removal of potash is apparently essential to good electrical characteristics, but improvement electrically has been attended in some cases by an actual increase in content of alkaline earths. This suggests that interchange of electrolytic impurities between the textile and the water is involved as well as actual removal of electrolytes by the water. Thus in general hard natural waters, i.e., those containing calcium and magnesium salts, have proved as good or better than soft waters when used in economically small amounts. Very exhaustive extraction with distilled water gives excellent results, though not vastly superior to washing with very dilute solutions of alkaline earth salts. Sufficiently complete and accurate analyses of samples of textiles brought into equilibrium with washing liquids and of the kind and quantity of electrolytes in the corresponding liquids have yet to be made to determine the precise importance of the composition of the saline residues. Non-saline electrolytes have also to be considered. This matter requires extended study and the experimental data are reserved for future publication.

The commercial value of such treatments of insulating yarns has proved to be very substantial. The utilization of the products forms the subject matter of another paper[14] from the Bell Telephone Laboratories.

Another common property of textiles is known as the Evershed effect. Evershed [15] found in various insulating materials, including textiles, that insulation resistance does not obey Ohm's law but is less if a larger measuring voltage is used. Evershed's finding as to cotton has been verified by us. This result is easily obtainable if conditions are maintained so that little polarization occurs. But if extensive polarization is allowed to take place the reverse effect is observed and the ultimate resistance is higher in proportion to the voltage used. The conditions which favor polarization are, of course, considerable voltages, prolonged application, high relative humidities, and short paths through the insulation. Evershed's work apparently

[14] Glenn, H. H., and Wood, E. B., This Journal.
[15] Evershed, S., *Inst. Elec. Eng. Jl. 52*, 51 (1914).

did not involve any special attention to the time of application of voltage.

Evershed's explanation of decreased insulation resistance with increased voltage involves the assumption that much of the water contained in insulations is originally in the form of isolated pools and therefore of no conductive effect at the instant of application of voltage. In support of his theory of "dormant" water, he lays great stress upon his observation that the volume of water present in the insulating materials is far in excess of that which would be required to furnish the observed conductivity if the water were in the form of continuous filaments of uniform cross section. This argument seems impressive and conforms to our own ideas of the distribution of water in textiles. However, according to Evershed, this pool water is electrokinetically spread out into conducting films under electric stress, thus accounting for decreased resistance with increased voltage. A tendency to such movement of water cannot be denied. But it is difficult to harmonize Evershed's conception of electroendosmotic movement of water as the predominating phenomenon with all the facts regarding textiles with which we have had experience; for example, with the fact that the Evershed effect is greatest when the electrolyte content of the textile is high. Electroendosmose in systems designed for its ready detection usually diminishes with increasing electrolyte content except when the concentrations are very small.[16] Further a decrease of resistance with increase of voltage has been noted in other systems in which electrolysis is unquestionably involved and in which electrokinetic redistribution of water seems improbable.

Though sufficient support for an alternative cannot be furnished at present, electrokinesis does not constitute the sole possible explanation of the Evershed effect. The analogy between the moist fiber and an electrolytic cell conforms to a number of other corollary facts about the Evershed effect which are discussed in a more specialized paper by one of the authors.[3c] The various properties which have been discussed are in agreement with the view that the cardinal principle of conduction in textiles is the electrolysis of aqueous solutions.

Distinctive Characteristics of Each Fiber Species

The several kinds of fibers exhibit a number of curious contrasts in the relation of electrical behavior to hygroscopic properties, some of which at first glance appear contradictory. For convenience in dis-

[16] Powis, Frank, *Zeit. Physik. Chem. 89*, 91, 1914 and Burton, E. F., *Coll. Symp.*, Monograph IV, 132, 1926.

cussion, let us classify the commercial fibers into two main groups: (1) the animal fibers, and (2) the vegetable fibers, and a subgroup (2a), the cellulose ester fibers of which the so-called cellulose acetate silk is the sole representative of commercial importance at present. It will be seen by reference to Fig. 1 that over the entire range of relative humidity the animal fibers, silk and wool, absorb more water

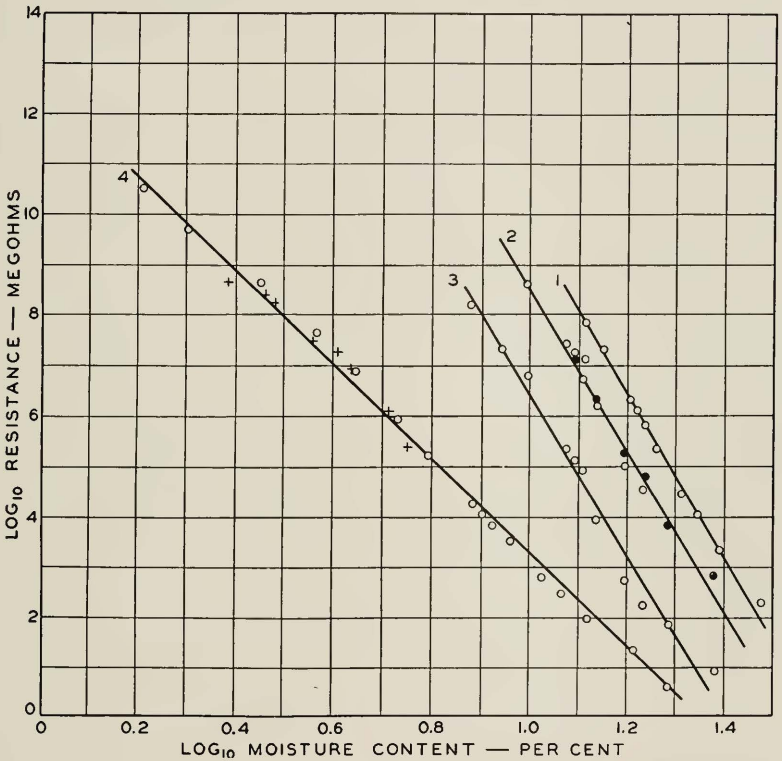


Fig. 4—Insulation resistance as a function of moisture content of textiles
1. Wool yarn
2. Silk threads purified [sample 1 (o); sample 2 (●)]
3. Silk threads impure
4. Cotton threads (o); cellulose acetate threads (+)

than the natural vegetable fibers. This is true whether we deal with fibers in their natural impure state or after a washing process which has been shown to improve greatly the electrical characteristics of both types of natural fibers. Cellulose acetate absorbs less water at any given humidity than either class of natural materials.

We have seen that for any given kind of fiber there is an orderly dependence of electrical properties upon the moisture content of the

fiber and in turn upon the relative humidity of the atmosphere. The more water present in any given fiber, the poorer are the electrical properties. If the amount of water in fibers were the sole determinant of electrical characteristics we would expect the animal fibers to be, at a given humidity, the poorer electrically of the classes enumerated above. But this is emphatically not the case. With respect to electrical properties, we find that the vegetable fibers are inferior to the more hygroscopic animal fibers and are also inferior to the least hygroscopic variety of commercial fiber, viz., cellulose acetate. To make the existing contrast clear we have plotted in Fig. 4 for the several textiles, Log Insulation Resistance *vs.* Log Moisture Content. When so plotted the values for each kind of fiber fall approximately on straight lines throughout the range of actual measurement. The relative position of the curves for animal fibers to the right and above that for cotton [17] means that the animal fibers have the better insulating qualities in spite of higher hygroscopicity.

The slopes of these lines have an even greater significance for they indicate the relative sensitivity of the fibers to an increment of moisture. Since the slope for the animal fibers is greater, it is evident that the animal fibers are more sensitive electrically to moisture than cotton. Under a given set of conditions they are not only wetter than cotton, but are more sensitive to the effects of further increments of moisture and yet they have a higher insulation resistance.

In one respect alone can we say that cotton is preferable as an insulating material. It has the merit of being more nearly uniform in behavior under a variety of weather conditions. When the amount of moisture taken up from the surrounding atmosphere by silk or wool is doubled, the electrical leakage increases by a factor of 50,000 to 100,000, while that for cotton rises by a factor of only 600.

The position and slope of the values for cellulose acetate are of great interest as the curve coincides with that for cotton, indicating that moisture affects these two fibers in a very similar manner. The essential electrical difference between these two appears to be satisfactorily accounted for by the fact that cellulose acetate absorbs less water than cotton at any given relative humidity of the atmosphere. The conversion of cotton which is essentially cellulose into cellulose acetate by the process of acetylation, has, as could be predicted on chemical grounds, reduced its hygroscopicity but apparently has not modified its structure greatly. Cellulose acetate is therefore put in a subgroup rather than in an independent classification.

[17] The threads referred to are approximately, but not precisely, of the same size. This variable is by no means sufficient to account for the higher position of the animal fibers as compared with cotton.

The reader will have been led to ask several questions. Why do the several kinds of fibers differ so much in absorption of water and particularly why does not a given amount of water affect them all alike electrically? He will want to know if the mere fact of animal, vegetable, or artificial origin is associated with a particular type of electrical or hygroscopic behavior. He will wonder whether there is any fundamental resemblance between the silk which is rapidly extruded by a worm and the hair which grows so slowly on a sheep's back. He will inquire whether there is sufficient justification for classifying other vegetable fibers with cotton. To these questions only partial and to some extent speculative answers can be given.

We are justified on chemical grounds in classifying the fibers in the same way which we have found to be convenient for discussion of their hygroscopic and electrical properties. How much importance should be attached to this correspondence between the chemical and electrical classifications cannot be determined at present. However, the correspondence seems suggestive and deserving of a brief discussion. The first class, that of animal fibers, has a common chemical nature in that they consist largely of proteins. Proteins are molecular aggregates of colloidal size composed in turn of simpler substances known as amino acids. Each of the constituent amino acids contains both an acidic and a basic group, so that in acid solutions they behave as bases and in alkaline solutions they behave as acids. This so-called amphoteric property is due to the presence of an acidic oxygen nucleus and a basic nitrogen atom, which are almost invariably adjacent to one another. Amphoteric properties persist in the proteins which are formed by union of many amino acids in a single molecule. This is illustrated by the fact that either amino acids or proteins in a solution which is subjected to a d.c. voltage tend to migrate to positions intermediate between the electrodes where the acidity is such that they are equally ionized as bases and as acids. The combination of adjacent acidic and basic groups within a single molecule gives them a salt-like property which may be significant. It is reasonable to associate the hygroscopic quality of the proteins with these groups as the molecules are usually without groups of polar character other than the paired groups mentioned.

That their common protein character is responsible in some way for the properties of principal interest from the insulating standpoint is rendered the more probable by the close resemblance of silk and wool, as shown by the approximate parallelism of their curves in Fig. 4. This resemblance is shared in considerable measure by other hairs than wool.

The second class of fibers, coming from the vegetable world, are alike in being composed of cellulose, a substance like the protein in having a high molecular weight but unlike it in that its polar groups are hydroxyls which have a faintly acidic rather than amphoteric nature. These are the groups in cellulose with which water is likely to associate itself. Such data as are available concerning vegetable fibers other than cotton, notably linen, ramie, manila hemp, and wood pulp, indicate a strong resemblance not only chemically but hygroscopically and electrically.

The subclass embracing only cellulose acetate as a commercial fiber is chemically more neutral and non-polar in type than other cellulose fibers, with which fact it is reasonable to associate its lower hygroscopicity and consequent better electrical characteristics. It is probable that cellulose nitrate and cellulose ethers will be found to fall in this class but artificial silks other than cellulose acetate absorb more water and appear on chemical grounds to be better classified with the cellulose fibers of natural vegetable origin.

It cannot be decided from available information whether the similarities and differences in electrical properties among the textiles are traceable directly to chemical similarities and differences or indirectly to physical (colloidal) structures which in turn are determined by chemical composition. In either case it is possible to account for the high sensitivity of all the fibers to moisture and the variation in sensitivity from species to species by assumptions as to the distribution of water in them. Water which collects in any isolated form in the material will have little electrical effect compared with that which forms continuous filaments. The distinction we are making is essentially the same as that of Evershed when he referred to part of the water as "dormant," though we do not attach the same importance as he does to electrokinetic redistribution of water under electric stress. Each increment of water may be considered as undergoing partition into two portions, one causing a large increase of conductivity and one having a negligible effect, in a ratio determined in some fashion by the structure or nature of the material and the humidity of the atmosphere. The ratio of the two portions which are in equilibrium via the surrounding atmosphere will be subject to constant readjustment under changing conditions.

The fact that the electrical characteristics of the two classes of fibers as affected by moisture appear to be specific properties of the substances involved suggests some highly regular distribution pattern of conducting water paths determined by the chemical or physical (colloidal) structure of the material. Such a regular pattern may

involve only water condensed upon the surfaces of the elements of structure in such a way that the thickness of the film varies regularly from point to point through the material. Accumulation of water at thick points would have little electrical effect, while that in thin portions would be very significant. An alternative regular mode of distribution would involve water in part dispersed in solution or chemical combination within the units of structure of the material and in part in fairly uniform thin films on their surfaces, in which case the latter would have the major electrical consequence. While such a regular form of distribution seems preferable, it is perhaps not the only way of accounting for the electrical properties observed.

The curves as shown in Fig. 4 for both cotton and silk are straight lines within the experimental error but the assumption is not justified that they can be projected as straight lines to zero humidity. At the lower ranges of humidity the resistance of silk is so high as to exceed the limits of our present technique of measurement. Over some range below 40 per cent relative humidity, it may well be that the sensitivity of the animal fibers to increments of moisture is less than that of the cellulose fibers. If so, the break in the curve would have great interest in connection with determining the mode of water distribution and in turn the colloidal structure of the materials in question. It is hoped that an extension of the study in this direction will be possible in the future.

Summary

Attention is called to the practical importance of water in all insulators and especially to the extreme electrical sensitivity to moisture of textiles as a class.

Significant amounts of electrolytic impurities occur in many insulators.

In textiles in the presence of moisture such impurities are responsible for very conspicuous features of electrical behavior. Data are given showing their effect on the insulation resistance of cotton and silk.

The insulation resistance of textile fibers in moist air rises greatly with duration of d.c. voltage, accompanied by many evidences of electrolysis of aqueous solutions of impurities in the textile.

The instantaneous insulation resistance of fibers decreases with increase of the measuring voltage as previously shown by Evershed. However, this fact does not necessarily support his idea of kinetic redistribution of water in textiles, as this behavior is also compatible with the nature of electrolytic conduction.

Electrolytic impurities may be washed out of textiles and sub-

stantial practical improvements effected thereby. The increase of resistance is of the order of 50 times.

Fibers are classified according to their electrical behavior in a manner which is also in harmony with their chemistry as follows:

(*1*) *Animal fibers.*

These are of protein nature and are characterized by high moisture content at ordinary humidities and by great electrical sensitivity to further increments of moisture, yet possess excellent insulating properties under usual atmospheric conditions.

(*2*) *Vegetable fibers.*

These are of cellulosic nature and are characterized by lesser moisture absorption and lesser electrical sensitivity to further increments of moisture, yet possess relatively poor insulating properties over the range of prevalent atmospheric conditions.

(*2a*) *Cellulose acetate.*

This absorbs little water and accordingly has excellent insulating properties under like conditions.

The differences in electrical behavior of the two main classes of textiles are believed to be due to differences in the space patterns according to which water is distributed within the individual fibers. The patterns are probably determined directly or indirectly by the chemical composition of the fibers and associated with the colloidal structure.

The authors wish to acknowledge their indebtedness to their colleagues, whose names appear as authors of kindred papers, for their advice and assistance. Also we wish especially to thank Dr. Homer H. Lowry, whose discernment has stimulated the development of evidence necessary to several of the more important deductions.

Purified Textile Insulation for Telephone Central Office Wiring

By H. H. GLENN and E. B. WOOD

This paper outlines methods by which silk and cotton insulation can be purified and improved. It gives the results of tests on the insulation properties of these materials before and after purification and explains the testing procedures. One of the findings is that the purified cotton may be substituted for ordinary commercial silk.

IN a contemporary paper, *The Predominating Influence of Moisture and Electrolytes upon Textiles as Insulators*, Messrs. Williams and Murphy have shown that the electrical properties of textiles are closely associated with the moisture content and impurities in the textiles. In particular, water-soluble salts become ionically conducting in the presence of moisture and the ions migrate along the paths of initially low resistance to the electrodes with which they react chemically, causing serious corrosion. The resulting corrosion products, themselves electrolytes, accelerate the process of current transfer and may easily lead to a complete failure of the insulating textile at the point of greatest concentration. Conversely, if the impurities are removed, the insulating properties of the textile are improved initially and, furthermore, are not subject to cumulative deterioriation due to concentration of conducting salts and electrolytic corrosion products at the weaker points. It is the purpose of this paper to show how these principles are borne out by field observations and laboratory tests, and to show in a general way the extent to which the insulating properties of silk and cotton can be improved commercially with particular application to telephone central office wiring.

Since the early days of telephone development work, silk and cotton have been the standard insulating materials for wire insulation in telephone central office apparatus, supplemented in later years by enamel insulation. Relatively low voltages have always been used in the telephone plant, 24 to 48 volts being the usual voltages which are carried continuously in cables, while intermittent a.c. and d.c. potentials generally do not exceed 100 to 150 volts. Therefore it has been generally accepted that telephone cables once installed and properly protected from accidental high voltages, could be depended upon to have a substantially indefinite life. In general the

[1] Presented at the Winter Convention of the A. I. E. E., New York, N. Y., Jan. 28–Feb. 1, 1919; Abridgment published in *A. I. E. E. Journal*, February, 1929, p. 146.

insulation of these cables has been satisfactory, but breakdowns have occurred which could not be attributed to faulty operating conditions or to manufacturing defects. A study of this subject showed that it was possible under certain conditions to get discolored or faded spots in the insulation and corresponding corroded or pitted spots in the tinned copper conductors. It was also observed that the textile insulation at such spots showed a strong concentration of water soluble salts. Also, cables in which such conditions occurred measured relatively low in insulation resistance with the current leakage concentrated at these points. These observations led to the conclusion that silk and cotton would be decidedly improved as insulating materials if they were made less susceptible to deterioration under telephone service conditions.

Aside from the consideration of improving silk and cotton to assure greater insulation stability, considerable thought has been given to the possibility of improving the insulating characteristics of cotton to such a degree that it could be substituted for the more expensive silk. The importance of this work with respect to its bearing on the cost of telephone service can be better appreciated from the fact that about 2,000 pounds of silk are required daily to provide for the growth of the country's telephone requirements, which if replaced with cotton would reduce raw material costs by a very substantial sum.

The desirability of reducing the quantity of silk required in the telephone plant does not arise entirely from this phase of the economic question. The problem of supply and demand has at times entered into the matter. For example, shortly after the close of the World War the supply of insulating silk was limited and the price prohibitively high. Substantially the same condition arose a few years later, which leads to the conclusion that silk is inherently much more subject to violent fluctuations in available supply and cost than cotton. Therefore, with demands for telephone equipment rapidly increasing, we have decidedly greater assurance of an adequate supply of insulating material at reasonable cost if cotton instead of silk is used.

Purification Process

With the foregoing as an introduction to indicate the economic advantages to be gained by improving the electrical characteristics of cotton and silk, the following is intended to show what has been accomplished by the commercial application to silk and cotton thread of the processes referred to by Messrs. Williams and Murphy for removal of objectionable impurities.

Since such impurities are soluble in water, it will be inferred that

the purifying process consists in a thorough washing with water. In effect, this is the case. The process, however, for both silk and cotton, being based on substantially complete removal of the ionically conducting salts, especially those of sodium and potassium, prescribes the use of water of low saline content. It also means that the washing

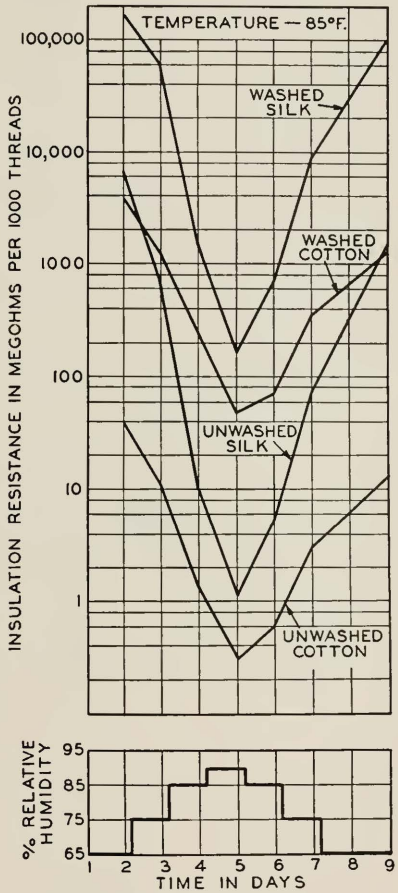


Fig. 1—Typical d.c. resistance characteristics of washed and unwashed silk and cotton threads of equal size.

is best accomplished by a continual flow which after passing through the textile is considered to be contaminated and is not used again.

Where cotton is to be dyed and washed, the washing consists in an additional operation applied to the cotton immediately following the dyeing operation without the necessity of drying between processes.

Characteristics of Purified Insulations

Obviously, the first consideration in the insulation of electrical conductors is to provide an insulating medium of sufficient dielectric strength to withstand the working potentials to which it is subjected. Also, the d.c. insulation resistance must be high enough

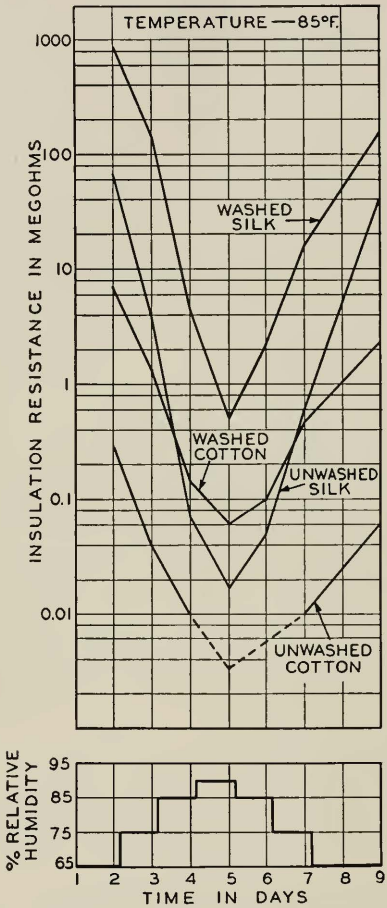


Fig. 2—D.C. insulation resistance of 50 ft. of twisted pair wire insulated with double servings of equal thickness.

to prevent undue d.c. energy loss. A comparison of the electrical resistance of the cotton and silk at relative humidities ranging upward from 65 per cent to 90 per cent and down again to 65 per cent, before and after washing, is shown by the graphs in Fig. 1 as determined by samples prepared and tested by the method and testing apparatus

shown in Figs. 6, 7 and 8 and described later. The same comparison is shown in Fig. 2 except that these graphs show the insulation resistance of wire insulated with the washed and unwashed textiles. In addition to the insulation resistance requirement, it is required that the energy losses at talking and carrier current frequencies must be maintained at the minimum point consistent with the space limitations permitted for the conductors. The effect of purification of the textiles on this characteristic expressed in capacitance and conductance, measured at 1,000 cycles per second between the wires of twisted pairs is shown in Fig. 3 and Fig. 4. The data represented by these graphs converted into transmission loss units are illustrated in Fig. 5. As the same thickness of insulation was used in all cases, the graphs are on a comparative basis. It should be noted that the graphs are illustrative of the effects of purification on the electrical properties of cotton and silk as insulation and should not be considered as applying quantitatively to telephone circuits.

From a telephone transmission point of view, perhaps the most significant fact to be observed is the large reduction in capacitance and conductance at relative humidities of 75 per cent and higher. These characteristics which largely determine transmission efficiency are relatively low for both silk and cotton at 65 per cent and below, but in commercial textiles in general use for insulating purposes they increase very rapidly as the relative humidity increases. The characteristics of purified textiles are not as markedly different from those of unpurified textiles at 65 per cent relative humidity as at higher humidities, but their rate of increase as the humidity increases is greatly reduced. This fact is of particular importance in the maintenance of a standard level of voice transmission through toll offices where suitable repeater gains and balance must be maintained. Losses, if fixed in value and not excessively large, can be compensated for, but if they change with every change in atmospheric moisture content the compensation problem becomes serious.

Method of Testing

Two fundamental characteristics of silk and cotton made it necessary to do a large amount of experimental work before a practicable shop test method could be established to determine whether or not the textiles were washed to the point of meeting the requirements established. One of these characteristics is the high electrical resistance of both washed and unwashed textiles at the lower relative humidities and the other the extreme sensitivity to change, with minor change in relative humidity especially at the higher humidities. The first

mentioned characteristic precludes the use of any but measuring instruments of the highest degree of sensitivity and makes desirable the use of comparatively high humidities, and the second characteristic

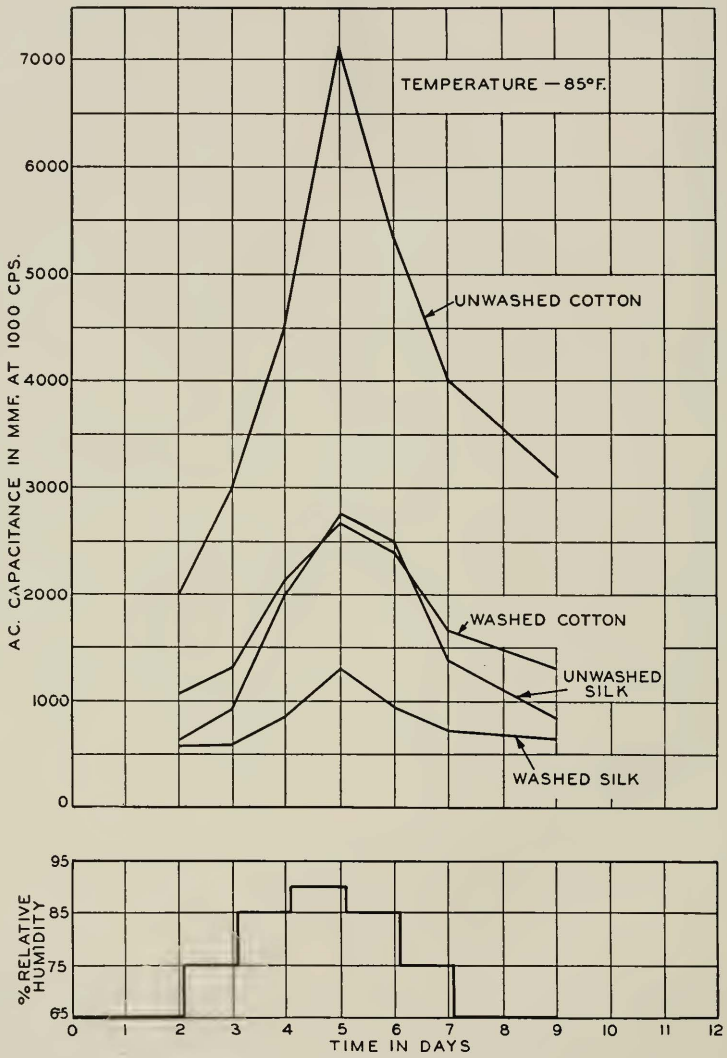


Fig. 3—A.C. capacitance of 50 ft. of twisted pair wire insulated with double serving of equal thickness.

means that the specimen must be tested under exceedingly well controlled relative humidity conditions. Furthermore, the problem is complicated by the polarization effect discussed in the paper by Williams and Murphy, and the fact that this effect varies in magnitude

with humidity and with the degree of purity of the textiles. The problem was finally solved by the development of the test equipment shown in Figs. 6, 7 and 8.

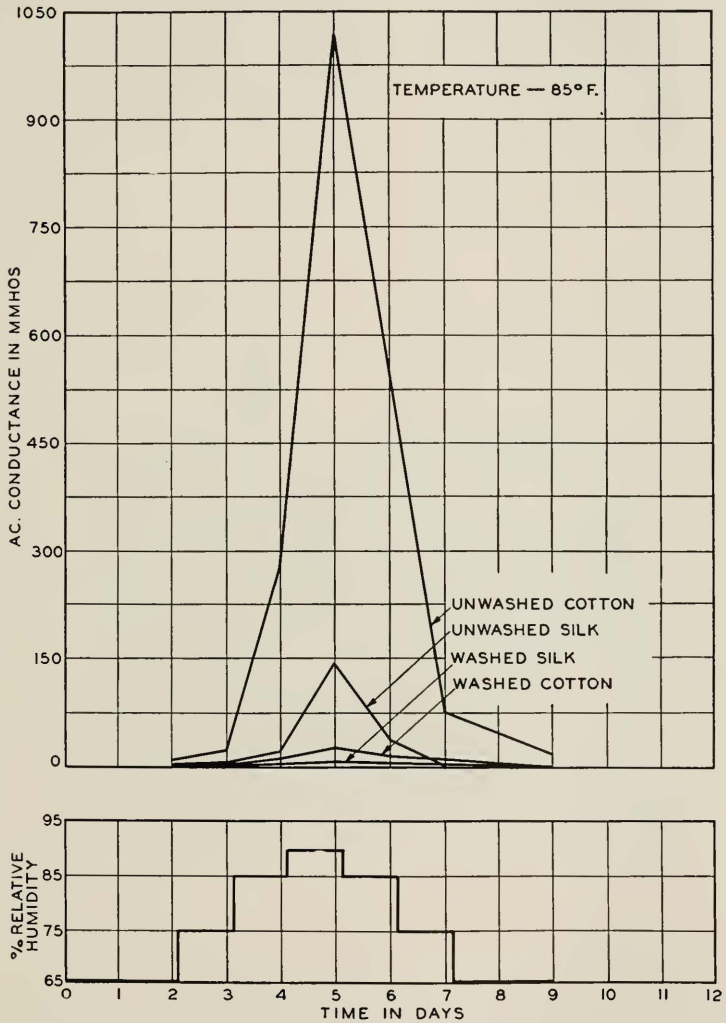


Fig. 4—A.C. conductance of 50 ft. of twisted pair wire insulated with double servings of equal thickness.

Figs. 6 and 7 show a heat-insulated glass tank of about one cubic foot capacity fitted with an insulating cover in which holes normally closed with stoppers are used to introduce the test samples. The humidity is maintained by means of sulphuric acid or a saturated

salt solution in the bottom of the tank and constant temperature within very narrow limits is maintained in the tank by placing the entire assembly inside a cabinet or oven automatically controlled to ± 0.5 degrees Fahrenheit. Due to the heat insulation it has been

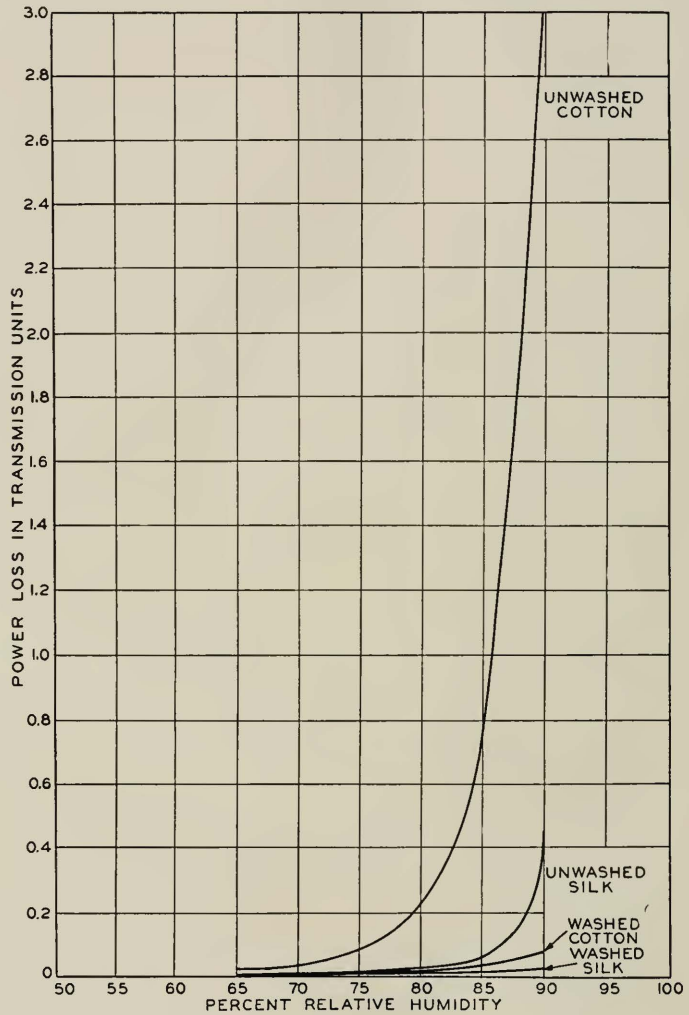


Fig. 5—Transmission loss in 50 ft. of twisted pair wire insulated with double servings of equal thickness.

found that temperature variations within the tank are reduced to the vanishing point for all practical purposes.

This is very important as it has been found that fluctuations in the temperature of the test chamber introduce large errors in the insulation

resistance of the samples. The errors, however, are attributed not to temperature effects on the samples but to variations in relative humidity produced by the temperature changes and the considerable time required for equilibrium to be restored after such changes occur.

Another source of error in textile testing is found in the fact that the values of insulation resistance are affected by the humidity condition to which the sample has been exposed prior to the test.

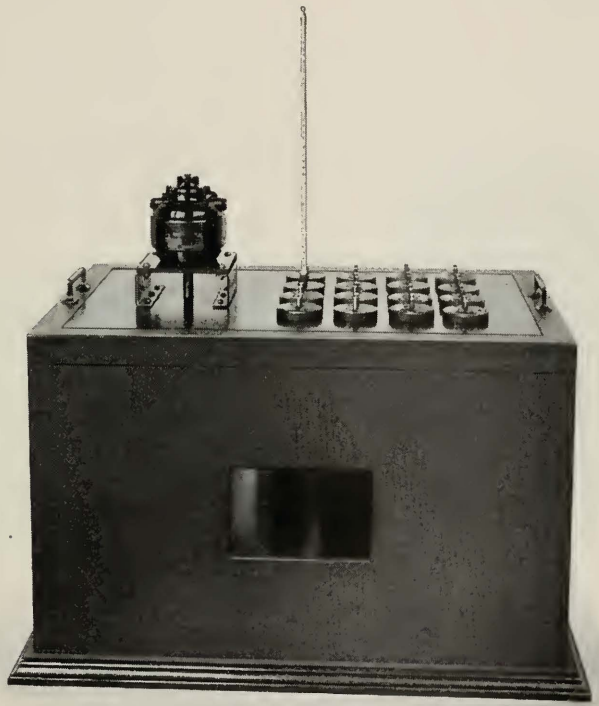


Fig. 6—Humidity cabinet for conditioning samples.

To avoid error from this source, all test samples are conditioned by drying in a desiccator at the approximate temperature of the test tank before being placed in the tank.

The samples are prepared by winding a number of turns of the textile around the electrodes inserted in the stoppers as shown in Fig. 8. Care is taken not to handle the textile itself during the winding process as perspiration from the hands is likely to contaminate the thread. Samples are left in the tank over night as there is considerable

evidence to show that several hours are required for them to come to complete equilibrium. A temperature of 100 degrees Fahrenheit and relative humidity of 75 per cent has been found suitable for cotton testing and 100 degrees Fahrenheit and 87 per cent relative humidity for silk.

The number of turns of yarn or thread wound around the electrodes will vary with the size of the thread since the winding space is fixed and a single layer of thread is applied. For No. 30/2 cotton approximately 90 turns, 180 parallel threads, have been found to give satisfactory readings. This same space accommodates about 256 turns, 512 parallel threads, of No. 62/1 spun silk.

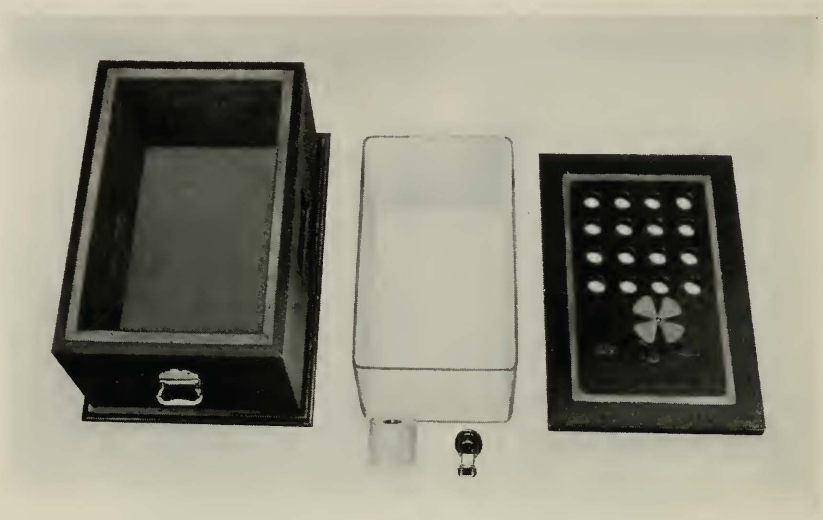


Fig. 7—Humidity cabinet disassembled.

The distance between the electrodes is not particularly critical. That is, it is not important, for example, whether the distance is $\frac{5}{8}$ in. or $\frac{3}{4}$ in. It is important, however, that having decided upon a certain separation, say $\frac{3}{4}$ in., this separation be accurately maintained for all electrodes if the readings are to be comparative. Of course if the separation is too great, an unreasonable number of turns of textile is required to bring the resistance of the sample within the range of the galvanometer. On the other hand, if the separation is too small, the error due to variation in separation for different sets of electrodes increases in magnitude.

In actual practise, $\frac{3}{4}$ in. separation with a winding space of 2 in. accommodating, as mentioned above, about 90 turns of No. 30/2

cotton has been found to be fairly satisfactory. This arrangement gives galvanometer readings of the order of 2000 megohms for washed silk and 1,000 megohms for washed cotton as compared with 12 meghoms and 5 megohms for unwashed silk and cotton respectively. It is obviously necessary to maintain a high degree of insulation resistance between the electrodes. This is accomplished by using hard rubber for the stoppers in which they are mounted and preventing surface leakage by coating the end of the stoppers with ozokerite wax.

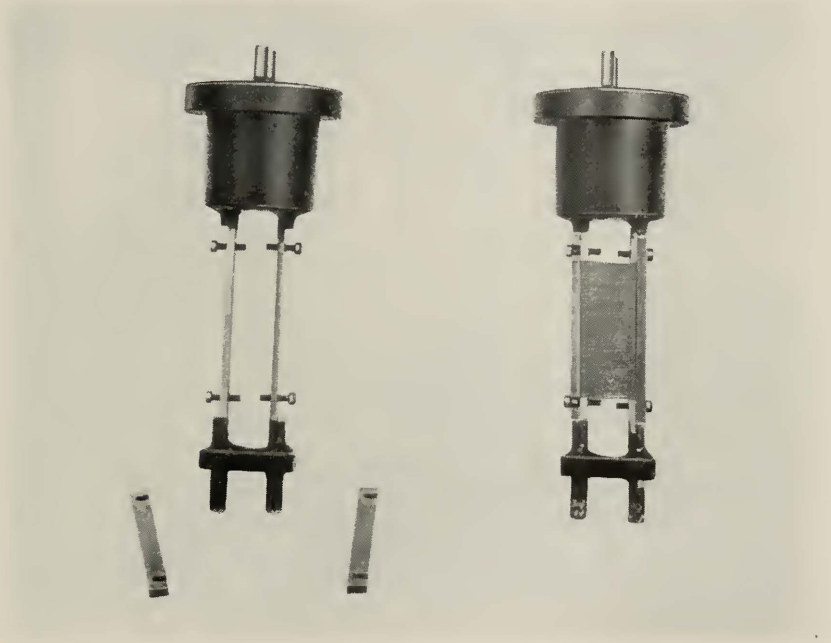


Fig. 8—Electrodes on which samples are wound for test.

The electrodes themselves are gold or platinum plated to prevent oxidation or corrosion. Observing these precautions, it is possible to obtain readings sufficiently consistent to distinguish not only between washed and unwashed textiles but to determine differences in degree of purification in various lots of washed textiles.

The question may be asked as to why 75 per cent relative humidity and 100 degrees Fahrenheit were selected for cotton and 87 per cent relative humidity and 100 degrees Fahrenheit for silk. These values were, within reasonable limitations, more or less arbitrarily selected and further experience may show that some other values are preferable.

However, for the following reasons, 75 per cent and 87 per cent at 100 degrees Fahrenheit were selected as offering definite promise of giving consistent results.

The main considerations in the choice of humidity conditions were, first, that the humidity should be high enough so that insulation resistance measurements of sufficient accuracy could be made using a band of threads as described above and a commercial galvanometer of reasonably high sensitivity; second, that the humidity be lower than that at which polarization effects would introduce serious error. The humidities chosen are within the range found suitable for cotton and silk under these limitations. Furthermore, these conditions are readily obtainable by the use either of saturated salt solutions or sulphuric acid solutions, thereby increasing the flexibility of the test. The temperature of 100 degrees Fahrenheit was chosen arbitrarily as one which could be maintained in the shop at any time of the year without artificial cooling.

Application to Apparatus

From an economic standpoint the most important conclusion to be drawn from the graphs is that cotton can be improved by washing to such an extent that it becomes a better insulator than the ordinary commercial insulating silk in general use. Since the cost of washing silk and cotton is nominal, usually less than 5 per cent of the cost of the material, the engineer given purified textiles may either take advantage of marked improvement in quality of electrical characteristics by using washed silk, or may substitute washed cotton for silk and realize substantial economies without degrading the product. As an example of how this applies to Bell System apparatus, central office distributing frame wire with annual requirements of more than 400 million conductor feet is now insulated with two coverings of silk where three were formerly required. The resultant wire is superior electrically to the old wire and the annual saving in silk amounts to about 70,000 pounds.

As another example, telephone cords of various types have been reduced substantially in cost with no impairment in quality by substituting two washed cotton braids for the cotton and silk braids formerly used. Altogether, various types of textile insulated wire aggregating annual requirements in excess of two billion conductor feet have either been changed to employ washed textile insulation or are scheduled for change as soon as possible because of corresponding economies in manufacturing cost or improvement in electrical properties.

The foregoing is intended to show what has been accomplished on a commercial scale at reasonable cost in the way of improving the insulating properties of silk and cotton. There still exists a rather wide margin in insulating properties between washed silk and washed cotton at high humidities which further study may show can be reduced. The graphs do not show the magnitude of improvement in cotton which has been obtained occasionally in laboratory experiments which leads us to hope that presently it may be possible to process cotton in a way that will result in its having electrical properties equal to those of washed silk for many practical purposes.

The question naturally arises as to the permanence of the improvement effected by the purification process. We have attempted to answer this question by periodic tests of washed silk and cotton insulated wire over an extended time, the test samples being exposed to ordinary room conditions where they could accumulate the normal quantity of dust. The results show no tendency for the insulation to revert to the constants of unwashed insulation. This appears logical since there is no particular reason to expect contamination by accumulation of such impurities as sodium or potassium salts from ordinary exposure to the air. Furthermore, in service, telephone office wiring is protected from the effects of dust by braided textile coverings or by the application of waxes or varnishes where the individual wires are exposed.

Conclusion

The discussion has been confined primarily to telephone central office cabling where silk and cotton are used in the cable core without impregnation. However, it is believed that the whole subject of purification of textiles becomes of general interest when it is stated that the improvements obtained by washing are not nullified by the supplementary use of impregnating waxes or varnishes. That is, the improvement in dielectric properties and reduced electrolysis obtained by washing and by impregnating are apparently substantially additive. While the studies have not proceeded far enough to cover comprehensively all of the better known impregnating waxes, asphalts, varnishes, etc., they have proceeded to the point where we can say that this is the case for the beeswax-paraffine waxes and certain asphaltic compounds. These findings are in line with the generally known fact that impregnation of textiles with wax compounds does not prevent, though it does retard, the absorption of moisture which in the presence of soluble salts causes conducting paths to be established, probably through the embedded textile fibers. Consequently,

such materials as fabric base insulating tapes, varnished linens and cambrics, electromagnet coil winding insulation, all being sensitive electrically to moisture, should be benefited to a substantial degree by purification of the fibrous components.

Therefore, while there is still much to be learned about the behavior of silk and cotton with respect to their electrical characteristics under various treatments and conditions, the study has progressed to the point where the following statements can be made.

1. The removal of water soluble salts which are present in both silk and cotton not only results in a very decided improvement in their insulating properties, but reduces the sensitivity to change of the a.c. characteristics with changes in atmospheric moisture conditions.

2. The improvement which can be realized is great enough to permit the substitution of washed cotton for silk where ordinary commercial silk has been found to give satisfactory results.

3. The use of purified textiles in cables carrying continuous d.c. potential will reduce electrolysis and consequently prolong the useful life of such cables about in proportion to the extent to which the purification process is carried.

In presenting the foregoing discussion, the authors wish to acknowledge their indebtedness to engineers of the Western Electric Company whose work in cooperation with silk suppliers has been largely responsible for the development of commercial methods of purifying insulating silk. Acknowledgment must also be made of the importance of the fundamental and research work which underlies the engineering result briefly described by this paper.

Telephone Apparatus Springs[1]

A Review of the Principal Types and the Properties Desired of These Springs

By J. R. TOWNSEND

This article describes the types of springs employed in telephone apparatus and enumerates the engineering requirements both from the standpoint of mechanics and the quality of materials desired. The chemical and physical requirements of the spring materials are given. The importance of fatigue is emphasized and the endurance limit is given for spring brass, nickel silver and phosphor bronze.

THE proper functioning of telephone apparatus springs depends upon careful design and selection of material. In many instances the springs must occupy small space and maintain delicate adjustment throughout the life of the apparatus with a minimum of attention. Whereas the physical size of these springs is small, the forces are sometimes necessarily large due to space limitations and this requires careful choice of material. It is believed that the use of these small springs is not unique to the telephone art and a discussion of materials and methods of test used should have broad interest and application.

There are three general classes of springs employed in telephone apparatus. For the purposes of this discussion the springs may be classified as follows:

Springs of sheet non-ferrous metal
Springs of clock spring steel
Helical springs.

Sheet Non-Ferrous Springs

Sheet non-ferrous springs usually consist of punched and formed parts made from brass, nickel silver, or phosphor bronze. These serve as electrical contact carrying members that are deflected or operated either electromagnetically or mechanically. Such springs are essentially cantilever springs clamped at one end and bearing near the other end one or more precious-metal contacts that are spot welded in place. The precious-metal contacts are employed to reduce contact resistance and the destructive effects of arcing when circuits are broken. The apparatus employing such springs are keys

[1] Presented at the Annual Meeting, New York, N. Y., Dec. 3 to 7, 1928, of The American Society of Mechanical Engineers, 29 West 39th Street, New York, N. Y.

of the switchboard type, relays, jacks, and interrupters. Certain apparatus known as switches employ these springs as brushes or wiping members. Here precious-metal contacts are not usually employed, but the ends of the springs must serve as contacting and wearing parts. Other springs are employed statically or, in other words, are required to maintain constant pressure for long periods without interruption. All springs that are attached to, or form part of, electrical circuits are soldered to connecting wires. This is usually done by soldering the wire or connection to a lug or projection that forms part of the spring. These lugs in most instances are on the opposite end of the clamped area from the operated end. The springs are almost always clamped between strips of phenol fiber because of its good insulating properties, mechanical strength, and permanency of form. A typical example of the use of the more common type of these springs is illustrated by Fig. 1 showing the familiar switchboard key.

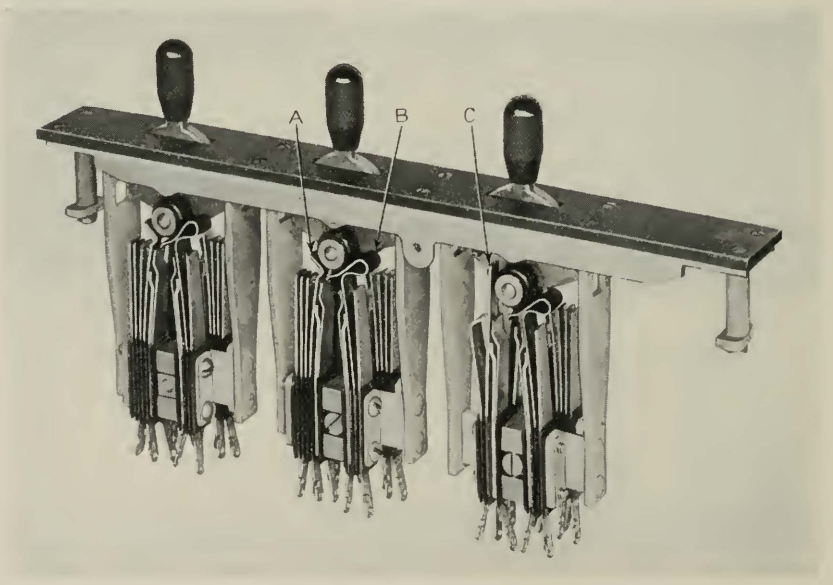


Fig. 1—Common type of switchboard key illustrating use of sheet-metal springs: *A*, plunger spring with crimp; *B*, crook spring; *C*, straight plunger spring.

The properties required of these small springs which are numerous and vary for each type of application are summarized as follows:

The proportional limit must not be too high to prevent the spring's being adjusted by flexing it with a tool to the point where it takes a set and occupies a position where it provides the desired operating

pressure. In other words, there must be room enough to flex the spring to the point where it will take a set within the space provided. A spring of high proportional limit such as one of clock-spring steel may be bent nearly double without being permanently deformed. Obviously, such a spring could not be adjusted in the key shown by Fig. 1.

The modulus of elasticity should be within the range of 12,000,000 to 20,000,000 lb. per sq. in. in order that the load-deflection rate will not be too steep to permit reasonable ease of adjustment by hand.

The endurance limit must be high enough to permit satisfactory operation throughout the life of the apparatus. In cases where space limitation is a factor, reference must be made to the stress-cycle graph made for the material to determine if the spring may be expected to stand up in service.

Creep or deformation under sustained load must not take place since the material will lose tension. Brass, nickel silver, and phosphor bronze may be expected on the basis of years of experience to hold adjustments when stressed up to approximately their proportional limits. Other materials when considered for telephone apparatus springs are investigated to determine their "creep" characteristics.

Season cracking takes place with highly stressed brass under severe atmospheric conditions, and for this reason springs that are required to hold their pressures without being deflected for long periods are not made from this material. Nickel silver will also season crack under still higher stress, and phosphor bronze least of all. In designing these springs, generous fillets and easy curves are employed to prevent the building up of localized high stresses that may lead to season cracking under sustained load and fatigue failure under repeated flexure.

When these springs are used as wipers in electrical circuits where arcing can occur, brass and nickel silver are not employed for the reason that the heat of the arc breaks down the material, volatilizes the zinc, and disintegrates the metal. For this reason, phosphor bronze, which does not contain zinc and has superior wear resistance to brass and nickel silver, is employed.

In addition to the foregoing properties, these springs must be resistant to atmospheric corrosion and capable of being readily soldered with soft solder. Nickel silver, as may be seen from Table 2, is superior to brass in its mechanical properties and in addition may be readily spot welded. As a result of years of experience it has been found that springs made of this material are capable of maintaining adjustment in a satisfactory manner in service. Phosphor

bronze has still greater wear resistance than brass or nickel silver and superior spring properties.

The chemical compositions of brass, nickel silver, and phosphor bronze spring materials used for telephone apparatus are shown in Table 1 and the mechanical properties are given in Table 2. The range of tensile strength and hardness shown comprises the specification limits. A more detailed description of the properties of sheet brass has been given elsewhere.[2]

TABLE 1

CHEMICAL COMPOSITION OF NON-FERROUS SHEET SPRING MATERIALS

	Min., per cent	Max., per cent
Brass		
Copper	64.5	67.5
Lead	0.0	0.3
Iron	0.0	0.05
Zinc	Remainder	
Nickel Silver		
Copper	53.50	56.50
Nickel	16.50	19.50
Zinc	25.50	28.50
Iron	...	0.35
Phosphor Bronze		
Copper	91.0	...
Tin	7.50	8.50
Phosphorus	0.05	0.25
Iron	...	0.10
Lead	...	0.02
Antimony	...	0.01
Zinc	...	0.20

TABLE 2

MECHANICAL PROPERTIES OF NON-FERROUS SHEET SPRING MATERIALS

Brass

Thickness	Temper	B. & S. gage nos. hard	Tensile strength		Rockwell hardness "B" scale		Proportional limit	Modulus elasticity
			Min.	Max.	Min.	Max.		
0.40 in. and thicker....	Hard	4	68,000	78,000	78	85	30,000	14×10^6
Below 0.40 in.	Hard	4	68,000	78,000	75	83		
0.040 in. and thicker....	Spring	8	86,000	95,000	88	92	30,000	14×10^6
Below 0.040 in.....	Spring	8	86,000	95,000	85	89		
0.040 in. and thicker....	Extra spring	10	89,500	98,500	89	93		
Below 0.040 in.....	Extra spring	10	89,500	98,500	86	90		

[2] "Physical Properties and Methods of Tests for Sheet Brass," by H. N. Van Deusen, L. I. Shaw, and C. H. Davis. Proceedings A.S.T.M., 1927.

Nickel Silver

0.40 in. and thicker....	Hard	4	92,000	106,500	69	77	60,000	20×10^6
Below 0.40 in.	Hard	4	92,000	106,500	66	75		
0.040 in. and thicker....	Extra hard	6	102,000	115,000	75	82	60,000	20.0×10^6
Below 0.040 in.....	Extra hard	6	102,000	115,000	72	79		
0.040 in. and thicker....	Spring	8	108,000	120,000	78	84		
Below 0.040 in.....	Spring	8	108,000	120,000	75	81		
0.040 in. and thicker....	Extra spring	10	111,000	123,000	80	85		
Below 0.040 in.....	Extra spring	10	111,000	123,000	77	82		

Phosphor Bronze

0.040 in. and thicker....	Extra hard	6	97,000	111,500	76	82	55,000	15×10^6
Below 0.040 in.....	Extra hard	6	97,000	111,500	73	79		
0.040 in. and thicker....	Spring	8	105,000	118,500	79	85		
Below 0.040 in.....	Spring	8	105,000	118,500	76	82		
0.040 in. and thicker....	Extra spring	10	109,500	122,000	81	86		
Below 0.040 in.....	Extra spring	10	109,500	122,000	78	83		

Note: The proportional limit and modulus figures are representative. They are only slightly changed by cold work.

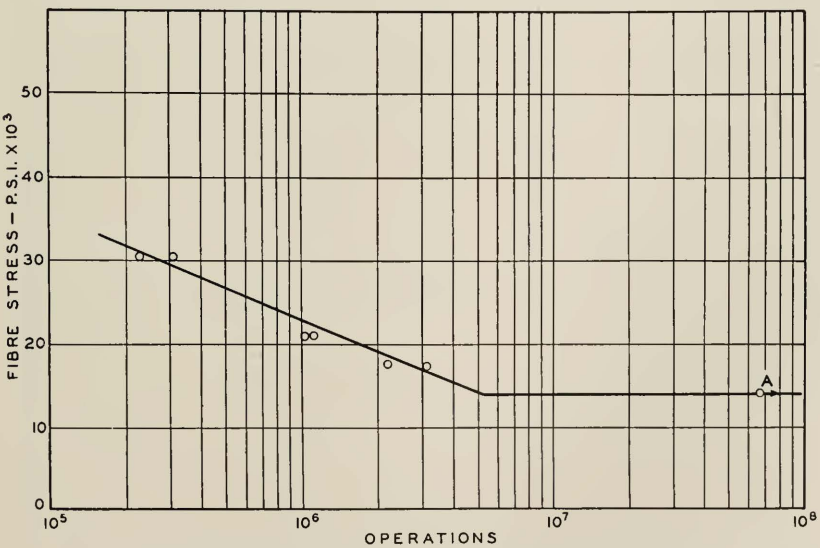


Fig. 2—Fiber stress vs. fatigue—brass sheet—24 gage—4 Nos. hard: *A*, average of three specimens taken from test without failure.

Fatigue Properties of Sheet Non-Ferrous Spring Materials

Because of the important bearing of fatigue of springs for telephone apparatus, this property has been carefully studied. The stress-

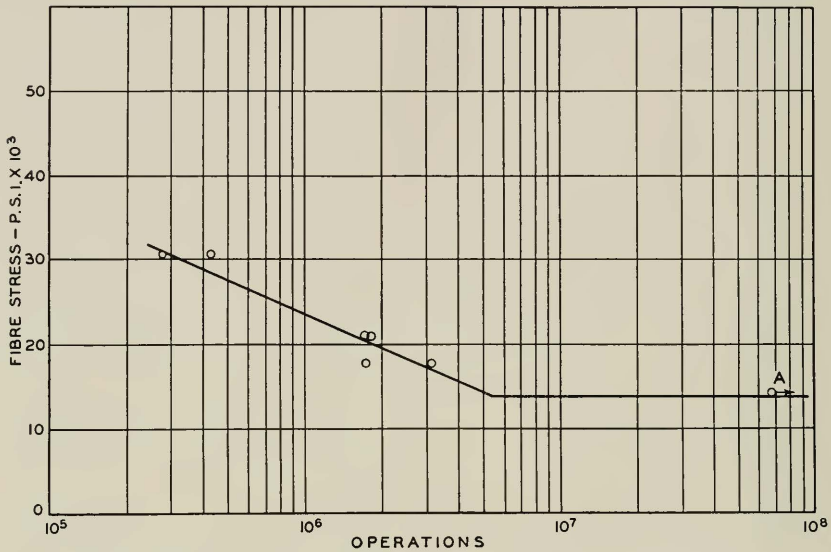


Fig. 3—Fiber stress vs. fatigue—brass sheet—24 gage—10 Nos. hard: *A*, average of three specimens taken from test without failure.

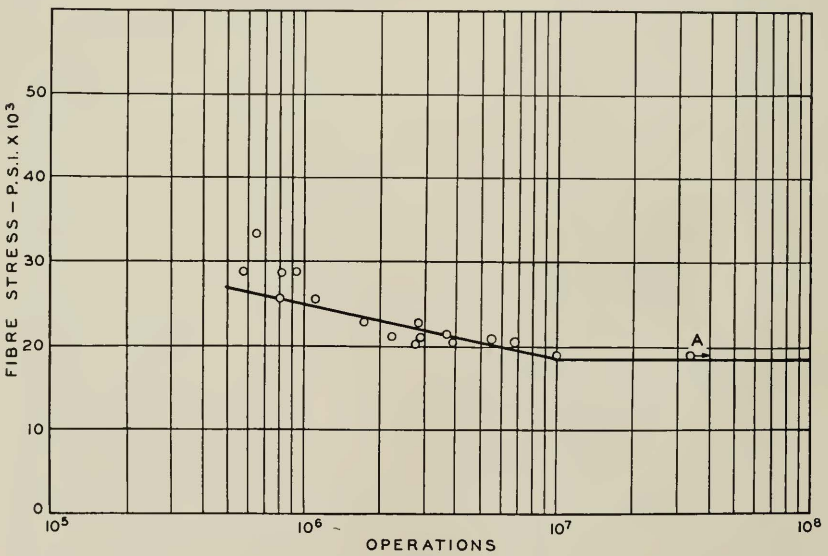


Fig. 4—Fiber stress vs. fatigue—nickel silver sheet—24 gage—4 Nos. hard: *A*, average of three specimens taken from test without failure.

cycle graphs shown by Figs. 2 to 6 give typical results. The design of specimen is shown by Fig. 7. These specimens were alternately stressed at the rate of 700 cycles per minute.

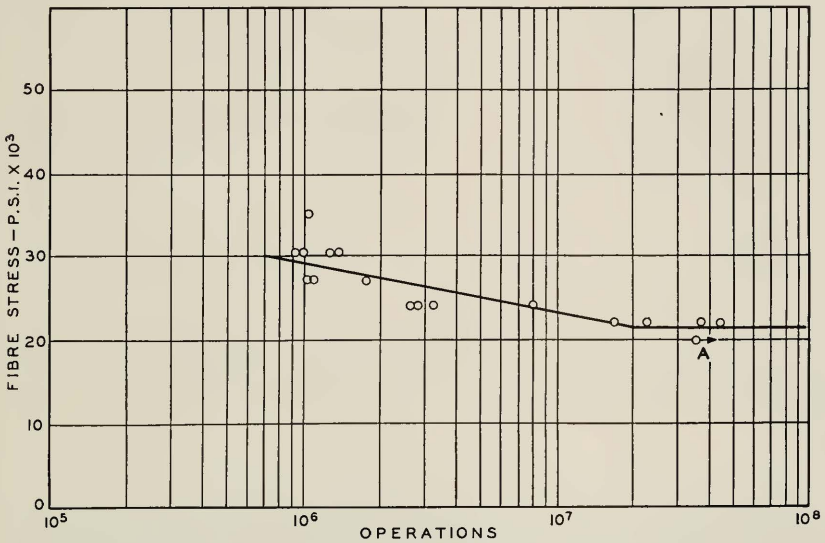


Fig. 5—Fiber stress vs. fatigue—nickel silver sheet—24 gage—10 Nos. hard: *A*, average of three specimens taken from test without failure.

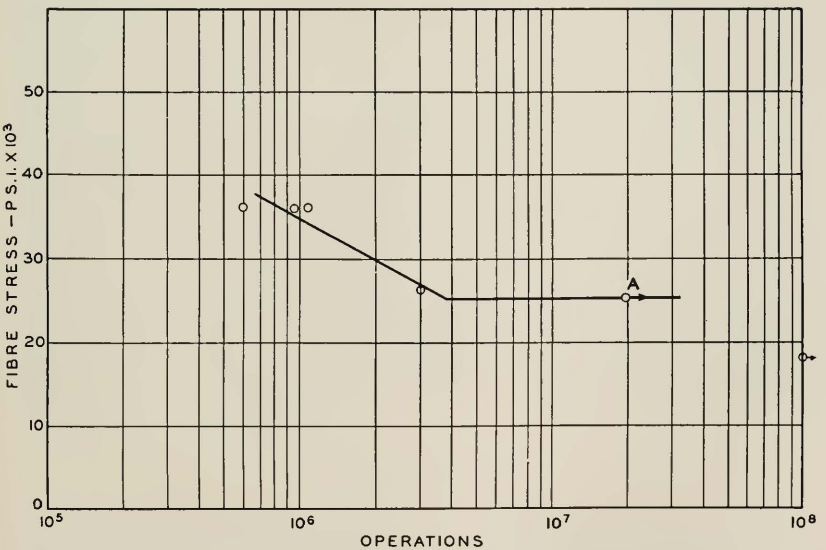


Fig. 6—Fiber stress vs. fatigue—phosphor bronze sheet—24 gage—10 Nos. hard: *A*, average of two specimens taken from test without failure.

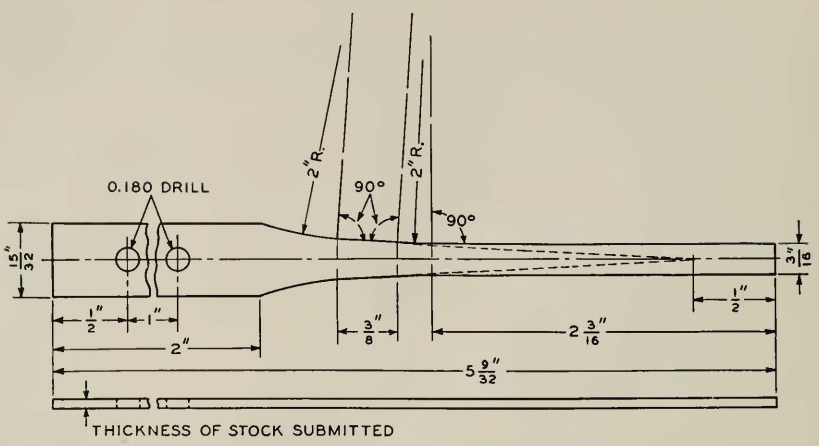


Fig. 7—Test specimen for fatigue study.

(The piece must be free from scriber and tool marks. The 2 in. radius fillets must join the tapered portion without an abrupt break. The intersection of the tapered sides produced must be ½ in. from the end.)

Clock-Spring-Steel Springs

These springs are used as vibrating elements in interrupters, where a high fatigue endurance is required, and for springs that must be worked at high pressures and rapid buildup. The material has the chemical composition shown by Table 3 and mechanical properties given by Table 4. This is a carbon-steel, heat treated and then cold rolled, and is by nature brittle. To guard against excessive brittleness and at the same time provide a high strength cold rolled steel, a bend test has been developed.

TABLE 3

Chemical Composition of Clock Spring Steel

	Min., per cent	Max., per cent
Carbon	0.85	1.15
Manganese	0.25	0.60
Silicon	..	0.22
Sulphur	..	0.025
Phosphorus	..	0.03
Nickel, chromium, tungsten	..	0.10
Vanadium	Optional	0.25
Iron	Remainder	

TABLE 4

Mechanical Properties of Clock Spring Steel

Ultimate strength	250,000 to 290,000 lb. per sq. in.
Proportional limit	160,000 to 215,000 lb. per sq. in.
Modulus of elasticity	27.6×10^6

The bend test requires that when the material is bent back parallel to itself to form a "U" and further compressed between flat parallel surfaces (for example, between the jaws of a vise), at a rate not to exceed 0.3 inches per minute, the material shall break along a straight line making approximately a 90 degree angle with the axis of the strip when the distance between the inside of the legs of the "U" is 25 to 16 times the thickness of the material. It must not break before the distance is reduced to 25 times the thickness of the metal. This test can be conveniently applied by drawing the looped material through two of a series of graduated slots.

Music Wire Springs

Tinned and plated music wire is extensively used for compression springs in telephone apparatus. Here the spring is in the form of an open helix. Fig. 8 shows a type widely used. These springs are either tinned or plated with nickel. It has been observed that the plating baths adversely affect the fatigue characteristics of music wire and this is under investigation at the present time.

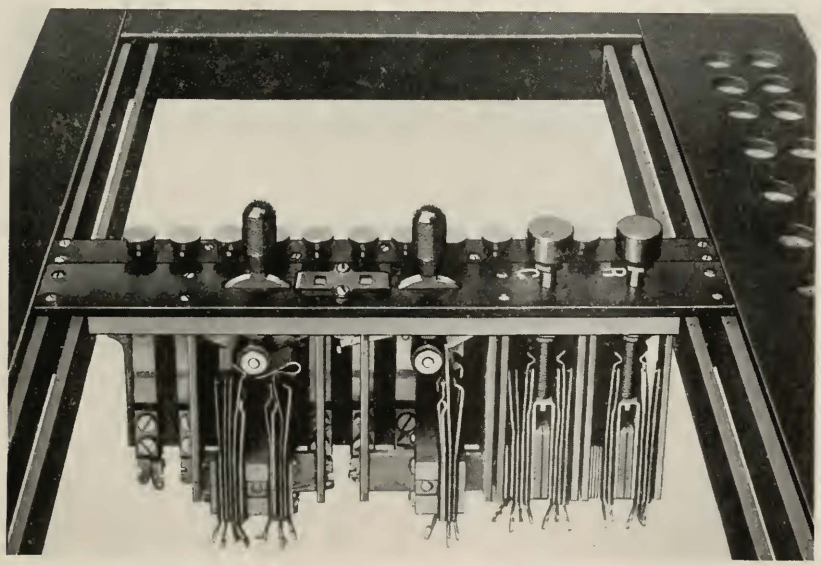


Fig. 8—Switchboard keys employing sheet metal springs and helical compression springs

The average tensile properties of the music wire employed are: proportional limit—217,000 lb. per sq. in.; ultimate tensile strength —350,000 lb. per sq. in. No chemical or tensile requirement is

placed on this wire, the main feature of concern being brittleness. The elongation measured in an 8 in. length must be between 1 and 4 per cent for wire up to and including No. 27 gage and $1\frac{1}{2}$ to 7 per cent for No. 28 gage wire and heavier. The wire is also subject to a kinking test in which No. 30 wire and smaller shall kink without breaking. This last test indirectly controls the tensile properties of the wire.

Summary

A review has been given of the more common types of telephone apparatus springs. Vast quantities of these springs are employed in the telephone plant and the numerous factors that must be considered with regard to selecting material have been reviewed. The materials most generally used have abundant competitive sources of supply and the quality is carefully controlled by tests that are designed to be easy to apply and effective in evaluating the properties desired.

Effect of Signal Distortion on Morse Telegraph Transmission Quality

By J. HERMAN

In applying telegraph transmission measuring apparatus to the development and maintenance of telegraph circuits, it is desirable to correlate quantitative measurements of telegraph signal distortion with quality of telegraph transmission. Accordingly, a series of tests has been carried out in order to determine this relationship for the case of manual operation using the American Morse Code. These tests are described and the results, together with the conclusions reached, are given in summarized form.

I. Introduction

WHENEVER telegraph signals are transmitted over a circuit they become more or less distorted depending upon the type of circuit, the adjustment of the apparatus and the speed of transmission. An adequate knowledge of the relationship between the possible types of distortion and the satisfactoriness of the telegraph services rendered over various circuits is evidently of considerable importance. This is true both in the design of telegraph circuits to insure the necessary quality of transmission and in the operation and maintenance of these circuits to insure that they are giving the service for which they are designed.

Considerable data both qualitative and quantitative, bearing on this matter have been collected in the past in connection with development work and as a result of operating experience in the Bell System. Recently some tests were made to correlate quality of telegraph transmission with quantitative measurements of signal distortion on manual telegraph circuits employing the American Morse code. This paper presents and discusses the results of these tests.

Commercial telegraph operation over land lines in the United States, is carried on almost exclusively by two well known methods, manual Morse and printing telegraph. In the first method, the signals are sent by hand in accordance with the Morse code and received by ear by listening to the clicking of a sounder. In the second method, the signals are sent mechanically under the control of a typewriter keyboard and received so as to cause the selection and printing of the proper character by mechanical means. Although the question of permissible distortion in transmission is of importance for both methods, it is to be expected that the answer will be considerably more involved for the manual Morse case since the human element is so large a factor in this case.

II. Description of Tests

1. *Preliminary Considerations*

Telegraph transmission quality may be considered perfect when the received signals at one end of the circuit correspond exactly to the signals as sent at the other end of the circuit. Any departure in the length of a signal or part of a signal at the receiving end is, therefore, a quantitative measure of the degradation of the quality of transmission and has been termed telegraph distortion. Since telegraph signals are composed of dots, dashes, and spaces, the measurement of degradation in transmission quality consists of measuring the lengths of these signal elements at the receiving end of the circuit and comparing them with their lengths at the sending end of the circuit. Means for doing this, together with a discussion of the various types of distortion affecting the lengths of telegraph signals, have been given in a recent paper.[1]

As brought out in the paper referred to, distortion of telegraph signals may be divided into three components, namely, bias, characteristic distortion and fortuitous distortion. Each of these may be either positive or negative, depending upon whether the distortion causes a lengthening or a shortening of the signal part under consideration. The three components may be described briefly as follows:

Bias consists of a substantially uniform lengthening of the marks and a corresponding shortening of the spaces of telegraph signals, or vice versa. The first condition is called positive (marking) bias and the reverse condition negative (spacing) bias. It is usually due to lack of symmetry in the marking and spacing battery voltages or in the adjustment of repeating relays.

Characteristic distortion is distinguished from other types of distortion by the fact that it is a function of the signal combination as well as of the electrical and mechanical characteristics of the circuit. For example large inductance in a circuit may prevent the signaling current from building up to its full value during a short impulse following a long impulse of opposite polarity, thereby causing a decrease in the length of the short impulse which would be called negative characteristic distortion.

Fortuitous distortion is an erratic lengthening and shortening of marks and spaces such as that due to the superposition of extraneous interfering currents upon the signaling currents in the line, or due to chattering and sparking at the contacts of repeating relays.

In the case of machine telegraphy such as printing telegraph, the

[1] Nyquist, Shanck and Cory, *Trans. A. I. E. E.*, Vol. XLVI, 1927, p. 231–240.

change in length in the signal elements which will cause errors in reception can generally be determined from the construction of the machine. It has a fairly definite value for a particular type of machine and is largely independent of the type of distortion which produces it.

In the case of Morse telegraphy where the signals are received by listening to the clicks of a sounder, the effect of distortion is more complicated. Morse operators do not interpret signals entirely by the length of individual signal elements as does a telegraph printer. They interpret them by the general sound of the clicks due to the succession of dots and dashes making up a particular letter or word of the code. Consequently, it is to be expected that for the same total change in length of signal elements caused by one or more of the three types of distortion, the effect upon operators will differ, depending upon the type of distortion and upon the particular operators who are receiving. It is this phase of Morse telegraph operation with which the present paper deals.

To carry out the investigation, namely, to obtain data on the effects of various kinds and combinations of distortion on the accuracy of reception by telegraphers and their opinions as to the satisfactoriness of the received signals, several methods of procedure were suggested. These were carefully considered and the following appeared to be the most suitable.

A circuit, simulating a commercial telegraph circuit, was to be constructed and arrangements devised for impressing any desired amount of distortion or combination of distortion upon the signals transmitted over this circuit. The manner of introducing the distortions was to simulate as nearly as practicable the manner in which they would occur on real lines and the methods employed for sending and receiving the signals were to simulate closely actual operating conditions.

As regards the type of test messages to be used by the operators, it was thought that neither plain English nor unpronounceable code would be satisfactory, the former because distorted letters and words could be supplied by the operators from the context and the latter because it would be unnecessarily difficult to send and receive. Consequently, messages intermediate between these extremes were decided upon and were obtained by using text from a foreign language with which the operators were unfamiliar.

For sending the messages, a semi-automatic key (vibroplex), carefully adjusted to be unbiased and to operate at a speed of 13.5 d.p.s. (dots per second) was provided. Although this speed of operation corresponds to fairly rapid Morse sending, the rate of transmission of words during the tests was fairly low, being about 25 five-letter words

per minute. This resulted from the unintelligible nature of the text which required a wider and more careful spacing of letters and words than would have been the case with a more intelligible text.

For measuring telegraph transmission, the methods outlined in the paper referred to previously were used. A distortion measuring device which measured total distortion and the components separately, and which employed test signals known from past experience to be representative and suitable for obtaining good data on telegraph transmission, was provided. A speed of transmission of 15 d.p.s. was chosen for the test signals in order that the data obtained could be directly compared with distortions as measured in development work and in field transmission measurements. Consequently, the results of the tests could be made of immediate practical utility.

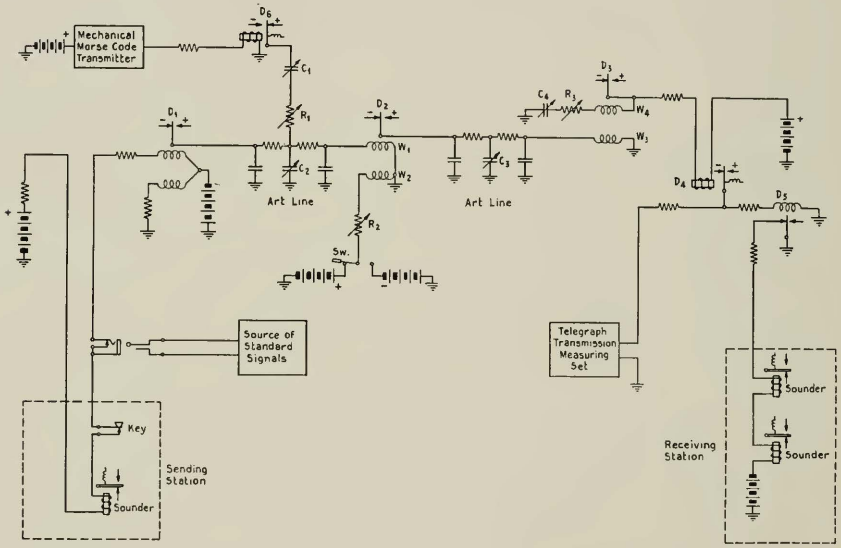


Fig. 1—Diagram of Test Circuit.

2. *Test Circuit*

A diagram of the test circuit is shown in Fig. 1. The telegraph circuit consisted of two artificial lines connected by a repeating relay and arranged to transmit in one direction only.

The sending operator transmitted into a local circuit at the point marked "sending station." This local circuit could also be connected to the source of standard signals which consisted of a distributor operated at a constant speed by means of a "phonic-wheel" motor.

Signals sent into the local circuit operated a sounder, a polar trans-

mitting relay D_1 and passed over the first section of artificial line to the polar receiving relay D_2. From D_2 the signals were repeated directly into the second section of artificial line and operated the polar receiving relay D_3. The latter relay repeated the signals into another local circuit where they were received by the neutral relay D_4. The neutral relay repeated the signals in polar form to the transmission measuring set and also to the polar relay D_5. The latter relay operated two neutral sounders which were located near the two receiving operators.

3. *Method of Introducing Distortion*

Bias was introduced into the circuit by passing a direct current through winding W_2 of relay D_2. The direction of this current could be reversed by means of the switch SW which was arranged to connect either a positive or a negative battery to the relay winding. Consequently, either a positive or a negative bias could be introduced, the amount being controlled by means of the variable resistance R_2. For amounts of bias greater than about 35 per cent the bias was introduced in two sections, by connecting winding W_4 of relay D_3 in series with the winding W_2 of relay D_2. This change was found necessary to prevent failure of the system whenever large amounts of bias were desired.

Negative characteristic distortion was introduced by means of the condensers C_2 and C_3. By increasing the value of capacity in these condensers, any desired amount of distortion up to about 70 per cent could be introduced into the circuit. The effect produced by the condensers is similar to that caused by the capacity to ground of a long cable circuit, or of intermediate composite sets and similar apparatus having condensers connected from the line wires to ground.

Positive characteristic distortion was introduced by means of transient currents in a circuit into which was connected a separate winding of one of the receiving relays. The currents flowed through the winding in such a direction as to tend to hold the relay armature against the particular contact to which the armature had been operated. The circuit consisted of the relay winding W_4, the resistance R_3, and the condenser C_4, connected in series from the armature of relay D_3 to ground. By increasing the values of resistance and capacity to a sufficient extent the duration of the charging current could be made appreciably long. Consequently, a reversal of current in winding W_3, due to the telegraph signals was not able to operate the relay armature to the opposite contact until the condenser C_4 had become sufficiently charged so that the charging current, which flowed through winding W_4, was reduced to a value below that of the line current. The time constant of the holding circuit was such that the amount of charge on

the condenser C_4, at the beginning of a particular signal element, was determined largely by the length of the mark or space immediately preceding this element and to some extent by the ones preceding that. As a result, this mark or space was lengthened an amount depending upon its position in the signal combination, thereby simulating the effect of positive characteristic distortion. It was necessary to introduce this type of distortion in two sections when values of distortion greater than about 35 per cent were desired. This was accomplished by connecting winding W_2 of relay D_2 in a holding circuit similar to that of winding W_4 of relay D_3.

Fortuitous distortion was introduced by means of a telegraph crossfire arrangement. It consisted of the neutral relay D_6 whose windings were connected into a local circuit and operated by means of a mechanical Morse code transmitter. The contacts of the relay had positive and negative batteries connected to them and the armature was connected to the telegraph circuit by means of condenser C_1 and series resistance R_1. The magnitude of the maximum fortuitous distortion was determined by the magnitude of the crossfire impulses, this being controlled by means of the condenser and resistance. Since the combination of the crossfire impulses with the telegraph signals was erratic, it is obvious that positive and negative distortions ranging from zero up to the maximum value were obtained at one time or another.

The methods described above for distorting the signals simulated closely, except possibly in the case of positive characteristic distortion, the conditions which occur on actual telegraph circuits. For this reason, the results obtained in the tests may be taken as reasonably representative of results which would be obtained on actual circuits.

4. *Method of Making Operating Tests*

Two series of tests were made. The first series covered all types of distortion individually and in combinations while the second series covered only those individual types of distortion the effects of which, as shown by the first series of tests, appeared questionable and for which check tests were desired.

Testing was done on alternate days between the hours of 9 a.m. and 4 p.m. Six operators were available, two for sending and the remaining four for receiving. These were divided into two groups and appeared alternately for the tests.

The method of testing was briefly as follows: the sending operator was provided with a vibroplex key and with copies of messages to be transmitted. These messages were taken chiefly from Hungarian

books and were arranged in the form of sentences and paragraphs. The words were unintelligible to the operators, which prevented them from supplying distorted parts of the messages from the context. As an additional precaution, words which occurred frequently and might have been memorized were omitted or changed arbitrarily.

At the receiving station, which was located in a different room from the sending station, the two receiving operators were provided with separate sounders and typewriters. They copied each message simultaneously and wrote their own opinions of the condition of the circuit at the end of each message. Care was taken to prevent the two operators from exchanging opinions.

On each day of the tests certain preliminary adjustments were made. The vibroplex key was first adjusted to vibrate at 13.5 d.p.s. by "beating" in a local circuit with signals from a rotary interrupter. The interrupter was then set to give signals at 15 d.p.s. and the transmission measuring set adjusted. After these adjustments had been made, standard signals from the interrupter were transmitted over the telegraph circuit and the latter adjusted for zero bias and distortion by means of the measuring set. The sending operator was then asked to send a few sentences over the circuit and the receiving operators listened to these signals and adjusted their sounders to be unbiased.

There was some question as to whether this adjustment of sounders by the operators really produced unbiased operation of the sounders. Some tests were, therefore, made in which the sounders were adjusted so that they just failed to operate properly on signals containing the same amount of large positive or negative bias. The results obtained in this case were almost identical with those obtained when the operators adjusted the sounders.

Immediately before a test message was transmitted over the circuit, distortion was introduced. Rapid measurements of distortion were made on standard signals by means of the transmission measuring set to determine when approximately the desired amount of distortion had been introduced into the circuit. An accurate measurement was then made and recorded together with the number of the test and the number of the message to be transmitted. These numbers were also recorded on the received messages and served to identify the transmission measurements corresponding to the messages received by the operators.

The various received messages were analyzed both for accuracy of reception and nature of errors produced. In determining the accuracy of reception, the number of misinterpretations in each received message was subtracted from the total number of characters in the correct

copy of the message and the difference expressed as a percentage of the latter. The nature of the errors was determined by translating the errors into the corresponding Morse symbols and comparing them with the Morse symbols for the correct characters.

5. *Transmission Measurements*

The amount of distortion present in the telegraph signals after passing over the circuit was measured by means of a telegraph transmission measuring set, as mentioned previously. This is a device for measuring any departure in the length of signal elements from their normal value. The signals used for the measurements were standard signals, obtained from a mechanical interrupter. The latter consisted of a printing telegraph distributor, operated at a constant speed and arranged to send out a particular signal during each revolution of the distributor brush arm. The procedure for the measurements consisted in connecting the source of standard signals to the sending loop of the circuit, and measuring the distortion of the various elements of this standard signal at the receiving end of the circuit by means of the measuring set.

The types of standard signals used, consisted of reversals and signals similar to the Morse letters *C* and *E*. Bias was measured with all three signals, average characteristic distortion with the *C* and *E* signals only, and maximum distortion including fortuitous with the *C* signal only.

The speed of signaling for transmission measurements was 15 d.p.s. and the values of distortion given in the following discussion refer to this speed. Since the speed of transmission by operators was 13.5 d.p.s., it is evident that the distortion which the circuit impressed upon the operators' signals was somewhat less than the values given. The distortion at the lower speed may be obtained approximately, if desired, by assuming that the per cent distortions are directly proportional to the speed of transmission.

It should also be understood that, in addition to the distortion which is due to the condition of the circuit, an appreciable amount of distortion is introduced into the signals by the sending operator. The amount of such additional distortion is a variable quantity depending on the characteristic sending of a particular operator. Neither its magnitude nor effect was determined in the present investigation which was limited to the effect which the condition of a circuit had upon the reception of signals transmitted and received by good telegraph operators.

III. Discussion of Results

1. *General Effect of Distortion*

A. Effect Upon Accuracy of Reception.—The effect of distortion upon accuracy of reception, as shown by the results of the tests, differs in character for different types of distortion. In accordance with the effects produced, the various types of distortion may be divided into two general classes. The first class consists of types of distortion which produce only a small change in the accuracy of reception nearly up to the point where the circuit actually fails. The second class consists of types of distortion which produce a rapid decrease in accuracy of reception when the distortion is increased beyond a certain moderate value. Of the various types of distortion encountered on telegraph circuits, negative bias and fortuitous distortion fall into the first class while positive bias and positive and negative characteristic distortion fall into the second class.

Combinations of various types of distortion in which one type of distortion predominates appear generally to fall into the same class as the predominating type of distortion in the combination. There is, however, a marked tendency for many combinations of distortion to fall into the first class, especially, when the various distortions in a combination are about equal in magnitude.

The accuracy of reception for a particular circuit condition differed considerably with operators. For the case of zero distortion in a circuit, most of the operators consistently reached accuracies between 98 and 100 per cent while other operators failed to reach accuracies higher than 88 per cent. This difference is due partly to the fact that some of the receiving operators were more experienced than others and partly to poor sending by some of the sending operators. It is of interest, however, that the general shape of the distortion versus accuracy curve for a given condition is unaffected by this difference.

B. Effect Upon Opinion of Operators.—In general, the various operators were in fair agreement as to the point at which a circuit became unsatisfactory for commercial operation. This point corresponded closely with the decrease in accuracy of reception for some types of distortion but differed widely from it for other types of distortion. In those cases for which the opinion of operators disagreed with their accuracy of reception, the operators usually pronounced the circuit unsatisfactory at values of distortion considerably lower than those required to cause an appreciable decrease in their accuracy of reception. The reason given for condemning such a circuit at the low value of distortion was the peculiar sound of the signals. This, they

claimed, required an unusual amount of concentration on their part, causing fatigue and making it difficult to receive over the circuit for a long period of time.

The type of distortion for which the disagreement between opinion and accuracy of reception was most pronounced, is negative characteristic distortion. The operators considered 25 per cent of this type of distortion to be about the maximum allowable value for commercial operation, whereas about 50 per cent was required to cause an appreciable reduction in their accuracy of reception. These values were checked a number of times with different operators and appear to be well established. A similar condition occurred in the case of fortuitous distortion for which the corresponding values are 50 per cent and 85 per cent, respectively. In this case, however, the infrequent occurrence (about four times a minute) of the maximum value of this type of distortion probably accounts for the small effect on the accuracy of reception.

These results would seem to indicate that the accuracy of reception is not a complete criterion of the allowable amount of distortion in a telegraph circuit. It is undoubtedly true that operators who work over a circuit, often do not wait until they make errors before calling the repeater attendant to fix the circuit. They usually take such steps as soon as they notice any appreciable distortion or have any difficulty in receiving signals.

In explanation of this point it may be said that when an operator is receiving perfect signals at a speed below the maximum rate at which he can work, he has what may be termed a "margin of attention" which may be utilized in interpreting defective signals. Before the point is reached where his accuracy of reception is impaired, he experiences a reduction in the "margin of attention" and on that account may pronounce a circuit unsatisfactory, since operation with little or no "margin of attention" soon produces a mental strain. The reception of signals having frequent distortions of say 30 per cent may very likely decrease the "margin of attention" to a greater extent than the reception of signals having only occasional distortions of 60 per cent or even greater. Also, if the speed of transmission during the tests had been slower or faster, the "margin of attention" would have been greater or less, and the point at which the operators considered the circuit unsatisfactory would have been different.

Because of the above it is thought that in establishing allowable limits of distortion for commercial telegraph circuits, the opinion of operators must be given consideration in addition to the accuracy of reception.

2. *Effect of Bias*

A. Positive Bias.—The relationship between the amount of positive bias and the accuracy of reception by operators is given in Fig. 2, in which is plotted the average of the results obtained in three tests.

It will be seen from an inspection of this curve that the accuracy falls off rapidly between 40 per cent and 50 per cent bias in which region the majority of the operators also called the circuit unsatis-

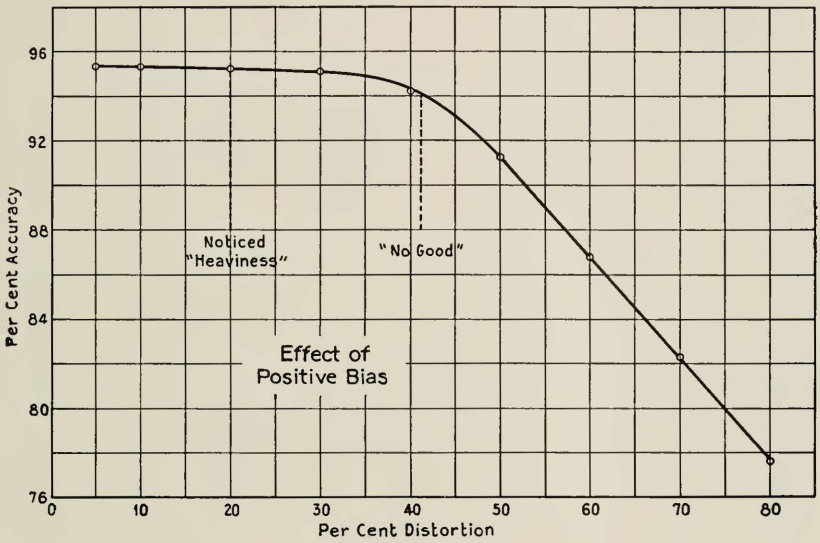


Fig. 2.

factory. The three most interesting facts, as depicted on this curve, are as follows: The accuracy of reception changed very little up to a bias of about 40 per cent but decreased rapidly above this value. The operators remarked upon the bias when it reached a value of about 20 per cent, and considered the circuit unsatisfactory with a bias of about 40 per cent.

An analysis of the errors made by the operators for values of bias from 35 per cent to 70 per cent showed that most of the errors were due to the omission of letters. This indicates that the effect of positive bias was such as to confuse the operators and cause them to miss characters while trying to interpret some peculiar sounding character which had gone before. There was very little indication of errors due to interpretations of dots as dashes even for a bias greater than 50 per cent. Most of the interpretations were of a miscellaneous nature, although there were a few errors which appeared to be due to a dropping out of spaces between signal elements.

B. Negative Bias.—The curve of Fig. 3 shows the effect of negative bias. As in the case of positive bias above, this curve is also an average of three test curves.

With this type of distortion the accuracy of reception is influenced only slightly with increase in distortion over wide limits. There is a gradual decrease in accuracy of reception up to about 65 per cent distortion and a more rapid decrease beyond this value.

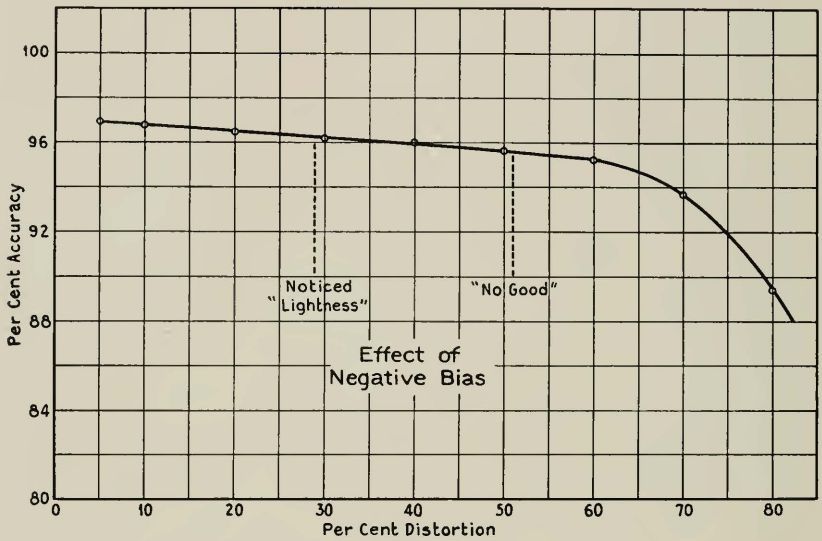


Fig. 3.

It was thought that possibly the difference in the effect of positive and negative bias was due to the adjustment of the sounders. If the operators had adjusted the sounders to be biased slightly heavy for the condition of zero bias over the circuit, then the sounders would have failed sooner on a circuit with positive bias than on one with an equal amount of negative bias. Some tests were, therefore, made for which the sounders were adjusted so as to fail to operate properly for equal values of large positive and negative bias in the circuit but the results obtained were the same. It may be concluded, therefore, that the effect is mainly of a psychological nature and is not due to a difference in the adjustment of the apparatus.

The noteworthy features in the results of tests with negative bias are as follows: The accuracy of reception decreased gradually with increase of distortion up to about 65 per cent, and fairly rapidly above this value. The operators remarked upon the bias when it reached a

value of about 30 per cent and considered the circuit unsatisfactory at about 50 per cent.

An analysis of the errors made by the operators for values of distortion from 30 per cent to 70 per cent indicated that a large number of errors were due to confusion, as in the case of positive bias. For large values of bias there were some errors which indicated an interpretation of dashes as dots, although most of the errors were miscellaneous interpretations, together with a large number of letters added and omitted.

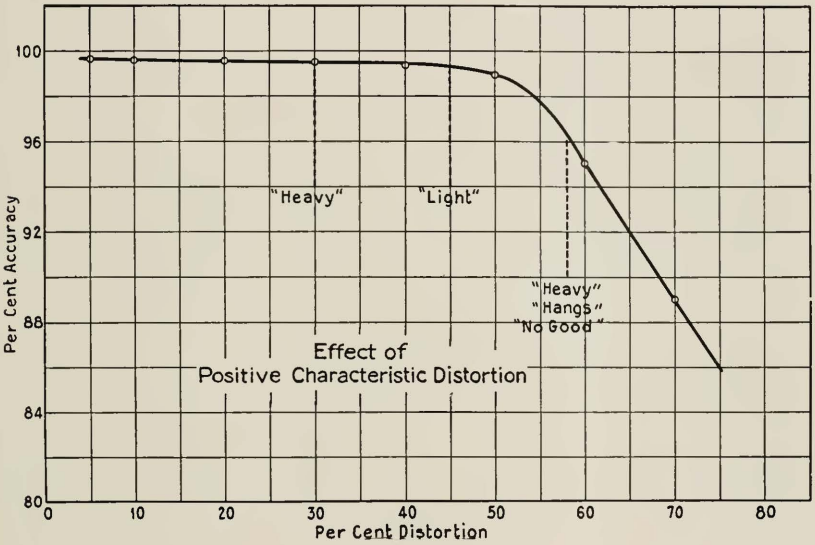


Fig. 4.

3. *Effect of Characteristic Distortion*

During the tests with positive and negative characteristic distortion, a peculiar effect of moderate amounts of such distortion on the opinion of operators was noticed. In general, operators tend to call a circuit "heavy," "light," or "unsteady." If the amount of characteristic distortion is not large enough to warrant the term "unsteady," some operators may call the circuit "light" while others may call it "heavy." In a few cases the operators stated that the bias changed constantly from one letter or word to another. The latter characterization is probably the more accurate and accounts to some extent for the difference of opinion. When the amount of distortion becomes large enough so that the operators consider a circuit "unsteady," they also call it "light" in most cases.

A. Positive Characteristic Distortion.—The curve of Fig. 4 shows the effect of positive characteristic distortion as determined by averaging the results of two test curves and a number of qualitative check tests. The effects of this type of distortion upon reception are briefly as follows: The accuracy of reception remained practically constant up to a distortion of about 50 per cent and decreased rapidly above this value. The operators remarked upon the distortion at a value of about 30 per cent and considered the circuit unsatisfactory at about 50 per cent.

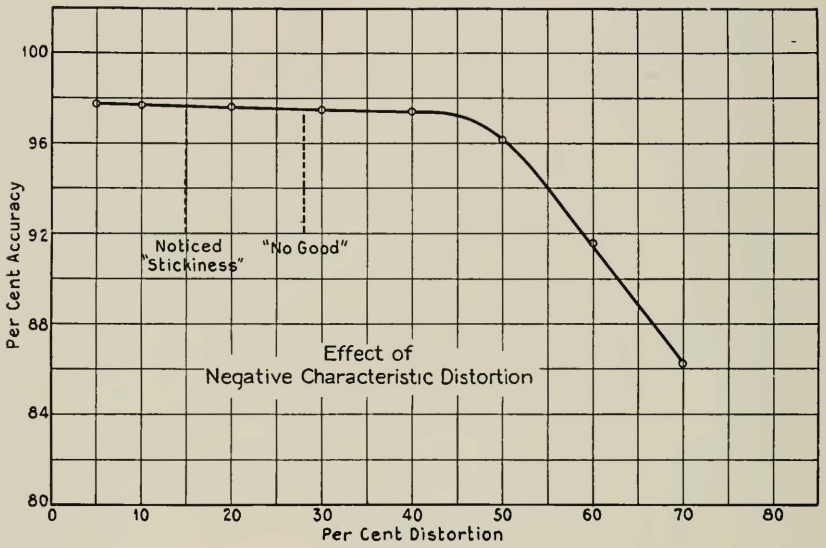


Fig. 5.

The errors produced by positive characteristic distortion were analyzed for values of distortion from 35 per cent to 65 per cent. The results obtained indicated that most of the errors for distortions greater than 50 per cent were due to a lengthening of dots at the beginning of letters and a dropping out of dots at the middle and at the end of letters. As a consequence a large number of errors were due to the following misinterpretations: H interpreted as B; S as G; N as T; J as K; C, A, J or F as M. In addition an appreciable number of errors were due to miscellaneous interpretations together with a large portion of letters added and omitted.

B. Negative Characteristic Distortion.—The effect of negative characteristic distortion is illustrated by the curve of Fig. 5 which is an average of three test curves and several qualitative check tests. This type of distortion is of considerable interest due to the fact that a

value of distortion lower than that with any other type or combination of distortion causes a circuit to be unsatisfactory, according to the opinion of the operators. Moreover, the opinion of the operators disagrees considerably with the effect which the distortion has upon their accuracy of reception, as shown by the curve.

A summary of the effect of negative characteristic distortion as given on the curve of Fig. 5 is as follows. The accuracy of reception remained nearly constant up to about 45 per cent distortion and then commenced to decrease rapidly. The operators remarked on the distortion at about 15 per cent and considered the circuit unsatisfactory at about 30 per cent.

The nature of the errors made by operators for this type of distortion was analyzed for values of distortion from 15 per cent to 70 per cent. Consistent misinterpretations occurred for distortions greater than 40 per cent and were similar to those obtained with positive characteristic distortion. The most common errors were as follows: J interpreted as K or M; N as T; G as M; and K as M. In addition, there were the usual miscellaneous interpretations, together with a large number of letters added and a smaller number of letters omitted.

4. *Effect of Fortuitous Distortion*

The curve of Fig. 6 which is an average of three test curves, illustrates the effect of fortuitous distortion upon the accuracy of reception. The distortions plotted are the maximum values which were equalled or exceeded about three or four times a minute. It appears from the curve that large amounts of fortuitous distortion have very little effect upon the accuracy of reception. In fact the accuracy is practically unchanged over the range of distortion from zero to 75 per cent. This is not very surprising in view of the fact that such distortions occur only about three or four times per minute, and may affect signals without destroying their identity. Even with fortuitous effects of such large values as to cause the breaking up of signals, only a relatively small decrease in the accuracy of reception was produced in certain tests.

The errors made by the operators for this type of distortion were largely of a miscellaneous nature. There were very few misinterpretations which were repeated consistently, although most of the errors indicated a shortening of dashes to dots and a dropping out of dots and spaces. There were also a considerable number of letters omitted and a few letters added.

Briefly summarized, the effects of fortuitous distortion upon reception are as follows. The accuracy of reception remained practically

unchanged up to a distortion of about 75 per cent and decreased slowly for distortions above this value. The operators first remarked upon the distortion at a value of about 35 per cent and considered the circuit unsatisfactory at about 50 per cent.

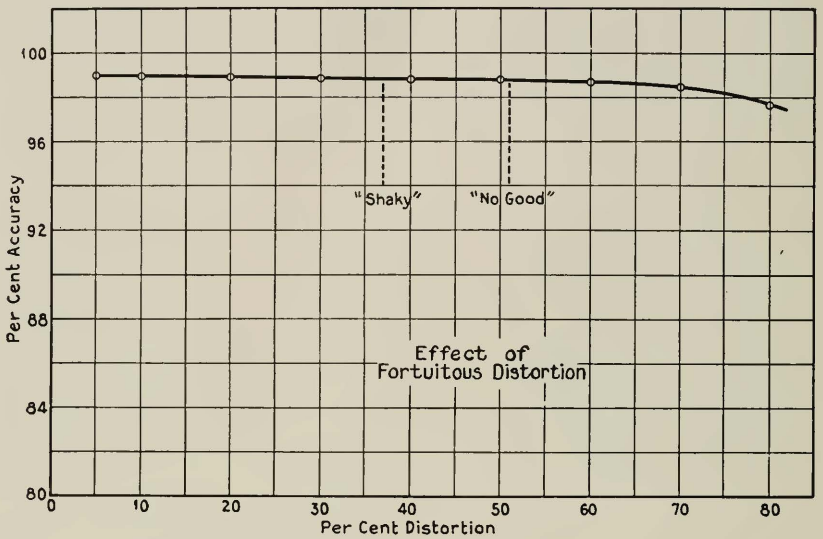


Fig. 6.

5. *Effect of Combined Distortions*

A large number of tests were made with various combinations of distortions, the results of which were also plotted in the form of curves. These curves are not included here because of lack of space but the results obtained are discussed below.

The maximum permissible distortion is generally much higher when the distortion occurs in combinations of various kinds than when it occurs singly. This is true when judged either from the standpoint of accuracy of reception or opinion of operators. Moreover, it appears that the shape of the distortion versus accuracy curve is determined largely by the predominating type of distortion.

It would seem from the above that a measurement of maximum distortion only, is not sufficient to determine the condition of a telegraph circuit. For example, a circuit may have a maximum distortion of about 50 per cent. If this is made up of 35 per cent negative characteristic distortion and 15 per cent fortuitous distortion the circuit would be considered very poor by the operators. If, however, it were made up of 15 per cent negative characteristic distortion and 35 per cent fortuitous distortion, the circuit would be considered fairly good. Similar deductions can be drawn from other combinations of distortion.

There was also some indication that a circuit which contains an appreciable amount of one type of distortion may be improved by adding a certain amount of some other type of distortion, even though the total distortion is increased thereby. This was brought out in a certain test where 25 per cent negative characteristic distortion by itself made the circuit unsatisfactory. By adding 15 or 20 per cent positive bias the quality of the signals was improved, as shown by the accuracy of reception and by the opinion of the operators, even though the maximum distortion was increased from 25 per cent to 40 or 45 per cent.

The nature of the errors made by operators for combined distortions in the circuit varied greatly for different combinations of distortion. Positive bias combined with moderate amounts of negative characteristic distortion gave errors similar to those for negative characteristic distortion alone. On the other hand negative bias combined with moderate amounts of negative characteristic distortion gave errors of a miscellaneous nature, such as were obtained with negative bias alone. Combinations of fortuitous distortion with moderate amounts of positive bias gave errors which indicated a consistent lengthening of marks and dropping out of spaces, as for example, interpretations of A as M; N as T and J as K. Combinations of fortuitous distortion with negative bias, however, gave errors which were chiefly of a miscellaneous nature. In addition to the above there was an appreciable number of letters added and omitted for all combinations of distortion.

A brief summary of the effects of combined distortion upon reception is as follows. The curves of accuracy versus distortion have the same general shape as the curves for the predominating type of distortion taken by itself, and the maximum permissible distortion, as judged either by the accuracy of reception, or by the opinion of the operators, is generally higher than for the various combined distortions taken by themselves. The values of total distortion at which the accuracy of reception began to decrease, ranged from about 55 to 100 per cent. The operators considered the circuits unsatisfactory at values of distortion ranging from about 50 to 65 per cent.

IV. Summary

There are given below conclusions which have been drawn tentatively from the results of the present tests. Further tests are desirable in order to confirm the results obtained thus far. For the present, however, it is thought that the ideas as to the effect of distortion upon

manual Morse operation outlined below, can be of material use as a general guide.

1. The effect of distortion upon accuracy of reception differs in character for different types of distortion. For some types of distortion the accuracy decreases rapidly when the distortion exceeds a certain value, which indicates that there is a rather definite limiting value of distortion. For other types of distortion, the accuracy of reception decreases very little nearly up to the point where the circuit actually fails.

2. The amount of distortion which appears to be tolerable, both from the standpoint of accuracy of reception and from the opinion of operators appears to be larger for combined distortions than for the individual types of distortion. This would seem to indicate that the maximum distortion by itself, without at least a general idea as to the components making up the distortion, is not sufficient to indicate how satisfactory the circuit will be for handling manual telegraph service.

3. The effect of distortion upon the operators themselves, as indicated by the errors which they made, appears to be such as to cause hesitation and a resulting confusion while interpreting distorted signal combinations. Most of the errors in the case of large distortions consisted of miscellaneous interpretations with letters frequently added or omitted. For some types of distortion, especially those for which the accuracy of reception begins to decrease rapidly at a certain value of distortion, consistent misinterpretations were made. These usually occurred after the accuracy of reception had decreased appreciably, and were of a nature such as would be expected from the type of distortion present in the signals.

4. The opinion of the operators as to the amount of distortion above which a circuit is unsatisfactory for commercial operation, is in reasonable agreement with the effect on their accuracy of reception for some types of distortion; for other types of distortion there is considerable disagreement. In general, the operators pronounce a circuit unsatisfactory before the point is reached where the accuracy of reception decreases appreciably.

The agreement which exists between the two criteria is shown in the following table. The values in the first column of the table are based on the opinion of the operators, while those in the second column are based on their accuracy of reception and are those values at which the accuracy has decreased 2 per cent from the initial value. In every case the value given is the average for all the tests made with a particular type of distortion.

Kind of Distortion	Limiting Values From Opinion of Operators	Limiting Values From Curves of Accuracy of Reception
1. Positive bias	40%	45%
2. Negative bias	50%	65%
3. Positive characteristic distortion	58%	55%
4. Negative characteristic distortion	28%	53%
5. Fortuitous distortion	51%	85%
6. Various combinations of distortion	50 to 65%	55 to 100%

5. Judged from the *opinion of operators*, all types and combinations of distortion with the exception of negative characteristic distortion and positive bias become objectionable at values of about 50 per cent. Negative characteristic distortion appears to become objectionable at about 30 per cent and positive bias at about 40 per cent. On the other hand, judged from the *accuracy of reception* of the operators, all distortions, with the possible exception of positive bias, become objectionable only at values above about 50 per cent. In the case of combined distortions there is considerable disagreement when large amounts of fortuitous distortion are present. The reason for this is discussed under Section III "Discussion of Results—4. Effect of Fortuitous Distortion."

The disagreement between opinion and actual performance is of considerable interest, since it shows that operators will condemn a circuit due to the presence of distortion even though their accuracy of reception is not materially influenced thereby. It is probable, therefore, that in establishing allowable limits of distortion for commercial telegraph circuits, both the accuracy of reception and the opinion of operators will have to be considered.

A Braun Tube Hysteresigraph

By J. B. JOHNSON

In this paper apparatus for observing hysteresis loops of magnetic materials is described. It combines a cathode ray oscillograph with a vacuum tube amplifier and an electrical integrating circuit consisting of condenser and resistance. The device describes the B–H curve for alternating magnetization in the frequency range of five to perhaps several thousand periods per second. The specimens may be either long strips or closed rings. Alternating flux as low as one maxwell may be readily observed.

The operation of the apparatus is analyzed so as to account for the effects of finite time constants of the amplifier and integrator, of conductance in the condensers, of demagnetization by current in the search coil and by the stray fields of coils and specimen, and of eddy currents in the specimen.

THE use of the cathode ray oscillograph for delineating magnetic hysteresis curves has proved a convenience in a number of studies of magnetic phenomena.[1] The essential advantage of the method lies in the speed with which the hysteresis loop is traced. Complete curves are drawn in rapid succession by the oscillograph tube, and any change in these curves resulting from altered mechanical or magnetic conditions of the specimen can immediately be observed and recorded. Furthermore, the area bounded by these curves represents the total energy loss in the specimen corresponding to the particular kind of magnetic cycle that is used, and not the hysteresis loss alone as is the case in the curves derived by the slower point-by-point methods.

In the present article is described an apparatus combining a cathode ray oscillograph with an electric circuit, for magnetic measurements. A fairly extensive analysis of the operation of the device is presented in order to show how accuracy may be maintained in the measurements and the probable errors estimated.

The apparatus has been in use since 1924, particularly for observing the magnetic properties of various alloys. It is so designed that by suitably choosing the circuit constants it can be used for obtaining the hysteresis curves of magnetic materials from the saturation value down to fields where the amplitude of alternating flux is about one maxwell. The purposes for which the apparatus has been employed are illustrated by the hysteresis curves reproduced in Plates I and II.

[1] K. Ångström, *Phys. Zeits.*, 1, p. 121, 1899; *Phys. Rev.*, 10, p. 74, 1900.
E. Madelung, *Ann. d. Phys.*, 17, p. 861, 1905; *Phys. Zeits.*, 8, p. 72, 1907.
C. W. Waggoner and F. A. Molby, *Phys. Rev.*, 17, p. 427, 1921.
Y. Niwa, J. Matura, and J. Sugiura, "Researches of the Electrotechnical Laboratory (Ministry of Commerce, Tokyo)," No. 144, May, 1924.

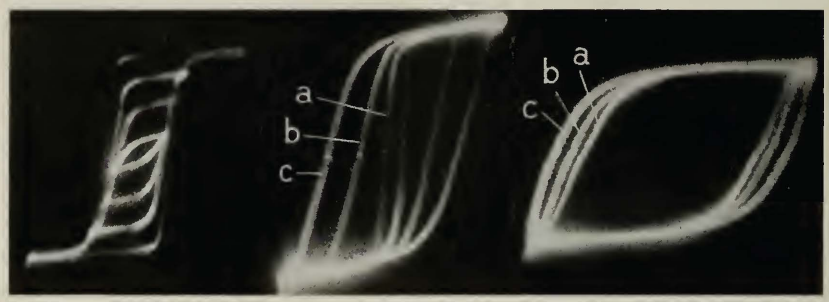
The hysteresis curves of Plate I (1) are those of a ribbon of Armco iron. The curves were made with 25 cycle magnetization and at four different magnetizing fields, the highest field being sufficient to saturate the sample. The effects of frequency and of permeability on the eddy current loss in materials are shown in the next two figures of this plate. Two similar ribbons were used. The one was permalloy magnetized to saturation, Plate I (2); the other, Plate I (3), was Armco iron at a low magnetizing field. The curves were made at three different frequencies, those marked with the letters *a*, *b* and *c* corresponding to magnetization at 25, 200 and 1,000 cycles, respectively. A comparison of the magnetic properties of pure iron and permalloy is made in the two figures, (4) and (5). In each figure the letter *a* indicates the curve for iron, *b* that for permalloy. The former figure was made at a field strength high enough to saturate the iron, the latter at a field strength which saturated the permalloy but not the iron. The last figure of this plate, (6), shows the influence of tension on the magnetic properties of a ribbon of permalloy containing 65 per cent nickel. The curve *a* was obtained without tension, *b* with a moderate tension on the ribbon. The curves of Plate II reproduce the magnetization cycles of a sample of Armco iron at various temperatures, from room temperature to the recalescence point. At the temperature of about 790° C., ferromagnetic properties were gone and the only flux that is indicated existed in the uncompensated air space of the search coil.

I. Description of Method

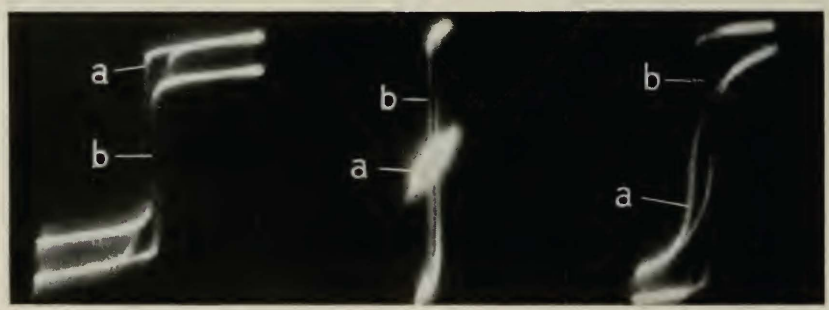
A change in magnetic flux in a specimen is usually measured either as a change in the field in a relatively short air gap, or as the time integral of the potential set up in a search-coil surrounding the specimen. The second method is illustrated by the use of the ballistic galvanometer as an integrating instrument, employed in most magnetic measurements. In the present case, however, the integration is accomplished by a purely electrical circuit. The integrating element consists of a resistance and condenser in series with the search coil surrounding the sample of material.[2]

[2] This type of integrating circuit is one of at least ten simple combinations of resistance, capacity and inductance which can be used for obtaining the cyclic integral of current or potential. Some of these have been described in connection with hysteresis measurements by E. L. Bowles, *Jl. A. I. E. E.*, 42, p. 849, 1923; O. E. Charlton and J. E. Jackson, *Jl. A. I. E. E.*, 44, p. 1220, 1925; W. Kaufmann, *Zeits. f. Phys.*, 5, p. 316, 1921. W. Kaufmann and E. Pokar, *Phys. Zeits.*, 26, p. 597, 1925. K. Krüger and H. Plendl, *Zeits. f. Hochfr.*, 27, pp. 155–161, 1926; W. Neumann, *Zeits. f. Phys.*, 51, p. 355, 1928. The circuit chosen here was used by Bowles and by Charlton and Jackson in connection with a mechanical oscillograph, and by Krüger and Plendl in connection with a Braun tube.

PLATE I. Hysteresis curves of iron and permalloy under various conditions.



1 2 3



4 5 6

1. Armco iron at various fields.
2. Permalloy at (*a*) 25˜, (*b*) 200˜, (*c*) 1,000˜; high field.
3. Armco iron at (*a*) 25˜, (*b*) 200˜, (*c*) 1,000˜; low field.
4. Armco iron (*a*) and 78 per cent permalloy (*b*); high field.
5. Armco iron (*a*) and 78 per cent permalloy (*b*); low field.
6. 65 per cent permalloy, (*a*) without tension, (*b*) with tension.

PLATE II. Magnetic cycle of Armco iron at various temperatures.



a. 20° C. *b*. 380° C. *c*. 555° C.



d. 690° C. *e*. 733° C. *f*. 757° C.



g. 775° C. *h*. 788° C. *i*. 795° C.

The principle of the arrangement whereby the cyclic magnetization curve is recorded on the oscillograph tube, will be made clear by reference to Fig. 1. The specimen of magnetic material M, is mag-

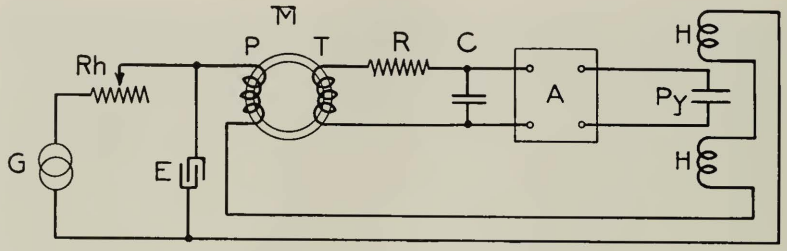


Fig. 1—Elementary diagram of the circuit.

netized by alternating current flowing through the primary winding P. The varying magnetic flux in the specimen induces in the secondary winding or search coil T a voltage which is proportional to the *rate of change of flux.* This voltage is applied to the integrating circuit consisting of the resistance R and the condenser C. When the resistance is large compared with the impedance of the condenser, the current in this circuit is limited largely by the resistance and it is, therefore, proportional to the voltage applied by the search coil. The charge on the condenser, and therefore the voltage across its terminals, is proportional to the time integral of the current in the circuit. The *voltage of the condenser* is, therefore, proportional to the *magnetic flux in the sample.*

This voltage is amplified by the distortionless amplifier A, the output side of which is connected to one pair of deflector plates of the oscillograph tube. While the deflection of the indicating spot thus follows the *flux* in one direction, deflection in a line at right angles to this and proportional to the *magnetizing field* is produced by the magnetic field of the deflector coils H which are connected in series with the magnetizing winding P. The spot then traces out a path on the screen during each cycle of current which is the hysteresis diagram for the sample, and which by suitable calibration yields quantitative results.

The voltage on the integrating condenser at any time is given by the relation

$$e = 10^{-8} \int_{-\infty}^{t} \frac{NS}{RC} \frac{dB}{dt}\, dt = \frac{NS}{RC} B \times 10^{-8} \text{ volts}, \qquad (1)$$

where R and C are the resistance and capacity of the integrator, N is the number of turns in the search coil, S is the cross-sectional

area of the sample and B the flux density. The voltage e is amplified and applied to the deflector plates of the oscillograph. This part of the apparatus is calibrated in terms of a small alternating voltage of known amplitude applied to the amplifier, which produces a deflection of d cm. per volt on the oscillograph tube. An ordinate b on a hysteresis curve therefore indicates an induction of

$$B = \frac{b}{d}\frac{RC}{NS} \times 10^8 \text{ gauss.} \qquad (2)$$

The calibration for magnetizing field is done by measuring the deflection produced by a known direct current passed through the coils H, and the magnetizing field in the coil P is then calculated from the dimensions of the coil in the usual way in terms of the current indicated by the oscillograph.

The greatest difference between the actual circuits used and the simple one shown in Fig. 1 is that the amplifier is constructed with the push-pull arrangement in order to reduce distortion. This makes necessary the maintenance of symmetry on the two sides of the circuit so that the apparatus is really two similar circuits in parallel as shown in Fig. 2.

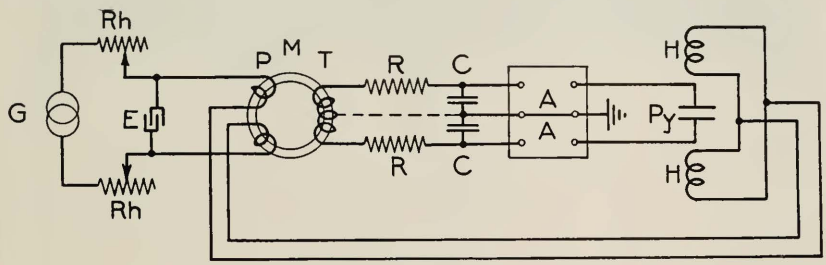


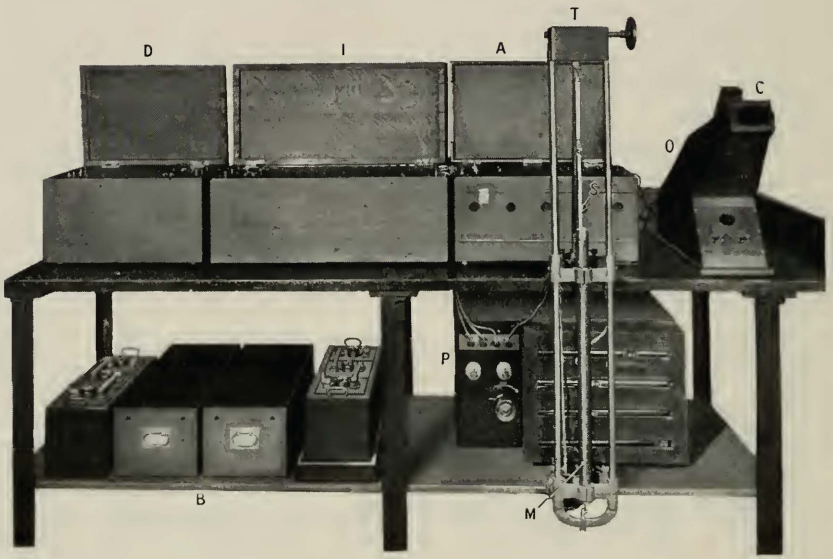
Fig. 2—Diagram of the symmetrical circuit.

The fact that the integrator and amplifier are thus connected makes no difference in the final results of the calculations, if for N is used the total number of turns on the search coil and for the other constants only the values on one side of the system.

II. Description of the Apparatus

For greater flexibility the apparatus has been made in a number of separate units which will now be described. A photograph of the complete assembly is reproduced in Plate III.

PLATE III. Apparatus assembly.



A—Amplifier
B—Batteries
C—Camera
D—Calibration Set
I—Integrator
M—Magnetizing Coil
O—Oscillograph Tube Box
P—Power Unit
S—Sample
T—Tension Rack

1. The Integrator.

The integrator unit contains two banks of paper condensers of 100 mf. each, variable in steps of 1 mf., 3 mf., 6 mf., and 9 steps of 10 mf. The two resistances are each variable by factors of about two from 12,000 ohms to 1,000,000 ohms. These combinations are thought to cover any requirements that are likely to arise.

2. The Amplifier.

Four stages of Western Electric 102-D tubes on each side, resistance-capacity coupled, make up the amplifier. The time constant of each coupling unit (stopping condenser and grid leak) is 8 seconds, which is sufficient for frequencies down to 10 p.p.s. or less. The amplifier is suitably shielded against electromagnetic, mechanical and acoustic shocks.

3. The Oscillograph.

A box contains the cathode ray oscillograph tube [3] and its controls, except the batteries. The coils *H* are mounted on a hard rubber holder which can be clamped securely to the tube. A camera attachment contains a Dallmeyer F. 1.9 lens with which photographs of the pattern can be made on plates or film pack.

The oscillograph tube is connected to the amplifier unsymmetrically since one side of the amplifier leads to the common deflector plates and the batteries of the oscillograph tube, the other side only to a deflector plate. The oscillograph and its batteries must, therefore, be so placed that the conductance and capacity to ground and to the rest of the apparatus are small. When this is done no distortion results from the lack of symmetry.

4. The Calibration Set

The calibration box contains the apparatus for applying a small sinusoidal voltage of known amplitude to the amplifier and oscillograph. The set takes current from the same source that supplies the magnetizing current. A low-pass filter admits only current of the fundamental frequency, and a thermocouple provides for measuring the output current of the filter. This current passes through a potentiometer from which the calibrating voltage is applied to the amplifier. The potentiometer is variable in several steps, each reducing the current amplitude by a factor of about two.

[3] Western Electric 224-B tube. J. B. Johnson, *J. O. S. A.* and *R. S. I.*, 4, p. 701, 1922.

5. The Power Unit.

The alternating magnetizing current is derived from a low speed dynamotor. The machine operates on 24 volt storage battery power, and delivers at low load, nearly sinusoidal current of 24 volts *amplitude*. Resistance in the field circuit permits regulation of the frequency between 10 and 30 cycles, while rheostats in each side of the output of the machine serve to regulate the current for the magnetizing coil.

Mounted in the power unit and connected directly across its output terminals there is an electrolytic condenser of about 1,200 mf. (*E* of Figs. 1 and 2). This condenser serves three important functions: *a*. It smoothes out ripples in the magnetizing current which would, if they were large enough, produce secondary loops in the hysteresis curve. *b*. It makes the power unit a source of potential having low impedance to a.c. so that the magnetizing current is in part determined at any moment by the counter e.m.f. of the magnetizing coil, rather than wholly by external impedances. This being so, the magnetizing current is retarded in the steep parts of the hysteresis curve of the sample where the counter e.m.f. is great. Eddy currents do not, therefore, build up in the sample nearly so much when the condenser is in the circuit as when it is not. *c*. The spot on the oscillograph tube being thus slowed down on the steep parts of the hysteresis curve and correspondingly speeded up on the saturation parts results in a much more uniform brightness of pattern which can be observed more readily.

6. Magnetizing Coils, Search Coils, and Samples.

The samples which have been used are mostly in the form of either flat rings stamped from sheet metal, or long straight strips of thin tape. The rings are about four inches in diameter so as to fit into a toroidal furnace made for magnetic testing.[4] When the furnace is used the windings are applied on the furnace, 45 turns for the magnetizing winding and 90 turns for the search coil. When used without the furnace a similar number of turns are wound directly on the sample. The number of turns in the search coil being small the cross-sectional area of the sample is made correspondingly large, about 1/4 square inch in order to apply sufficient voltage to the integrator.

The apparatus for use with the single ribbon samples consists of straight magnetizing and search coils. A glass tube about $1\frac{1}{4}$ inches in diameter carries two parallel windings 30 inches long, for producing the magnetizing field. At the center of the length of this coil and

[4] The furnace will be described by Mr. G. A. Kelsall in a forthcoming number of the Journal of the Optical Society of America and Review of Scientific Instruments.

side by side are placed two equal search coils, connected in series opposing. These coils are 10 inches long, and in one assembly carry 30,000 turns each, in another 7,000 turns each. The sample, about 40 inches long, is placed in the axis of one search coil, the other coil serving to compensate for the air space of the first. This assembly is mounted in a brass frame to which is attached a windlass and spring scale for applying tension to the sample when so desired. The effect of the earth's magnetic field on the specimen is made negligible by either passing a direct current through one of the magnetizing windings in series with a high inductance, or by passing a direct current through the two coils from a battery in parallel with the electrolytic condenser.

III. Mathematical Analysis of the Apparatus.

The performance of the apparatus will be analyzed for two simple cases representing extremes in the properties of the magnetic material. In the first case it will be assumed that the flux in the search coil is proportional to the magnetizing field, a condition that is approached with hard materials at low field strengths. In the second case the assumption will be made that the material becomes saturated perfectly in the positive or negative direction, as the magnetizing field passes through zero. This is the limiting case of soft materials at high alternating field strengths.

1. Circuit Corrections for the Case of Small Magnetizing Fields.

The circuit of the integrator is represented in Fig. 3. The condenser

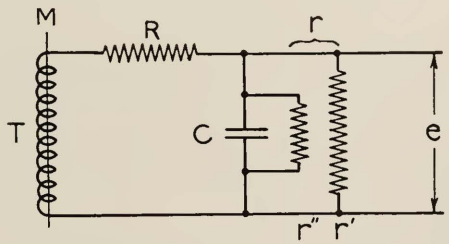


Fig. 3—Equivalent circuit of the integrator.

is shunted by an effective resistance r made up of the grid leak resistance r' and the condenser resistance r''. The conductance of condensers is a function of frequency and temperature and possibly of other factors. As a function of frequency it may be written

$$G = \frac{\omega C}{Z_1} + \frac{\omega^2 C}{Z_2} + \cdots. \tag{3}$$

If the terms in powers of ω higher than the first are neglected, the resistance is

$$r'' = \frac{1}{G} = \frac{Z}{\omega C}, \tag{4}$$

where Z is the resistance of a one farad condenser at $1/2\pi$ cycles per second.

In terms of the rate of change of induction, B, in the sample and the output voltage e of the integrator, the differential equation of the circuit in Fig. 3 is

$$\frac{de}{dt} + \omega Je = -NS\frac{dB}{dt}, \tag{5}$$

where

$$J = \frac{1}{RC\omega}\left[1 + \frac{R}{r'} + \frac{RC\omega}{Z}\right].$$

Let us now assume that the magnetizing field is sinusoidal,

$$H = H_1 \cos \omega t;$$

that

$$B = \mu H, \ \mu \text{ being considered constant;}$$

and that

$$e = a \cos \omega t + b \sin \omega t.$$

Making these substitutions in equation (5) and solving for a and b by equating coefficients of like terms, we get

$$e = -\frac{NS}{RC}\mu H_1 \frac{1}{1 + J^2} \ [\cos \omega t - J \sin \omega t]. \tag{6}$$

The value of J is less than .01 in this apparatus so that J^2 can be neglected compared with unity. The negative sign is merely a matter of convention, since the curve can be turned by reversing one pair of terminals. The sign of the $\cos \omega t$ term will therefore be taken positive henceforth, and we have

$$e = \frac{NS}{RC}B\left[\cos \omega t - \frac{1}{RC\omega}\left(1 + \frac{R}{r'}\right)\sin \omega t - \frac{1}{Z}\sin \omega t\right]. \tag{7}$$

The value of e therefore departs from that for a perfect integrator by two terms, one depending upon the time constant of the integrator, the frequency, and the value of the grid leak resistance of the first amplifier stage; the other term depending only upon the conductance of the integrating condenser.

This voltage e is applied to the first stage of the amplifier, resulting in a voltage e_1 at the output of this stage which is then applied to the second stage, and so on. Fig. 4 represents the first stage of the

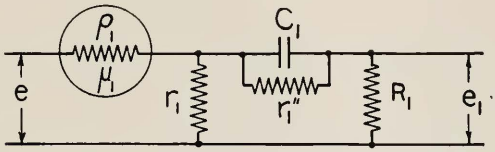


Fig 4—Equivalent circuit of one stage of the amplifier.

amplifier, ρ_1 being the internal resistance of the vacuum tube whose amplification constant is μ_1, r_1 the external plate resistance, R_1 the grid leak resistance for the next tube, C_1 the coupling capacity having the effective leak resistance r_1''. Solving for e_1 by the same method that was used for e and omitting negligible terms gives the result

$$\left.\begin{aligned} e_1 &= \frac{NS}{RC}\frac{B\mu_1 R_1 r_1}{(R_1 + r_1)(\rho_1 + r_1) - r_1^2} \\ &\left[\cos \omega t - \left(\frac{1}{r'C\omega} + \frac{1}{RC\omega} + \frac{1}{R_1 C_1 \omega} + \frac{2}{Z}\right)\sin \omega t\right] \end{aligned}\right\}. \tag{8}$$

Similarly, the output from s stages of the amplifier is

$$e_s = \frac{NS}{RC} B\left[\prod_1^s M_s\right]\left[\cos \omega t - \left(\frac{1}{r'C\omega} + \sum_0^s \frac{1}{R_s C_s \omega} + \frac{s+1}{Z}\right)\sin \omega t\right], \tag{9}$$

where

$$M_s = \frac{\mu_s R_s r_s}{(R_s + r_s)(\rho_s + r_s) - r_s^2} \doteq \mu_s \frac{r_s}{\rho_s + r_s}.$$

The product from 1 to s contains only amplifier constants, while the summation from 0 to s means that the values for the integrator are included.

According to equation (9) the circuit introduces errors which at low fields makes the spot describe an ellipse instead of a straight line with positive slope. This ellipse is traced in the clockwise direction, the opposite direction to that in which the apparatus traces hysteresis loops. Since in all practical cases the hysteresis loop at low frequencies approaches an ellipse traced in the counter-clockwise direction, the effect of finite time constants and condenser resistance is to make the figure traced by the apparatus narrower than it ought to be.

Fig. 5 illustrates the nature of this distortion. The straight line aa' is the curve for the ideal material and perfect apparatus for which

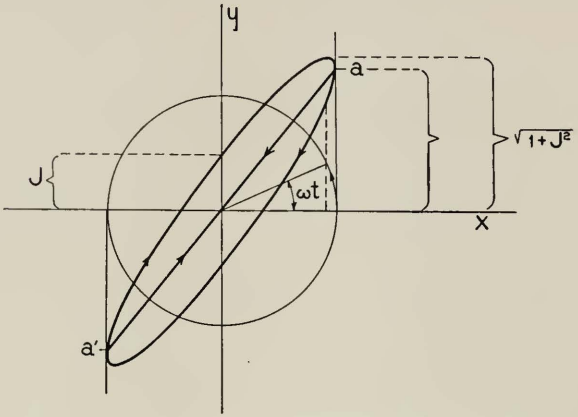


Fig. 5—Distortion produced by integrator and amplifier at low flux densities.

the maximum amplitude in the B direction may be taken as unity. The ellipse is the curve

$$\left.\begin{aligned} x &= \cos \omega t, \\ y &= \cos \omega t - \sum_0^s J_s \sin \omega t, \end{aligned}\right\} \tag{10}$$

which the apparatus traces. The point of tangency of the ellipse with a line parallel to the B axis is at the point a and always remains there. The intersections with the B axis are at $\pm \Sigma J_s$ and the height exceeds that of the point a by the usually very small quantity $\sqrt{1 + (\Sigma J_s)^2} - 1$.

In the present system the constants are: $R_s = 2 \times 10^6$ ohms; $C_s = 4 \times 10^{-6}$ farad; $r' = 2 \times 10^6$ ohms. We may set $R = 10^5$ ohms, $C = 20 \times 10^{-6}$ farad, $\omega = 100$, and assume four stages of amplification. For the type of condensers used in the integrator and amplifier, $Z = 250$. The vertical width of the ellipse is then 3 per cent of the total height, of which .5 per cent is caused by the time constant of the integrator, .5 per cent by the time constant of the four amplifier couplings (including the output circuit), and 2 per cent by the conductance of the condensers in the circuit. The conductance therefore contributes the largest error in the present system, showing that the time constants have been made large enough for practical purposes.

2. Circuit Corrections for the Case of Large Magnetizing Fields.

When the induction in the specimen changes abruptly between negative and positive saturation as the magnetizing field passes

through zero, it may be represented by the Fourier series

$$B = B_m \frac{4}{\pi} \sum_{1}^{n} (-1)^{n-1} \frac{1}{m} \cos m\,\omega t, \tag{11}$$

where B_m is the saturation value, $n = 1, 2, 3$, etc., and $m = 2n - 1$. It is assumed that the magnetizing field has the constant fundamental frequency $\omega/2\pi$, but it need not be sinusoidal. Each term of this series can be treated as was the single term in the expression of B for the small fields. When this is done for the integrator and amplifier and the terms are again summed, the output voltage is given by

$$\begin{aligned} e_s &= \frac{NS}{RC} B_m \frac{4}{\pi} \left[\prod_{1}^{s} M_s \right] \sum_{1}^{n} \left[(-1)^{n-1} \left\{ \frac{\cos m\,\omega t}{m} \right.\right. \\ &\quad \left.\left. - \left(\frac{1}{r'C\omega} + \sum_{0}^{s} \frac{1}{R_s C_s \omega} \right) \frac{\sin m\,\omega t}{m^2} - \frac{s+1}{Z} \frac{\sin m\,\omega t}{m} \right\} \right] \\ &= \frac{NS}{RC} B_m \left[\prod_{1}^{s} M_s \right] \left[\alpha - \left(\frac{1}{r'C\omega} \sum_{0} \frac{1}{R_s C_s \omega} \beta - \frac{s+1}{Z} \gamma \right) \right], \end{aligned} \tag{12}$$

where

$\alpha = 1$ from $\omega t = -\frac{\pi}{2}$ to $+\frac{\pi}{2}$;

$\alpha = -1$ from $\omega t = \frac{\pi}{2}$ to $3\frac{\pi}{2}$; (Fig. 6)

$\beta = \omega t$ from $\omega t = -\frac{\pi}{2}$ to $+\frac{\pi}{2}$;

$\beta = (\pi - \omega t)$ from $\omega t = \frac{\pi}{2}$ to $3\frac{\pi}{2}$; (Fig. 7)

$\gamma = \frac{4}{\pi} \sum_{1}^{n} (-1)^{n-1} \frac{\sin m\,\omega t}{m}$; (Fig. 8)

$m = 2n - 1$.

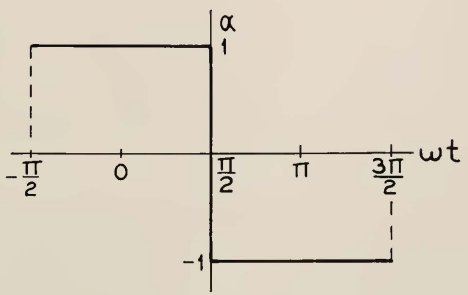


Fig. 6—The value of α during one cycle.

The value of γ is finite for all values of ωt except $\pi/2$ and $3(\pi/2)$ where it becomes infinite. Near these values, therefore, the equation fails for the reason that the higher powers of ω were neglected in the expression for the condenser conductance (Eq. 4).

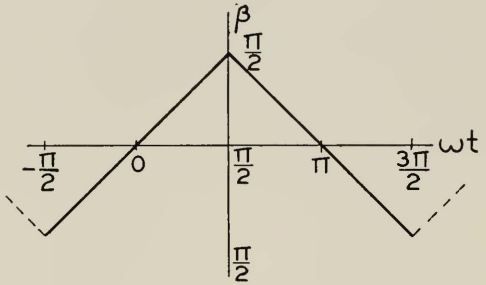


Fig. 7—The value of β during one cycle.

Let the heavy line $abob'a'$, Fig. 9, be the trace of the ideal curve

$$\left.\begin{aligned} x &= \cos \omega t, \\ y &= \frac{4}{\pi} \sum_{1}^{n} (-1) \frac{\cos m \omega t}{m}. \end{aligned}\right\} \tag{13}$$

The curve followed by the value of e_s in equation (12) is then of the form shown by the light line $ac'a'ca$, which again is traced in the

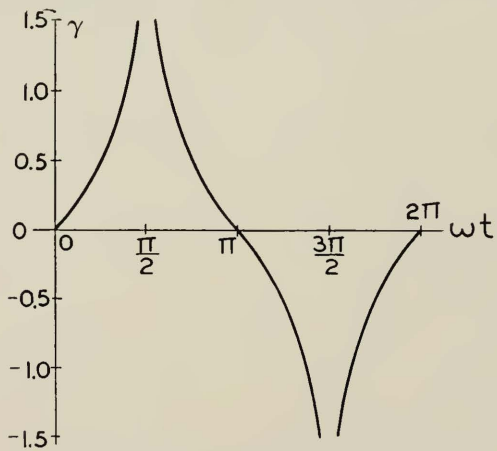


Fig. 8—The value of γ during one cycle.

clockwise direction. When the apparatus is used for an actual material having positive hysteresis, the loop is made narrower at the *II* axis by this distortion, but the amount of the decrease in width

can not be determined analytically without more accurate knowledge of the condenser conductance as a function of frequency. The value of equation (12), therefore, lies chiefly in giving the distortion to be

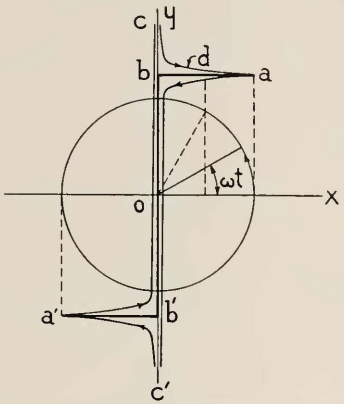


Fig. 9—Distortion produced by the integrator and amplifier for an ideally saturated material.

expected along the saturation parts of the curve, where the curve is made narrower in the B direction and in extreme cases crosses itself as shown in Fig. 10.

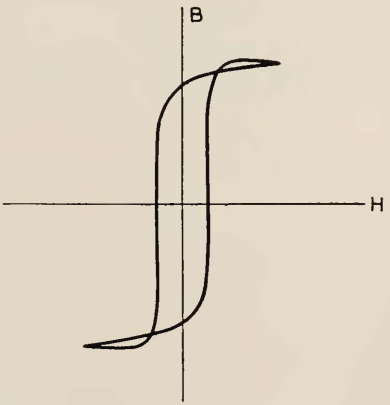


Fig. 10—Distortion produced by the integrator and amplifier with a normally saturated material.

The amount of the distortion will be estimated for the same circuit constants as were used above, but with only three stages of the amplifier. The distorted curve coincides with the ideal one at the points a and a', Fig. 9. At the point d, where $\omega t = \pm 60°$, we have $\alpha = 1$, $\beta = 1.05$ and $\gamma = .84$ for $\omega t = 60°$; and $\alpha = 1$, $\beta = -1.05$,

$\gamma = -.84$ for $\omega t = -60°$. Calculation by equation (12) gives for the vertical spread of the curve at this point 2.3 per cent of the double height of the curve, of which 1 per cent is contributed by the time constant factors and 1.3 per cent by the condenser conductance. It is seen that of the errors discussed in this and the preceding section, those caused by the conductance of the condensers are the larger.

3. Demagnetization by Current in the Search Coil.

The voltage induced in the search coil by the change of flux in the sample creates a current in the search coil circuit. The magnetic field set up by this circuit in the search coil opposes changes in the field due to the current in the magnetizing winding and so makes the hysteresis loop appear wider than it actually is.

At low fields the effect may be calculated with some degree of accuracy. Let the field induced by the magnetizing coil be $H_1 \cos \omega t$, and let the search coil contain a sample which has the initial permeability μ and the cross section S. The field set up by the search coil current is then

$$-\frac{4\pi}{10}\frac{N}{l}\cdot\frac{NS}{R}\frac{dB}{dt}\times 10^{-8}$$

and the total field is, very nearly,

$$H = H_1\left(\cos \omega t + \frac{4\pi}{10}10^{-8}\frac{N^2S}{lR}\mu\omega \sin \omega t\right).$$

The induction is

$$B = \mu H = \mu H_1\left(\cos \omega t + \frac{4\pi}{10}\times 10^{-8}\frac{N^2S}{lR}\mu\omega \sin \omega t\right), \qquad (14)$$

while the magnetizing field which the tube indicates due to the influence of the deflector coils is $H_1 \cos \omega t$. The spot therefore traces an ellipse in the *counter-clockwise* direction, so as to add to the effect of hysteresis in the material.

Assuming the values $N = 30{,}000$, $l = 30$, $R = 100{,}000$, $\mu = 10{,}000$ (permalloy), $S = .005$, $\omega = 100$, the width of the ellipse along the B axis is about 6 per cent of its total height. This is a considerable error, so that when working with so high a permeability a high value of amplification may have to be used while the factor N^2S/lR is decreased.

When the magnetizing field is such as to saturate the sample the calculation of the effect of search coil current is more uncertain.

Let the true cycle of magnetization of the sample be that shown in Fig. 11. Let the magnetizing field be sinusoidal, $H' = H_1 \cos \omega t$, so

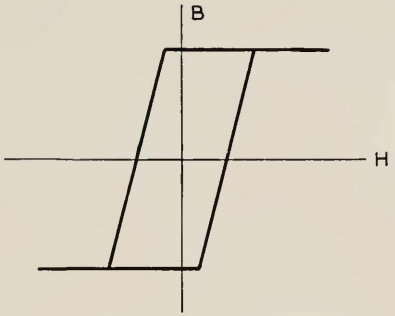


Fig. 11—Simplified hysteresis loop.

that the rate of change of the field in the region where the induction is variable is nearly

$$\frac{dH'}{dt} \doteq \omega H_1.$$

When the induction is varying there is a reverse magnetizing field set up by the current in the search coil of the strength

$$\Delta H = \frac{4\pi}{10} 10^{-8} \frac{N^2 S}{lR} \frac{dB}{dt} = \frac{4\pi}{10} 10^{-8} \frac{N^2 S}{lR} \frac{dB}{dH} \frac{dH}{dt}.$$

The value of dH/dt is not known, since it itself involves ΔH, but its maximum value is $(dH'/dt) = \omega H_1$. The reverse field is, therefore, at the most

$$\Delta H = \frac{4\pi}{10} 10^{-8} \frac{N^2 S}{lR} \frac{dB}{dH} \omega H_1. \tag{15}$$

The calculation of ΔH has been done for a number of cases. With the ring shaped samples having search coils of 90 to 135 turns the value of ΔH was .1 per cent of the coercive force H_c for iron, and 4 per cent of H_c for permalloy. When a permalloy having a thin hysteresis loop with the slope $(dB/dH) = 200{,}000$ is used in the search coil of 30,000 turns, the value of ΔH is nearly as large as H_c so that the curve is approximately doubled in width. In such a case the factor N^2S/lR must be made smaller, preferably by decreasing N^2 since this does not involve a corresponding increase in the amplification.

It must be remarked that these formulæ give the maximum errors, and that in fact the errors are considerably less. The condenser across the magnetizing coil permits the magnetizing current to take other than sinusoidal form, so that the value of dH/dt is much smaller than ωH_1.

4. Demagnetizing Factor of Magnetizing Coil and Sample.

The errors under this heading pertain to straight coils and samples and include the open end correction of the magnetizing coil, the error due to the non-uniformity of the magnetizing field of the open ended coil, and the demagnetizing factor of the finite length of sample.

Calculations by L. W. McKeehan and measurements by P. P. Cioffi [5] on coils and samples of nearly the same dimensions as those used in this apparatus, have shown that the errors in H and B which result from neglecting the above factors amount to less than three per cent. The exact correction to be applied for any particular sample is not readily calculated, but the sample can be assumed to be sufficiently long if displacing it a few centimeters in the coils results in no appreciable change in the hysteresis loop.

5. Eddy Currents and Magnetic Viscosity in the Sample.

Currents induced in the sample flow in such a direction as to oppose the applied magnetizing force, with the result that to reach any given density of magnetization a higher applied field is required when eddy currents are present. The loss of energy in eddy currents is added directly to the hysteresis loss, so that the dynamic hysteresis curve is different from the static one. Qualitatively this difference manifests itself as a widening in the H direction of the recorded loop at intermediate frequencies, and a shrinkage in the B direction when the frequency is so high that skin effect prevents the induction from reaching its full value before the applied field has receded appreciably from its maximum.

In dealing with the distortion introduced by eddy currents, the procedure must be governed by the nature of the problem in hand. If the total energy loss corresponding to a given magnetic cycle is sought, no eddy current correction is required. If it is desired to reproduce as nearly as possible the static hysteresis loop of a specimen, then the frequency of magnetization should be chosen low enough so that, with the aid of the condenser in parallel with the magnetizing coil, the eddy current effect is not appreciable. If, finally, the object of the work is to detect differences between the static and the dynamic hysteresis curves, the presence of eddy currents must be reckoned with. The following analysis, based upon a simplified static magnetization cycle and an assumed sinusoidal time variation of the magnetizing field, makes possible a rough estimation of the apparent increase in the coercive force H_c introduced by the eddy currents.[6]

Let Fig. 12 represent the cross-section of a rectangular lamination

[5] P. P. Cioffi, *Jl. Opt. Soc. Am.* and *Rev. Sc. Inst.*, 9, p. 53, 1924.

[6] A similar result has been obtained and used by Neumann, l.c.

of iron having the thickness d and the width c. The sample is being magnetized in the direction perpendicular to the plane of the paper. The magnetization produces eddy currents and we shall assume that the elements of induced current flow in circuits like the shaded element

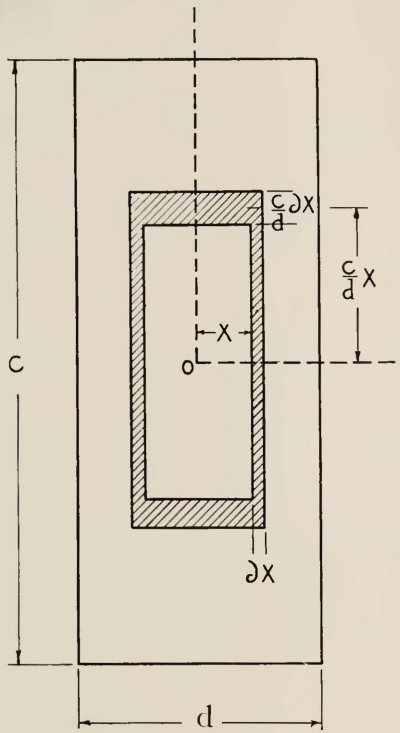


Fig. 12—Cross-section of specimen.

of area at the distance x, $-x$ on the horizontal axis and the distance $(c/d)x$, $-(c/d)x$ on the vertical axis, with the corresponding widths dx and $(c/d)dx$. The current density i_x which is induced in the elementary circuit is

$$i_x = \frac{e}{r} = 10^{-8} \frac{d\varphi_{cx}}{dt} \frac{1}{\rho} \frac{cd}{4(c^2 + d^2)} \frac{1}{x}. \tag{16}$$

The flux φ_{cx} is the total flux enclosed by the elementary circuit, and ρ the specific resistance of the iron. Within the elementary area itself the flux $d\varphi_x$ is the sum of the fluxes produced by the applied field and by the field due to the eddy currents external to the elementary area. If it is now assumed that the applied field is just passing through

the value H_c and that for a small variation in H about this point the flux is proportional to $H - H_c$, the flux threading the elementary circuit is

$$\varphi_{cx} = \int_0^x d\varphi_x = \int_0^x 8\,\frac{c}{d}\,x dx\,\frac{dB}{dH}\left[H - H_c - \frac{4\pi}{10}\int_x^{d/2} i_x d_x\right]. \quad (17)$$

The value of dB/dH is obtained from the static loop. It does not depend upon x or t and is for the moment assumed independent of H. The total flux φ in the sample is, therefore,

$$\varphi = 8\,\frac{c}{d}\,\frac{dB}{dH}\,(H - H_c)\int_0^{d/2} x dx - 8\,\frac{c}{d}\,\frac{4\pi}{10}\,\frac{dB}{dH}\int_0^{d/2} x dx \int_x^{d/2} i_x dx \quad (18)$$

$$= cd\,\frac{dB}{dH}\,(H - H_c) - \varphi_2. \quad (19)$$

The term φ_2 may now be expanded by successively substituting in equation (18) the value of i_x from equation (16). The result is the series

$$\varphi = cd\,\frac{dB}{dH}\,(H - H_c) - \frac{\pi}{4}\,\frac{10^{-9}}{\rho}\,cd^3\,\frac{(dB)^2}{(dH)}\,\frac{dH}{dt} + \varphi_r. \quad (20)$$

where φ_r is the sum of higher terms and where the factors have been simplified by regarding d small compared to c.

Now the coercive force obtained by the dynamic method is the value H_a of the field at the point of the curve where the total flux φ is zero. The apparent increase in coercive force caused by eddy currents can at this point be obtained from equation (20), giving

$$\Delta H_c = H_a - H_c = \frac{\pi}{4}\,\frac{10^{-9}}{\rho}\,d^2\,\frac{dB}{dH}\,\frac{dH}{dt}. \quad (21)$$

We may now inquire what, in a qualitative way, is the distortion of the hysteresis loop predicted by equation (21).

a. The widening of the loop is proportional to the square of the thickness of the lamination.

b. If the magnetizing force is kept sinusoidal the value of dH/dt is proportional to H_m. Since also dB/dH increases with H_m the value of ΔH_c increases at a greater rate than the first power of H_m.

c. For narrow loops and sinusoidal magnetizing current ωH_m is a sufficiently close approximation to dH/dt, while with loops which are wide because of hysteresis loss $(dH/dt) = \omega\sqrt{H_m^2 - H_c^2}$. If now the curve is widened by eddy currents, these values of dH/dt are too large and the correction to be applied is smaller than that given by inserting these values in equation (21).

d. For small values of ΔH_c this correction is proportional to the frequency of the magnetizing field, through the factor dH/dt. When the correction becomes appreciable compared to H_m the effect considered under *d* becomes active and ΔH_c increases at a slower rate than the first power of the frequency.

e. The value dB/dH is not constant as was assumed but depends upon H and attains a maximum near $H = H_c$. The effect of eddy currents is to make the magnetization non-uniform over the cross-section of the sample, and the average value of dB/dH is then considerably less than the maximum value. This effect again makes the value of ΔH_c increase at a slower rate than the first power of the frequency and of H_m.

f. When conditions are such that ΔH_c is small, a sufficient approximation to its value is obtained by deriving dB/dH from the dynamic loop so that the correction can be made without reference to the static curve.

Besides the eddy current effect, widening of the hysteresis loop has also been ascribed to what has been called magnetic viscosity, a property of the material which causes the induction to lag behind the actual magnetizing force. The existence of this phenomenon has been affirmed by some experimenters and denied by others. Not enough is known about it to make its discussion here pertinent.[7]

6. Non-Linearity of the Amplifier.

The push-pull construction of the amplifier compensates to some extent for curvature of the tube characteristics. In order to have full compensation, the tubes should be matched so that the tubes of each pair work on a part of their characteristic having the same steepness and curvature.

Furthermore, the last pair of tubes, the tubes working into the oscillograph, should have a practically straight characteristic over about 30 volts on each side of the average plate voltage, or in the case of 102-D tubes about 1.5 volts on each side of the average grid voltage.

7. Observational Errors.

Because of the width of the fluorescent trace on the screen, there is a probable error of about .2 mm. in measuring the width or height of the hysteresis loop on the tube. This corresponds to probable errors of the order of 1 per cent in B and 2 per cent to 3 per cent in H_c. Probable errors of 1 per cent are also introduced in the calibration

[7] For a recent consideration of time-lag in magnetization see R. M. Bozorth, *Phys. Rev.*, 32, p. 124, 1928.

measurements. The thermocouple reading involves an error of perhaps 1 per cent. The wave shape factor of the calibrating voltage must in general be considered, but in this apparatus the distortion is probably made negligible by the filter.

The deflections on the tube are not absolutely proportional to the deflecting forces, but since the calibrations are made over approximately the same distances as the hysteresis measurements the error thus introduced must be small.

Altogether, the probable observational errors may amount to as much as 4 per cent.

The Receiving System For Long-Wave Transatlantic Radio Telephony[1]

By AUSTIN BAILEY, S. W. DEAN, and W. T. WINTRINGHAM

Transmission considerations and practical limitations indicate that in the lower frequency range, frequencies near 60 kc are best suited for transatlantic radio-telephone transmission. A radio receiving location in Maine gives a signal-to-noise ratio improvement over a New York location equivalent to increasing the power of the British transmitter about 50 times.

Various types of receiving antennas are briefly discussed. The wave-antenna is selected as being most suitable for long-wave radio telephony. The various factors affecting wave-antenna performance and methods for measuring the physical constants of wave-antennas are discussed in detail. High-frequency ground conductivities determined from wave-antenna measurements are given. Combination of several antennas to form arrays is found to be a desirable means of decreasing interference. The use of a wave-antenna array in Maine decreases the received noise power by an additional 400 times. If the receiving were to be accomplished near New York using a loop antenna, we would have to increase the power of the British transmitting station 20,000 times to obtain the same signal-to-noise ratio. Comparisons of calculated and observed directional diagrams of wave-antennas and wave-antenna arrays are presented and discussed.

The transmission considerations governing the design of a radio receiver for commercial telephone reception are outlined.

Mathematical discussions of the wave-antenna, antenna arrays, quasi-tilt angle, and probability of simultaneous occurrence of telegraph interference are given in the appendices.

EARLY in October, 1915, engineers of the Bell System stationed in Paris heard the words "good night Shreeve," which had been transmitted from Arlington. That date then marks the inception of transatlantic radio-telephone receiving. The progress which has been made in the radio-telephone receiving art since these first experiments is demonstrated by contrasting the homodyne receiver and the non-directional antenna then used with the present commercial receiving system employing double-demodulation of single side band signals and an extensive array of wave-antennas forming a highly directional system. In the pages which follow we shall endeavor to give some of the engineering considerations upon which the design of the present receiving system was based.

Choice of Frequency

In the early development of long-distance radio telegraphy, the strength of the received signal was the principal factor upon which the selection of the operating frequency was based. After the development of the vacuum-tube amplifier, however, the following considerations each became important, especially so for a telephone circuit:

[1] Published in *Proc. of I. R. E.*, Dec., 1928, pp. 1645–1705.

1. The signal-to-noise ratio at the receiving location, which in turn is dependent upon four factors—
   (*a*) The efficiency of the transmitting set,
   (*b*) The efficiency of the transmitting antenna,
   (*c*) Attenuation in the radio path,
   (*d*) Variation of radio noise with frequency;
2. Band width of the transmitting antenna;
3. Receiving antenna efficiency;
4. Available space in the frequency spectrum.

1. *Signal-to-Noise Ratio at the Receiving Location.* At the time that the transatlantic radio-telephone development was undertaken, engineers of the Western Electric Company Engineering Department (now Bell Telephone Laboratories) had developed a form of water-cooled vacuum tube capable of generating efficiently large amounts of power at any frequency up to perhaps several hundred kilocycles.[2] Therefore transmitter efficiency, although a major problem in itself, imposed no restriction on the frequency for the telephone circuit.

For transmission over a given path, utilizing a particular transmitting antenna with constant power supplied to it, there will be, in general, a frequency at which the greatest signal-to-noise ratio is obtained. To illustrate this point, we have chosen the problem of transmission from an antenna of the type used at the Rocky Point station of the Radio Corporation of America in U. S. A. to a receiving station in England, a distance of approximately 5,000 kilometers. The approximate variation with frequency of loss resistance, radiation resistance, and efficiency of this antenna is shown in Fig. 1. The loss resistance at 60 kilocycles was determined by engineers of Bell Telephone Laboratories, while the data in the lower frequency range were published by Alexanderson, Reoch, and Taylor.[3] The radiation resistance was calculated from the measured effective height of the antenna. It is seen in Fig. 1 that the antenna efficiency increases with frequency throughout the range we are considering, first rapidly and then more slowly.

For a constant power radiated, radio attenuation tends to cause a decrease in the average received signal strength as the frequency is increased. This effect is in the opposite direction to the effect of antenna efficiency, so that for a given power supplied to the antenna the field strength at a given distance will be a maximum at a certain

[2] W. Wilson, "A New Type of High Power Vacuum Tube," *Bell System Tech. Jour.*, *1*, 4; July, 1922. *Elec. Comm.*, *1*, 15; August, 1922.

[3] E. F. W. Alexanderson, A. E. Reoch, and C. H. Taylor, "The Electrical Plant of Transocean Radio Telegraphy," *Trans. A. I. E. E.*, *42*, 707; July, 1923.

frequency. In Fig. 2 we have shown the calculated field strength at 5,000 kilometers for a power of 85.9 kilowatts supplied to the Rocky Point antenna, using efficiency data of Fig. 1 and the radio transmission formula given by Espenschied, Anderson, and Bailey.[4] Since this curve reaches a maximum near 18.5 kilocycles, the reason for the operation of early transatlantic radio-telegraph circuits in the range 10 to 30 kilocycles becomes apparent in light of the limitation then placed on the receiving systems.

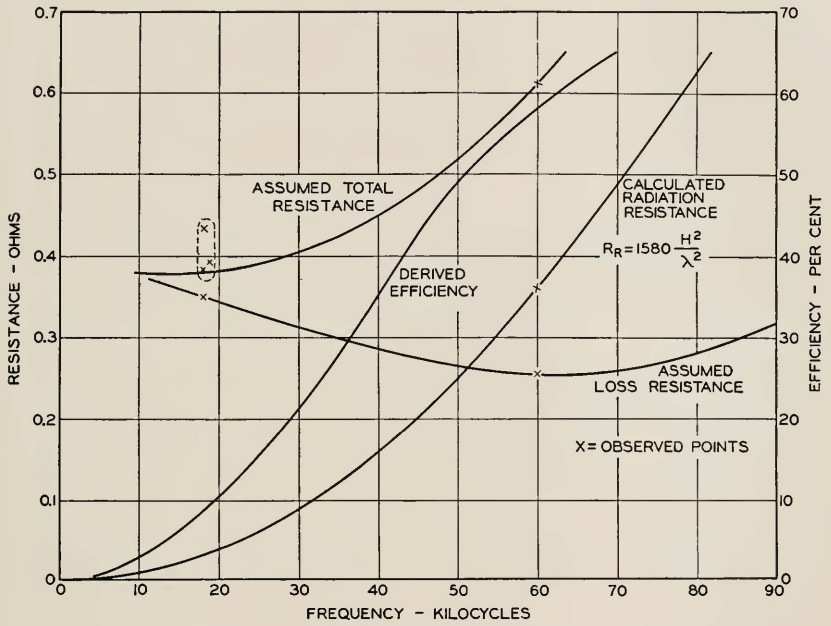


Fig. 1—Assumed resistance and efficiency of Rocky Point antenna. (Effective height 75 meters.)

Systematic measurements of radio noise by the warbler method,[5] begun early in 1923, have yielded important information on the variation of noise with frequency.[4] From measurements begun by engineers of Bell Telephone Laboratories and continued by engineers of the International Western Electric Company at New Southgate, England, during 1923 and 1924, the average daylight noise curve, in Fig. 2, was obtained. It is seen that the noise decreases with increasing

[4] Lloyd Espenschied, C. N. Anderson, and Austin Bailey, "Transatlantic Radio Telephone Transmission," *Bell System Tech. Jour.*, *4*, 459; July, 1925. *Proc. I. R. E.*, *14*, 7; Feb., 1926.

[5] Ralph Bown, C. R. Englund, and H. T. Friis, "Radio Transmission Measurements," *Proc. I. R. E.*, *11*, 115; April, 1923.

frequency, at first rapidly and then more slowly, being almost constant after passing the frequency of 40 kilocycles.

From the values of signal and noise so obtained, the signal-to-noise ratio has been computed, and is also plotted in Fig. 2. The curve of signal-to-noise ratio reaches a maximum near 44 kilocycles which would seem to be the optimum frequency for daylight transmission from the Rocky Point station to England. This is not strictly the case, however, since there is some evidence that a phenomenon exists which makes frequencies in the vicinity of 40 kilocycles particularly poor for the transatlantic path. Data published by Anderson [6] tend

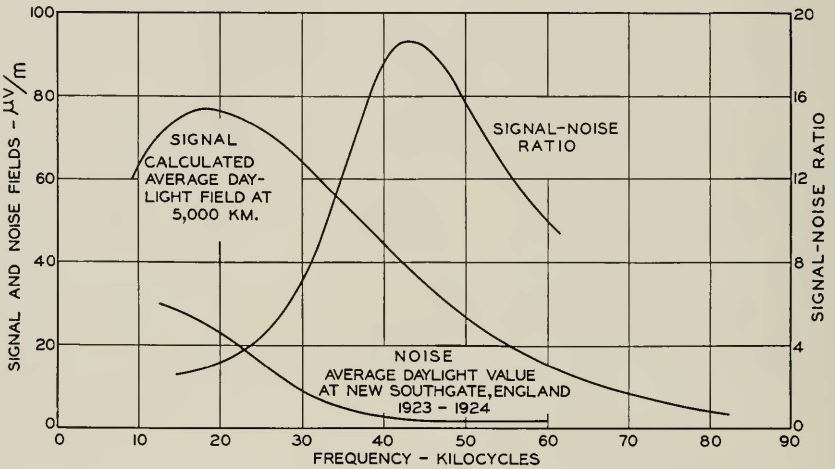


Fig. 2—Variation of signal, noise, and signal-noise ratio with frequency. Transmission from U. S. A. to England. 85.9 kw. supplied to antenna of Rocky Point characteristics.

to show that the field strength is distinctly subnormal in the vicinity of 44 kilocycles and remains approximately constant from that frequency up to about 60 kilocycles, where the observed values agree fairly well with the calculations. (See later in this paper.)

2. Band Width of the Transmitting Antenna. Since the output of the transmitting set is at a high power level, the circuits coupling it to the antenna must be of the simplest type to reduce the loss to a minimum. In view of this requirement, the antenna constants largely determine the band width of the antenna system. At frequencies much lower than 60 kilocycles it was not possible to secure a sufficient width of band even for commercial telephony from the Rocky Point

[6] C. N. Anderson, "Correlation of Long Wave Transatlantic Radio Transmission with other Factors Affected by Solar Activity," *Proc. I. R. E., 16*, 297; March, 1928. In connection with reference above see Fig. 19, p. 315.

antenna, but at this frequency reasonably satisfactory results are obtained.

3. Receiving Antenna Efficiency. The use of directional receiving antennas is essential to satisfactory and economic results over such distances as the transatlantic radio path (see later in this paper). The directivity of an antenna system of a given kind, size, and cost in general increases with frequency, since the directivity is a direct function of the ratio of the dimensions of the antenna system to the wave-length employed.

4. Available Space in the Frequency Spectrum. Each of the above factors operates to make the frequency of 60 kilocycles about the best which could be used in the present state of the art for this transmission path. Fortunately this frequency was so located in the radio spectrum that a band of the desired width free from interference could be obtained.

It has been noted that the radio noise as shown in Fig. 2 varies very little with frequency above 40 kilocycles. There is some doubt as to whether or not this accurately represents the actual state of affairs, since the measurement sets used for measuring the noise would not satisfactorily measure much below one microvolt per meter on account of tube noise. At frequencies of 40 kilocycles and above, especially in the winter, there are many days during which the radio noise is practically absent. On these days the measurements tended to approach the minimum determined by the set noise. The fact that many such readings were incorporated in the average probably tends to mask the true variations of radio noise with frequency in this range. On the other hand, however, they indicate a very real limitation which tends to operate against the use of frequencies higher than about 60 kilocycles unless fields were increased by increase in transmitting power. This would be particularly true during the sunset and sunrise dips and during periods of abnormally poor transmission when the fields fall much below the average. If the set noise limitation could be removed it is quite possible that frequencies above 60 kilocycles would become more useful. Higher frequencies for radio telephone use would be particularly advantageous because of the greater band width which could be obtained from the transmitting antenna and because of the greater directivity which could be obtained in the receiving system at the same cost.

Selection of a Satisfactory Receiving Location

The selection of a suitable receiving location is based upon three major considerations; namely, maximum received signal-to-noise ratio,

reasonably suitable terrain for receiving antenna construction, and adequate wire connection facilities between the location and the more densely populated areas.

Since about 10 per cent of the populations of the United States and the British Isles are located within a radius of 40 miles of New York and London, respectively,[7] it was natural to decide upon making those cities the terminal points. It would hence be desirable to locate the receiving stations near and with good wire circuits to those cities.

Very early in the history of radio communication [8] it was, however, realized that in the United States a decrease of radio noise was obtained by a northerly location of the receiving station and, for receiving from European stations, the northern location is further advantageous, since higher field strengths result from the reduced transmission distance. The Radio Corporation of America had already taken advantage of this improvement by locating a receiving station at Belfast, Maine.

To obtain quantitative information on this matter, the American Telephone and Telegraph Company made comparative measurements of noise as received on loop antennas at Riverhead, New York; Green Harbor, Massachusetts; and Belfast, Maine; the loops were so oriented as to give maximum receptivity in the direction of England. Although these tests were only continued for a few months at each location, they left no doubt that the absolute level of the noise was less at the northerly locations.

In Fig. 3, there is shown the diurnal variation of improvement in noise conditions (in TU) for average days of each month at Belfast over Riverhead. The average hourly improvement was determined by averaging the ratios of practically simultaneous observations of noise at the two locations for each hour during any one month and taking a three-hour moving average of the result to reduce the effect of purely local phenomena at either of the two stations. The data for the two half years were taken on slightly different frequencies as is indicated on the figure. Unfortunately, during the month of July only two weeks data were taken on each of the frequencies, namely, 52 and 65 kilocycles, and these data were taken a year apart, namely in 1924 and in 1925. In order to give some idea of the location noise improvement for the month of July we have averaged in the same way the four weeks data thus obtained, and plotted the result as a broken line. Fortunately, the improvement of the more northerly location is, in general, large during the overlapping business day of England and the United States.

[7] "New York's New 10,000,000 Zone," *Literary Digest*, *95*, 12, p. 14; Dec. 17, 1927.

[8] G. W. Pickard, "Static Elimination by Directional Reception," *Proc. I. R. E.*, *8*, 358; October, 1920.

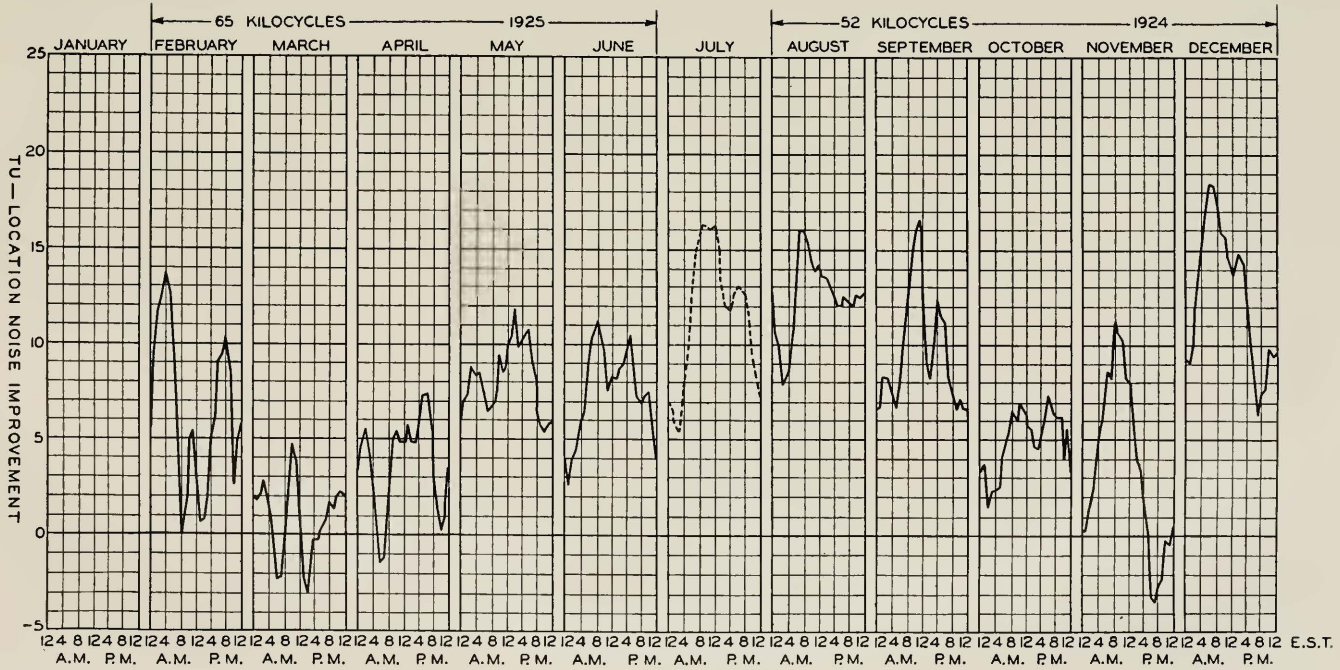


Fig. 3—Transatlantic radio noise measurements. Diurnal variations of location noise improvement (in transmission units) of Belfast, Maine, over Riverhead, New York. Three-hour moving averages of simultaneous observations.

It is apparent that the improvement is a maximum in the middle of the summer when the noise is high, and in the middle of the winter when the field strengths are usually abnormally low. This is important, since the greatest improvement is needed at each of these times.

The monthly averages of variations of noise and of signal have previously been published,[4, 6, 9] and the generalizations given above can be confirmed by reference to these articles.

For calculating daylight radio transmission, several formulas have been proposed.[10, 11, 12] In Fig. 4 the heavy curve was calculated

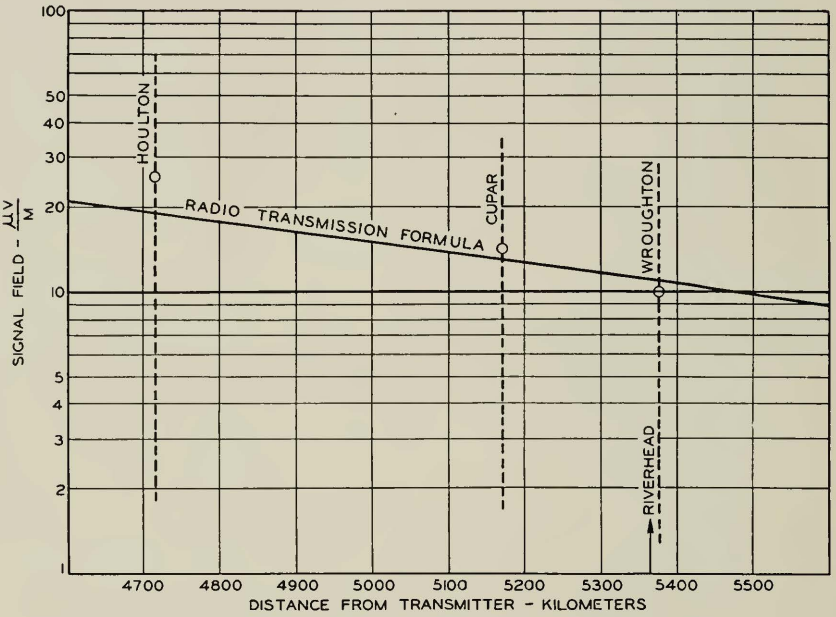


Fig. 4—Transatlantic radio daylight field strength. Average of hourly observations during 1927. Corrected to 50 kw. radiated power—frequency 60 kilocycles

from the empirical formula given by Espenschied, Anderson and Bailey,[4] and assumes a radiated power of 50 kilowatts. The great circle distance from the transmitting stations used in the transatlantic radio-telephone circuit to various receiving stations is indicated by the name of the receiving station. The average of daily averages

[9] Ralph Bown, "Some Recent Measurements on Transatlantic Radio Transmission," *Proc. Natl. Acad. of Sci.*, *9*, 221; July, 1923.

[10] A. Sommerfeld, "Ueber die Ausbreitung der Wellen in der drahtlosen Telegraphie," *Ann. d. Phys.*, *28*, 665; 1909.

[11] L. F. Fuller, "Continuous Waves in Long-Distance Radio Telegraphy," *Trans. A. I. E. E.*, *34*, pt. 1, 809; 1915.

[12] L. W. Austin, "Quantitative Experiments in Radiotelegraphic Transmission," *Bull. Bureau of Std.*, *11*, 69; Nov. 15, 1914.

of hourly measurements of the field strength made at Houlton and Wroughton during the time that the transatlantic path was entirely in daylight during 1927 is indicated by points on this figure. The data for Cupar are less complete since this station was not in regular daily operation until May, 1927. The range of variation between the maximum daily average and the minimum daily average for each receiving location is given by the limits of the dotted vertical line. (It is interesting to note that at a frequency of 60 kilocycles and for distances in the order of 5,000 kilometers any of the radio-transmission formulas referred to above will give a computed value lying within the range of variation of average daylight readings.)

The improvement in signal-to-noise ratio obtained by locating the receiving station in Maine instead of in New York is easily seen by reference to Figs. 3 and 4. The improvement due to decrease of noise, during that time of year when improvements are most needed on account of high noise values, is about 10 TU. The improvement due to increase of the average received daylight signal by decrease of the distance is calculated to be 5 TU. During 1927, this improvement was actually observed to be 8 TU. We may, therefore, state in round numbers that the total improvement realized by locating the receiving station in Maine instead of New York was equivalent to a fifty-fold increase of the power radiated by the British transmitting station.

The British General Post Office, during 1926, carried out a set of measurements of field and noise at various locations in the United Kingdom. Those tests led them to the same conclusions as regards the advantage to be obtained by locating their receiving station at some more northerly point.[13] They decided upon a location near Cupar, Scotland, and comparisons made daily from 1230 to 2300 GMT indicate that this location is better for receiving than Wroughton, England. The geometric mean of the improvement in signal-to-noise ratio for the more northerly location during the months May to September, 1927, inclusive, and for the daily period given above is 6.4 TU. This is equivalent to an increase of between four and five times in power from the American transmitting station.

Since such relatively large improvements were to be obtained by northerly locations of the receiving station it seemed best to take advantage of this fact and locate the receiving station in America at some place in the state of Maine. This decision led to further consideration of two factors mentioned above, namely, reliable wire

[13] A. G. Lee, "Wireless Section: Chairman's Address, *Jour. I. E. E.*, *66*, 12; Dec., 1927.

connections to New York and a suitable terrain for antenna construction. The first of these factors required a location along one of the main telephone trunk routes in Maine and the second, since we had decided upon the use of a wave-antenna [14] for reasons which will be given in the following section, demanded a rather large and reasonably flat land area available for pole-line construction. A location, although not altogether ideal, was decided upon near Houlton, Maine, about six miles from the Canadian border.

Choice of Receiving Antenna Systems

The number of fundamental types of receiving antennas that may be employed for long-wave reception is quite definitely limited. In fact all of the known practical receiving antennas may be considered as falling into one of three principal classes of structure; i.e., the vertical antenna, the loop or coil antenna, and the wave-antenna. The selection of the proper receiving antenna system quite evidently becomes a problem—first, of choosing the best type of antenna from one of these three classes and, second, of choosing a particular antenna structure in the class which is found to be best.

The factors governing the choice of a receiving antenna are as follows:

1. Directional Discrimination Against Static. Inasmuch as the signal to be received has a definite average value, the receiving system can only better the circuit in the amount that it improves the signal-to-noise ratio. A directional antenna system affords a means of reducing the received noise in relation to the desired signal.[8, 14] The directional characteristics of the principal antenna types are shown in Fig. 5.

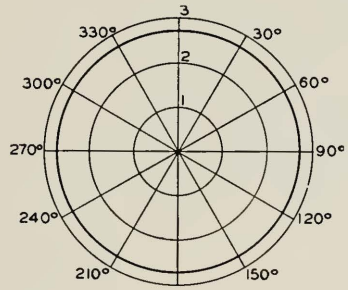
A measure of the directional discrimination of the various antenna types is the Noise Reception Factor (abbreviated *NRF*) which is defined as the ratio of the total noise current received from the antenna in question to that received from a vertical antenna under the conditions of continuous, constant distribution of noise sources about the antenna and of equal output currents for signals from the direction of maximum receptivity. The back end *NRF* is the noise reception factor for the arc between 90 degrees and 270 degrees from the direction of maximum receptivity.

On this basis, the choice rests quite unmistakably with the wave-antenna.

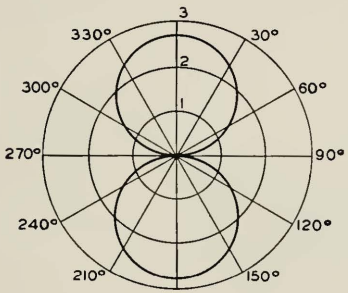
2. Transmission-Frequency Characteristic. Since the receiving antenna is to be used on a system for communication by speech, necessi-

[14] H. H. Beverage, C. W. Rice and E. W. Kellog, "The Wave Antenna," *Trans. A. I. E. E.*, *42*, 215; 1923.

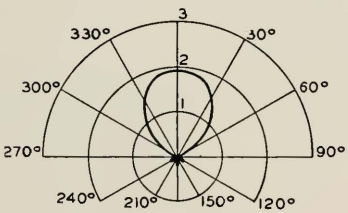
tating the transmission of a relatively wide band of frequencies, it must pass such a band without undue discrimination against any frequency contained therein. To utilize the vertical and the loop antennas



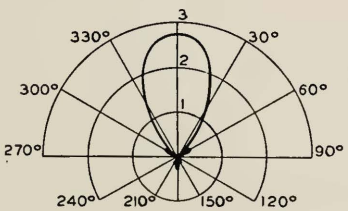
Vertical Antenna
Total N.R.F. = 1.000
Back End N.R.F. = 1.000



Loop Antenna
Total N.R.F. = 0.707
Back End N.R.F. = 0.707



Wave-Antenna
(One Wave-Length Long)
Total N.R.F. = 0.435
Back End N.R.F. = 0.058



Wave-Antenna
(Two Wave-Lengths Long)
Total N.R.F. = 0.361
Back End N.R.F. = 0.048

Fig. 5—Comparison of polar diagrams of simple antennas. (The unit for the radii is output current into the same resistance for a constant impressed field.)

efficiently, it is desirable that they be tuned, introducing the frequency discrimination of a tuned circuit. If these types of antennas be used

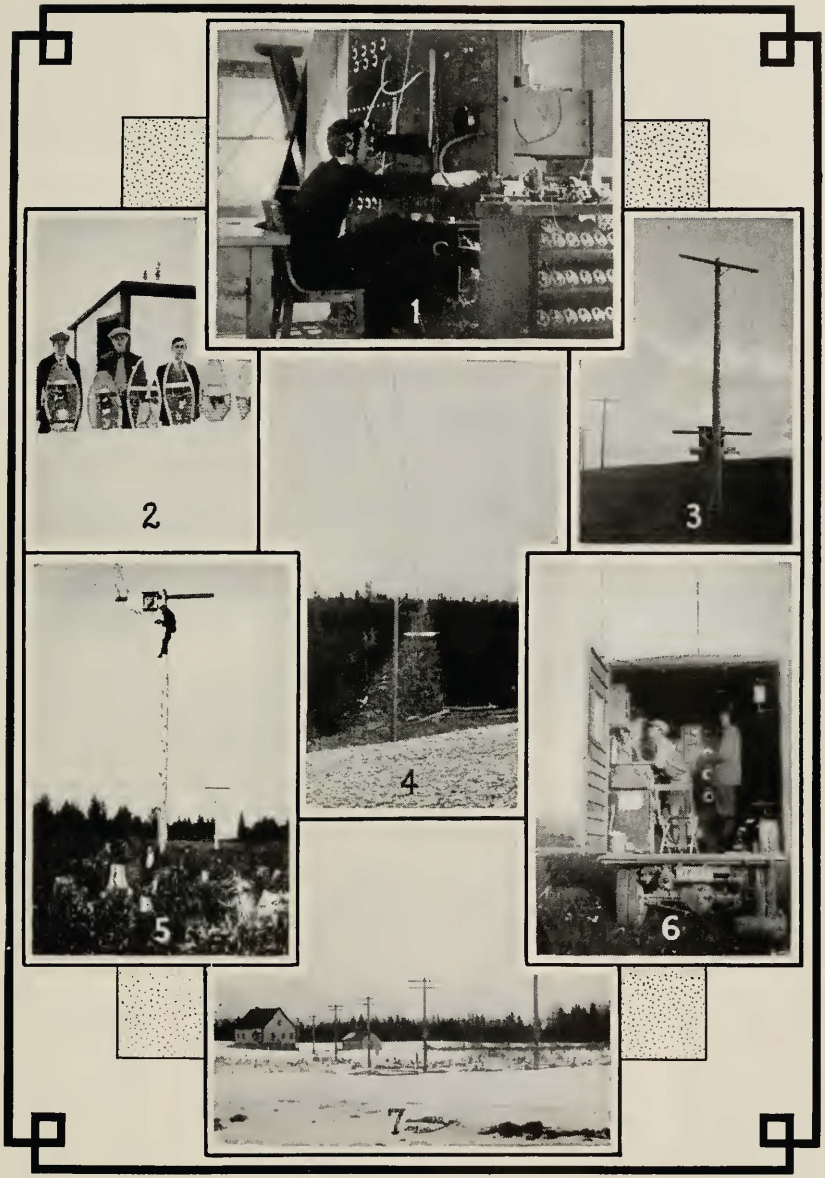
it is necessary, therefore, that the resonance characteristic be studied and means provided to eliminate excessive frequency discrimination within the desired band. On the other hand, the wave-antenna is an aperiodic structure and, in consequence, its transmission-frequency characteristic is so flat that it need not be considered.

3. Sensitivity. There are two factors which require that the output from the receiving antenna for a given field strength be as large as possible. First, if the receiving station be located at any position other than that at the terminal of the antenna, which is necessarily the case if more than one antenna be used in an array, the signal on the transmission line from the antenna to the station must be much greater than the noise currents induced into the transmission lines. If the antenna output be excessively small, it is impossible to balance the transmission lines so completely that this requirement is met. Second, the amount of gain that can possibly be used at the radio receiver is ultimately limited by the noise produced in an amplifier. (This is discussed more fully under "Power Output Required from the Radio Receiver" later in this paper.) To the first approximation, the sensitivity of each of the antenna classes under consideration is a direct function of its physical dimensions. There is, however, a limit to the sensitivity of each antenna class, for mechanical limits govern the maximum size of a vertical antenna, distributed capacity and mechanical considerations limit the loop, and in the wave-antenna a restriction occurs because of the peculiarity that the sensitivity reaches maximum values at definite lengths.

Since cost is likewise a factor governing the ultimate selection of an antenna system, the sensitivities may well be compared for antennas of equal cost. On this basis, a loop or a vertical antenna of effective height of fifty meters is directly comparable with a wave-antenna one wave-length long. By reference to Fig. 5, where the scale is the same for all the directional diagrams, it becomes evident that the sensitivities of all three classes of antennas are of the same order of magnitude, being slightly greater for the vertical antenna and the loop than for the one-wave-length wave-antenna.

4. Stability. The sensitivity and frequency-transmission characteristics of the antenna must be substantially constant during changes of weather and seasonal conditions. The antenna classes which require tuning are slightly poorer than the wave-antenna in this respect.

5. Reproducibility. Further improvement in directional discrimination against noise is obtained by using several similar antennas in an array. The loop and the vertical antennas probably are best for



1—Measuring field strength.
2—Outside an antenna terminal hut.
3—Pole box for reflection transformer.
4—The wave-antenna *A* at Houlton.
5—Measuring ground connection impedance at a temporary location.
6—The sixty kilocycle portable transmitting station.
7—Transmission line O-B with receiving station in background.

combining in arrays because several of either type of antenna can be made identical with one another. Wave-antennas combined in an array, however, give satisfactory results.

Although each of these factors governing the choice of the receiving antenna system is important, their relative importance is indicated by the order in which they have been presented. In view of the low noise reception factor of the wave-antenna, its lack of frequency discrimination, and its inherent stability, the wave-antenna was selected for the fundamental type of antenna to be used at the receiving station at Houlton.

The Wave-Antenna

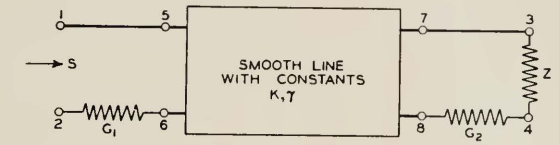
Among the types of antennas which may be considered for use in long-wave radio communication, the wave-antenna [14] possesses several characteristics which single it out as being unique. The most important of these are:

1. The length of a wave-antenna is directly comparable to and of the same order of magnitude as the wave-length of the signals for which it is designed.
2. Considering the straight horizontal wire comprising the wave-antenna as a grounded transmission line, a termination, equal to the characteristic impedance, is applied to each end of that line. The wave-antenna then becomes an essentially aperiodic antenna.
3. The major response of a properly designed wave-antenna is to the horizontal component of the impressed electric field. The propagated electric wave must therefore have an electric component parallel to the surface over which the wave-antenna is constructed.
4. On the basis of the preceding consideration, the design of a wave-antenna definitely excludes elevation of the antenna above ground to any extent greater (a) than is physically necessary to provide safe clearance and (b) than that height where the loss in the antenna considered as a transmission line reaches a nominal value. Practically, the wave-antenna is constructed as a high-grade telephone line, on 30-foot poles.

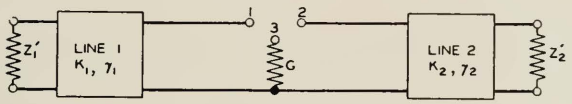
It is evident that the major electrical characteristics which distinguish the wave-antenna are intimately connected with the character of the surface over which the antenna is built, and with the details of construction of the wave-antenna. The performance of a wave-antenna at any specified location then can only be determined by constructing such an antenna and measuring its constants. The measurements made in determining the characteristics of any particular wave-antenna are outlined in the following paragraphs.

1. Ground-Connection Impedance. It is shown in Appendix 1 that the wave-antenna is considered to be a smooth line with uniformly distributed constants. This assumption is met to a sufficient degree in practice, but, unfortunately, it is impossible to connect to the four terminals of the practical line, since the connections to the ground side of the line must be made by burying wires in the earth rather than

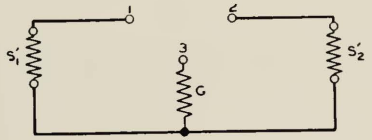
connecting to a discrete terminal which is the real ground. As is shown in Fig. 6a, the actual wave-antenna may still be considered as a smooth line, but between the terminals of the wave-antenna and the terminals that are available at the physical ends of the wave-antenna ground-connection impedances exist. To determine the constants of the wave-antenna, these impedances must be evaluated and taken into account as follows: In Fig. 6a, an impedance Z is applied to the avail-



(a)



(b)



(c)

Fig. 6.

able terminals of the wave antenna 3–4 and the impedance S measured at the available terminals 1–2; under this condition, the actual terminal and input impedances of the wave-antenna are respectively:

$$Z' = Z + G_2 \tag{1}$$

$$S' = S - G_1 \tag{2}$$

where G_1 and G_2 are the ground-connection impedances at the two ends of the antenna.

Figs. 6b and 6c illustrate the method that was used to determine the ground-connection impedance. In Fig. 6b, lines 1 and 2 represent two smooth ground-return transmission lines extending in opposite directions from the ground connection for ½ kilometer or more, the lines being terminated at the distant end in impedances Z_1' and Z_2', respectively. In practice one of these lines was the wave-antenna

and the other a temporary line of insulated wire laid along the surface of the ground.

For the purpose of analysis, each of the lines may be replaced by its input impedance. This simplification is shown in Fig. 6c, where

$$S' = \frac{K \tanh \gamma s + Z'}{1 + \frac{Z'}{K} \tanh \gamma s}. \tag{3}$$

The impedance between terminals 1 and 3 is:

$$S_1 = S_1' + G. \tag{4}$$

The impedance between terminals 2 and 3 is:

$$S_2 = S_2' + G. \tag{5}$$

The impedance measured between terminals 1 and 2 in parallel and terminal 3 is:

$$S_0 = G + \frac{S_1' S_2'}{S_1' + S_2'}. \tag{6}$$

Eliminating S_1' and S_2' from equations (4), (5), and (6) and solving for G:

$$G = S_0 - \sqrt{\frac{1}{2}[(S_1 - S_0)^2 + (S_2 - S_0)^2 - (S_1 - S_2)^2]}. \tag{7}$$

By building out either line 1 or line 2 with added series impedances until

$$S_1 = S_2 = S_{12} \tag{8}$$

the expression for the ground-connection impedance simplifies greatly, and incidentally the precision of the determination becomes greater because the number of measurements involved is less. Under this condition

$$G = 2S_0 - S_{12}. \tag{9}$$

This latter case is the one that was actually used in measuring the ground impedances.

Since the distribution of ground currents about the buried ground may be different under each of the three conditions that are measured, there is undoubtedly some error in measuring the ground-connection impedance by this method. This error is a second-order effect, however, so that the values determined are reliable within the precision

that the method allows, involving as it does, differences between measurements of high-frequency impedance.

All of the impedance measurements were made using a high-frequency bridge designed and constructed by Mr. C. R. Englund of Bell Telephone Laboratories. This bridge is similar to that described by Shackelton [15] except that the standards used consist of a calibrated condenser and a decade resistance. Impedances having capacitive reactance are measured by direct comparison with the standards, while impedances having inductive reactance are tuned with the standard condenser to parallel resonance and the resonant combination compared with the decade resistance. Impedances involving extremely small reactances, either positive or negative, are built out with a condenser in parallel to a value that may be measured conveniently.

2. Characteristic Impedance and Propagation Constant. Since the early days of transmission line study, the characteristic impedance and the propagation constant have been determined by two impedance measurements at the near end of the line with the far end of the line open- and short-circuited, respectively.[16] For two reasons, this method has not been used in our determination of the fundamental antenna constants: first, it is impossible to apply a short to the real terminals of the wave-antenna due to the presence of the ground-connection impedance; and, second, with lines multiple quarter wave-lengths long the input impedance, as a result of resonance in the line when it is open-circuited or grounded, attains either extremely large or extremely small values which could not be measured accurately with the available testing equipment.

To obviate these difficulties, Mr. C. R. Englund, of Bell Telephone Laboratories, developed a method of determining the characteristic impedance and the propagation constant of the wave-antenna by measuring the input impedance with two known finite terminations at the far end. Under this condition it may be shown that the characteristic impedance is given by the expression:

$$K = \sqrt{\frac{(S_1 - G_1)(S_2 - G_1)(Z_1 - Z_2) + (Z_1 + G_2)(Z_2 + G_2)(S_2 - S_1)}{(S_2 - S_1) + (Z_1 - Z_2)}} \quad (10)$$

and that the propagation constant is given by:

[15] W. J. Shackelton, "A Shielded Bridge for Inductive Impedance Measurements at Speech and Carrier Frequencies," *Bell System Tech. Jour.*, *6*, 142; Jan., 1927.

[16] Bela Gati, "On the Measurement of the Constants of Telephone Lines," *The Electrician*, *58*, 81, Nov. 2, 1906.

$$\gamma = \frac{1}{s}\tanh^{-1}\left[K \frac{(S_2 - S_1) + (Z_1 - Z_2)}{(Z_1 + G_2)(S_1 - G_1) - (Z_2 + G_2)(S_2 - G_1)} \right], \quad (11)$$

where the symbols in equations (10) and (11) have the following meanings:

Z_1 = the first termination applied to the available terminals at the far end ot the line (ohms)

Z_2 = the second termination applied to the available terminals at the far end of the line (ohms)

S_1 = the impedance measured at the available near-end terminals corresponding to the termination Z_1 (ohms)

S_2 = the impedance measured at the available near-end terminals corresponding to the termination Z_2 (ohms)

G_1 = the ground-connection impedance at the near end of the line (ohms)

G_2 = the ground-connection impedance at the far end of the line (ohms)

s = length of the line (kilometers)

K = characteristic impedance (ohms)

γ = propagation constant (hyps per kilometer)

3. Effective Height. The effective height of a wave-antenna is defined as the ratio of the voltage produced at any specified point in the antenna to the potential gradient of the electromagnetic field producing that voltage. If the constants of the antenna system are known, the effective height at any point in the antenna system may be calculated from the value at any other point in the system.

A convenient way to measure an effective height of a wave-antenna and obtain a value which may be easily correlated with wave-antenna theory is to introduce in series with the initial-end terminating impedance a voltage which produces the same output current from the antenna as is produced by an electromagnetic wave. The ratio of this induced voltage to the potential gradient of the electromagnetic field has been called "the effective height referred to the characteristic impedance." For small values of the quasi-tilt angle, the total potential gradient of the electric field is very closely equal to the vertical component of the electric field, so that within the precision of measurement we may write:

$$H_\theta = \left| \frac{E_K}{E'} \right|, \quad (12)$$

where

H_θ = Effective height of the wave-antenna referred to the characteristic impedance (kilometers)

E' = The potential gradient of the vertical component of the impressed field (volts per kilometer)

E_K = The electromotive force introduced in series with the characteristic impedance at the initial end of the wave-antenna producing the same current at the distant end as the impressed field (volts)

4. Quasi-tilt Angle and Ground Resistivity. The measured effective height of a wave-antenna is a function of four constants:

1. The length of the antenna;
2. The height of the antenna;
3. The propagation constant of the antenna;
4. The ratio of the component of the electric wave parallel to the surface over which the antenna is constructed to the vertical component of the electric wave.

In general, the first three of these constants are different in value for antennas constructed at different locations, but they may be varied over a limited range by changing the construction and dimensions of the wave-antenna. The comparison of effective heights, therefore, does not readily yield information regarding the relative suitability of various locations for wave-antenna systems. The ratio of the horizontal component to the vertical component of the impressed field is, however, a constant whose value is dependent solely upon the ground conditions at the location (assuming a fixed frequency for the comparison).

In case the time phase between the horizontal component and the vertical component of the impressed field were zero, the ratio of these two components would represent the tangent of the angle of forward inclination of the propagated wave front. In general, the phase angle between the two components is not zero, so that such a simple relation does not hold. It is convenient, however, to call the ratio of the two components of the impressed field the tangent of the "quasi-tilt angle," where the "quasi-tilt angle" becomes the real tilt angle in the limiting case.

In terms of the effective height, the antenna constants, and the vertical component of the impressed field, the current produced at the far end of the wave-antenna is (using the nomenclature of Appendix 1 and to the same degree of approximation as equation (12)):

$$|I_\theta| = \left|\frac{H_\theta E' \epsilon^{-\gamma S\lambda'}}{2K}\right| \tag{13}$$

or

$$|I_\theta| = II_\theta \epsilon^{-\alpha S\lambda'} \left|\frac{E'}{2K}\right|. \tag{14}$$

In terms of antenna constants alone, it is shown in Appendix 1 that the current produced at the far end of the wave-antenna is:

$$|I_\theta| = |I_{E'\theta} + I_{F'\theta}|, \tag{125}$$

where

$$I_{E'\theta} = -\frac{1}{\epsilon^{j\delta} \tan T} \frac{S\lambda' F}{2K}(a + jb), \tag{15}$$

$$I_{F'\theta} = \frac{S\lambda' F'}{2K}(c + jd), \tag{16}$$

also

$$-\frac{F'}{E'} = \epsilon^{j\delta} \tan T. \tag{301}$$

In equations (15) and (16), $(a + jb)$ and $(c + jd)$ are abbreviations defined as follows:

$$(a + jb) = \frac{h}{S\lambda'}(1 - \epsilon^{-[\alpha S\lambda' + j2\pi S(m - \cos\theta)]})\epsilon^{-j2\pi S\cos\theta} \tag{17}$$

and

$$(c + jd) = \cos\theta \frac{1 - \epsilon^{-[\alpha S\lambda' + j2\pi S(m - \cos\theta)]}}{\alpha S\lambda' + j2\pi S(m - \cos\theta)} \epsilon^{-j2\pi S\cos\theta}. \tag{18}$$

If we equate expressions (14) and (125), and solve for $\tan T$:

$$\tan T = \frac{ac + bd + \sqrt{(ac + bd)^2 - (c^2 + d^2)(a^2 + b^2 - g^2)}}{c^2 + d^2}, \tag{19}$$

where

$$g = \frac{II_\theta}{S\lambda'} \epsilon^{-\alpha S\lambda'}. \tag{20}$$

It is pointed out in Appendix 3 that the phase angle δ may be expressed as a function of the quasi-tilt angle T and the dielectric constant k. For that reason, the determination of T must be made in two steps. The procedure is as follows: first, it is assumed that the component of the total received current due to the vertical component of the impressed field is zero, i.e.,

$$(a + jb) \equiv 0.$$

Under this condition:

$$T = \tan^{-1} \frac{g}{\sqrt{c^2 + d^2}}. \tag{21}$$

Using Fig. 20 of Appendix 3, the value of δ corresponding to this value of T is determined (generally $\delta = \pi/4$). Second, T is revaluated from (19) using the value of δ so obtained.

The ground resistivity is evaluated from the value of the quasi-tilt angle by using Fig. 20 of Appendix 3.

5. Directional Characteristics. The measurement of the directional characteristics of a wave-antenna or a wave-antenna system consists entirely of measuring the effective height of the antenna for several directions of wave propagation, and determining the relative directional receptivity of the antenna in these directions by dividing the effective height for each direction by the value obtained for the direction of the axis of the antenna. For this purpose, the effective height at the output of the antenna system is most convenient to measure and use. This constant is defined as the ratio of the voltage at the input of the radio receiver to the field strength producing this voltage. It is exactly related to the effective height referred to the characteristic impedance (defined in the preceding subsection of this paper) by the real part of the total transfer constant between the termination at the initial end of the antenna and the input terminals of the radio receiver, and an additional factor of one-half because the voltage at the radio receiver is measured across the proper termination.

In certain receiving station locations, it is possible to utilize for determining the relative directional receptivity the regular transmission from existing radio transmitters operating at or very close to the frequency for which the directional characteristic is desired. At sites less favorably located with regard to existing transmitters, the directional characteristic may be measured by transmitting test signals from a portable transmitter, located successively in the several directions for which data are desired, and at least 15 wave-lengths from the antenna system.

A distinctly different method of measuring the directional characteristics of an antenna is based on a statistical study of the reduction of noise obtained by its use. While it is difficult to evaluate the directional characteristic exactly by this method, data showing the comparative decrease in noise with the wave-antenna as against a loop or a vertical antenna are of great value in predicting the improvement in a radio circuit to be obtained by its use. As a converse to these results, the statistical combination of the improvement given by the

wave-antenna, and a measured directional diagram, yields information on the direction of arrival of static.

Data on wave-antenna characteristics have been taken at several widely separated locations. Two antenna systems have been constructed by the British General Post Office—one at Wroughton, Wiltshire, in southern England, and one at Cupar, Fifeshire, in southeastern Scotland. We likewise have data on our antenna system at Houlton, Maine. The character of the earth under each of these antenna systems is different, resulting in widely different quasi-tilt angles and antenna directional characteristics.

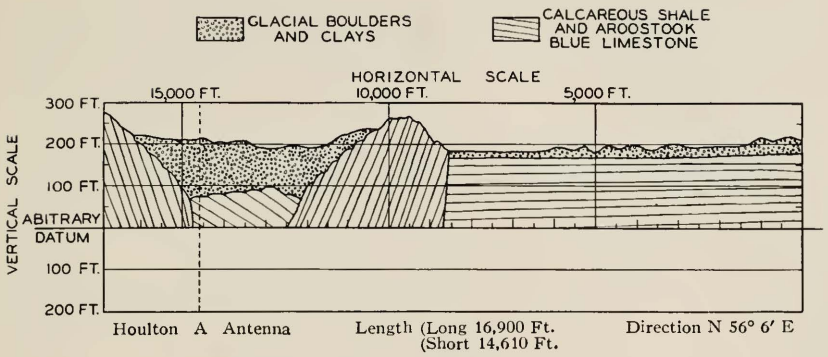
The probable geological formations under individual antennas at each of the three antenna sites mentioned in the preceding paragraph are shown in Fig. 7. The data for the British locations were compiled from the published reports of geological surveys conducted by the British Government, and the data for the Houlton location were taken from the "Soil Survey of the Aroostook Area, Maine," published by the U. S. Department of Agriculture. In Table I, the ground constants are given for these three locations, determined by the method given in Section 4, "Quasi-tilt Angle and Ground Resistivity":

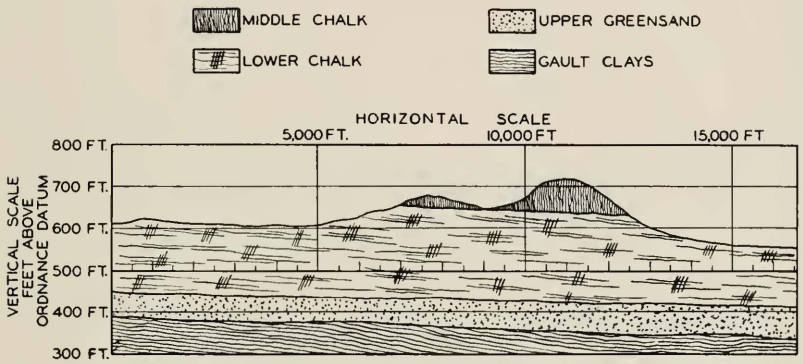
TABLE I

Location	Characteristic Sub Soil	Quasi-Tilt Angle at 60 kc Radians	Ground Resistivity Ohms per cm^3
Wroughton	Chalk	0.011	3630.
Cupar	Sandstone	0.017	8670.
Houlton	Limestone	0.047	66300.

Fig. 8 shows the directional characteristics, calculated by the method given in Appendix 1, of wave-antennas erected over the geological formations shown in Fig. 7. In these directional characteristics, it is important to notice that a decrease in quasi-tilt angle increases the relative importance of the component of the received current due to the vertical component of the impressed field (abbreviated to the "vertical effect"). It is evident from Fig. 8 that the relative directional receptivity for the arc between $\theta = 90$ degrees and $\theta = 270$ degrees is smaller and that the effective height is much greater for the antenna at Houlton, for which the ground resistivity is higher than for the other two antennas.

Measured relative directional receptivities are also shown in Fig. 8. The values for the Cupar antenna were determined by using trans-





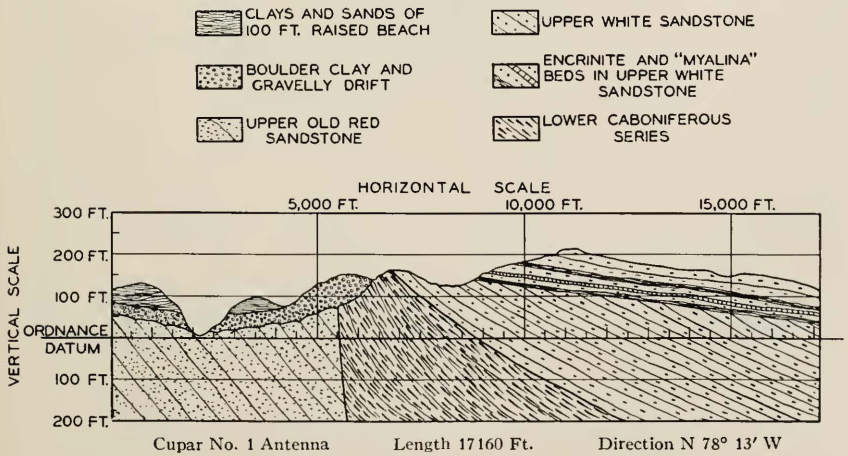


Fig. 7—Cross section of probable geological formation under several wave-antennas.

mission from the several European transmitting stations which are designated on this figure. The measurements on the Houlton antenna system were made using a portable two-kilowatt transmitter located

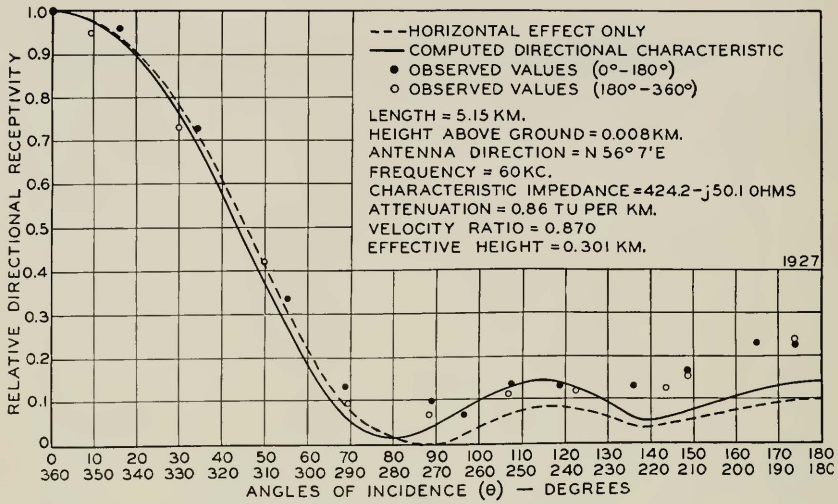


Fig. 8*a*—Relative Directional Receptivity of Houlton Antenna "A" Uncompensated (Long)

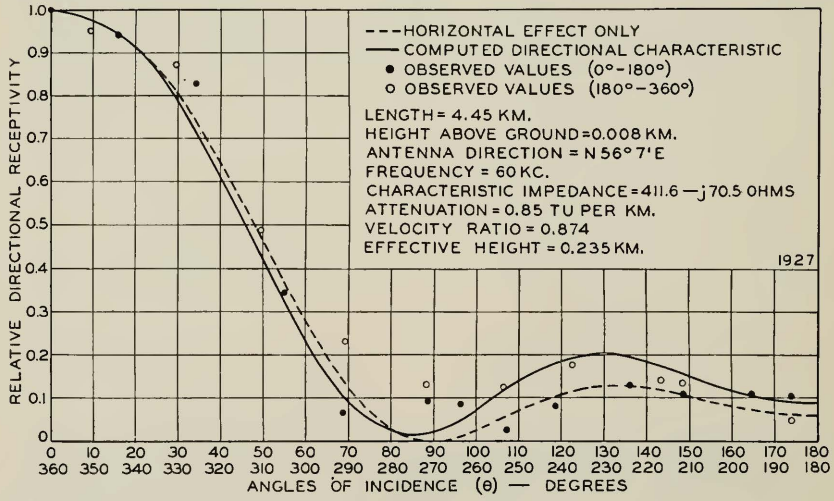


Fig. 8*b*—Relative Directional Receptivity of Houlton Antenna "A" Uncompensated (Short)

successively at each of the 22 positions shown on the map, Fig. 9. The authors wish to thank Mr. G. D. Gillett for his efficient operation of this transmitter during the summer of 1927.

It is seen that the agreement between the measured and the computed directional characteristic is much better for the shortened Houlton *A* antenna than it is for the same antenna 0.70 kilometer longer.

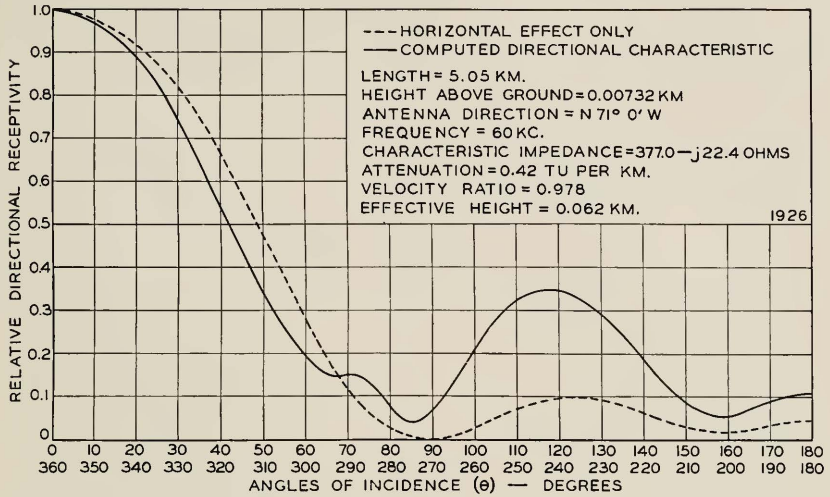


Fig. 8*c*—Relative Directional Receptivity of Wroughton, England—South Antenna

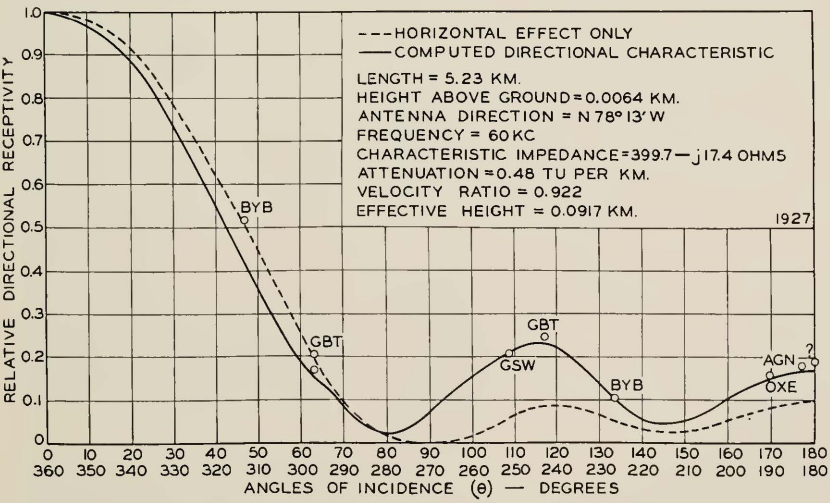


Fig. 8*d*—Relative Directional Receptivity of Cupar, Scotland—Antenna No. 1

The reason for this can be appreciated by reference to Fig. 7, where it is shown that the far end of the long antenna is at the top of a rocky hill; while after shortening, the far end is in a swamp, at the same

average elevation as the remainder of the antenna. The elimination of this sharp rise over rocky ground serves principally to remove an irregularity in the constants of the wave-antenna near the end, so that the entire antenna may be considered more nearly a smooth line. This makes the antenna function more satisfactorily as a unit of an array in connection with other antennas constructed nearby.



Fig. 9.

6. Wave-Antenna Arrays. Since 1899, when S. G. Brown [17] proposed the use of two vertical antennas, separated in space by an appreciable portion of a wave-length and excited at a half-period phase difference, as a means of directional transmission, the use of arrays of antennas

[17] R. M. Foster, "Directive Diagrams of Antenna Arrays," *Bell System Tech. Jour.*, *5*, 292; April, 1926. Also see references listed in Foster's paper.

for directional transmission and reception has become increasingly important. Antenna arrays may be divided into two general classes: (1) arrays of antennas having dissimilar directional characteristics, and (2) arrays of antennas the directional characteristics of which are identical. The array formed by the use of a loop and a vertical antenna to form the familiar "cardioid" is representative of the first class of antenna arrays. Foster [17] has pointed out that the ideal wave-antenna may be considered as an array of an infinite number of loop antennas, extending for the length of the wave-antenna, and hence an antenna array of the second class. (An ideal wave-antenna has no attenuation and a velocity of propagation equal to the velocity of radio propagation in free space.)

An important difference between arrays of dissimilar antennas and arrays of identical antennas lies in the following peculiarity of these two types. In general, the directivity of dissimilar antennas may be increased with no loss in desired signal receptivity by combining them in arrays with little or no separation between the individual antennas. To obtain an increase in directivity by using several identical antennas in an array, however, without too great a sacrifice in desired signal receptivity, the array must cover a space comparable to and of the same order of magnitude as the wave-length of the signals for which it is designed.

It has been stated earlier in this paper that the fundamental form of wave-antenna consists of a single straight horizontal wire, terminated to ground at each end in its characteristic impedance. If the input circuit of a radio receiver be connected across the termination at the end of the antenna most distant from the desired transmitter (the far end of the antenna) this simple form of wave-antenna can be used as a directional receiving system. If arrangements are made to bring the output from the initial end of the wave-antenna to the radio receiver as well as the output from the far end, the simple wave-antenna immediately becomes available for use as two identical antennas in an array. The ends of these two antennas from which the outputs are taken are separated by the length of the antenna and their axes are parallel but in the opposite sense. If before combining these two output currents, that from the initial end of the antenna is changed in phase and magnitude by the proper amount, it is possible to produce a null point of reception in any desired direction. The name "compensation" has been applied to the use of a single wave-antenna to form this array.[14] Since this null point is produced by balancing the back-end current from one antenna of the array (relative to its directional diagram) against the front-end current from the other antenna,

the null point does not remain in the directional characteristic over a band of frequencies.

A directional diagram of a single antenna compensated to produce a null point at $\Theta = 161.4$ degrees (the bearing of the Rocky Point transmitter relative to the axis of the antenna) is shown in Fig. 10. This diagram was calculated, by the method outlined in Appendix 2, from the average of the measured constants of Houlton antennas *A*, *B*, and *D*. In this same figure, measured points are indicated, these points being the average of observations on these three antennas.

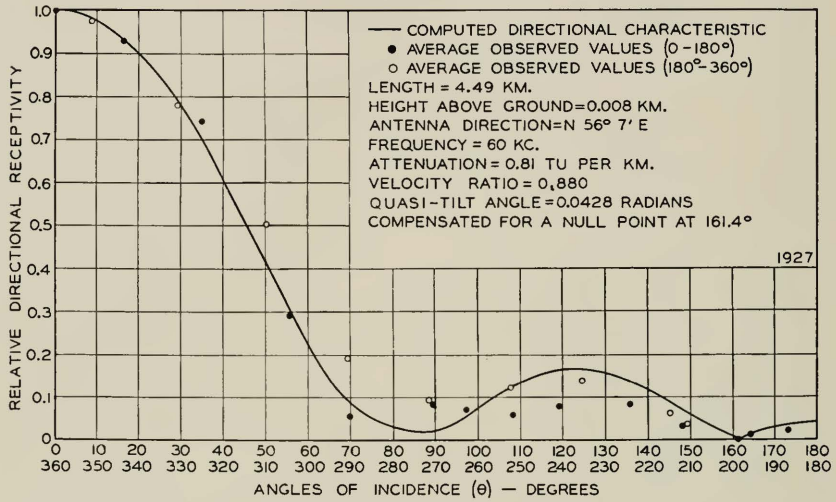


Fig. 10—Wave-antenna directional characteristic. Relative directional receptivity of compensated average Houlton antenna. (Short.)

Beverage, Rice, and Kellog [14] have shown that there are important practical advantages to be gained by constructing the wave-antenna as a two-wire line, and using the metallic circuit acquired thereby as a transmission line to bring the output from one end of the antenna to the radio receiver. The circuits used to bring the output currents from the two ends of the wave-antenna to the radio receiver are shown in Fig. 11. In this case, the radio receiver is located at the initial end of the antenna, so that the predominant desired signal currents are transmitted over the metallic circuit of the wave-antenna to the radio receiver, while the compensation currents are taken directly from the initial end termination when this form of array is used.

To obtain a greater reduction in the Noise Reception Factor (defined under "Directional Discrimination Against Static" earlier in this paper) than is given by compensation, two or more parallel wave-

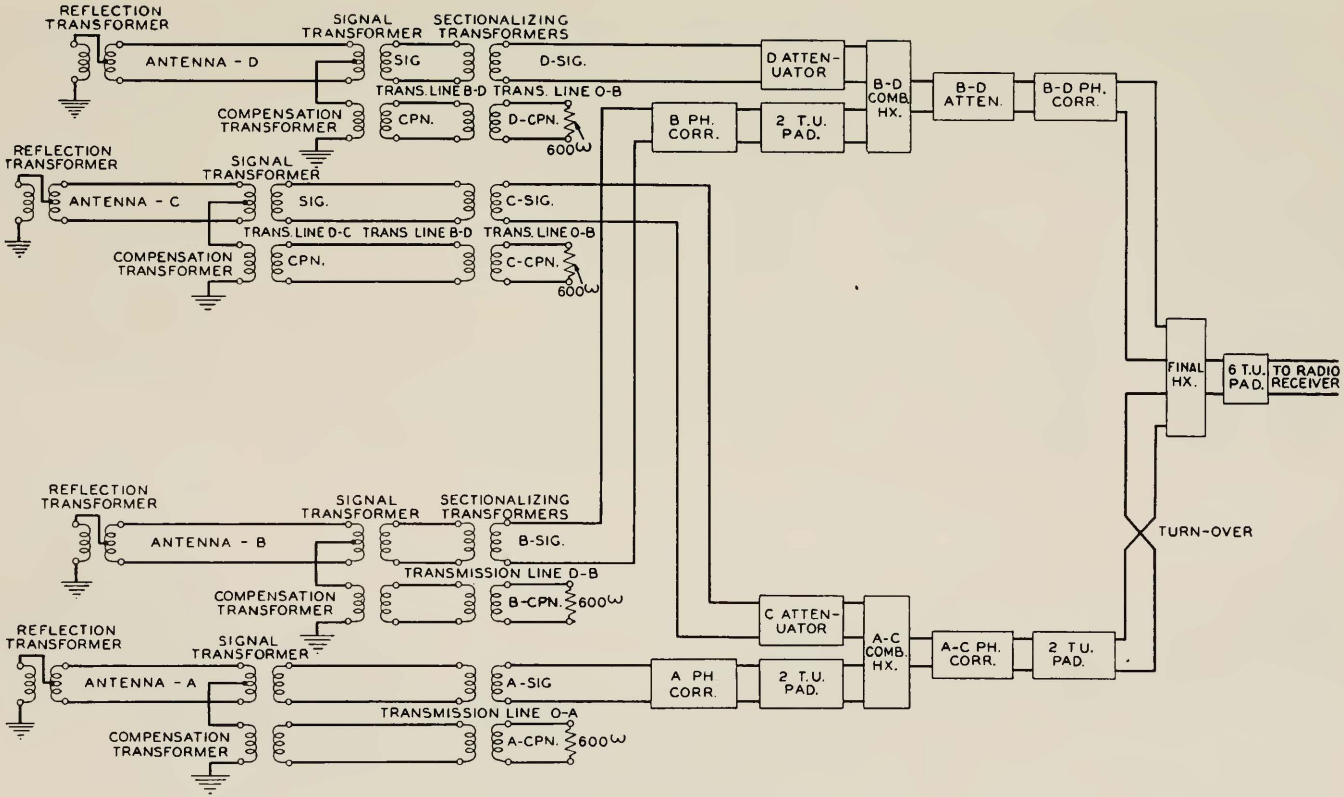


Fig. 11—Houlton antenna system.

antennas are used in either a lateral array, a longitudinal array, or a combination of the two.

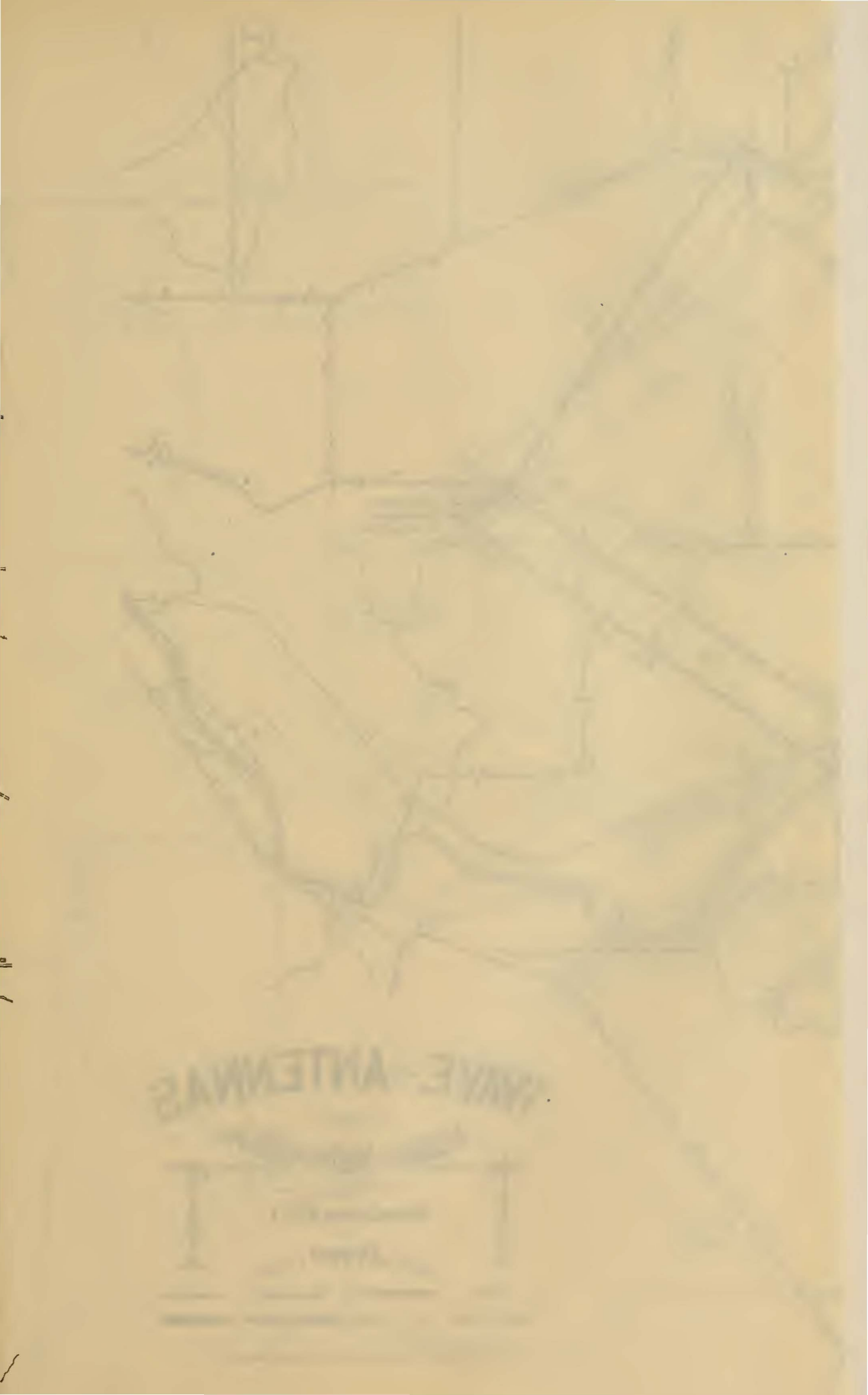
In the lateral array, the initial ends of the wave-antennas are spaced in the direction perpendicular to the axes of the antennas. Since it extends over space in the lateral direction, unless there be undue sacrifice in desired signal, the lateral array can only reduce the width of the directional diagram.

In a true longitudinal array, the antennas are coaxial, but their initial ends are separated by an appreciable fraction of a wavelength. If the wave-antennas forming this type of array overlap one another, then the mutual impedance between them would greatly modify their individual characteristics. In practice, a small amount of lateral spacing between the units of a longitudinal array is necessary to make the mutual impedance negligible. When this type of array is properly designed, the reduction in directional receptivity due to the array is principally in the back-end direction.

The physical layout of the Houlton antenna system is shown in Fig. 12, and the circuits serving to connect the antennas to the radio receiver are shown in Fig. 11. The same letters are used for corresponding line sections in both of these figures. At the time that the directional characteristics of the Houlton antennas were measured, the antenna system comprised only three antennas, *A*, *B*, and *D*. Antenna *A* at that time extended from pole A-33 to pole A-117. Two arrays could then be used: antennas *B* and *D* forming a lateral array, and antennas *A* and *B* forming a modified longitudinal array. In normal operation using either of these two arrays, the transducers in the antenna output circuits were adjusted to combine equal amplitudes of the desired signals from the two antennas in phase with one another.

Using as the unit antenna for the arrays a directional diagram derived from the average constants of the antennas *A*, *B*, and *D*, the directional diagrams of these two arrays have been computed. Fig. 13 shows the computed directional characteristic of the lateral array and Fig. 14 the computed directional characteristic of the modified longitudinal array. On each of these figures, the measured points are shown.

The three antennas *A*, *B*, and *D* represented an uneconomical antenna system inasmuch as but two of the antennas could be used simultaneously in an array. To utilize fully these three antennas, at the same time increasing the discrimination against noise, the fourth antenna *C* has been constructed. To use these four antennas, they are arranged in pairs to form two lateral arrays, and the two lateral arrays arranged in a longitudinal array. The resultant total array



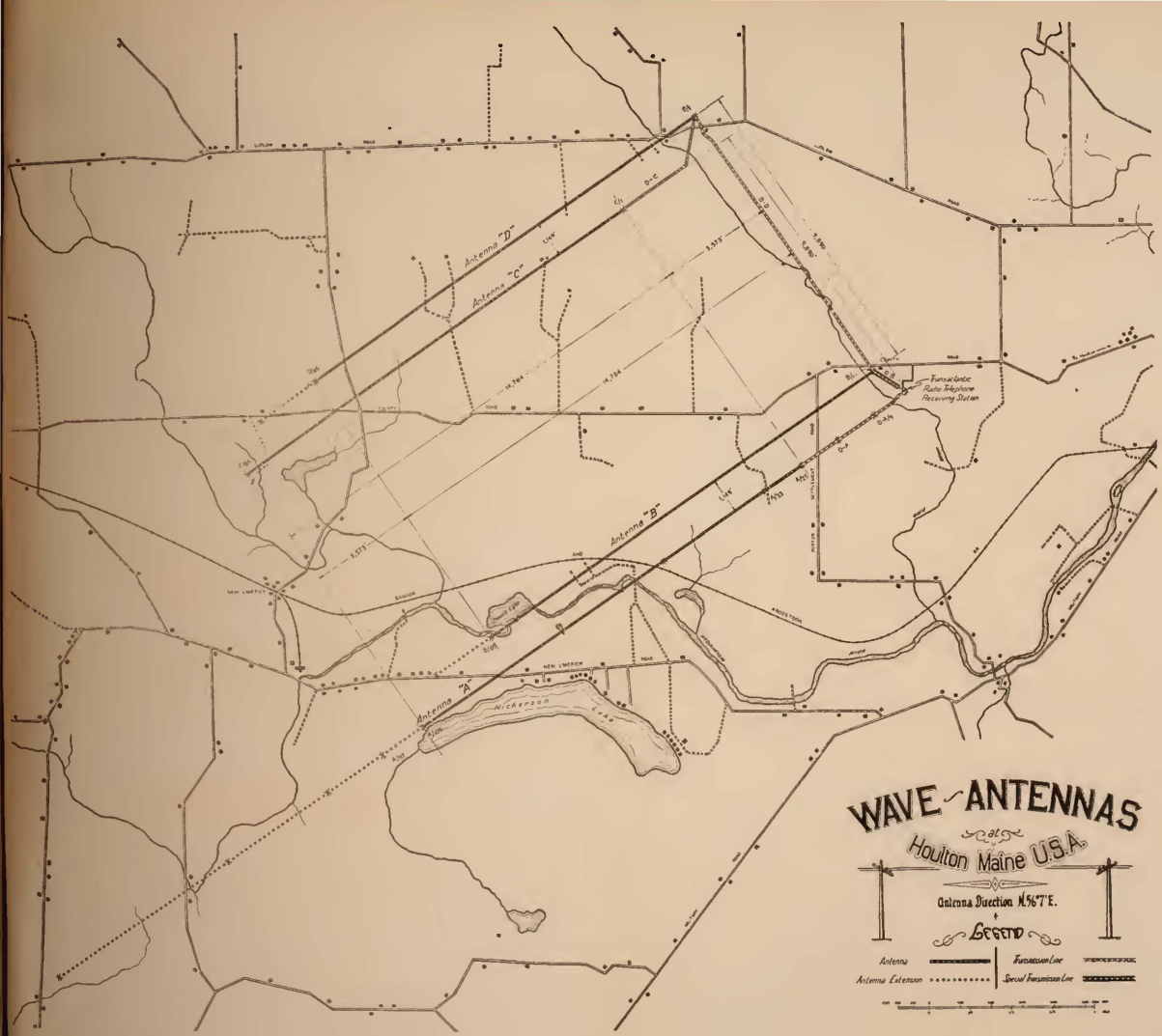

WAVE-ANTENNAS
at
Houlton Maine U.S.A.
Antenna Direction N.56°E.
LEGEND
Antenna
Antenna Extension
Transmission Line
Special Transmission Line
Antenna "A"
Antenna "B"
Antenna "C"
Antenna "D"
Nicatous Lake
Forest Lodge
Radio Telephone
Receiving Station

Fig. 12

quite evidently combines the narrowing of the directional diagram due to the lateral array and the reduction of the back end area of the

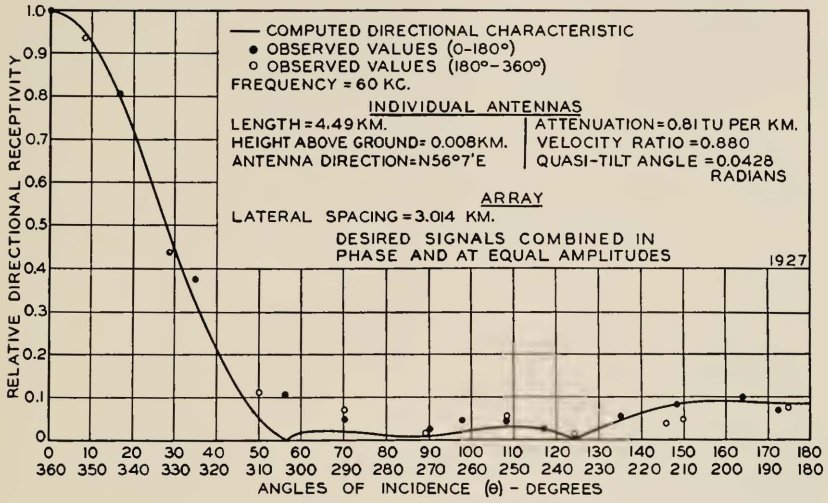


Fig. 13—Wave-antenna array directional characteristic. Relative directional receptivity of lateral array of two Houlton antennas. (Short)

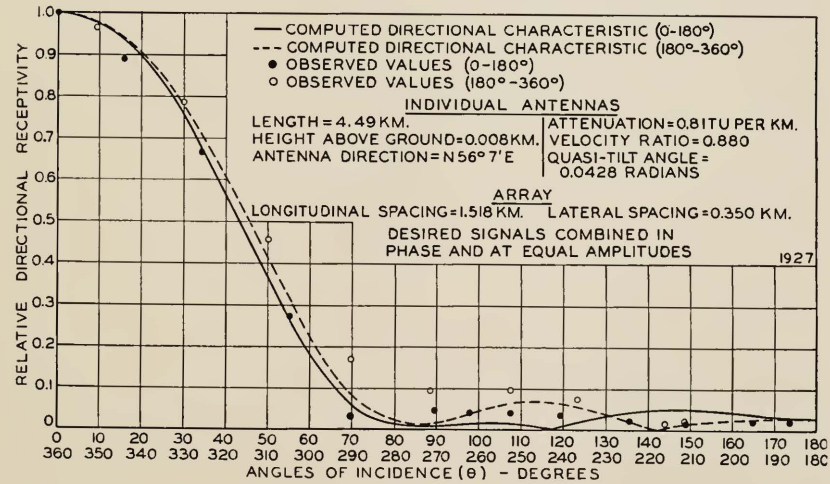


Fig. 14—Wave-antenna array directional characteristic. Relative directional receptivity of modified longitudinal array of two Houlton antennas. (Short)

directional diagram caused by the longitudinal array. A map of this array is shown in Fig. 12.

The circuits for combining four antennas of an array of the type

quite evidently combines the narrowing of the directional diagram due to the lateral array and the reduction of the back end area of the

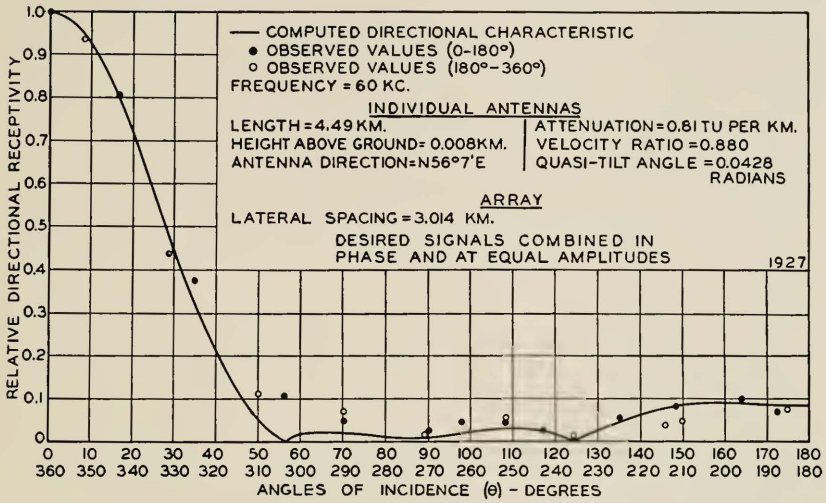


Fig. 13—Wave-antenna array directional characteristic. Relative directional receptivity of lateral array of two Houlton antennas. (Short)

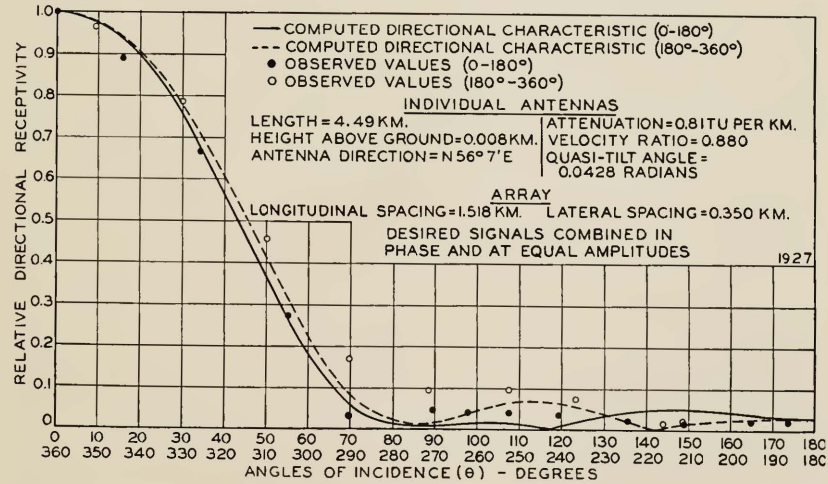


Fig. 14—Wave-antenna array directional characteristic. Relative directional receptivity of modified longitudinal array of two Houlton antennas. (Short)

directional diagram caused by the longitudinal array. A map of this array is shown in Fig. 12.

The circuits for combining four antennas of an array of the type

described in the preceding paragraph are shown in Fig. 11. Antennas *A* and *C* form one lateral array; antennas *B* and *D* form the second. Since antennas *C* and *D* are further removed from the station than *A* and *B*, phase correctors are inserted in the circuits from *A* and *B* to compensate for the phase change in the transmission lines from *C* and *D*, so that the desired signals are combined in phase. The combination of the 2 TU fixed pads and the variable attenuators makes it possible to correct for the attenuation in the transmission lines to the more distant antennas. These several output currents are actually combined in hybrid coils, since this method of combination prevents the antennas from reacting one upon another through the combining system.

After the antennas are combined in pairs to form two lateral arrays, the lateral arrays are combined in the longitudinal array.

The change of phase of space waves between one antenna and the next in an array is a linear function of frequency, and that on the metallic transmission lines practically so. By using phase correctors which have a phase change linear with frequency,[18] the outputs of the antennas in the array may then be combined to produce a null point or a reduction in receptivity, as a result of the array, which retains the same position in the directional diagram for every frequency within a finite band. The longitudinal array at Houlton is designed and combined to produce such an invariable null point in the direction 161.4 degrees relative to the axis of the wave-antenna array. At this angle of incidence, it is evident that the space waves arrive at the lateral array of antennas *A* and *C* before arriving at the lateral array of antennas *B* and *D*. To bring these undesired signals in phase, therefore, phase shift must be introduced into the output of the first of these arrays. Part of this phase shift is supplied by the metallic transmission lines and part by the phase correctors in the combining equipment. At this point, the undesired signals remain in phase as the frequency is varied, so that a turn-over (reversal) inserted in the circuit to the lateral array of antennas *A* and *C* before the array is combined produces the null point which is invariable with variation of the frequency. Under these conditions, the phase of combination of the desired signals, incident at zero angle, varies as the frequency of the desired signals varies. To minimize the effect of this change in phase over the desired frequency band, the spacing of the antennas in the longitudinal array must be so chosen that the desired signals combine very nearly in phase at the middle of the frequency band. For that

[18] O. J. Zobel, "Distortion Correction in Electrical Circuits with Constant Resistance Recurrent Networks," *Bell System Tech. Jour.*, 7, 438; July, 1928.

reason, the longitudinal spacing in the modified longitudinal array was decreased at the same time that the fourth antenna was constructed.

At the time that the extension of the antenna system was undertaken, the measured directional characteristics of the antennas *A*, *B*,

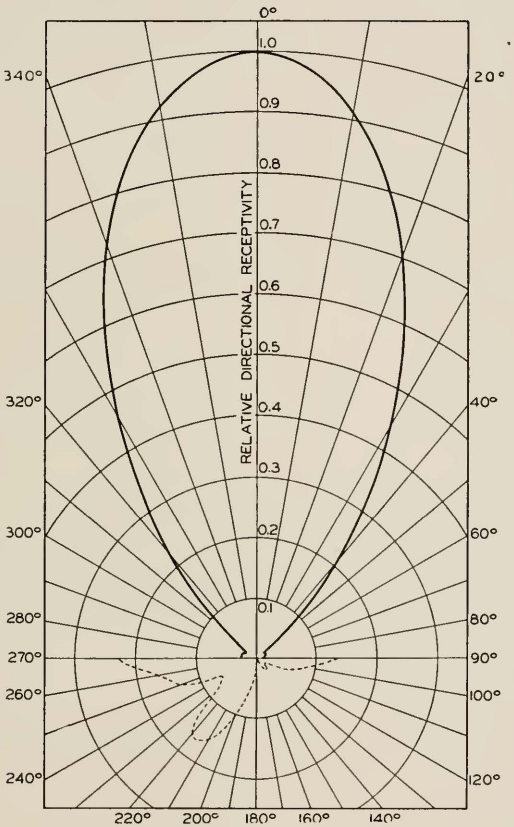


Fig. 15—Wave-antenna array directional characteristic. Calculated relative directional receptivity of array of Houlton antennas *A*, *B*, *C*, and *D*. (From average measured unit antenna characteristic.) Dotted curve—magnified × 10.

and *D* were available, so that the unit directional diagram for use in determining this array characteristic was taken as the average measured characteristic of these three antennas. The calculated directional diagram of the complete Houlton antenna system is shown in Fig. 15. It should be noticed that the scale for the back-end dotted curve is ten times as great as that for the full-line curve for the major lobe.

It is believed that the directional diagram, shown in Fig. 15, represents about the ultimate that can be done economically in a general reduction of back-end area and narrowing of the diagram by means of wave-antennas. Future extensions or redesign of the array at Houlton must be based on the reduction of the relative receptivity in distinct directions determined either by statistical study of the noise received by the antenna system or actual measurements of the direction of arrival of the noise which limits the operation of the transatlantic radio-telephone circuit.

The Radio Receiver

A description of the design and performance of the radio-telephone receiving set will constitute another paper. The radio receiving equipment employed in connection with the antenna systems was developed and constructed by Bell Telephone Laboratories.

The major transmission requirements upon which design must be based are as follows:

1. The limiting values of the signal field to be received;
2. The output power of the receiving antenna for a given signal strength;
3. Power output required from the radio receiver;
4. The type of telephone transmission to be received;
5. The frequency band to be received;
6. The nature and strength of interference from other radio stations and from noise; The selectivity required to reduce undesired modulation
   a. In amplifiers,
   b. In demodulators;
7. Stability of frequency, gain, and transmission-frequency characteristic.

1. Limiting Values of the Signal Field to be Received. The range of daily averages of signal field at 60 kilocycles for all daylight path hours is shown in Fig. 4. The fields, as previously published data indicate,[4, 6, 9] vary diurnally between much wider limits. At night the field frequently approaches, as a maximum, the value calculated on the basis of the inverse distance law. During sunrise and sunset dip periods the field frequently goes to a value less than one microvolt per meter with even 50 kilowatts radiated from a transmitter 5,000 kilometers away. Suppose we take as being approximately correct values, field strengths of 0.4 microvolts per meter as the lower limit and 400 microvolts per meter as the upper limit. We then have determined that the receiving set should have a variation of gain of 60 TU.

2. The Output Power of the Receiving Antenna for a Given Signal Strength. From the observed constants of a Houlton wave-antenna and the assumed value of 0.4 microvolts per meter received at zero degrees to the antenna direction as the lowest field, we calculate, using equations (125), (126), and (127) in Appendix 1, that the power supplied to the reflection transformer terminals is 3.716×10^{-6} microwatts. This power must suffer loss as a result of the transmission back to the receiving station over transmission lines and as a result of the necessity of providing flexibility in the operation of the apparatus used to combine the output of the antenna in question with the output of other antennas before it reaches the input terminals of the radio receiving set. (See Fig. 11.) This loss is such that the power at the input terminals of the radio receiving set from a single antenna and for the minimum signal field is very nearly equal to 3.7×10^{-7} microwatts. With the combining system actually used, the input to the radio receiver from all four antennas will be 12 TU above this value or 5.9×10^{-6} microwatts.

3. Power Output Required from the Radio Receiver. The value of output power required from the radio receiver is really governed by considering the whole radio circuit as a part of a long-distance telephone system. An overall loss of 10 TU has been found satisfactory for long toll circuits. If the telephone lines connecting the circuit terminals to the transmitting and receiving stations have an equivalent of 0 TU then we can place the 10 TU loss in the radio portion of the circuit. If we then supply on a single frequency within the voice-frequency band a power of 1 milliwatt to the input terminals of the radio transmitter, to get a 10 TU equivalent in the radio circuit we must obtain 0.1 milliwatt at the output of the radio receiver.

In the preceding section we determined that the minimum input would be 3.7×10^{-7} microwatts from a single antenna and hence the maximum gain required in the radio receiver to raise this power to the specified 100 microwatts output is 84 TU.

Within amplifiers using three-electrode vacuum tubes, noise is generated in two ways: (*a*) by thermal agitation [19] in the conductor of the input circuit; and (*b*) by "Schottky Effect" [20] in the vacuum

[19] J. B. Johnson, "Thermal Agitation of Electricity in Conductors," *Phys. Rev.*, *32*, 97; July, 1928.
Harry Nyquist, "Thermal Agitation of Electric Charge in Conductors," *Phys. Rev.*, *32*, 110; July, 1928.
J. B. Johnson, "Thermal Agitation of Electricity in Conductors," *Nature*, *119*, 50; Jan. 8, 1927.

[20] Walter Schottky, "Atomare Schwankungsvorgänge an Glühkathodenoberflächen," *Physik. Zeitschr.*, *27*, 701; Nov. 1, 1926.
T. C. Fry, "The Theory of the Schroteffekt," *Jour. Frank. Inst.*, *199*, 203; Feb., 1925.

tubes themselves. Since the transatlantic radio-telephone circuit is so operated that the strength of the voice waves, or "electrical volume," is constant at the output of the radio receiver,[21] the maximum allowable noise at this point in the circuit is likewise constant. Good engineering practice specifies that the continuous "tube noise" should be more than 40 TU below the signal or less than 0.01 microwatt for the specified receiver output of 100 microwatts when using any gain up to the maximum of 84 TU. (With uniformly distributed noise over the voice-frequency band, this is equivalent to about 400 noise units.[22])

4. The Type of Telephone Transmission to be Received. The "single-sideband, suppressed-carrier" type of telephone transmission, invented by John R. Carson,[23] has long been used in the Bell System in carrier systems on wire circuits.[24] Since the advantages of single sideband in radio transmission have been described by Hartley,[25] and in the radio transmitter by Heising,[26] we shall only briefly review the benefits arising from its use.

Transmission of two sidebands with the carrier suppressed represents an improvement over the "carrier and two-sideband" method ordinarily used in "broadcasting" since all of the transmitter power may be concentrated in the intelligence-bearing frequencies. By transmitting only one sideband, further advantages are gained since the frequency space occupied is slightly more than halved for the same grade of circuit, the distortion at the output of the receiver is decreased, and practical simplifications may be made at the transmitting and receiving stations.[27] If the radio transmitter radiates equal power in each of the above-mentioned suppressed carrier transmission schemes and if the radio receiver accepts only the intelligence-bearing fre-

J. B. Johnson, "The Schottky Effect in Low Frequency Circuits," *Phys. Rev., 26,* 71; July, 1925.

[21] S. B. Wright and H. C. Silent, "The New York-London Telephone Circuit," *Bell System Tech. Jour., 6,* 736; October, 1927.

[22] The noise unit is an arbitrary unit used in the Bell System for comparison of any noise with a certain arbitrary source of noise known as a noise standard. The output of the noise standard may be attenuated to produce the same interfering effect on speech as the noise being measured. See W. H. Harden, "Practices in Telephone Transmission Maintenance Work," *Bell System Tech. Jour., 4,* 26; Jan. 1925, for details of making such comparisons.

[23] U. S. Patents Nos. 1,343,306 (1920); 1,343,307 (1920); 1,449,382 (1923), to J. R. Carson.

[24] E. H. Colpitts and O. B. Blackwell, "Carrier Current Telephony and Telegraphy," *Trans. A. I. E. E., 40,* 205; 1921.

[25] R. V. L. Hartley, "Relation of Carrier and Sideband in Radio Transmission," *Proc. I. R. E., 11,* 34; Feb., 1923.

[26] R. A. Heising, "Production of Single Sideband for Transatlantic Radio Telephony," *Proc. I. R. E., 13,* 291; June, 1925.

[27] J. R. Carson, "Signal-to-Static Interference Ratio in Radio Telephony," *Proc. I. R. E., 11,* 271; June, 1923.

quencies in each case, then the signal-to-noise ratio will be the same,[25] provided the resupplied carrier is in frequency synchronism in both systems and in addition in phase synchronism with the suppressed carrier in the two-sideband system.[27]

When receiving single-sideband transmission the carrier suppressed at the transmitting station is resupplied in the radio receiver. Since this carrier will demodulate both sidebands with equal efficiency, the opposite sideband must be eliminated before demodulation to prevent the noise in this sideband from appearing in the voice-frequency output. If the noise power in either sideband is p, then the noise power without opposite sideband suppression is $2p$ and if we reduce the noise power from the opposite sideband to $0.1p$ the total received noise will be reduced

$$10 \log_{10} \frac{2p}{p + 0.1p} = 2.59 \text{ TU}.$$

The maximum possible reduction in noise is 3.01 TU,[28] so that for engineering purposes a 10 TU suppression of the noise in the opposite sideband may be considered adequate. For other reasons to be brought out later in this paper the opposite sideband loss must be greatly in excess of this value.

Provided the resupplied carrier used to demodulate the single sideband suppressed carrier signals is large relative to the signal magnitude [25] at that point in the circuit where we choose to supply it, the only other requirement is that its frequency be correct. Since a displacement of the resupplied carrier 50 cycles above or 20 cycles below the zero of the equivalent voice-frequency band is sufficient to give an appreciable decrease in speech intelligibility, its frequency should be maintained within the smaller of these two limits or within plus or minus 20 cycles of the correct value. It is interesting to note that an absolute variation of only one-tenth of this amount can be observed on music and that speech naturalness is similarly affected.

5. Frequency and Frequency Band to be Received. To utilize the power available at the transmitter most effectively, it is essential to transmit only those frequencies contributing most to received intelligibility. The energy of speech lies largely below 500 cycles while the frequencies most important for intelligibility lie between 400 and 2,600 cycles.[29] By limiting the band transmitted to speech frequencies above 400 cycles some saving is obtained in the transmitter power

[28] J. R. Carson, "Selective Circuits and Static Interference," *Bell System Tech. Jour.*, *4*, 265; April, 1925.

[29] W. H. Martin and Harvey Fletcher, "High Quality Transmission and Reproduction of Speech and Music," *Trans. A. I. E. E.*, *43*, 384; 1924.

required. The range of frequencies transmitted may then extend from 58.9 to 61.1 kilocycles with the suppressed carrier at 58.5 kilocycles, and the radio receiver must be designed to accept this band of frequencies. The transmission-frequency characteristic of the overall radio receiving set should not vary more than ± 2 TU within the band specified above to give a good telephone communication circuit.

6. Selectivity Requirements. The selectivity required in the receiving set is such that when the desired signal is at the assumed minimum value no deleterious effects will be caused by undesired signals.

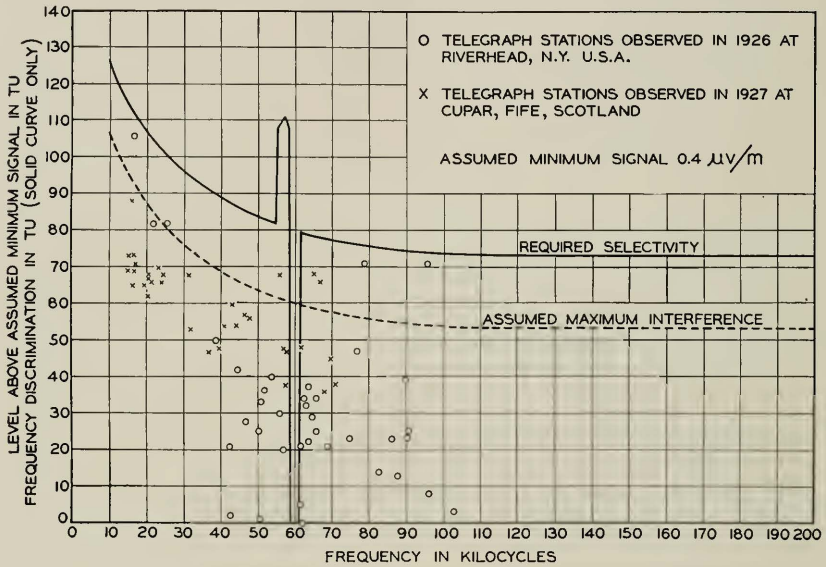


Fig. 16—Selectivity requirements for the long wave transatlantic radio-telephone receiving system.

In Fig. 16 there are shown measured daylight field strengths of various existing radio-telegraph stations as observed at Riverhead, New York, and at Cupar, Scotland. Since these measurements could not be indefinitely extended in frequency nor could they take into account all stations which might exist in this range, they may be considered only as a guide in obtaining a curve of the maximum telegraph interference to be expected. These data, in Fig. 16, have been expressed as ratios (in TU) to the minimum desired signal to be received. It is important to note that the directional selectivity of the receiving antenna system used materially decreases the relative magnitude of many of these interfering signals, particularly at the higher frequencies. The American receiving station is now located in

Houlton, Maine, instead of Riverhead, New York, and this increase in distance from many of the American high-power transmitting stations decreases somewhat their field strengths. In view of these factors the "Assumed Maximum Interference," although it is not greater than any observed station field strength, is, in fact, greater than the interfering signals when the outputs from the actual receiving antennas are used instead of field-strength observations.

6a. Selectivity Requirements Imposed by the Use of Amplifiers. To operate a vacuum tube as an amplifier with negligible distortion the peak voltage applied to its grid must be less than a limiting value so that the tube always operates over the practically linear portion of its characteristic. If no discrimination were provided against unwanted signals, we would be placed in the peculiar situation of having to supply ample tube capacity in the radio receiver to care for the combined load produced by perhaps 100 telegraph stations each of which, on the average, may have a received signal strength 1,000 times the assumed minimum signal. An easy way to decrease the load produced by interference is to insert a filter at the input of the receiving set, which will reduce the required capacity of the first tube. Additional selectivity following the first tube still further reduces the load of undesired signals on the following tubes as more of the capacity of those tubes is used for the desired signals.

Now for design purposes let us assume that the load capacity of each tube is at least 6 TU greater than the capacity required in the tube for the performance of its functions on the desired signal. The undesired signals may then be allowed to produce on the tube grid a voltage equal to that of the desired signal.

Since each of the undesired signals shown in Fig. 16 are about 60 TU stronger than the minimum desired signal, they must be reduced by that amount to make them each no greater than the desired signal.

It is shown in Fig. 21 of Appendix 4 that, as a result of unit random input voltages from 100 operating radio-telegraph stations, a peak voltage will be produced equal to or greater than 10 such units during less than 0.1 per cent of the time. If the undesired telegraph station signals were all of the same magnitude as the desired signal then the voltage which they would produce would be 20 TU above the voltage of the desired signal.

From purely load considerations then, the total required suppression of every interference-bearing frequency outside of the desired signal receiving band will be

$$60 + 20 = 80 \text{ TU}.$$

6b. Selectivity Requirements Imposed by the Use of Demodulators. In all demodulators, the range of desired output frequencies should not be included in the input frequency band because the input frequencies amplified appear in the output of the demodulator as the first order modulation product. The output band of the demodulator should be at a lower frequency than the input band in order to reduce the number of undesired modulation products in the output and in order to obtain the benefit of greater selectivity from circuits operating at lower frequencies. Of course, if all the required selectivity can be conveniently put before the first demodulator, there is no valid reason why multiple demodulation should be used.

In all demodulators except the final demodulator of a radio receiver, the band of frequencies allowed to pass into the demodulator should not be greater in width than the absolute value of the lowest desired frequency in the demodulator output. This requirement is apparent when we consider second order modulation products of interference within the band accepted by the demodulator. Suppose we assume the use of double demodulation and choose 30 kilocycles as the lowest desired frequency in the output of the first demodulator. Then if the band impressed upon the demodulator be more than 30 kilocycles in width, two interfering signals within the band might together give a difference frequency of 30 kilocycles producing load in subsequent stages and possibly tone or noise in the output circuit.

Second order modulation between two signals, one lying within the band accepted by the demodulator and one outside it, may also give rise to interference, due to the difference frequency falling in the output band of the demodulator. Assume that one interfering signal lies within the band accepted by the demodulator and is + 60 TU referred to the minimum desired signal at the grid of the first tube. An equal signal at a frequency outside the band and subject to the selectivity provided for meeting the load requirement will be − 20 TU referred to the minimum desired signal at the same point. Since the second order output from a demodulator is approximately proportional to the product of the grid voltages producing it,[30] we may write (in TU):

Relative desired signal = (0) + (Beating oscillator voltage),
Relative interference = (+ 60) + (− 20) = + 40.

Tests have shown that an interrupted tone, similar to telegraph interference, which is heard at a frequency of 1,100 cycles in a tele-

[30] J. R. Carson, "A Theoretical Study of the Three-Element Vacuum Tube," *Proc. I. R. E.*, 7, 187; April, 1919.

phone receiver is about the most serious frequency of interference to received speech on a telephone circuit, and that frequencies above and below 1,100 cycles are of somewhat less importance. Signals at the equivalent 1,100-cycle frequency produce a type of interference which good engineering practice requires should be reduced at least 50 TU below the desired signal. (This amount of interference is equal to about 500 noise units at the − 10 TU transmission level.[22])

To satisfy this requirement, the relative desired signal should be 50 TU greater than the relative interference at the output of the demodulator. Assigning a minimum magnitude to the beating oscillator voltage of 90 TU above the minimum desired signal voltage on the grid of the demodulator reduces this type of interference sufficiently. (Using a balanced demodulator arrangement, this value might be reduced some 20 TU).

Since any two signals at frequencies entirely outside the band accepted by the demodulator are suppressed some 80 TU, we need not consider their second-order modulation products.

If we use double demodulation in a radio receiver as is assumed in the first part of this section, then we must consider other products of modulation with the beating oscillator for frequencies distant from the accepted band. Space will not permit us to more than mention these, but since the frequencies to be suppressed are distant from the frequencies to be received, their suppression is relatively simple.

In the final demodulator of a radio receiver we must tolerate a certain amount of distortion due to the intermodulation of input frequencies. By limiting the band width into the final demodulator to the same width as the desired output band, the distortion due to intermodulation with interference lying outside the desired band is eliminated. By supplying a large amount of carrier to the final demodulator and by using a balanced demodulator the amount of noise and distortion due to intermodulation of frequencies lying inside the desired band is reduced.

If the desired signal band extends from 58.9 to 61.1 kilocycles and is an upper sideband corresponding to voice frequencies from 400 to 2,600 cycles then, in effect, we must supply a carrier at 58.5 kilocycles to produce the proper voice frequencies in the output circuit. This carrier frequency will also demodulate the frequencies below it in such a way as to produce audible signals and for this reason ample protection must be supplied against the opposite sideband if stations are likely to exist in that range. Calculations show that this is the case; for if 100 stations are distributed at random over the 190-kilocycle range between 10 and 200 kilocycles, then the probability that at least

one station lies between 55.5 and 58.5 kilocycles is 0.796. If the assumed maximum interference at the equivalent 1,100 cycles in the opposite sideband, as indicated by the dashed line in Fig. 16, is 61 TU above the minimum signal and we wish to have it 50 TU below, as previously stated, then we require a selectivity of 61 TU plus 50 TU or 111 TU for this frequency. For other tone-producing frequencies of the opposite sideband similar selectivity requirements have been set up and the resultant for frequencies from 55.5 to 58.5 kilocycles is shown by the solid curve in Fig. 16.

7. *Stability.* As mentioned in Section 4 above, the carrier for a single sideband receiver must be resupplied at the correct frequency. All of the oscillators in the radio link must have sufficient frequency stability to maintain the voice frequencies at the receiver output correct within 20 cycles per second over long periods of time. Suppose we allow 10 cycles per second variation in frequency to exist at the transmitter and an equal amount at the receiver, then the variation in frequency at the receiver must never exceed 0.017 per cent if the resupplied carrier is at 58.5 kilocycles. Certain advantages in stability of the resupplied carrier can be obtained by the use of double demodulation in the receiving set and these will be discussed in another paper.

Variations in the efficiency of the transatlantic radio transmission path for long wave-lengths occur with time of day and season, but during any individual all-daylight transmission period the transmission efficiency of the path is fairly constant. If the gain of the receiver is constant, then, during this important period of the day, the minimum of circuit adjustments will be required. It is hence desirable that the gain of the entire receiving set be made to hold constant within $\pm$ 2 TU for all variations of temperature and of voltage of battery supply, within the operating limits.

It is almost self-evident that the transmission-frequency characteristic through the radio receiver should not vary with temperature and time. Changes of this nature should not exceed 0.5 TU within the transmission band nor 5 TU outside of the transmission band. Design of stable filters and vacuum-tube circuits are essential to produce this result.

The authors have endeavored, in the limited space of the preceding pages, to show what radio transmission considerations must be taken into account in properly designing a receiving system for a commercial radio-telephone circuit. A rather detailed discussion has been necessary to present an accurate picture of the various factors entering into the production of the very essential and highly directional long-wave receiving antenna system employed.

The cooperation of the engineers of the Wireless Section of the British General Post Office, particularly Col. A. G. Lee and Mr. I. J. Cohen, in the measurements made on wave-antennas in England and Scotland, is greatly appreciated and we take this occasion to thank them for having made possible the obtaining of these data. All of our early work in connection with wave-antennas and our initial field trials of lateral and longitudinal arrays of wave-antennas were carried out using wave-antennas located at Belfast, Maine, and Riverhead, New York. These antennas were made available through the courtesy of the Radio Corporation of America, and the authors wish to express to Mr. H. H. Beverage of that organization their appreciation for his interest and assistance during the tests.

APPENDIX 1

The Wave-Antenna

Fundamentally, the wave-antenna consists of a straight horizontal wire, terminated to ground at each end in its characteristic impedance.[14] The determination of the receptivity characteristics of the wave-antenna consists in determining the current flowing in the terminal impedances of the antenna resulting from a field impressed along the antenna.[31]

The wave-antenna is shown in Fig. 17, consisting of a line of length s extending from $x = 0$ to $x = s$. In the nomenclature of the following discussion, letters with no primes refer to the antenna, letters with a single prime ($'$) to the impressed field, and letters with a double prime ($''$) to the resultant field. The wave-antenna is in an impressed electromagnetic field which is defined by the quantities ϕ', V', f_w', and f_g' where

ϕ' = impressed magnetic flux between the lower surface of the wire and the surface of the ground (per unit length);
V' = impressed electric force between the wire and the ground;
f_w' = the impressed electric force *along* the lower surface of the wire;
f_g' = the impressed electric force *along* the surface of the ground.

The total field about the antenna is the sum of this impressed field and a secondary field due to the currents and charges produced in the circuit by the impressed field, so that

$$\begin{aligned} \phi'' &= \phi' + \phi, \qquad f_w'' = f_w' + f_w, \\ V'' &= V' + V, \qquad f_g'' = f_g' + f_g, \end{aligned} \tag{101}$$

[31] J. R. Carson and R. S. Hoyt, "Propagation of Periodic Currents over a System of Parallel Wires," *Bell. System Tech. Jour.*, *6*, 495; July, 1927.

where ϕ, V, f_w, and f_g are the components of the secondary field set up by the currents and charges in the system and ϕ'', V'', f_w'', and f_g'' represent the resultant field about the system.

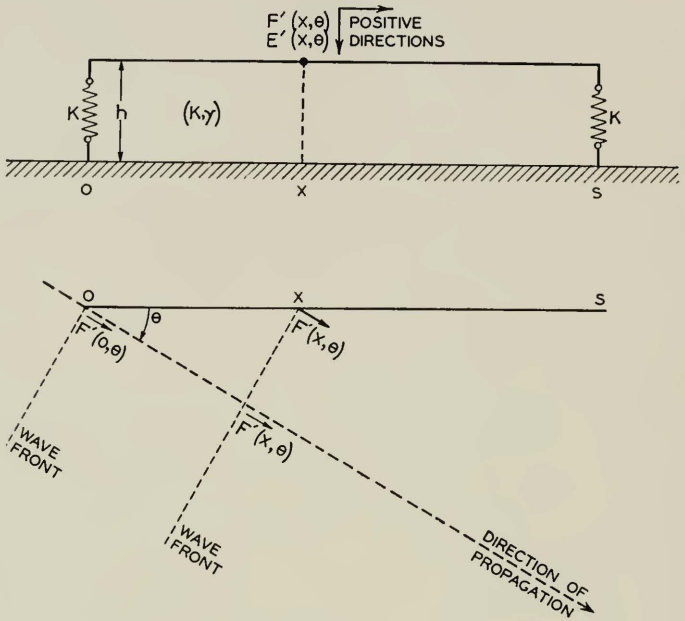


Fig. 17.

As a result of the impressed field, a current I flows in the wire, and a corresponding superposed current distribution is induced in the ground. If the internal impedance of the wire be z_w and that of the ground be z_g, the resultant longitudinal electric force along the wire may be written

$$f_w'' = Iz_w = f_w' + f_w \tag{102}$$

and similarly the resultant longitudinal electric force along the ground is

$$f_g'' = (-Iz_g + f_g') = f_g' + f_g. \tag{103}$$

The second curl law applied to the periphery of the rectangle formed by the vertical at x, the wire, the vertical at $(x + \Delta x)$, and the ground yields

$$zI - f_g' + \frac{dV''}{dx} = -\frac{d\phi''}{dt}, \tag{104}$$

where z is the total series impedance of the wire and the ground circuit and is

$$z = z_g + z_w \tag{105}$$

A summation of the voltages around the above defined rectangle yields

$$f_w' - f_g' + \frac{dV'}{dx} = -\frac{d\phi'}{dt}. \tag{106}$$

Subtracting (106) from (104) we get

$$zI - f_w' + \frac{dV}{dx} = -\frac{d\phi}{dt}. \tag{107}$$

If we write Q as the charge, C as the capacity to ground, and L as the external inductance, each per unit length of the wire, equation (107) becomes

$$zI + L\frac{dI}{dt} + \frac{1}{C}\frac{dQ}{dx} = f_w', \tag{108}$$

but the line current is decreased by the amount of the charging current and the leakage current

$$-\frac{dI}{dx} = \frac{dQ}{dt} + I_Y, \tag{109}$$

where I_Y is the leakage current per unit length of the wire. If the admittance of the leak to ground be designated as Y, the leakage current is

$$I_Y = YV'' = Y(V' + V). \tag{110}$$

Since we are interested only in the steady state, the operator d/dt may be replaced by $j\omega$. Substituting the expression (110) for I_Y into (109) and differentiating with respect to x yields

$$-\frac{d^2I}{dx^2} = \frac{dQ}{dx}j\omega + Y\frac{dV'}{dx} + \frac{Y}{C}\frac{dQ}{dx}. \tag{111}$$

By means of (111) we may eliminate Q from (108)

$$(z + jL\omega)I - \frac{1}{Y + jC\omega}\frac{d^2I}{dx^2} = f_w' + \frac{Y}{Y + jC\omega}\frac{dV'}{dx} \tag{112}$$

and if

$$K = \sqrt{\frac{z + jL\omega}{Y + jC\omega}}, \tag{113}$$

$$\gamma = \sqrt{(z + jL\omega)(Y + jC\omega)}, \tag{114}$$

where K is the characteristic impedance and γ the propagation con-

stant of the antenna circuit, equation (112) may be written

$$\frac{K}{\gamma}\left(\gamma^2 - \frac{d^2}{dx^2}\right) I = f_w' + \frac{YK}{\gamma}\frac{dV'}{dx}. \tag{115}$$

When the boundary conditions are applied, equation (115) defines the value of the current I in the wave-antenna in terms of the impressed electromagnetic field specified by V' and f_w'. By equation (101) the resultant voltages at the ends of the antenna are:

$$V''(0) = V'(0) + V(0), \tag{116}$$

$$V''(s) = V'(s) + V(s). \tag{117}$$

To this point, the solution of the wave-antenna problem has been in a rigidly analytic form. While it is possible to determine completely the received current by following through this method of solution, the problem can be greatly simplified and a physical picture of the problem gained by a synthetic process.

The synthetic method of attack consists of replacing the impressed field by a set of electromotive forces identically equivalent to the impressed field in the sense that it produces the same currents and charges.[31]

The proposed set of electromotive forces is as follows:

A. A distributed longitudinal electromotive force f_w' per unit length in the wire, i.e., an electromotive force $f_w'dx$ in each element of length dx;

B. A distributed vertical electromotive force, V', in the superposed shunt admittance Y between the wire and ground, i.e., an electromotive force V' in each elemental admittance path Ydx;

C. In each end of the wire, $x = 0$ and $x = s$, localized series electromotive forces, equal respectively to minus and plus the impressed voltages at those points; i.e., equal to $-V'(0)$ and $+V'(s)$ respectively.

The electromotive force of *A* is suggested by (107), that of *B* by (109) and (110), that of *C* by the terminal conditions expressed in (116) and (117). In the case of a wave-antenna constructed to maintain high insulation resistance, the conductance portion of the superposed admittance Y can be made negligibly small. Under this condition, the susceptance part of this admittance can be combined with the linear capacitance of the wire to alter the propagation constants (K and γ) of the antenna and the voltages induced in the superposed shunt admittances neglected.

By reference to Fig. 17, the impressed field may be identically defined at each point along the antenna.

The longitudinal electromotive force in each element of the wire is

$$\begin{aligned} f_w'dx &= F'(x, \theta) \cos \theta \, dx, \\ f_w'dx &= F'(0)\epsilon^{-\gamma' x \cos\theta} \cos \theta \, dx. \end{aligned} \tag{118}$$

The impressed voltage at the point x along the antenna is

$$\begin{aligned} V'(x) &= h \cdot E'(x, \theta), \\ V'(x) &= h \cdot E'(0)\epsilon^{-\gamma' x \cos\theta}. \end{aligned} \tag{119}$$

In (118) and (119) $F'(0)$ and $E'(0)$ represent the horizontal and vertical components respectively of the impressed electric field at the end of the antenna $x = 0$, and h represents the height of the antenna above ground. For the purpose of this discussion, it will be assumed that F' and E' are not dependent upon θ. The current produced at the receiving end s by the horizontal component of the impressed field is given by

$$I_{F'\theta} = \int_0^s \frac{F'(0)\epsilon^{-\gamma' x \cos\theta} \cos \theta \, dx}{2K} \epsilon^{-\gamma(s-x)}, \tag{120}$$

from which

$$I_{F'\theta} = \frac{sF'(0) \cos \theta}{2K} \frac{\epsilon^{(\gamma-\gamma' \cos\theta)s} - 1}{(\gamma - \gamma' \cos \theta)s} \epsilon^{-\gamma s}. \tag{121}$$

The current produced at the receiving end s by the vertical component of the impressed field is evaluated as follows:

$$I_{E'\theta} = \frac{V'(s)}{2K} - \frac{V'(0)}{2K} \epsilon^{-\gamma s} \tag{122}$$

and by combination of (119) and (122)

$$I_{E'\theta} = \frac{hE'(0)}{2K} [\epsilon^{(\gamma-\gamma' \cos\theta)s} - 1] e^{-\gamma s}. \tag{123}$$

Zenneck's theory of wave propagation [32] has been developed by Breizig [33] to show that the horizontal and vertical components of the impressed field are related by the expression

$$-\frac{F'}{E'} = \epsilon^{j\delta} \tan T. \tag{124}$$

[32] J. Zenneck, "Ueber die Fortpflanzung ebener electromagnetischer Wellen längs einer ebenen Leiterfläche und ihre Beziehung zur drahtlosen Telegraphie," *Ann. der Phys., 23,* 846; June, 1907.

[33] Franz Breizig, "Theoretische Telegraphie," Braunschweig, 1924. 2d ed., pp. 482–487.

The total current produced at the receiving end s by the impressed field is

$$I_\theta = I_{F'\theta} + I_{E'\theta} \tag{125}$$

and by application of (124) the constituents of the total current are

$$I_{F'\theta} = \frac{S\lambda' F'}{2K} \cos\theta \frac{1 - \epsilon^{-[\alpha S\lambda' + j2\pi S(m - \cos\theta)]}}{\alpha S\lambda' + j2\pi S(m - \cos\theta)} \epsilon^{-j2\pi S \cos\theta}, \tag{126}$$

$$I_{E'\theta} = -\frac{S\lambda' F'}{2K} \frac{h}{S\lambda'} \frac{1}{\epsilon^{j\delta} \tan T} (1 - \epsilon^{-[\alpha S\lambda' + j2\pi S(m - \cos\theta)]}) \epsilon^{-j2\pi S \cos\theta}. \tag{127}$$

In (125), (126), and (127), the symbols have the following meanings

Symbol	Definition	Unit
I_θ	The total current produced at the receiving end of the antenna s by an impressed field propagated at an angle θ from the axis of the antenna.	amperes
$I_{F'\theta}$	The portion of I_θ produced by the horizontal component of the impressed field.	amperes
$I_{E'\theta}$	The portion of I_θ produced by the vertical component of the impressed field.	amperes
F'	The horizontal component of the impressed field. (Positive direction in the direction of propagation along the ground.)	volts per kilometer
E'	The vertical component of the impressed field. (Positive direction downward.)	volts per kilometer
δ	Phase angle between the horizontal and vertical components of the impressed electric field.	radians
T	"Quasi-tilt angle" of the impressed electric field.	radians
K	The characteristic impedance of the wave-antenna.	ohms
γ	The propagation constant of the wave-antenna.	
α	The real part of the propagation constant of the wave-antenna or the attenuation constant.	napiers per kilometer
β	The imaginary part of the propagation constant of the wave-antenna or the phase constant.	radians per kilometer
γ'	The propagation constant of the space waves.	
α'	The real part of the propagation constant of the space waves (assumed equal to zero).	napiers per kilometer
β'	The imaginary part of the propagation constant of the space waves.	radians per kilometer
s	The length of the wave-antenna.	kilometers
h	The height of the wave-antenna above ground.	kilometers
$S = s/\lambda'$	The length of the wave-antenna.	space wave-lengths
$\lambda' = 2\pi/\beta'$	The wave-length of the space waves.	kilometers
$V = 2\pi f/\beta$	Apparent velocity of propagation of waves along the wave-antenna.	kilometers per second
V'	The velocity of propagation of the space waves ($= 3 \times 10^5$ km per second).	kilometers per second

V/V' $m \equiv V'/V =$	Velocity ratio.	numeric
β/β'	Reciprocal of the velocity ratio.	numeric
$j = \sqrt{-1}$		
θ	The angle between the axis of the wave-antenna and the direction of propagation of space waves measured in a clockwise direction.	
R.D.R. $= \frac{I_\theta}{I_0}$	Relative directional receptivity.	numeric

APPENDIX 2

Antenna Arrays

The directional discrimination yielded by a single antenna can be increased by utilizing several such antennas in an array.[17] In Fig. 18,

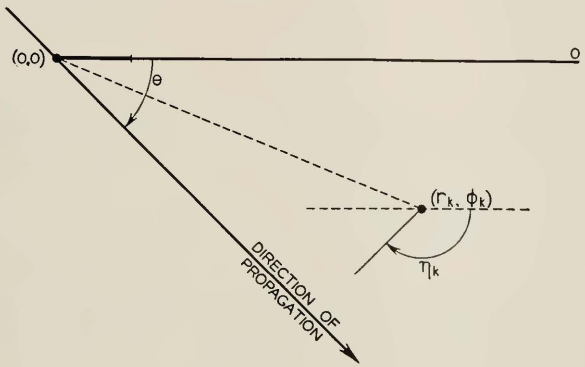


Fig. 18

a general array of n antennas is indicated, of which only the first and the k'th are portrayed.

Each antenna in the array is completely specified by the coordinates of the initial end of the antenna, the angle between the zero axis of the coordinate system and the axis of the antenna, and the current delivered at the receiving end of the antenna for a given electric field impressed on the antenna at each angle of incidence with the antenna. Literally, the first and the k'th antennas are specified as follows:

	First Antenna	*k'th Antenna*
Coordinates of initial end of antenna...	(0,0)	(r_k,ϕ_k)
Direction of antenna.....	0	η_k
Current delivered by antenna for a constant electric field propagated in the direction θ	$I_{\theta 0}$	$I_{\theta k}$

For the purpose of this discussion, it is sufficiently accurate to assume that the propagation of space waves over the area covered by the array only involves phase retardation, i.e.,

$$\gamma' = j\beta'. \tag{201}$$

The output of the k'th antenna is transmitted through a linear transducer having a transfer constant P_k to a common point where it is combined with the outputs of the other antennas of the array. The current from the k'th antenna at the point of combination is therefore

$$J_{k\theta} = I_{\theta k}\epsilon^{-j\beta'[r_k/V']\cos(\theta-\phi_k)}\epsilon^{-P_k}, \tag{202}$$

where

$$\theta_k = \theta - \eta_k \tag{203}$$

and

$$\beta' = \frac{2\pi V'}{\lambda'}. \tag{204}$$

The total current received from the n antennas of the array is equal to the sum of the currents received from the individual antennas, or

$$J_\theta = \sum_{k=1}^{k=n} I_{\theta k}\epsilon^{-j[2\pi r_k/\lambda']\cos(\theta-\phi_k)}\epsilon^{-P_k}. \tag{205}$$

Equation (205) gives the total current received from *any* array of antennas for any direction of wave propagation in a horizontal plane. This general expression is not adapted to ready determination of directional characteristics of antenna systems, but it may be simplified by placing the following restrictions on the individual antennas forming the array and their space relations in the array:

(1) The antennas are all alike. This restriction may be defined by the expression:

$$I_{\theta k} = I_{\theta(k+1)}.$$

(2) The axes of the antennas are parallel, as defined by the expression

$$\eta_k = 0 \text{ or } \pi.$$

(3) The initial ends of the antennas are equally spaced along straight lines in each subgroup and the subgroups are equally spaced along straight lines. All of the subgroups are identical.

The general antenna array conforming to these restrictions is shown in Fig. 19. In this figure, there are q groups of antennas equally spaced by the distance a along a line 90 deg. from the zero axis. In each of these q groups of antennas, there are p antennas, divided into

two series, those for which $\eta = 0$ being numbered 1, 3, $\cdots$, $(2l - 1)$, $\cdots$ $(p - 1)$ and those for which $\eta = \pi$ being numbered 2, 4, $\cdots$, $2l$, $\cdots$, p, the initial ends of the second series being removed by a distance s from the initial ends of the first series, along the axes of the antennas of the first series.

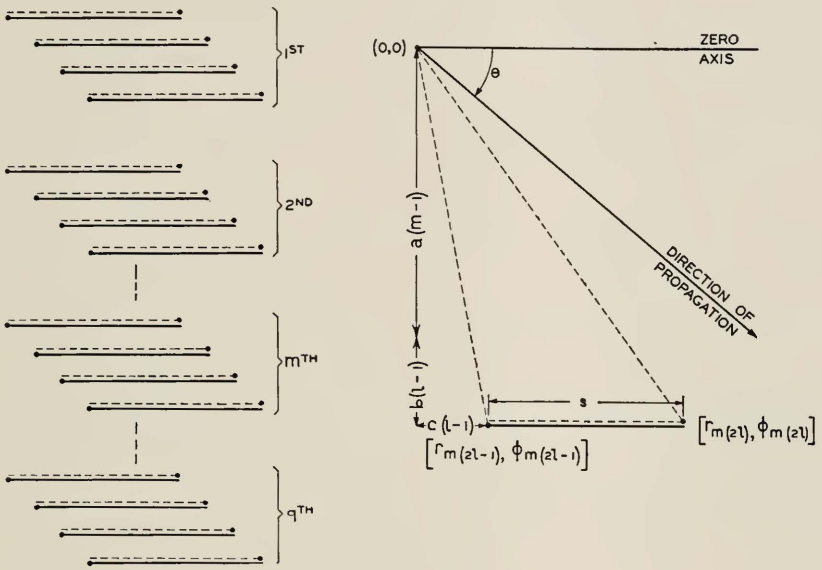


Fig. 19

Equation (205) applied to this general array gives for the total current received from the array

$$J_\theta = I_\theta \sum_{m=1}^{m=q} \sum_{2l=2}^{2l=p} \epsilon^{-j\frac{2\pi r_{m(2l-1)}}{\lambda'}\cos[\theta-\phi_{m(2l-1)}]} \epsilon^{-P_{m(2l-1)}} + I_{\theta-\pi} \sum_{m=1}^{m=q} \sum_{2l=2}^{2l=p} \epsilon^{-j\frac{2\pi r_{m(2l)}}{\lambda}\cos[\theta-\phi_{m(2l)}]} \epsilon^{-P_{m(2l)}}, \quad (206)$$

where

$$r_{m(2l-1)} = \sqrt{[a(m-1) + b(l-1)]^2 + [c(l-1)]^2}, \quad (207)$$

$$r_{m(2l)} = \sqrt{[a(m-1) + b(l-1)]^2 + [c(l-1) + s]^2}, \quad (208)$$

$$\phi_{m(2l-1)} = \tan^{-1}\left[\frac{a(m-1) + b(l-1)}{c(l-1)}\right], \quad (209)$$

$$\phi_{m(2l)} = \tan^{-1}\left[\frac{a(m-1) + b(l-1)}{c(l-1) + s}\right]. \quad (210)$$

In a double summation, the result is independent of the order in which the summations are taken. If then we write

$$u_\theta = \sum_{2l=2(m=1)}^{2l=p} \epsilon^{-j\frac{2\pi r_{m(2l-1)}}{\lambda'}\cos[\theta-\phi_{m(2l-1)}]}\epsilon^{-P_{m(2l-1)}}, \tag{211}$$

$$v_\theta = \sum_{m=1(2l=2)}^{m=q} \epsilon^{-j\frac{2\pi r_{m(2l-1)}}{\lambda'}\cos[\theta-\phi_{m(2l-1)}]}\epsilon^{-P_{m(2l-1)}}, \tag{212}$$

$$w_\theta = \sum_{2l=2(m=1)}^{2l=p} \epsilon^{-j\frac{2\pi r_{m(2l)}}{\lambda'}\cos[\theta-\phi_{m(2l)}]}\epsilon^{-P_{m(2l)}}, \tag{213}$$

$$y_\theta = \sum_{m=1(2l=2)}^{m=q} \epsilon^{-j\frac{2\pi r_{m(2l)}}{\lambda'}\cos[\theta-\phi_{m(2l)}]}\epsilon^{-P_{m(2l)}}. \tag{214}$$

The expression for the total current may be written

$$J_\theta = I_\theta u_\theta v_\theta + I_{\theta-\pi} w_\theta y_\theta. \tag{215}$$

If the transducers in the circuits from each antenna of a pair are so related that

$$P_{m(2l)} - P_{m(2l-1)} = P_c, \tag{216}$$

the expression for the total current becomes

$$J_\theta = u_\theta v_\theta [I_\theta + I_{\theta-\pi}\epsilon^{-j[2\pi s/\lambda']\cos\theta}\epsilon^{-P_c}]\epsilon^{-P_1}. \tag{217}$$

The directional diagram in terms of relative directional receptivity is

$$RDR = \frac{J_\theta}{J_0} = \frac{u_\theta}{u_0} \times \frac{v_\theta}{v_0} \times \left[\frac{I_\theta + I_{\theta-\pi}\epsilon^{-j[2\pi s/\lambda']\cos\theta}\epsilon^{-P_c}}{I_0 + I_{-\pi}\epsilon^{-j[2\pi s/\lambda']}\epsilon^{-P_c}}\right]. \tag{218}$$

Since there has been no assumption to this point of the character of I_θ, the significance of the coefficients u_θ and v_θ may be determined by assuming

(1) $I_\theta = I_0$, which is the directional characteristic of a vertical antenna

(2) $s = 0$

(3) $P_c = \infty$.

Consideration of (218) in light of (211) and (212) under these conditions leads to the conclusion that

$$\frac{u_\theta}{u_0} \quad \text{and} \quad \frac{v_\theta}{v_0}$$

are the relative directional receptivities of two arrays of vertical antennas placed at the initial ends of the antennas comprising the desired array. If, then, we designate the relationship between antennas indicated by the expression

$$J_c = [I_\theta + I_{\theta-\pi}\epsilon^{-j[2\pi s/\lambda']\cos\theta}\epsilon^{-P_c}] \tag{219}$$

as *compensation* [14] and recognize that this expression gives the directional characteristic of a compensated antenna, we may formulate the rule that the directional characteristic of an array of similar parallel unit antennas is equal to the product of the directional characteristic of the unit antenna and the directional characteristic of an array of unit vertical antennas placed at the initial ends of the unit antennas forming the array, the product being taken point for point as the angle of incidence increases. The relative directional receptivity of each fundamental array of vertical antennas is termed the array factor, so that similarly, the relative directional receptivity of an array of similar parallel unit antennas is given by the product of the relative directional receptivity of the unit antenna and the array factor. This method may be extended to the solution of a complicated array such as that shown in Fig. 19, by determining the relative directional receptivity for groups of unit antennas, then determining the array factor for these groups taken as unit antennas. Expressed literally for a complex array of this type:

$$RDR_{\text{array}} = [A_1 \times A_2 \times \cdots \times A_n]RDR_{\text{unit antenna}}, \tag{220}$$

where $A_1 \cdots, A_n$ are the array factors for the fundamental groups into which the complete array may be divided.

APPENDIX 3

Wave Tilt and Ground Conductivity

In Zenneck's [32, 33] exposition of the relation between the horizontal and vertical components of a plane electric wave propagated along a horizontal surface between two media, it is demonstrated that these two constituents of the wave in the upper medium (1) are related by the expression

$$-\frac{F'}{E'} = \epsilon^{j\delta}\tan T = \sqrt{\frac{\frac{9\times 10^{11}}{\rho_1} + j\frac{1}{4\pi}\omega k_1}{\frac{9\times 10^{11}}{\rho_2} + j\frac{1}{4\pi}\omega k_2}}, \tag{301}$$

where

Symbol	Definition	Unit
F'	The horizontal component of the electric wave in medium 1. (Positive direction in the direction of propagation along the interface.)	volts per kilometer
E'	The vertical component of the electric wave in medium 1. (Positive direction downward.)	volts per kilometer
ρ_1	Specific resistivity of medium 1.	ohms per centimeter cube
ρ_2	Specific resistivity of medium 2.	ohms per centimeter cube
k_1	Dielectric constant of medium 1 and equal to unity for a vacuum.	numeric
k_2	Dielectric constant of medium 2.	numeric
f	Frequency	cycles per second
ω	$2\pi f$	

Our primary interest is in the case where the first medium is air, and the second medium is the earth beneath an antenna system. In this case the constants of the media may be given the values:

$$\rho_1 = \infty \quad \text{(air)},$$
$$\rho_2 = \rho \quad \text{(earth)},$$
$$k_1 = 1 \quad \text{(air)},$$
$$k_2 = k \quad \text{(earth)}.$$

Substituting these values into the general equation (301)

$$-\frac{F'}{E'} = \epsilon^{j\delta} \tan T = \frac{1}{\sqrt{k}} \left[\frac{\left(\frac{fk\rho}{18 \times 10^{11}}\right)^2}{1 + \left(\frac{fk\rho}{18 \times 10^{11}}\right)^2} \right]^{\frac{1}{4}} \epsilon^{j[1/2 \tan^{-1}(18\times 10^{11}/fk\rho)]}. \quad (302)$$

At this point it is desirable to indicate the significance of the term "quasi-tilt angle" as applied to T. It is seen that (tan T) is the absolute magnitude of the ratio of the horizontal and vertical components of the electric field. In the case that the time phase between the two components of the field is zero (i.e., $\delta = 0$), T would represent the angle of forward inclination of the propagated wave front. In general, δ is unequal to zero and hence the angle of inclination of the major axis of the ellipse traced by the electric vector is less than T, but it still remains convenient to express the ratio of the magnitudes of the two components of the field as the tangent of an angle. This angle cannot be called the wave tilt, however, but the term "quasi-tilt angle" may safely be applied to it.

The ground constants may be determined from measurement of the "quasi-tilt angle" as the following development shows:

Equation (302) may be written as two equations

$$\tan T = \frac{1}{\sqrt{k}} \left[\frac{\left(\frac{fk\rho}{18 \times 10^{11}}\right)^2}{1 + \left(\frac{fk\rho}{18 \times 10^{11}}\right)^2} \right]^{\frac{1}{4}}, \tag{303}$$

$$\delta = \frac{1}{2} \tan^{-1} \left(\frac{18 \times 10^{11}}{fk\rho} \right). \tag{304}$$

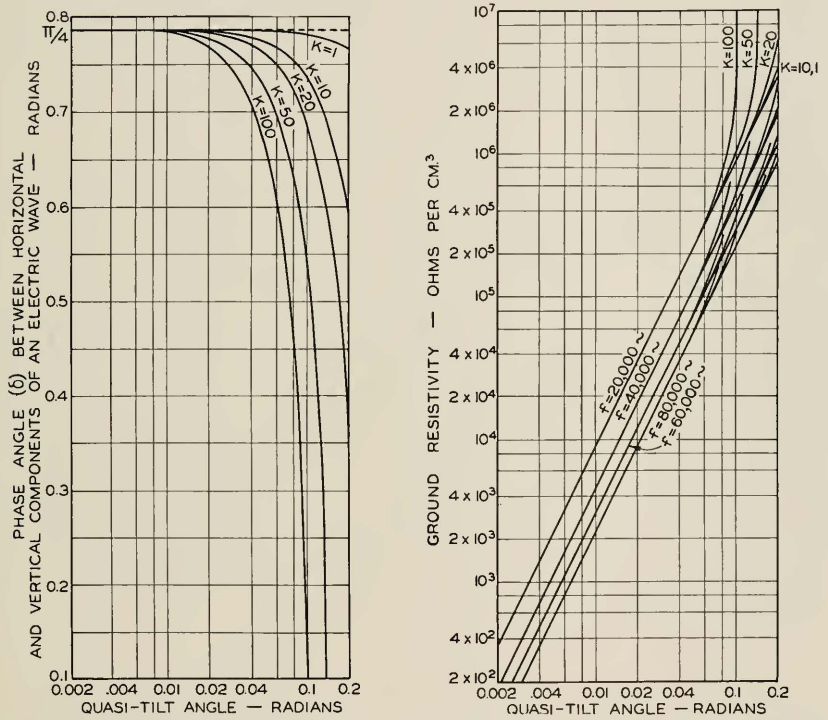


Fig. 20—Relation between quasi-tilt angle, ground resistivity, and phase angle between horizontal and vertical components of an electric wave. (By Zenneck's formula.)

Solving equations (303) and (304) as simultaneous equations for δ in terms of k and T and for ρ in terms of f, k, and T yields

$$\delta = \frac{1}{2} \cos^{-1} (k \tan^2 T), \tag{305}$$

$$\rho = \frac{18 \times 10^{11}}{f} \frac{\tan^2 T}{\sqrt{1 - k^2 \tan^4 T}}. \tag{306}$$

These two expressions have been evaluated for the extreme range of values of k that would be met in practice (k between 1 and 100) and for values of T between 0.002 and 0.2 radian and are plotted in Fig. 20. The figures for dielectric constant given by Fleming [34] show that for earth, the maximum value of k to be expected is below 20. It is evident, therefore, that δ is negligibly different from $\pi/4$ for values of T below 0.05 radian in the vicinity of an antenna which is constructed over land. Also Fig. 20 shows that the specific resistivity is practically independent of k for the same range of T. Fortunately, the measured values of T lie within these limits, so that the time phase difference between the horizontal and vertical components of an electric wave, and the ground resistivity may be evaluated with but slight error from measurements of the quasi-tilt angle.

APPENDIX 4

Probability of Voltages Greater than any Specified Value Resulting from the Simultaneous Reception of Several Radio-Telegraph Stations in a Restricted Frequency Range

In order to determine the required load capacity of vacuum tubes for a radio receiver, it is necessary to obtain some estimate of the voltages from interfering signals which may occur at the input of the radio receiver and during how much of the time certain specified voltages are exceeded.

If we assume that there are N telegraph stations within a restricted frequency range, that each station contributes equal unit voltage at the receiver, and that the probability of the key being closed at any one station is constant, then the probability that exactly n stations have their keys depressed at the same time is

$$P_n = \frac{N!}{n!(N-n)!} K^n (1-K)^{(N-n)}, \tag{401}$$

where K is the fraction of the total time that each station has its key depressed.

In order to determine the probability that n stations will produce a voltage equal to or greater than any specified value x we have followed Rayleigh's problem of random phases as explained in Volume 6 of his "Scientific Papers," page 618. While the conditions are not all satisfied it can be shown that they are approximately satisfied for

[34] J. A. Fleming, "Principles of Electric Wave Telegraphy and Telephony," Longmans, Green and Co., 1916. 3d edition, p. 800.

the great majority of possible combinations and for small time intervals. The formula of Rayleigh gives the probability that the resultant of n vectors lies within an arbitrary interval $(r - dr/2,\ r + dr/2)$. Since we will assume sinusoidal voltages in the actual problem under consideration we require the probability that the projection of the resultant on the real axis is greater than a given value of x. This can be calculated by changing the polar coordinates of Rayleigh's formula to rectangular coordinates and integrating with respect to y from $-\infty$ to $+\infty$ and then with respect to x from x to $+\infty$.

The integrated formula then becomes

Probability of a voltage greater than $x = P_x$

$$P_x = A_1\Gamma\left(\frac{1}{2}, \frac{x^2}{n}\right) + A_2\Gamma\left(\frac{3}{2}, \frac{x^2}{n}\right) + A_3\Gamma\left(\frac{5}{2}, \frac{x^2}{n}\right) + A_4\Gamma\left(\frac{7}{2}, \frac{x^2}{n}\right) + A_5\Gamma\left(\frac{9}{2}, \frac{x^2}{n}\right), \tag{402}$$

where

$$A_1 = \frac{1}{2\sqrt{\pi}}\left(1 - \frac{3}{16n} - \frac{5}{24n^2} + \frac{105}{16.32n^2}\right),$$

$$A_2 = \frac{1}{2n\sqrt{\pi}}\left(\frac{3}{4} - \frac{25}{64n}\right),$$

$$A_3 = \frac{1}{2n\sqrt{\pi}}\left(-\frac{1}{4} + \frac{155}{192n}\right),$$

$$A_4 = \frac{1}{2n\sqrt{\pi}}\left(-\frac{47}{144n}\right),$$

$$A_5 = \frac{1}{2n\sqrt{\pi}}\left(\frac{1}{32n}\right)$$

and

$$\Gamma(p, u\sqrt{p}) = \Gamma(p)[1 - I(u, p - 1)],$$

in which

$$p = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \frac{7}{2}, \frac{9}{2} \text{ and } u\sqrt{p} = \frac{x^2}{n}.$$

Having found u, the I functions of $(u,\ p - 1)$ can be obtained from Pearson's "Tables of the Incomplete Γ-Functions." $\Gamma(p)$ for the values of p given above is found to be

$$\sqrt{\pi},\ \frac{1}{2}\sqrt{\pi},\ \frac{3}{2}\cdot\frac{1}{2}\cdot\sqrt{\pi},\ \frac{5}{2}\cdot\frac{3}{2}\cdot\frac{1}{2}\cdot\sqrt{\pi},\ \frac{7}{2}\cdot\frac{5}{2}\cdot\frac{3}{2}\cdot\frac{1}{2}\cdot\sqrt{\pi}.$$

The probability of exactly n stations being on at the same time

multiplied by the probability that exactly n stations will give a voltage equal to or greater than x equals the probability of obtaining a voltage equal to or greater than x from just n stations.

Hence the summation from $n = 1$ to $n = N - 1$ of these probabilities for a given value of x will give the probability of obtaining a

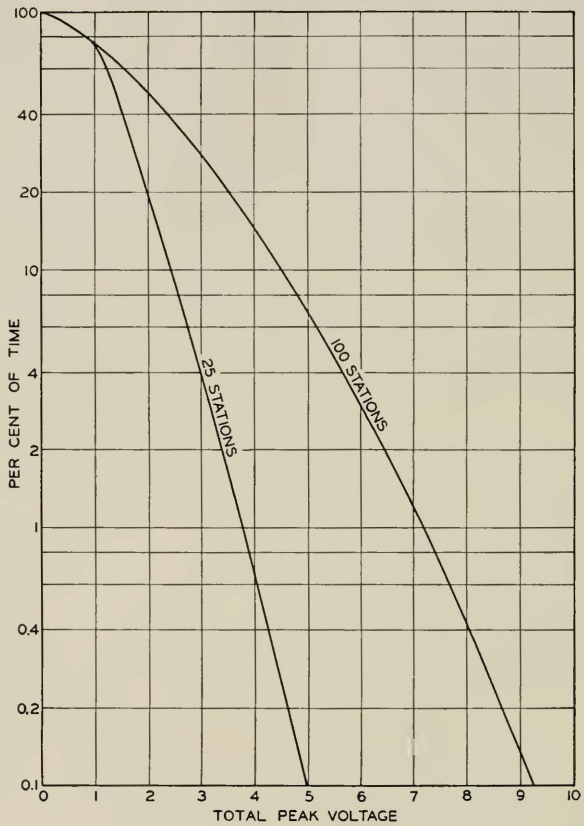


Fig. 21—Voltages resulting from several unit voltages each applied 15 per cent of the time and with random phase and frequency

voltage equal to or greater than x from all of the N stations in the restricted frequency range considered or

$$P_{xN} = \sum_{n=1}^{n=N-1} P_n P_x. \tag{403}$$

In a vacuum tube large negative voltages are equally as important as large positive voltages in producing distortion. Equation (403) has

been derived for positive values greater than x but negative values greater than $-x$ are equally probable and therefore the fraction of the time that the absolute value of voltage is equal to or greater than x is $2P_{xN}$ or

$$P_{|x|N} = 2P_{xN}. \tag{404}$$

Specific cases which approximate the existing conditions of long-wave transatlantic reception have been calculated from equation (404) and are shown in Fig. 21. These curves are based on the following assumptions:

1. That the number of stations lying in the restricted frequency range is $N = 100$ and $N = 25$.
2. That each station contributes unit peak voltage to the input of the radio receiver.
3. That each station has its key depressed 15 per cent of the total time during any day. $K = 0.15$.
4. That transmissions from all stations are random.

Considerable valuable assistance in the preparation of this appendix has been obtained from Dr. F. H. Murray of this department and the authors wish to express their appreciation of this aid.

Oscillographs for Recording Transient Phenomena [1]

By W. A. MARRISON

In this paper, oscillographs developed for recording transient phenomena are described which obtain automatically records of amplitude, wave form, frequency, duration, and time of any electrical disturbance for which they are adapted. Two instruments are described for recording very short or very long transients: these may be used in combination. At power frequencies satisfactory records may be made on film or sensitized paper with a two-watt lamp. The instruments and their performance are illustrated by photographs and oscillograms.

OSCILLOGRAPHS are described which were developed primarily for recording transient phenomena of which the time of occurrence is neither known nor subject to control. The specific apparatus described was designed primarily for recording transient inductive disturbances in communication lines from neighboring power circuits. When the design of this apparatus was begun, there was no satisfactory way for determining the duration, frequency or wave form of such disturbances, although apparatus was available by means of which the approximate magnitude of such transients could be determined, and, by constant supervision, the time of their occurrence. It was with the idea of determining part or all of these factors automatically in a single record that the oscillographs to be described were developed.

Transients in general may be of various types. They may have components in a large range of frequencies, they may occur in a large range of amplitudes and may be very long or very short or intermittent. Attention was directed toward recording devices which would obtain records of any disturbances in excess of a predetermined magnitude regardless of the time of occurrence. For practical reasons it was necessary also to give attention to the cost of operation, the power consumed, and the amount of servicing in operation.

To meet these requirements two somewhat different types of oscillograph were developed. One is capable of making records of short duration having uniform resolution throughout. By its use the wave shape of the first half cycle of a transient is recorded as clearly as that of any subsequent wave. The other instrument makes long continuous records and may be arranged to record a disturbance of any reasonable duration. The former instrument makes records in polar coordinates on a sheet of film rotating in its plane and will be called a "polar oscillograph." The latter records in rectangular coordinates

[1] Presented at the Regional Meeting of the Middle Eastern District of the A. I. E. E. at Cincinnati, Ohio, March 20–22, 1929.

on long strips such as motion picture film and will be called a "continuous-film oscillograph." [2]

Features Common to Both Oscillographs

Since both of these oscillographs were designed for recording the same sort of phenomena and for operating under somewhat similar conditions, they have a number of features in common.

As in most oscillographs the optical system consists of a light source, a mirror capable of being vibrated about an axis in its plane with an amplitude proportional to the signal to be recorded, and a lens system, including the mirror, to form an image of the light source on light sensitive film moved in a direction perpendicular to the plane of vibration of the mirror.

The light source consists of a concentrated filament flashlight lamp placed as close as possible to a pinhole aperture in such a way that the aperture, as viewed from the vibrator side, appears to be completely

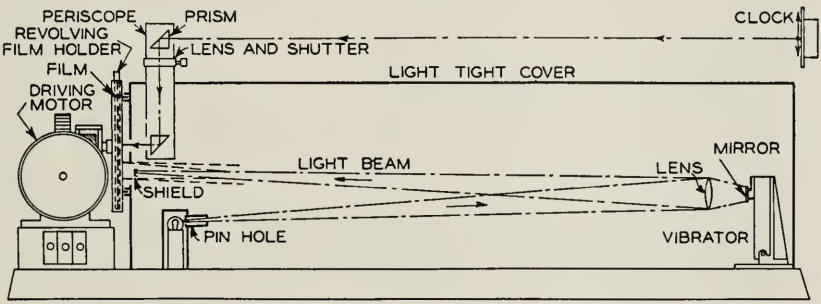


Fig. 1—Essential elements of polar oscillograph.

filled by the lighted filament. No condensing lens is used because of the small size of the bulb which permits the filament to be brought close to the aperture. The filament is brighter than its image and the use of a condensing lens in this instance would waste light unnecessarily by reflection and absorption, and would make the optical system larger.

The vibrator is of the moving-iron balanced armature type similar to a driving element frequently used in loud speakers. The armature is attached by means of a stiff rod to a mirror free to vibrate about an axis in its plane in such a way that, as the armature of the element vibrates, the mirror vibrates at a relatively large angular amplitude. With this type of vibrator it has been possible to employ a mirror half an inch in diameter and still retain a satisfactory frequency range and

[2] This instrument is frequently called a "Movie Oscillograph."

sensitivity. The use of a large mirror makes it possible to use either less sensitive film or a less intense light source than ordinarily would be required for a given recording frequency and film speed. With a half inch mirror it has been found practicable to use a lamp requiring only about two watts for recording on par-speed film.

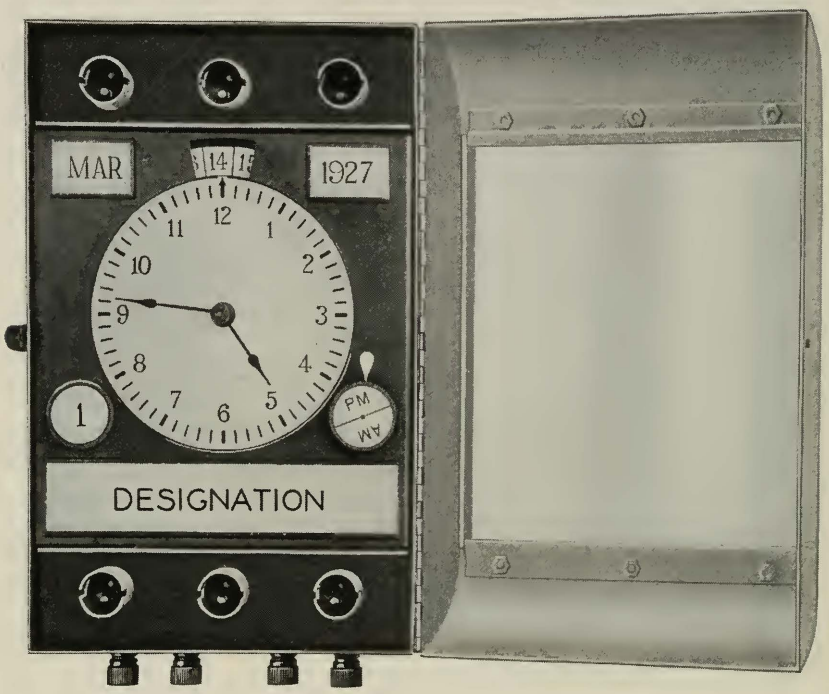


Fig. 2—Calendar clock used with oscillographs. Door is open to show lamp sockets.

A plane mirror is used on the vibrator, and a single meniscus lens mounted in front of it serves, in virtue of the reflection, as a symmetrical lens in forming an image of the pinhole on the film. When greater resolution along the time axis is required than can be obtained with this simple system, a cylindrical lens with short focus is placed in the light path just in front of the film.

Each oscillograph is equipped with a camera for the purpose of photographing a clock on the oscillogram to indicate the exact time of occurrence of the disturbance recorded. Any other information it is desired to associate with the records made by a particular oscillograph may be recorded photographically along with the clock. In several cases calendar clocks have been employed indicating the day and the month, and indicating whether the time is A.M. or P.M.

A schematic diagram of the optical system of the recorder and of the camera, as used in the polar oscillograph, is shown in Fig. 1. Some of the mechanism essential for operation is omitted for the sake of clearness.

One of the calendar clocks used in conjunction with these oscillographs is shown with cover open in Fig. 2. A space is left below the clock face for a card on which may be written identifying or other information relative to records that may be obtained. Lamps are mounted within the cover to illuminate the clock when necessary.

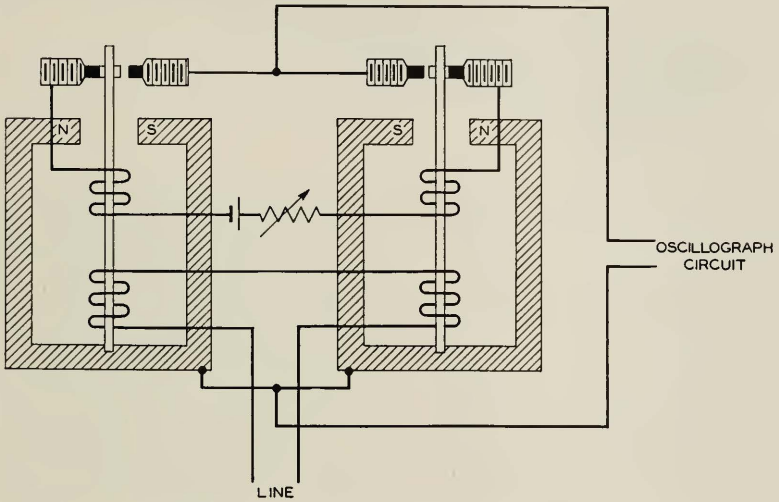


Fig. 3—Circuit of high-speed line-relay.

Both oscillographs are equipped with automatic devices which enable them to make records of transients for which they are intended without the attention of an operator. These automatic features will be described in some detail in the following discussion of the individual oscillographs.

One part, however, a high-speed "line-relay" is common to both. It consists of a pair of high-speed polar relays, the windings of which may be connected into a line in such a way that, depending on the polarity, one or the other will be operated by any pulse of sufficient magnitude. The relays may be so biased that they operate only on pulses in excess of any given magnitude. They are connected so that when they do operate they remain operated until reset by some external means. Contacts on the relays are connected to the apparatus to be controlled so that a single positive or negative pulse will put that

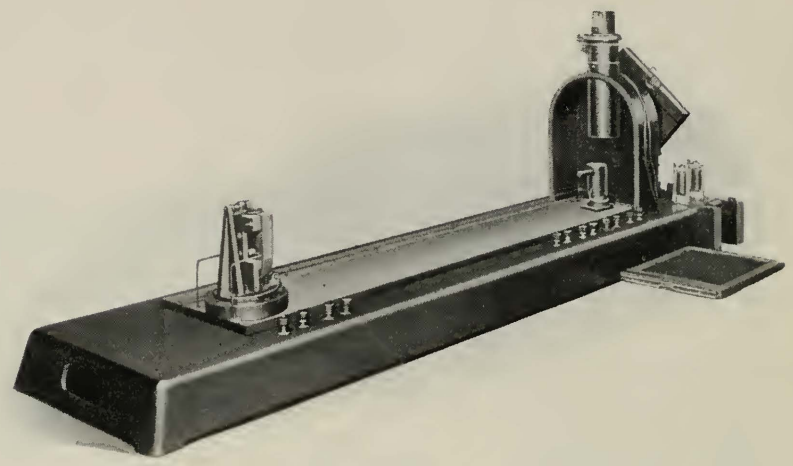


Fig. 4—Polar oscillograph, showing film rotor, periscope, lamp housing, and vibrator.

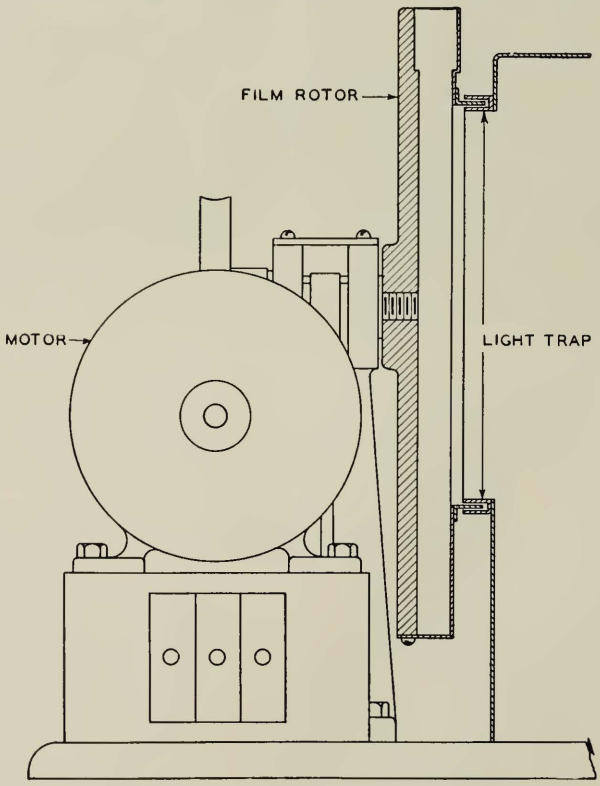
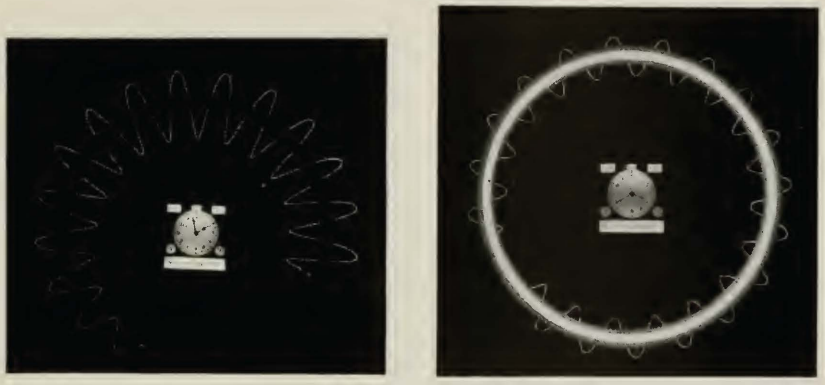


Fig. 5—Scale drawing of rotating light trap.

apparatus in operation. The time elapsed between the arrival of a pulse and the closing of the operating contact is less than 0.01 second. A schematic diagram of the line relay is shown in Fig. 3.

A polar oscillograph is shown in Fig. 4 with the light-tight cover removed to show the optical system. The film is held in a standard film holder in a rotating member at the extreme right of the picture. The use of standard film holders facilitates loading in daylight as in an ordinary camera. The film is rotated by a small motor geared to the rotating member. The rotating member is separated from the remainder of the oscillograph by a circular light trap which permits free rotation while shielding the film from external light. The circular light trap used is illustrated in Fig. 1 and in Fig. 5. With this arrangement films may be exposed for days at a time under ordinary light conditions without appreciable fogging.



A B

Fig. 6—Oscillograms illustrating the use and omission of a light shield over the zero line.

The flashlight lamp is housed in the small light-tight box on the base near the film rotor. Excessive scattering of the light is prevented by a small tube, with a diaphragm near the end, directed toward the vibrator mirror. The vibrator and mirror and the lens of the optical system are mounted on the base near the other end.

The chief value of this oscillograph lies in its ability to record with good resolution from the very beginning of a transient, regardless of the time at which it occurs, and regardless of the angular position of the film at which it begins. To accomplish this, the lamp is lighted continuously during the time a transient is expected and a narrow shield is placed in the light path of just sufficient width to prevent

light from reaching the film when the vibrator is at rest. In this way fogging of the film is prevented during the time when no current is flowing into the vibrator but a record is made of any disturbance of sufficient magnitude to move the spot off the shield. A record made in this way appears like an ordinary oscillogram except that a narrow,

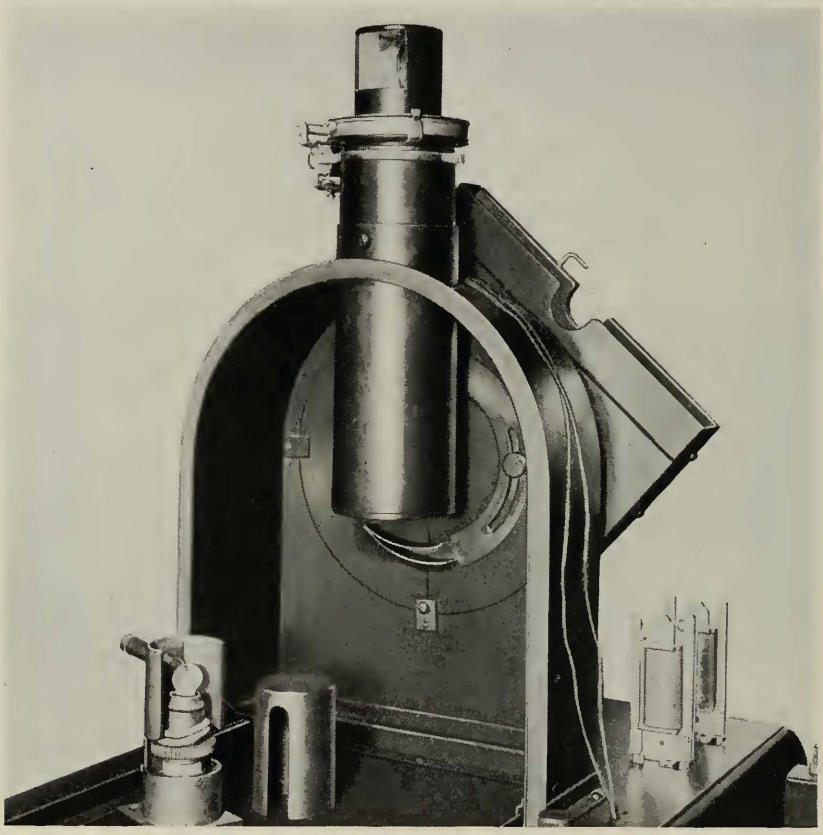


Fig. 7—Shields to prevent fogging on polar oscillograph.

clear space is left where a zero line is usually obtained. This is illustrated in Fig. 6-A.

In the absence of a shield the film soon becomes so badly fogged that an oscillogram made upon it is useless. Fig. 6-B shows the fogging obtained with an exposure of one minute on the zero line without a shield.

If it is desired to record only disturbances in excess of a certain magnitude, the width of the shield may be increased so that no exposure

occurs until the vibrator moves at more than a predetermined amplitude. In Fig. 7 a shield for this purpose is shown which may be adjusted to cover a portion in the center of the record from the width of the spot to about half an inch, or removed from the field entirely. As may be seen from the illustration, this is accomplished by moving the shield in guides about the axis of film rotation.



Fig. 8—Oscillogram illustrating the use of removable shield.

There is a disadvantage in using a very wide shield of the type described. If, for example, a disturbance occurs which is just great enough to be recorded, the major portion of the wave is hidden by the shield and all that can be deduced from the record are the peak amplitude, frequency and time. This difficulty can be avoided readily, however, by attaching a shield to the armature of an electromagnet so that it can be removed from the light path when the magnet is energized. The magnet may be operated by the high-speed line-relay, which is adjusted to operate when the disturbance exceeds a certain amount, in this case the same amount that moves the light spot at an amplitude greater than the width of the shield. The oscillogram shown in Fig. 8 was made with a removable shield of this type. The shadow of the shield is indicated in the first four cycles but does not appear during the remainder of the oscillogram. For convenience in interpreting the results, a zero line is automatically recorded immediately after the recording of the oscillogram.

In order to reduce fogging due to stray light a large shield is placed between the film and the light source, having a vertical slit just large

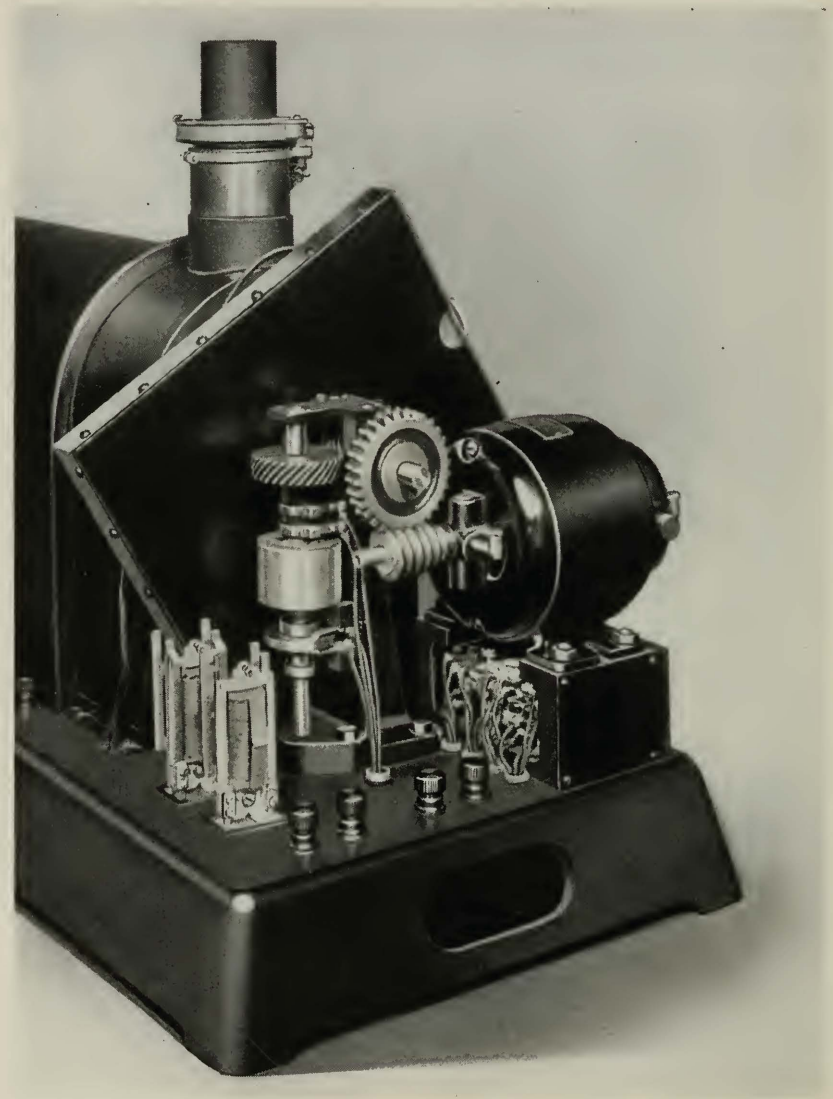


Fig. 9—Device on polar oscillograph to limit recording time.

enough to allow the vibrating beam of light from the mirror to pass through to the film. This device reduces the fogging due to stray

light to a small fraction of what otherwise would be obtained. The vertical slit can be seen in Fig. 7 behind the variable shield. A slight corona-like fogging on either side of the circular shadow of the shield, as shown in Fig. 8, is obtained after a few hours of exposure. In practice a film may be exposed for twenty-four hours or more without the fogging becoming so serious as to obscure a record.

To avoid confusion due to overlapping records a device is used to stop recording after one complete revolution of the film. It is put in operation at the beginning of a transient by the high-speed line-relay and allows recording to continue for one complete revolution, regardless of the angular position of the film at which it begins.

This device is shown on the oscillograph in Fig. 9. A vertical shaft driven at half the speed of the film carries a magnetic clutch fastened rigidly to it and a commutator which idles on the shaft except when engaged by the clutch. This commutator has one insulating segment and one conducting segment, each of angle about 180 degrees. Contacts, controlling the current to a relay winding, are normally on the insulating segment of this commutator, but when the high-speed line-relay operates, the magnetic clutch is energized and the commutator is rotated until the contacts touch the conducting segment, thus operating the relay. One contact on the relay automatically releases the magnetic clutch, preventing further rotation of the commutator. Another contact interrupts the current going to the light source. Since the filament of the lamp is small the light is extinguished in a very short time and, of course, recording is stopped immediately.

After the exposed film has been replaced, and it is desired to put the oscillograph in operation again, the clutch and commutator are restored to their original condition by operating a key which energizes the clutch magnet and releases it again automatically after one-half revolution of the vertical shaft.

If it is desired to make a record covering more or less than one revolution it can be arranged simply by changing the gear ratio between the film shaft and the vertical commutator shaft. The film speed may be changed either by changing the gear ratio between the motor and the film driving shaft or by varying the speed of the motor.

The camera for photographing a clock is shown in Fig. 4. It consists of a lens and shutter, shown at the top, and a periscope consisting of two right-angled glass prisms mounted in the vertical tube. The periscope places the image in the center of the film and since there are two reflections the image will be the same as if none were used. The shutter is equipped with an automatic release which is operated a definite time interval after the beginning of an oscillogram, the time

being determined by means of a slow acting relay or by a sequence switch. It is desirable to stop the film from rotating before photographing the clock which may be done by means of the same relay that turns off the lamp, or by means of a sequence switch.

Main Features of Continuous-Film Oscillograph

A picture of the continuous-film oscillograph is shown in Fig. 10. It differs from the polar oscillograph mainly in the form in which records are obtained. As previously stated, records are made in

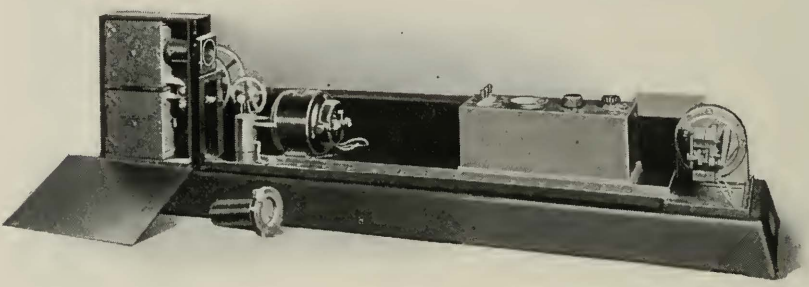


Fig. 10—Continuous-film oscillograph with covers removed.

rectangular coordinates on a strip of film and may, therefore, be of any length depending only on the length of film available and on the size of the storage magazines. The oscillograph shown makes records on motion picture film or sensitized paper of the same width. The film is

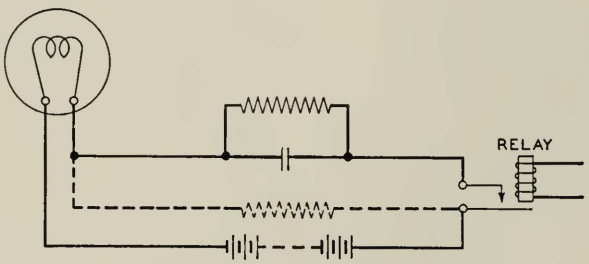


Fig. 11—Circuit for lighting lamp quickly.

stored in standard magazines for motion picture film holding up to 200 feet. It is advanced by means of a motion picture sprocket driven through gears and a magnetic clutch from a variable speed motor. The optical system is practically identical with that used on the polar oscillograph.

With an oscillograph of this type it is not practicable to allow the film to be moving all the time on account of waste of film and the maintenance difficulties. In order to avoid this and still make it possible to begin recording soon after the beginning of a disturbance,

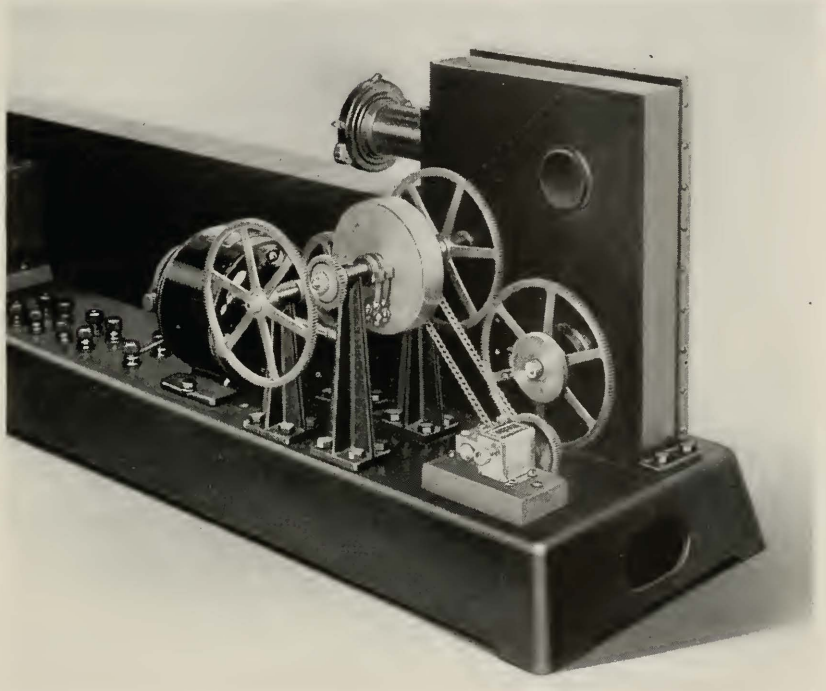


Fig. 12—Quick acting magnetic clutch on continuous-film oscillograph.

the motor and associated gears are left running the whole time during which a transient may be expected, and, when a disturbance occurs, a quick acting magnetic clutch engages the film driving shaft with the motor which puts the film in motion very quickly. The line relay lights the oscillograph lamp at the same time. The whole recording mechanism may be put in operation within 0.02 second, thus insuring a good record of any but a very short transient.

Normally the lamp would require several hundredths of a second to become lighted to full brilliancy if operated at normal voltage. However, with a voltage several times normal and by the use of the circuit shown in Fig. 11, it is possible to bring it to full brilliancy within 0.01 of a second without danger to the lamp. When the circuit is closed by

the relay the condenser is charged suddenly to the applied voltage, the charging current passing through the lamp filament. This current, at the outset, is several times the normal current for the lamp and brings it to full brilliancy quickly. The resistance shunting the condenser has such a value that normal current flows through the lamp filament in the steady state, so, once lighted, the lamp remains at normal brilliancy as long as the circuit is closed. The lamp may be lighted in even less time if a small current is left flowing through the filament continuously in order to keep it hot but not hot enough to be luminous.

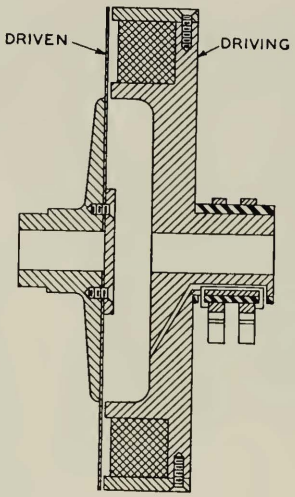


Fig. 13—Sectional scale drawing of magnetic clutch.

A resistance of suitable value connected as indicated by dotted lines in Fig. 11 will accomplish this result.

The magnetic clutch, while especially designed to operate quickly, accelerates the sprocket and film without shock in order to avoid danger of tearing the film and to reduce wear and tear on the mechanism. The clutch is shown mounted on the oscillograph in Fig. 12. It is also shown diagrammatically in Fig. 13 to illustrate its construction and operation. The annular coil in the driving member is connected, through slip-rings and contacts on the high-speed relay, to a battery. When current flows in this coil a steel diaphragm on the driven member is drawn against the annular electromagnet, traction being obtained at the outer edge of the diaphragm. Due to the small clearance between the diaphragm and the electromagnet the diaphragm is drawn into contact very quickly, and due to the small moment of

inertia of the driven member it is rapidly accelerated to maximum speed. Since this is a friction type of clutch, the acceleration is gradual and does not submit the sprocket and film to shock as would a toothed clutch.

The delay in recording after the beginning of a disturbance depends on the time of operation of the high-speed relay plus that of either the clutch or the lamp, whichever is the longer. The relay requires only a few thousandths of a second to operate and both the clutch and the lamp may be adjusted to operate in less than one hundredth of a second. With this apparatus, therefore, it is possible to record all of a disturbance except that part which occurs during about the first 0.02 of a second. If desired, the lamp can be arranged to operate in considerably less time than the clutch, in which case the first part of the record will not be resolved but will indicate the amplitude of the disturbance which is frequently the most desired information. In the case of 25-cycle or 60-cycle disturbances the maximum of even the first half cycle may be recorded in this manner.

The film driving mechanism is arranged so that the film may be advanced at any speed in a wide range, from a few inches per minute to about a foot per second. This is accomplished by means of a set of change gears and by changing the speed of the driving motor, or by both in combination.

As with the polar oscillograph, a camera is included for the purpose of recording the time automatically on the oscillograph film. This camera may be seen in Fig. 10. It is similar to that on the polar oscillograph with the difference that only one prism is used. This has the advantage, when recording is done on paper, that the image obtained through a lens and a single reflection is not reversed. The shutter is equipped with an automatic release that can be associated with slow acting relays or a sequence switch to take care of photographing the clock on the right portion of the film.

Operation

There are a great many ways in which the oscillographs, described above, may be used. An arrangement is described in which two polar oscillographs and one continuous-film oscillograph have been used in conjunction for studying transients which are likely to occur at any time during long continuous periods. It was desired to determine the magnitude with considerable accuracy and at the same time to determine the time of occurrence, duration, frequency and wave form.

The arrangement of oscillographs is shown diagrammatically in

Fig. 14. One of two polar oscillographs is connected in the circuit being investigated so that it is in condition to record the first part of any transient that should occur. A high-speed line-relay associated with it is arranged to put in motion the sequence switch which takes care of a number of operations consisting chiefly in starting the continuous-film oscillograph, in substituting the second polar oscillograph for the first after a certain small time interval, and in operating the camera shutters at the proper times.

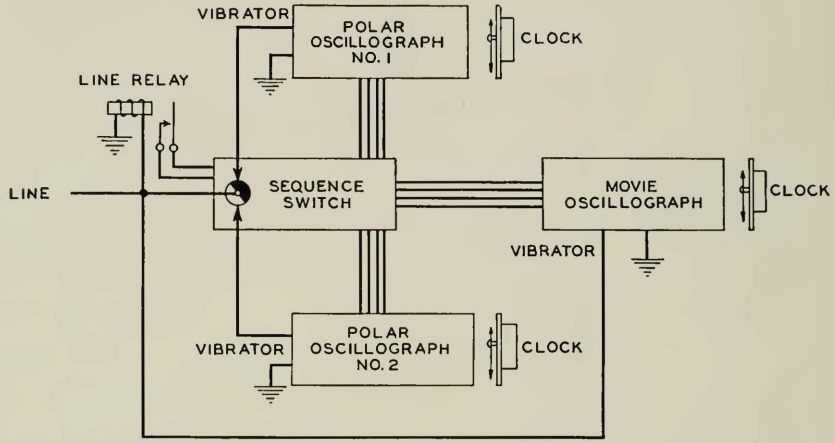


Fig. 14—Arrangement of oscillographs for recording any transient in a line. Two polar oscillographs and one continuous-film oscillograph with control equipment and clocks are arranged so that a transient of any duration occurring at any time will be recorded with a record of the time of occurrence.

Used in this way the first polar oscillograph obtains a record, having considerable resolution, of the first part of a transient while the other oscillograph obtains a record of the complete transient with the exception of the first few cycles which are, however, obtained on the polar machine. The sequence switch is adjusted so that the continuous oscillograph is put in operation before the polar machine stops recording so that, with the two records, complete information of the disturbance may be obtained. Two polar oscillographs are used in this way so that, in case two transients occur close together, a record of one of them will not be lost during the time required for reloading. The sequence switch connects the line to the spare polar oscillograph automatically and gives the operator ample time to make any necessary adjustments on the remaining machine. The arrangement is symmetrical and either polar oscillograph may become the spare.

An example of the record of a transient obtained with a polar and a

continuous-film oscillograph used together is shown in Fig. 18. The polar oscillogram is similar to that shown in Fig. 8. It is obvious that the continuous oscillograph began recording about five cycles after the beginning of the transient while, of course, the polar oscillograph began recording immediately. The long record, however, continues twenty-five or thirty cycles beyond the end of the polar record and shows the manner in which the transient ended. Space does not permit of showing the clock that was photographed on the strip record.

When certain factors are known about the disturbances to be recorded the arrangement may be somewhat simplified. If, for example, it is known that any transient to be recorded will be of very short duration, the continuous oscillograph need not be used. If, on the other hand, it is known that the first cycle or two of the disturbance will be of no importance in the record, the polar oscillograph may be dispensed with.

The continuous film type of oscillograph offers some decided advantages over the polar type in that a large number of records can be made at one loading. Largely because of this it is possible to make the oscillograph entirely automatic in operation, causing it to record, without any attention whatever, all the transients in a circuit as they occur, until the supply of film is exhausted. Such an oscillograph may be left permanently connected into a circuit in which transients are expected, and at the end of any period the film that has been advanced into the "exposed" magazine will show on development records of the magnitude, frequency, and wave form of the disturbances and of the time of occurrence of each.

For automatic operation of the oscillograph a sequence switch of some sort must be used to insure that the various automatic operations are performed in the proper sequence. In Fig. 15 a schematic drawing illustrates the essential elements of such an arrangement. When standing by, ready to record a transient, the motor on the oscillograph is running but does not engage the film advancing mechanism because the magnetic clutch is normally deenergized. On the arrival of a transient the sequence of operation is as follows:

The line-relay operates, and locks in operated position in virtue of the bias current through the outer winding being removed by the opening of the back contact. The lamp is lighted through the resistance and condenser combination RC, and, at the same time, the motor is engaged with the film driving mechanism through the magnetic clutch. Since the vibrator is connected continuously to the line, recording begins immediately. Relay SR operates at the same time that the clutch is energized and starts the motor of the sequence switch

which rotates the cams in the direction of the arrow. After a predetermined amount of film has been advanced, cam 3 rebiases the line relay and restores it to the original non-operated condition (unless the disturbances on the line continue beyond this time). This releases relay SR but the cams continue to turn because the contact operated by cam 1 is in parallel with that on the relay, which allows the motor

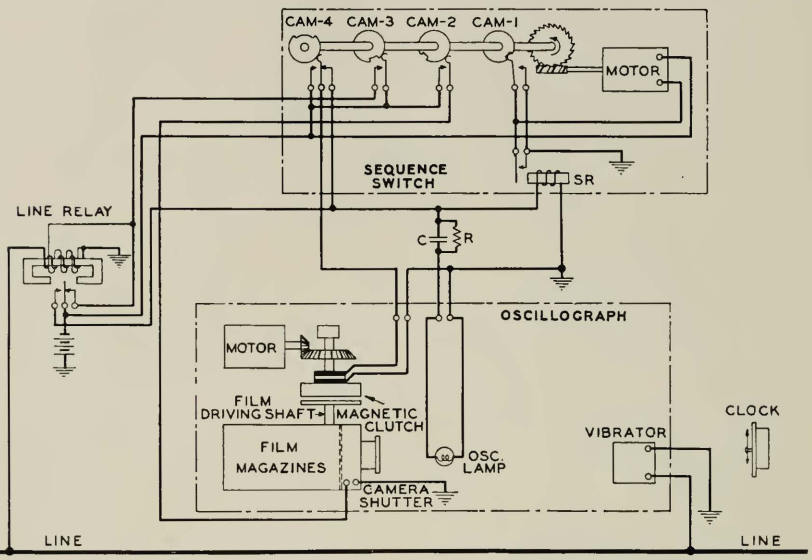


Fig. 15—Schematic arrangement of continuous-film oscillograph with sequence switch, line-relay, and clock, arranged for automatic operation.

to run during one complete revolution of that cam. When the line-relay is restored the film stops and the lamp circuit is opened. After cam 3 stops the recording and resets the line-relay, cam 2 operates the camera shutter which photographs the clock on the film at the end of the oscillograms. As the cam shaft continues to revolve, cam 4 operates the magnetic clutch for a short time without lighting the lamp. This advances the film far enough so that the beginning of the next record will not be superposed on the image of the clock. When the cam shaft has completed one revolution, cam 1 stops the sequence switch motor by opening the contact associated with it.

After this sequence of operations everything is exactly the same as before except that a certain length of film has been advanced into the used-film magazine and a complete record made up on it. The length of the record may be adjusted by an adjustment of cam 3.

In case the disturbance lasts until after the amount of film allotted

to each record has been used, the line-relay will remain operated and the film will continue to be exposed during one whole cycle of the sequence switch. Thus a continuous record can be made of any disturbance, however long, provided there is sufficient film.

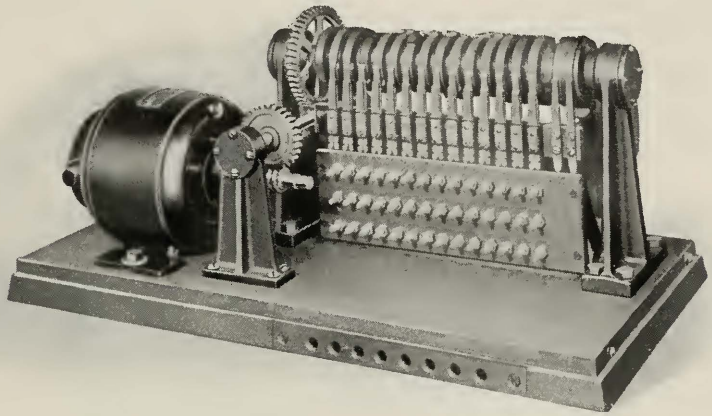


Fig. 16—Sequence switch.

Fig. 16 shows a sequence switch that has been used in the arrangements of both Fig. 14 and Fig. 15. With it a great many arrangements of automatically controlled equipment may be set up besides those described, permitting the oscillographs to be used in many different ways.

A modification of the continuous-film oscillograph which appears to have some novel and useful features is shown in Fig. 17. It is adapted

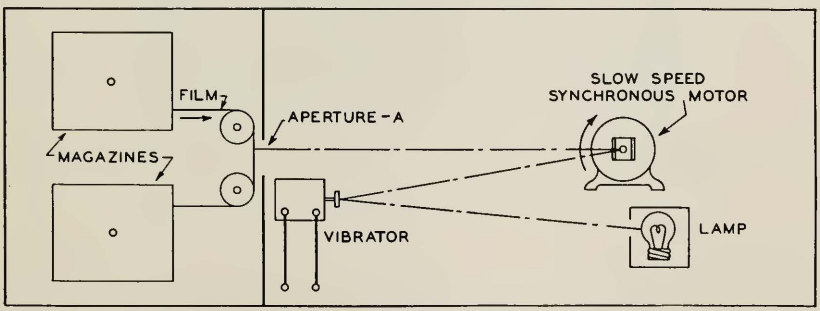


Fig. 17—Schematic drawing of sampling oscillograph. This oscillograph records one wave out of a number at regular intervals, say one cycle in sixty, with considerable resolution in order to record slow variations in wave form.

especially for sampling a wave at regular short intervals instead of making a continuous record or merely a record of unusual disturbances.

It involves, in addition to the usual optical system and means for advancing the film, an additional mirror in the light path between the vibrator and the film, rotating synchronously with the current or voltage to be recorded, about an axis perpendicular to both the direction of motion of the film and the axis of the vibrator mirror. The function of the rotating mirror is to sweep the light beam along the oscillograph film past an aperture A in such a way that the effective film speed during exposures is many times the actual film speed, and to permit of exposure during only a small part of the total time.

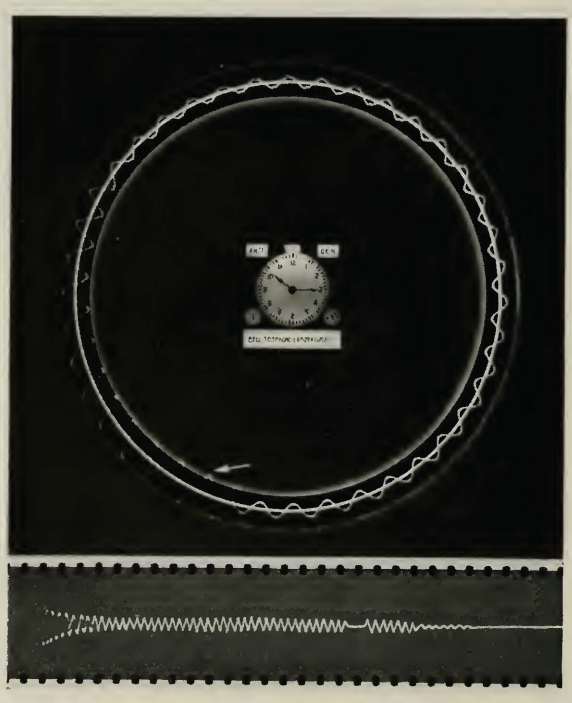


Fig. 18—Sample of records made by polar and continuous-film oscillographs used together.

As an example, suppose that the mirror makes one revolution in two seconds and that the wave to be recorded has a frequency of 60 cycles per second. If the distance of the rotating mirror from the film is 8.5 inches, one cycle of the wave recorded will be spread over approximately one inch of film. If a rotating mirror with a single facet is used, and if the aperture is just one inch wide, the actual film speed should be one inch in two seconds and every one hundred and twentieth wave will be recorded. If two facets 180 degrees apart are used on the

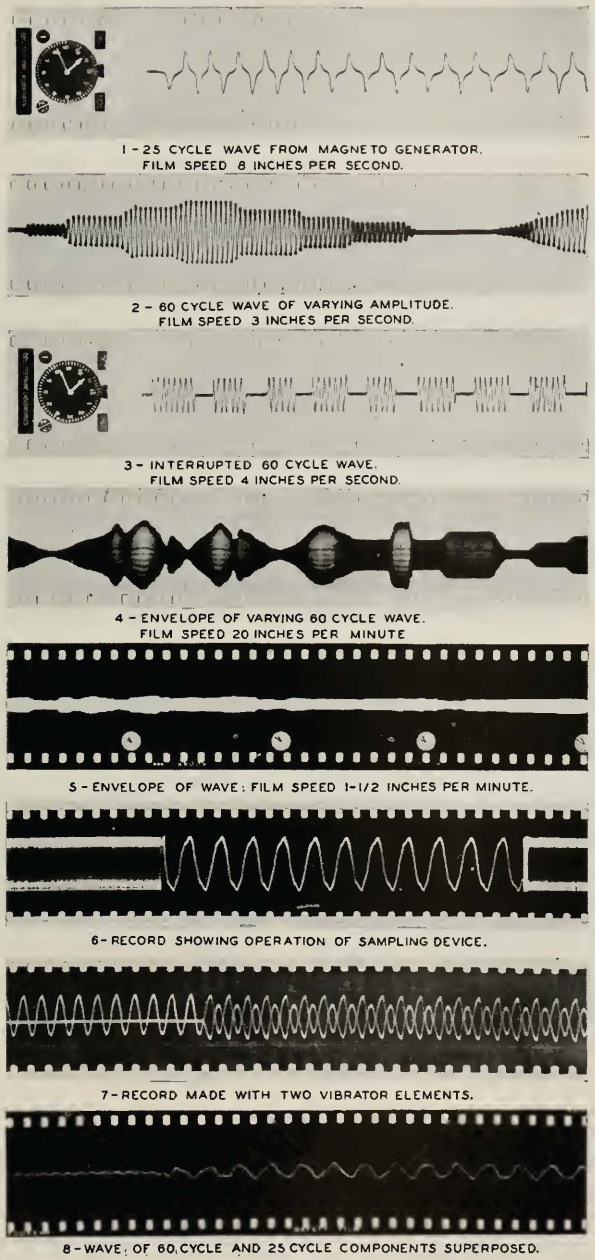


Fig. 19—Samples of continuous oscillograms.

rotating mirror, and if the film speed is doubled, one wave in every 60 will be recorded.

Individual facets on the revolving mirror may be inclined to the axis of rotation in order that successive exposures may be made from different vibrator elements. For example, three facets suitably mounted could be employed to show in succession sample waves of

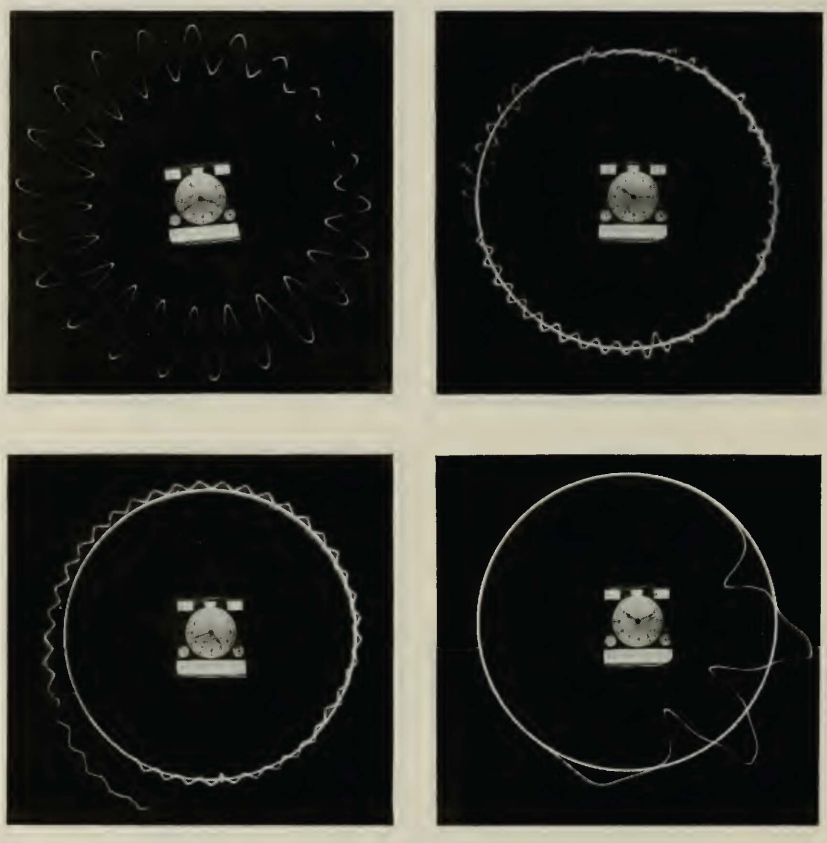


Fig. 20—Samples of polar oscillograms.

current from the three circuits in a three-phase line, the motor driving the mirror being operated synchronously in phase with the three-phase voltage.

The advantage in this recording method lies in the ability to obtain a good record of slow changes with good resolution and without the use of a large amount of film.

Another method of sampling which gives a somewhat different kind

of information may be used with a continuous-film oscillograph with the usual form of optical system. The film is run at very slow speed in order to obtain normally an envelope of the wave. At intervals the speed of the film is increased to a value sufficient to resolve the wave and show the actual wave shape. A record of the time may be made on the film by photographing a clock at regular intervals, say once a minute. A record made in this way is shown in Fig. 19, No. 6.

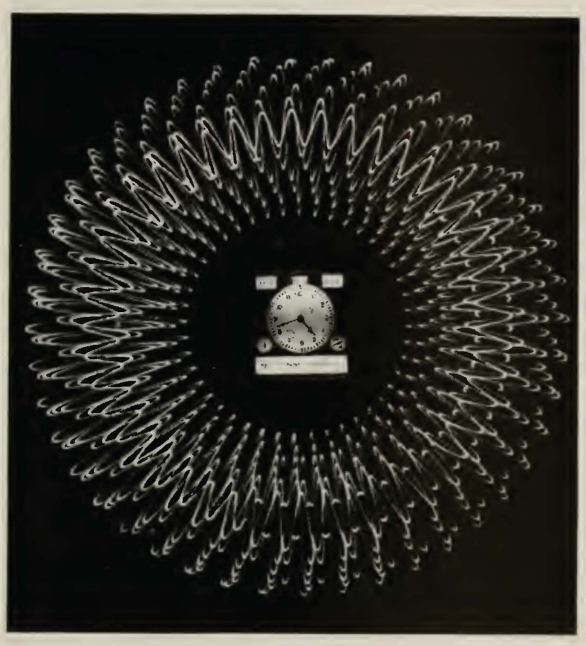


Fig. 21—Amplitude calibration of polar oscillograph.

Frequently there is an advantage in recording two or more variables simultaneously. A record obtained with two vibrators is shown in Fig. 19, No. 7.

Performance

The limitations of an oscillograph lie mostly in the vibrator and, to a smaller degree, in the optical system and photographic emulsion used. The frequency characteristic of the vibrator up to 800 cycles is quite uniform, permitting records of disturbances having components in this range to be made with little distortion. This range includes the first 13 harmonics of 60 cycles and the first 32 harmonics of 25 cycles.

The vibrator may be wound to have any impedance in a wide range. If wound to have a high impedance it is especially suited for recording

voltage waves, and if wound to have low impedance it is better suited for recording current waves. The sensitivity, as usually expressed, in millimeters deflection per milliampere of input, varies approximately as the square root of the impedance of the winding. The ratio of deflection to input is constant over a considerable range due to the balanced structure of the motor element.

As noted previously the oscillographs described are intended for recording in a comparatively low frequency range. In the range given there has been no difficulty in obtaining good records with a two-candle-power flashlight lamp. This of course, is principally due to the large size of the mirror on the vibrator.

Samples or records made with the oscillographs described are shown in Figs. 6, 8, 18, 19, 20 and 21. Those in Figs. 6, 8 and 18 have already been mentioned. Fig. 19 shows eight continuous oscillograms and Fig. 20 shows four polar oscillograms illustrating some of the possibilities of these instruments. Fig. 21 is a response calibration at constant frequency of a polar oscillograph.

A number of field applications of oscillographs of both types have been made with satisfactory results. In some cases where cooperative studies were being made, the oscillographs have been used for recording transient neutral currents in power systems as well as to record voltages induced in telephone circuits by power system transients. Experience with the oscillographs in these field installations has suggested a few improvements of a mechanical nature and certain rearrangements of parts to increase the convenience of operation. These changes are now being embodied in a new design. It is hoped that it will be possible in a later paper to describe these features and to give the results of field experience more fully than can be done at this time.

Contemporary Advances in Physics, XVIII

The Diffraction of Waves by Crystals

By KARL K. DARROW

This is an elementary introduction to the phenomena of diffraction of waves by crystals, one of the most striking and important discoveries of the last twenty years of physics. These phenomena have proved that X-rays and electrons are partly of the nature of waves, and have supplied the best available methods of measuring their wave-lengths; while on the other hand, the study of the diffraction-pattern of a crystalline substance makes it possible to determine the arrangement and the interrelations of the atoms with a precision and fullness heretofore unimagined, which has already yielded knowledge of great value in all the fields of science and promises immeasurably more.

THE diffraction of waves by crystals was discovered in 1912, the very year in which the first of the revolutionary new theories of the atom was being thought out by Bohr. But while since then the atomic theory has undergone mutation after mutation—until one can hardly guess any more what is stable and what is unstable, what it is expedient to retain and what should be forgotten— the consequences of that other discovery have steadily and serenely broadened out. Already they have penetrated into more fields of science than the deductions from the new atomic theories. Eventually their effects, not only on physics but on mineralogy and chemistry and engineering practice and even on biology, may well become so great that diffraction by crystals will prove the most valuable instrument for research which the physicists of our time have presented to the world.

The discovery was not an accident, but a rarely perfect example of theoretical foresight. A mathematical physicist, von Laue, was pondering the theory of diffraction by ruled gratings and the other standard instruments of optical research. He was at a university (Munich) where Roentgen was professor and interest in X-rays was intense. There was a controversy then over the question whether these rays are waves or corpuscles. In those days, the antithesis was absolute; people thought that either answer must exclude the other; they did not realize that in ten or fifteen years they would be accepting both. Towards 1912 the weight of evidence seemed to be forcing the wave-theory from the scene. There was however one piece of evidence which could be interpreted in its favor, provided that the wave-length of the rays was of the order 10^{-8} centimeter. It was also known,

though the knowledge was at that date very recent, that the number of atoms in a cubic centimeter of an ordinary crystal is such that the average distance between them must be of that same order. It was also known, as it had been for many years, that the atoms of a crystal must be arranged in a regular order—a pattern, a network, or a lattice; like soldiers on parade, except that the atoms parade in three dimensions instead of two. Laue was aware of all these facts; and one day it occurred to him, taking them all together, that if a beam of X-rays was truly wave-like a crystal would *diffract* it—would split it into a multitude of diverging beams, themselves grouped in a pattern so curious and so symmetrical that if such a pattern were indeed observed it could not be fortuitous, but by itself must prove the assumption.

The experiment was performed by two of Laue's colleagues on "the experimental side," Friedrich and Knipping, with a crystal of zincblende and a beam of X-rays from an ordinary X-ray tube. The multitude of diverging beams made their appearance: the diffraction pattern was exactly as predicted.

From this magnificent point of departure the advance was early and rapid, in two directions. Crystals being able to diffract X-rays, the phenomena could be used as sources of information either about the rays or about the crystals. The former field was dominated the sooner. In the fifteen years since the discovery, the technique of using crystals to analyze X-ray spectra and measure the wave-lengths of X-rays has been carried near to perfection. Nearly all the rays which atoms can emit from their electron-shells, many of those which proceed from their nuclei, have now been measured; and the advantage to atomic theory is immense. It is true that one can no longer say that except for diffraction by crystals we should not know that the X-rays are wave-like or what their wave-lengths are; for now physicists are beginning to map the X-ray spectra with optical ruled gratings. But the crystals were the first to present us with these data, and in most cases, I suppose, they are still the best.

By contrast, the field in the other direction—the exploration of the arrangement of atoms in crystals by means of their diffraction patterns—seems unlimited. Newton's "ocean of undiscovered truth" is not too strong a metaphor. The crystalline state, it transpires, is universal. It is not confined to the lovely glassy specimens of the mineralogical museum, with their smooth facets, sharp edges and pointed pyramids—the jewels of Nature, from which the jewels of art are made by perfecting or perverting the original design. The vivid geometry of such as these is a signal of a regular, a "crystalline" ar-

rangement of the atoms; but where the advertisement is wanting, the inner order may be none the less precise. Nearly every solid substance owns it; metal, brick, stone and sand, wood, cotton, wool and bone approach in varying degrees to crystalline perfection; so do films of grease and films of liquid, and even in the middle of a liquid mass there are traces of regular arrangement. The diffraction patterns disclose all this, revealing the fine details of crystalline structure even where the eye sees nothing but a shapeless mass.

Even with the beautiful finely-formed crystals of the minerals in museums, the diffraction-pattern teaches more than the crystallographer could learn without it. I would not disparage the crystallographers. Perhaps there are few physicists who realize how far they went before the time of X-rays, and certainly any who thinks that it was Laue's work which showed the world how to tell whether a crystal is cubic has specialized in his science not wisely but too well. Organic chemists also made many inferences about the arrangement of atoms in large organic molecules, which the X-rays are now beginning to verify. But the X-rays in these few years have carried us clear beyond the farthest reach of inference to which chemist or crystallographer could have aspired.

Another service of diffraction is its disclosure of the ways in which the tiny crystals making up an ordinary piece of metal are distributed, their orientations in particular; these are liable to variations whenever the metal is twisted or extended or rolled or hammered or annealed, and it may some day be possible to explain from them the variations of the mechanical properties of the mass.

Yet another service of diffraction, and a very great one to the physicist, is the information which it gives about the individual group of atoms—sometimes indeed about the individual atom—which is repeated over and over again to form the crystal. The beams of the diffraction-pattern shooting off in their various directions may be regarded as the beams proceeding in those same directions from any individual group of atoms, tremendously amplified by the cooperation of the other groups which go with it to make up the entire crystalline network. The amplification-factor can often be estimated; and dividing the observed intensities by it, one obtains an idea of the diffraction-pattern which one group of atoms by itself would form, and from this in turn may infer something about atoms.

Diffraction-patterns are not formed exclusively with X-rays; crystals may build them out of waves of another sort. Towards 1924 Louis de Broglie suggested that electricity and matter are partially wave-like in nature. The philosophy of physicists had changed since 1912,

and it was no longer necessary to lay down an "either . . . or" alternative. It was conceivable that in spite of all the evidence that a stream of negative electricity through a vacuum consists of particles, yet in some ways it might act as though it consisted of waves. In 1925 Elsasser did remark that such a stream might be diffracted by a crystal; for the wave-lengths which de Broglie had assigned to electrons, moving with such speeds as are customary in technical vacuum tubes, were again of that order of magnitude 10^{-8} cm., which had suggested to Laue that X-rays might be diffracted by crystals. Early in 1927 Davisson and Germer looked for the diffraction-pattern of negative electricity with a crystal of nickel, as Friedrich and Knipping nearly fifteen years before had looked for that of electromagnetic radiation with a crystal of zincblende. The techniques were very different, and for a time there was confusion due to the refraction of the electron-waves in the crystalline substance; but even at the start they found a pattern very like that which was predicted, and when the influence of refraction was understood, the discrepancies were cleared away. So they proved that negative electricity is partly of the nature of waves, and initiated the spectroscopy of electrons; and in examining how the diffraction-pattern was affected by films of gas on the surface of the crystal, Davisson and Germer were the first to perform a crystal analysis by electron-waves.

The accepted model for an ideal crystal—accepted now these last two hundred years—is an array of objects or particles much too small to be seen, all exactly alike, all oriented exactly alike, and spread out in three-dimensional space with perfect regularity of arrangement. These particles are marshalled in ranks and files like soldiers on parade, except that the parade is in three dimensions instead of two, as if on every floor of some colossal building a regiment were drawn up. Were the arrangement only in two dimensions, I could find numberless other examples—the pattern of printed wallpaper, a chessboard, the meshes of a handkerchief, the array of jacks on a telephone switchboard, the sections or the townships into which mid-western prairie country is divided, the unvaried multitude of windows on the walls of many a skyscraper. But in three dimensions the only similes which present themselves are a honeycomb, and cannonballs piled in a heap such as are set around old war memorials, and the girders of a steel-frame building as we see them before the walls cover them over. All these pictures fall very far short of suggesting the millions upon millions of particles which are conceived to constitute even the smallest of visible crystals, or their continual trembling in thermal agitation.

The particles are said to be arranged upon a *lattice*. The lattice is

an abstraction, like a coordinate-frame, or the network of meridians and circles of latitude which intersect upon a map. It is usually conceived as a network of three sets of parallel planes, the planes of any set following one another at even intervals of spacing. For convenience of drawing, each plane is sketched as a network of two sets of parallel lines (Fig. 1). The intersections of three planes are the

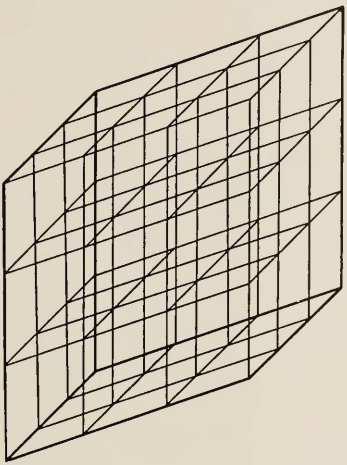


Fig. 1—A space-lattice. Intersections of three lines are lattice-points.

lattice-points; the intersections of two planes are *axes of the lattice.* Axes of the lattice run in three directions, and along each there is a constant spacing from one lattice-point to the next. The three may be at right angles to one another, and the spacings along all three may be the same, in which case the lattice is *cubic;* but this need not be the case. All these ideas are required to make definite the notion of *regularity of arrangement* which as I have mentioned is the distinction of a crystal. They will probably become clearer in what follows.

Around each lattice-point a particle is placed. I say *around* rather than *at,* for it is desirable to think of the particles as rather bulky, each containing a lattice-point at some definite place within itself, and filling an appreciable part of the space extending from that point to its neighbors. Moreover it is desirable to think of them as quite irregularly-shaped objects, not as spheres, the way they are drawn in Fig. 2. Some crystals do have such properties that the picture of spherical particles is nearly adequate; but usually, if we were to assume that the particle has the full and complete symmetry of the sphere, we should be restricting the model to such an extent that it could not be adapted to the properties of the actual crystals. The observations

of crystallographers on the forms which crystals assume and the ways in which they act on light lead to conclusions about the symmetry of these elementary particles; and it is found that while they usually have some degree of symmetry, they do not have that full degree which the sphere represents. In later sketches, therefore, I have followed Bragg in representing the particles by perfectly unsymmetrical figures, like large commas (Figs. 4 and 5). Only, they should be unsymmetrical in three dimensions instead of only two; the reader may conceive each comma as being rough on one side, smooth on the other.

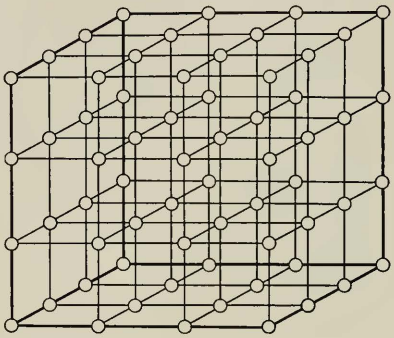


Fig. 2—A cubical space-lattice, with particles of full spherical symmetry indicated at the lattice-points.

The next obvious question is: what is the relation between these particles, and the atoms of the crystal if it is a crystal of some chemical element, or the "molecules" if it is a crystallized chemical compound? One might expect that they would be the same; but one would usually be wrong. When for instance gaseous potassium or argon condense into crystals, it takes four atoms of the argon gas to make up the elementary particle of the argon crystal, while that of the potassium crystal consists of two atoms.[1] With chemical compounds the crystal particle may be the molecule, part of the molecule or a group of several molecules. I will consequently use for it hereafter the colorless name *atom-group*, with occasional variation by the conventional but not very descriptive term *unit cell.*

Now to visualize the diffraction-pattern, let us conceive the crystal set at the centre of a transparent bulb painted inwardly with some fluorescent matter, so that wherever X-rays or electrons fall upon its

[1] Provided that we conceive the lattices as cubic, which is customary. There happens to be in these cases something like what in other fields of physics is called a "degeneracy": one and the same lattice may be viewed either as cubic or (with a different choice of axes) as non-cubic, and with the latter conception the "particle" turns out to be a single atom.

inner surface it will shine. The crystal should be very small in comparison with the bulb, and the incident beam of waves extremely narrow. We suppose that this beam comes in at a window, and follows a diameter of the bulb, and the unscattered part after flowing through the crystal goes out through another window. If the rays are plane-polarized X-rays, the electric vector will remain constantly parallel to some direction at right angles to the beam. If they are unpolarized X-rays, the electric vector will run or swing rapidly around in the plane at right angles to the beam. Thus far it is customary to use in crystal analysis X-rays which either are unpolarized, or have the slight degree of polarization usually imprinted on such rays at their excitation in a discharge-tube or a Coolidge tube.[2] We do not positively know whether electron-waves are polarized, but we cannot alter their degree of polarization whatever it may be, and it seems probable that they are not.

This "imaginary bulb" is very nearly realized in practice, although instead of a fluorescent screen it is customary to use a photographic film on which the imprints of the diffraction beams are permanent spots. The film is usually flat instead of spherical, so that the rings presently to be mentioned are distorted from circles into ellipses. Often an ionization-chamber (for X-rays) or a Faraday chamber (for electrons) is swung around in arcs over a spherical surface centred at the crystal, and the current which it reports is plotted as a function of its position; the curves then display peaks wherever the chamber is so placed as to capture a beam. In what follows I shall use the terms "diffraction beam," "diffraction peak" and "diffraction spot" almost as synonyms.

On the inner coating of the bulb, then, we shall see the diffraction-pattern of the crystal. It will be an assemblage of luminous spots arranged in a symmetrical array recalling the symmetry of the crystal lattice. In making this statement I am anticipating what is presently to be proved, or rather to be deduced from the fact that the atom-groups of the crystal are marshalled on a lattice. Indeed, if the incident waves are monochromatic or nearly so, we shall not see even spots unless by happy accident or by a careful choice of wave-length. Most waves are not diffracted by a three-dimensional crystal lattice. If we could reduce our crystal to a single plane of atom-groups forming a two-dimensional network, there would be a pattern of spots for any wave-length whatever. If we could isolate a single row of atom-groups, there would be rings of light on the coating of the bulb,

[2] Because the electrons which produce the rays in falling onto the target of the tube are all moving along parallel lines when the impinge upon it.

instead of only spots. And if we could remove all but one of the groups, and then in some magical way magnify the luminescence which the waves that this survivor scatters produce at the wall of the bulb, then we should see the wall shining all over with a continuous brightness, sinking to zero perhaps nowhere, perhaps at occasional points.

Reverse the process, starting from the solitary atom-group. The intensity of the scattered waves of which it is the source varies continuously with direction, and the brightness of the wall of the bulb varies correspondingly from point to point. If this brightness is proportional to that intensity, its distribution over the spherical wall is the scattering-pattern or diffraction-pattern of the atom or group of atoms. However it is immeasurably too faint to see.

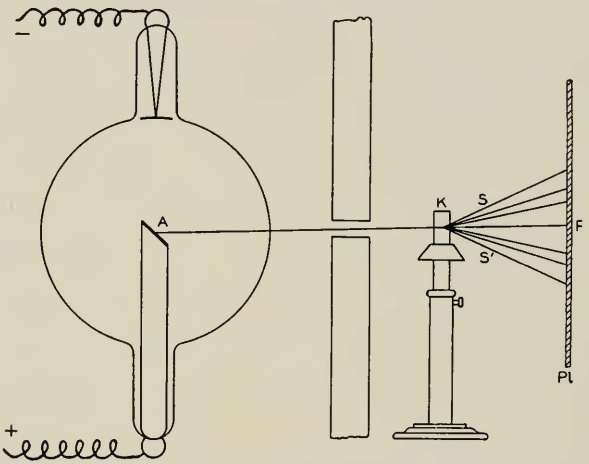


Fig. 3—Illustrating how a beam of waves is split by a crystal (at K) into diffraction beams forming spots on a plane screen (P). (Apparatus of the first experiment on X-rays by Friedrich and Knipping.)

Add now to the solitary group a great number of similar and similarly-oriented groups of atoms, forming altogether a long and evenly spaced row. The luminescence now fades out everywhere except over certain rings upon the wall, but along these rings it is enhanced. The newly-added atom-groups have amplified the waves scattered by the original one along the directions leading to these rings; in compensation they have effaced those scattered in other directions. These are effects of interference.

Add next a great number of such rows of similar and similarly-oriented groups, forming altogether a lattice in a plane. The rings now all fade out except for occasional spots, which however are much

brighter than they were before. Interference between the waves scattered from the various atom-groups has exalted the brightness of certain points on the wall of the bulb, to the detriment of all the rest; it has amplified the diffraction-pattern of the single group in certain directions, and destroyed it in the others. The spots due to the atoms of a single plane have been observed with electron-waves, but never with X-rays (Figure 6).

Add finally a great number of such planes, thus building the three-dimensional crystal. Now except for certain wave-lengths the spots vanish altogether. For those exceptional waves, however, they are intensified into visibility. One group of atoms by itself would have produced a complete diffraction-pattern—at least we suppose that it would—but never one intense enough to be perceived. But when it is joined with an enormous number of its peers in a lattice, they all conspire to enhance not indeed the entire pattern, but the intensities at certain of its points. The physicist, if he varies the direction in which the rays fall upon the crystal or the wave-length of the rays or both, can observe a great number of these intensities which the crystal has so obligingly amplified for him; and out of them he can reconstruct the entire diffraction-pattern of the individual atom, which but for this amplification would have been forever out of his reach.[3]

I must not leave the impression that the diffraction-pattern of the atom-group is amplified equally in all the directions in which the lattice amplifies it at all. Amplification depends on direction. If the observer sees two spots produced by diffraction-beams inclined say at 45° and at 60° to the primary beam, he is not to infer that the ratio of their brightnesses is the ratio of the intensities of the waves scattered at 45° and at 60° by the individual group. A correction-factor must be employed to translate one ratio into the other. Later on we will consider this factor. Meantime, not forgetting it but not yet taking it into account, we will calculate the directions in which amplification occurs.

We start as before from the solitary atom-group. In Fig. 4 this is depicted as a two-dimensional figure, irregular and utterly without symmetry. It ought to be conceived as a three-dimensional, perfectly unsymmetrical mass. This depiction is meant to imply that if a stream of waves strikes the atom-group on any side, the intensity of the waves scattered in any direction—what above I called the diffraction-pattern of the group—is or may be a thoroughly unsym-

[3] If the atoms or atom-groups of a substance would orient themselves all alike without at the same time spacing themselves at regular intervals, and without being too much crowded together, the entire pattern would be amplified instead of certain spots only.

metrical function of direction. Also it may depend on the side of the group on which the primary waves impinge, or, in better words, on the direction whence they come. Of course, in any particular case, it might turn out to be a symmetrical function of direction, or it might be the same function whichever the side of the atom which the primary waves encounter; but we should not rely in advance on either of these simplicities. I must qualify this statement somewhat: the crystallographers have their ways of learning more or less (and sometimes a good deal) about the symmetry of the atom-group, and thus foretelling certain aspects of its diffraction-pattern. Moreover in many cases it is possible to make in advance a good estimate of the absolute intensity of the scattered waves. It will do no harm to remember these encouraging facts; nevertheless it will be best for us to keep our ideas fluid by supposing an atom-group of absolute asymmetry, the scattering-pattern of which is one of the ultimate objects of the quest.

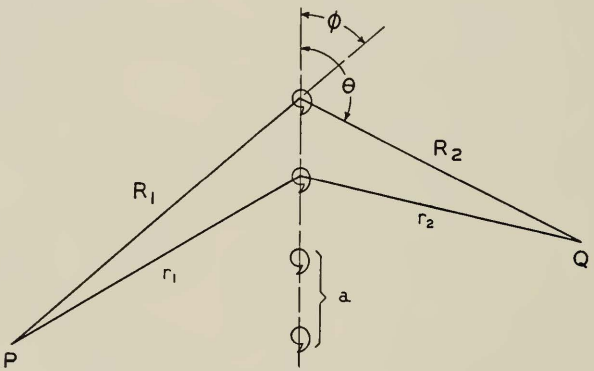


Fig. 4—Illustrating diffraction by a single file of atom-groups.

I must state with all emphasis that whatever may be foretold or measured in the scattering-pattern for electromagnetic waves (X-rays) need not always be valid in the scattering-pattern for electron-waves.

Consider now the amplification which ensues when other atom-groups are added to the first one, to form with it an evenly-spaced row.

Denote by a the distance between corresponding points in adjacent groups (Fig. 4). Suppose the primary waves to emanate from a point P, while the observation of the scattered intensity is made at a point Q (on the wall of the bulb, to continue the picture). In practice the distance a is so submicroscopically small, that P and Q are practically infinitely far away; yet it is easiest to begin by thinking of them as nearby, and passing to the limit. Designate then by R_1, R_2 the

distances from any atom-group A to P and Q, by r_1 and r_2 the distances to these points from the adjacent group B. We shall not assume that R_1, r_1 are coplanar with R_2, r_2, though owing to the flatness of the page the drawing must make them seem so. Denote by ϕ, θ the angles between the direction of the row of atom-groups and the directions in which the primary and the scattered waves advance, respectively—these latter being the directions from A *away from* P and *towards* Q, respectively.

Under what condition will the waves scattered by the atom-groups A and B reinforce one another best at Q? Best reinforcement will occur, greatest enhancement of the effect of either atom by the presence of the other, when the waves from both arrive at Q in identical phase. This will occur when Q is so located that the path from P to Q *via* A either is equal to the path from P to Q *via* B, or differs from it by an integer number of wave-lengths.[4]

$$(R_1 + R_2) - (r_1 + r_2) = (R_1 - r_1) + (R_2 - r_2) = n\lambda; \qquad n = 0, \pm 1, \pm 2, \ldots \quad (1)$$

Let P and Q recede to infinity; in the limit:

$$R_1 - r_1 = a\cos\phi, \qquad R_2 - r_2 = -a\cos\theta \quad (2)$$

and the condition for optimum reinforcement at Q is this:

$$a(\cos\theta - \cos\phi) = n\lambda. \quad (3)$$

In all directions for which θ satisfies this equation, there will be maximum amplification of the diffraction-pattern of A (or B) by the presence of B (or A). For every such value of θ, there will be a ring on the wall of the bulb. If the two atom-groups by themselves could make the wall fluoresce brightly enough, we should see annular fringes. They would be broad and hazy, for though the cooperation between the scattered waves is best at the definite angles determined by equation (3), it is also very good over quite a range of nearby angles. Equation (3) would give the locations of the central rings of the broad fuzzy bright fringes. Thus, when visible light is sent through a pair of parallel similar slits in a screen, one sees hazy fringes superposed on the diffraction-pattern which either slit by itself can produce; and the formula analogous to equation (3) locates the central lines of these.

[4] This statement implies the tacit assumption that there is a constant phase-difference (whether it is zero or not is of no importance) between the primary waves striking an atom-group and the scattered waves leaving it—the very important assumption of *coherence*.

This allusion to parallel slits in a screen will simplify the next steps. It is well known that when to a pair of parallel slits, new ones just like them are added at equal intervals one after the other, the fringes are not displaced. The bright fringes shrink, the dark ones widen, but their central lines remain unshifted. As more and more slits are added, as more and more lines are ruled on a metal surface to constitute a grating, the dark fringes encroach steadily on the bright ones, and it becomes easier to locate the central lines of these latter with precision. As they grow narrower, they brighten, the energy which was lavished over a wide angular range being gathered into a small one as it is progressively increased by the addition of new slits or rulings. So there is a double gain. In the limit, nothing remains but the central lines, and these are brilliant. And in the limit, these lines are still located where the formula derived for only two slits predicted that the maxima of brightness should be found.

Now in the same way, when to a pair of like and likewise-oriented atom-groups additional such groups are added so as to form an evenly spaced row, the annular fringes on the wall of the ensphering bulb are not changed in location but in distinctness. The bright fringes contract into brighter rings, the dark ones broaden into (relatively) dark bands. *The multitude of the atoms sharpens the diffraction-pattern.* In the limit, the bright rings are very sharp and brilliant, and they are still located at exactly the angles predicted by equation (3) derived for two atom-groups only. Remember however that a ring may not be equally bright all around its circuit. It is only an enhancement of the scattering-pattern of the individual atom-group; and this in general will vary from one point to another.

So much for the single row or file of groups of atoms! We must now pass to two dimensions, and predict the diffraction by a plane in which groups are arranged in a network. In the plane, rows of atoms lie side by side at equal spacings. We might start with one of the rows, and estimate how its diffraction-pattern is amplified by the cooperation of a second and then a third and a fourth and eventually an infinite number of added rows laid parallel with it at equal intervals. Owing to this equality of intervals and the new periodicity which it entails, the diffraction-rings of the pattern of the individual rows will be amplified not uniformly, but only at certain points; precisely as owing to the equal intervals between the atom-groups of the single row these amplified the pattern of the individual group not uniformly, but only over certain rings. However we can reach this result in another way, by considering three atom-groups forming a triangle, as formerly we considered two forming a pair.

Start then as heretofore from a solitary atom-group A, and add to it two others B and C (Fig. 5). They must not all three be collinear; as a rule it is best that B and C should be the two groups nearest A.[5] As before P stands for the source of the primary waves, Q for the point (on the wall of our imaginary bulb) where the scattered waves are to be measured. The question is: under what condition do the scattered waves from A and B and C all three reinforce one another best at Q? under what condition do the waves from all three groups arrive at Q in identical phase?

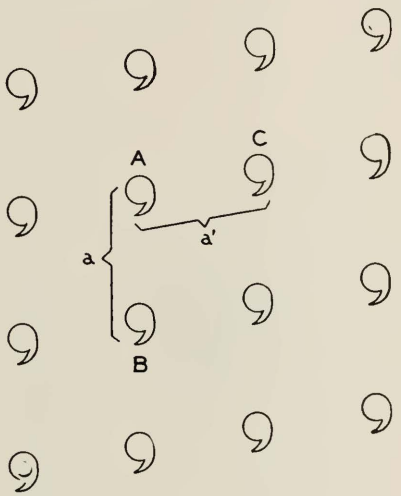


Fig. 5—Illustrating diffraction by a plane of atom-groups in regular array.

Evidently we have only to restate the condition that the waves from A and B arrive in identical phase, and supplement it with one exactly like it for the waves from A and C. We have only to repeat equation (3), and beside it write another like it in which a is replaced by a', the distance from A to C; ϕ by ϕ', the angle between the direction of the primary waves PA and that from A to C; and θ by θ', the angle between the direction of the scattered waves AQ and that from A to C. When the waves from all three atom-groups reinforce one another, both equations prevail:

$$a(\cos\theta - \cos\phi) = n\lambda, \tag{3}$$

$$a'(\cos\theta' - \cos\phi') = n'\lambda. \tag{4}$$

[5] The choice of groups to serve as B and C is purely a question of expediency. The same results are reached whichever two we choose, so long as they are not collinear with A; but the results when reached are in a form which depends upon the choice, and is most convenient when AB and AC are parallel to the crystallographic axes in the plane.

In these equations n and n' stand for integers but they need not stand simultaneously for the *same* integer. In all directions for which θ and θ' satisfy equations (3) and (4), there will be maximum amplification of the diffraction-pattern of any atom-group by its pair of neighbors.

Now equation (3), with various integer values 0, 1, 2, $\cdots$, substituted for n one after the other, described a system of rings on the wall of the bulb—parallel rings like latitude-circles, with the poles at the points where the bulb is intersected by the line drawn through its centre parallel to AB. Likewise equation (4), with various integer values for n', describes a system of rings oblique to the first, having *its* poles at the points where the diameter drawn parallel to AC

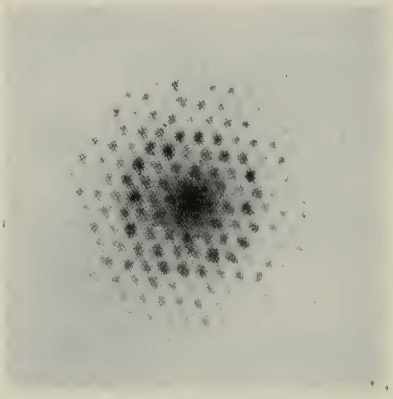


Fig. 6—Diffraction pattern of electron-waves attributed to the array of atom-groups in the superficial plane of a mica crystal. (S. Kikuchi; *Japanese Journal of Physics.*)

through the centre of the bulb reaches the wall. There are intersections between the rings of the two systems, and these intersections are the points—the discrete, the finitely numerous points—where the diffraction-pattern of the atom-group is most greatly amplified. So long as there are but three of the groups, these points are merely the centres of broad, hazy, bright (of course, in practice, utterly invisibly dim) blotches. But when to the three groups we add enormously many others to form the extensive two-dimensional network of which a section is depicted in Fig. 5, interference eats away the edges of these patches and enhances their centres, and in the limit nothing remains of the diffraction-pattern except brilliant dots at the intersections of the rings.

The next and last step follows immediately. The crystal is built

up of planes of atom-groups laid parallel to one another at equal intervals. Suppose the space above and below the plane of Fig. 5 to be thus stratified; select, from the stratum just above or just below that figured on the page, an atom-group close to A, preferably the closest. Call it D; let a'' stand for the distance from A to D, ϕ'' for the angle between the direction in which the primary waves advance and the direction AD, θ'' for the angle between the direction from A towards Q and that from A to D. When the scattered waves from all *four* groups $ABCD$ reinforce one another best at Q, all these three equations are valid simultaneously:

$$a(\cos\theta - \cos\phi) = n\lambda, \tag{3}$$

$$a'(\cos\theta' - \cos\phi') = n'\lambda, \tag{4}$$

$$a''(\cos\theta'' - \cos\phi'') = n''\lambda, \tag{5}$$

n'' standing for a third integer, which may or may not be the same as either of the other two. In all directions for which θ, θ', θ'' conform with equations (3), (4), and (5), there will be maximum amplification of the scattering-pattern of any atom-group by its triad of neighbors.

Now in thus adding a third equation to the previous one and two, we have made the conditions so severe that save in exceptional cases they are quite unfulfillable. Equations (3) and (4) confined the amplification-effects for which we seek to the points of intersection of a few rings belonging to two families, oblique to one another upon the wall of the bulb. Equation (5) supplies a third family of rings oblique to both, having for its poles the points where the diameter parallel to AD reaches the wall of the bulb. Agreement of phase between the waves scattered from A and B and C and D, optimum amplification, can occur only if and where a ring of the third family cuts rings of the first and the second just where these happen to cut one another. And when the set of four atom-groups $ABCD$ is repeated over and over again in an extensive crystal lattice, these points of optimum amplification are the only points where the amplification is great enough to enhance the diffraction-pattern into visibility. Visible luminous spots will appear on the coating of the wall of the bulb, only if three rings one from each family happen to intersect at the same point—*only by coincidence*, in the popular sense of the word.

To bring about such a coincidence, there are four practicable ways.

(I) If the incident beam is monochromatic, or comprises a narrowly limited range of wave-lengths, we can rotate the crystal—presenting it to the beam under varying aspects, and so in effect varying the

direction of incidence (the angles ϕ of the equations). Now and then the desired coincidence occurs.

This is Bragg's method. Instead of the fluorescent screen there is usually a photographic film bent to follow an arc, or an ionization-chamber swinging over an arc, of the surface of the imaginary bulb.

(II) If the incident beam is monochromatic or nearly so, and its wave-length can be varied continuously, we can keep the crystal motionless and the direction of incidence constant, and yet count on the coincidence turning up occasionally (Figs. 13, 15 and 16).

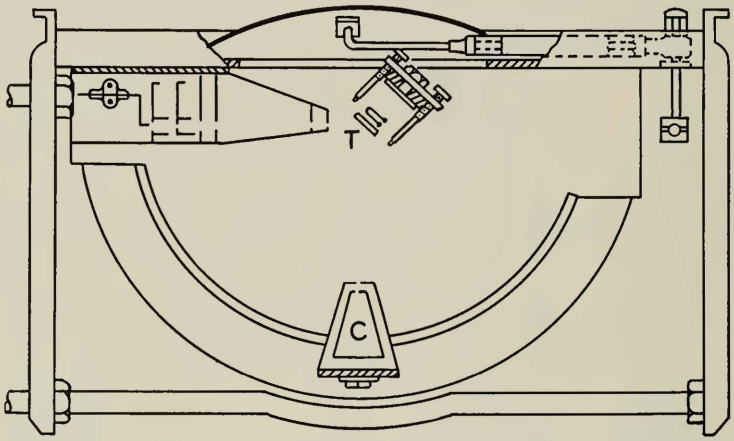


Fig. 7—Illustrating how a crystal (T) is mounted so that it may be oriented in various ways relatively to the oncoming beam of electron-waves (emerging from the electron-gun on the left) while a collector (C) is moved around to catch the diffraction beams. (Davisson and Germer.)

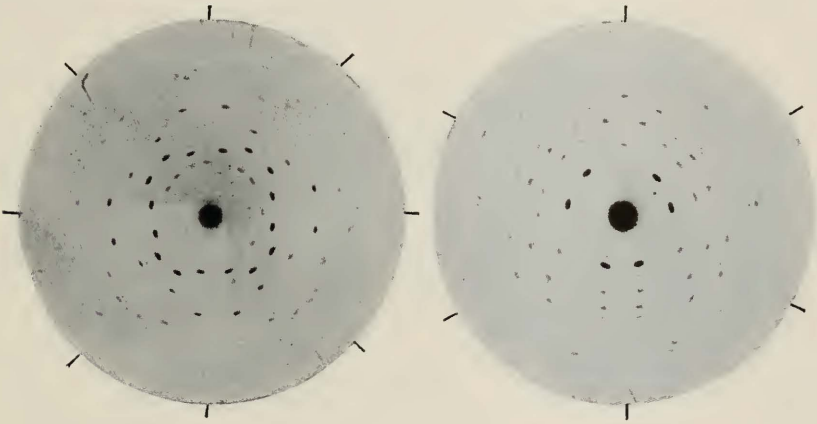
This is the method used with electron-waves by Davisson and Germer, the method whereby it was first shown that free negative electricity possesses some of the qualities of waves. To waves of this kind it is especially adapted, as their wave-lengths can easily be varied by varying the voltage-rise which endows the electrons with their speed. Instead of a fluorescent screen a Faraday chamber is used which swings in an arc over the surface of the imaginary bulb.

(III) If the incident beam is a mixture of wave-lengths covering a very wide range, we can keep the crystal motionless and the direction of incidence constant, and yet count on the coincidence turning up for some of the wave-lengths (Figs. 8, 9, 10, 11 and 12).

This is Laue's method, the first to be applied to the analysis of crystal structure, and the one whereby it was first proved that X-rays are waves—as people said at the time, which was 1912. Nowadays

we say that it was proved that X-rays possess some of the qualities of waves.

To make the spots appear a fluorescent coat or a photographic film is spread on a flat screen, instead of the spherical bulb. The locations on the screen can be deduced from those on the imaginary bulb by simple projection. Different spots are likely to be due to components of different wave-length in the primary beam, which will probably not be equally intense—a thing to be remembered when deducing the diffraction-pattern of the atom-group.



Figs. 8, 9—Diffraction-spots ("Laue patterns") obtained when a beam including waves of many wave-lengths is directed against a fixed single crystal. These are the historic patterns obtained with X-rays and a zincblende crystal by Friedrich and Knipping. The two correspond to different orientations of the crystal relative to the primary beam.

(IV) If the incident beam is monochromatic or nearly so we can present to it not a single stationary crystal but a confused mass of tiny crystals oriented in every way whatever. The desired coincidence will certainly occur for some among the crystals.

This is the "powder method" invented and applied to X-rays by Hull and by Debye and Scherrer, and applied to electron-waves by G. P. Thomson. The term implies that the chaotic mass of little crystals is obtained by pulverizing a large one; but small pieces or thin films of ordinary metals are likely to present quite as complete a chaos. The diffraction-pattern when formed on the wall of the bulb or on a flat screen set normal to the direction of the primary waves, consists not of separate spots but of continuous rings (Figs. 17–20).

Such in outline are the four great methods for the analysis of waves by crystals and of crystals by waves. Each has its own field; each

is beautifully adapted to certain problems as nature or art present them, not so useful or altogether useless for others. So for instance, the first is enormously the best for the spectroscopy of X-rays; the second is the outstanding method for long electron-waves (of the



Fig. 10—Laue pattern of a mica crystal. (R. M. Bozorth.)

order of an Angstrom) and thin films of matter; the third is very useful in the analysis of reasonably large crystals formed by intricate chemical compounds, and the fourth is invaluable for substances like the metals of which the crystals are often very small and tangled up together. But the fourth does not carry the investigator so far into the delicate details of crystal structure, as do the third and the first, when large crystals are available; the third and the first are impotent, when large crystals are not to be had; and the second would be exceedingly toilsome with X-rays. The complete crystal-structure laboratory now contains apparatus for the first method, the third, and the fourth. Perhaps it will not be long before the second also is demanded.

Before entering into details I wish to comment on four assumptions which have crept unsignalized into these deductions.

(i) *The assumption of the unlimited perfect crystal.* As I have already said sufficiently, a crystal composed of only a few atom-groups would produce broad hazy spots instead of sharp ones, as a grating with

only half-a-dozen rulings would cast a spectrum of indistinct wide fringes instead of fine sharp "lines." It takes a multitude of atom-groups or rulings to produce the efficient destructive interference which etches out the borders of these blotches and leaves the centres standing up in high relief,—which in technical language makes the resolving-power high. In actual crystals, are there atom-groups enough?

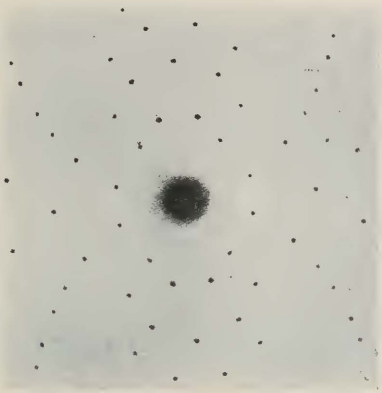


Fig. 11—Laue pattern of an iron crystal, showing that these crystals are built on a cubic lattice, though they are seldom so shaped as to reveal the fact. (G. L. Clark.)

In actual crystals, there are usually plenty. Infinite resolving-power or infinite sharpness of diffraction-pattern would of course imply infinitely many groups arrayed at perfectly regular intervals; but infinity and perfection are not workable ideas in physics. The practical question is, whether the actual defects of the diffraction-pattern from infinite sharpness arise because the crystal lattices are not prolonged enough, or from other causes. Usually they are due to other causes, which are numerous; for instance, appreciable breadth of the primary beam and appreciable size of the crystal. In the rare cases where the fuzziness of the diffraction-spots betokens that the crystal lattice is limited, the fact is often of scientific importance. The size of exceedingly small—submicroscopic—crystals can be determined in this way; the dimensions of colloid particles, and of the crystals in metals, are estimated thus.

(ii) *The assumption of stationary atoms.* This of course is faulty, for the atoms are in thermal agitation. Their oscillations make the spots of the diffraction-pattern somewhat hazy, and alter furthermore their relative intensities. Out of this circumstance the physicists

have drawn some profit; they have estimated the amount of the thermal agitation and determined how it alters as the temperature goes down. It decreases, of course; not however in such a way as to imply entire standstill at absolute zero, but quite the contrary.

(iii) *The neglect of absorption.* Since the energy of the scattered waves is drawn from that of the primary beam, this cannot retain its amplitude unaltered as it progresses through a crystal lattice; and calculations based on the assumption that all the atom-groups of a crystal scatter equally cannot be correct, for they do not all have

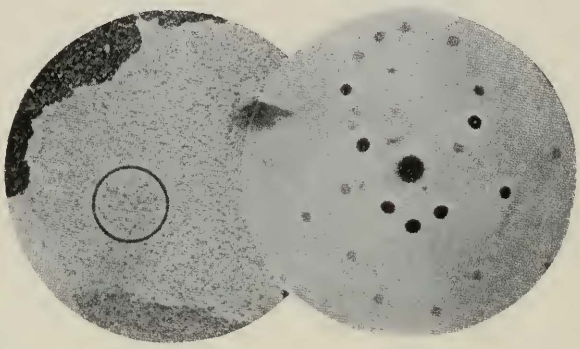


Fig. 12—Laue pattern of an aluminium crystal, introduced to show that if the primary beam does not happen to follow an axis of the crystal lattice there are still diffraction-spots though not so regularly arranged as in the previous cases. The crystal appeared under the microscope as an irregular patch on the metal surface (left-hand figure; the circle shows where the primary beam struck.) (Czochralski.)

incident waves of the same amplitude to scatter. This might well be serious, in a large crystal. It turns out however that large crystals are seldom if ever perfect; instead, they are likely to consist of smaller crystals tilted with respect to one another. The tilts are very small, but they are sufficient to suspend the consequences which should strictly follow from the assumption that absorption may be neglected.

(iv) *The neglect of refraction.* It has been tacitly assumed that there is neither bending of path nor change of wave-length when primary waves enter a crystal lattice or scattered waves emerge. Refraction however entails both of these phenomena; and refraction will in general occur. However with X-rays it is so slight (the refractive index is so nearly unity) that it need seldom be allowed for in crystal analysis, and must indeed be looked for with care and skill if it is to be detected. The like is true for short electron-waves. With the long electron-waves first studied, however, the refraction is considerable; and during the interpretation of the earliest data, there was serious confusion.

We will now work out the details of the diffraction-pattern of a *cubic* lattice. This kind of lattice is much the easiest to treat, and Nature has kindly bestowed it on many of the commonest crystals.

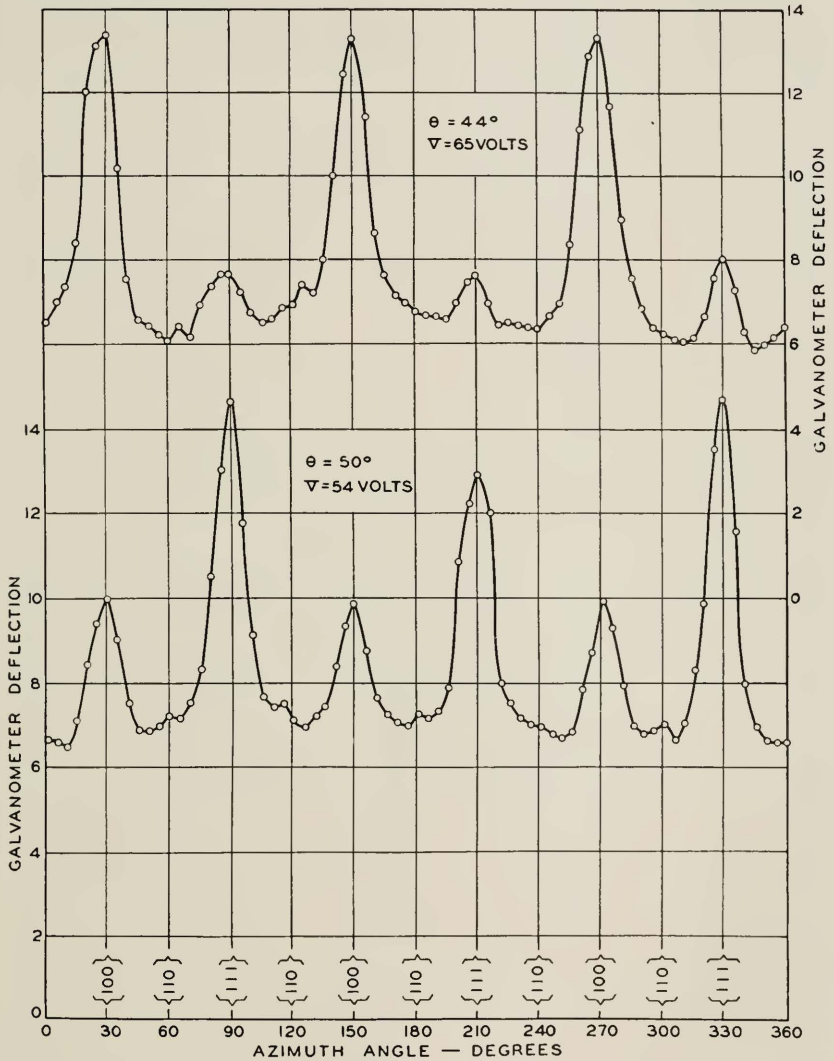


Fig. 13—Diffraction-peaks obtained with a fixed crystal and monochromatic electron waves (the two curves correspond to different wave-lengths chosen because they satisfy the condition for producing diffraction-beams) by moving a collector around so as to capture one beam after another. (Davisson and Germer.)

In the cubic lattice, the nearest neighbors of any atom-group lie at equal distances from it along three directions at right angles to one another. Moreover any atom-group in the entire (perfect) lattice

can be reached from the one first chosen by a sequence of three displacements, one along each of the three directions and each an integer multiple of the minimum distance, or "spacing," or "grating-constant," or "edge of the unit cube." In other words, any atom-group can be brought into coincidence with any other by a sequence of three such shifts. This way of putting it brings out the point that all the groups in the lattice are oriented alike.

Denote by a the magnitude of the spacing. Install a system of rectangular coordinates with its origin at some one atom-group, say A of our previous picture, and its axes along the three directions stated. Among the six neighbors of A we pick out three to serve as B, C, D of our previous picture; say the three groups shifted from A through the interval a in the *positive* senses of the three axes, so that the coordinates of the four shall be:

$$A(0, 0, 0); \qquad B(a, 0, 0); \qquad C(0, a, 0); \qquad D = (0, 0, a).$$

The directions of the diffraction-spots are to be deduced from the general equations (3), (4) and (5), with all the simplifications from which we benefit thanks to the lattice being cubic. The three spacings are now all equal; but the greatest advantage is, that the various cosines which figure in the equations are now direction-cosines, and we can avail ourselves of the theorems to which direction-cosines conform. There are two of these which we shall use: the theorems that the sum of the squares of the direction-cosines of any line is unity, and that the cosine of the angle between any two lines is the sum of the three products formed by multiplying together corresponding direction-cosines of the two. Denote by α_1, α_2, α_3 the direction-cosines of the primary, by β_1, β_2, β_3 those of the diffracted beam. The first three are the quantities $\cos\phi$, $\cos\phi'$, $\cos\phi''$ of equations (3), (4) and (5); the second three are $\cos\theta$, $\cos\theta'$, $\cos\theta''$. To bring the notation fully into harmony with usage I further write h_1, h_2, h_3 for the integers n, n', n''.

Then by translating equations (3), (4) and (5) into the new notation and adding two more supplied by the first of the foregoing theorems, we form a family of five equations:

$$a(\beta_1 - \alpha_1) = h_1\lambda, \tag{6a}$$

$$a(\beta_2 - \alpha_2) = h_2\lambda, \tag{6b}$$

$$a(\beta_3 - \alpha_3) = h_3\lambda, \tag{6c}$$

$$\alpha_1^2 + \alpha_2^2 + \alpha_3^2 = 1, \tag{6d}$$

$$\beta_1^2 + \beta_2^2 + \beta_3^2 = 1. \tag{6e}$$

They involve seven variables: the wave-length, the direction-cosines of the primary or oncoming beam, the direction-cosines of the scattered or outgoing beam.

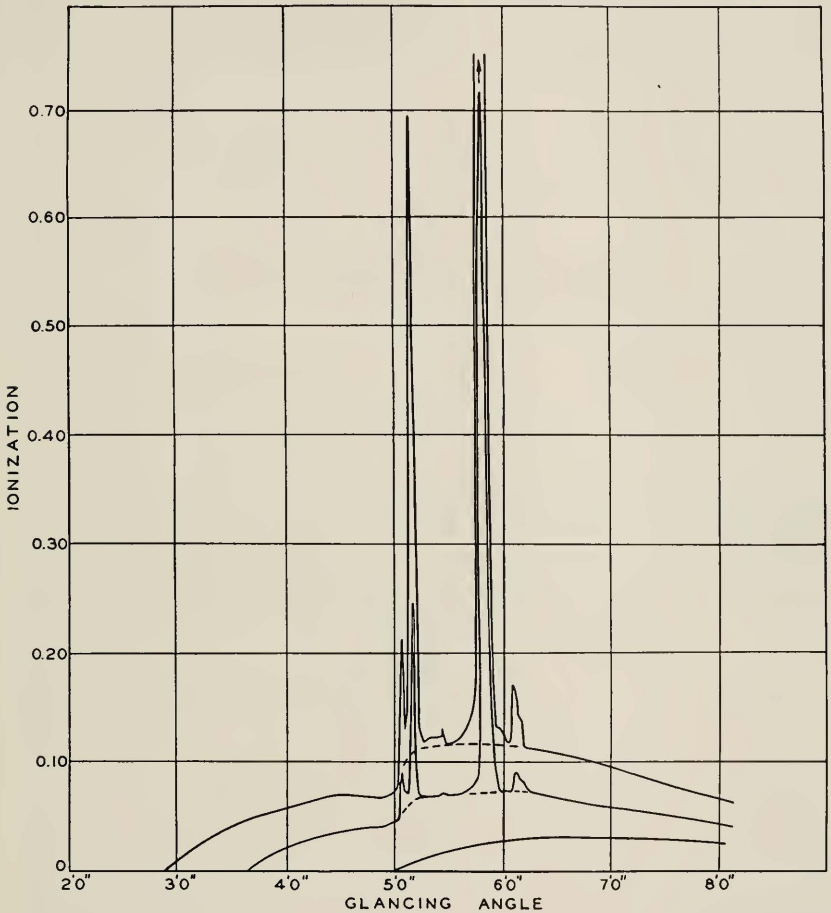


Fig. 14—Diffraction-peaks obtained with X-rays and a revolving crystal. The X-rays contained very intense waves of several different wave-lengths, and the collector was shifted continually so that it would capture a certain strong diffraction-beam produced by each of these in turn. (D. L. Webster.)

Now that we have this family of equations, the features of the four great methods of crystal analysis can be restated in few words. In the methods of Laue and of Davisson (II and III) the crystal and the primary beam are fixed in space, which amounts to prescribing values for α_1, α_2, α_3; then only four equations are left ($6d$ has fallen

out) and they contain four variables (λ, β_1, β_2, β_3) so that diffraction-spots can appear only for special sets of values of these four, to be obtained by solving the equations. In the original method of Laue, a wide range of values of λ is constantly provided, and the screen extends over a wide range of values of β_1, β_2, β_3; consequently the chance of observing spots is good. In the method of Davisson and

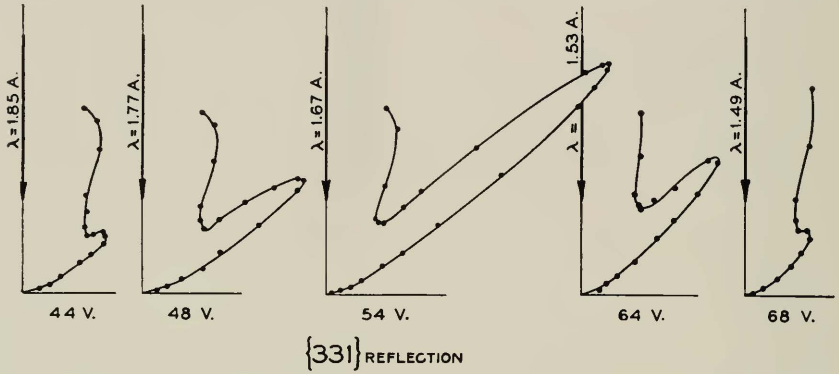


Fig. 15—Gradual appearance and disappearance of a diffraction beam as the mean wave-length of the primary waves passes through one of the values compatible with equations (6). (Davisson and Germer.)

Germer the different values of λ are realized in succession and a movable Faraday chamber searches out the peaks—a process much more long-drawn-out, but whenever one finds a peak one knows the wave-length to which it is due. With the other two methods the value of λ is prescribed, and consequently two at least of the direction-cosines α must be variable; therefore the crystal with its implanted coordinate-frame must be revolved, or else a multitude of crystals oriented every way must be placed in the path of the incident beam.

In the foregoing passage it seems as if I had taken for granted that the integers h_1, h_2, h_3 have fixed unchangeable values. As a matter of fact they may be any three integers at all. Strictly speaking, there is a different quintet of equations for every conceivable triad of integral values of the "indices" h. One might infer that in Laue's experiment the screen would be found completely covered with spots due to all the different triplets. However it turns out that only the spots for which all the integers are small stand out strongly enough to be seen. Meanings for these integers must now be found; but before finding them I will deduce two more equations out of the quintet.

Squaring and adding the left-hand members of equations (6*a*, 6*b*, 6*c*), doing the same with the right-hand members, equating the sums and

substituting from (6*d*, 6*e*), we obtain:

$$2 - 2(\alpha_1\beta_1 + \alpha_2\beta_2 + \alpha_3\beta_3) = \frac{\lambda^2}{a^2}(h_1^2 + h_2^2 + h_2^2). \qquad (7)$$

Now by the second of the theorems concerning direction-cosines, the quantity in parentheses on the left is none other than the cosine of the angle between the direction of advance of the primary beam, and the direction of the scattered waves which go to form the spot or

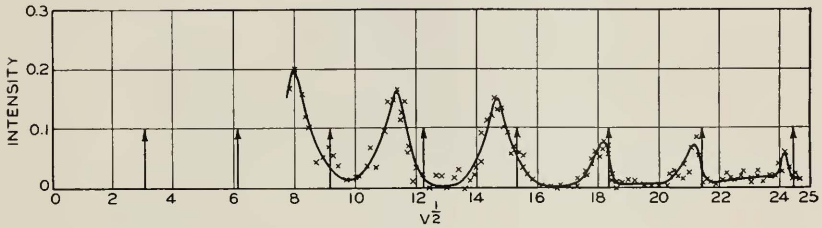


Fig. 16—Diffraction-peaks obtained with a fixed crystal and a fixed collector, *i.e.* with a constant value of the angle of deflection Φ, by varying the wave-length. (Davisson and Germer.)

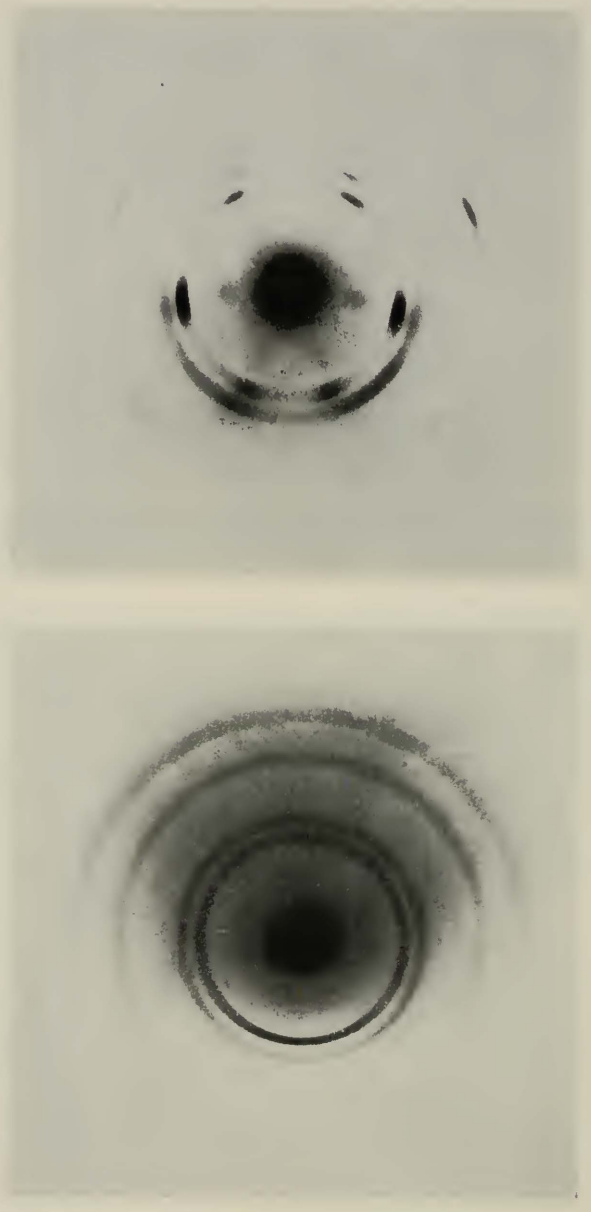
diffraction-maximum of the indices h_1, h_2, h_3. If we conceive the diffraction-beam as the path of a portion of the energy which came with the primary stream and was deflected out of it, then this is the angle of deflection. Call it Φ. We have:

$$1 - \cos\Phi = \frac{\lambda^2}{2a^2}(h_1^2 + h_2^2 + h_3^2), \qquad (8)$$

$$\sin\frac{1}{2}\Phi = \frac{\lambda}{2a}\sqrt{h_1^2 + h_2^2 + h_3^2}. \qquad (9)$$

As the reader will observe, there is no allusion here, explicit or implicit, to the orientation of the crystal. This is therefore the appropriate equation for the "powder method," in which crystals turned every way are presented all together to the primary stream, and no one knows the orientation of any particular one—indeed the individuals are often too small to be seen.

Equation (9) describes a cone, having for its origin the mass of assembled crystals, for its axis the direction of the primary beam, for its apical semi-angle the angle Φ. Such a cone intersects any sphere centred at the crystals (our imaginary bulb), or any plane at right angles to the primary wave-stream, in a *ring*. There are in principle as many of these rings as there are triads of integers (h_1, h_2, h_3) except that when two different triads have the same value of the



Figs. 17, 18—"Powder method" diffraction-rings obtained with X-rays and masses of small crystals of a nickel-iron alloy built on a cubic lattice. The rings are uniformly dark for one sample because the crystals were oriented quite at random, while for the other there were certain preferred orientations and the rings are "spotted." (R. M. Bozorth.)

sum-of-squares ($h_1^2 + h_2^2 + h_3^2$), their rings coincide. They appear as dark circles on a photographic film so placed as to coincide with such a sphere or such a plane, exposed and subsequently developed. Each of these circles consists of the diffraction-spots with the appropriate indices cast by the various crystals. If the crystals are few, one sees the individual spots (the ring looks ragged and spotty, like a star-cluster); if they are few and small the spots are hazy; if instead of being turned at random they favor certain orientations, the circles are not evenly dark all the way round. But these are matters for later study.

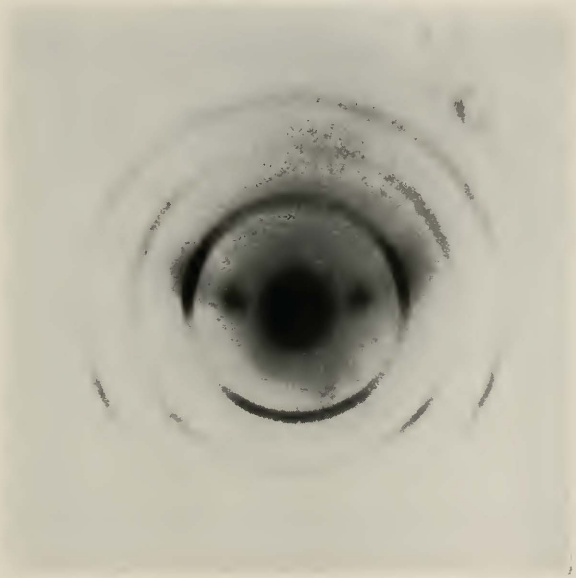


Fig. 19—"Powder method" diffraction-rings obtained with X-rays and a nickel-iron alloy. Like the alloys used in Figs. 17, 18, the crystals of this are built on a cubic lattice but with a differently-shaped atom-group, whence the changed appearance. (R. M. Bozorth.)

If we measure the radii of the first few rings and calculate from them the values of $(1 - \cos \Phi)$ for the corresponding cones (a simple matter of geometry) we should find that these stand to one another as $1 : 2 : 3 : 4 \cdots$—provided, that is, that the lattice is cubic and the incident waves are nearly monochromatic. This is verified by experience for X-rays and electron-waves. The first ring consists of spots having the indices (1, 0, 0) or (0, 1, 0) or (0, 0, 1); the indices for the second ring are (1, 1, 0) or (0, 1, 1) or (1, 0, 1), while those for the third are (1, 1, 1). The reader can easily guess the indices

which yield other small integer values for the sum (h_1, h_2, h_3)—but not in every case. There is no triad of integers of which the squares add up to seven, and there is none for which they add to fifteen. If the radii of the rings are plotted as a function of their ordinal number, there are breaks in the smooth curve beyond the sixth and the thirteenth, as if two were absent from the regular sequence. To express it more graphically than accurately, the seventh and the fifteenth rings are missing. Other rings also may be wanting, for the atoms in the atom-groups may be so disposed that in certain directions the individual groups scatter no waves whatever; and even if the lattice were so proportioned that waves in such a direction would be tremendously amplified, there would be nothing to amplify and no diffraction-spot. But this also is a subject for later study.

If we measure the radius of any ring and calculate sin $\frac{1}{2}\Phi$, and then change the wave-length and repeat the process, the values so obtained for the sine should stand to one another in the ratio of the wave-lengths. If the waves in question are electron-waves, then since their wave-length is inversely as the speed of the electrons, the radius of each ring should vary inversely as the speed of the electrons falling upon the crystals.[6] This was verified by G. P. Thomson. Knowing the values of the spacing a of the metal crystals which he had used—these values having been deduced in earlier days from the diffraction-circles produced by X-rays of known wave-length, scattered by the crystals of such metals—Thomson also determined by equation (8) the actual wave-lengths of the electron-waves. With X-rays the powder-method is seldom used to evaluate wave-lengths, Bragg's being the better when available; it serves chiefly for the study of lattices. But with X-rays and electrons both, the splendid array of the diffraction circles which spring forth when a beam of either kind is sent against a mass of tiny crystals is the most easily adducible, the simplest and perhaps the most vivid and striking evidence that there is something wave-like in the nature of either.

So much for the fundamental equation of the powder method! We will now derive, from equations $(6a \cdots 6e)$, the fundamental equation of the method invented by Bragg.

We have seen that the diffracted beam forming the spot with

[6] More precisely, the radii should vary inversely as the momentum of the electrons. The true formula for the wave-length as function of the electron-speed is probably

$$\lambda = (h/m_0 v)\sqrt{1 - v^2/c^2}$$

instead of $\lambda = h/m_0 v$ (here v stands for the speed and m_0 for the mass-at-zero-speed of the electrons); but as yet the speeds employed have not been great enough nor the measurements exact enough to distinguish between the formulae.

indices (h_1, h_2, h_3) is inclined to the primary beam at a certain angle Φ for which we have found the formulæ (8) and (9). We may conceive that this *deflection* is due to a *reflection* of part of the incident wave-motion from a mirror or mirrors traversing the crystal, so tilted that their plane bisects the angle Φ. The picture would be legitimate even if there were nothing physical corresponding to these "mirrors"; but we shall presently see that they are not imaginary. The normal to their plane makes supplementary angles θ_0 and $\theta = \pi - \theta_0$ with

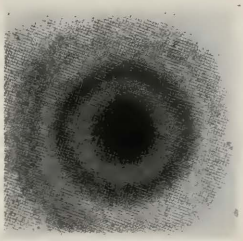


Fig. 20—"Powder method" diffraction-rings obtained with electron-waves and a thin film of gold. (G. P. Thomson, *Proc. Roy. Soc.*)

the primary and the diffracted beams, respectively; and $(\theta_0 - \theta)$ is the angle of deflection Φ.[7] Combining these statements, and choosing the positive sense of the normal so that it shall make an acute angle with the diffracted beam, we find:

$$\theta = \tfrac{1}{2}\pi - \tfrac{1}{2}\Phi; \qquad \theta_0 = \tfrac{1}{2}\pi + \tfrac{1}{2}\Phi. \tag{10}$$

We wish to deduce the direction-cosines of the normal—denote them for the moment by the symbols γ_1, γ_2, γ_3—from the conditions that it makes the angles θ_0 and θ with the rays having the direction-cosines α_1, α_2, α_3 and β_1, β_2, β_3 respectively. These conditions are thus expressed by the aid of equation (9):

$$\alpha_1\gamma_1+\alpha_2\gamma_2+a_3\gamma_3=\cos\theta_0=-\sin\frac{1}{2}\Phi=-\frac{\lambda}{2a}\sqrt{h_1^2+h_2^2+h_3^2}, \tag{11}$$

$$\beta_1\gamma_1+\beta_2\gamma_2+\beta_3\gamma_3=\cos\theta=+\sin\frac{1}{2}\Phi=+\frac{\lambda}{2a}\sqrt{h_1^2+h_2^2+h_3^2}, \tag{12}$$

and the reader can easily show by means of equations (6*a*, 6*b*, 6*c*) that they are satisfied when:

$$\gamma_1=\frac{h_1}{\sqrt{h_1^2+h_2^2+h_3^2}}\,;\ \gamma_2=\frac{h_2}{\sqrt{h_1^2+h_2^2+h_3^2}}\,;\ \gamma_3=\frac{h_3}{\sqrt{h_1^2+h_2^2+h_3^2}}\,. \tag{13}$$

[7] It is the custom to say that the normal to a mirror makes *equal* angles with the incident and the reflected beams, but this manner of statement implies that the positive senses along the directions of the beams are defined oppositely—one *with*, the other *against* the sense in which the waves are advancing. The convention adopted here is the more logical, and leads to more symmetrical equations.

If therefore a diffraction-spot (h_1, h_2, h_3) can with any reason be attributed to a reflection of part of the incident energy from mirrors traversing the crystal, then these mirrors must be so tilted that their normal is pointed in the direction defined by equation (13).

This equation being attained, we are prepared to discern the physical meanings of the integers h.

One of their meanings is evident. They state the "order" of the diffraction-spot, in the sense in which that word is used in describing the spectra cast by optical gratings. An ordinary ruled grating supplied with light of a single wave-length forms, it may be, six or seven or even more different "lines" on the focal plane of the lens installed in front of it. These correspond to diffraction-spots on the surface of our imaginary bulb. Take any one of these lines, say the nth. The paths of the light from the source *via* consecutive rulings of the grating to this nth line differ by precisely n wave-lengths; the contributions of any two adjacent rulings to the wave-motion at this line differ in phase by n complete cycles. It is named the line, or the image, or the diffraction-maximum *of the nth order.* There may be lines of positive, of negative and of zero order; the line of zero order is commonly called the "direct image" of the source.

Now in the same sense a diffraction-spot with the indices h_1, h_2, h_3, cast by a crystal, is of three orders simultaneously; these indices are its orders with respect to the three principal directions of the crystal lattice. Referring to our quartet of atom-groups $ABCD$: the paths of the waves from the source *via* A and B to the diffraction-spot differ by h_1 wave-lengths, those *via* A and C differ by h_2 and those *via* A and D by h_3 wave-lengths. The (000) spot is in the prolongation of the incident stream, and is called the "direct image."

This is one important meaning of the indices h. Happily there is another which is much more picturesque—happily indeed, for otherwise it is likely that the art of crystal analysis would have developed more slowly than it has, while the art of X-ray spectroscopy might not even have begun for years. It does not always happen, perhaps indeed it rarely happens, that the earliest formulation of a theory is the one best adapted to make it widely understood and useful. Someone other than the founder meditates upon the idea, and tries to re-express it to himself, and hits upon a novel way of stating it—a new aspect to it, possibly, or perhaps nothing more than a new distribution of the emphasis, laying the greatest stress upon some feature which to his forerunner seemed minor. Then suddenly the theory appeals to the world of physicists in general. Experimenters see what can be done with it, how it can easily be tested and how it

can be applied after the tests are passed; and progress for a time is furiously fast. Something much like this befell the theory of the diffraction of waves by crystals. The form in which I have thus far developed it (except while describing the powder method) is the one in which it was clothed by the brilliant inspiration of Laue. However it was W. L. Bragg who made the theory well known and widely used all the world over, by singling out and featuring the fact that *each diffraction-beam is due to its own special set of atom-strata in the crystal, which reflect it as light is reflected by a pile of parallel mirrors.*

The integers h are then no other than the indices, by which the crystallographers denote these strata. For, to the student of crystallography, the strata are very real— as much so as the rows of atom-groups, by referring to which I developed the theory of the diffraction-spots in Laue's way. We started with the individual group and went on to the row, and then constructed the plane by laying rows down side by side; but we might have started with planes, and defined the rows of atom-groups as the lines or edges where two planes intersect, and located individual atom-groups at corners where three planes meet. This is the historical way; for the planes are the prominent feature of any well-developed crystal. The smooth flat facets which are the boundaries of every well-formed crystal are parallel to important strata, they are themselves examples of important strata. In studying a crystal otherwise than by diffraction, the first step is to measure the directions (relative, of course) of all the available facets. Before the invention of analysis by diffraction, this was often the last step also; but if the crystals available have grown up really well, it is a very long step. Having taken it, the crystallographer proceeds to visualize the crystal as a region of space which is intersected and partitioned by flocks of planes, long sequences of evenly-spaced planes parallel to the facets; and he locates the atom-groups at their intersections. How then shall we harmonize these inferences of his with the implications of the diffraction pattern?

A good way to unify the two procedures is to explain the notation by which the crystal planes are named. In doing this, I shall often speak of planes "containing atom-groups," meaning in the strict sense planes containing lattice-points around which atom-groups are placed.

Return to the cubic lattice and to the basic set of four atom-groups $ABCD$, so chosen that the lines AB, AC, AD are three edges of the fundamental or "unit" cube. Complete the cube by adding four more atom-groups $EFGH$. We will pick out the planes which contain three or four of these eight groups. They will be the most populous with atom-groups of all the planes traversing the cube or

"cell"; hence the most populous of all the planes traversing the crystal. I have spoken above of the "important" strata of the crystal, meaning those which are likely to be parallel to facets. The word "important" is vague, and it is better that it should remain somewhat vague; but, in a rough way, the more populous a stratum is the more likely it is to be "important" in that sense, and also in the sense that it produces strong diffraction-spots. We will therefore consider chiefly the planes, which cut through the unit cube in such a way as to traverse three or four of the atom-groups. Put the origin of coordinates at A, the x, y, z axes through C, B, D respectively.

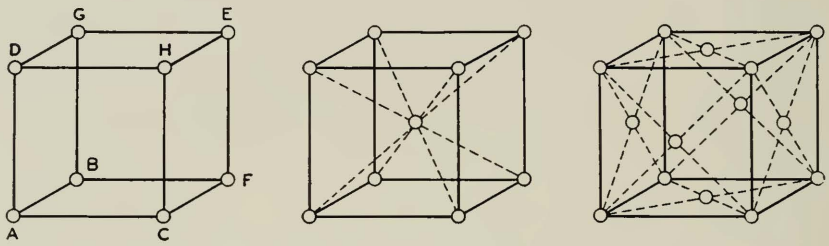


Fig. 21—Illustrating the cubic lattice.

Commence with the plane containing BCD, normal to the direction AE which is one of the principal diagonals of the cube. We wish a symbol for this plane that shall describe its orientation, which is its important feature,—a symbol that shall denote not only the plane BCD, but equally all those which are parallel to it, such as FGH and the parallel planes through A and E and other unit cells. We might use the three direction-cosines of the normal to this family of planes; or we might use the three intercepts of BCD on the three coordinate-axes, multiplied by an arbitrary factor to convert them into convenient integers and show that we are not more concerned with BCD than with any other member of the family; or we might use the reciprocals of these three intercepts, also multiplied by some convenient arbitrary factors. The last choice is the standard one. The three intercepts are a, a, a; their reciprocals are $1/a$, $1/a$, $1/a$; we multiply these by a and obtain (1, 1, 1) as the symbol for not only the plane or stratum BCD, but all the strata parallel to it, including for example any facets that may be formed upon such strata. The reader will easily identify three other families of planes for which the symbols are $(-1, 1, 1)$; $(1, -1, 1)$; and $(1, 1, -1)$. Some substances with cubic lattices form crystals in the shape of octahedra; the surfaces of these are strata with such symbols. Usually the commas are left out of the symbols, and the minus signs printed over the digits.

Now try the plane containing the atom-groups *BCGII*. Its intercepts on the x and y axes are both equal to a, but it is parallel to the z-axis, a fact which is described by giving its z-intercept as infinity. The reciprocals of its intercepts then are $1/a$, $1/a$, 0; we multiply by a and obtain (1 1 0) as the symbol for all the strata parallel to the one containing the atom-groups *BCGII*. The reader can easily identify members of the (0 1 0) and (0 0 1) families, and ascertain how many new families of planes he can get by reversing the signs of some or all of the indices. There are six altogether; one sometimes sees facets with these indices, beveling off the edges of natural cube-shaped crystals.

Next consider the plane containing *CFHE*. It is parallel to the yz-plane and distant from it by a, so that its intercepts are a, ∞, ∞, and its symbol is (1 0 0). One sees immediately that (1 0 0) and (0 1 0) and (0 0 1) are the symbols for the three families of planes which comprise these which we have taken for our coordinate-planes. When a substance with a cubic lattice forms crystals which are cubes, their facets belong to these families.

What could a symbol such as (2 1 0) imply? It would stand for a family of planes, one of which would have for its intercepts the values $a/2$, $a/1$, $a/0$ or $\frac{1}{2}a$, a, ∞. This plane would be parallel to the z-axis and would traverse the atom-groups B and G, and would slant across the cell in such a way as to pass through the wall *ACDH* midway between its vertical edges. Continuing it and the lattice in imagination, one sees that at the far angle of the *next* cell it would traverse another pair of atom-groups, and at the far angle of every second cell thereafter it would do the like. It is therefore a fairly "important" stratum, though not so populous as those of which the indices are zeros and ones exclusively. It might form facets, and would be likely to give a noticeable diffraction-spot. But in general as the indices mount up, the importance of the plane declines.

It is evident that if any set of indices is multiplied by any constant, the new set thus obtained corresponds to the same family of planes. To choose one set definitely for each family, we may agree to adopt the triad of integers having no common divisor. Thus all three of the symbols (963), (642) and (321) refer to the same family of planes; we always choose the last one.

Given these the so-called "Millerian" indices of a family of planes, what are the direction-cosines of their common normal?

Denote the three indices by H_1, H_2, H_3. The question is: what are the direction-cosines $\gamma_1\gamma_2\gamma_3$ of the normal to a plane, of which the intercepts on the coordinate axes are a/H_1, a/H_2, a/H_3? The an-

swer is given by a standard formula: the three direction-cosines are:

$$\gamma_1 = \frac{H_1}{\sqrt{H_1^2 + H_2^2 + H_3^2}};$$
$$\gamma_2 = \frac{H_2}{\sqrt{H_1^2 + H_2^2 + H_3^2}}; \qquad (14)$$
$$\gamma_3 = \frac{H_3}{\sqrt{H_1^2 + H_2^2 + H_3^2}};$$

The common factor a has vanished, as it should.

Compare these expressions with those in equation (13). One sees instantly that the strata (H_1, H_2, H_3) are so oriented that these strata could serve as mirrors to reflect the primary beam towards the diffraction-spot of which the indices are $h_1 = H_1$, $h_2 = H_2$, $h_3 = H_3$; or towards the spots of which the indices are (nH_1, nH_2, nH_3), where n stands for any integer. Or: the diffraction-spot (h_1, h_2, h_3) may be conceived as due to a reflection of part of the wave-motion in the incident stream, by the atom-layers of which the symbol is n_1/C, n_2/C, n_3/C; C standing for the greatest common divisor of h_1 h_2 h_3.

This is part of the principle which Bragg deduced from Laue's theory, but not the whole of it.

I have shown in an earlier article of this series [8] (and the reader can easily work out) that when a beam of plane waves falls successively on two plane parallel surfaces which reflect a part and transmit a part of it, the two reflected beams are in phase with one another and the resultant is maximum, when the following relation prevails between the wave-length λ of the light, the distance d between the mirrors, and the angle θ of incidence and of reflection:

$$n\lambda = 2d \cos \theta, \qquad n = 0, 1, 2, 3 \cdots. \qquad (15)$$

When instead of a pair there is an endless or a very long sequence of mirrors spaced at equal intervals, the result is much the same as when a pair of atoms is supplemented by a very long row. The angles defined by equation (15) are now not merely the angles of maximum reflection; they are the *only* angles where reflection is at all appreciable. If a pile of parallel semi-transparent mirrors is to reflect to any notable extent, their thicknesses and the wave-length and the angle of incidence of the light must be very carefully adjusted according to equation (15) with some integer value for n.

[8] Number 15 (October, 1928), p. 24. The formula there given contains an index of refraction which I here equate to unity, and an additive constant which vanishes if the phase-change at reflection is the same at each of the reflecting surfaces, which is here the case.

Now the principle emphasized by Bragg is this, in full: *any diffraction-spot results from reflection by the strata having the same indices as the spot, at the angle of selective reflection defined by equation (15).*

The proof of this statement depends on the following formula [9] for the distance between adjacent strata of the family having the Millerian indices $H_1H_2H_3$:

$$d = a/\sqrt{H_1^2 + H_2^2 + H_3^2}. \tag{16}$$

Substituting into equation (12) the value of a given by this formula, we find that:

$$\cos\theta = \frac{\lambda}{2d}\frac{\sqrt{h_1^2 + h_2^2 + h_3^2}}{\sqrt{H_1^2 + H_2^2 + H_3^2}} \tag{17}$$

and since the quantities H are integers without a common divisor, while the quantities h_1 are integers for which $h_1/H_1 = h_2/H_2 = h_3/H_3$, the ratio of the two radicals must be an integer.

The mirrors which I introduced at a previous page as a way of accounting for the spots do actually exist. They are the strata into which the groups of atoms fall. Each one by itself, however, reflects so little that in effect there is no reflection to speak of, unless and until an entire procession of parallel strata is brought into play. Earlier we located the diffraction-beams as the directions in which the scattered waves from four adjacent atom-groups enhance one another most by constructive interference. Multiplication of atom-groups beyond the first four merely made the beams sharper and more intense. Alternatively we may now locate these beams as the directions in which the reflected waves from two adjacent parallel strata enhance each other most. Multiplication of strata beyond the first two intensifies and sharpens them.

This theorem of Bragg's thus gives a remarkably helpful picture of "the way of a crystal with a beam of waves"; a picture most valuable, when one has a single large crystal of which the surfaces are natural facets. Suppose for instance one has a cubic crystal of rocksalt, one of

[9] Let O and P stand for two planes of the family $(H_1H_2H_3)$, one being drawn through the origin, the other through any other lattice-point. The components of the vector r from the origin to this other lattice point must be integer multiples l_1a, l_2a, l_3a of the spacing a. The projection of this vector on the direction of the normal drawn from the origin to the plane P is equal to $(\gamma_1 l_1 + \gamma_2 l_2 + \gamma_3 l_3)a$; the values of $\gamma_1\gamma_2\gamma_3$ are to be taken from (14). This projection is equal to the distance D between the planes O and P, for which we therefore have:

$$H_1l_1 + H_2l_2 + H_3l_3 = \frac{D}{a}\sqrt{H_1^2 + H_2^2 + H_3^2}.$$

The least value (except for zero) which the quantity on the left can and does assume is unity; whence equation (16).

its faces being parallel to the (100) strata. If one has a monochromatic beam of X-rays incident on that face and revolves the crystal so that θ varies steadily from zero to 90°, then for the several angles given by equation (15) with various values of n diffraction-beams spring forth. One may set up a photographic plate beside the crystal, and find the imprints of all the beams upon it after the rotation is completed; or alternatively one may revolve an ionization-chamber at double the angular speed of the crystal, so that whenever a beam shoots out the chamber is in the right place to capture it, and then the curve of ionization-current versus angle θ shows a peak for every value of θ corresponding to an integer value of n. If the incident beam comprises many wave-lengths, one finds their spectrum spread out in the ionization-current curve; three examples are shown in Fig. 14, where each of the sharp tall peaks is due to the first-order diffraction-beam of a monochromatic wave very intensely represented in the primary wave-mixture. Or one may hold the crystal and the collector still and vary the wave-length, obtaining a peak wherever λ is such that for some integer value of n the equation (15) is satisfied; the curve of Fig. 16 was obtained in this way, using waves of negative electricity.

The process of measuring wave-lengths of X-rays is usually conducted by this method, using a crystal such as rocksalt for which the density is very accurately known. For if we know the density of the crystal we know how many atoms it contains in a given volume; and if we know in addition how many atoms constitute the atom-group which is repeated over and over again to form the crystal, we can compute by simple division what is the volume of the unit cell, and what therefore is the spacing from one atom-group to the next—the edge of the elementary cube. If we then set up the crystal so as to get reflections from the 100 face we know that the edge of the cube is the quantity d which figures in the equation (15); and measuring then the values of θ corresponding to several diffraction beams we can identify the corresponding values of n, and so evaluate λ. Diffraction of X-rays by ruled gratings can now be called upon in confirmation—or in correction, as certain recent data indicate.

The determination of the number of atoms in the atom-group is the delicate point of this computation; and perhaps it is to be accounted a piece of luck that with the first crystals used in the spectroscopy of X-rays the guess was easy and was rightly made. The routine of determining it is but a part of the general process of learning from the diffraction as much as possible about the atom-group; and this deserves an article to itself, or many. Two examples of this process however are interesting, useful and very simple.

I remarked near the beginning of the article that while it would be the simplest possible thing to suppose that the particles arranged in a cubic lattice have full spherical symmetry, yet this supposition is as a rule too simple for the facts. In particular it is too restricted to explain the diffraction-pattern of any one of the numerous elements which crystallize on cubic lattices. One might then be forced to assume that the atoms themselves do not have spherical symmetry. But luckily this is unnecessary; for it happens that if with each lattice-point of the cubic lattice we associate a properly-spaced and properly-oriented *group* of spherical atoms—in some elements a pair, in other elements a group of four—the difficulties vanish. The diffraction-patterns are explained, and there is no outstanding conflict with the data assembled by the crystallographers; for both of these arrangements, like that in which each lattice-point is occupied by a single spherical atom, possess full cubic or isometric symmetry in the crystallographic sense of those words.

In the first of these permitted arrangements, the two atoms associated with each lattice-point are so placed and so spaced, that if we label them, say, A and B, the atoms A by themselves form one single cubic array, and the atoms B by themselves form another simple cubic array with an atom B in the very centre of each cube composed by atoms A—and vice versa. This is the "body-centred cubic" arrangement. It is depicted in the middle drawing of Fig. 21. The atom at the centre of the cube may be associated with any one of the eight corner atoms to form a pair: this pair is then repeated over and over again on the cubic lattice to form the crystal. The alkali metals, iron, and several other elements are addicted to this arrangement.

In the second of the arrangements, the four atoms forming a group are so placed and so spaced that if we call them A, B, C, and D, the atoms of each letter form a cubic array; and these four cubic arrays are interlocked in such a fashion, that the cubes of any one of these arrays have in the centres of all their faces atoms belonging to the others. Thus in the righthand sketch of Fig. 21 the atom at any corner may be associated with the atoms in the centres of the three faces which meet at that corner, and these four form the atom-group which is repeated over and over again on the cubic lattice to build the crystal. Many of the metallic elements have adopted this "face-centred cubic" arrangement, the noble metals for example, and argon also.

How does one recognize from the diffraction pattern which of these arrangements exists in a cubic lattice? At this point I will not give an exact answer: but the principle is simple. Even as on an earlier page it was shown that for certain directions of diffraction adjacent

atom-groups *reinforce* the scattering from one another, so it may be shown that for certain directions the different atoms of a single atom-group *destroy* the scattering from one another. One has only to write down the condition that the distance from source P to fieldpoint Q *via* one atom of the group differs by an odd-integer multiple of $\frac{1}{2}\lambda$ from the distance *via* the other; or if there are four atoms in the group, that the waves scattered to Q from the four are so balanced in phase that they annul one another.

Now if it should turn out that one or more of the diffraction-spots expected from the cubic lattice fall exactly where the effects of the atoms of each individual group cancel each other out, then those spots will be lacking. For the spots are due to amplification of the diffraction-pattern of the individual atom-group; but if at the location of a predicted spot this pattern sinks to a vanishing intensity, there is nothing to amplify. Well! owing to the neat and accurate way in which the spacings between atoms of a group are related to the spacings between the atom-groups, this sort of coincidence occurs for a respectable fraction of the diffraction-spots—or, in the powder method, it occurs for several of the diffraction-rings. From the missing spots or rings therefore one identifies which style of atom-group prevails in the cubic crystal. And there is much more yet to be learned; but that will be material for another article.

Abstracts of Technical Articles From Bell System Sources

Scattering of Quanta with Diminution of Frequency.[1] KARL K. DARROW. In this article the author points out that certain phenomena of X-rays recently reported were illustrations of the general process of scattering of light with change in frequency, which had just begun to attract attention owing to important observations made by Raman and others with visible and ultra-violet light. The content was amplified and restated in Dr. Darrow's article entitled "Contemporary Advances in Physics, XVII—The Scattering of Light with Change of Frequency," which appeared in the January, 1929, issue of the *Bell System Technical Journal.*

Dissociation of Molecules as Disclosed by Band-Spectra.[2] KARL K. DARROW. This lecture was a contribution to a Symposium on Atomic Structure of the American Chemical Society. It is an elementary account of the way in which the band-spectra of molecular gases are interpreted so as to disclose the laws and details of the dissociation of their molecules into atoms, a process of great scientific and some practical importance.

Using Inspection Data to Control Quality.[3] H. F. DODGE. This paper outlines a method of using inspection data to improve the technique of controlling at economic levels the quality of product in the various stages of manufacture. Essentially, the method rests on the application of statistical methods of analysis, employing the viewpoint that every batch of manufactured product constitutes a sample from a much larger universe and as such is subject to random of chance variations in quality. The variations in quality as observed in inspection data may thus be the result of either chance causes or of fundamental production causes whose presence is undesirable.

Speech and Hearing.[4] HARVEY FLETCHER. This book is concerned mainly with the results of Bell System research work on speech and hearing. These results, however, can be understood and appreciated better when their relationship to similar work is shown. Conse-

[1] *Science*, Vol. 68, November 16, 1928, pp. 488–490.
[2] *Chemical Reviews*, Vol. V, December, 1928, pp. 451–466.
[3] *Manufacturing Industries*, Volume XVI, November, 1928, pp. 517–519, and December, 1928, pp. 613–615.
[4] D. Van Nostrand Co., Inc., New York, 1929.

quently, copious references to the experimental results of other workers have been included. The material is grouped under four headings: (1) Speech, (2) Music and Noise, (3) Hearing, and (4) Perception of Speech and Hearing.

The first part is concerned with the mechanism of speaking, the classification of the fundamental English speech sounds, and with the wave forms of such sounds. It includes a description of various types of apparatus which can be used for making permanent records of speech waves and gives a large number of accurate wave pictures of the speech sounds together with the power contained in such waves.

In the second part similar data are given for musical sounds and noise.

The third part begins with a discussion of a theory of hearing which is proposed to explain the experimental facts of audition. This is followed by a discussion of the known facts of audition such as the limits of audition, the minimum perceptible differences in sound, masking effects, binaural effects, methods of testing the acuity of hearing, etc. Along with this discussion is given a description of the apparatus and experimental methods used for determining these facts.

The fourth part is concerned with those phases of the subject that involve personal judgment, that is, the psychological element. A scale for measuring the loudness and the pitch of complex sounds is defined. Experimental data are given which show how these two subjective quantities depend upon external physical quantities. Methods of measuring the recognition of speech sounds are described and experimental results using such methods are given to show the effect of various types of distortion upon the ability of persons to recognize such distorted sounds.

Elementary Differential Equations.[5] THORNTON C. FRY. In this book Dr. Fry has covered the field of differential equations as usually offered in elementary courses in universities and technical schools. The mathematical ideas are first presented as mathematical entities in themselves and not as the symbolic formulation of physical concepts. With this accomplished, these ideas are broadened and illustrated by live scientific examples and problems, which are drawn from a wide variety of fields. The inclusion of such technical material does not presuppose a wider knowledge of technical subjects than the reader can reasonably be expected to possess, nor does it interfere with the clarity of the mathematical presentation.

[5] D. Van Nostrand Company, Inc., New York, 1929.

Optical Conditions for Direct Scanning in Television.[6] FRANK GRAY and HERBERT E. IVES. This paper discusses the conditions for securing the maximum amount of light in a photoelectric cell placed behind a television scanning disc when an image is formed on the disc by a lens. Results obtained with a large scanning disc and a lens forming images of sunlit objects are described.

A Camera for Making Parallax Panoramagrams.[7] HERBERT E. IVES. This paper describes a camera for making transparencies which when viewed through an opaque line grating show stereoscopic relief through a wide range of distances and angles. The essential feature of the camera is a mechanical coupling by means of which the camera lens, the sensitive plate and grating, and the object photographed, are kept in line as the camera moves from one side to the other of the normal from the camera track to the object.

European Factory Methods and Equipment in the Manufacture of Metals.[8] DAVID LEVINGER. In this paper the author outlines his observations of the metal-working industries of Europe, based on a three months' tour of eight countries during the summer of 1927, in which seventy-five industrial establishments were visited in England, France, Germany, Belgium, Holland, Italy, Austria and Switzerland.

Electrical Conduction in Textiles. Part I—The Dependence of the Resistivity of Cotton, Silk and Wool on Relative Humidity and Moisture Content.[9] E. J. MURPHY and A. C. WALKER. The data reported show that the resistivity of cotton is about 10^{12} times greater at 1 per cent humidity than at 99 per cent, that it is an exponential function of relative humidity in the range 20–80 per cent and a power function of moisture content over the whole range investigated. By means of the equations expressing these relationships the resistance of a cotton sample can be calculated for any moisture content (or the relative humidities corresponding to it) provided a measurement has been made at a single moisture content. The curves for the logarithm of resistance vs. relative humidity (or moisture content) for samples of cotton containing different amounts of electrolytic material are parallel, low electrolyte content corresponding to high resistance. Similar but less extensive measurements were made on silk and wool.

[6] *Journal of the Optical Society of America and Review of Scientific Instruments*, Vol. 17, December, 1928, pp. 428–434.
[7] *Journal of the Optical Society of America and Review of Scientific Instruments*, Vol. 17, December, 1928, pp. 435–439.
[8] *Mining and Metallurgy*, Vol. 9, November, 1928, pp. 483–486.
[9] *Journal of Physical Chemistry*, Vol. 32, December, 1928, pp. 1761–1786.

The results indicate that the conductivity of a textile is practically completely determined by three factors, the amount of absorbed water, its specific conductance (as determined by the amount of electrolytic material present in the textile) and its distribution.

The Effect of Gases on the Resistance of Granular Carbon Contacts.[10] P. S. OLMSTEAD. This paper describes a method whereby reproducible measurements of the resistance of granular carbon contacts can be made. The experimental arrangements were such that the resistance could be measured as a function of gas pressure, applied voltage, or time.

Note on the Determination of the Ionization in the Upper Atmosphere.[11] J. C. SCHELLENG. This paper describes a method of estimating the distribution of ionization in the upper atmosphere, based upon measurements of the effective height determined by interference or echo experiments. These two types of experiment are shown to give identical results.

Lead-Tin-Cadmium as a Substitute for Lead-Tin Wiping Solder.[12] EARLE E. SCHUMACHER and EDWARD J. BASCH. In this paper data are presented which show that certain lead-tin-cadmium alloys may be advantageously substituted as solders for lead-tin alloys. Data are given showing the physical and chemical properties of these alloys.

New Specifications for Raw Materials.[13] J. R. TOWNSEND. In this article the author points out that the annual demand for new telephone apparatus by the Bell System requires a steady flow of materials of the proper quality and uniformity into its manufacturing plants. To meet this demand, a new set of engineering specifications has been inaugurated to control these raw materials. A notable example of this specification work is the preparation of Rockwell hardness and tensile strength requirements for sheet brass, nickel silver and phosphor bronze. The Western Electric Company, the Northern Electric Company and one of the suppliers, the American Brass Company, cooperated in this work. Rolling series were prepared covering all grades, thicknesses and tempers. The requirements were based on the data furnished by producer and consumer, and on experience over a long period with commercial material.

[10] *Journal of Physical Chemistry*, Vol. 33, January, 1929, pp. 69–80.
[11] *Proceedings of the I. R. E.*, Vol. 16, November, 1928, pp. 1471–1476.
[12] *Industrial and Engineering Chemistry*, Vol. 21, January, 1929, pp. 16–19.
[13] *Instruments*, Vol. 1, December, 1928, pp. 519–521.

Contributors to this Issue

Austin Bailey, A.B., University of Kansas, 1915; Ph.D., Cornell University, 1920; Instructor in Physics, Cornell University, 1915–18; Signal Corps, U.S.A., 1918–19; Fellow in Physics, Cornell University, 1919–20; Corning Glass Works, 1920–21; Assistant Professor of Physics, University of Kansas, 1921–22; Department of Development and Research, American Telephone and Telegraph Company, 1922–. Dr. Bailey's work while with the American Telephone and Telegraph Company has been largely along the line of methods for making radio transmission measurements and of long wave radio problems.

Karl K. Darrow, B.S., 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, 1917–25; Bell Telephone Laboratories, 1925–. Dr. Darrow has been engaged largely in writing studies and analyses of various fields of physics and the allied sciences. Some of his earlier articles on Contemporary Physics form the nucleus of a recently published book entitled "Introduction to Contemporary Physics" (D. Van Nostrand Company).

C. J. Davisson, B.S., University of Chicago, 1908; Ph.D., Princeton University, 1911; Instructor in Physics, Carnegie Institute of Technology, 1911–17; Engineering Department of the Western Electric Company, 1917–25; Bell Telephone Laboratories, 1925–. Dr. Davisson's work since coming with the Bell System has related largely to thermionics and electronic physics.

S. W. Dean, A.B., Harvard University, 1919; Cutting and Washington, Inc., 1917; Ensign, U. S. Naval Reserve Force, 1918; Radio Corporation of America, 1919–25; Department of Development and Research, American Telephone and Telegraph Company, 1925–. Mr. Dean's work has been chiefly in connection with long wave transatlantic radio systems.

H. H. Glenn, B.S., Pennsylvania State College, 1909; Engineering Department, Western Electric Company, 1909–25; Bell Telephone Laboratories, 1925–. Mr. Glenn is engaged in apparatus development work with particular reference to tinsel cords, insulated wire and switchboard cable studies.

J. Herman, E.E., Lehigh University 1920; Department of Development and Research, American Telephone and Telegraph Company, 1920–. Mr. Herman has been engaged chiefly in telegraph transmission development work.

J. B. Johnson, B.S., University of North Dakota, 1913; M.S., University of North Dakota, 1914; Ph.D., Yale, 1917; Engineering Department, Western Electric Company, 1917–25; Bell Telephone Laboratories, 1925–. Dr. Johnson has been engaged in the development and application of special vacuum tubes and gaseous discharge devices.

Marrison, W. A., Royal Flying Corps, later Royal Air Force, Canada, 1917–18; B.S. in Physics, Queens University, Canada, 1920; A.M. in Physics and Mathematics, Harvard University, 1921; Western Electric Company, 1921–25; Bell Telephone Laboratories, 1925–. Mr. Marrison is engaged in the study of picture transmission and methods for the production of constant frequency.

E. J. Murphy, B.S., University of Saskatchewan, Canada, 1918; McGill University, Montreal, 1919–20; Harvard University, 1922–23; Engineering Department, Western Electric Company, 1923–25; Bell Telephone Laboratories, 1925–. Mr. Murphy's work is largely confined to the study of the electrical properties of textiles and of electrical conduction in dielectrics in general.

J. R. Townsend, Engineering Department, Western Electric Company, 1919–25; Bell Telephone Laboratories, 1925–. Mr. Townsend has been largely concerned with the testing of telephone apparatus and more recently with the development of requirements of test for metallic materials.

R. R. Williams, B.S., University of Chicago, 1907, M.S., 1908; Research Chemist, Bureau of Science, Philippine Islands, 1908–15; Bureau of Chemistry, U. S. Department of Agriculture, 1915–18; Engineering Department, Western Electric Company, 1918–25; Bell Telephone Laboratories, 1925–. Mr. Williams has done extensive research work on submarine cable insulation. Since 1925, as Chemical Director, he has been in charge of the Chemical Laboratories of the Research Department of the Bell Telephone Laboratories.

W. T. Wintringham, B.S. in Electric Communication Engineering, Harvard University, 1924; Department of Development and Research, American Telephone and Telegraph Company, 1924–. Mr. Wintringham's work has been largely along the line of long wave radio transmission and measurement problems.

E. B. Wood, B.S., Princeton University, 1915; M.A., Princeton, 1916; Teaching Staff, Cascadilla School, 1916–17; Captain, Coast Artillery Corps, U. S. Army, 1917–19; Teaching Staff, Pratt Institute, 1919–20; Engineering Department, Western Electric Company, 1920–25; Bell Telephone Laboratories, 1925–. Mr. Wood's work has been connected mainly with the development of central office wire and cable and humidity testing equipment.

The Bell System Technical Journal

July, 1929

Magnetic Alloys of Iron, Nickel, and Cobalt [1]

By G. W. ELMEN

SYNOPSIS: Recent investigations of magnetic properties of alloys of iron, nickel and cobalt have resulted in the discovery of materials of remarkable magnetic properties previously unknown. In a brief review, early experiments that led to the discovery of these materials and the magnetic properties of the entire field are discussed. Those groups of alloys of outstanding scientific and technical importance such as the permalloys and the perminvars and special heat treatment required for development of special magnetic properties are taken up in detail. A theory is suggested to account for some of the magnetic characteristics, and a few of the practical applications of these materials are described.

IF the three ferromagnetic metals—iron, nickel and cobalt—are melted together in various proportions, all of the resultant alloys are magnetic when the materials used are reasonably pure. The magnetic properties of the alloys, however, vary greatly with composition and heat treatment. Some alloys have been found to be nearly non-magnetic; others have higher permeability than any hitherto known magnetic material and may be magnetically saturated in the earth's field. Still others are superior at high field strengths or excel in constancy of permeability at low fields.

In the Bell Telephone Laboratories we have been especially interested in these alloys and have made a fairly complete survey of the magnetic properties of the whole field of compositions of these metals. In this address I wish to tell you how our interest in those alloys was aroused, what our procedure was in carrying out the investigation, and what some of the principal results were. I shall also refer to some of the particular applications which we have made as a result of our discoveries.

I first became interested in these alloys in 1913. At that time I was looking for a magnetic material which would be more suitable for certain uses in the electrical communication field than the iron then used. Of the large number of alloys investigated, several contained principally iron and nickel. One of these was particularly interesting. The composition of this alloy was approximately 70 per cent nickel and 30 per cent iron; it was a commercial alloy used as a special resistance material. In the hard worked condition in which it was

[1] *Journal of the Franklin Institute*, Vol. 207, May, 1929, pp. 583–617.

supplied, its magnetic properties were not even as good as iron in a similar magnetic condition, but when heat treated it was superior to the iron at low field strengths, the region in which I was especially interested.

This alloy differed from iron in another important characteristic. Experience with iron had shown that the best magnetic quality was obtained when the material was heated to a high temperature, and then cooled slowly to room temperature. It was considered particularly important to cool slowly in order to give the iron time to pass through its transformations and to allow it to build up a large grain structure. When the nickel-iron alloy was heat treated in this manner, it was found to have lower permeability than it has when cooled fairly rapidly.

The discovery that rapid cooling was required for this alloy to give the best magnetic quality was one of the major contributions from our early work. It showed us that one of the important factors in developing the magnetic properties of new alloys was the determination of the rate of cooling for the best magnetic quality for each alloy.

Another difference between this alloy and iron relates to the energy loss caused by the hysteresis at low flux densities. In this range iron has lower hysteresis loss when it is in a mechanically hard condition than when it is well annealed. As the flux density increases, however, the hysteresis loss of the hard material increases more rapidly than does that of the annealed and at medium and high flux density, the mechanically hard iron is much inferior. Tests on the nickel-iron alloy show that both for high and low flux densities the hysteresis loss was a great deal higher for the mechanically hard material than for the one heat treated to give the best magnetic quality.

The discovery of the unusual magnetic properties of this 70–30 per cent nickel-iron alloy gave us the lead that we were looking for, and started our investigation of the magnetic properties of the whole series of nickel-iron alloys. We found, as we had reason to expect, that the 70–30 per cent alloy was one of a large group of alloys in the same series which had similar magnetic properties. In fact substantially all of the alloys containing more than 30 per cent nickel had similar characteristics except that some of them were not as sensitive to heat treatment as the first alloy we had tested.

Because of the technical possibilities of the nickel-iron alloys, we spent several years in their investigation and their commercial application. We were especially concerned with increasing the resistivity of a number of these alloys, and with this in view we added other

elements to the iron and nickel in order to determine their effect both on the resistivity and the magnetic properties. Copper, chromium, molybdenum, tungsten, and tantalum are a few of those we added; also the magnetic element cobalt. The striking results obtained with the addition of cobalt to the iron-nickel alloys led us to make a com-

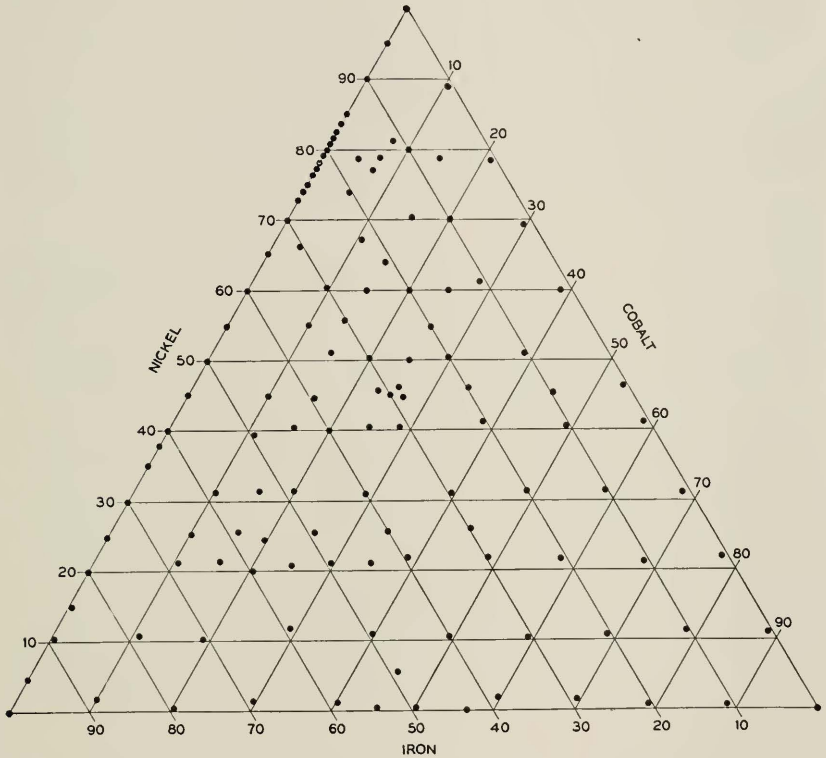


Fig. 1—Composition diagram—Ni-Fe-Co series. Dots show compositions tested.

plete survey of the whole field of alloys containing the three magnetic metals.

A convenient way of showing graphically the number of alloys we investigated and the distribution of compositions of these alloys is afforded by the equilateral composition triangle in Fig. 1. In this triangle the metals—iron, nickel, and cobalt—are represented by the corners; the binary alloys by points on the three sides, and the ternary alloys by points within the area of the triangle. Each alloy we investigated is indicated by a dot and the location of each dot indicates the proportion in per cent of the three metals in the alloy. A glance

at the diagram will give you an appreciation of the completeness with which the field was covered.

In preparing these alloys we used the best commercial materials available. These contained small amounts of impurities, some of which affect unfavorably the magnetic properties. It was considered beyond the scope of this investigation, in which such a large number of alloys were required, to attempt to remove these impurities completely or even partially. Moreover we were especially interested in alloys which could be reproduced on a commercial scale, without too great cost due to refinements of raw materials or to methods of preparing the alloys.

Throughout the investigation we have followed a standard procedure for preparing our samples. We have also followed a standard procedure for heat treating and in magnetic measurements. The result is that we have accumulated a large mass of data over a number of years, all of which may be significantly compared.

The alloys were cast from Armco iron, electrolytic nickel, and a very high grade of commercial cobalt containing only small amounts of impurities. They were melted together in the desired proportions in a silica crucible in a high frequency induction furnace. The metal was cast into bars 18 in. long and $\frac{3}{4}$ in. in diameter. The bars were rolled or swaged into $\frac{1}{4}$ in. rods; then they were drawn into wire and flattened and trimmed into $\frac{1}{8}$ in. by .006 in. tape. The material was annealed several times in the reduction process, for the cold working hardened the alloys rapidly and made them difficult to work.

To prepare the tape for heat treatment and subsequent magnetic measurements, about 30 ft. of it was wound spirally into a ring of 3 in. inside diameter, the ends being spot welded to the adjacent turns. Care was taken to wind the rings loosely to prevent the turns from sticking to each other during annealing.

For convenient comparison of the magnetic properties of the alloys with those of the metals from which the alloys were cast, sample lots of the iron, nickel, and cobalt which we used in making the alloys, were melted, cast, and prepared in the same manner as the alloy test samples.

In Fig. 2 the various steps through which the alloys had to pass in the mechanical reduction to tape are shown. Between each step in the reduction, as indicated by a sample, the alloy had to be annealed to remove the mechanical hardness. A sample ring wound from the finished tape, ready for heat treatment, is also shown in the figure.

The next step in the process of preparing these alloys for magnetic measurements is the heat treatment. Early experience with the

nickel-iron alloys indicated the need for a variety of such treatments, but of course it was impossible to apply to every alloy such a complete variety of heat treatments, that all of the combinations of heat treatment and composition could be known. Our procedure was to apply three types of heat treatment.

A number of rings of a given composition were packed in a nichrome pot. The pot was placed in an electrical resistance furnace, the temperature of the furnace raised to between 900° and 1,000° C. and held at that temperature for one hour. The current was then turned

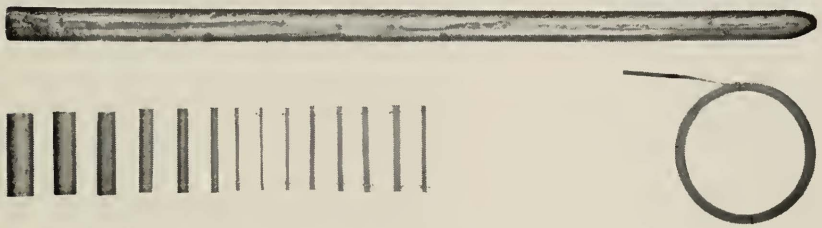


Fig. 2—Cast alloy bar and intermediate stages of samples in reduction to tape. Also ring from tape ready for heat treatment.

off and the pot cooled with the furnace. Ten hours were required for the furnace to cool to the temperature of the room. Between 700° C. and 400° C. the rate of cooling was approximately 1.5° C. per minute.

At least two rings of each composition were always annealed together. One of these rings received no further heat treatment. The second ring was heated for 15 minutes in a furnace held at 600° C., then removed and cooled rapidly on a copper plate in the air. With this cooling the rate was approximately 20° C. per second. In some cases a third annealed ring was heated 24 hours at 425° C.

In the discussions and in the figures, the rings which received the first heat treatment only are referred to as "annealed," those reheated to 600° C. and rapidly cooled as "air quenched," and those held for a long time at 425° C. as "baked." The magnetic measurements were made on these rings.

The magnetic induction, or flux density, was determined for a large number of magnetizing forces, beginning at a few thousandths of a gauss and increasing in uniform steps up to 100 gauss. Magnetization curves were plotted from these measurements. The induction was also determined for each composition at a magnetizing force of 1,500 gauss.

The permeabilities were computed from the induction measurements,

and were plotted either against the flux density or the magnetizing force, depending on which graph illustrated best the characteristics of the material. At the lower ends these curves were extended to zero field strengths. Their intercepts on the permeability axis are the initial permeabilities. The maximum permeabilities were also obtained from these curves.

For determining hysteresis loss, two methods were used. In some cases hysteresis loops were plotted from ballistic measurements, and in other cases a direct determination of hysteresis loss was made from the apparent alternating-current resistance of a winding wrapped around the sample. Ordinarily the hysteresis loop was obtained for one condition only in which the flux density was varied between plus and minus 5,000 gauss. For some of the alloys of special interest, a large number of loops were obtained for different magnetizations, the maximum flux densities varying from 100 or less to 5,000 gauss.

In illustrating the magnetic properties I will be forced to limit the discussion to a few outstanding values for each alloy, and by comparing these, obtain a general view of the relation of the magnetic properties to composition. The values I have selected are the intrinsic inductions for two magnetizing forces, 50 and 1,500 gauss respectively, the initial and the maximum permeabilities and the hysteresis loss for a maximum flux density of 5,000 gauss. For a number of alloys which represent regions of composition with magnetic properties of special interest, curves for magnetization and permeability will be shown. A number of hysteresis loops for different maximum flux densities will also be given.

In illustrating graphically the relations between the magnetic properties and compositions of ternary alloys, it is convenient to plot these quantities in the form of solid diagrams. Such a diagram is shown in Fig. 3 constructed for initial permeabilities of the alloys in the annealed condition. In this figure the composition triangle, Fig. 1, is used as the base. On this triangle, verticals are erected proportional to the numerical values of the initial permeability. The ends of these verticals give a contour of the upper face of the figure. With a sufficient number of alloys this surface represents fairly accurately the values of the initial permeabilities for all compositions. The edges of the surface give the initial permeabilities of the binary alloys and the rest of the surface those of the ternaries.

The coordinates of the composition triangle, the intersections of which give the compositions of the alloys in 10 per cent variations, are projected to the surface and are represented by the narrow black lines. The heavy black lines on this surface are contours such as

you see on topographical maps and represent the intersections with the surface of planes parallel to the base at definite elevations. Each line locates the compositions of those alloys which have equal values of initial permeability.

In this solid diagram of the initial permeabilities for the alloys in annealed conditions, there are three regions of composition which are

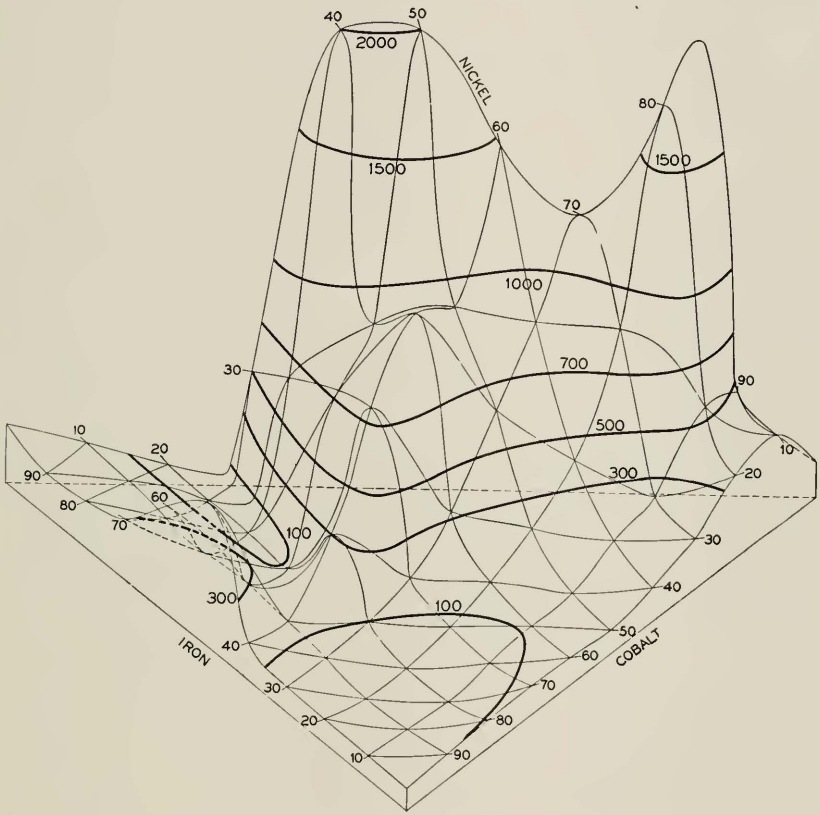


Fig. 3—Initial permeabilities for annealed alloys.

of special interest. The first of these is in the binary iron-nickel series represented by the back face of the diagram, which may be referred to as the nickel-iron plane. The nickel percentage is indicated by the numbers at the intersections of the permeability coordinates with the surface. The left-hand corner represents 100 per cent iron and the right-hand corner 100 per cent nickel, with initial permeabilities of 250 and 200, respectively. Between these limits of composition

the initial permeability varies from approximately 100 to more than 2,000. With small additions of nickel to iron, the permeability drops gradually until the added amount is approximately 28 per cent of the composition. From that point there is a rapid rise in the permeability

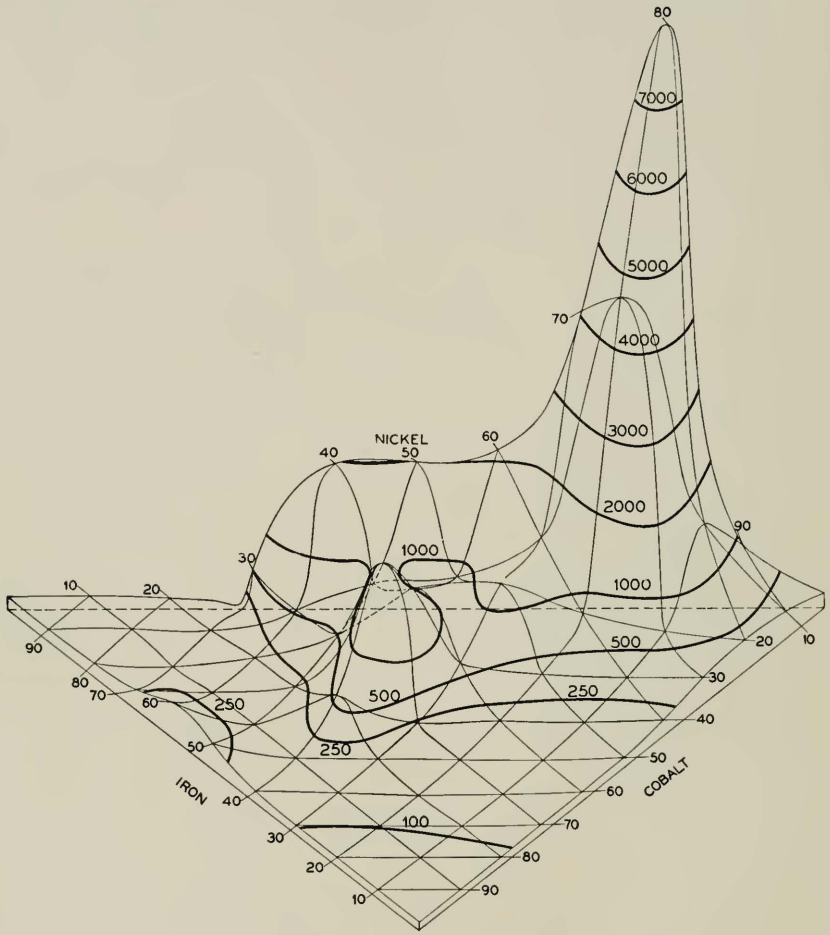


Fig. 4—Initial permeabilities for air-quenched alloys.

with the increase in the nickel content reaching a maximum of over 2,000 for the alloy containing approximately 45 per cent nickel. It then decreases as more nickel is added, reaching a second minimum of about 1,200 for 70 per cent nickel and increases again reaching a second maximum of approximately 1,900 at 83 per cent nickel.

Another region of special interest is found in the ternary series in

front of the first maximum in the nickel-iron series. If you follow the coordinate from 40 per cent nickel, you will note an elevation between the 700 and 1,000 permeability contour lines. The ternary alloys of this region are remarkable, not because of their high permeabilities although they are higher than for the surrounding region, but because of the constancy of permeability and low hysteresis loss at low magnetizing forces.

Another region of interest is located in the iron-cobalt plane between 40 per cent and 70 per cent iron. This is the plane in the figure on which the iron percentages are marked. The initial permeability for the highest point is over 600, more than twice the initial permeability of Armco iron.

When the alloys are air quenched, the initial permeabilities change in some regions very materially and give us an entirely different looking solid diagram as shown in Fig. 4. Because of the very high values of the permeabilities a different scale was used in this figure from that used in Fig. 3. It is interesting to note how the rapid cooling has affected some of the binaries in the iron-nickel plane. The rise in permeability begins at about 28 per cent, the same as for the annealed alloys, and the initial permeabilities are substantially the same, up to about 45 per cent nickel. Between 45 per cent and about 90 per cent nickel, the permeabilities have increased to a remarkable degree. The valley we saw in the region 45–75 per cent nickel for the annealed alloys has disappeared, and from 55 per cent upward there is a rapid increase reaching a peak value of approximately 8,000 for the composition containing $78\frac{1}{2}$ per cent nickel, an increase of over four times the permeability of the annealed alloy of the same composition.

Air quenching also increases the initial permeabilities of the group of ternary alloys which in the annealed condition showed a maximum characterized by a low hysteresis loss and a substantially constant permeability. This increase in permeability, however, is accompanied by a material increase in the hysteresis loss at low flux densities and a decrease in the constancy of permeability.

In contrast to the alloys just discussed, the iron-cobalt series in the neighborhood of 50 per cent iron gives a lower permeability upon rapid cooling. In this respect they behave the same as iron and cobalt.

The maximum permeabilities for the annealed alloys shown in Fig. 5 give a surface very similar to that for the initial permeabilities. The most interesting difference is the rapid decrease in permeability of iron as indicated at the left-hand corner of the diagram by the

addition of small amounts of cobalt or nickel, 10 per cent of either being sufficient to reduce the maximum permeability 85 per cent. All of the iron-nickel alloys between 35 per cent and 85 per cent nickel have as high or higher permeability than Armco iron.

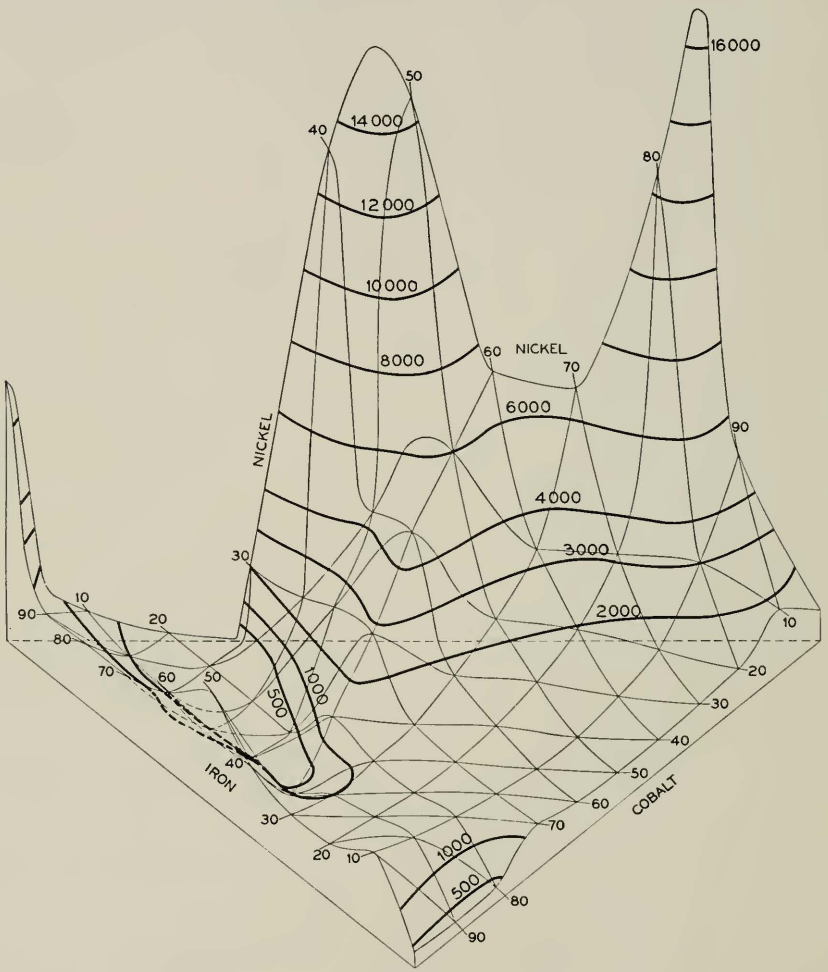


Fig. 5—Maximum permeabilities for annealed alloys.

I have not prepared a diagram for the maximum permeabilities for the air-quenched alloy. However, I can say that the surface obtained resembles the one shown for the initial permeabilities for the rapidly cooled alloys, although, of course, the maximum permeabilities are of different magnitudes. As high as 120,000 has been obtained for the alloy containing 78.5 per cent nickel.

The group of alloys of high permeability for low magnetizing forces in the iron-nickel series we have named permalloy.[2] The permalloys, therefore, include nickel-iron alloys containing more than approximately 30 per cent nickel. In referring to specific compositions in this group we have found it convenient to use as a distinguishing prefix for each composition its nickel content. For example, 78.5 permalloy is the alloy containing 78.5 per cent nickel and 21.5 per cent iron. When other elements are added to the permalloys, the chemical symbol and the percentage of the added element are also added to

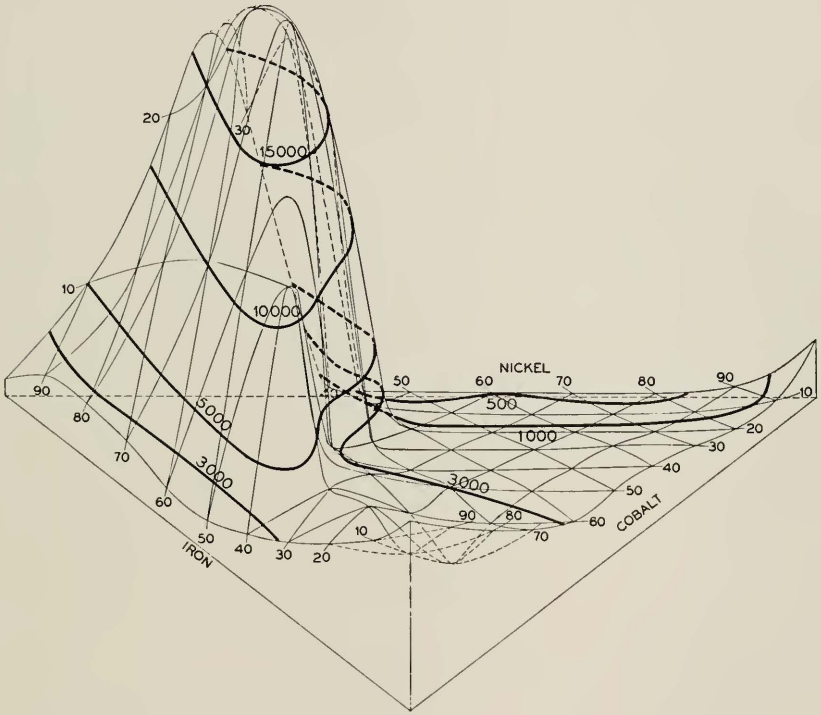


Fig. 6—Hysteresis losses, ergs per cm.[3] per cycle, for maximum flux density of 5,000 gauss for annealed alloys.

the prefix. For example, 3.5–78.5 Mo-permalloy is an alloy containing 3.5 per cent molybdenum, 78.5 per cent nickel, and 18 per cent iron.

The ternary alloys of constant permeability and extremely low hysteresis loss at low flux densities we have also grouped together under a common name, perminvar.[3] The limits of composition for the perminvars are less easily defined than for the permalloys, because the transition in magnetic properties is not as marked with small

[2] H. D. Arnold and G. W. Elmen, *Journal of Franklin Inst.*, May, 1923, p. 621.
[3] G. W. Elmen, *Journal of Franklin Inst.*, Sept., 1928, p. 317.

changes in composition. We have found that compositions between 10 per cent and 40 per cent iron, 10 per cent and 80 per cent nickel, and 10 per cent and 80 per cent cobalt have marked perminvar characteristics.

The hysteresis losses of the alloys in the annealed condition are plotted in Fig. 6 for a maximum flux density of 5,000 gauss. This

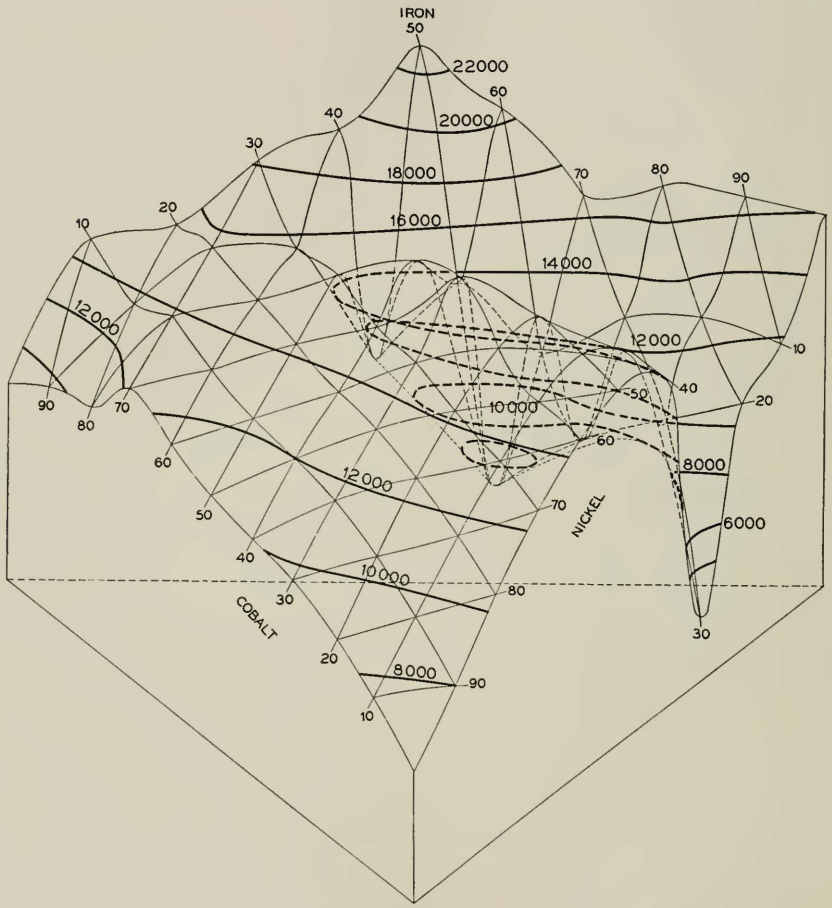


Fig. 7—Intrinsic inductions for annealed alloys at $H = 50$.

diagram illustrates again the abrupt change in the magnetic properties as we pass from alloys with large iron content to those in which nickel predominates. In this figure the lowest energy losses are those for alloys in the neighborhood of the 78.5 permalloy composition. In the iron-cobalt series the 50 per cent cobalt alloy has the lowest hysteresis loss.

The intrinsic inductions, which are those parts of the inductions contributed by the magnetic material, are shown in Fig. 7 for a magnetizing force of 50 gauss. In the figure the triangle has been turned through 120° clockwise from its position in the previous figures, placing the iron-cobalt alloys in the back of the diagram. For this

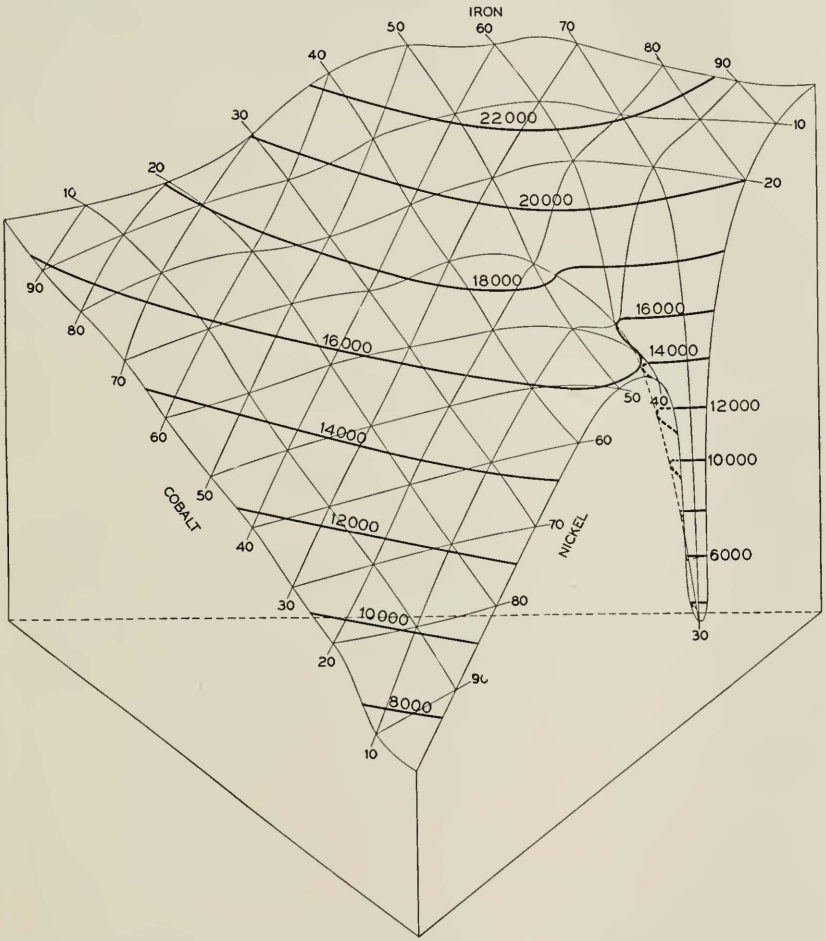


Fig. 8—Intrinsic inductions for annealed alloys at H = 1,500.

magnetization the 50 per cent iron-cobalt alloy is superior to any of the others.[4] Another interesting part of this diagram is the deep depression at about 30 per cent nickel, in the iron-nickel plane—now the right-hand front face of the diagram. This depression extends back for some distance into the ternary alloys.

[4] This was also found by Ellis, Engineering and Science Series No. 16, June, 1927, Rensselaer Polytechnic Institute.

The intrinsic inductions are also shown for a magnetizing force of 1,500 gauss in Fig. 8. Here we find that the irregularities of the surface in the previous figure have largely disappeared and the surface has become fairly flat. The inductions for the alloys in the neighborhood of 34 per cent cobalt and 66 per cent iron have now increased so that they are practically the same as for the 50 per cent composition.

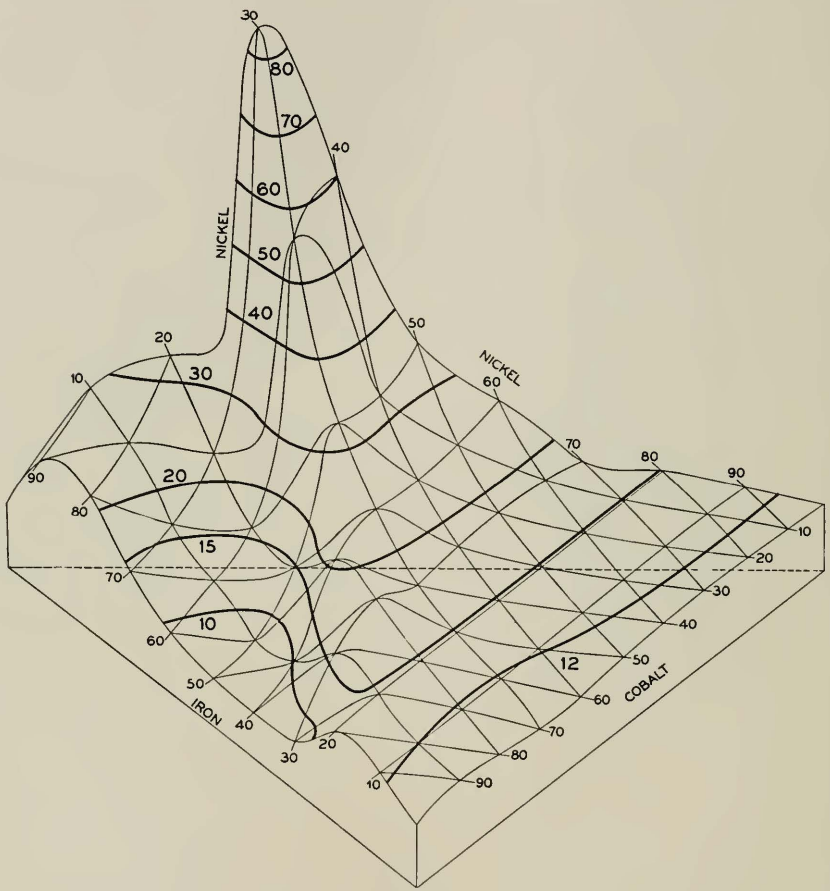


Fig. 9—Resistivities, microhm-cm., for annealed alloys.

The valley in the neighborhood of 30 per cent nickel has largely disappeared except for the alloys without cobalt which are much harder to saturate.

In Fig. 9 a solid diagram has also been constructed for the resistivity of these alloys in the annealed condition. It is interesting to note the high resistivities of some of the nickel-iron alloys—again in the

back of the diagram—and the low values for some of the cobalt-iron alloys in the 45–70 per cent cobalt range of compositions.

In the discussion of the magnetic properties of these alloys, I have pointed out that there are three regions of composition which are of special interest from the standpoint of applications because of their unusual magnetic properties. One of these is in the iron-cobalt series. A representative alloy of this group is the 50 per cent cobalt composition. In Fig. 10 the magnetization curves for this alloy and for

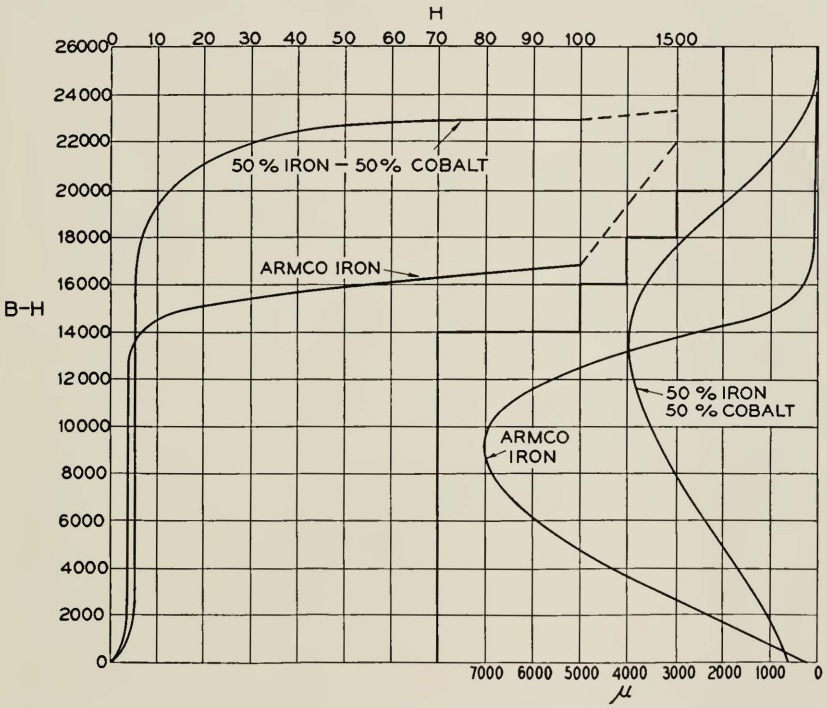


Fig. 10—Magnetization and permeability curves for the 50 per cent iron, 50 per cent cobalt composition and for Armco iron.

a sample of iron are shown on the left and the permeability curves plotted against the same scale of intrinsic inductions on the right. The scale for the magnetizing force of the magnetization curve is at the top of the figure. Below an induction of 1,000 and above 13,000 the permeability of the alloy is higher than for iron. Its initial permeability is 600 and the maximum permeability is 4,000. For a magnetizing force of 100 gauss the intrinsic induction is nearly 23,000 gauss. For Armco iron the initial and maximum permeabilities are 250 and 7,000, respectively, and the intrinsic induction for a magnetizing force of 100 gauss is 17,000.

The perminvars are the next group with interesting characteristics. Their remarkable constancy of permeability at low magnetizing forces and their extremely low hysteresis loss at low flux densities make them of unusual interest. The composition 45 per cent nickel, 25 per cent cobalt and 30 per cent iron is a typical alloy for this group. Permeability curves for this alloy for three types of heat treatment are plotted against magnetizing force in Fig. 11. The insert in the upper right-hand corner shows the lower parts of these curves plotted to a larger scale. For the baked alloy the permeability is constant at 300 with

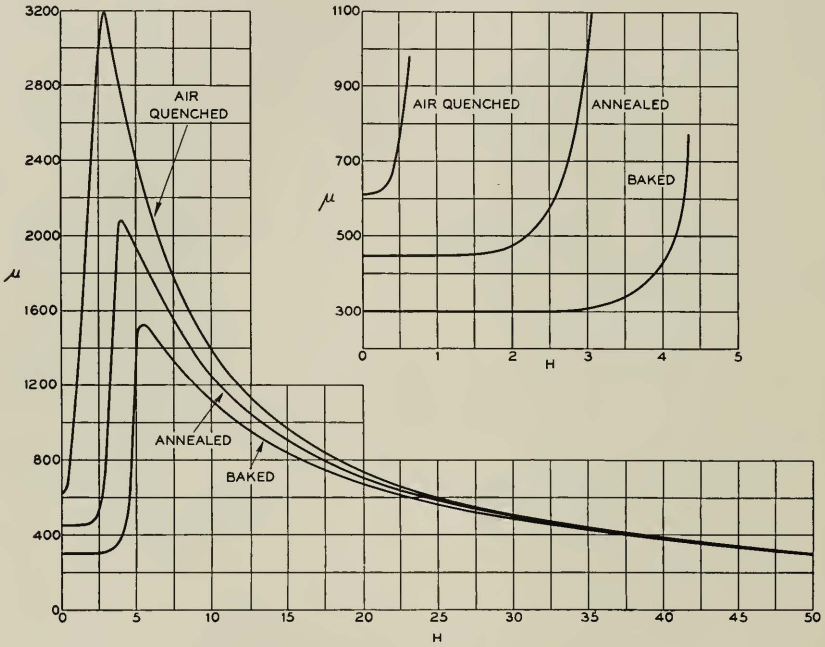


Fig. 11—Permeability curves for a perminvar alloy containing 45 per cent Ni, 25 per cent Co, 30 per cent Fe.

an increase of the magnetizing force from zero to $2\frac{3}{4}$ gauss. The initial permeability for the air-quenched alloy is more than twice that of the baked, but the constancy of permeability is a great deal less. This illustrates the close relation of these two properties. As the magnetizing force approaches 40 gauss the differences in the permeabilities, resulting from different rates of cooling, disappear.

The hysteresis loops for this composition are also of unusual interest. Fig. 12 illustrates the hysteresis characteristics for this alloy, in air-quenched and baked condition, for three maximum flux densities, 750, 1,000 and 5,000 gauss, respectively. For the lowest flux density of

the baked sample the ascending and the descending branches of the loop coincide and the loop is represented by a straight line. For the next higher flux density the loop, for the same heat treatment, begins to have a measurable area. At low values of induction, however, the two branches of the loop for the baked alloy approach each other and often come together completely at the origin. The complete loop for a maximum flux density of 5,000 gauss also shows this peculiar shape for the baked alloy. The loop is constricted in the middle, the two branches almost passing through the origin. The air-quenched alloy also shows tendency of constriction but much less than the baked. The areas of the two loops show that for this value of maximum flux

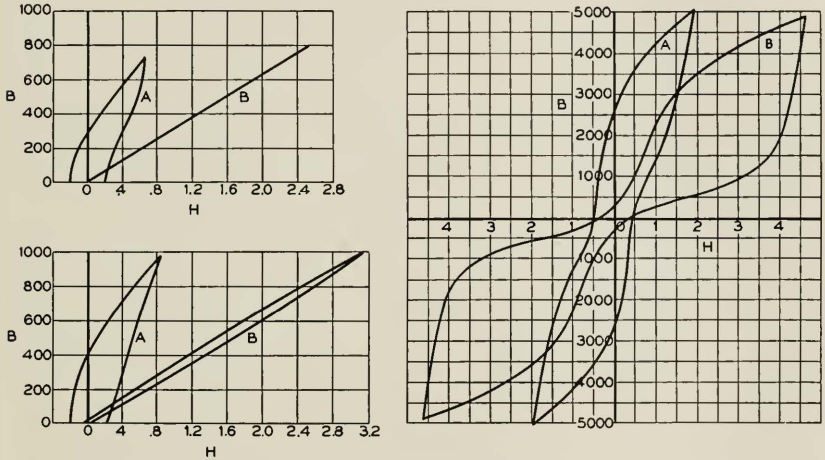


Fig. 12—Hysteresis characteristics for a perminvar alloy containing 45 per cent Ni, 25 per cent Co, 30 per cent Fe. A = air quenched; B = baked.

density the hysteresis loss for the baked sample is greater than for the air-quenched. This condition is the reverse of that for lower flux densities.

The last group of alloys of special interest are the permalloys. In Fig. 13 I have taken the curves for initial permeabilities of the iron-nickel series from the solid diagrams for the air-quenched and the annealed conditions and plotted them on the same scale of coordinates. A curve of initial permeabilities for a series of baked alloys is also plotted in this figure. The baking process for these alloys differed from the one we usually employed in that each alloy was baked until no decrease in the permeability resulted from further baking. For some alloys several weeks were required before this condition obtained. The time was shortest for compositions between 60 per cent and 80

per cent nickel. On both sides of this range the time necessary for stabilization increased both with increasing and decreasing nickel content. With nickel content of less than 42 per cent and more than 88 per cent, approximately, no difference sufficiently large to be attributed to the baking could be observed, even after the alloys had been baked for several weeks.

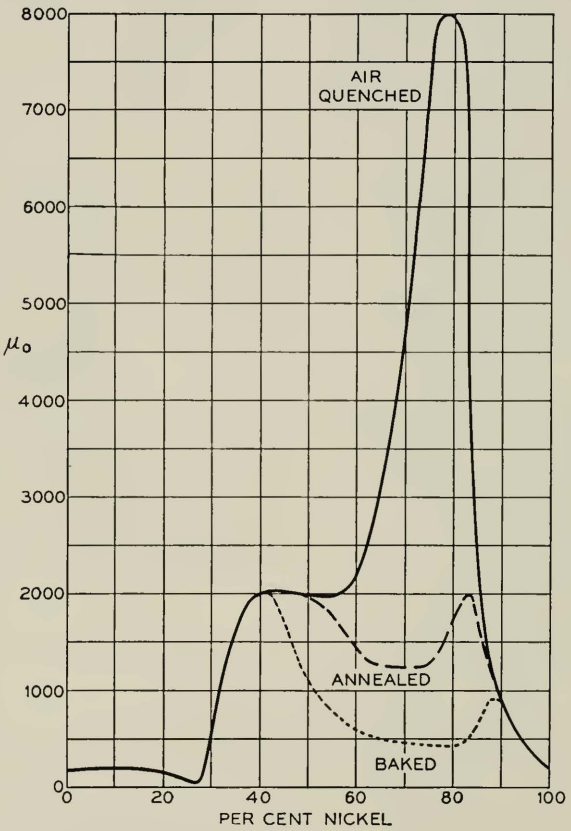


Fig. 13—Initial permeabilities for the Fe-Ni series.

The 78.5 permalloy composition in this series is of special interest because this alloy when air quenched gives the highest initial and maximum permeabilities. The magnetization curves for this composition are plotted in Fig. 14 for annealed and air-quenched samples, and for a sample of annealed iron. The lower part of this graph is shown in the insert on a larger scale. The curves are plotted for magnetizing forces between 0 and 10 gauss only. In these curves

the induction for the annealed sample remains lower than for the air-quenched sample, but as the force increases there is a tendency for the two curves to approach each other. This continues with still further increase in the magnetizing force and beyond 50 gauss the two curves coincide.

The permeability curves for these samples are plotted in Fig. 15, illustrating the great difference in their maximum permeabilities. For the annealed and the air-quenched samples, the initial perme-

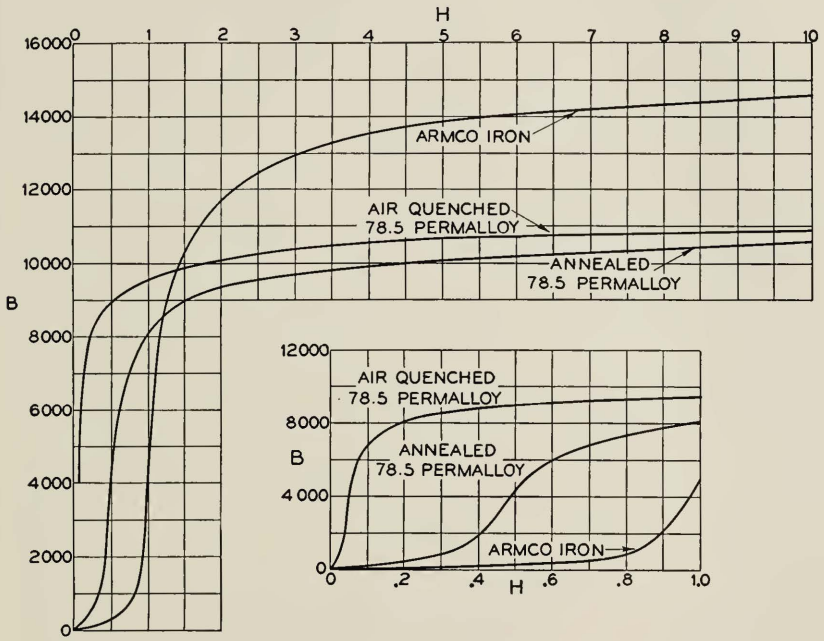


Fig. 14—Magnetization curves for 78.5 permalloy and for Armco iron.

abilities are 2,000 and 9,000, and the maximum permeabilities 10,000 and 87,000, respectively. The initial and the maximum permeabilities for the iron sample are 250 and 7,000, respectively.

The effect of air quenching on the energy loss is illustrated in Fig. 16 in which hysteresis loops for a maximum flux density of 5,000 gauss are plotted for the air-quenched and the annealed samples. The energy loss for the rapidly cooled sample is only 35 per cent of that for the annealed.

In deciding on a rate for air quenching we selected the rate which gave the highest initial permeability for the 78.5 permalloy. In Fig. 17 I wish to illustrate how small changes in this rate affect the initial

and the maximum permeabilities for this composition. The scale for maximum permeability is on the left and for initial permeability on the right. The cooling rate is in degrees centigrade per second. It is interesting to note that the highest initial permeability was obtained

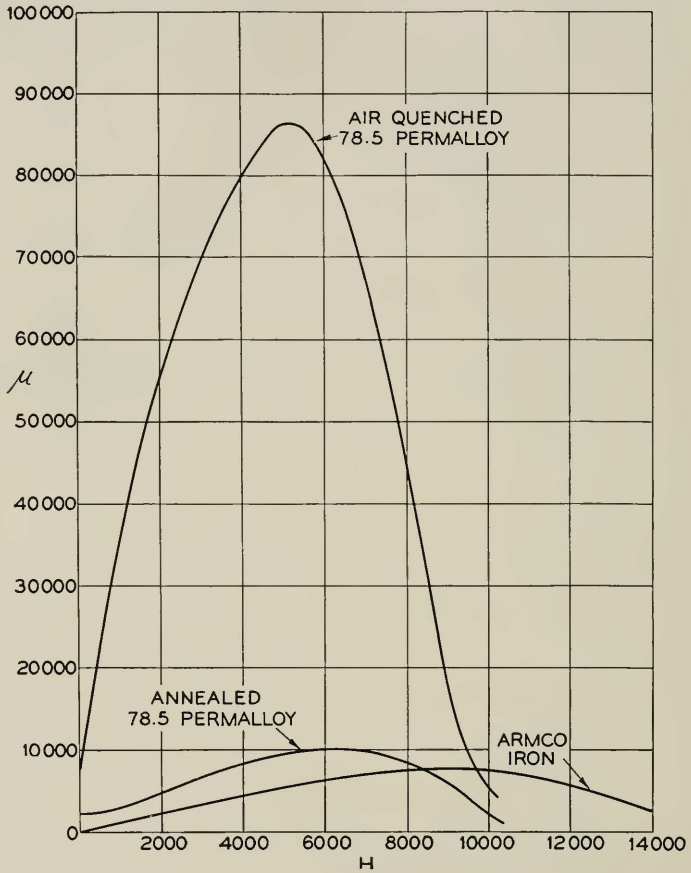


Fig. 15—Permeability curves for 78.5 permalloy and for Armco iron.

when the cooling rate was approximately 20° C. per second, while a rate four times as rapid gave the highest maximum permeability. Ten thousand and 120,000 were the highest initial and maximum permeabilities, respectively, for these particular samples. These values are not the highest we have obtained for this composition. Test samples from other castings have given as high as 13,000 for initial permeability and upwards of 400,000 for maximum.

Fig. 18 illustrates the manner in which the permeability at a low magnetizing force for a rapidly cooled 78.5 permalloy composition is affected by passing it through a temperature cycle between room temperature and 650° C. The permeability was measured for a constant magnetizing force of .003 gauss as the temperature was passed through this cycle. The rate at which the sample was heated and

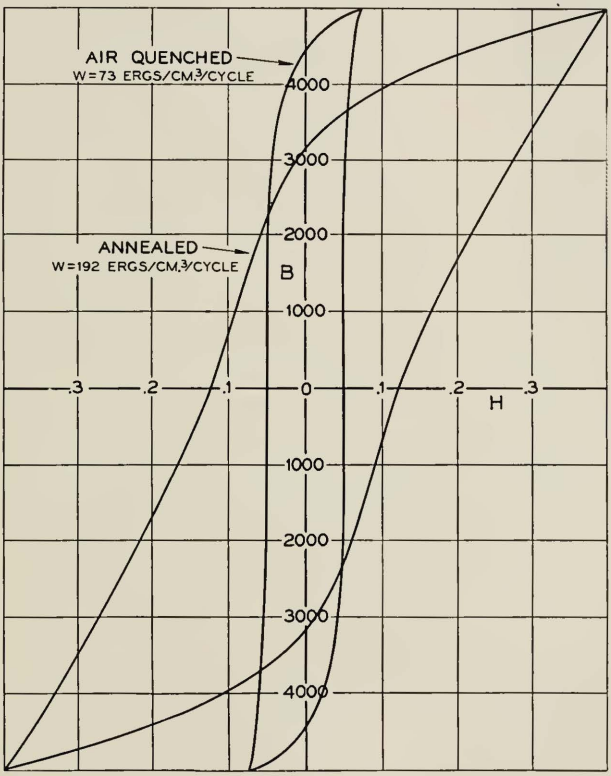


Fig. 16—Hysteresis loops for 78.5 permalloy.

cooled was rather slow, requiring upwards of two hours to complete the heat run.

The permeability of the air-quenched sample was 7,000 at the start, increasing rapidly up to about 315° C. With further increase in temperature there is a rapid drop in permeability until 500° C. is reached. With further increase in temperature the permeability rises very rapidly to a sharp peak at about 530° C. and then decreases, the alloy becoming non-magnetic at 590° C. With decreasing temperature the alloy again becomes magnetic at about the same temperature, at

which the magnetism disappeared and the curve for decreasing temperature is the same as for increasing temperature over a short range. At 500° C. the curve for decreasing temperature does not follow its original path but continues to drop until the temperature reaches about 425° C. From that point on until room temperature is reached the return curve is nearly horizontal.

At room temperature the permeability is only 2,000, a drop of 5,000 from its value at the beginning of the run. With the alloy in this condition, if a second cycle is run, the permeability both for increasing

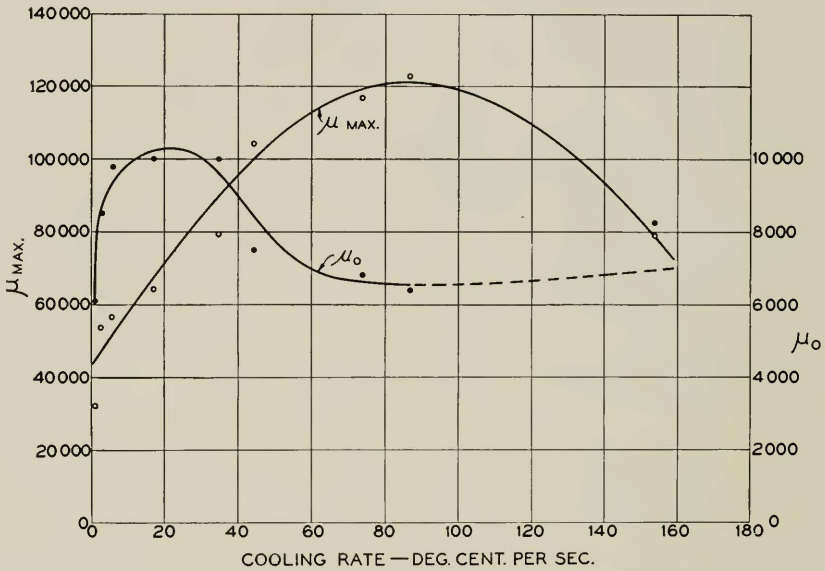


Fig. 17—Initial (μ_0) and maximum (μ_{max}) permeabilities for 78.5 permalloy for different rates of air quenching from 600° C.

and decreasing temperature will be substantially the same as the one in the figure for decreasing temperature. By heating the sample to 600° C. and air quenching, the magnetic properties at the beginning of the test are restored.

The connection between the permeability of this composition and the heat treatment in the temperature range below 600° C. is also illustrated by a series of tests in which annealed rings were air quenched from temperatures below 600° C. The rings were placed in a furnace and heated at 600° for 15 minutes. The temperature was then decreased slowly to 550° C. and held until the alloys had reached a constant condition. One of the samples was then taken out of the furnace and air quenched. The temperature of the furnace was then

dropped another step and the process of stabilization and air quenching repeated for the next ring. This was repeated for a number of temperatures until there was substantially no change between two rings quenched from successive temperatures. Another series of annealed rings was slowly heated in successive steps to the same temperatures, stabilized, and then air quenched.

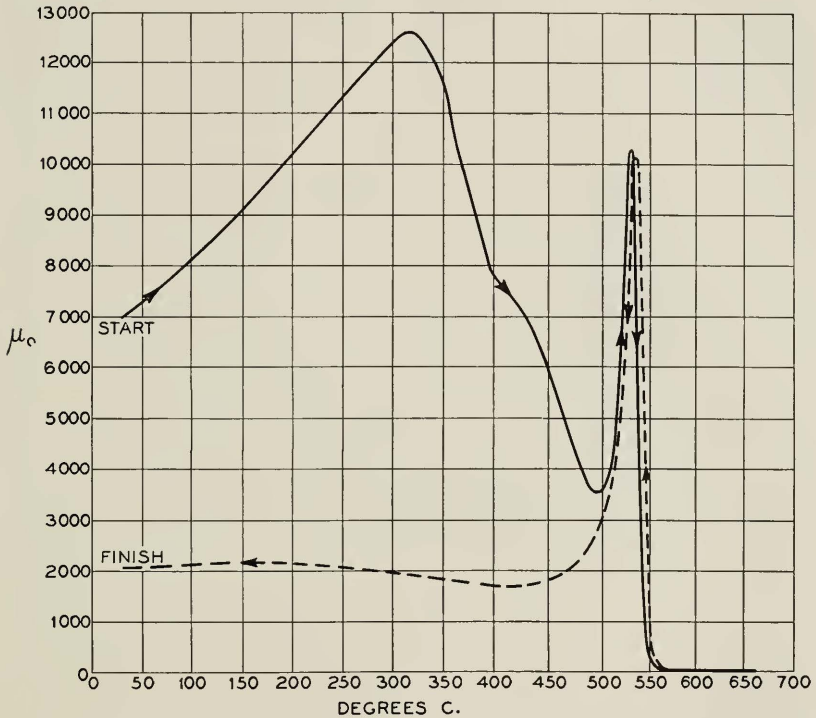


Fig. 18—Temperature-permeability curve for 78.5 permalloy: magnetizing force .003 gauss.

A comparison of the initial permeabilities of these rings showed that each ring attained a definite permeability, characteristic of the temperature from which it was air quenched whether it reached that point from a lower or a higher temperature, provided it was stabilized before air quenching. The air quenched rings stabilized between 500° and 400° C. showed the greatest change in permeability. The time required for stabilization at the higher temperatures was very short, only a few minutes being required in the neighborhood of 500° C., but it increased progressively as the temperature was lowered, and in the lower end of the range several days were required for the alloy to reach a constant condition.

Tests on other compositions showed that those alloys in which the magnetic properties depended on the rate of cooling below 600° C. gave results similar to those for the 78.5 permalloy. The temperature range was also found to be substantially the same as for the 78.5 permalloy, although in some cases there were indications that small changes occurred above 600° and below 400° C.

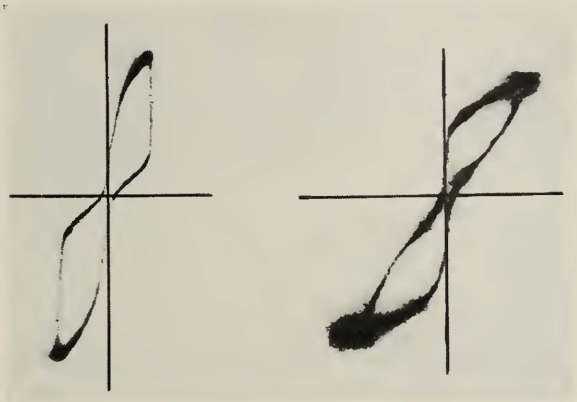
This connection between heat treatment and magnetic properties led me to conclude that the differences in these properties in the alloys which had been heat treated differently below 600° C. were caused by constitutional changes in the alloys. Such changes are common in alloys in the solid state, often at low temperatures. The progressive change in the permeability as the temperature of the alloy decreases slowly below 600° C., the gradual increase in the time required for a change to complete itself as the temperature drops, and the prevention of the change by rapid cooling through this temperature range support this conclusion.

It is well known that some alloys are in the state of homogeneous solid solutions at high temperatures, but segregate into two or more phases as the temperature drops. Such segregation ordinarily takes place in a definite temperature range, and is progressive in nature. It is a change of this type which I picture as taking place in these alloys during slow cooling. At the upper end of the critical temperature range for each alloy, the homogeneous solid solution begins to segregate into constituents of different composition. This segregation continues until the temperature has dropped to a point where no further changes take place. Rapid cooling prevents this segregation and the alloys remain after cooling in a metastable condition.

We would suppose that if such a change takes place, confirmatory evidences might be found. I shall refer, briefly, to some of our attempts to obtain evidence to confirm our speculations as to the nature of these magnetic changes.

The resistivity was measured for rapidly cooled and for baked alloys both of permalloy and perminvar compositions. These measurements showed that for both types of alloys, baking reduced the resistivity. For example, the resistivity of a 78.5 permalloy sample was measured after baking and after air quenching. The resistivities for the two conditions were 14.2 and 15.8 microhm-cm., respectively. Upon rebaking the resistivity again dropped to 14.5, about the same as before it was air quenched. This change is in accordance with what would be expected if a segregation took place with annealing. A homogeneous solid solution has the highest resistivity and segregation tends to lower it.

Other interesting evidence is obtained from the study of the hysteresis loops. For the perminvar alloys the constricted loops are very marked and easily produced. In the permalloys containing between 60 and 80 per cent nickel also there is a tendency to constriction in the slowly cooled alloy, not so marked but sufficiently prominent to lead me to believe that the same general changes occur in both groups of alloys, differing only in the nature of the segregates and the ease with which segregation occurs. Now we know that homogeneous magnetic materials have a characteristic type of hysteresis loop. For such materials there is no constriction in the middle, but the widest part of the loop is generally at that point. We also can



a *b*

Fig. 19—Hysteresis loops: *a*, perminvar (45 per cent Ni, 25 per cent Co, 30 per cent Fe); *b*, bi-metallic rod. Loop traced with cathode ray oscillograph.

construct hysteresis loops which have constrictions by making up cores of several materials in a parallel or parallel-series arrangement.[5] This is illustrated by the hysteresis loop *b* in Fig. 19. This loop is traced by a cathode ray oscillograph for a bi-metallic rod, 15 in. long consisting of a core of .04 in. diameter unannealed piano wire and a .006 in. wall permalloy tube, heat treated to give high permeability and fitting closely to the wire. Curve *a* is a loop similarly traced for a perminvar core. Though the magnetic circuit conditions for the two cores are not the same, the marked similarity of the two loops favors the view that the constricted loop of the perminvar results from segregation.

It is interesting to note in this connection that the examination by X-ray crystal analysis methods of these alloys has not given evidence

[5] E. Gumlich, *Arch. f. Elektrotechnik*, Vol. 9, p. 153, 1920.

of a segregation such as other evidence leads me to think must occur. Perhaps the reason for this is that the different constituent metals are so closely related that small changes in the structure cannot be detected by this means, or perhaps the reason is that the size of the groups of atoms making up the constituents is too small to be detected by X-ray methods.

While the variation in the permeability and in the other magnetic characteristics which could be affected by heat treatment may be explained by the theory of segregation, this theory does not explain why these alloys have such high permeabilities at low magnetizing forces.

Nor does this theory explain the unexpected magnetic characteristics, such as high saturation values of induction, or the low electrical resistance which characterize a large proportion of the alloys in the iron-cobalt series. It has been suggested by Weiss [6] that when the saturation values of an alloy are higher than they are for the constituent metals, it is an indication of the existence of an intermetallic compound. On this ground he has accounted for the high saturation values he found for an iron-cobalt alloy containing 34 per cent cobalt. In our investigation, which was not carried up to the high magnetizing forces used by Weiss, the 50 per cent cobalt alloy gave us as high flux densities as any in the series for magnetizing forces upwards of 1,500 gauss.

It has been found in the study of intermetallic compounds that one indication of their existence is a low resistivity. It is generally believed that if an alloy has lower resistivity than any of its constituent metals, an intermetallic compound exists. In our measurements of the resistivity of the iron-cobalt series we found that the alloys with lowest resistance were those containing between 25 per cent and 60 per cent iron. There is a rather abrupt decrease in the resistivity as the percentage of iron increases beyond 25 per cent. Beyond 50 per cent iron there is a gradual increase with a maximum at about 85 per cent iron. From these measurements we would conclude that if an intermetallic compound exists, it is of a higher cobalt percentage than that suggested by Weiss. From our measurements of the resistivity of the alloys in this series the most probable intermetallic compound would be one containing approximately 66 per cent cobalt of the chemical formula $FeCo_2$.

The data which I have presented are those for the compositions of iron, nickel and cobalt. In addition to these alloys, we have studied the effects of adding non-magnetic elements to numerous alloys of

[6] *Transactions of the Faraday Society*, Vol. 8, p. 149.

particular compositions. I cannot at this time go into a detailed discussion of the results with these, but I shall mention briefly some of them. We found that the addition of some of the non-magnetic metals to both the permalloys and the perminvars made those alloys less sensitive to heat treatment. The resistivity was also generally increased. For some compositions the addition in small percentages

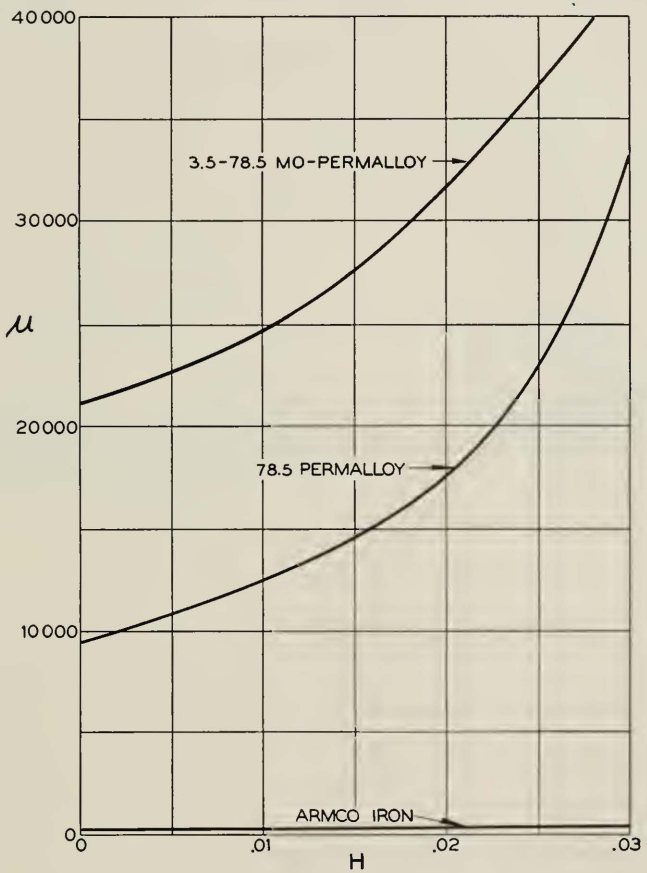


Fig. 20—Permeability curves for permalloys.

of a non-magnetic metal, particularly chromium and molybdenum, increases the permeability at low magnetizing forces. This is illustrated in Fig. 20 where I have plotted the permeability curves for Armco iron, 78.5 permalloy, and a 3.5–78.5 Mo-permalloy. You will note that the initial permeability of the molybdenum-permalloy alloy is 21,000 as compared with 9,000 for the permalloy without

molybdenum. As a rule the maximum permeability was generally decreased when a non-magnetic element was added and so were the saturation values of induction.

As our investigation was undertaken primarily for the purpose of searching for magnetic materials which could be used to advantage in the electrical communication field, it may be of interest to describe a few of the principal uses to which some of these alloys have been put.

Of the three groups of alloys which are of special technical interest, the permalloys are now used extensively in electrical communication circuits. Perhaps its most spectacular use is for continuous loading of submarine telegraph cable.

The term loading is used in the electrical communication art to designate a system of adding inductance to a transmission circuit for the purpose of overcoming the unfavorable transmission characteristics resulting from the electrical capacity of the circuit. This system has



Fig. 21—Sample of submarine loaded cable, showing the loaded conductor.

been used in telephone transmission circuits for over a quarter of a century. The standard method used for telephone circuits, in which inductance coils are placed at equally spaced intervals along the transmission line, was not considered practical for deep sea cables. The only suitable method from a mechanical standpoint was a continuous loading in which a magnetic material is distributed uniformly along the whole length of the cable. Before permalloy was developed the best magnetic material available was iron. It could not be used economically for long submarine cables because of its low permeability. With permalloy having between 40 and 50 times the permeability of iron in the range of magnetic field strength encountered in such cables, it was found that beneficial results could be attained and cables of more than five times the carrying capacity of the old type could be built.

The first permalloy loaded submarine telegraph cable was laid in 1924 between New York and the Azores, a distance of approximately 2,300 nautical miles. A sample of the deep sea section of this cable

is shown in Fig. 21. Numerous other loaded cables have been laid since 1924; the total mileage of loaded submarine telegraph cables is now upwards of 16,000 nautical miles.

Another purpose for which permalloy is now used extensively is for loading coils for telephone transmission circuits. Before the introduction of permalloy, finely powdered, insulated and compressed iron dust was used for the cores of these coils. Permalloy has now replaced iron for loading coil cores, and upwards of a million cores per year are used by the Bell System.[7] For these cores the permalloy is used in the form of compressed insulated powder. Some of the advantages in using permalloy result from its lower hysteresis losses and higher permeability. Taking advantage of these qualities in the design of the coils, it has been possible to reduce materially their

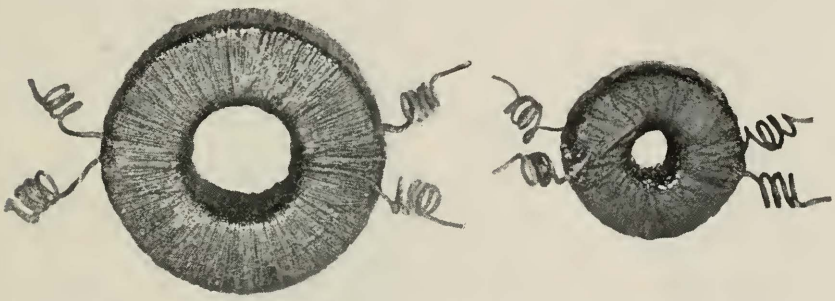


Fig. 22—Compressed powdered core loading coils: left, electrolytic iron core; right, permalloy core.

sizes. This is illustrated in Fig. 22 where two standard loading coils for use in the same circuits are shown. One of these has a permalloy core and the other a core of iron. The former is approximately one-third of the size of the latter.

When these coils are used in service usually a large number are placed in containers which are placed at different points in vaults or on poles along the telephone circuits. In Fig. 23 two such iron cases with their cable connecting stubs are shown; each contains 200 coils. The small case contains the permalloy coils, and the large case the iron core coils. The use of permalloy has reduced the combined weight of the coils and the cases from approximately 1,700 lbs. to about 700 lbs.[8]

[7] The use of permalloy core loading coils is increasing very rapidly. Recent figures show that approximately 2,000,000 of these coils will be required by the Bell System per year during the next few years.

[8] The decreased size of the coils and containers has resulted in a very substantial saving in the cost of loading.

Another use of permalloy in the telephone plant has been found in the case of relay cores. Relays are used to connect and disconnect mechanically the telephone circuits at the central office and at other points. These relays are magnetically operated and their efficiency depends largely on the magnetic quality of the core material. Certain groups of these relays are required to operate under very severe circuit conditions where the operating currents are small and where



Fig. 23—Cast-iron cases each containing 200 loading coils: left, permalloy core coils; right, iron core coils.

the relays are required to distinguish between very small changes in current values. For such circuits, relays with permalloy cores are now used extensively.

The permalloy alloys are also used extensively for cores in high quality audio transformers, in telephone receivers and earphones, and in electrical measuring instruments.

The permalloy series was first developed, and is also the one which

was first used, for commercial purposes on a large scale. The other alloys are still in the commercial development stage, but interesting results have been obtained which make us feel confident that both the perminvars and the iron-cobalt alloys will take their places beside permalloy as important magnetic materials in electrical communication.

A Test for Polarization of Electron Waves by Reflection[1]

By C. J. DAVISSON and L. H. GERMER

A homogeneous beam of electrons is directed at 45° incidence against a {111}-face of a nickel crystal. The beam regularly reflected from this face impinges upon a second similar face at the same incidence angle. A Faraday collector is set to receive electrons regularly reflected from the second crystal, but only such electrons are accepted into the collector as have survived the two reflections without appreciable loss of kinetic energy. The collector and second crystal are rigidly joined, and may be rotated about the axis of the beam proceeding from the first to the second crystal. Measurements of the intensity of the twice reflected beam have been made at bombarding potentials from 10 to 160 volts. Within this range selective reflections (intensity maxima) are observed at 20, 55, 77, 103 and 120 volts.

These five selectively reflected beams have been separately tested for polarization by measuring the current received by the collector as a function of the azimuth of the movable system. If electron waves are polarized by reflection the intensity of the twice reflected beam should be greatest when the planes of incidence of the two reflections coincide, and least when they stand normal to one another. No such variation of the current to the collector is observed within the limits of error of the measurements—about one half of one per cent of the total current. *Our observation is that electron waves are not polarized by reflection.*

THE experiment described in this article was undertaken to determine whether or not a beam of electron waves is polarized by reflection from the surface of a nickel crystal. It is similar in certain respects to the experiment with double Nörrenberg mirrors by which one demonstrates the polarization of light by reflection from glass, and in others to the experiment by which Barkla established that X-rays may be polarized. It resembles most closely, however, the variation of the Barkla experiment performed by Mark and Szilard in which the first of the radiators was a crystal and a Bragg reflection beam proceeded to the second radiator. A homogeneous beam of electrons is directed at 45 degrees incidence against a {111}-face of a nickel crystal, and the beam proceeding in the direction of regular reflection from this crystal is then reflected at the same angle of incidence from a second similar crystal. A double Faraday box is placed to receive electrons which have been regularly reflected from the second crystal, but only such electrons are allowed to enter the collector as have retained all or nearly all of their kinetic energy through the two reflections; those which have lost more than a small fraction of their kinetic energy are excluded by a retarding potential of suitable strength.

The second crystal and the collector are joined rigidly together, and may be rotated about an axis which coincides with the axis of the beam proceeding from the first to the second crystal. It is possible, there-

[1] *Phys. Rev.*, Vol. 33, May, 1929, pp. 760-772

fore, to vary the dihedral angle between the plane of incidence of the second reflection and that of the first. There are two positions of the movable system for which these planes coincide. For these "parallel" positions the current entering the collector should be at a maximum provided the electron beam is unpolarized initially and becomes asymmetric at reflection; for the intermediate "transverse" positions the current should be at a minimum. In the analogous experiment in optics the intensity I of the twice reflected beam satisfies the formula

$$I = I_0(1 + p \cos 2\theta),$$

where θ represents the azimuth angle of the movable system measured from either of its parallel positions, and p an amplitude coefficient which serves as a convenient measure of the polarization effect.

In the experiment with electrons our procedure has been to measure the intensity of the twice reflected beam for various values of θ— though chiefly for the values corresponding to the cardinal positions— to assume the same form of relation between intensity and angle as in optics, and to evaluate the coefficient p.

The reflection of electrons from a crystal surface is, like that of X-rays, "selective in wave-length"; the intensity of the reflected beam attains maximum values at various critical wave-lengths or speeds of bombardment. This effect is, of course, accentuated in a beam which has suffered two reflections. In the test for polarization we have made observations in the range of bombarding potentials from 10 to 200 volts, chiefly at five different electron speeds at which there are intensity maxima of the beam twice reflected.

Preliminary observations indicated that at each of these critical speeds the intensity of the reflected beam is, to a first approximation, independent of angle. The acutal values found for p were some of them positive and some negative, and none greater absolutely than 0.02, which was about the order of uncertainty involved in the determinations of the collector currents. These results were described in a letter to the Editor of "Nature." [2]

In the present article the experiment is described more fully, and additional data are adduced from which it is concluded that the value of p, if different at all from zero, cannot be greater than 0.005.

The principal parts of the apparatus are the gun for supplying a homogeneous beam of electrons, the two crystal reflectors, and the collector. These are contained in two metal boxes or enclosures shown in longitudinal sections in Fig. 1. The right hand or gun enclosure contains the electron gun and the first reflector, and is attached rigidly

[2] C. J. Davisson and L. H. Germer, *Nature*, 122, 809 (1928).

to the framework of the apparatus. The left-hand or collector enclosure contains the second reflector and the collector, and is supported from the frame of the apparatus through bearings by which it can be rotated about a horizontal axis. Communication between the enclosures is through the right-hand bearing which is hollow. The sections of the enclosures at right angles to the plane of the drawing are square.

The electron gun is similar in construction to the one described in an earlier paper [3] to which we refer for the details. The apertures are circular and those which define the beam are 2 mm. in diameter.

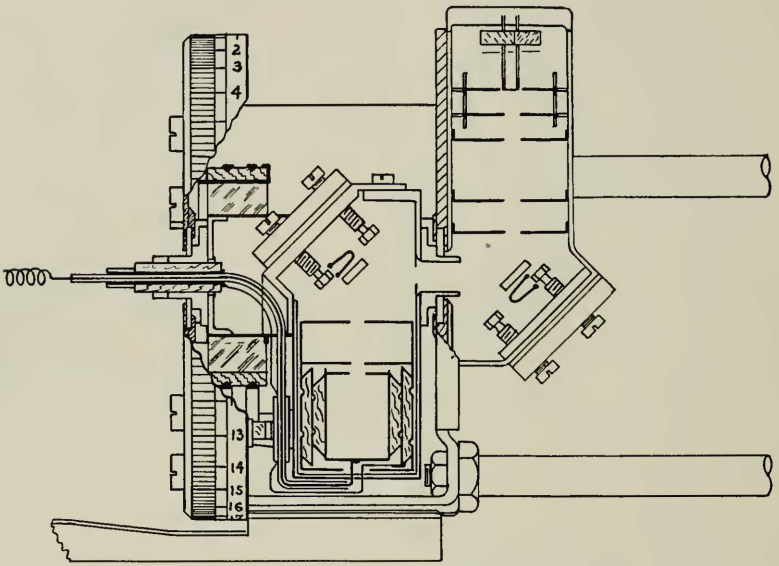


Fig. 1—Cross-section of the experimental apparatus—0.8 actual size.

The reflectors were cut from a single crystal of nickel formed by the slow freezing of pure nickel in vacuum. Their faces, which were polished to fairly good optical flats and then lightly etched by acid, are approximately 4 × 4 mm. in extent. The normals to these faces diverge from one of the {111}-directions of the crystal structure by only about 10 minutes of arc.

The reflectors are so mounted that for each of them the incident beam lies in what we have designated as a {111}-azimuth of the crystal structure, as illustrated in the schematic diagram, Fig. 2. This adjustment may be unimportant, but was made because it has not yet been established that the selectivity of reflection is independent of the azimuth of the incident beam. The {111}-azimuth was chosen rather

[3] C. J. Davisson and L. H. Germer, *Phys. Rev.*, 30, 705 (1927).

than any other because our earlier observations on electron reflection were made with the incident beam in this azimuth, and several of the critical electron speeds for 45 degrees incidence were already known.

Each reflector is attached to a triangular frame which is supported from the diagonal wall of the enclosure through three adjusting screws. Two only of each set are shown in Fig. 1. The frames to which the crystals are attached and other accessory parts have been omitted from the drawing in the interest of clearness.

Small tungsten filaments, mounted one behind each of the reflectors, are supported by stiff wires from quartz plates which are clamped to the outer walls of the enclosures. Electrons emitted by these filaments

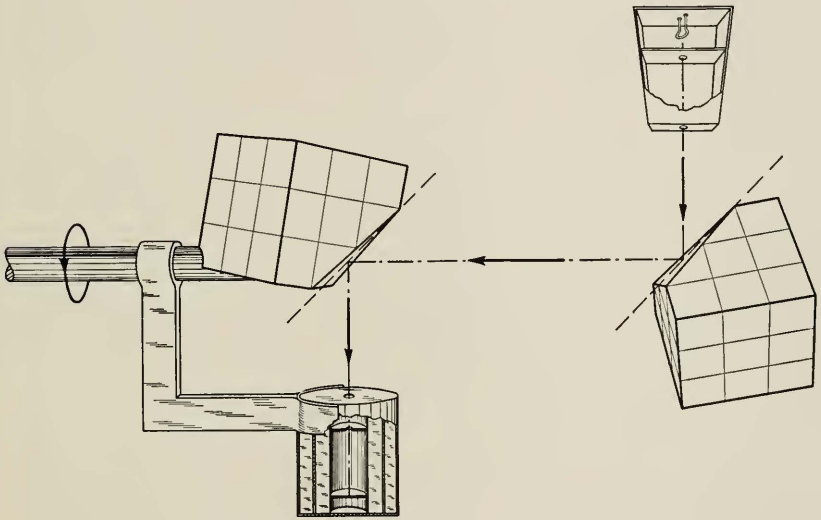


Fig. 2—Schematic diagram illustrating the principle of the experiment.

are used for heating the reflectors by bombardment. The reflectors are not insulated from the enclosures, which in fact contain no insulating material whatever except that incorporated in the gun and the collector.

The metal parts of the collector comprise an inner and an outer box of circular cross-sections and a cylindrical guard electrode of intermediate diameter. These parts are separated by cylinders of pyrex glass, and the assembly constitutes a unit which fits into the end of the collector enclosure. The aperture in the outer box is circular and 2 mm. in diameter; that in the inner box is of the same form but of slightly greater diameter. The guard cylinder is interposed to intercept the leakage current which would flow otherwise from the outer to the inner box. It was anticipated that the electron current entering

the collector would be excessively small and that this leakage current, unless guarded against, might prove an intolerable disturbance.

The lead wire from the inner box is guarded from the frame of the apparatus at all points of support within the tube by electrodes connected with the guard cylinder. This lead wire and the wire from the guard electrodes leave the tube through remote seals as indicated in Fig. 3. The isolation of the latter of these seals was a matter of convenience rather than of precaution.

Four electrical connections are required to parts of the movable system—two to the filament and one each to the collector and to the guard electrodes. These are maintained, with the exception of that to the collector, through platinum tipped molybdenum brushes which bear upon platinum rings. The connection to the collector is through a flexible spiral of tungsten wire lying in the axis of rotation.

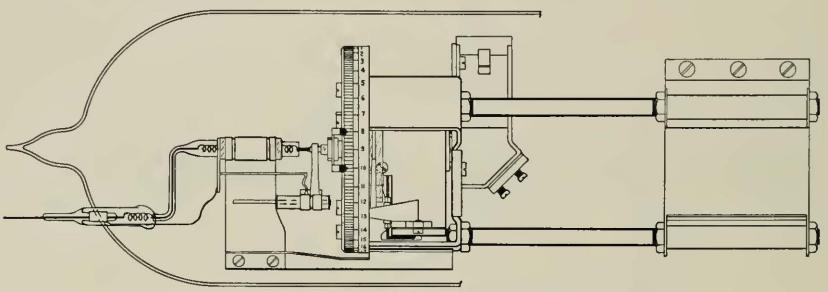


Fig. 3—Outside view of the experimental apparatus.

It may be well to add to this description of the tube a few words in regard to the adjustment of the reflectors to their proper positions and orientations. Each reflector is attached, as has been already mentioned, to a triangular frame which is supported through three adjusting screws from the diagonal wall of the enclosure. By turning these screws the reflector can be rotated through small angles about any axis parallel to the wall, and by the same means its distance from the wall can be varied. These adjustments were sufficient for locating the beam reflected from the first mirror in the axis of rotation, and that reflected from the second in the axis of the collector. They were not sufficient, however, for meeting the further requirement that the incident beam should lie in a {111}-azimuth of the crystal structure. For this adjustment we relied upon orienting the reflector correctly with respect to the triangular frame at the time of its attachment. A mosaic of sharply defined triangular etch pits was visible under the microscope on the surface of the crystal reflector, and it was only necessary to relate these properly to the triangle formed by the frame

to insure the desired orientation of the crystal with respect to the incident beam. Short wires attached to the reflector and protruding from it were rested upon the frame, and the reflector was then turned until the triangles on its surface stood in opposition to the triangle formed by the frame, this being the necessary relation. This adjustment was made with the frame and reflector mounted on the movable stage of a tool maker's microscope. One of the wires was then electrically welded to the frame. The adjustment was disturbed slightly by this operation, but the disturbance was corrected for by bending slightly the attached wire before proceeding with the second weld. This alternation of adjustment and welding was continued until all wires were attached. As finally adjusted the orientation of the reflector may have been wrong by one or two degrees, but hardly by more.

In adjusting the first reflector for position two conditions sought were, first that the intersection of the axis of the gun with the axis of the movable system should lie in the surface of the reflector, and second that the normal to the reflector should bisect the angle formed by these axes. These were attained by removing the collector enclosure from the frame of the apparatus and the filament from the gun, and collimating the collector enclosure bearings with the images of the gun apertures formed by the reflector. For making the similar adjustment of the second reflector an aperture was formed in the center of the rear wall of the collector so that a view of the reflector might be had along the collector axis. The gun enclosure which had been detached from the frame of the apparatus during the adjustment of the second reflector was then replaced, and the adjustment of the two reflectors was checked by directing a beam of light along the axis of the gun and finding that the twice reflected beam proceeded accurately along the axis of the collector.

The preparation of the tube—the preheating of the metal parts, the baking, the exhausting, and the sealing off—was the same essentially as described in an earlier paper to which the reader is referred for particulars. (*Phys. Rev.*, loc. cit.)

In operation, the tube is mounted in a cradle with its axis inclined 30 degrees from the horizontal, so that an auxiliary tube lying in the axis and containing charcoal may be kept submerged in liquid air. The movable system swings to the lowest part of its arc, and its angular position with respect to the frame of the apparatus is read against the circular scale shown in Fig. 3. To alter this azimuth angle θ the tube is rotated about its axis; actually the "movable system" remains at rest relative to the earth, and all other parts are rotated.

No means were provided for measuring the current of electrons incident upon the first crystal. We had found, however, from a preliminary investigation of the characteristics of the gun, that currents of the order 2×10^{-4} amperes could be obtained from it. It was known also from these tests that the electrons ejected from the gun are very nearly homogeneous in speed. Given this value for the current in the primary beam, it was possible from our previous observations on the regular reflection of electrons at 45 degrees incidence to estimate the order of magnitude of the current of full speed electrons which might reach the collector after two such reflections. The estimated magnitudes were from 10^{-12} to 10^{-11} amp, and the currents of selectively reflected electrons actually observed have had values within this range.

In measuring these small currents we have had the use of a direct current vacuum tube amplifier designed and built by Dr. J. M. Eglin. It is the type of amplifier described recently by Wynn-Williams,[4] but embodies certain improvements described by Dr. Eglin at a recent meeting of the American Physical Society.[5] Conditions for observing were best when the amplification factor was about 2,000, so that the currents actually measured were of the order of 10^{-8} amp.

A few preliminary observations were made before heating the crystals to free their surfaces from adsorbed gas. The relation between the current entering the collector and the bombarding potential for a fixed angle θ was quite different in these first tests from that observed after the crystals had been heated. The principal feature of this initial current-voltage relation is a strong maximum at 20 volts. Tests were made for polarization with the crystals in this condition but no evidence of such a phenomenon was obtained.

The current-voltage curve characteristic of reflection from the crystals in a thoroughly cleaned condition is shown as Curve A in Fig. 4. The data from which this curve has been plotted were obtained with the faces of the reflectors parallel to one another as illustrated in Figs. 1 and 2. It will be convenient to designate this position of the movable system as the position $\theta = 0°$. The "parallel" positions are then the positions $\theta = 0°$ and $\theta = 180°$ and the "transverse" positions are those for which $\theta = 90°$ and $\theta = 270°$. A curve similar to Curve A is obtained whatever value is chosen for θ. In this and in all other tests the inner box of the collector was maintained at a potential 2 volts above that of the midpoint of the filament.

Curve B of Fig. 4 exhibits, on a different scale of ordinates, the relation between current and voltage observed for angle of incidence 45

[4] C. E. Wynn-Williams, *Proc. Camb. Phil. Soc.*, 23, 811 (1927).
[5] J. M. Eglin, *Phys. Rev.*, 33, 113 (1929).

degrees in our earlier experiments on the single reflection of electrons incident in the {111}-azimuth. The locations of the maxima of this curve are indicated in a diagram which forms a part of a report of these experiments.[6]

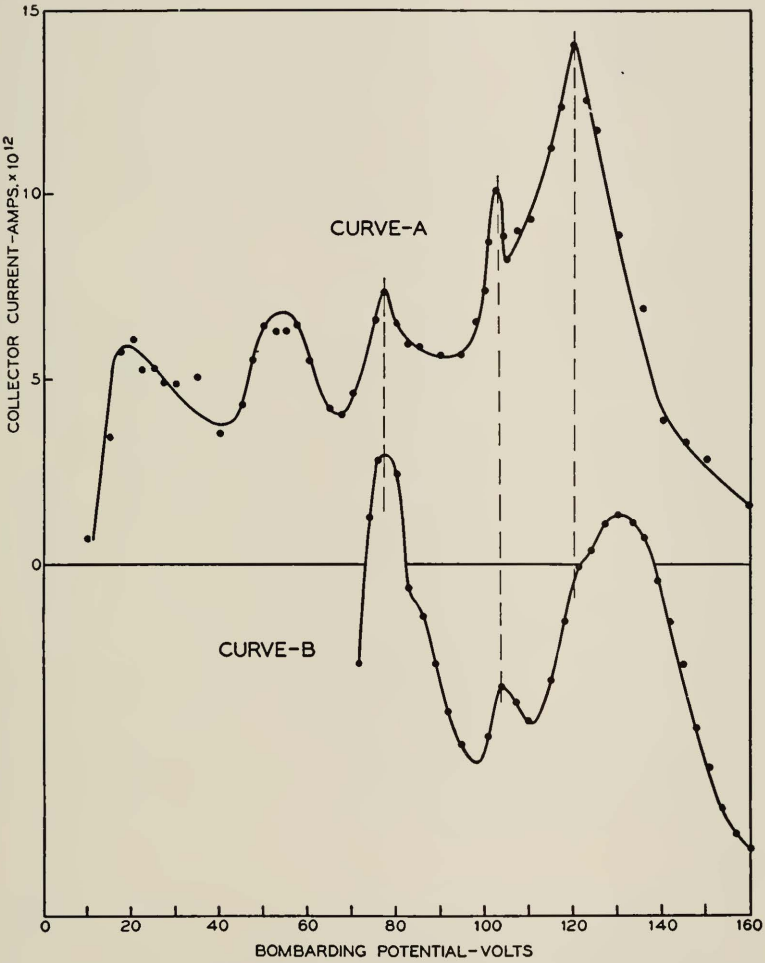


Fig. 4—Variation with bombarding potential of the intensity of beams reflected at 45°. Curve A is the doubly reflected beam of this experiment. Curve B is the singly reflected beam previously reported. (*Proc. Nat. Acad.*, loc. cit.)

The agreement between the curves of Fig. 4 is on the whole satisfactory; each displays three maxima in the voltage range common to both, two of which occur at the same voltages in Curve *A* as in Curve *B*. The voltages at which the third maxima occur—those on the ex-

[6] C. J. Davisson and L. H. Germer, *Proc. Nat. Acad. Sci.*, 14, 619 (1928), Fig. 3.

treme right—differ by about 10 volts. We believe that the position of this maximum is given correctly by Curve A, and that in Curve B it is shifted to the right owing to an eccentricity of the tube used in the earlier experiments. It will be noted that the position of the maximum in Curve A is marked in Curve B by a shoulder which protrudes from the side of the peak. It is with respect to the positions of the maxima only that the curves of Fig. 4 may be legitimately compared, the ordinates in the two cases being proportional to different quantities. Those of Curve A are proportional to the current of full speed electrons entering the collector, while those of Curve B are roughly proportional as has been explained (loc. cit.), to the ratio of full speed electrons, entering the collector to the corresponding current of electrons of all lower speeds.

There is some doubt in our minds as to whether the maximum in Curve A which occurs at 20 volts truly indicates a maximum in the reflecting power of the crystals for electrons of corresponding speed. The current to the collector is determined primarily by the product of the primary current by the square of the coefficient of reflection, so that a maximum in the collector current must correspond to a maximum in the reflecting power if the current in the primary beam is almost or quite independent of voltage, but not otherwise. This condition is known to be reasonably well satisfied in the range of bombarding potentials above 30 or 40 volts. Below this range, however, the current from the gun is limited partly by space charge, and its variation with voltage is rapid. A maximum in the current to the collector in this region must therefore be regarded with a certain suspicion; it may be due to a maximum in the reflecting power of the crystal with which, however, it will fail to coincide in voltage, or it may signify only that the reflecting power has a trend opposite to that of the primary current. We are not, however, greatly concerned in this investigation with the interpretation of this maximum, nor even of the other maxima of Curve A.

Measurements have been made of the intensity of the twice reflected beam as a function of the angle θ for bombarding potentials corresponding to the five maxima of Curve A. In some cases intensities have been measured at intervals of 5 or 10 degrees around the entire circle; but for the most part measurements have been made only at the cardinal positions $\theta = 0$, 90, 180 and 270 degrees. The total number of measurements of this kind is about 500. The complete data for bombarding potential 77 volts, corresponding to the third maximum of Curve A, and for $\theta = 270$ degrees are given in Table I.

TABLE I

Bombarding Potential 77 volts, Azimuth angle $\theta = 270$ degrees. R_0 = zero reading of galvanometer, R = galvanometer reading. $(R - R_0) = D$ = deflection, $\overline{D}$ = arithmetic mean of deflections. $|D - \overline{D}| = S$ = deviation from mean, $\overline{S}$ = mean deviation.

R_0	R	D	S	R_0	R	D	S
33.0 mm.				38.8 mm.			
(33.6)	36.6	3.0	.10	(39.15)	42.1	2.95	.15
34.2				39.5			
(34.55)	37.4	2.85	.25	(39.9)	43.2	3.3	.20
34.9				40.3			
(35.4)	38.6	3.2	.10	40.5			
35.9				(40.55)	43.9	3.35	.25
28.0				40.6			
(27.9)	30.5	2.6	.50	34.4			
27.8				(34.7)	37.8	3.1	.00
26.6				35.0			
(26.75)	29.7	2.95	.15	(35.4)	38.5	3.1	.00
26.9				35.8			
(27.2)	30.3	3.1	.00	(36.6)	39.8	3.2	.10
27.5				37.4			
36.5				(37.8)	40.7	2.9	.20
(36.75)	39.9	3.15	.05	38.2			
37.0				(38.4)	41.9	3.5	.40
(36.5)	40.0	3.5	.40	38.6			
36.0				25.5			
(36.2)	39.4	3.2	.10	(26.0)	28.9	2.9	.20
36.4				26.5			
(36.8)	40.0	3.2	.10	(27.1)	30.1	3.0	.10
37.2				27.7			
(37.3)	40.3	3.0	.10	(28.2)	31.5	3.0	.20
37.4				28.7			
33.6				(29.3)	32.5	3.2	.10
(33.85)	37.2	3.35	.25	29.9			
34.1				29.4			
(34.35)	37.1	2.75	.35	(29.8)	33.0	3.2	.10
34.6				30.2			
(35.15)	37.9	2.75	.35	(31.0)	34.1	3.1	.00
35.7				31.8			
(36.15)	39.5	3.35	.25	(31.85)	34.8	2.95	.15
36.6				31.9			
(36.75)	40.0	3.25	.15	(32.45)	35.5	3.05	.05
36.9				33.0			
37.4				33.8			
(37.8)	40.9	3.1	.00	(34.4)	37.6	3.2	.10
38.2				35.0			
(38.5)	41.6	3.1	.00	(35.45)	38.5	3.05	.05
38.8				35.9			

Number of observations, $N = 36$; $\overline{D} = 3.104$; $\overline{S} = 0.154$

$$\text{Probable error, } \Delta = 0.845 \frac{\overline{S}}{N^{1/2}} = .022 \text{ approx.}$$

$\therefore$ Deflection, $D = 3.104 \pm .022$.

That the deviations from the mean value of the deflections are distributed in this, and in other cases, in close accordance with the normal error function is illustrated by diagrams displayed in Fig. 5.

The value obtained in Table I for the deflection at $\theta = 270°$ is shown again in Table II, together with the values similarly obtained for the same bombarding potential at the other cardinal positions.

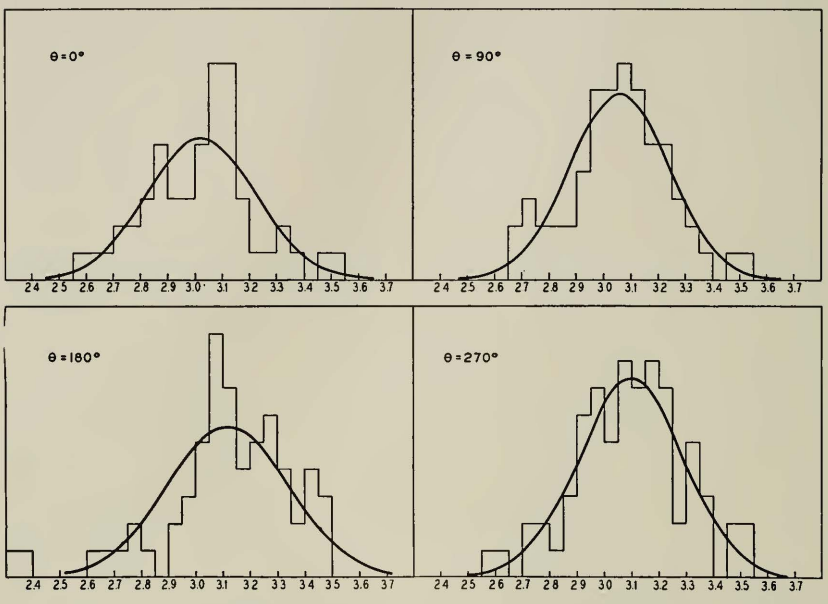


Fig. 5—Plots of all data taken at 77 volts for the four cardinal positions—$\theta = 0°, 90°, 180°, 270°$. The solid curves represent calculated normal error function curves. The data plotted here are summarized in Table II.

TABLE II

Bombarding Potential 77 volts. Wave-length 1.39 A.

Angle θ	No. of Obs.	Deflection
0 deg	26	3.023 ± .026
90	31	3.057 ± .022
180	31	3.116 ± .027
270	36	3.104 ± .022

The values of these deflections and their probable errors have the characteristics of four measurements of one and the same quantity. There are certain values of deflection common to two of the error ranges but none common to three. This is the situation most likely to be met with if we are measuring the same quantity in each case; the maximum number of overlapping ranges should be one half the total number of ranges. It is of some interest, however, to pretend that the different

values found for the deflections correspond to actual differences in the current to the collector, and to attempt to evaluate the amplitude coefficient of the polarization effect. We shall find actually that observations at four positions only are insufficient to determine this constant precisely.

It will be appreciated that differences in the collector current at different angles may arise from mechanical defects in the apparatus—improper alignments, etc.—as well as from polarization, and that in general we shall require for the expression of D as a function of θ a complete Fourier series such as

$$D = D_0 \left[1 + \sum_1^\infty a_n \cos (n\theta + \alpha_n) \right].$$

The data available for evaluating the constants of this series consist of four values of D, corresponding to the four values of $\theta - 0°$, $90°$, $180°$ and $270°$. We designate these four values respectively by D_1, D_2, D_3 and D_4. From the four simultaneous equations formed by writing these pairs of values of θ and D into the series we obtain the following relations:

$$D_1 + D_2 + D_3 + D_4 = 4D_0 \left[1 + \sum_1^\infty a_{4n} \cos \alpha_{4n} \right],$$

$$D_1 - D_3 = 2D_0 \sum_0^\infty a_{2n+1} \cos \alpha_{2n+1},$$

$$D_4 - D_2 = 2D_0 \sum_0^\infty (-1)^n a_{2n+1} \sin \alpha_{2n+1},$$

$$D_1 - D_2 + D_3 - D_4 = 4D_0 \sum_0^\infty a_{4n+2} \cos \alpha_{4n+2}.$$

If we make the definite assumption that all periodic terms of orders greater than the second may be neglected, these reduce to the relations

$$D_1 + D_2 + D_3 + D_4 = 4D_0,$$

$$D_1 - D_3 = 2D_0 a_1 \cos \alpha_1,$$

$$D_4 - D_2 = 2D_0 a_1 \sin \alpha_1,$$

$$D_1 - D_2 + D_3 - D_4 = 4D_0 a_2 \cos \alpha_2,$$

from which we may obtain expressions for a_1, α_1 and $a_2 \cos \alpha_2$, but not, unfortunately, for a_2 and α_2 separately; the fourth observation is used up in fixing D_0 in which we have no interest. Observations at one additional angle would have been sufficient to resolve a_2 and α_2, but this was not appreciated at the time the measurements were made.

If we write Δ_1, Δ_2, etc. for the probable errors involved in the measurements of D_1, D_2, etc. we find on solving for amplitudes and phases, and compounding errors [7] that

$$a_1 = \frac{[(D_1 - D_3)^2 + (D_4 - D_2)^2]^{1/2}}{2D_0} \pm \delta(2 + a_1^2)^{1/2},$$

$$\tan \alpha_1 = \frac{D_4 - D_2}{D_1 - D_3} \pm \frac{(\Delta_1^2 + \Delta_3^2)^{1/2}}{D_1 - D_2}(1 + \tan^2 \alpha_1)^{1/2},$$

$$a_2 \cos \alpha_2 = \frac{D_1 - D_2 + D_3 - D_4}{4D_0} \pm \delta(1 + a_2^2 \cos^2 \alpha_2)^{1/2},$$

where

$$\delta = (\Delta_1^2 + \Delta_2^2 + \Delta_3^2 + \Delta_4^2)^{1/2}/4D_0.$$

Substituting into these formulas the values of D and Δ contained in Table II, we find

$$a_1 = 0.0169 \pm .0080,$$

$$\tan \alpha_1 = -0.50 \pm .45$$

$$(136° < \alpha_1 < 177°),$$

$$a_2 \cos \alpha_2 = -0.0018 \pm .0040.$$

The last of these quantities includes the amplitude of the polarization effect as one of its components. To make this explicit we may restrict a_2 and α_2 to represent the amplitude and phase angle of variations of twice the fundamental frequency due to mechanical imperfections only, and use p to represent the amplitude of the polarization effect. We may then write, since the phase angle associated with p is zero,

$$p + a_2 \cos \alpha_2 = -0.0018 \pm .0040,$$

and from this we wish to infer that p is itself a small quantity, the same in order of magnitude as $(p + a_2 \cos \alpha_2)$.

It may be urged, of course, that nothing in regard to the value of p is to be inferred from the value of $(p + a_2 \cos \alpha_2)$, and this in a strictly mathematical sense is true enough; the individual terms may both be large, and the small value of their sum may be entirely fortuitous. While one must recognize this as a possibility, he must recognize also that the likelihood of the occurrence of chance compensations of such perfection in the case not only of this beam, but in the cases of the others as well, is extremely small. The values found for $(p + a_2 \cos \alpha_2)$ for all five beams have been set down in Table III. It will be seen that,

[7] In calculating the probable errors of these functions we have disregarded the small differences in precision involved in the measurements of the various deflections.

with the possible exception of the value for the 103 volt beam, all are equal sensibly to zero.

TABLE III

Beam No.	Bombarding Potential	No. of Obs.	a_1	α_1	$p + a_2 \cos \alpha_2$
1	20 volts	111	0.013 ± .012	133° to 194°	0.0089 ± .0058
2	55	108	0.015 ± .013	107° to 264°	−0.0025 ± .0065
3	77	124	0.017 ± .008	136° to 177°	−0.0018 ± .0040
4	103	30	0.065 ± .011	113° to 127°	0.0230 ± .0057
5	120	146	0.021 ± .004	102° to 121°	0.0053 ± .0020

The large values found for both $(p + a_2 \cos \alpha_2)$ and a_1 in the case of the 103 volt beam are due, we believe, to some departure from the usual conditions of the experiment which occurred while observations on this beam were being made. Fewer measurements were made on this beam than on any of the others, and the discordant values found for its constants are traceable to an exceptionally low value—based on eight readings only—which was obtained for $D_2(\theta = 90$ degrees). The discordance of this value with those obtained for the other deflections is evident from the figures set down in the last column of Table IV. These are the differences between the various deflections and the mean of the deflections D_1, D_3 and D_4. It will be noted that the departure of the value obtained for D_2 from this mean is three times as great as that for any of the others. The fact that this single unusual departure is responsible for the exceptionally large value not only of $(p + a_2 \cos \alpha_2)$ but of a_1 as well, is reason, we think, for regarding it as accidental. We believe, therefore, that we are justified in disregarding the result obtained in this case, and in concluding from the values found for $(p + a_2 \cos \alpha_2)$ for the other beams that the amplitude of the polarization effect is zero within the limits of uncertainty of our measurements—that is, within about one half of one per cent.

TABLE VI

103 Volt Beam

θ	No. of Readings	Deflections	$D - \frac{D_1 + D_3 + D_4}{3}$
0 deg.....	7	$D_1 = 4.829 \pm .070$	−0.17
90.....	8	$D_2 = 4.481 \pm .052$	−0.52
180.....	9	$D_3 = 5.133 \pm .031$	+0.13
270.....	6	$D_4 = 5.033 \pm .061$	+0.03

. . . . . .

Experiments designed to test for the polarization of electrons by

reflection have been made also by Cox, McIlwraith and Kurrelmeyer, by Joffé, and by Wolf. The experiment by the first-named three [8] is similar in principle and arrangement to our own; the intensity of a beam of electrons which has been twice reflected through 90 degrees is measured while the second reflector and collector are revolved about the direction of incidence of the second reflection. But in other respects the experiments differ. The electrons constituting the primary beam are β-rays from a sample of radium, the reflectors are plates of polycrystalline gold, and the collector is a point-discharge electron counter. The authors report that the shielding between the electron source and the counter was inadequate to suppress entirely an effect due to the gamma radiation, and further that rapid changes in the characteristics of the discharge point made it difficult to obtain consistent data. The results which they publish are ratios of the current received by the collector in one of the "parallel" positions to that received in one or the other of the "transverse" positions, and the ratios of the currents received in the two "transverse" positions. The values found for the first of these ratios depart from unity by much more than the probable error, and show a bias in favor of polarization. The authors do not point this out, however, but lay emphasis instead upon a rather slight departure from unity of the values obtained for the second ratio—that of the currents in the two transverse directions.

The experiment by Joffé is mentioned by Darwin [9] in a short article on the Sixth Congress of Russian Physicists which was held last summer. Darwin remarks that at one of the meetings Joffé reported that he had looked for a polarization of electrons by reflection, but had failed to detect such an effect. So far as we are aware no report of this work has been published.[10]

In the experiment by Wolf [11] a beam of low speed electrons (accelerating potentials of about 10 volts) is deflected in a magnetic field and caused, while still in the field, to impinge at 45 degrees incidence upon a target which in various tests was a plate of brass, a cleft crystal of galena and a crystal of copper. The currents to the target and to an enclosing electrode are measured as the target is revolved about the direction of incidence, and are found to be independent of azimuth. This result is susceptible of two interpretations at least; it may mean that the incident beam is not polarized by the magnetic field, or it may mean that none of the targets serves as an analyser. The latter interpretation, which leaves unanswered the question of polarization in a magnetic field, is consistent with the result which we have obtained.

[8] Cox, McIlwraith & Kurrelmeyer, *Proc. Nat. Acad. Sci.*, 14, 544 (1928).

[9] Darwin, *Nature*, 122, 630 (1928).

[10] A brief account of these experiments has appeared recently in the Comptes Rendus; Joffé and Arsenieva, C. R. *188*, 152 (1929).

[11] Wolf, *Zeit. f. Phys.*, 52, 314 (1928).

The question of the result to be expected from the wave theory of the electron in experiments of the kind here described has recently been considered by Darwin.[12] The conclusion which Darwin reaches is that a beam of electrons initially unpolarized will remain unpolarized after diffraction by a grating provided the forces in the grating responsible for the scattering are electric rather than magnetic, and that therefore any experiment designed to detect polarization by successive reflections from crystals can lead only to a negative result. The result of our experiment is in accord with this prediction.

It is a pleasure to express our best thanks to Mr. G. E. Reitter for the great care with which he constructed the special apparatus used in this experiment, and to Mr. C. J. Calbick for valuable assistance in collecting and reducing the data.

[12] Darwin, *Proc. Roy. Soc.*, 120, 631 (1928).

A Generalization of Heaviside's Expansion Theorem

By W. O. PENNELL

The expansion theorem is one of the most frequently used methods of evaluating operational forms arising from the operational calculus developed by Heaviside. The original theorem, however, is applicable, in general, only to expressions containing integral powers of the operator d/dt. This paper describes an extension to, or a generalization of the original expansion theorem whereby, in general, operational forms with either fractional or integral powers of the operator can be evaluated. A number of operational equivalents are given to be used with the theorem, one of which is the equivalent used by Heaviside. Examples of the application of the theorem to electric circuit problems are shown.

THE well known expansion theorem given by Heaviside in Vol. II of his "Electromagnetic Theory" may be stated as follows:

An operational equation of the form $h = Y(p)/Z(p)$, may under certain well known restrictions on the functions Y and Z, have as its solution

$$h = \frac{Y(0)}{Z(0)} + \sum_n \frac{Y(p_n)}{p_n Z'(p_n)} e^{p_n t}, \qquad n = 1, 2, 3 \cdots. \tag{1}$$

p is the differential operator d/dt, and p_1, p_2 $\cdots$ are the roots of $Z(p) = 0$. $Z'(p_n)$ is the result of substituting p_n for p in $d(Z(p))/dp$. The theorem is true only when no root is zero and all roots are unequal. $Y(p)$ and $Z(p)$ must contain p to positive integral powers only. Various proofs of this theorem have been given and perhaps the simplest depends upon the expansion of $Y(p)/Z(p)$ by partial fractions.

The expansion theorem is valuable in the solution by operational methods, of problems in mathematical physics, and especially electric circuit theory problems.

Generalization of the Expansion Theorem

The generalization of this theorem may be stated as follows: Under certain circumstances it may be possible to write the operational equation

$$h = \frac{Y(p)}{Z(p)} \qquad \text{as} \qquad h = \frac{N(q)}{D(q)},$$

where q is a function of the operator p. With suitable restrictions on the functional forms of N and D the solution of the operational equation is given by

$$h = \frac{N(0)}{D(0)} + \sum_n \frac{N(q_n)}{q_n D'(q_n)} \psi(t, q_n), \tag{2}$$

where $\psi(t, q_n)$ is the equivalent of the operational expression $q/(q - q_n)$

and $q_1, q_2, \cdots$ represent the roots of $D(q) = 0$. If q is the differential operator, that is if

$$q = p = \frac{d}{dt},$$

then as is well known

$$\frac{q}{q - q_n} = \frac{p}{p - p_n} = e^{p_n t}$$

and (2) becomes the Heaviside expression (1).

A proof of the generalized theorem equation (2), is as follows:

By a theorem of partial fractions:

$$\frac{N(q)}{D(q)} = \frac{N(q_1)}{(q - q_1)D'(q_1)} + \frac{N(q_2)}{(q - q_2)D'(q_2)} + \cdots + \frac{N(q_n)}{(q - q_n)D'(q_n)} \tag{3}$$

where $q_1, q_2, \cdots q_n$ are the roots of $D(q) = 0$. The above theorem is true when $D(q)$ and $N(q)$ are rational polynomials and $N(q)$ is of a lower degree than $D(q)$. Further limitations are that no root can be zero and all roots must be unequal.

In writing the above identity in terms of operators it is tacitly assumed that the operators obey the three fundamental laws of algebra, the associative, commutative and distributive laws.

Now

$$\frac{1}{q - q_n} = -\frac{1}{q_n} + \frac{q}{q_n(q - q_n)} \cdot \tag{4}$$

Substituting (4) in (3)

$$\frac{N(q)}{D(q)} = \frac{N(q_1)}{(-q_1)D'(q_1)} + \frac{N(q_2)}{(-q_2)D'(q_2)} + \cdots + \frac{N(q_n)}{(-q_n)D'(q_n)} + \sum_n \frac{N(q_n)}{q_n D'(q_n)} \left(\frac{q}{q - q_n} \right) \tag{5}$$

$$= \frac{N(0)}{D(0)} + \sum_n \frac{N(q_n)}{q_n D'(q_n)} \psi(t, q_n), \tag{6}$$

where

$$\frac{q}{q - q_n} = \psi(t, q_n).$$

The expression fails where $N(0)/D(0)$ is infinite. When the operator

$$q = p = \frac{d}{dt}$$

then

$$\frac{p}{p - p_n} = e^{p_n t}$$

and (6) becomes the Heaviside Expansion theorem.

Although the above proof of (6) is for cases where $D(q)$ is a polynomial, if $D(q)$ is a transcendental function which can be expanded by the process shown, the equation will still hold. It is shown in treatises on trigonometry that tan at, cot at, 1/sin at, 1/cos at, 1/sinh at, 1/cosh at, tanh at, and coth at, all can be expanded in an infinite series of partial fractions which are identical [1] with the expansions obtained by applying the process of equation (3).

Equivalents to be Used in Generalized Theorem

In applying this theorem the following operational equivalents are useful:

Equivalent No. 1:

Let

$$q = p = \frac{d}{dt}.$$

Then

$$\frac{q}{q-a} = \frac{p}{p-a} = e^{at}. \tag{7}$$

This is the equivalent used in the expansion theorem by Heaviside.

Equivalent No. 2:

Let

$$q = p^{1/2} = \left(\frac{d}{dt}\right)^{1/2}.$$

Then

$$\frac{q}{q-a} = \frac{p^{1/2}}{p^{1/2}-a} = e^{a^2t}[1 + \operatorname{erf}(at^{1/2})] \tag{8}$$

where

$$\operatorname{erf}(at^{1/2}) = \frac{2}{\sqrt{\pi}} \int_0^{at^{1/2}} e^{-\lambda^2} d\lambda.$$

Equivalent No. 3:

Let

$$q = p^{1/s} = \left(\frac{d}{dt}\right)^{1/s} \qquad s = \text{a positive integer.}$$

Then

$$\frac{q}{q-a} = \frac{p^{1/s}}{p^{1/s}-a} = e^{a^s t}[1 + \psi_1(t,a) + \psi_2(t,a) + \cdots + \psi_{s-1}(t,a)], \tag{9}$$

where

$$\psi_1(t,a) = \frac{1}{\Gamma(1/s+1)} \int_0^{at^{1/s}} \epsilon^{-\lambda^s} d\lambda,$$

[1] Except in some cases for the first term $Y(0)/Z(0)$.

$$\psi_2(t, a) = \frac{1}{\Gamma(2/s + 1)} \int_0^{a^2 t^{2/s}} e^{-\lambda^{s/2}} d\lambda,$$

. . . . . .

$$\psi_{s-1}(t, a) = \frac{1}{\Gamma\left(\frac{s-1}{s} + 1\right)} \int_0^{a^{s-1} t^{(s-1)/s}} e^{-\lambda^{s/(s-1)}} d\lambda.$$

Equivalent No. 4:

Let

$$q = p^2 = \left(\frac{d}{dt}\right)^2.$$

Then

$$\frac{q}{q - a} = \frac{p^2}{p^2 - a} = \cosh a^{1/2} t. \tag{10}$$

Equivalent No. 5:

Let

$$q = p^3 = \left(\frac{d}{dt}\right)^3.$$

Then

$$\frac{q}{q - a} = \frac{p^3}{p^3 - a} = (1/3)e^{a^{1/3}t} + (2/3)e^{-a^{1/3}t/2} \cos\left(a^{1/3} t \frac{\sqrt{3}}{2}\right). \tag{11}$$

Equivalent No. 6:

Let

$$q = p^4 = \left(\frac{d}{dt}\right)^4.$$

Then

$$\frac{q}{q - a} = \frac{p^4}{p^4 - a} = \tfrac{1}{2} \cosh (a^{1/4} t) + \tfrac{1}{2} \cos (a^{1/4} t). \tag{12}$$

Equivalent No. 7:

Let

$$q = (p + b)^{1/2} = \left(\frac{d}{dt} + b\right)^{1/2}.$$

Then

$$\frac{q}{q - a} = \frac{b}{b - a^2} - \frac{a^2}{b - a^2} e^{(a^2 - b)t} + \frac{a\sqrt{b}}{b - a^2} \operatorname{erf} (\sqrt{bt}) - \frac{a^2}{b - a^2} e^{(a^2 - b)t} \operatorname{erf} (a\sqrt{t}).$$

where $b \neq a^2$

Equivalent No. 8:

Let

$$q = \left(\frac{p}{p + b}\right)^{1/2}.$$

Then

$$\frac{q}{q-a} = \frac{1}{1-a^2}\left[e^{a^2bt/(1-a^2)} + ae^{-(bt/2)} I_0\left(\frac{bt}{2}\right) + \frac{ab}{(1-a^2)^2} e^{a^2bt/(1-a^2)} \int_0^t e^{(a^2+1)bt/2(a^2-1)} I_0\left(\frac{bt}{2}\right) dt \right]$$

where $a^2 \neq 1$

and $I_0\left(\frac{bt}{2}\right) = J_0\left(\frac{ibt}{2}\right)$ = Bessel Function of the first kind.

The above equivalents can be obtained by known operational methods and their derivation will not be given here.

In the application of the generalized theorem to electrical problems, equivalents No. 1, No. 2 and No. 3, especially No. 1 and No. 2, are the ones which will be most frequently used. Equivalents No. 4, No. 5, and No. 6, since they involve only integral powers of p are of use in reducing the labor of applying the original expansion theorem to expressions containing only these powers of the operator p or multiples of these powers. Their use in such cases is illustrated by example No. 3 below.

Equivalent No. 7 enables expressions like the following to be evaluated in closed form.

$$\frac{1}{p + c(p+b)^{1/2} + d}, \qquad \frac{1}{(p+b)^{3/2} + cp + d},$$

$$\frac{(p+b)^{1/2}}{(p+b)^{3/2} + cp + d}, \qquad \frac{1}{\cosh (p+b)^{1/2}}, \quad \text{etc.}$$

In applying equivalents No. 2 and No. 7 some of the following properties of the error function are often conveniently used.

$$\text{erf}\,(-t) = -\,\text{erf}\,(t) \qquad \text{and} \qquad \text{erf}\,(it) = \frac{2i}{\sqrt{\pi}} \int_0^t e^{\lambda^2} d\lambda;$$

also

$$\frac{d}{dt}\,\text{erf}\,[\psi(t)] = \frac{2}{\sqrt{\pi}} e^{-[\psi(t)]^2} \psi'(t)$$

and

$$\frac{d}{dt}\,\text{erf}\,(at^{1/2}) = \frac{e^{-a^2t}}{\sqrt{\pi}}\, at^{-1/2}.$$

The value of erf (t) for different values of t may be obtained from tables of the probability integral as for example Pierce Table of Integrals.

The values of erf (it) for values of t from .01 to 2 are given in a table in London Mathematical Society Vol. 29, 1897–98, page 519. The values of erf $(te^{i\pi/4})$ and erf $(te^{-i\pi/4})$ are given by the following formulæ:

$$\text{erf}\,(te^{i\pi/4}) = \sqrt{2i}[C(t\sqrt{2/\pi}) - iS(t\sqrt{2/\pi})], \tag{14}$$

$$\text{erf}\,(te^{-i\pi/4}) = \sqrt{2i}[-iC(t\sqrt{2/\pi}) + S(t\sqrt{2/\pi})], \tag{15}$$

where

$$C(t\sqrt{2/\pi}) = \int_0^{t\sqrt{2/\pi}} \cos\left(\frac{\pi t^2}{2}\right) dt,$$

and

$$S(t\sqrt{2/\pi}) = \int_0^{t\sqrt{2/\pi}} \sin\left(\frac{\pi t^2}{2}\right) dt,$$

and

$$C(-it\sqrt{2/\pi}) = -iC(t\sqrt{2/\pi})$$

and

$$S(-it\sqrt{2/\pi}) = iS(t\sqrt{2/\pi}).$$

Tables of the values of these two integrals known as the Fresnel Integrals are given in various handbooks such as Jahnke and Emde.

Examples of Application of Theorem

A few applications of the theorem will be given.

Example 1:

The operational solution for the current entering an infinitely long ideal cable with a given impressed voltage of the form Ee^{-at} is

$$K\frac{p^{3/2}}{p+a},$$

K being a constant and

$$p = \frac{d}{dt}.$$

To evaluate $p^{3/2}/(p+a)$ in closed form call $p^{1/2} = q$ then

$$\frac{p^{3/2}}{p+a} = \frac{q^3}{q^2+a}.$$

Since the theorem applies in general only when the degree of the numerator is less than that of the denominator we will write

$$\frac{q^3}{q^2+a} = q - \frac{aq}{q^2+a}$$

and

$$\frac{q}{q^2+a}=\frac{Y(q)}{Z(q)}=\frac{Y(0)}{Z(0)}+\sum_n\frac{Y(q_n)}{q_nZ'(q_n)}\psi(q_n,t),$$

$$\frac{Y(0)}{Z(0)}=0. \qquad \frac{Y(q_n)}{q_nZ'(q_n)}=\frac{q_n}{2q_n{}^2}=\frac{1}{2q_n},$$

$Z(q)=q^2+a$ and the roots of $Z(q)=0$ are $q_n=\pm ia^{1/2}$,

$\psi(q_n,t)=e^{q_n{}^2t}[1+\text{erf } q_nt^{1/2}]$ (see equivalent No. 2).

So

$$\frac{q}{q^2+a}=\frac{1}{2ia^{1/2}}e^{-at}[1+\text{erf }(ia^{1/2}t^{1/2})]-\frac{1}{2ia^{1/2}}e^{-at}[1+\text{erf }(-ia^{1/2}t^{1/2})]$$

$$=\frac{1}{ia^{1/2}}e^{-at}\text{ erf }(ia^{1/2}t^{1/2}).$$

Hence

$$\frac{q^3}{q^2+a}=q-\frac{aq}{q^2+a}=\frac{t^{-1/2}}{\Gamma(1/2)}-\frac{a^{1/2}}{i}e^{-at}\text{ erf }(ia^{1/2}t^{1/2}),$$

since

$$q=p^{1/2}=\left(\frac{d}{dt}\right)^{1/2}=\frac{t^{-1/2}}{\Gamma(1/2)}.$$

Example 2:

The operational expression for the current entering at time t in a cable of distributed resistance R and capacity C with an electromotive force sin ωt impressed is given by

$$I=\sqrt{C/R}\,\frac{\omega p^{3/2}}{p^2+\omega^2}$$

where

$$p=\frac{d}{dt}.$$

Put $q=p^{1/2}$. Then

$$\frac{p^{3/2}}{p^2+\omega^2}=\frac{q^3}{q^4+\omega^2}.$$

Here

$$\frac{Y(0)}{Z(0)}=0, \qquad \frac{Y(q_n)}{q_nZ'(q_n)}=\frac{q_n{}^3}{4q_n{}^4}=\frac{1}{4q_n}.$$

If

$$q^4+\omega^2=0, \quad q_n=\omega^{1/2}e^{i(\pi/4)}, \quad \omega^{1/2}e^{-i(\pi/4)}, \quad -\omega^{1/2}e^{i(\pi/4)}, \quad -\omega^{1/2}e^{-i(\pi/4)}.$$

So

$$\sqrt{C/R}\,\frac{\omega p^{3/2}}{p^2+\omega^2} = \omega\sqrt{C/R}\sum_n \frac{1}{4q_n}\,e^{q_n^2 t}[1+\operatorname{erf}(q_n t^{1/2})]$$

$$= \omega\sqrt{C/R}\left[\frac{1}{4\omega^{1/2}e^{i(\pi/4)}}\,e^{i\omega t}\{1+\operatorname{erf}(\omega^{1/2}e^{i(\pi/4)}t^{1/2})\}\right.$$

$$+\frac{1}{4\omega^{1/2}e^{-i(\pi/4)}}\,e^{-i\omega t}\{1+\operatorname{erf}(\omega^{1/2}e^{-i(\pi/4)}t^{1/2})\}$$

$$-\frac{1}{4\omega^{1/2}e^{i(\pi/4)}}\,e^{i\omega t}\{1+\operatorname{erf}(-\omega^{1/2}e^{i(\pi/4)}t^{1/2},\}$$

$$\left.-\frac{1}{4\omega^{1/2}e^{-i(\pi/4)}}\,e^{-i\omega t}\{1+\operatorname{erf}(-\omega^{1/2}e^{-i(\pi/4)}t^{1/2})\}\right]$$

$$= \frac{\omega^{1/2}}{2}\sqrt{C/R}[e^{i(\omega t-\pi/4)}\operatorname{erf}(\omega^{1/2}e^{i(\pi/4)}t^{1/2})$$

$$+e^{-i(\omega t-\pi/4)}\operatorname{erf}(\omega^{1/2}e^{-i(\pi/4)}t^{1/2})]$$

$$= \sqrt{\frac{2C\omega}{R}}[\sin(\omega t)S(\omega^{1/2}t^{1/2}\sqrt{2/\pi})$$

$$+\cos(\omega t)C(\omega^{1/2}t^{1/2}\sqrt{2/\pi})].$$

The last transformation is obtained by means of formulæ 14 and 15.

Example 3:

Evaluate

$$y = \frac{1}{p^4-3p^2+2}.$$

This can be solved by the expansion theorem in the usual way. A somewhat shorter method is to use the generalized theorem with the operator $q = p^2$. Then

$$y = \frac{1}{q^2-3q+2} = \tfrac{1}{2} + \sum_n \frac{1}{2q_n{}^2-3q_n}\,\psi(q_n,t)$$

$$q_n = 1, 2;\ \psi(q_n,t) = \cosh p_n{}^{1/2}t.$$

See equivalent No. 4. So

$$y = \tfrac{1}{2} + \tfrac{1}{2}\cosh t\sqrt{2} - \cosh t.$$

Example 4:

Evaluate

$$\frac{\sinh bp^{1/2}}{\sinh ap^{1/2}} = \frac{\sinh bq}{\sinh aq} = \frac{b}{a} + \sum \frac{\sinh bq_n}{aq_n\cosh aq_n}\,e^{q_n^2 t}[1+\operatorname{erf}(q_n t^{1/2})],$$

$$-\sinh aq = i\sin iaq.$$

The roots[1] of $\sin iaq = 0$ are

$$q_n = \frac{n\pi}{ia}, \qquad n = \pm 1, \pm 2, \qquad \text{etc.}$$

Substituting these values of q_n in above we get

$$\frac{\sinh bq}{\sinh aq} = \frac{b}{a} + \sum_{\substack{n=\pm 1 \\ n=\pm 2 \\ \text{etc.,}}} \frac{\sinh \frac{n\pi b}{ia}}{\frac{n\pi}{i} \cosh \frac{n\pi}{i}} e^{-(n^2\pi^2/a^2)t} \left[1 + \operatorname{erf} \left(\frac{n\pi}{ia} t^{1/2} \right) \right]$$

$$= \frac{b}{a} + \frac{2}{\pi} \sum_{n=1, 2, 3, \cdots} \frac{(-1)^n \sin \frac{n\pi b}{a}}{n} e^{-(n^2\pi^2/a^2)t}.$$

If $(\sinh bp^{1/2})/(\sinh ap^{1/2})$ is solved by the expansion theorem and the summation is extended over both positive and negative roots, the result is

$$\frac{b}{a} + \frac{4}{\pi} \sum_{n=1, 2, 3, \cdots} (-1)^n \frac{\sin \frac{n\pi b}{a}}{n} e^{-(n^2\pi^2/a^2)t}.$$

In other words the summation quantity is just double what it should be. In order to correct this in practice, those who have used the theorem for such cases have extended the summation only over the positive roots, notwithstanding the fact that in similar cases with integral exponents such as, for example, $1/\cosh ap$ the summation is extended over all the roots. The truth is the original expansion theorem is not applicable if either numerator or denominator contains p to a fractional form. In the above case were the problems to evaluate $(\sinh bp^{2/3}/\sinh ap^{2/3})$ the expansion theorem gives an entirely incorrect answer, while the correct answer is obtained from the extension to the theorem.

Example 5.

$$\frac{p^{1/3}}{p^{2/3} - 1} = \frac{q}{q^2 - 1} \qquad \sum \frac{q_n}{2{q_n}^2} \psi(t, q_n).$$

Here

$$\psi(t, q_n) = e^{q_n^3 t} \left[1 + \frac{1}{\Gamma(4/3)} \int_0^{q_n t^{1/3}} e^{-\lambda^3} d\lambda + \frac{1}{\Gamma(5/3)} \int_0^{q_n^2 t^{2/3}} e^{-\lambda^{3/2}} d\lambda \right],$$

$$q_n = \pm 1,$$

[1] Excluding the root $q_n = 0$ which is not used in this case.

$$\frac{q}{q^2-1} = \tfrac{1}{2}e^t\left[1 + \frac{1}{\Gamma(4/3)}\int_0^{t^{1/3}} e^{-\lambda^3}d\lambda + \frac{1}{\Gamma(5/3)}\int_0^{t^{2/3}} e^{-\lambda^{3/2}}d\lambda\right]$$

$$- \tfrac{1}{2}e^{-t}\left[1 + \frac{1}{\Gamma(4/3)}\int_0^{-t^{1/3}} e^{-\lambda^3}d\lambda + \frac{1}{\Gamma(5/3)}\int_0^{t^{2/3}} e^{-\lambda^{3/2}}d\lambda\right]$$

$$= \sinh t + \frac{\sinh t}{\Gamma(5/3)}\int_0^{t^{2/3}} e^{-\lambda^{3/2}}d\lambda + \tfrac{1}{2}\frac{e^t}{\Gamma(4/3)}\int_0^{t^{1/3}} e^{-\lambda^3}d\lambda$$

$$- \tfrac{1}{2}\frac{e^{-t}}{\Gamma(4/3)}\int_0^{-t^{1/3}} e^{-\lambda^3}d\lambda.$$

Example 6:

If the problem is to evaluate

$$\frac{\sinh b(p+c)^{1/2}}{\sinh a\,(p+c)^{1/2}}$$

Equivalent No. 7 is used. The details will not be worked out since they are quite similar to Example No. 4. The answer is

$$\frac{\sinh b(p+c)^{1/2}}{\sinh a(p+c)^{1/2}} = \frac{b}{a}$$

$$+\frac{2}{\pi}\sum_{n=1,\,2,\,3\,\cdots}(-1)^n\frac{\sin\left(\dfrac{n\pi b}{a}\right)}{n}\left[\frac{c}{c+\dfrac{n^2\pi^2}{a^2}} + \frac{n^2\pi^2}{a^2c+n^2\pi^2}e^{-[(n^2\pi^2/a^2)+c]t}\right].$$

If $c = 0$ the above equivalent reduces to the answer of Example No. 4.

Final Remarks

Operational methods were used by Euler and other mathematicians prior to Heaviside. Their use, however, depended in general upon a formal definition of the operator. Heaviside, on the other hand, adopted a different procedure. In the differential equation of the problem he replaced the operator d/dt by p and obtained the solution of the resulting algebraic equation. He then determined the significance of the operator by the condition that it should give the complete solution of the original differential equation subject to equilibrium boundary condition.

While Heaviside developed the operational calculus in a fairly workable and complete form he failed to correlate it or reconcile it with conventional mathematics or to put its theorems on a rigorous basis. The development since Heaviside's day has been due to a considerable extent to the engineer and mathematical physicist rather than to the pure mathematician.

There are now available a number of methods of evaluating operational forms, among which may be mentioned

The original Heaviside expansion theorem,
Operational Division which gives a series solution,
Contour Integration of the Bromwich-Fourier Integral,
Carson's Integral Equation.

It is thought that this extension to the expansion theorem will be of value as another way of evaluating in closed form certain operational expressions, especially those involving fractional exponents.

In preparing this paper, the author wishes to acknowledge his indebtedness to Mr. R. M. Foster for the contribution of Equivalent No. 7 in its present form and for notes regarding the evaluation of the error function. He is also indebted to Mr. J. R. Carson for reading a draft of the manuscript and for a number of helpful suggestions.

A High Precision Standard of Frequency [1]

By W. A. MARRISON

Synopsis: A new standard of frequency is described in which three 100,000 cycle quartz crystal-controlled oscillators of very high constancy are employed. These are interchecked automatically and continuously with a precision of about one part in one hundred million. They are checked daily in terms of radio time signals by the usual method employing a clock controlled by current maintained at a submultiple of the crystal frequency. Specially shaped crystals are used which have been adjusted to have temperature coefficients less than 0.0001 per cent per degree C.

TO meet the demands for increased precision in measurement and greater reliability of operation a new reference standard frequency system has been developed in the Bell Telephone Laboratories having an absolute accuracy that may be relied upon at all times to better than one part in a million. This reference standard is similar in many respects to one described by J. W. Horton and W. A. Marrison a little over a year ago,[2] but a number of important changes have been made which have contributed to increased accuracy and reliability.

The standard is based on the quartz crystal-controlled oscillator, with a synchronous motor-driven clock, used to determine its rate. It differs from others of the same general type in having a number of similar crystal-controlled oscillators which may be interchanged at will and which are intercompared continuously and automatically with a precision of one part in one hundred million. A number of improvements have been made in the crystal and mounting, and in the circuit, which justify this precision of measurement.

By far the most important element in a crystal-controlled oscillator is the crystal itself and great care was taken in selecting the type to be used in the new standard. A crystal was required as nearly independent as possible of ordinary variations in temperature and pressure and which could be mounted so as to vibrate freely. The effect of temperature appeared to be especially serious as the changes in frequency thus obtained with an ordinary crystal, even with the best commercial thermal regulators available, are greater than are caused by any other single factor in the new standard.

It has been known for some time that plates of quartz cut in the plane of the optic and electric axes usually have positive temperature coefficients and that plates cut in the plane of the optic axis but perpendicular to an electric axis have negative coefficients. It has

[1] Presented before Institute of Radio Engineers, April 3, 1929.

[2] "Precision Determination of Frequency," by J. W. Horton and W. A. Marrison, *Proceedings of Institute of Radio Engineers*, Vol. 16, pp. 137–154, Feb. 1928.

also been known that oscillations may be produced in a crystal either parallel to the impressed electric field or perpendicular to it, the so-called longitudinal and transverse effects. There is a certain amount of mechanical coupling between such different modes of vibration within the crystal, more or less close, depending upon the shape, and in particular depending upon the ratio of dimensions in the principal directions of vibration. In view of these facts it was thought probable that crystals could be produced with such coupling between the modes which have inherently positive and negative coefficients that the resultant temperature coefficient would be nil.

Series of crystals of rectangular and circular shape were made to test this fundamental assumption. It was found that the temperature coefficient does vary with the shape of a resonator and, in particular, that crystals may be proportioned so as to have a coefficient that is practically nil. The relations between the temperature coefficient



Fig. 1—Crystals used in preliminary temperature coefficient tests.

and the dimensions in the case of rectangular plates have been further studied in detail by F. R. Lack [3] of the Bell Telephone Laboratories.

In the first experiment performed with circular discs for this study a large disc was first cut and smaller ones cut from it, after measurement, to insure constancy of material, thickness and orientation with respect to the crystal axes. The parts remaining after three sizes of discs had been cut in this way, with the remainder of the slab from which they were obtained, are shown in Fig. 1. The slab is shown in the partly assembled original crystal in Fig. 2 to show the manner of cutting. With such circular discs it was found that at least one diameter could be found for which the temperature coefficient is very small throughout the entire room temperature range.

Low temperature coefficient crystals obtained in this way are subject to the usual mounting difficulties, namely that the friction on the mounting considerably increases the decrement, and by an amount

[3] "Observations on Modes of Vibrations and Temperature Coefficients of Quartz Plates," by F. R. Lack, presented before the Institute of Radio Engineers, April 3, 1929.

which may vary with time. A form of crystal is desired which can be mounted so that the parts vibrating at relatively large amplitude do not bear heavily on any portion of the mounting.

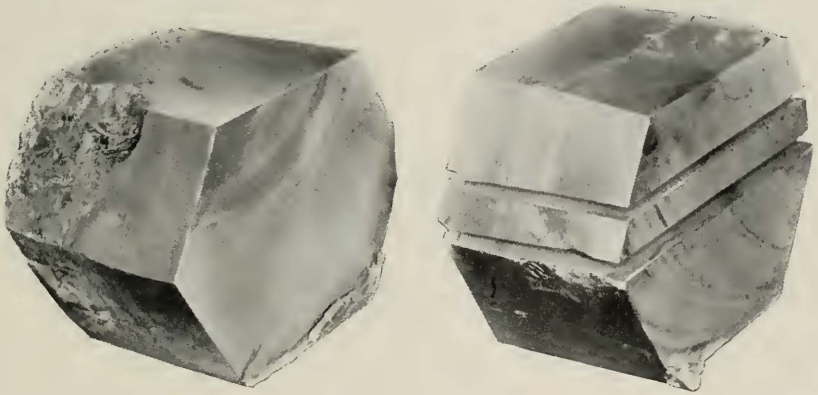


Fig. 2—Partly assembled crystal showing the relation of the slab to the crystal axes.

Further study of temperature coefficients showed that the rings remaining after the small discs had been cut from the larger one,

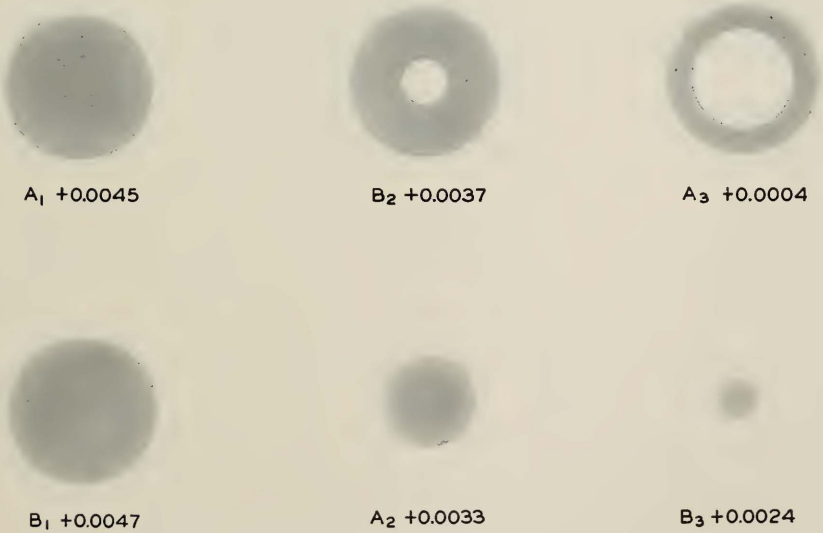


Fig. 3—Temperature coefficients of some discs and rings.

shown in Fig. 1, have a temperature coefficient lower than discs of the same diameter and thickness. This is further illustrated in Fig. 3 which gives the temperature coefficient of two discs and the four

parts remaining after holes of different diameters had been trepanned in them.

It is possible to make ring-shaped crystals having negligible temperature coefficients in a considerable range of frequencies, and, since the ring shape permits of an improved method of mounting in which there is very little friction on the holder, they have been adopted for use in the present standard. Such a crystal having a frequency of 100,000

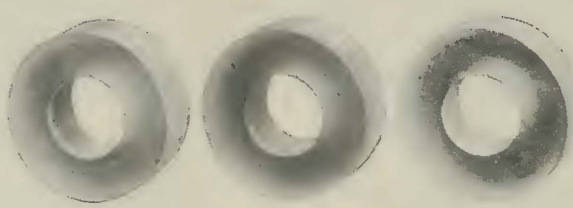


Fig. 4—Three 100,000-cycle low temperature coefficient rings used in frequency standard.

cycles is of substantial size and is reasonably easy to make and adjust. Three of the crystals used in the present standard, adjusted to 100,000 cycles, and having temperature coefficients less than one part in a million per degree C., are shown in Fig. 4.

The variation of frequency with temperature for one of the ring-shaped crystals is given in Fig. 5, showing that it is very small over

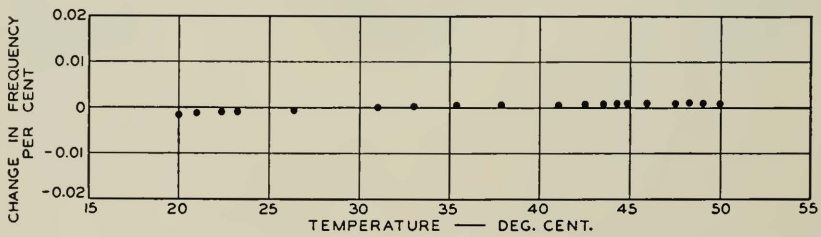


Fig. 5—Variation of frequency with temperature for a 100,000-cycle ring crystal adjusted for low temperature coefficient.

the usual room temperature range. All of the ring-shaped 100,000-cycle crystals made thus far are alike in having a coefficient which is small throughout this range.[4] The temperature coefficient of a disc of the same frequency having the same outside dimensions as the 100,000-cycle rings, is approximately 30 parts in a million per degree C., more than thirty times that of the adjusted crystal.

[4] Where an accuracy of the order of only one part in 100,000 is desired, as in some portable standards, such a crystal could be employed without any form of temperature control.

The manner in which the ring-shaped crystals are mounted in their operating position is shown in Fig. 6. The hole is shaped so that when the crystal hangs on a horizontal cylinder the point of contact is at a theoretical node for mechanical vibration. There is evidence of slight vibration where the central plane intersects the double conical hole but it is small in comparison with that obtained at the outer surface where a crystal is usually supported. The decrement of the crystal when so mounted is considerably less than when it is supported on one of its plane surfaces.

In the mounting the crystal is spaced from the electrodes and is kept approximately central by means of paper spacers on each side. The crystal is free to move laterally in a narrow region but, since it is

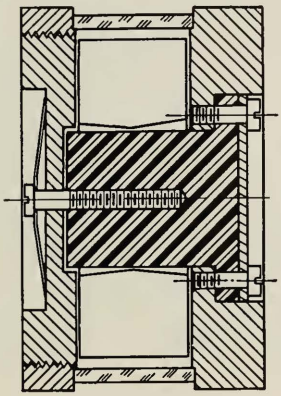


Fig. 6—Section of crystal mounting showing point support.

centrally located, the frequency is at a maximum value and hence a slight motion of the crystal to either side has only a second order effect on the frequency.

The variation of frequency with total electrode spacing is appreciable, but variations due to this factor are avoided by keeping the electrodes accurately spaced by means of a ring of pyrex glass. The temperature coefficient of pyrex is about one quarter of that of crystal quartz perpendicular to the optic axis, so the variation in spacing that is obtained is due almost entirely to the expansion of the crystal. The effect on the frequency due to the differential thermal expansion of the crystal and crystal holder is, however, less than one part in 10^7 per degree C. and so it may be neglected. If it is desired to eliminate this effect entirely a spacer should be used having the same temperature coefficient of expansion as quartz perpendicular to the

optic axis, but for practical purposes a material such as pyrex glass or fused quartz is entirely satisfactory.

The crystal holder is constructed so that a slight variation can be made in the total spacing between electrodes. Since the frequency varies with electrode spacing this can be used for making a slight adjustment of frequency. The thread on the adjustable plate is kept tight by spring tension to prevent the spacing from varying irregularly. A crystal holder and a 100,000-cycle crystal are shown in Fig. 7.

Even though the crystal and its mounting have very low temperature coefficients, it is desirable to control their temperature, for which

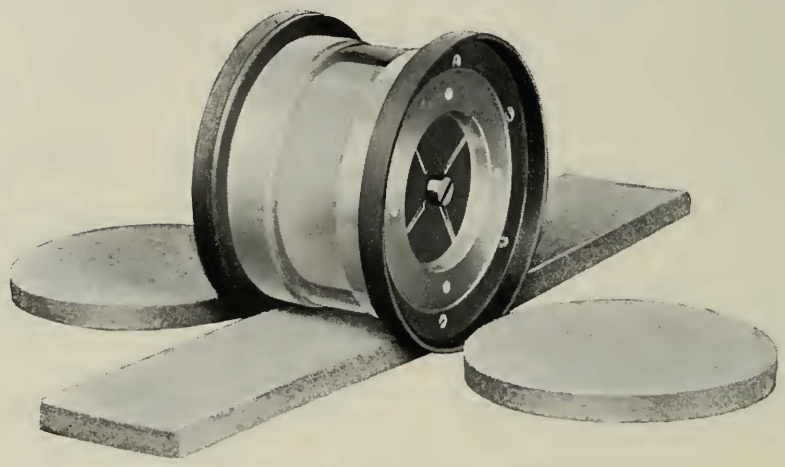


Fig. 7—Crystal mounting with crystal.

purpose the temperature controlling device shown in Fig. 8 has been constructed. It consists of a cylindrical aluminum shell with a wall about one inch thick, with a heater (not shown), and with a temperature responsive element in the wall to control the rate of heating. The aluminum shell has a metal plug that screws into the open end forming a chamber for the crystal which is then completely closed except for a small hole for electrical connections.

Since aluminum is a good thermal conductor the shell equalizes the temperature throughout the chamber and thus avoids the use of a fluid bath. The main heating coil is wound in a single layer over the whole curved surface of the aluminum cylinder, being separated from it only by the necessary electrical insulation. Auxiliary heating coils are wound also on the ends so as to distribute the heating as uniformly as possible. This, in effect, makes the short cylinder behave like a

section from an infinite cylinder. To protect the thermostat from the effect of ambient temperature gradients the heating coil has an outside covering consisting of four layers each of thin felt and sheet

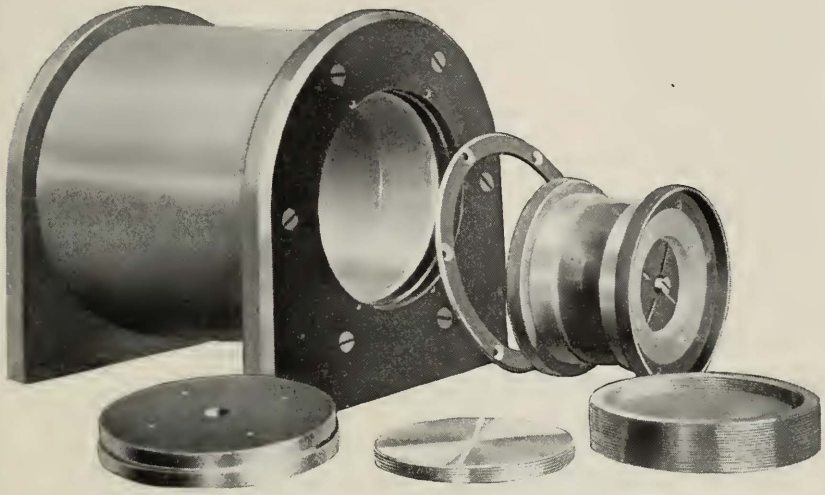


Fig. 8—Temperature control chamber with crystal mounting.

copper spirally wound so that alternate layers are of copper and felt, the innermost layer being of felt and the outer one of copper. This is the covering that appears on the complete device shown in Fig. 9.

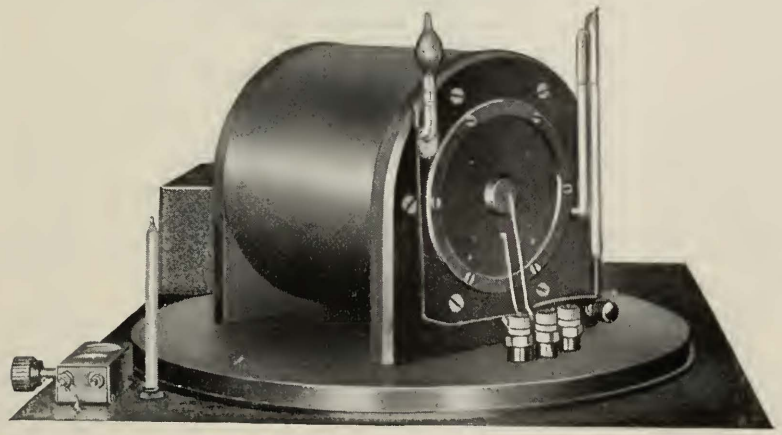


Fig. 9—Complete temperature control unit.

This covering is very effective in reducing surface gradients since the conductivity in directions parallel to, and perpendicular to, the surface differ by a large ratio.

The temperature of the shell rises and falls periodically by about 0.02° C. but even this variation is prevented from reaching the crystal in its mounting by a layer of felt about half a centimeter thick surrounding the crystal holder. At the period of thermostat operation obtained the temperature variations actually reaching the crystal are reduced more than a thousand-fold. The complete temperature controlling device is shown mounted in its operating position in Fig. 9. One of the mounted crystals wrapped in its felt protecting layer is shown in Fig. 10.

To protect the resonator from humidity and pressure variations it is kept under a bell jar at a pressure slightly below atmospheric.

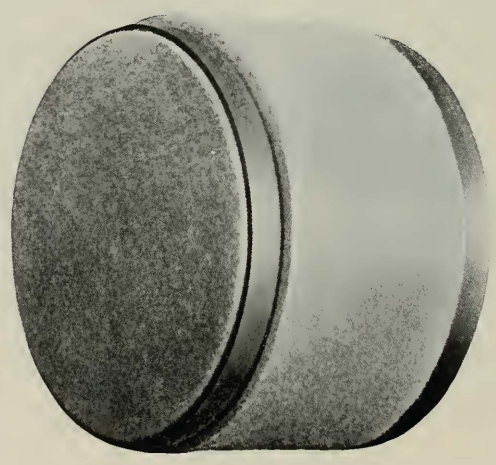


Fig. 10—Crystal mounting with felt insulation.

With the crystals used the frequency varies approximately one part in a million for 10 cm. of mercury change in pressure. It is aimed, therefore, to maintain the pressure constant to about $\pm$ 1 mm. A small mercury gauge within the bell jar indicates the pressure, which may be adjusted by a vacuum pump through a valve in the surface plate. The pressure within the bell jar is affected somewhat by the temperature, and in order to keep it within the required limits it is necessary to maintain a rough control of the temperature within the jar. The pressure gauge does not indicate a change in pressure due to a change in temperature but will indicate any slow leak into the jar that may develop. A thermometer within the bell jar indicates the temperature, from which the change of pressure, and the correction of frequency due to it, may be computed if desired.

Since the frequency varies with the pressure surrounding the crystal an approximate adjustment of the frequency may be made conveniently by an adjustment of pressure.

The circuit of the crystal-controlled oscillator and the first amplifier stages is shown in Fig. 11. The oscillator is of the familiar type in which the crystal electrodes are connected to grid and ground and in which a tuned plate circuit is used. The great advantage in being able to ground one electrode was the major consideration in choosing this circuit. With this circuit, as in the one described a year ago,[2] it has been found possible to choose plate tuning elements such that slight variations in either the inductance or the capacity have little effect upon the frequency. For certain values of inductance and

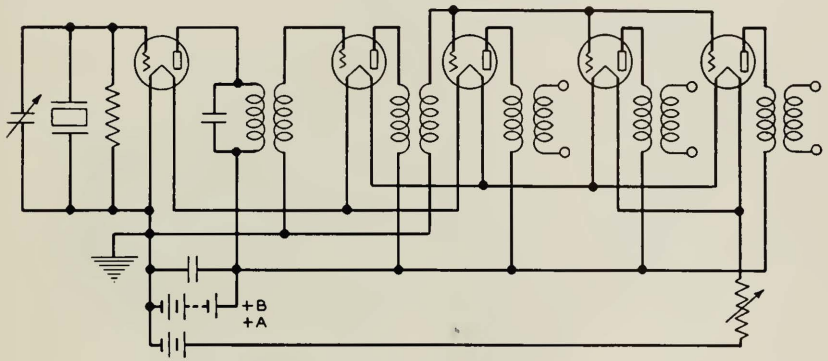


Fig. 11—Circuit of standard frequency oscillator.

capacity the frequency, as a function of their product, takes on a maximum value. The adjustment that gives the maximum value of frequency is used, therefore, so that slight variations, such as those due to temperature coefficient and aging of the tuning elements, will have a negligible effect on the frequency.

The output circuit of the oscillator is very loosely coupled to three independent output amplifiers. This arrangement provides three independent output circuits free from mutual interference and unable to react to an appreciable extent on the crystal oscillator.

The final adjustment of frequency is made with a small cylindrical condenser, having a capacity of about 5 mmf., connected in parallel with the crystal electrodes. The size of this condenser is chosen such that an adjustment of one division on the dial corresponds to a change of frequency of about one part in a hundred million. There are 100 divisions on the dial and a total of 10 turns may be made

corresponding to a total possible adjustment of about one part in 100,000. This condenser is shown at C in the circuit drawing.

The oscillator circuit, showing the tubes and transformers, the plate tuning elements, the filament and plate current meters, and the frequency adjusting condenser is shown in Fig. 12. The adjusting condenser is mounted between the meters and is controlled by the large knob and dial.

One complete oscillator unit, consisting of a 100,000-cycle crystal controlled oscillator with three independent 100,000-cycle outputs,

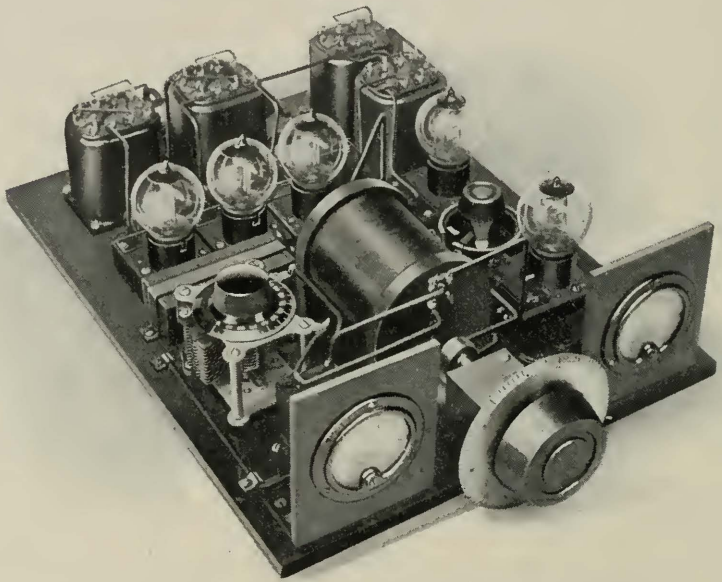


Fig. 12—Standard frequency oscillator without shield.

having a self-contained temperature and pressure controlled crystal, and having a temperature controlled, electrically shielded circuit, is shown in Fig. 13.

The submultiple generator circuit that is used to obtain outputs at 10,000 cycles and 1,000 cycles is shown in Fig. 14. It consists of an inherently unstable vacuum tube oscillator with the tube operating on the curved part of its characteristic. The frequency of this oscillator may be controlled readily by any frequency which is a small multiple or submultiple of it. In this instance the oscillator is controlled by an input having the frequency of its tenth harmonic, the controlling high-frequency input being resistance coupled into the

plate circuit of the lower frequency oscillator. The frequency of the controlled oscillator remains indefinitely at an exact submultiple of the controlling frequency.

Two such circuits are used, one to obtain current at 10,000 cycles and one for 1,000 cycles. Of course, additional amplifier circuits are required in order to supply outputs of considerable magnitude at

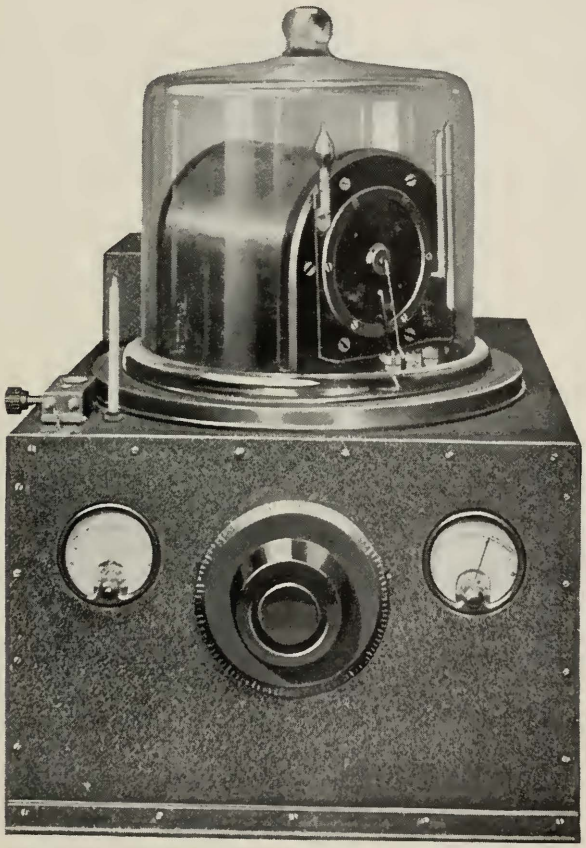


Fig. 13—One complete 100,000-cycle standard frequency unit.

these frequencies, and in such a way that there can be no reaction on the controlling circuits due to load variations or to stray currents at other frequencies fed backward through the output circuits.

A 1,000-cycle motor, operated by current controlled at the 100th submultiple of the standard, drives generators for producing current at 100 cycles and 10 cycles. There are available, therefore, frequencies

in decade steps from 100,000 cycles to 10, all controlled by the 100,000-cycle primary oscillator.

The 1,000-cycle motor is geared to a clock in such a way that, when the controlling frequency has its nominal value exactly, the clock keeps accurate time. In order to check the frequency of the system, therefore, it is only necessary to observe changes in rate of the clock so controlled.

An error of 0.864 second per day in the rate of the clock corresponds to an error in the frequency controlling it of one part in 100,000. It is possible to check the rate of the clock visually with an accuracy of about 0.2 second from audible time signals but obviously this is not sufficiently accurate for our purpose, giving an accuracy of only about one part in 400,000 in a day's observation. In order to facilitate the

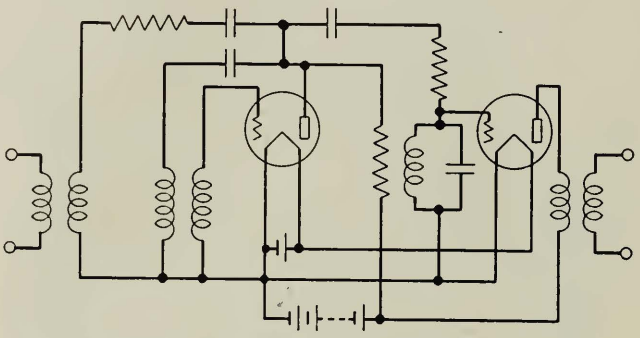


Fig. 14—Circuit of submultiple generator.

comparison with time signals, a contact operated by a cam driven by the 1,000-cycle synchronous motor makes a contact once each second, or to be exact, once for each 100,000 cycles of the primary oscillator. This contact operates one element of a two element recorder while time signals operate the other. Comparisons may thus be made by actual measurements on tape and can be made with greater accuracy than can be judged by eye.

The 1,000-cycle synchronous motor, with its two generators and induction starting motor geared to the clock, is shown in Fig. 15. In this figure the seconds contact mechanism may be seen on the vertical shaft intermediate between the shaft of the motor and the second-hand shaft of the clock.

The assembled rotor of this motor is shown in Fig. 16. The large disc is the 1,000-cycle motor rotor. The disc below it is a hollow steel flywheel filled with mercury used to reduce hunting. The small

rotor below the flywheel is the rotor of the unipolar 10-cycle generator. The disc above the motor rotor is the armature of the 100-cycle generator. The squirrel cage armature of an induction motor for starting is immediately above the 100-cycle generator rotor.

A single constant frequency generator is no longer sufficiently reliable as a standard of frequency of high precision and, as has been the practice where accurate time standards are maintained, three

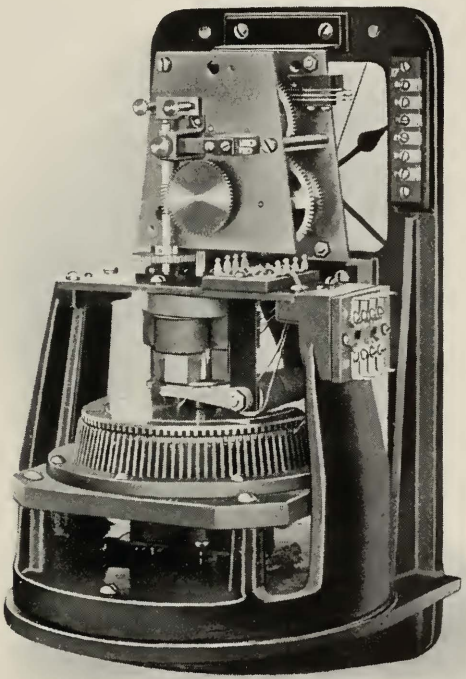


Fig. 15—1,000-cycle synchronous motor, generators, and clock.

similar units have been installed. Means are provided for interchecking them continuously and automatically with the highest precision justified. The use of three such generators with means for interchecking them makes it possible to determine very quickly if and when one generator fails to operate properly.

Only one submultiple generator, clock, and multiple output amplifier is provided, but a special 3-way switch is used by means of which any one of the three primary oscillators may be selected and used as the controlling unit. The oscillators may be interchanged in any order without interrupting the circuits controlled by them.

In the method used for automatic interchecking a fourth oscillator unit is used, identical with the other three except that the frequency is maintained at a slightly different value. The difference between the frequency of this oscillator and that of the other three is kept at about 1 cycle in 10 seconds. The number of beats between the fourth oscillator and each of the other three oscillators is recorded automatically during each 1,000 second interval. The number of beats thus recorded is approximately 100 during each interval.

In 1,000 seconds each oscillator generates approximately one hundred million waves. The numbers that are recorded are, therefore, the number of parts in one hundred million by which oscillator No. 4

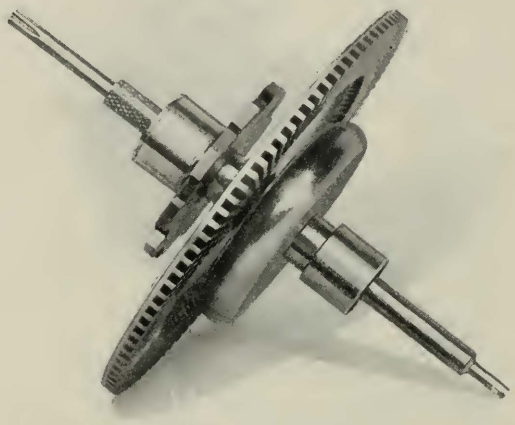


Fig. 16—Rotor of motor-generator.

differs from each of the three primary oscillators during the interval. If the numbers recorded during successive intervals remain the same, the oscillators either did not vary, or they all varied in the same direction by the same number of cycles. If the numbers recorded in successive intervals vary by say 1, 2 or 5, it means that the oscillators have drifted, relative to each other, by so many parts in one hundred million.

Designating the frequencies of the four oscillators by

$$f_1, f_2, f_3, f_4, \tag{1}$$

the three numbers recorded are

$$1{,}000\ (f_4 - f_1),\ 1{,}000\ (f_4 - f_2),\ 1{,}000\ (f_4 - f_3). \tag{2}$$

The mean of these numbers is

$$1{,}000\left(f_4 - \frac{f_1 + f_2 + f_3}{3}\right). \tag{3}$$

If we subtract each of the original three recorded numbers from the mean we obtain

$$1{,}000\left(f_1 - \frac{f_1 + f_2 + f_3}{3}\right) = \delta_1, \tag{4}$$

$$1{,}000\left(f_2 - \frac{f_1 + f_2 + f_3}{3}\right) = \delta_2, \tag{5}$$

$$1{,}000\left(f_3 - \frac{f_1 + f_2 + f_3}{3}\right) = \delta_3. \tag{6}$$

Thus we may compute readily the performance of each of the four oscillators referred to the mean of the three similar primary oscillators. It is obvious that the accuracy of intercomparison of the three similar oscillators does not in any way depend upon the constancy of oscillator No. 4. For convenience in reducing the results, however, it is controlled as carefully as the others.

The records and computed results for approximately ten hours are given in Table 1. During this time the largest relative variation between any two of the four oscillators taken in pairs was 5 parts in 10^8. The random variations between 1,000 second periods appear to be in the order of one or two parts in a hundred million. These random variations are superposed on slow drifts of a quasi-periodic nature probably caused by temperature changes in the circuit and amounting to less than one part in ten million. In addition to these effects a slow, steady drift is expected due to a settling-down of the oscillator circuit and the crystal in its mounting as well as due to aging of the vacuum tubes and even of the crystal itself. The effects of aging can, of course, only be determined after long continued operation.

It is preferable in some cases to refer the performance of each of the four oscillators to the mean performance of all four. This is in the event that all four oscillators are equally reliable in which case the mean of all four makes a better reference standard than the mean of any three. If we designate the numbers

$$1{,}000\ (f_4 - f_1),\ 1{,}000\ (f_4 - f_2), \text{ and } 1{,}000\ (f_4 - f_3)$$

TABLE 1

A Ten-Hour Record Obtained by Means of the Beat Recorder.

The columns δ_1, δ_2 and δ_3 indicate the difference between each oscillator and the mean of the three during each 1,000 second interval expressed in parts in one hundred million.

Serial	Mean	$(f_4 - f_3)$	δ_3	$(f_4 - f_2)$	δ_2	$(f_4 - f_1)$	δ_1
45	86	92	+6	98	+12	69	−17
44	86	92	+6	98	+12	68	−18
43	87	92	+5	99	+12	69	−18
42	87	93	+6	99	+12	70	−17
41	86	92	+6	99	+13	68	−18
40	87	94	+7	98	+11	68	−19
39	86	93	+7	99	+13	67	−19
38	87	94	+7	99	+12	67	−20
37	86	93	+7	99	+13	67	−19
36	87	95	+8	99	+12	66	−21
35	86	93	+7	98	+12	67	−19
34	87	95	+8	100	+13	66	−21
33	87	94	+7	100	+13	66	−21
32	87	95	+8	101	+13	66	−21
31	87	95	+8	101	+14	65	−22
30	87	94	+7	101	+14	66	−21
29	87	94	+7	101	+14	65	−22
28	87	95	+8	101	+14	66	−21
27	87	94	+7	102	+15	65	−22
26	87	95	+8	101	+14	66	−21
25	88	94	+6	103	+15	66	−22
24	88	94	+6	103	+15	68	−20
23	89	96	+7	103	+14	68	−21
22	89	95	+6	103	+14	68	−21
21	89	95	+6	103	+14	68	−21
20	90	95	+5	105	+15	69	−21
19	89	95	+6	103	+14	69	−20
18	89	95	+6	104	+15	68	−21
17	88	94	+6	102	+14	68	−20
16	89	95	+6	104	+15	69	−20
15	89	95	+6	102	+13	69	−20
14	89	95	+6	103	+14	69	−20
13	89	95	+6	102	+13	69	−20
12	89	94	+5	103	+14	70	−19
11	89	95	+6	103	+14	69	−20
10	90	95	+5	103	+13	71	−19
9	89	95	+6	103	+14	70	−19
8	89	95	+6	103	+14	69	−20
7	89	94	+5	103	+14	70	−19

by a, b, and c respectively, it can be shown readily that the numbers

$$a - \frac{a+b+c}{4}, \; b - \frac{a+b+c}{4}, \; c - \frac{a+b+c}{4} \text{ and } -\frac{a+b+c}{4} \qquad (7)$$

represent the difference between each of the oscillators Nos. 1, 2, 3 and 4 respectively and the mean of all four, expressed in parts, in one hundred million. This method treats all four oscillators symmetrically. The symmetry is evident if we substitute in (7) the values (2) assigned to a, b, and c, whence we get:

$$a - \frac{a+b+c}{4} = 1{,}000\left(\frac{f_1+f_2+f_3+f_4}{4} - f_1\right), \qquad (8)$$

$$b - \frac{a+b+c}{4} = 1{,}000\left(\frac{f_1+f_2+f_3+f_4}{4} - f_2\right), \qquad (9)$$

$$c - \frac{a+b+c}{4} = 1{,}000\left(\frac{f_1+f_2+f_3+f_4}{4} - f_3\right), \qquad (10)$$

$$- \frac{a+b+c}{4} = 1{,}000\left(\frac{f_1+f_2+f_3+f_4}{4} - f_4\right). \qquad (11)$$



Fig. 17—Four 100,000-cycle oscillators with auxiliary equipment.

The four oscillators used in the equipment described are shown in Fig. 17. The panels on which the controlling and measuring circuits are mounted are at the right of the picture. The 1,000-cycle motor-driven clock is at the top of the nearest panel. A schematic of the apparatus showing the general arrangement of parts is given in Fig. 18.

The circuit of one element of the modulator for producing low-frequency beats is shown in Fig. 19. The input circuits A and B are supplied from oscillator No. 4 and one of the other three, respectively. The plate circuit includes the windings of a balanced relay

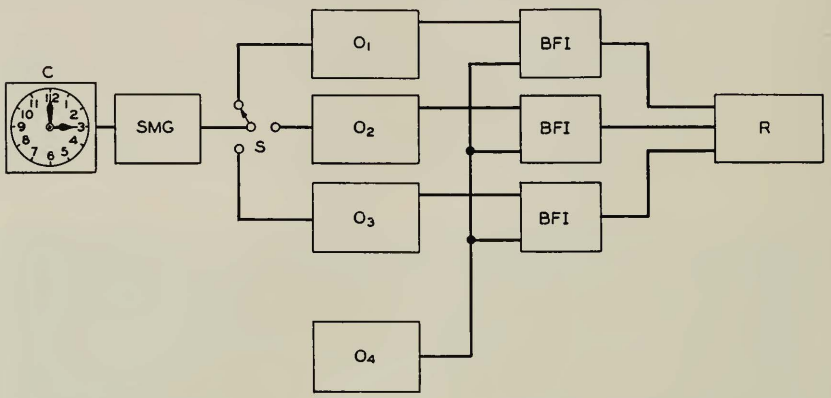


Fig. 18—Schematic of complete frequency standard system.

which makes a contact once for each cycle difference between the input frequencies at A and B and which operates the recording mechanism accordingly.

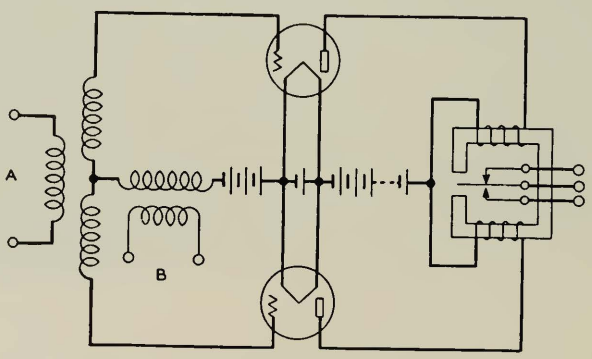


Fig. 19—Balanced modulator of beat frequency indicator.

The beat recorder is a counter arranged to count these relay operations for a definite time interval, in this case for 1,000 seconds, and then to print the total and reset to zero. Five such units are provided which print on a wide strip of paper similar to that used in

an adding machine. Three of the counters are used as outlined above. The fourth counter is to be used for recording the mean of these three numbers, computed automatically by an auxiliary device. The remaining one is a serial counter which is used to record the time, either directly, or by numbering the 1,000 second intervals consecutively.

The five element counter is shown in Fig. 20 with the cover removed to show part of the mechanism. The energy for actuating the counting and resetting mechanism is obtained from a small motor running continuously. These elements are operated at the proper times by clutches controlled by electro-magnets which are selected by the relays in the modulator circuits described above. The counting, printing,

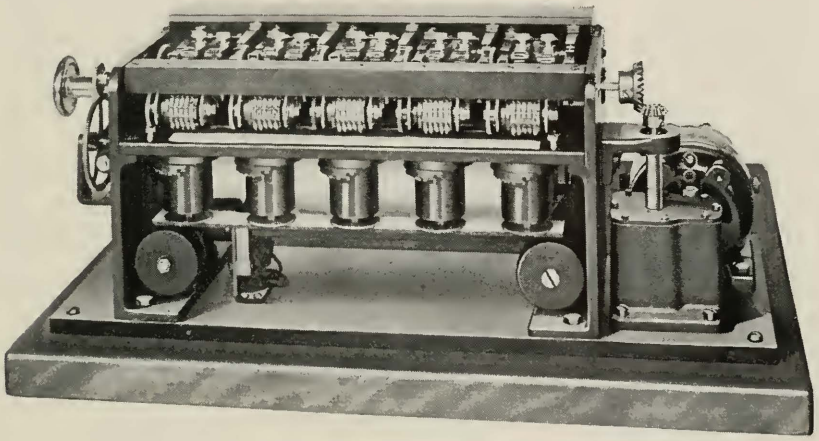


Fig. 20—Automatic beat counter.

and resetting operations are interlocked by means of cams and relays so that no counts may be missed through superposition of operations. The electrical circuit of one unit of the recorder is shown in Fig. 21.

The 1,000 second intervals are determined by a cam operated by the 1,000-cycle synchronous motor. It might be questioned whether one of the crystals being checked should be used to determine the 1,000 second intervals. No serious error arises from this, however, since the percentage variation in the interval due to a change in rate of the crystal is only one millionth of the percentage variation in the recorded beat number. Thus, using one crystal to determine the intervals, used in comparing its rate with other crystals, makes the final measurement subject to an error from this cause of only about 0.0001 per cent.

In order to determine extremely small relative variations in frequency between two oscillators a method of measurement is used in which the duration of individual beats between two oscillators may be measured with an accuracy of about one part in 10,000. This is equivalent to saying that the average beat frequency during each 10 second interval is measured with an accuracy of one part in ten thousand. Since the percentage error in the beat frequency is one million times greater than the percentage error in the primary frequency, to measure the beat frequency with an accuracy of one part

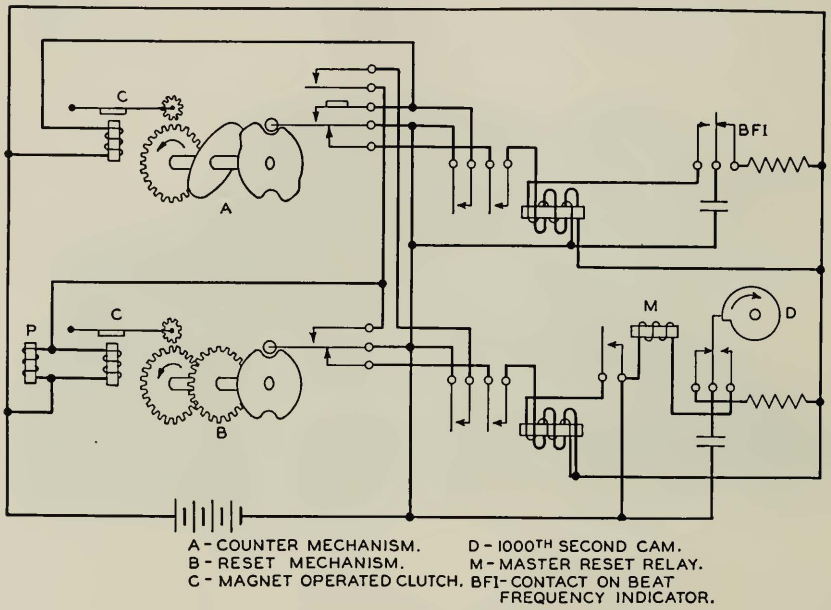


Fig. 21—Electrical circuit of one element of beat counter.

in ten thousand is equivalent to intercomparing the primary frequencies with a precision of one part in ten thousand million. By such a precise method of measurement a great deal can be learned about the nature of the variations that do occur.

The circuit of this device is shown in Fig. 22. Fig. 23 is a photograph of the actual apparatus used in obtaining the data given in Fig. 24. A circular transparent scale S, having 100 numbered divisions, is driven at 10 revolutions per second by a 1,000-cycle synchronous motor controlled by one of the crystals. A modulator operates the balanced relay at the difference frequency between the crystal controlled oscillators. This relay discharges condenser C_1

through the primary of an induction coil, which discharge produces a spark across the gap below the circular scale. The portion of the scale illuminated by the spark is photographed on slowly moving

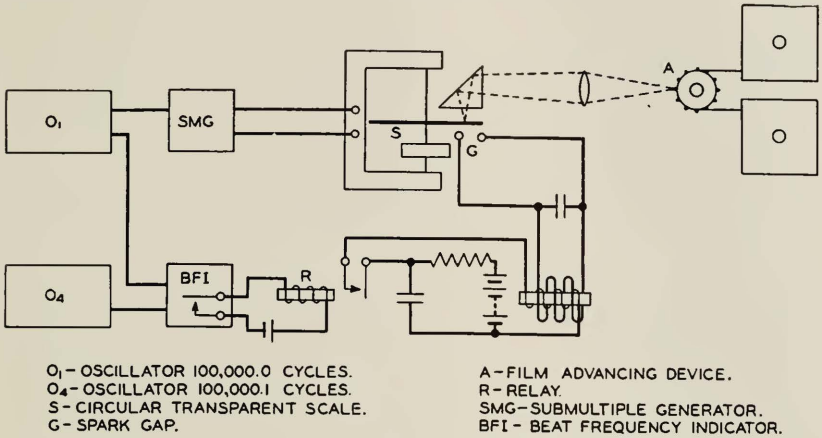


Fig. 22—Circuit of device for determining beat periods accurately.

film at *A*. In this manner, assuming a beat frequency of 0.1 cycle per second, 360 checks per hour may be obtained with practically no supervision.

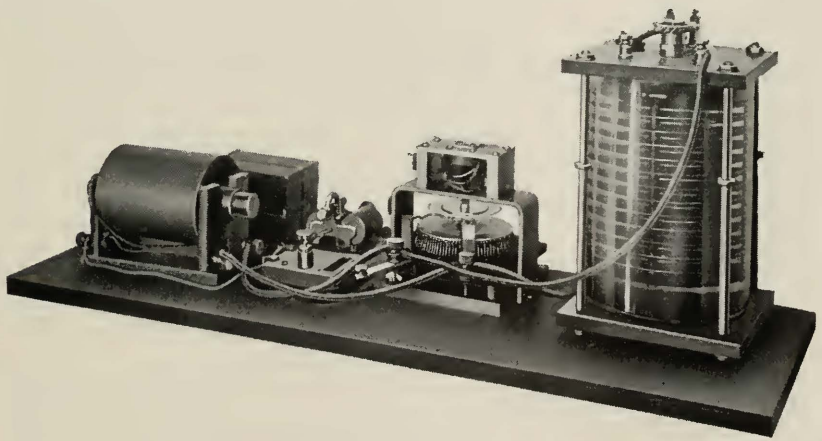


Fig. 23—Apparatus used for measuring beat periods.

If the sparks occur at exactly an even number of revolutions of the motor apart, the same portion of the scale will be photographed by each spark. If the intervals differ from such a value by 0.001 second, the successive photographic images will differ by one scale division.

Thus, the length of the beat periods may be read directly from the photographic records with an accuracy of 0.001 second, which determines the length of the 10 second periods with an accuracy of one part in ten thousand. Thus a variation of one division on the photographic record corresponds to a relative variation of one part in ten billion between the two frequencies, compared during an interval only ten seconds long. The whole number of revolutions of the scale may be determined readily by auxiliary means.

The two graphs in Fig. 24 show typical variations between two crystal oscillators operating under rather unfavorable conditions. One crystal was not in its sealed bell jar and one circuit was only

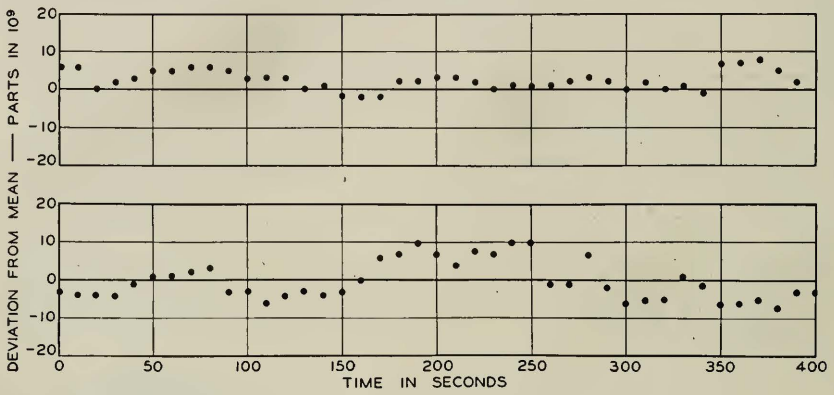


Fig. 24—Relative rates of two pairs of crystal oscillators showing small random variations in frequency.

partially shielded and was exposed to draughts of varying temperature. Even under these conditions, however, the variations from the mean did not exceed one part in 10^8 during the test.

The checking methods just described are intended primarily to indicate more or less rapid changes in frequency. Slow changes, and, of course, the absolute rate, can best be determined in terms of standard time. This is done, as previously indicated, by checking the rate of a clock controlled by one of the crystals against radio time signals. Unfortunately no long checks have been obtained as yet in this way, but several tests made over periods of a few days indicate a constancy of rate in the order of 0.01 second a day.

The measurements made so far indicate that the frequency of a crystal controlled oscillator such as described when suitably controlled, may be expected to be constant to at least one part in 10^7 over periods of seconds or over periods of days. It is hoped that it will be possible in the near future to present accumulated data on the performance of the frequency standard system described.

Observations on Modes of Vibration and Temperature Coefficients of Quartz Crystal Plates [1]

By F. R. LACK

The characteristics of piezo-electric quartz crystal plates of the perpendicular or Curie cut are compared with parallel or 30-degree cut plates with reference to the type of vibration of the most active modes, the frequency of these modes as a function of the dimensions, and the magnitude and sign of the temperature coefficients of these frequencies.

It is pointed out that the two principal modes of the perpendicular cut plate appear to be of the longitudinal type, the high-frequency mode being a function of the thickness while the low-frequency mode is a function of the width (along the electric axis). Both modes have a negative temperature coefficient of frequency. Of the two corresponding modes of the parallel cut plates a shear vibration is responsible for the high frequency. This frequency has a positive temperature coefficient. The low-frequency mode is of the longitudinal type and has a negative temperature coefficient.

Considering only the high-frequency vibration of these plates it is observed that there are characteristic variations of the frequency and temperature coefficient with the ratio of dimensions of the plate and the temperature, which are peculiar to the parallel cut plate. These variations can be attributed to a coupling of the shear and longitudinal modes.

It is then shown that if the parallel cut plate be treated as a group of coupled oscillatory systems with appropriate temperature coefficients the usual coupled system analysis will explain the curves of frequency *vs.* dimensional ratio, frequency *vs.* temperature, and temperature coefficient *vs.* dimensional ratio that are characteristic of this plate. This analysis offers an explanation of the low temperature coefficients which can be produced by a proper choice of the dimensional ratios.

WITH the increasing demands of the radio industry for a high degree of carrier-frequency stability, considerable attention has been focused recently on the piezo-electric quartz crystal as a circuit element in frequency generating systems. The low damping of these mechanical oscillators, combined with their piezo-electric properties makes them particularly suitable for frequency control where a high degree of constancy is required. The frequency stability of the quartz plates prepared in the usual manner, is however, often not sufficient for many of the demands for constant frequency. For instance such a crystal plate does not compare favorably as a substandard of frequency with a good astronomical clock. To meet the demands for frequency substandards as well as many other practical problems concerning frequency generation in the communication art, it becomes necessary to devise methods for improving the frequency stability of these crystal systems. This involves a study of the many factors upon which this stability depends.

A crystal plate constitutes an extremely complex vibration system with a large number of degrees of freedom which are for the most

[1] Presented April 3, 1929, before Institute of Radio Engineers.

part combinations of certain fundamental types of vibration. The ultimate frequency stability attained with a given crystal-controlled frequency generator is then a function of the equivalent electrical characteristics of the combination vibration set up in the crystal plate as well as the constants of the rest of the generator circuit. In particular, the frequency change in a crystal oscillator with changes in tube constants or attached load is a function of the equivalent electrical decrement of the vibration which the crystal happens to be executing. Further, the temperature coefficient of frequency of the crystal oscillator depends largely upon the temperature coefficient of frequency of the crystal vibration, which in turn depends upon the change with temperature of the various mechanical elastic constants that are called into play by this vibration.

The general relation between stress and strain, which in an ordinary isotropic medium involves only two constants, in crystal quartz requires six.[2] The choice of a particular constant or constants that enter into a given mode of vibration depends upon the orientation of the plate with respect to the original crystal axes, and the particular type of vibration, whether longitudinal, torsional, etc.

It is to be expected, therefore, that there will be a variation among the characteristics of the modes of vibration of plates cut in a different fashion, as well as between the different modes of a given plate. In practice we have found considerable difference in the magnitude of the electric and electrothermal constants, between the various modes of vibration of a given crystal plate, even when the vibration frequencies are within a few hundred cycles of each other.

To secure uniformity of results with respect to frequency stability it becomes necessary, therefore, to study the various possible modes of vibration of these crystal plates in detail, and set up certain criteria by which it will be possible to produce plates that will vibrate in a definite mode whose characteristics are known.

The theoretical aspects of this problem offer considerable difficulty, for it will be remembered that the classical case of the vibrations of an isotropic plate whose edges are free has as yet only been solved approximately,[3] and with the extension of the theory made necessary by the crystalline nature of quartz, the complexity of the problem is considerably increased, with the possibility of a complete solution very remote.

Using long rods or bars of crystal, instead of plates, other investi-

[2] Voigt's "Kristallphysik," pp. 749–755, or Love's "Mathematical Theory of Elasticity," Chap. VI.

[3] Rayleigh, "Theory of Sound," Chap. X and X_A.

gators [4] have been able to set up the three types of vibration (longitudinal, flexural, and torsional) common to isotropic bars. Moreover, the formulæ for these vibrations in isotropic material can be used to determine the frequency of the quartz rods to a good first approximation.

Returning to the problem of the plate, if the experimentally determined facts concerning plates of certain definite orientations are examined, it will be seen that they suggest the treatment of the plate as a special case of a bar. A résumé of these facts will illustrate this point and at the same time indicate the effect of orientation on the character of the modes of vibration.

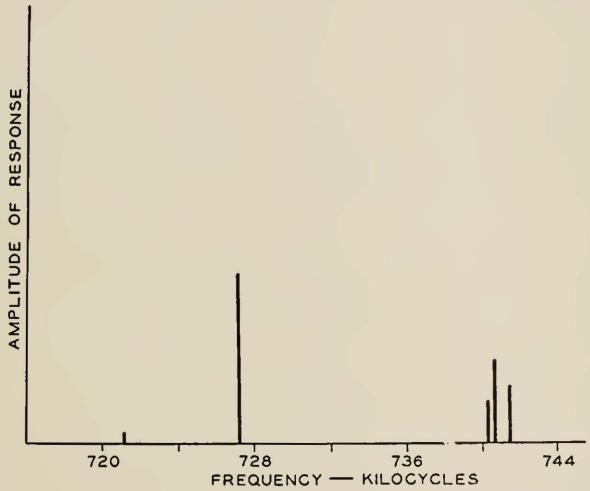


Fig. 1—Response frequencies of 32 x 47 x 2.760 mm. parallel cut crystal plate in the region of the major high frequency.

In general, a quartz crystal plate cut with any orientation with respect to the crystal axes will respond to a large number of frequencies. A plot of these frequencies showing their spacing and the relative magnitudes of response [5] may be termed the frequency spectrum of the plate. Fig. 1 shows part of the high-frequency region of such a spectrum. In these frequency spectra there are usually one or more frequencies at which the crystal will react with sufficient voltage to drive a vacuum tube in the usual crystal oscillator circuit.

[4] Cady, *Proc. I.R.E.* 10, p. 83, 1922. Harrison, *Proc. I.R.E.* 15, p. 1040, 1927 Giebe, *ZS. f. Phys.* 46, p. 607, 1928.

[5] The amplitude of response in this case is the maximum amplitude of current through the crystal at constant voltage, which in turn is a measure of the equivalent series resonant impedance of the crystal system.

The relation between these major response frequencies and the dimensions of the plate for the two principal orientations can be outlined as follows:

Curie or Perpendicular Cut

When the crystal plate is so cut that its major surfaces are parallel to the optic axis and perpendicular to an electric axis (the Curie or perpendicular cut, see Fig. 2) there are two major response frequencies, one high and one low.[6] The high frequency is a function of the

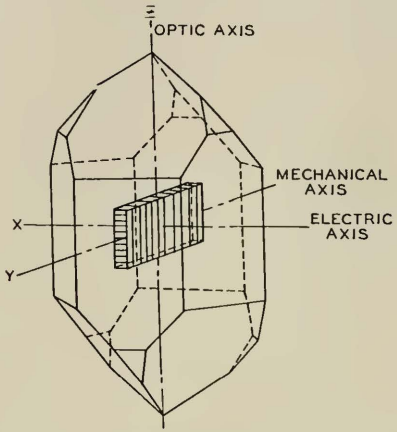


Fig. 2—Orientation of a perpendicular or Curie cut plate with respect to the crystal axes.

thickness of the plate and to a good approximation is given by the expression

$$f = \frac{K}{t}, \tag{1}$$

where t is the thickness in millimeters and $K = 2.860 \times 10^6$. If the plate could be considered as a bar of length t then the frequency of a simple longitudinal vibration would be given by the expression

$$f = \frac{1}{2t}\sqrt{\frac{E_{xy}}{d}}, \tag{2}$$

where E_{xy} is Young's modulus in the X-Y plane and d is the density. If the numerical values[7] of E_{xy} and d are substituted in the above

[6] For this discussion the low-frequency flexural vibration of the type described by Harrison will not be considered.

[7] For numerical values of the elastic constants and the density of quartz, see Sossman, "The Properties of Silica," the American Chemical Society Monograph Series.

expression it is found that the same value for K is obtained as that of equation (1).

The low frequency is a function of the width, the dimension parallel to the Y axis, and is given by the same expression as equation (1) with the same value of K, the width in millimeters being substituted for the thickness.

For this type of crystal plate there are then two possible major modes which appear to be of the longitudinal type and depend upon the same elastic constant. (Young's modulus in the X-Y or equatorial plane has the same magnitude in any direction.)

The temperature coefficient of both these frequencies is negative, which is in agreement with the temperature coefficient of Young's modulus for the equatorial plane.[8]

The Parallel or 30-degree Cut

When the crystal plate is so cut that its major surfaces are parallel to both the optic and electric axes (the parallel or 30-degree cut, see Fig. 3) this 30-degree shift in orientation from the perpendicular

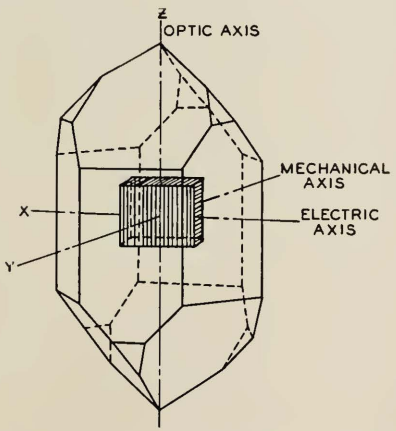


Fig. 3—Orientation of a parallel or 30-degree cut plate with respect to the crystal axes.

changes the characteristics in some important respects. As before there is a high and a low principal frequency, but in this case the high frequency sometimes occurs as a doublet (two response frequencies a kilocycle or so apart).

For thin plates of large area the high frequency is a function of the thickness of the plate and is given by the approximate expression

[8] Perrier & Mandrot, *Mem. Soc. Vaudoise Sci. Nat.* (1923), 1, pp. 333–364.

$$f = \frac{K}{t}, \tag{3}$$

where t is the thickness in millimeters and K is now 1.96×10^6. It will be noted that this constant differs from that found in the case of the perpendicular cut crystal. Moreover the temperature coefficient of this frequency is positive.

These facts lead one to believe that this is not a simple longitudinal vibration. Cady [9] has pointed out that if it be considered as a shear vibration in the X-Y plane the frequency can be calculated using the appropriate shear modulus.[10]

The low frequency is a function of the width, the dimension parallel to the electric or X axis, and is given by the same expression and constant as the frequencies of the perpendicular cut plate. It has the same characteristic negative temperature coefficient.

For these parallel cut plates there are then two possible major modes which, however, differ in type of vibration and sign of temperature coefficient.

Limiting this discussion to the high-frequency region, it is seen that these parallel and perpendicular cut plates have different frequency-thickness constants and temperature coefficients of opposite sign. On closer examination it is found that there is an additional difference which involves the variation of the magnitudes of these frequency-thickness constants and temperature coefficients with the ratio of width to thickness of the plate.

For the perpendicular cut plate the frequency-thickness constant changes but little with the size of the crystal. The same is true for the temperature coefficient, and from recent measurements on a number of sizes of plates the magnitude of this coefficient lies between minus 20 and minus 35 cycles in a million per degree centigrade.

The parallel cut plate, on the other hand, has a frequency-thickness constant which for any but thin plates of large area varies considerably with the width. The temperature coefficient also varies with the width, and is in addition a function of the temperature. This coefficient has a wide range of values whose limits are approximately plus 100 cycles in a million per degree centigrade and minus 20 cycles in some special instances, with all possible intermediate values including zero. Then, as has been mentioned before, these parallel cut

[9] Cady, *Phys. Rev.*, 29, p. 617, 1927.

[10] If it could be shown that the shear modulus of this plane had a positive temperature coefficient it would substantiate this assumption, but there is no information at present available regarding the effect of temperature on the elastic constants other than for the two values of Young's modulus.

crystals frequently have two high-frequency modes of vibration within a kilocycle or so of each other and will start on either of these modes if the circuit constants are changed slightly. These two modes usually have widely different characteristics.

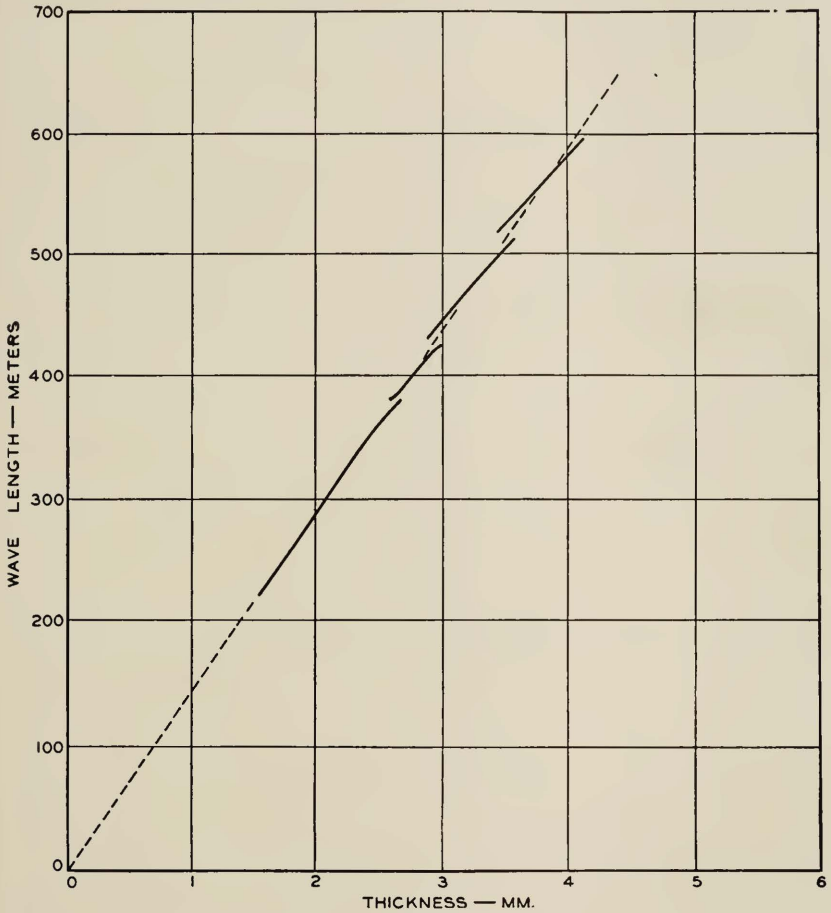


Fig. 4—The wave-length at which a 2.5 cm. square parallel cut crystal plate will operate in an oscillator circuit as its thickness is progressively reduced.

Apart from this seemingly erratic variation it has been the experience of this laboratory that the parallel cut crystal will oscillate more readily in the Pierce type of oscillator circuit. For this reason this type of crystal has been used for a number of purposes and these observed variations have been the object of considerable study.

As a result of this work an explanation has been evolved to account

for these variations. This explanation not only suggests reasons for the above mentioned phenomena, but, what is more important, it indicates the procedure by which it is actually possible to produce crystals having negligible temperature coefficients. Before outlining this theory the experimental facts which served as its foundation will be discussed in detail.

Frequency-Thickness Constant as a Function of Dimensions

When work on the production of parallel cut crystals in the broadcast frequency band was first started, it was found that it was very difficult to grind crystals for certain low frequencies using a 2.5 cm. square plate because of discrete jumps in frequency for a small reduction of thickness. Fig. 4 is a typical curve showing the wave-length [11] as a function of the thickness for a 2.5 cm. square crystal. This curve should be a straight line (for from equation (3) it is evident that $\lambda = K't$) but it will be noted that there are certain discontinuities

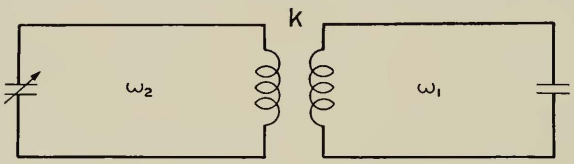


Fig. 5.

at the upper end. It was found that these discontinuities were present at frequencies that could be identified with harmonics of the frequency the crystal would have if it were vibrating in the direction of its length along the electric axis.

This was the first definite indication obtained in the Bell Telephone Laboratories that the longitudinal vibration of the crystal in the direction transverse to the applied field could affect the frequency supposed to depend only on the thickness. It was checked by further work on crystals of other dimensions, and in each case the position of these discontinuities was found to depend on the width of the crystal.

The presence of a resonant system whose frequency depends upon the width is evidently responsible for this phenomena, this system affecting the frequency of the vibration along the thickness through some form of mechanical coupling. At the suggestion of Mr. R. A. Heising of the Bell Telephone Laboratories, an explanation of these experimental facts was developed based on the treatment of the plate

[11] In plotting the change in rate of vibration of a crystal plate as a function of the dimension, it is more convenient to use wave-length instead of frequency because of the direct linear relation between the dimensions and the wave-length.

as a system of coupled circuits.[12] Consider the two coupled oscillatory circuits shown in Fig. 5 having the uncoupled angular frequencies ω_1 and ω_2, then the frequencies of the coupled system in the absence of damping will be given by the usual expression [13]

$$\omega = \frac{\sqrt{\frac{1}{2}(\omega_1^2 + \omega_2^2) \pm \frac{1}{2}\sqrt{(\omega_1^2 - \omega_2^2) + 4k^2\omega_1^2\omega_2^2}}}{\sqrt{1 - k^2}}, \qquad (4)$$

k being the coupling.

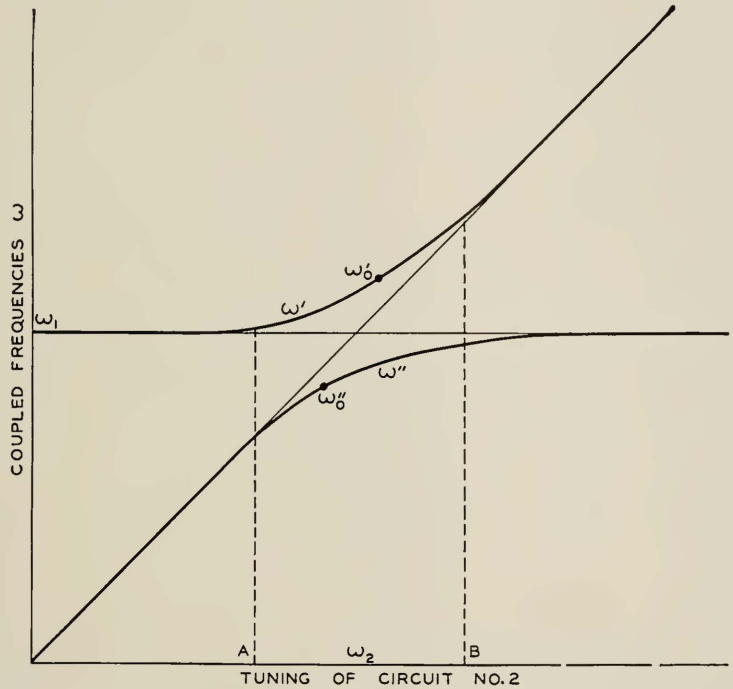


Fig. 6—Angular frequencies of a system of two coupled circuits as a function of the tuning of one circuit, the tuning of the other circuit being fixed.

If these two frequencies be plotted as a function of the tuning of the second circuit, i.e., ω_2, the familiar set of curves shown in Fig. 6 results.

Suppose now other circuits are added to the system as shown in Fig. 7, each additional circuit being fixed at a harmonic of the uncoupled frequency of circuit No. 2, and so linked with this circuit mechanically that the group is tuned as a whole.

[12] The term "circuit" is introduced here to describe a mechanical oscillatory system because many readers are accustomed to think in terms of electrical circuits.
[13] See Pierce, "Elec. Osc. & Waves," Chap. VII.

There are now two possible combinations depending upon which circuit or group of circuits is kept fixed while the other is varied. If the case in which the second group of circuits is kept fixed be examined first, it will be seen that as the frequency of circuit No. 1 is varied it will come into tune successively with each of the circuits of the second group. The result will be a series of coupling curves with the characteristic reaction illustrated by Fig. 6 repeated at each coincident point. If it be assumed that the coupling decreases as the order of the harmonic increases, then the magnitude of the reaction also decreases. This is illustrated by Fig. 8, which shows the coupling curves of such a system plotted in terms of the equivalent electrical wave-length.

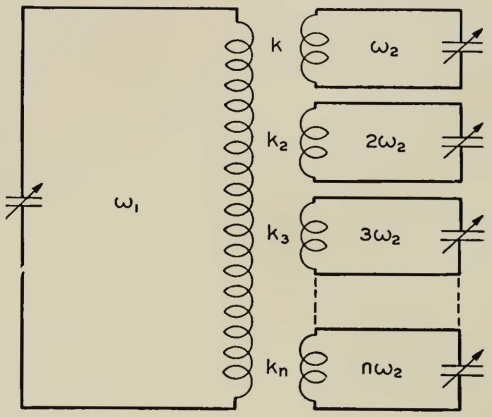


Fig. 7.

Returning to the crystal plate, if the vibration in the direction of the thickness be identified with circuit No. 1 while the width vibration and its harmonics be identified with circuit group No. 2, then Fig. 8 should represent what happens to the crystal wave-length as the thickness is reduced. Comparing Figs. 4 and 8, it is seen that this is true in a restricted region but that the wave-lengths which depend upon the width vibration do not continue much beyond the coupling region in the experimental curves. This is to be expected, for these wave-lengths which depend upon a harmonic of a vibration transverse to the applied field are more difficult to excite than the fundamental in the direction of the field. This particular point is discussed further in connection with temperature coefficients.

If now the case be examined in which the tuning of the second group of circuits is varied, it will be seen that the coupling curves

are slightly different in character. The curves for this case are illustrated by Fig. 9 which shows the wave-lengths of this system as a function of the tuning of circuit group No. 2. Fig. 10 shows the wave-lengths at which a parallel cut plate will oscillate plotted as a function of the width. The similarity between this experimentally

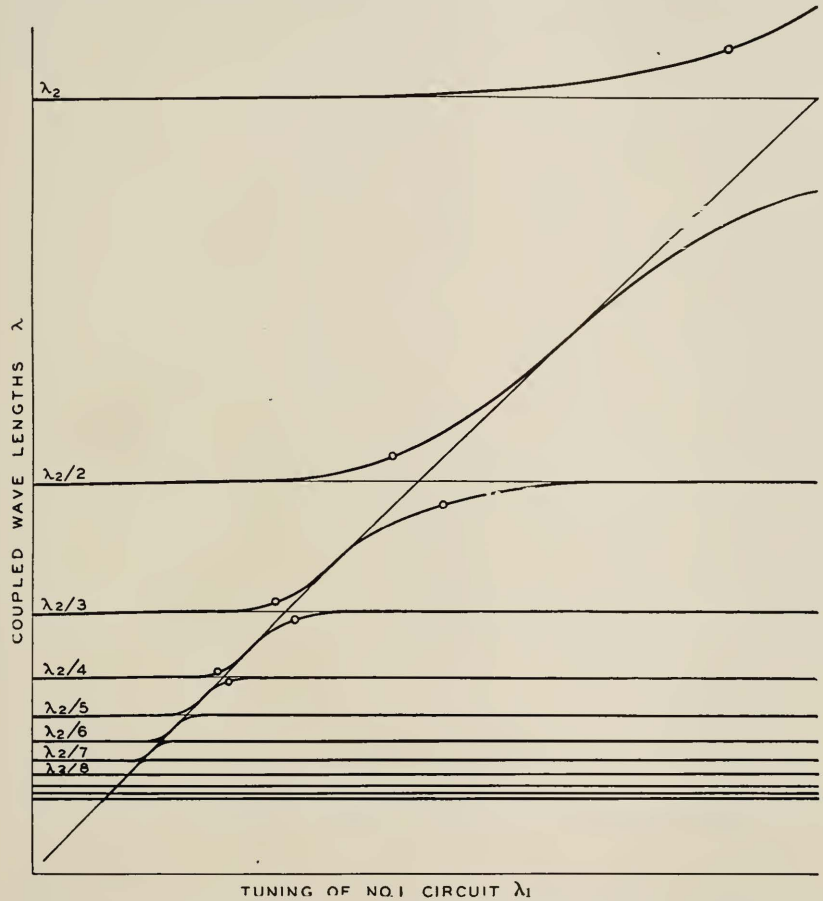


Fig. 8—Wave-lengths of the system of coupled circuits of Fig. 7 as a function of the tuning of circuit No. 1, the tuning of circuit group No. 2 being fixed.

determined curve and Fig. 9 is at once apparent. There is one anomalous segment of a curve between the 7th and 8th harmonics, the line AB; but it is possible that this is caused by the coupling of some third free period which has not been considered, perhaps a high order harmonic of a flexural vibration. In general, however, the curves of wave-length versus thickness for these crystal plates are of

such a character as to indicate that the analogy between the two systems of coupled circuits and the crystal modes of vibration is sufficiently good to serve as a useful guide.

If, then, these parallel cut crystal plates are considered as a system of coupled circuits the reason for the variation of the frequency-thickness constant with dimensional ratios and the presence of the frequency doublets is at once apparent. With the coupling at the various harmonics determined, the character of these variations can

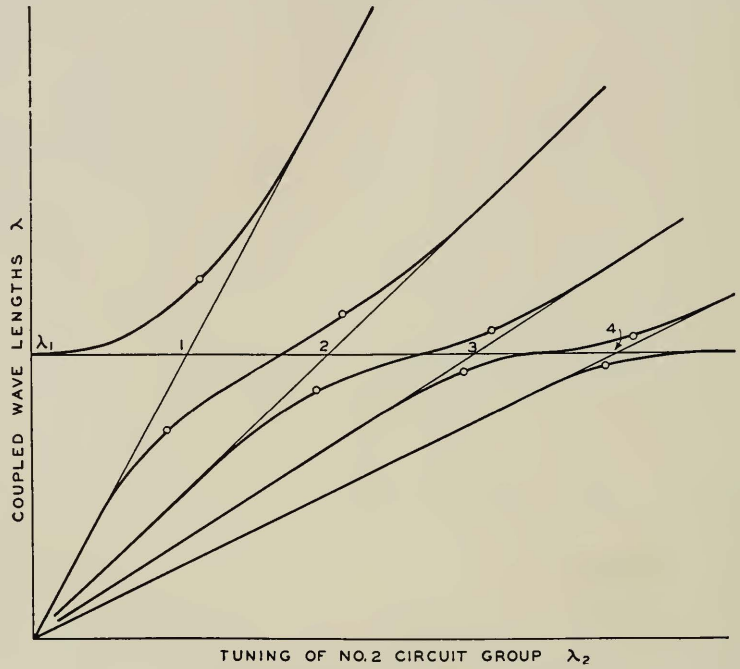


Fig. 9—Wave-lengths of the system of coupled circuits of Fig. 7 as a function of the tuning of circuit group No. 2, the tuning of circuit No. 1 being fixed.

be predicted. Given an experimentally determined series of coupling curves similar to Fig. 10, the coupling at the nth harmonic can be determined from the expression

$$k_n = \frac{\left(\frac{\lambda'}{\lambda''}\right)^2 - 1}{\left(\frac{\lambda'}{\lambda''}\right)^2 + 1}, \tag{5}$$

where λ' and λ'' are the wave-lengths of the coupled system at the point where $\lambda_1 = n\lambda_2$.

Temperature Coefficient as Function of Dimensions and Temperature

As mentioned above, when the temperature coefficients of these parallel cut crystals were studied it was found that there was considerable variation between plates having the same thickness but slightly different areas, and the temperature coefficient of a given plate was found to be a function of the temperature. To illustrate this last point a typical frequency-temperature curve for a parallel

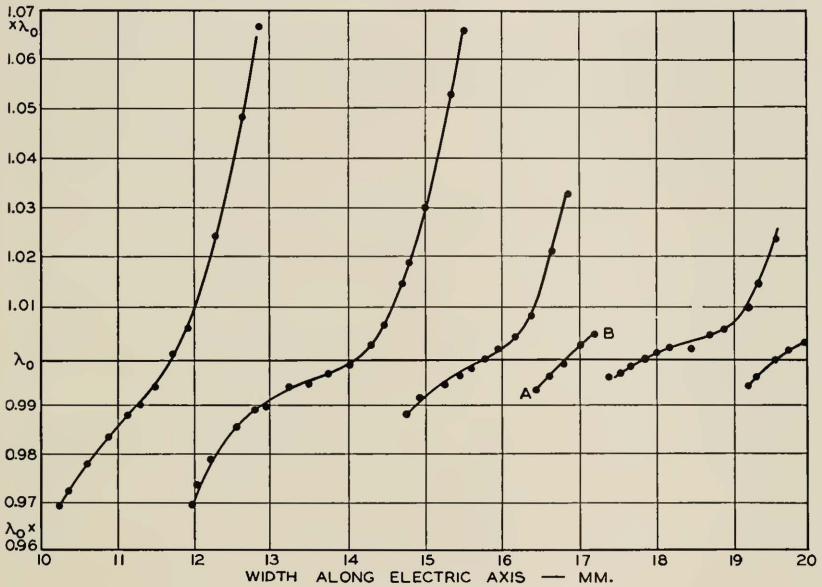


Fig. 10—The wave-lengths at which a parallel cut crystal will operate in an oscillator circuit as its width (the dimension along the electric axis) is progressively reduced, the other dimensions being fixed.
$\lambda_0 = 153 \times$ thickness.
Thickness along mechanical axis = 1.64 mm.
Length along optic axis = 54.8 mm.

cut crystal is shown in Fig. 11. It will be noted that the frequency increase is linear until a given temperature is reached, at which point the curve flattens off and then begins to reverse. Just beyond the point of reversal the frequency jumps to a new value and, if the curve is continued, the frequency increases again at the same rate as originally. This type of frequency-temperature curve is common to a large percentage of parallel cut crystals, the only difference being the width of the flat part of the curve and the temperature at which the discontinuity occurs.

Mr. W. A. Marrison of the Bell Telephone Laboratories first suggested that low temperature coefficients could be obtained with parallel cut crystals by utilizing the coupling of two modes of vibration having individual coefficients of opposite sign. Several of this type of low temperature coefficient crystals were produced by Marrison

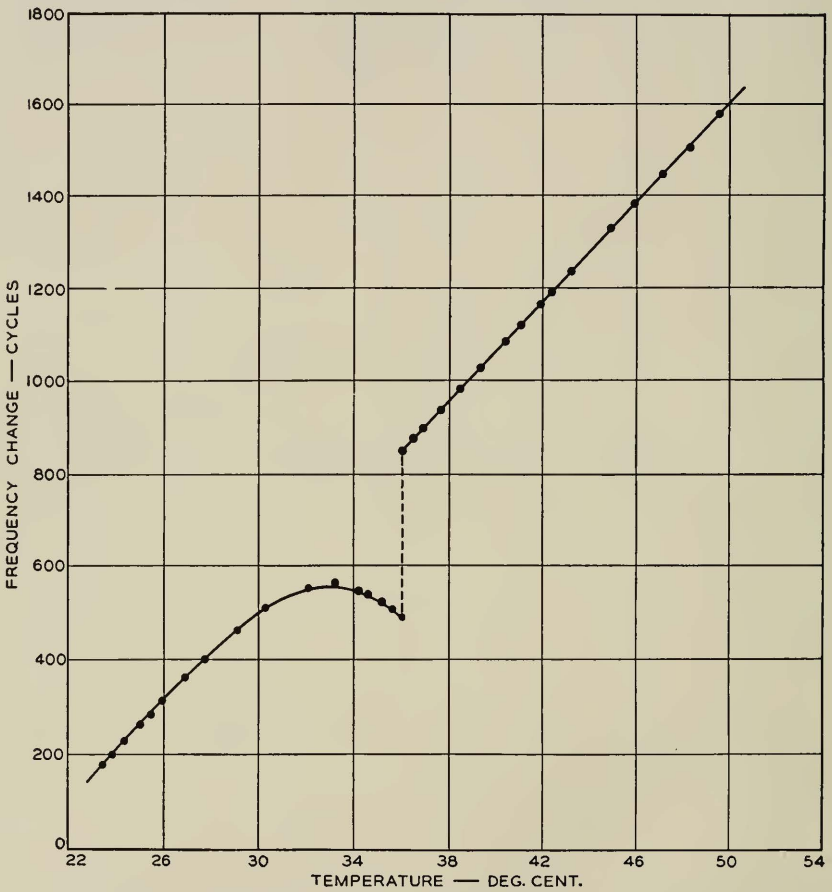


Fig. 11—The frequency change of a 32 x 47 x 2.760 mm. parallel cut crystal with temperature.

and are described in his concurrent paper "A High Precision Standard of Frequency."

If Heising's coupled circuit analysis is extended to include the effect of temperature and the proper temperature coefficients with due regard to relative magnitude and sign are identified with each circuit, the change in temperature coefficient with dimensional ratio and

temperature can be explained. In addition, the dimensional ratios or tuning points which yield zero temperature coefficients for a given temperature can be predicted if the coupling is known.

Referring again to Fig. 6, suppose the two coupled circuits have temperature coefficients of opposite sign, circuit No. 1 being positive and circuit No. 2 negative, for ω_2 less than ω_1 say at the point A, ω' has a positive and ω'' a negative temperature coefficient. For a value of ω_2 greater than ω_1 say at B, ω' now has a negative and ω'' a positive temperature coefficient, ω' and ω'' having interchanged rôles. Somewhere between therefore, both ω' and ω'' must have had a zero temperature coefficient. Returning to equation (4), if this expression for ω be differentiated with respect to the temperature, regarding k, the coupling as constant, and the result placed equal to zero, the condition that ω is independent of temperature [15] is obtained as follows:

$$\omega^2 = \frac{\omega_1^2\omega_2^2(m - n)}{(m\omega_1^2 - n\omega_2^2)}, \tag{6}$$

where

$$m = \frac{1}{\omega_1}\frac{d\omega_1}{dT} = \text{temperature coefficient of circuit No. 1,}$$

$$n = -\frac{1}{\omega_2}\frac{d\omega_2}{dT} = \text{temperature coefficient of circuit No. 2;}$$

now let $Q = n/m$ then equation (6) becomes

$$\omega^2 = \omega_2^2\frac{1 - Q}{1 - Q\left(\frac{\omega_2}{\omega_1}\right)^2}, \tag{7}$$

solving equation (7) for ω_2/ω_1 replacing ω^2 by its value from equation (4)

$$\left(\frac{\omega_2}{\omega_1}\right)^2 = \frac{k^2(1 - Q)^2}{2Q} + 1 \pm \sqrt{\left[\frac{k^2(1 - Q)^2}{2Q}\right]^2 + \frac{k^2(1 - Q)^2}{Q}},$$

which when k is small becomes

$$\left(\frac{\omega_2}{\omega_1}\right)^2 = 1 \pm \frac{k(1 - Q)}{\sqrt{Q}}. \tag{8}$$

This equation gives the tuning points, or the values of ω_2 at which the angular frequencies of the coupled system, ω' and ω'', will have

[15] Dr. F. B. Llewellyn of the Bell Telephone Laboratories is responsible for this analysis.

zero temperature coefficients, in terms of the ratios of the uncoupled temperature coefficients and the coupling.

Referring again to Fig. 6, ω_0' and ω_0'' represent the values of ω' and ω'' which would have zero temperature coefficient provided m is greater than n, that is, the temperature coefficient of ω_1 is greater in magnitude than that of ω_2. Carrying this idea over to the case of the group of circuits for which the curves of Figs. 8 and 9 are drawn,

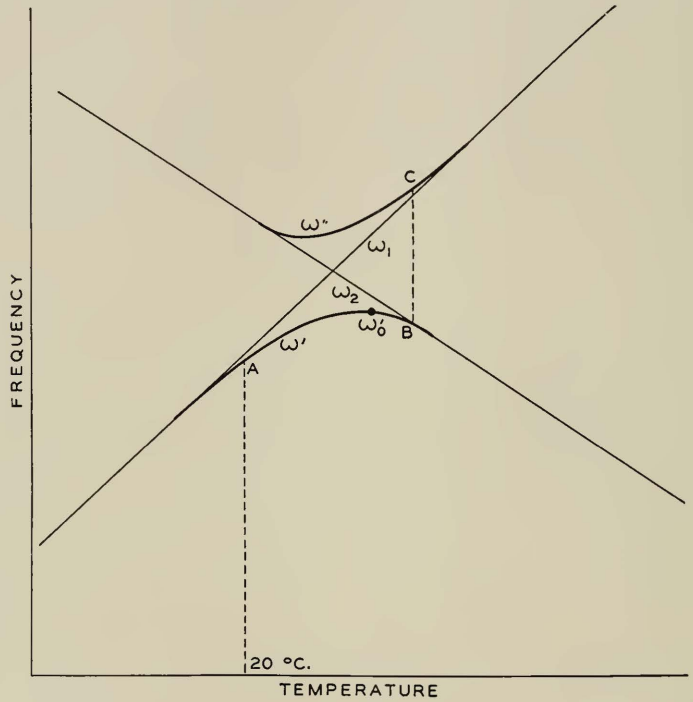


Fig. 12—Effect of temperature on the angular frequencies of a system of two coupled circuits having temperature coefficients of opposite sign.

there will be points of zero coefficient in the neighborhood of each coupling point as indicated by the circles on the curves.

The above conditions for zero temperature coefficient only apply if the coefficient as a function of the tuning be examined in the region of some given temperature. If the temperature is varied over a considerable range a small change in the tuning of both circuits is effected, one having its frequency raised, the other lowered. The result of this tuning on the frequencies of the coupled system can be illustrated by Fig. 12, which shows the tuning with temperature on a magnified scale. In this figure, the lines ω_1 and ω_2 represent the change

in frequency of these circuits with temperature if there were no coupling between them. The change of the frequencies of the coupled system with temperature is shown by the curves ω' and ω''. It will be seen that both these frequencies pass through regions of zero temperature coefficient.

Such frequency-temperature curves can be derived graphically by a construction similar to that shown in Fig. 13. This figure illustrates what happens when the tuning of both circuits No. 1 and No. 2 is varied, and consists of a series of the usual coupling curves (the

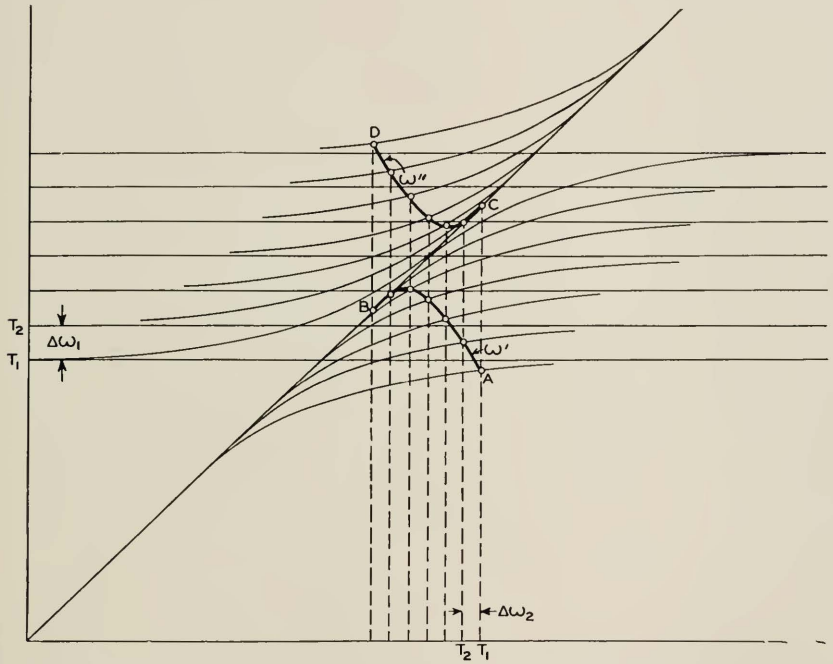


Fig. 13—Effect on the angular frequencies of a system of two coupled circuits as the individual circuits are tuned simultaneously in opposite directions.

coupled frequencies plotted as a function of ω_2), each set of curves of the series being drawn for a different value of ω_1. When the temperature is increased from T_1 to T_2, the uncoupled frequency of circuit No. 2 is reduced by an amount $\Delta\omega_2$ and the uncoupled frequency of circuit No. 1 is increased by an amount $\Delta\omega_1$. The result is the frequencies ω' and ω'' move from curve to curve in the direction shown by the lines AB and CD.

Now if the variation of the temperature coefficient of a crystal plate when used in an oscillator circuit be examined at a given tempera-

ture as the width is changed (which amounts to a change in the tuning of the transverse vibration), it will be seen that the experimental results are in accord with the above treatment. Fig. 14 shows the temperature coefficient of the two frequencies of a crystal plate at 58° C. as its width is progressively reduced in the neighborhood of the 5th harmonic of the transverse vibration. These curves show how the temperature coefficients change sign in this region. The dotted sections of the curves are extrapolated, for owing to the rapid reduction in activity once a coupled frequency acquires a negative coefficient,

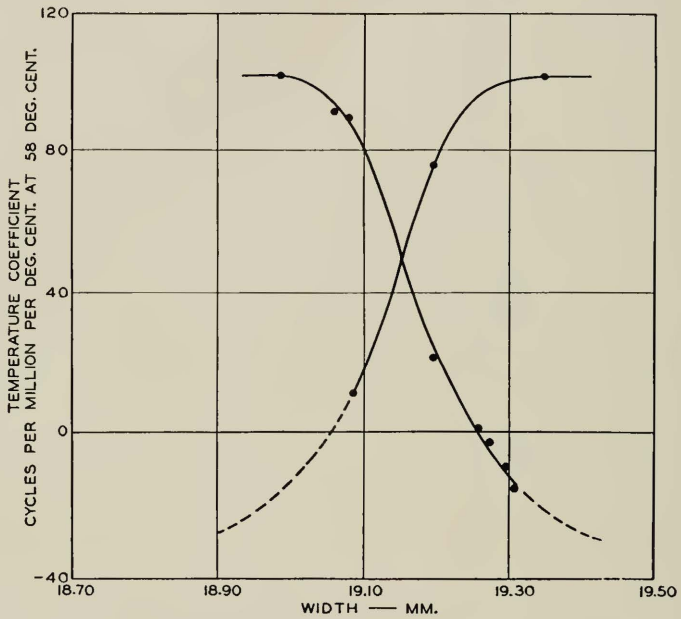


Fig. 14—The change of temperature coefficient of a parallel cut crystal at 58° C. as the width is progressively reduced in the region where the fifth harmonic of the vibration in the direction of the width coincides with the frequency of the vibration in the direction of the thickness.

data on the crystal plate used as an oscillator are difficult to obtain in this region.

Returning to the experimentally determined curve of frequency versus temperature for a parallel cut crystal plate shown in Fig. 11, this can also be explained with the aid of the above analysis. Referring to Fig. 12, if it be assumed that at 20° C. the crystal is oscillating with a frequency A, this is in the region where this particular frequency has a positive temperature coefficient. As the temperature increases the frequency increases in the direction of B, passing through a

maximum at the point ω_0' where it has zero coefficient, and then decreases. As the frequency decreases the activity of this particular mode decreases rapidly and finally the crystal "hops" frequency to point C on the ω'' curve. From this point on the frequency with

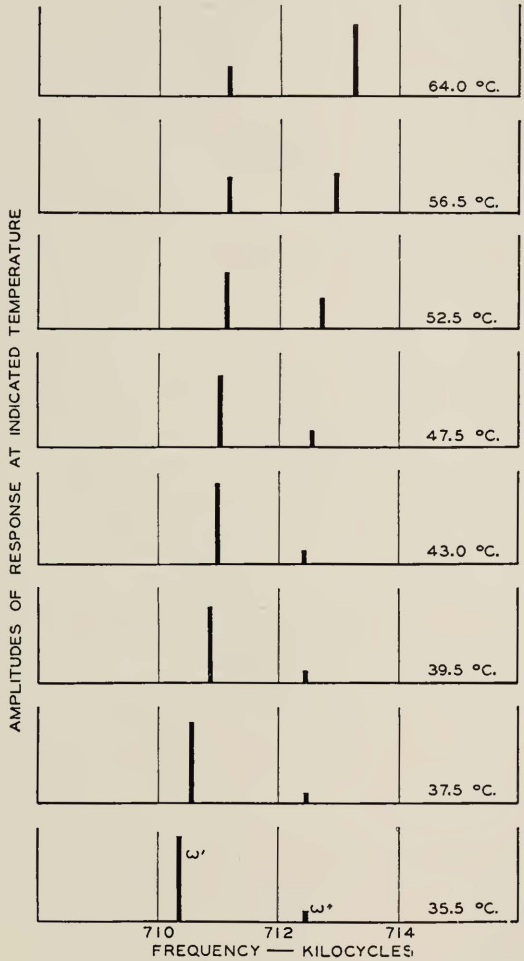


Fig. 15—The response-frequency spectra of a parallel cut crystal plate at different temperatures illustrating the interchange of activity between the two frequencies as the frequencies of the two modes of vibration pass through a coincident value.
Length of plate along optic (Z) axis = 47 mm.
Width along electric (X) axis = 19.35 mm.
Thickness along mechanical (Y) axis = 2.75 mm.

temperature increases, for this frequency has a positive temperature coefficient in this region.

If it were not for the decrease in activity of the period with the negative temperature coefficient it is to be expected that the crystal frequency, instead of "hopping," would continue to decrease with increase in temperature. In some instances (for low order of harmonics) the crystal frequency will decrease for a few degrees, and it

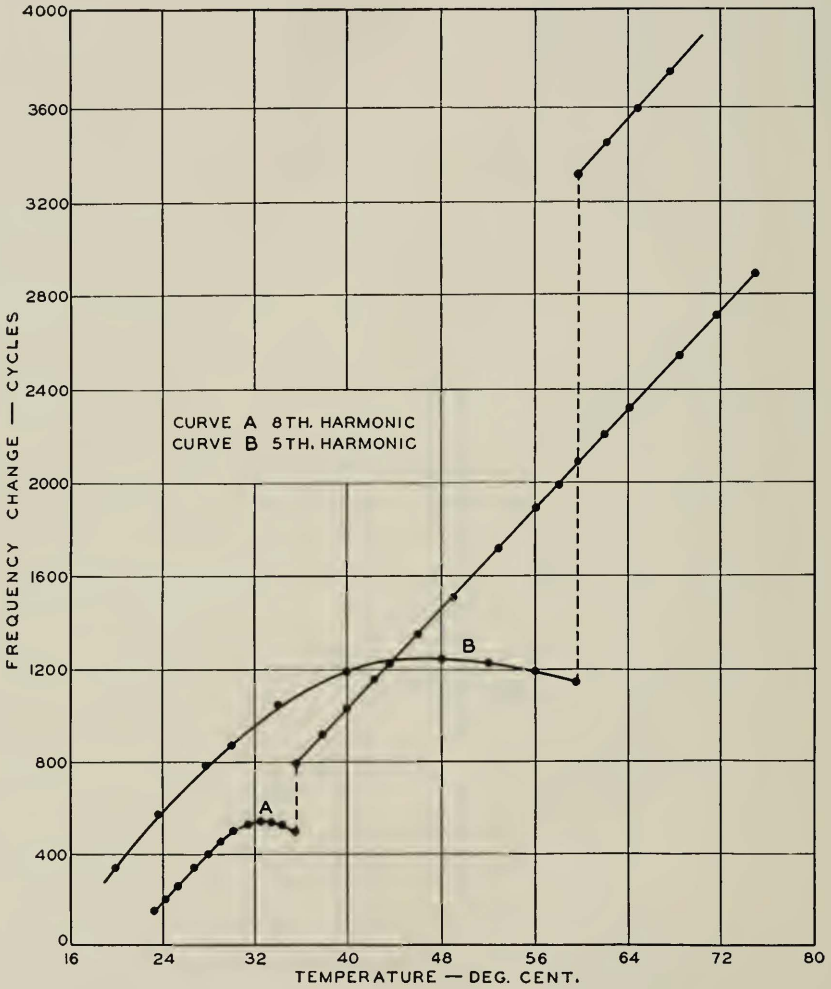


Fig. 16—The frequency change with temperature of two parallel cut crystals of different width.

Curve *A* region of eighth harmonic
(width = 32.0 mm.)
Curve *B* region of fifth harmonic
(width = 19.35 mm.)
The other dimensions of the plates are identical.

is found that the magnitude of the negative coefficient for this region approximates that to be expected for the transverse vibration alone. In general, however, a frequency jump occurs just after the zero temperature coefficient region is passed.

This interchange of activity of these two periods as they interchange temperature coefficients can be studied in detail by examining the changes in the spectrum of a crystal at different temperature levels. Fig. 15 shows a series of spectra of a crystal taken for different temperatures in the region of zero temperature coefficient, the dimensions of the crystal being unchanged. These spectra illustrate the rapid decrease in activity of the frequency ω' after it passes through zero temperature coefficient while at the same time ω'' increases and assumes the place of major activity vacated by ω'.

The assumption that the coupling increases with the decrease in the order of the harmonic finds confirmation in the experimentally determined facts as computed from curves of the type shown by Fig. 10. As the coupling increases the temperature range for which there is no frequency change with temperature increases, that is the region of zero temperature coefficient becomes extended. To illustrate this, Fig. 16 shows two curves of frequency versus temperature, one for the coupling of a fifth harmonic, the other for an eighth.

It would, of course, be desirable to extend the zero temperature coefficient range over the limits of temperature to be expected in normal operation. This necessitates tight coupling of the two modes which in turn demands a dimensional ratio in the neighborhood of unity. The cross sectional area of such a plate in the direction of its thickness and width approaches a square in shape which, for high-frequency crystals, is of very small dimensions.

Before concluding it should be noticed that since both modes of the perpendicular cut crystals have a negative temperature coefficient, it is to be expected that it would be impossible to obtain zero temperature coefficient crystals with this orientation. This seems to be true as far as our experience with those crystals is concerned.

Master Reference System for Telephone Transmission [1]

By W. H. MARTIN and C. H. G. GRAY

The telephone transmission system described here is the Master Reference System of the Bell System for the expression of transmission standards and the ratings of the transmission performance of telephone circuits. The transmitter and receiver elements of this system are reference standards for the ratings of the transmitting and receiving performance of terminal station sets.

A replica of this reference system, installed in Paris, has been adopted as the Master Reference System of the International Advisory Committee on Long Distance Telephone Communication in Europe. The establishment of these two master systems provides a common reference for the telephone transmission work of the Bell System and the telephone administrations which are members of this International Advisory Committee.

THE Master Reference System for Telephone Transmission, as its name indicates, is to serve as the fundamental circuit in the ratings of the transmission performance of telephone circuits. In describing this system, therefore, it will be advantageous to outline first the general considerations underlying the methods of determining and specifying these ratings and their applications.

The conversions and transfers of energy which constitute the process of telephone transmission result in general in a difference between the speech sounds at the sending end of the telephone circuit and the sounds reproduced at the other end in the ear of the listener. These reproduced sounds may differ from the original in three important respects; their loudness, their distortion or degree to which their wave shape departs from facsimile reproduction, and the amount of extraneous sound or noise which accompanies them. From the standpoint of telephony, the major importance of a difference between the original and reproduced sounds is determined by its effect on "intelligibility," that is, the degree to which the latter sounds can be recognized and understood by the listener when carrying on a telephone conversation. The tolerable departure of the reproduced from the original sounds is limited also by certain effects which are noticeable to the listener before they materially affect intelligibility, such as loss of naturalness.

Measurements of intelligibility are of utmost importance in rating the performance of telephone circuits, but they are unduly cumbersome for direct use in the detailed development and design of telephone circuits and their many parts, particularly where small effects are concerned. It has been desirable, therefore, to handle telephone

[1] Presented before A. I. E. E. Summer Convention, June 24–28, 1929.

transmission work in two steps. One natural division is suggested by the statement that intelligibility is a function of the relation between the output and input speech sounds, and of the psychological reaction of the listener to these output sounds. Because of the complex nature of the speech sound waves, however, it has been found more practicable to treat the transmission performance of telephone circuits in the following two parts: (1) the physical performance of the circuit, and (2) the relation between physical performance and intelligibility. The physical performance of a circuit is taken here to cover the transmission characteristics which can be specified in terms of the performance for single frequencies, a number of frequencies being taken to cover the range which is important for the reproduction of speech sounds. These measurements of physical performance cover such things as the response-frequency characteristic of the circuit over the range of speech frequencies, the distortion due to non-linear elements, phase distortion and the extraneous currents which cause noise. These determinations of physical performance do not include measurements of the speech sounds themselves, nor of the functioning of the talker and listener. This differentiation is advantageous in segregating the studies of speech sounds and of the psychological phases of the work, and permits the design of the operating plant and a large portion of the development work to be carried out on a physical basis.

The determination of the relation between intelligibility and the physical performance of a telephone circuit is a laborious process, because persons play the parts of generators and meters and a number of people must be used in both parts to take into account the normal ranges of their performance. The goal of this portion of the work has been, therefore, to establish suitable relations which will permit the determination of the intelligibility of a circuit by computations which start with the physical characteristics of the circuit. This work [2] has involved determinations of the capabilities of circuits having various kinds of physical characteristics to reproduce intelligible speech, investigations of the nature of speech sounds and hearing, and of people's customs in using the telephone.

Prior to the time when suitable means were available for measuring the physical performance of telephone circuits, and when the kinds of circuits in commercial use were quite similar in their distortion characteristics, the practice was adopted of rating the performance of a circuit by comparing it on a loudness basis with a reference circuit which was adjustable in attenuation, and whose distortion was closely similar to that of the commercial circuits. In such a comparison, a

[2] "Speech and Hearing," by H. Fletcher, published by D. Van Nostrand Co.

determination is made of the equivalence of loudness or volume of these two circuits by talking alternately over them and adjusting the reference circuit until the sounds coming out of the two receivers are judged to be equal. For the conditions where volume is the important controllable characteristic of telephone circuits, these loudness comparisons constitute a practicable and effective means of indicating the performance of these circuits.

The reference circuit adopted about twenty-five years ago for these loudness comparisons consisted of transmitters, receivers, station sets, cord circuits and a line, of types which were then used commercially. In this reference circuit the line was an adjustable artificial line simulating a 19 AWG. cable circuit having a capacity per loop mile of 0.054 mf. The amount of cable in this line to give a loudness

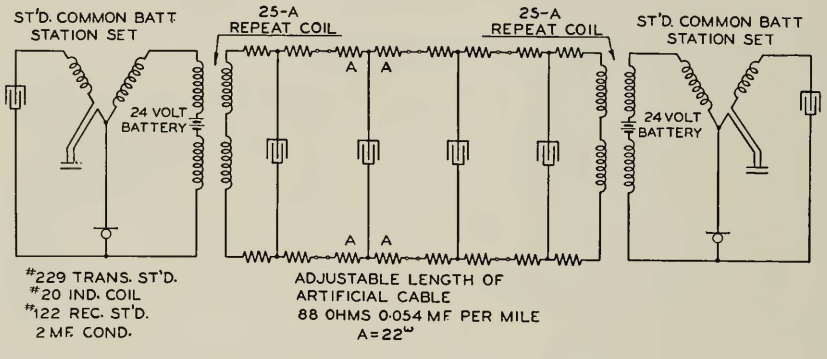


Fig. 1—Standard cable reference system.

balance was taken as the rating of the circuit under comparison. This reference system, shown schematically in Fig. 1, has been called the Standard Cable Reference System.

In addition to this rating of the performance of a telephone circuit, this standard cable reference system has had other applications. Certain settings of this reference system were selected as specifying the standards of transmission which were to be provided in the design and operation of commercial circuits for the several kinds of service, such as local and toll. The effect of introducing or changing any part in a commercial circuit was rated in terms of the amount of cable by which it was necessary to change the line of the reference system to produce the same effect on the loudness of the reproduced speech sounds. Likewise, the transmitters and receivers of this reference system were used as reference instruments for the comparison and rating of other transmitters and receivers.

This cable reference system has played a very important and necessary part in the development of telephone transmission, in that it has provided a ready means of rating the performance of the various parts of the system and of any changes, and made it possible to design commercial circuits to provide a predetermined grade of service. The performance of this system was specified by stating the kinds of apparatus and circuits used. The performance of the elements of the electrical portion of the system could be checked by voltage, current and impedance measurements, but for the transmitters and receivers reliance for constancy of performance was placed primarily upon the careful maintenance and frequent cross comparisons of a group of transmitters and receivers which were specially constructed to reduce some of the sources of variation in the regular product instruments. In this way, reasonable assurance of the performance of the reference system was secured. This system has been widely used both in this country and in other parts of the world, and the performances of the various systems have been kept in accord by frequent circulation of calibrated transmitters and receivers.

As the telephone art has developed, modifications have been found to be desirable in this reference system to make it more suitable for its purpose. Telephone instruments and circuits have been designed and used which have less distortion than existed in the corresponding parts of the cable reference system. For this reason it is desirable to have as a new reference system one with which the transmission over the most perfect telephone circuit or over some less perfect one may be simulated at will. The change of the unit of transmission from the mile of standard cable to the decibel [3] has brought about the need for a change in the line of the reference system.

In selecting a new reference system, it is obviously desirable to eliminate as far as possible the factors which are not subject to exact measurement, or which may possibly vary with time. For this reason, the elements of the new system have been chosen so that their performance may be definitely measurable at any time, and may remain as far as possible invariable. This applies also to those elements which are provided for insertion in the system when it is wished to produce some distortion which will make more easily possible a loudness balance between the reference system and the circuit under investigation.

A reference system such as that described here, in which the essential elements are so constructed as to reproduce speech with a high degree

[3] "Decibel" (db) is the name for the Transmission Unit which has superseded the "mile of standard cable." W. H. Martin, *Bell System Technical Journal*, January, 1929, and *A. I. E. E. Journal*, March, 1929.

of perfection, and with which provision may be made for modifying the speech in definite and reproducible ways, affords a convenient means for studying the capabilities of telephone circuits of different physical performances. These investigations, however, are outside the purpose of this paper, which is to describe the new reference system and its application in making volume ratings.

General Requirements

The outstanding conception of the new reference system is that its performance should be suitable to serve as a reference base line for indicating the performance of all telephone circuits and that the transmitter and receiver elements of the new system should provide similar base lines for the performance of electroacoustic converters. To meet these needs properly the performance of the system and its parts should be capable of being measured and definitely specified in terms of physical quantities. In this connection, there arises a matter which has been the subject of much discussion, namely, as to whether or not the specified performances of the reference transmitter and receiver should be those which are realized when used as telephone instruments. In regard to the transmitter, the difficulty comes in specifying the input when it is placed in front of the mouth of the talker. This is due to the non-uniformity of the sound field from the speaker's mouth, the nature of the waves of speech sounds and the reflections of these waves from the transmitter. In regard to the receiver, the difficulty is due to determining the output when the receiver is held to the ear. To obviate these, it has been decided to specify the performance of the transmitter in terms of the electrical output for a given pressure on the diaphragm of the transmitter and of the receiver in terms of pressure set up in a simple closed chamber for a given electrical input to the receiver. These conditions are definite and reproducible.

The most important requirement, then, for the reference system is that the physical performance of the system and of its component parts should be capable of being measured and definitely specified in terms of physical quantities. If this requirement is realized, a system providing the specified performance can be set up wherever desired. This has been the main criterion in the design of the master reference system for telephone transmission which is described here.

The second main requirement is that specifiable and predeterminable changes can be made with respect to the performance which is selected as the reference. These changes must be capable of varying the relation between the loudness of the reproduced sounds with respect to the initial sounds, the distortion of the wave shape of these repro-

duced sounds, and also the amount of noise accompanying these reproduced sounds.

For convenience in specifying these requirements, it has been found desirable to impose another requirement, namely, that the system be capable of giving a performance which is as free as possible from distortion and noise. It should be noted that two kinds of distortion must be taken into account: that due to unequal efficiency for sounds of different frequencies and that due to non-linearity causing unequal efficiency for sounds of different magnitudes. This requirement is also of advantage in insuring that the reference system and its parts will have less distortion than any circuit or instrument with which it may be compared. This will permit the simulation of the distortion of such instruments or circuits by the insertion of distortion in the reference system.

For convenience in use, it is highly desirable that the performance of the reference system and its parts be constant for a reasonable time under normal operating conditions.

General Features

The master reference system [4] employs a transmitter and receiver which are capable of a high degree of freedom from distortion. The transmitter is of the condenser type and the receiver is of the moving coil type. Both these instruments are materially lower in efficiency than commercial types of apparatus, but this condition is compensated for by the use of multi-stage vacuum tube amplifiers. These instruments, together with their associated amplifiers, constitute reference standards for converters between acoustic and electrical energy. The third necessary element of a telephone transmission system, namely the line, is provided by a network of resistance elements. Such a line can be made to provide uniform attenuation over a wide frequency range, and can be made to control the magnitude of this attenuation over a large range. This line is taken as giving a reference performance for lines.

The specification of the performance of such a system is based on the principle of the thermophone, which is a converter of electrical energy into acoustic waves by means of the heat generated by the passage of an electrical current through a resistance. From a knowledge of the form and physical constants of this resistance element, of the medium in which it is used, and of the electrical input to the element, the acoustic pressure generated in a chamber of known size

[4] A discussion of a preliminary model of this system was given in "A Telephone Transmission Reference System," by L. J. Sivian, *Electrical Communication*, October, 1924.

can be determined by theoretical considerations.[5] The performance of the condenser transmitter is determined by making its diaphragm a wall of a simple closed chamber in which the thermophone is placed. By this means a known pressure wave of any frequency over the range desired can be impressed upon the diaphragm of the transmitter. The voltage output of the transmitter for a specified circuit condition is then measured. From this measurement the ratio of the voltage output to the acoustic pressure on the diaphragm is established for that instrument and circuit condition. With the performance of the transmitter thus established, the performance of the receiver element of the reference system is measured by acoustically coupling the receiver to the condenser transmitter, so that the receiver actuates the transmitter, and then determining the relation of the pressure generated by the receiver in the coupler to the voltage input to the receiver. The performance of the line element is determined by well-known means. The performance of the whole system can then be expressed in terms of the pressure produced by the receiver with respect to the pressure on the diaphragm of the transmitter.

The performance of this system is practically free from distortion for the energies which it is required to handle, and probably materially excels in this respect that of any previous system. With this system, volume relations between output and input sounds can be varied over a wide range with practically no accompanying distortion. In comparing such a system, however, with commercial systems, it is advantageous also to be able to control distortion. This is particularly the case when using the instruments of the master system for rating the volume efficiency of commercial types of transmitters and receivers. To facilitate this, arrangements are made for the introduction into the amplifiers associated with the transmitter and receiver, of networks which may be designed to give a variation of efficiency with frequency which corresponds to that obtained with commercial apparatus. These networks and their distortion effect can, of course, be definitely specified. The line element of this master system can be replaced by a line or network giving any type of distortion desired. Also, known amounts of extraneous currents to produce noise can be introduced into this circuit without otherwise appreciably affecting its performance. This is accomplished by connecting a relatively high impedance source of voltage of the desired wave shape across a circuit element of relatively low resistance.

This system provides a performance which is definitely known, and

[5] The theory of the thermophone as a precision source of sound is outlined in papers by H. D. Arnold and I. B. Crandall, *Physical Review*, July, 1917, and by E. C. Wente in the *Physical Review*, April, 1922.

which can be varied over a range of volume and distortion. It well meets the requirements and represents a material advance over the standard cable reference system which it now replaces. In order to tie together ratings established in terms of the old system with those of the new system, comparisons of the two have been made to determine the adjustments of the new system which make its loudness performance correspond with that of the old.



Fig. 2—Master reference system for telephone transmission with associated calibration apparatus.

Description of Master Reference System

The master reference system with associated calibration apparatus is shown in Fig. 2. This equipment, mounted on steel panels and racks, and arranged as shown, is installed in a room, shielded from acoustical and electrical disturbances, at the Bell Telephone Labora-

tories in New York City. In Fig. 2 the transmitter, line, and receiver of the reference system are shown on the three racks at the right. The calibration apparatus, consisting of an oscillator, thermophone, vacuum tube voltmeter and volume indicator are shown on the other three racks. A schematic diagram of the master reference system is shown in Fig. 3.

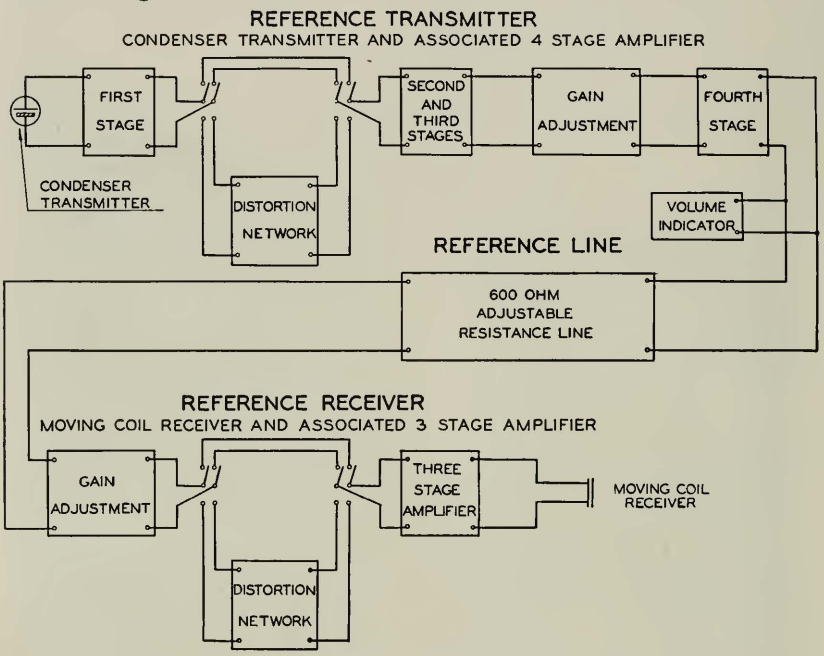


Fig. 3—Schematic diagram of master reference system.

The reference transmitter consists of a condenser transmitter and associated four-stage amplifier. The condenser transmitter is simple and rugged in construction. It has been available for work of this nature for several years during which time considerable experience has been obtained with it. In this system the condenser transmitter is held in an adjustable mounting and is equipped with a guard by means of which the speaker can keep his lips at a fixed distance from the diaphragm. This guard consists of a wire ring 4.7 cms. in diameter, held parallel to and at a distance of 4.1 cms. from the diaphragm of the transmitter by three wire supports. A volume indicator with meter visible to the speaker enables him to maintain an approximately constant talking intensity. In the condenser transmitter,[6] shown in

[6] The theory and operation of this transmitter are discussed in papers by I. B. Crandall, *Physical Review*, June, 1918, and E. C. Wente, *Physical Review*, July, 1917, and *Physical Review*, May, 1922.

Fig. 4 and in cross-section in Fig. 5, a thin highly stretched duralumin diaphragm is mounted close and parallel to a steel plate grooved and

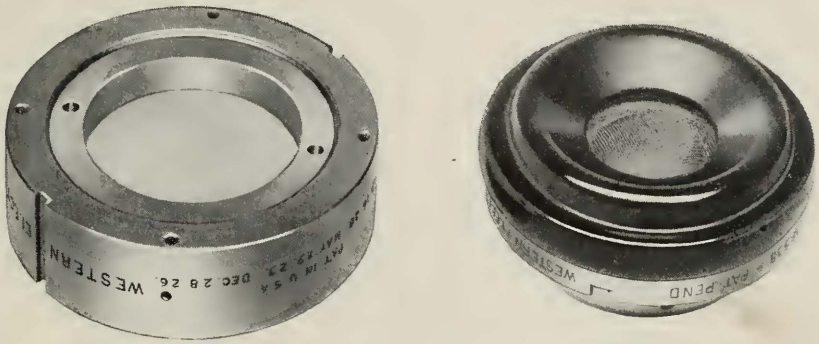
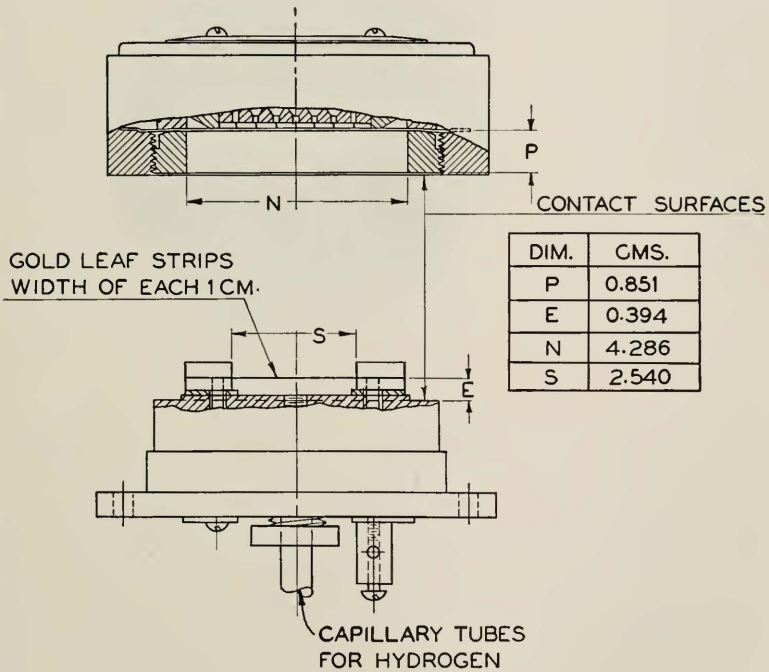


Fig. 4—Condenser transmitter and moving coil receiver.

perforated for air damping. This diaphragm and plate form the electrodes of a condenser polarized by a battery through a high



DIM.	CMS.
P	0.851
E	0.394
N	4.286
S	2.540

Fig. 5—Schematic of condenser transmitter and thermophone.

resistance. Sound vibrations impinging on the diaphragm actuate it and produce variations in the capacity of this instrument. The

resulting alternating potentials, faithfully representing the sound pressure waves, are amplified by the associated amplifier. Between the first and second stages of this amplifier, provision has been made for the introduction of distortion networks to simulate the distortion of any transmitting system. Between the last two stages are attenuating networks permitting adjustment of the relation between transmitter input and transmitter output over a range of 22 db in steps of



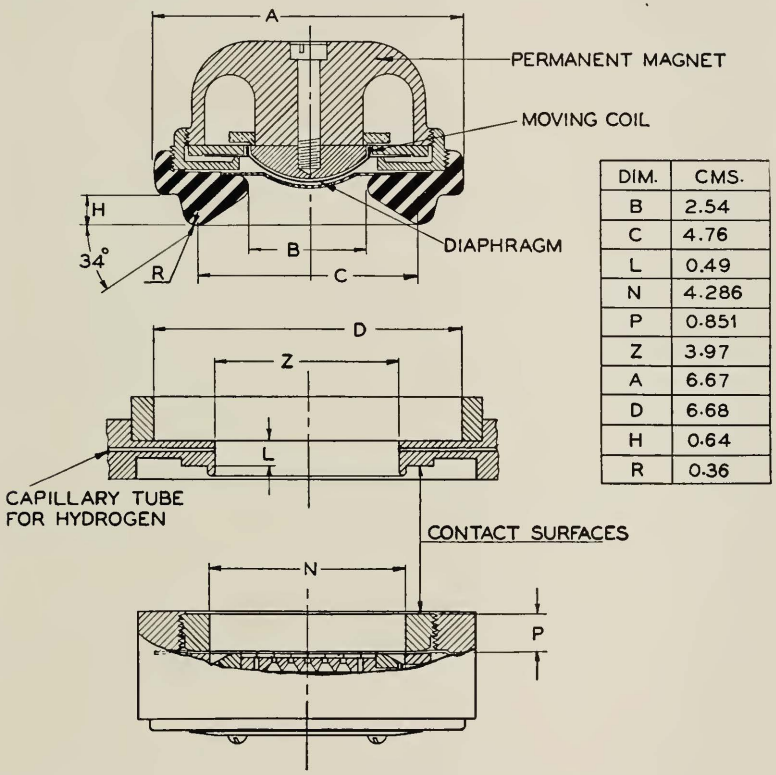
Fig. 6—Thermophone in chamber for calibrating transmitter.

0.1 db. The output impedance of the amplifier is 600 ohms with negligible phase angle.

The reference line consists of a series of balanced resistance networks. This line has a characteristic impedance of 600 ohms matching the output impedance of the reference transmitter and the input impedance of the reference receiver, thereby eliminating any reflection effects at these junctions. By means of suitable controls the attenuation may be varied over a range of 101 db in steps of 0.2 db.

The reference receiver consists of a moving coil receiver and associated three-stage amplifier. Means are provided for adjusting the relation between the input to the receiver and its output over a range

of 22 db in steps of 0.1 db. As in the case of the reference transmitter, provision has been made for the introduction of distortion networks to simulate the distortion of any receiving system. In the moving coil receiver[7] shown in Fig. 4 and in cross-section in Fig. 7, a coil of aluminum ribbon, by vibrating relatively to a fixed permanent magnet, actuates a clamped, unstretched, thin duralumin diaphragm to which



DIM.	CMS.
B	2.54
C	4.76
L	0.49
N	4.286
P	0.851
Z	3.97
A	6.67
D	6.68
H	0.64
R	0.36

Fig. 7—Schematic of moving coil receiver, coupler and condenser transmitter.

it is attached. Air damping in this receiver is secured by an arrangement somewhat similar to that employed in the condenser transmitter. The structure is simple, rugged, and, from the standpoint of freedom from distortion, its performance is comparable to that of the condenser transmitter. The central surface of the diaphragm is protected by a meshed wire screen instead of by the more usual type of receiver cap. This construction avoids resonance effects which

[7] This instrument, which is similar in many respects to one described by E. C. Wente and A. L. Thuras in the *Bell System Technical Journal*, January, 1928, will be discussed in a future paper by the same authors.

would otherwise occur in the confined air space between diaphragm and cap. The contour of the ear piece, resembling that of the familiar telephone receiver, permits the listener to readily center the receiver on his ear.

In order that an individual part of any telephone circuit may be compared with the corresponding element of the master reference system, the system is arranged so that the reference transmitter, reference line or reference receiver, may be replaced by the corresponding part to be rated.

Calibration of Master Reference System

To specify adequately the performance of the master reference system, in terms of definite physical quantities, apparatus is associated with the system for making electroacoustic calibrations of the reference transmitter and reference receiver, and electrical calibrations of the circuits. In general, the method employed in these calibrations is to adjust the setting of an attenuator to obtain a deflection on the galvanometer of a vacuum tube voltmeter, equal to the deflection produced when measuring the output of the element under calibration. This avoids the necessity for an absolute calibration of the measuring device.

The source of alternating currents, used for calibrating purposes, is an oscillator, capable of producing currents with a harmonic output usually less than 3 per cent of the fundamental.

The measuring equipment used for making the above calibrations consists of a two-stage vacuum tube voltmeter in conjunction with a tuned circuit connected across its input. By means of this tuned circuit, harmonics of the fundamental frequency to be measured are attenuated by at least 20 db.

The calibration of the condenser transmitter is made with a thermophone, the gold leaf thermal elements being shown in Fig. 6. In Fig. 5 is shown a cross-section of a condenser transmitter and a thermophone. Petrolatum is used to form a seal between the instrument and the thermophone block. The air in this small enclosed chamber, formed by the walls and diaphragm of the condenser transmitter and the thermophone block, is replaced by hydrogen. Since the velocity of sound in hydrogen is approximately four times that in air, the frequency range over which measurements may be made before standing wave effects are experienced is extended by the use of this gas. In addition, the efficiency of the thermophone is increased, the constant of diffusivity for hydrogen being greater than that for air. Both alternating and direct currents are passed through the gold leaf

thermal elements, the direct current being sufficiently large to make negligible the double-frequency effect resulting when alternating current is supplied to the thermophone. The alternating current passing through the thermal-elements (gold leaf being selected because of its low heat capacity) causes variations in their temperature. Periodic expansions and contractions of the surrounding gas, resulting from the varying heat transfer occasioned by the periodic temperature changes of the thermal elements, constitute sound waves of precisely determinable pressure. These sound waves actuate the transmitter and the ratio of the voltage output to the sound pressure gives the transmitter calibration. The process involved is as follows: Referring to Fig. 8.

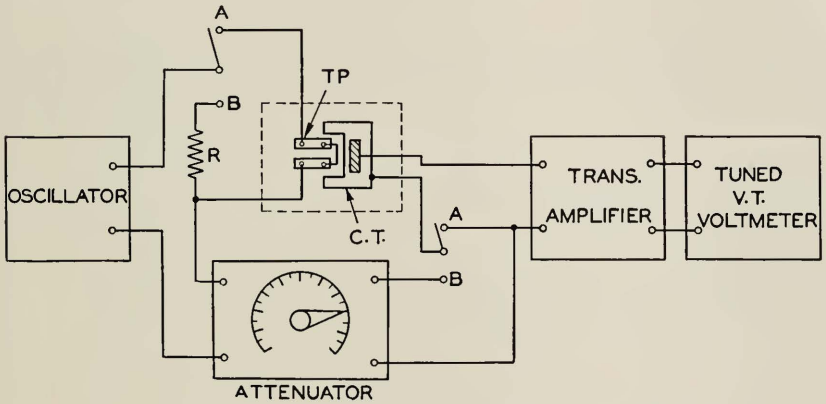


Fig. 8—Circuit for calibrating condenser transmitter.

E_T = Voltage generated by condenser transmitter per bar (one dyne/cm.^{2}).
I_1 = Alternating current through thermophone and attenuator.
P = Pressure developed by thermophone per volt alternating current across it.
V_{VM} = Voltage across voltmeter.
A = Ratio of voltage delivered to 600-ohm load by the transmitter amplifier to voltage impressed in series with condenser transmitter.
V_0 = Voltage across attenuator input.
R_1 = Input impedance of attenuator.
R = Thermophone resistance.
N = Attenuator setting in db.
TP = Thermophone.
CT = Condenser transmitter.

With the switches in Fig. 8 in position "A," the alternating current input to the thermophone and the amplifier gain are adjusted to obtain a convenient voltmeter deflection (V_{VM}). Then:

$$V_{VM} = I_1 R P E_T A.$$

Operating the switches to position "B" and manipulating the attenuator to attain the same voltmeter deflection as before we obtain

$$V_{VM} = V_0 A \times 10^{(-0.05N)} = I_1 R_1 A \times 10^{(-0.05N)}.$$

Then:

$$E_T = \frac{R_1 \times 10^{(-0.05N)}}{RP}.$$

The moving coil receivers are also calibrated in a sealed chamber in an atmosphere of hydrogen by an arrangement shown in cross-section

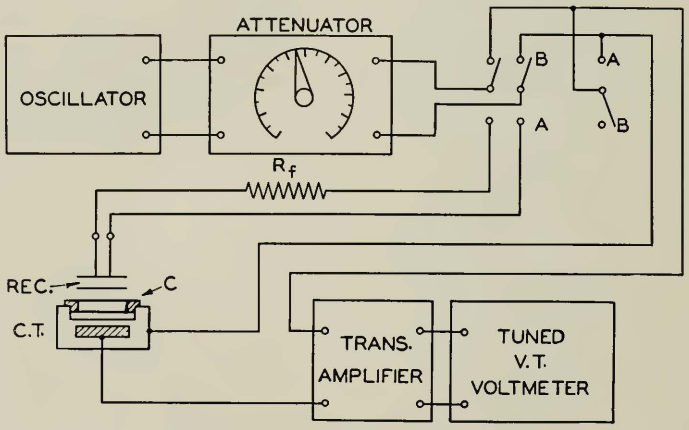


Fig. 9—Circuit for calibrating receiver.

in Fig. 7. A condenser transmitter, calibrated in the above manner, is actuated by the sound output of the receiver under test driven by the oscillator. The process involved is as follows: Referring to Fig. 9.

V_{VM} = Voltage across voltmeter.
V_0 = Voltage across attenuator input.
V_R = Voltage across receiver terminals

$$= \frac{V_0 R_R}{R_R + R_f} \times 10^{(-0.05N_a)}.$$

R_R = Impedance of receiver plus cord.
R_f = Fixed resistance.
P_R = Pressure developed by receiver per volt across receiver terminals.
E_T = Voltage developed by condenser transmitter per bar (one dyne/cm^2.).
A = Ratio of voltage delivered to 600-ohm load by the transmitter amplifier to voltage impressed in series with condenser transmitter.
N_a = Attenuator reading with switches on "A."
N_b = Attenuator reading with switches on "B."
C = Coupler (Fig. 7).
CT = Condenser transmitter.

With the switches indicated in Fig. 9 in position "A" the attenuator is adjusted to position N_a to produce a convenient deflection on the voltmeter (V_{VM}). Then:

$$V_{VM} = V_0 \frac{R_R}{R_R + R_f} P_R E_T A \times 10^{(-0.05N_a)}.$$

Operating the switches to position "B" and adjusting the attenuator to some position N_b to reproduce the above voltmeter deflection V_{VM} we obtain:

$$V_{VM} = V_0 A \times 10^{(-0.05N_b)}.$$

Then:

$$P_R = \frac{(R_R + R_f) \times 10^{0.05(N_a - N_b)}}{R_R E_T}.$$

The calibrations of the purely electrical elements of the circuit are made by measurements of input, output, and impedance.

Fig. 10 shows, for particular amplifier adjustments which are discussed below, the frequency response characteristics of the reference transmitter, reference receiver and the complete reference system with 0 db in the line. The characteristic of the reference transmitter and also of the reference receiver, in each instance, is that of the instrument and associated amplifier combined. However, as the frequency response of each of the amplifiers is uniform within 2 db from about 50 to 10,000 cycles per second, the curves shown are essentially the calibrations of the instruments determined as described above. The primary purpose of these characteristics is to show the performance of these elements for definitely specified physical conditions. Consequently, no corrections are included in these curves for the effect on

the sound field of the speaker when talking into the condenser transmitter. Neither have corrections been applied for the effect of leakage between the listener's ear and the earpiece of the receiver nor

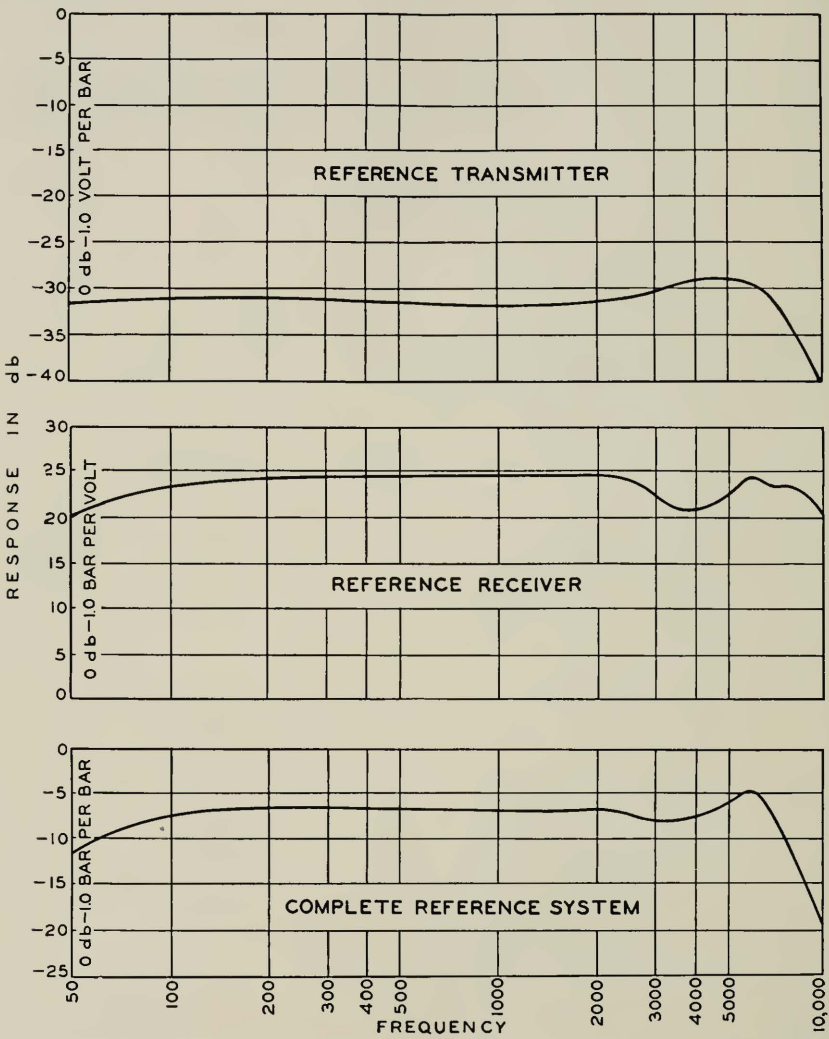


Fig. 10—Response characteristics of reference transmitter, reference receiver and complete reference system with 0 db in the line.

for the difference between the volume enclosed in the sealed chamber when making the calibration as compared with the volume enclosed in the ear canal when the receiver is held to the ear. The reference

line, consisting of balanced resistance networks, has a uniform frequency response characteristic. The characteristic of the complete system, therefore, except as its level is affected by the attenuation in the reference line, is that of the reference transmitter and reference receiver combined, there being no reflection effects at the junctions of the reference line with either the reference transmitter or receiver.

Fig. 11 shows the effect of introducing distortion networks in both the reference transmitter and reference receiver. The distortion net-

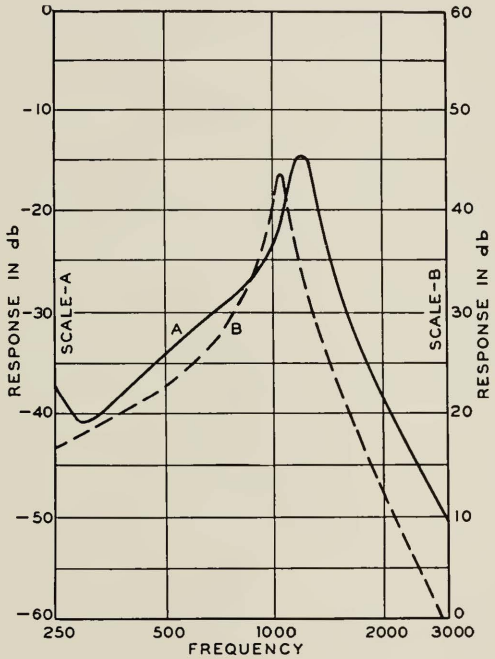


Fig. 11—Response characteristics of reference transmitter and receiver with distortion networks.
A = Reference transmitter—0 db—1.0 volt per bar.
B = Reference receiver—0 db—1.0 bar per volt.

work introduced in the reference transmitter simulates the distortion in the transmitter, station set and cord circuit, of the standard cable reference system. Similarly, the distortion network introduced in the reference receiver simulates the distortion of the receiver, station set and cord circuit, of the standard cable reference system. These two networks are designated, respectively, Transmitter Distortion Network No. 1 and Receiver Distortion Network No. 1. The distortion of any transmitter, receiver, or station set may be similarly simulated

by the design and introduction in the master reference system of a proper network. The advantage of such a procedure will be evident to those familiar with the difficulties of making voice volume balances between instruments of widely different frequency characteristics. The volume efficiency of any instrument can also be reproduced by adjustment of the controls in the amplifiers of the reference transmitter or receiver as the case may be. The results of voice volume balances made with the distorted reference system can be referred to the undistorted master reference system by applying a correction for the volume effect of the distortion network in the master reference system, this effect being determined independently as a rating of the distortion network.

Experience has shown that amplifiers properly designed and constructed remain essentially constant with time. Such changes in gain as may occur because of replacement of vacuum tubes or other apparatus may be readily compensated for by adjustments of potentiometers provided for this purpose. Condenser transmitters are affected to some extent by variations in temperature and barometric pressure. The magnitude of variation at any frequency between 50 and 10,000 cycles per second from these causes may be as much as 2.5 db, although under normal operating conditions such as those experienced in buildings in this climate, the variation is usually not more than 1 db. Since changes of this character are gradual in nature, their magnitude can be readily determined by a thermophone calibration. Corrections for any change in sensitivity of the instrument may then be made by adjustment of the controls in the amplifier associated with the condenser transmitter to maintain the proper gain in the reference transmitter. Similarly, any variations in the response of the moving coil receiver may be compensated for by adjustment of the gain controls in the amplifier associated with this receiver. The magnitude of the variations, at any frequency between 50 and 10,000 cycles per second, in the moving coil receiver may be as much as 3 db although usually variations of less than 1.5 db are observed. The calibration of the condenser transmitters, and indirectly of the moving coil receivers, is dependent upon the gold leaf thermophone, whose pressure characteristic is computed from physical measurements. Results obtained with thermophones can be held within about 0.5 db of the average obtained by using a group.

Comparison of Master Reference System and Standard Cable Reference System

Since the master reference system is to replace the standard cable reference system for volume ratings, the two systems have been compared by means of voice volume balances so that data obtained in the future may be directly comparable in this respect with those obtained in the past. The respective elements of the two systems, as well as the systems as a whole have been compared. The station set and cord circuit at the sending end of the standard cable system are taken as comprising the transmitting element of that system and, likewise, the corresponding apparatus at the listening end as the receiving element. In these measurements the reflection gain at the junction of the standard cable line and the output terminals of the transmitting element of the standard cable reference system has been taken as part of the transmitting efficiency of that element. Similarly, the reflection gain at the junction of the standard cable line and the input terminals of the receiving element of the standard cable reference system has been taken as part of the receiving efficiency of this element.

In the voice calibration of the master reference system to determine the adjustments which make this system equivalent on a volume basis to the standard cable reference system, eight series of voice tests were made. The purpose of the first five series of these tests was to determine the adjustments which make the master reference system with transmitter distortion network No. 1 and receiver distortion network No. 1 inserted in their respective elements, equivalent on a volume basis to the standard cable reference system. The purpose of the sixth and seventh series of tests was to determine the volume effects of the insertion of these distortion networks in the reference transmitter and receiver. The eighth series of tests was a direct comparison of the master system without distortion networks and the standard cable system, and so serves as an overall check on the determinations of the preceding tests.

In the first series of tests a comparison was made of the reference receiver with its distortion network and the receiving element of the standard cable system by interchanging the two in the standard cable system, with 24 miles of artificial cable in the line. The receiving element of the master system was adjusted so that its sound output was judged to be equal for this condition to that of the receiving element of the standard cable system. In the second series of tests a similar comparison was made of the reference transmitter with its distortion network and the transmitting element of the standard

cable reference system. These transmitting elements were connected in turn to a 24-mile standard cable line terminated by the reference receiver with its distortion network, the receiver being adjusted in accordance with the results of the first series of tests. In the third and fourth series of tests the master reference system with the distortion networks in both the transmitter and receiver was compared with the standard cable reference system, the line of the master system serving as the adjustable element. These two series of tests were similar except that 24 miles of standard cable were used in one series of tests and 14 miles of standard cable in the other. The adjustments for the reference transmitter and reference receiver in these tests were those determined from the first and second series. From the results obtained in the third and fourth series of tests, the magnitudes were determined of the reflection gain at the junction of the standard cable line with either the reference transmitter or the reference receiver, and of the volume equivalent of 1 mile of standard cable in terms of db. The reference transmitter and reference receiver, each with its distortion network, were then readjusted in accordance with these data. The fifth series of tests served as a check on the results of the previous four series. In this fifth series, the master reference system with transmitter and receiver distortion networks and 24 db in the line was compared with the standard cable reference system with 24 miles of standard cable in its line.

In the sixth series of tests the adjustment of the reference receiver without distortion was determined to make it equivalent on a volume basis to the reference receiver with distortion. The reference transmitter with distortion and the reference line with 24 db formed the rest of the system during these comparisons. In the seventh series of tests the sound output of the master reference system without distortion was compared to the sound output of this same system with distortion networks in the reference transmitter and reference receiver, the reference transmitter being adjusted to obtain a balance. During these tests 24 db was kept in the reference line. The final series of tests, serving as a check on all of the above determinations, consisted of a comparison between the master reference system without distortion and the standard cable reference system, the line of the master system being adjusted in making this comparison.

In making voice tests to determine the above settings, the speaker, whose position with respect to the transmitters was kept constant, called standard testing sentences in a conversational tone. A given intensity of calling was maintained by watching the deflections on a volume indicator connected across the output of the transmitting

element. An observer, located in a quiet room removed from the systems or instruments under test, determined when the two systems under comparison gave output sounds of equal loudness. Each testing team, consisting of a speaker and observer, made 12 balances. In order to attain a suitable precision in the final results of these tests, a large number of testing teams were used, the number being largest for the tests where the difference in the quality of the output sounds of the two systems under comparison was greatest. For example, for the first series of tests, six teams were used, for the fifth 25 teams, and for the eighth series, where the master reference system without distortion was compared with the standard cable system, balances were made with 37 teams. In all, over 2,000 individual balances were made. The standard deviation of the determination for each of the first five series is of the order of 0.5 db and for each of the last three series about 1 db.

The response characteristics of the transmitter and receiver elements of the master reference system, when adjusted on the basis of the results of the voice tests, to be equivalent on a volume or loudness basis to the corresponding parts of the standard cable reference system, are shown in Fig. 10. The mean values weighted from the standpoint of importance for volume are 0.027 volt per bar for the reference transmitter, 16 bars per volt for the reference receiver and 0.43 bar per bar for the complete master reference system with 0 db in the reference line. Further consideration is being given to the values of these response characteristics of the reference transmitter and receiver to be adopted as standards.

The response characteristics of the reference transmitter with transmitter distortion network No. 1 and of the reference receiver with receiver distortion network No. 1, when these elements are adjusted on the basis of the above voice tests to be equivalent on a volume or loudness basis to the corresponding parts of the standard cable reference system, are shown in Fig. 11.

Application of the System

The results of articulation tests over the master reference system when adjusted for optimum volume are practically equivalent to those obtained in direct air transmission in a quiet room. This system and replicas of it, are particularly adapted for use in making articulation studies, since they provide an approximately ideal system with which the loudness of the output sounds can be varied distortionlessly over a wide range and in which distortion networks of various types and controlled amounts of noise can be introduced. In this

way the effects on articulation of various kinds of physical performance of a telephone circuit can be investigated.

The master reference system itself will be used chiefly for the important work of rating working standard systems and instruments. These working standards can be simpler than the master system and can be provided in any number required to handle the rating of commercial circuits and apparatus.

The working standard may be of several forms. It can be similar to the master reference system, simplified in its detailed construction but capable of calibration by the means employed for calibrating the master reference system. Such a system would probably find employment in laboratories and factories where the volume of testing is sufficient to justify the use of such apparatus. Another form may include electrostatic or electrodynamic instruments which are not capable of being measured by the calibration equipment of the master reference system. A third form which the working standard may take is that involving the use of transmitters, receivers and station sets such as have been used in the standard cable reference system. These latter types of working standards can be calibrated by volume comparisons with the master system or with the first type of working standard. They will find their chief field of usefulness at such points as shops for the repair and recovery of station apparatus, where the volume of work is not sufficient to justify the expense involved in maintaining more elaborate working standards.

European Master Reference System

In Europe the recommendation of technical standards for telephony is a function of the Comité Consultatif International des Communications Telephoniques a Grande Distance (C.C.I.), which is composed of representatives of the various European telephone administrations. In 1926, at the invitation of the C.C.I., representatives of the Bell System met in London with a committee appointed by the C.C.I. to consider the adoption of a transmission reference system. This committee recommended that the C.C.I. adopt as their master reference system a system essentially the same as the one described in this paper, and that such a system, which would be a replica of one in New York, be installed in Paris in the laboratory of the C.C.I. and be known as the European Master Reference System. This recommendation was adopted by the C.C.I.

Subsequently, some improvements were made in the system, and two duplicate systems, each with its associated calibrating apparatus, have been constructed. One of these is now in the Bell Telephone

Laboratories in New York and the other in the laboratory of the C.C.I. in Paris. The C.C.I. further recommended that primary and working standard systems, used in the telephone administrations adhering to the C.C.I., be calibrated in terms of the Master Reference System. The establishment of these two master systems insures the use of a common base line for the expression of transmission standards, and for the ratings of the transmission performance of telephone circuits in the two continents where the telephone system has had its greatest development.

Shielding In High-Frequency Measurements [1]

By J. G. FERGUSON

In this paper the purpose and usefulness of shielding in high-frequency measurement are outlined. General principles of electrostatic shielding are developed as applied to simple impedances and to networks of impedances, particularly to bridge networks. Practical applications of these principles to the shielding of adjustable impedances, and in the construction of actual bridge circuits, are described.

SHIELDING of high-frequency measurement apparatus has for its immediate object the control of certain electromagnetic and electrostatic couplings unintentionally introduced in the usual high-frequency circuit. These couplings are represented by stray admittances between the various parts of the system, either direct or to ground, and mutual impedances resulting from stray magnetic fields. In general, the control of these couplings is exercised for the purpose of attaining an accuracy of test that cannot be obtained so readily in other ways.

When we speak of electromagnetic and electrostatic coupling, it should be understood that these are simply component parts which together make up the total coupling which exists. They cannot be considered as existing independently of each other, and we cannot consider the shielding for one of these components without taking into consideration the effect on the other.

However, for the frequencies and impedances ordinarily used in communication work, at least, this interdependence is small enough to allow us to consider the shielding problem for each type by itself, without getting into practical difficulties due to this connection. As a result, two types of shielding which are known as electromagnetic, and electrostatic shielding have been developed.

It may be argued that by extensive separation of the physical parts of circuits and apparatus, any couplings can be decreased in value and in consequence errors caused by them can be reduced, thus eliminating any need for shielding. But there are obvious limits to the extent to which this method can be employed practically. In the case of electrostatic coupling to ground, it is scarcely of any value, and in any case excessive separation of the parts of a circuit introduces other errors due to the length of the wiring involved. Accordingly, it is usually necessary, where the maximum accuracy is desired, to have recourse to shielding.

[1] Presented before the A. I. E. E. Summer Convention, June 24–28, 1929.

The principles involved in the application of electromagnetic and of electrostatic shielding are quite different. In the case of electromagnetic shielding, the methods used have for their object the elimination of all couplings from the unit shielded to all other apparatus; thus, if perfect shielding were possible, the unit would have no coupling to any other parts of the circuit. It is never possible to accomplish this in the case of electrostatic shielding. Any electrical apparatus will have electrostatic coupling to any other apparatus in the vicinity and particularly to ground. The addition of shielding always introduces additional electrostatic coupling from the apparatus to the shield and the shield usually has more coupling to other equipment and to ground than the apparatus had before shielding it. Consequently, the principles of electrostatic shielding are mainly a matter of controlling this coupling in such a way that it has the least harmful effect in the circuit even at the expense of increasing its actual magnitude, rather than a matter of eliminating it entirely. For this reason electrostatic shielding requires much more extensive consideration and this paper is, therefore, devoted mainly to it, the principles of electromagnetic shielding being covered only briefly.

Principles of Electromagnetic Shielding

The necessity for electromagnetic shielding is limited practically to wound apparatus such as coils and transformers. It may be reduced to a minimum by using high permeability core material wherever possible in coils and transformers, and by using some form of closed core such as the toroidal type. By these means stray fields may be reduced to a relatively low figure. However, there are cases where the remaining coupling may be objectionable and it is then necessary to use shielding to reduce still further the amount of these stray fields.

Two types of shielding may be used. A high permeability material may be used for the purpose of short circuiting the stray field. The principles of this method of shielding are described fully in another paper and will not be considered further here.

In the case of air-core coils which are often of the solenoidal type, since the advantage of using the toroidal form is less in this case, and for coils used at very high frequency where heavy magnetic material is not so effective, shields of non-magnetic material may be used to confine the field by the effect of eddy currents. For these shields, a material of high conductivity is used, usually copper, and the principal consideration is the spacing of the shield from the coil rather than the thickness of the shield itself.

There is always a loss in efficiency due to the losses in the shield, and this loss is greater, the closer the shield is placed to the coil, that is, the stronger the field in which the shield is placed. However, even with solenoidal air-core coils very effective shielding may be attained by moderately thick copper shields spaced about the distance of a diameter from the coil.

Principles of Electrostatic Shielding

In both theory and practice all measurements assume that between different terminals or junction points of the system there are impedances having values known to a degree of definiteness consistent

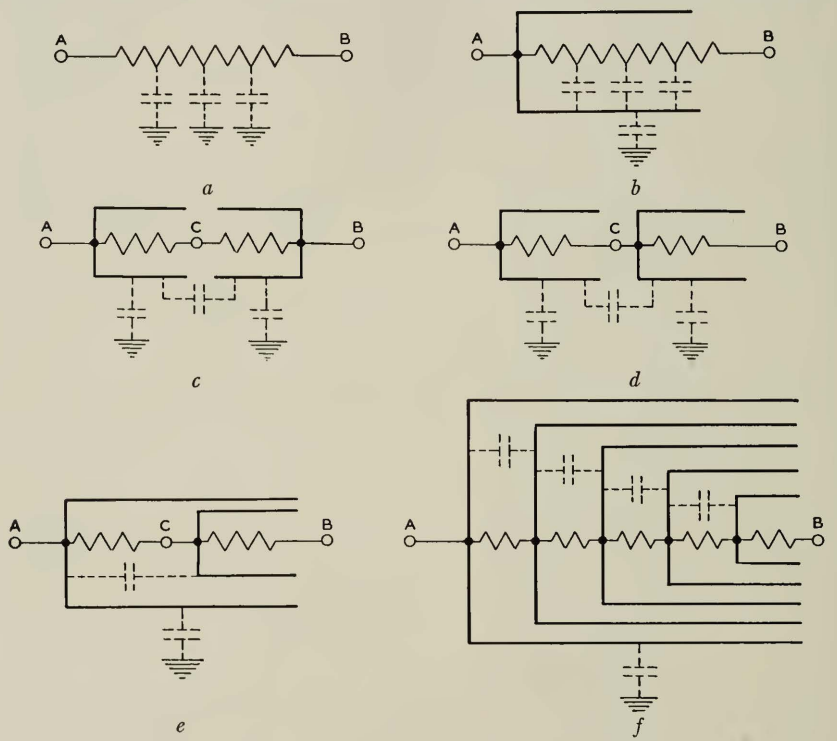


Fig. 1—Methods of Shielding Series Impedances.

with the accuracies sought in the test. In an unshielded circuit it will generally be the case that the elements connected by the various terminals or junction points will not provide impedances so definitely known or in other words will not carry all of the current flowing between the points in question.

Resistors

In the case of the simple resistor such as pictured schematically in Fig. 1*a* there are admittances from different parts of the conductor

to other parts of the whole system and in particular to ground. These, of course, act to modify the effective impedance between the terminals and as they vary with the location of the resistor, the result is that its effective impedance is variable and known only for the location in which it has been calibrated. One of the first objects to be accomplished by shielding is to remedy this type of indefiniteness of value. This is done by mounting the elements within a shield of conducting material and in fixed space relation thereto, as shown in Fig. 1*b*. Thus, the circuit element has direct admittances only to the shield and as these are of fixed value the terminal to terminal impedance becomes independent of the location of the shielded element.

If, then, we connect the shield to any fixed point in the circuit element such as one terminal, all of the current transferred by the shield admittances passes to or from the circuit at this particular point. This concentration of admittance enables the ready evaluation of the effect produced by it when the element is used in conjunction with others in a complete measuring system. We may summarize all of this to form a fundamental rule of shielding, viz.: "the association of an element of a system with a shield so that all admittances from the element to other parts of the system or to ground are confined to one terminal."

If it is possible to connect such an element in a circuit so that the terminal to which the shield is connected is grounded, all variable admittances will be eliminated completely.

Series Impedances

In the case of two impedances in series as shown in Fig. 1*c*, shielding may be accomplished by connecting one shield to terminal A and the other shield to B. In addition to the effects described for a single impedance, there will then be admittance between the two shields which will depend on the position of the apparatus. This admittance is slightly more objectionable than admittance from shield to ground since, while we may ground either A or B, there will always be an admittance from one shield to ground which will be variable.

The shields may also be connected as shown in Fig. 1*d*, in which case the admittance between shields appears across the first impedance. Now if we extend the shield connected to A to include the other shield as shown in Fig. 1*e*, we have introduced a fixed admittance across AC and have variable admittances to ground from A. The admittance across AC is not objectionable in the case of a capacitor since it may be considered simply as an addition to it, but it has the effect of increasing the phase angle of a resistor and in the case of an inductor

it increases the effective inductance and resistance variation with respect to frequency.

If this combination of impedances can be grounded at A we have a complete system having no variable admittances. The principle may be extended to include any number of series elements, the effect being to place admittances across all of the elements but one, and to enclose the whole system in one outer shield. Such a system for five elements in series is shown in Fig. 1*f*.

Parallel Impedances

The shielding of parallel impedances is comparatively simple since any number may be shielded individually and the shielding all connected to the same point. In reducing the shielding of multiple impedances to the simplest form the question arises whether it is sufficient to include them in a single shield or whether in addition

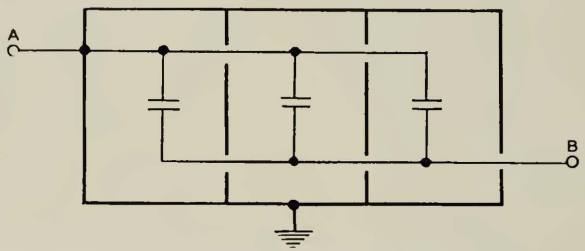


Fig. 2—Method of Shielding Parallel Impedances.

they should be shielded from one another. If they are not shielded from one another, there will be distributed admittances between them which may cause errors. Preferably each should be shielded individually. Fig. 2 shows such a shielding system for capacitors in parallel.

By following the procedure outlined above it is comparatively simple to apply shielding to any combination of impedances in series or in parallel in such a way that we will have all admittances to external conductors from the shielded elements concentrated at terminals or junction points of the system.

Circuit Shielding

In many cases it is impossible to connect the above combinations in a given circuit so that the outer shield is grounded. In such cases it is necessary to determine from the position of the network in the system the effect of admittances from the shield to other shields and to ground. To illustrate let us take the simple bridge circuit shown

in Fig. 3. Each of the four impedances constituting the arms may be considered as any combination of individual impedances. With the shields connected as shown, the total admittances are reduced to three; namely, between B and D and from B and D to ground. These admittances do not affect the bridge balance and, therefore, are not objectionable. However, if we add input and output circuits and follow the same system of shielding, we get the result shown in Fig. 4. In this case it is impossible to concentrate all of the admittances at B and D. Neglecting for the present the ground at D, we have added variable admittances from A to B, to D and to ground. The only way of overcoming this difficulty is to use double shielding as

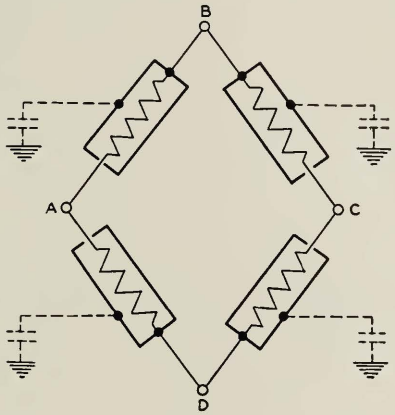


Fig. 3—Bridge Network Using Shielded Impedances.

shown, adding an outer shield to the impedance across AC and connecting it to D. This puts a fixed admittance across AD, but as we have not made any distinction between the four arms of the bridge, this admittance may generally be placed across an arm where it can be taken care of satisfactorily. If in addition we ground D, the admittances reduce to a single one from B to ground.

Admittance to Ground of Unknown Impedance

From the above it would appear that the general bridge circuit is susceptible of a simple complete solution, since the shielding shown in Fig. 4 is equally applicable to all cases. This would be true if the unknown impedance to be measured in the circuit had no admittance to ground. This is usually not the case. We generally have an additional requirement that the potential condition with respect to ground of the impedance during the measurement be defined in some way.

If the impedance can be connected across one arm of the bridge and its value is desired with one terminal grounded, the circuit shown is satisfactory. However, these are special conditions, and where the impedance to be measured forms only part of the total series impedance of an arm, or where the potential requirements are different, such as the requirement that the coil be measured with its terminals at equal potential to ground, the bridge shielding becomes a more serious problem.

In general, the question of selecting the most suitable system of electrostatic shielding for a specific test circuit, resolves itself into a determination of the most advantageous location of the admittances which, as described above, have been arranged to terminate at certain terminals or junction points. The facts which need to be taken into

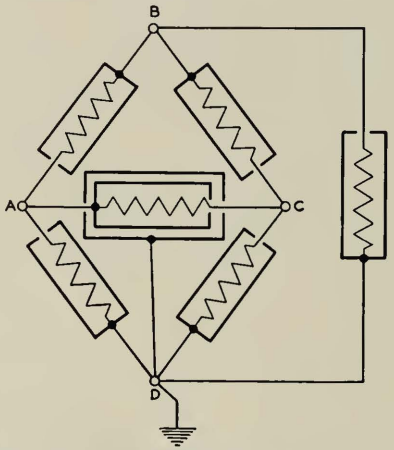


Fig. 4—Completely Shielded Bridge Network.

consideration are usually so varied that no general rules can be established. A few typical examples in which shielding is applied with considerable success will, therefore, be taken and the selection of suitable shielding for these circuits discussed.

Examples of Electrostatic Shielding

Adjustable Resistor

An adjustable resistor usually takes the form of a dial box in which there are from one to six dials arranged in series in decade formation. Each decade considered by itself is no more difficult to shield than a single resistor. The admittance of the shield, however, has a different effect at each step, which means that the phase angle varies with the

setting of the dial. If the admittance to the shield is small, this effect will not be very great and in any case it is always the same for a given setting and hence may be included in a calibration.

In shielding several decades in series, admittances between decades are introduced. Effects due to these admittances can be taken care of completely by the use of nested shields as already shown in Fig. 1*f*. For a resistance box of five or six dials, this type of shielding becomes

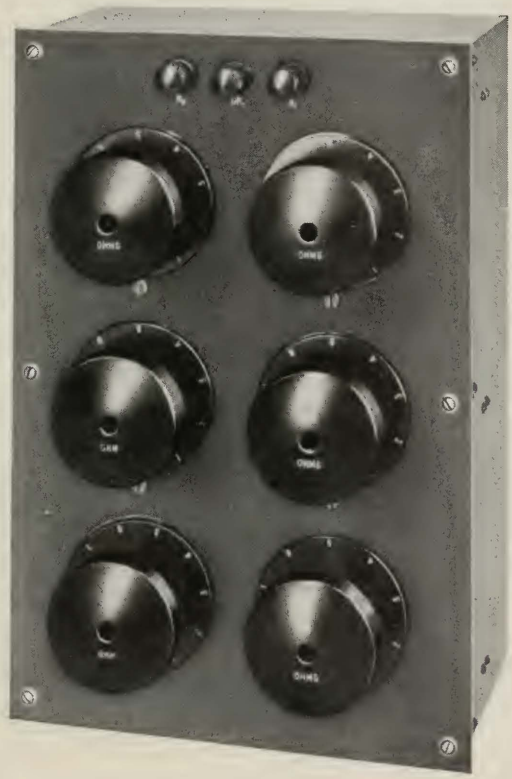


Fig. 5—Six-Dial Shielded Adjustable Resistor.

prohibitive from a size and cost standpoint and in consequence such shielding is usually not attempted. The use of a single shield for all decades of a resistor means that the impedance of two or more dial settings is not exactly equal to the sum of the impedances of each setting by itself. If the difference is appreciable the only alternative to the expensive type of shielding mentioned above is the use of a calibrated value for every combination of dial settings. This error in additions is smaller the lower the resistance, and usually may be neglected for values below 100 ohms.

The decades are ordinarily connected in series in ascending order of magnitude. The shield should always be connected to the low end, that is, to the terminal to which the decade of smallest value is connected. The reason for this may be explained as follows: the admittance from any decade to the shield is a function of the dimensions of the dial switch rather than of the resistance value. Consequently, all dials have approximately equal admittance to ground. It is desirable that the total admittance to ground be a minimum across the higher resistance settings. This requires that the low resistance dials be connected between the high dials and the shield, that is, the shield should be connected to the low end of the box.

In the actual constructon of such a resistor it is essential that the shielding be complete, particularly at the dials. Since the effect of the hands in operating the dials is more variable than any other coupling, it is of very little value to place an unshielded dial box in a metal shield which allows admittance from the hand of the operator to the circuit.

An example of a six-dial resistor in a complete single shield is shown in Fig. 5. The box itself and the panel are of metal, and the construction of the dials is such that there is a continuous metal shield between each dial head and the switch proper which it controls.

As stated earlier, the effect of the shield on the performance of the resistor is to increase the phase angle of the higher resistance values. In the case shown the admittance introduced by the shielding is of about the same value as the total admittance distributed in the coils themselves and from the coils to the switch parts.

Adjustable Inductor

The same considerations apply to an inductor as to a resistor except that on account of the larger physical size of the former, larger admittances are associated with it and for that reason it is usually necessary to use nested shields. Fig. 6 shows a standard inductor consisting of three decades and an inductometer using four shields. The three top panels have been removed showing the method of nesting the shields, and the construction used to bring the dial controls through the shields. The shielding of this unit is not complete in that the fourth shield (the outer one) does not extend over the top. This is allowable as the admittance from the third shield to ground is across the inductometer and any variation in it has little effect, particularly as the final balance is obtained with the inductometer, thus eliminating variations due to the hands of the operator, which occur in operating the other dials.

The admittance between shields is considerable but due to the method of construction the largest admittance is across the smallest inductance and there is no intershield admittance added across the highest decade. Accordingly, the effect is not as serious as might be thought at first glance.

In the case shown, the inductometer and the lowest decade are generally used only in combination with the higher dials and under these conditions they are not used at sufficiently high frequencies for the admittance shunting them to have much effect. The greatest effect of the admittance introduced by the shielding usually occurs when the second highest dial is used at a high frequency with the high dial set on zero. However, the admittance introduced across

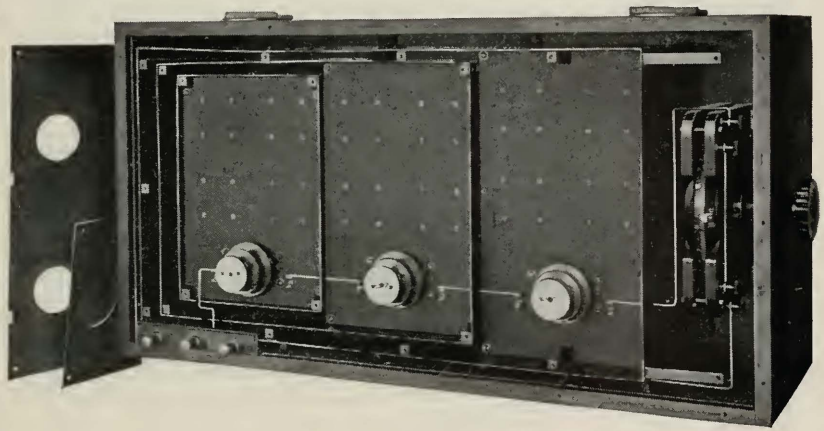


Fig. 6—Three-Dial Shielded Adjustable Inductor.

this decade is not appreciably larger than the distributed admittance across the coils. The shielding, therefore, does not limit the range of these inductance standards to any great extent.

Adjustable Capacitor

The units of an adjustable capacitor are practically always connected in parallel and the problem of shielding them is that of shielding a single capacitor. It is desirable to shield the decades from each other if the capacitances are small as this facilitates calibration and is easily effected. Where the capacitance is large, say over 10,000 $\mu\mu$f., this precaution is unnecessary. The capacitance introduced from the shield to the units has the effect of increasing slightly the value of each dial setting. The form of construction of the shielding is similar to that of the resistor shown in Fig. 5.

Bridge Circuits

The general principles of bridge shielding have been discussed by Campbell [2] and the equal ratio-arm comparison bridge has been discussed in detail by Shackelton.[3] The mechanical construction of the bridge itself exclusive of standards is simplified by the fact that there are comparatively few dials to be brought through the shielding.

This bridge with the standards described above may be used for a wide variety of measurements. A rather simple modification is the so-

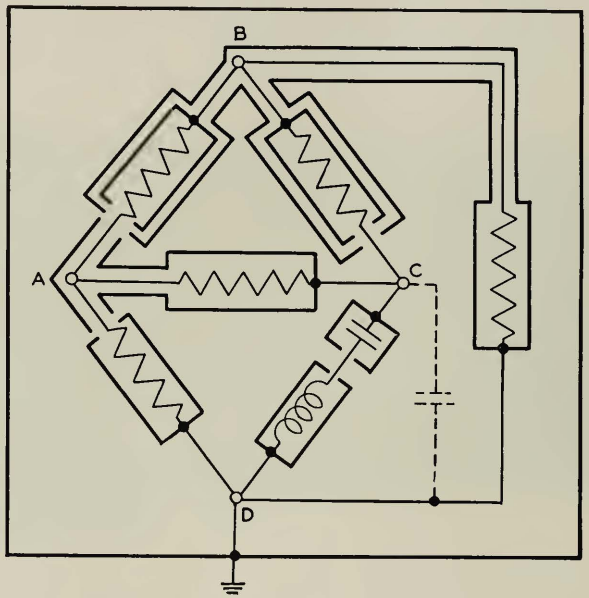


Fig. 7—Shielded Resonance Bridge Network.

called resonance bridge, in which the bridge unit is an equal ratio-arm comparison type, and a resistance is balanced in one impedance arm against a capacitance and an inductance connected in series in the other impedance arm. The balance is usually effected by adjusting the resistance and capacitance.

The shielded circuit of such a bridge is shown in Fig. 7. The capacitance from C to D introduced by the shielding may be compensated for in the usual way by the addition of an equal capacitance across AD. In this circuit the coil is usually measured under the

[2] G. A. Campbell, "The Shielded Balance," *Electrical World and Engineer*, April 2, 1904, p. 647.

[3] W. J. Shackelton, "A Shielded A-C. Inductance Bridge," *A. I. E. E. Journal*, Feb., 1927.

condition of one terminal at ground potential. Thus D is shown trapped to the ground shield. For this case, the shielding may be simplified considerably. Fig. 8 shows the mechanical construction of the combined resistance and capacitance standard used with the bridge unit for these measurements. The unit is shown with the top of the outer shield removed. The capacitance, in accordance with the shielding diagram, is double shielded, while the resistance requires only a ground shield.

Another bridge of the comparison type but using capacitance ratio arms is described by Kupfmuller [4] with a complete description of the shielding involved.

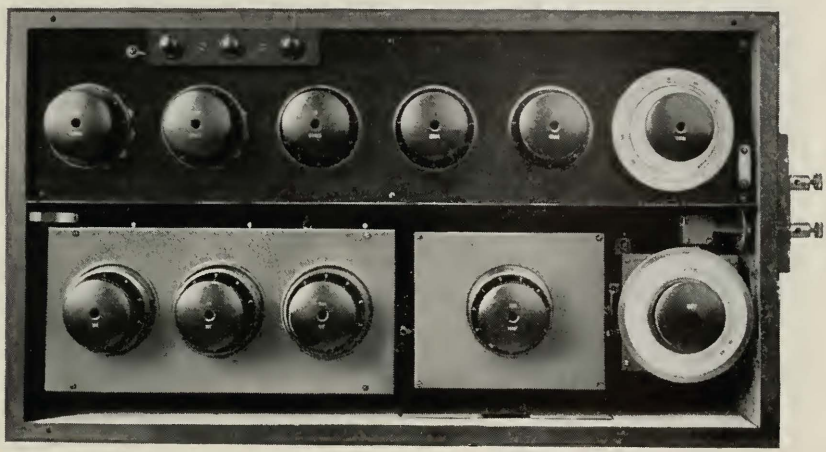


Fig. 8—Shielded Resonance Unit, Top Panel Removed.

A bridge of the general comparison type having self-contained standards of resistance and capacitance is shown with top panel removed in Fig. 9. It may be used for a wide variety of measurements such as series or shunt resonance, and direct comparison, by using the switches controlled by the small dials on the extreme left and right to throw the standards into various combinations.

Another bridge circuit which is interesting from the shielding point of view is the Owen bridge.[5] This is a skew bridge in which the ratio arms are 90° out of phase instead of being equal. For this reason any admittance introduced in one arm by the shielding cannot be compensated for by any equal admittance in another arm. Two

[4] Von K. Kupfmuller, "Über eine Technische Hochfrequenz Messbrucke," *Elek. Nach. Tech.*, September, 1925, pp. 263–270.

[5] D. Owen, "A Bridge for the Measurement of Self Inductance," *Proc. Phys. Soc. London*, October, 1914.

methods of taking care of this admittance may be used. It may be concentrated in the arm consisting of a capacitance and considered part of it, or an admittance across one impedance arm may be compensated for by a resistance across the other impedance arm. Both methods have been used in the construction of these bridges. A more detailed discussion of the shielding involved in this type of bridge is contained in a previous paper by the present author.[6]

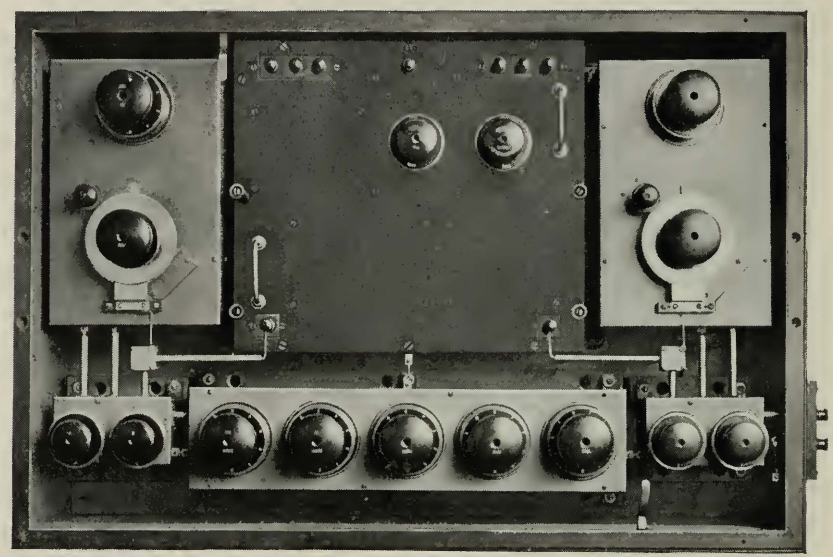


Fig. 9—Shielded Comparison Type Bridge, Top Panel Removed.

Details of Construction

We have discussed so far the admittance introduced by the shielding without going into details as to the form which this admittance takes although it has been broadly assumed that it is principally due to capacitance. Since it generally forms an integral part of the measuring circuit, it is obvious that as much consideration should be given to it as to the rest of the circuit. While the admittance is due essentially to capacitance, the necessary supports introduce a certain amount of conductance which causes some difficulty in obtaining compensation.

For instance, in the typical equal ratio-arm bridge circuit where the admittance across one arm requires compensation in the other arm, it is a simple matter to use an adjustable condenser for the compensation of capacitance. However, if the conductance is left un-

[6] J. G. Ferguson, "Measurement of Inductance by the Shielded Owen Bridge," *Bell System Technical Journal*, July, 1927, pp. 375–386.

compensated for it may cause considerable error, particularly in the measurement of high impedances at high frequencies. For this reason it is desirable that all shields be supported by insulating material of the highest quality such as hard rubber, glass or quartz and that only the minimum amount necessary for satisfactory mechanical support be used.

The wiring it will be noticed in Fig. 9 is shielded by brass tubing. This shielding is insulated from the conductor by means of bushings, only enough being used to insure that the conductor and shield do not change their relative positions with respect to each other. The insulating bushings used most generally are either hard rubber or glass beads.

Even after taking these precautions it has been found necessary for the highest precision work at the highest frequencies, to introduce a conductance compensator in the form of a small adjustable condenser in which the dielectric is an insulating material such as phenol fiber. By this means the amount of conductance in one arm may be varied to obtain correct compensation. The balance, once obtained, does not vary appreciably with frequency. Such an adjustment is used with the bridge shown in Fig. 9.

In bridge input and output transformers, which must generally be double-shielded, the shielding is rendered more difficult due to the requirement of a low conductance between shields. This demands a much more expensive construction than the simple requirement of complete electrostatic shielding.

Limitations of Shielding

Having discussed the uses and advantages of shielding, it may not be amiss to discuss briefly some of the limitations. As already brought out, the introduction of shielding always brings with it some additional admittance. Since this admittance is a function of frequency it is natural that shielding should introduce more trouble, the higher the frequency. However, it is also equally true that the stray admittances due to lack of shielding introduce more trouble, as the frequency is increased.

In general, it may be said that if shielding a circuit is found to have a definite advantage at moderately high frequencies, it will have an advantage up to the maximum frequency at which the circuit is used. The principles outlined already apply over the whole range of communication frequencies. Where shielding is found to result in frequency limitations, it is due to the added admittance introduced with it and not due to inherent defects in the principles involved.

The shielding may, in special cases, limit the maximum frequency at which the circuit will operate; but in such cases it can usually be taken for granted that even if the circuit would operate at higher frequencies without shielding, the accuracy of the results would be highly questionable. Examples of limitations of shielding may be given using the apparatus already described. Take the case of the resistor shown in Fig. 5. The admittance across the resistances results in objectionably high phase angles and an effective change in the resistances at very high frequencies. While this effect would be present even though no shielding were used, the shielding increases it and therefore limits the maximum frequency at which the apparatus can be used from the standpoint of this type of error. The same limitation occurs in the case of the inductance standards only it is more serious due to the large physical size of these standards. The exact type of limitation here is that the individual units increase in inductance due to the admittance across them to such an extent that the difference between them cannot be bridged by the next lower decade, thus rendering it impossible to obtain certain values of inductance by any dial combination.

In the case of a symmetrical bridge the principal limitation is the shunting effect of the admittance introduced across the impedance arms. This becomes so large that, at frequencies in the order of 100 kilocycles, difficulties are encountered in measuring the current through the unknown impedances by the method of measuring the total current input to the bridge. If the meter question is eliminated, the actual loss in sensitivity, which is the only other objectionable feature of this admittance, may be made of no serious consequence up to frequencies as high as 2,000 kilocycles. In all other respects, the shielding functions as satisfactorily at this frequency as at the ower frequencies.

Auxiliary Equipment

While auxiliary apparatus such as oscillators and detectors is not strictly speaking measuring apparatus, its operation is essential to the satisfactory operation of the measuring circuit and so a few words may be added regarding the shielding of this apparatus.

Provided they are separated sufficiently from the measuring circuit and from each other, there is no need to shield the oscillator or detector from the standpoint of operation of the measuring circuit. However, for maximum flexibility it is desirable that they be so constructed that no special precautions are necessary in placing them relative to the measuring circuit. If, as is usually done, the individual apparatus included in these circuits is adequately shielded it is only necessary to

place the completed equipment in an electrostatic shield to avoid all coupling to the measuring circuit or from one to the other. This is usually done by mounting the apparatus on a metal panel and placing it in a wood box with a sheet metal lining. By this means it is possible to place the auxiliary apparatus as close to the measuring circuit as desired, without introducing errors.

As far as the internal shielding of the auxiliary apparatus is concerned, the same rules hold as for other apparatus described. However, in the case of vacuum tube equipment, wherever any gain is introduced, coupling between parts of the circuit must be reduced in proportion to the gain introduced between these parts. This is usually accomplished quite readily by suitable mechanical design and may be insured by placing each stage in a separate grounded shield, although this is seldom necessary.

Conclusion

It has been impossible to go into very great detail in this brief paper on the subject of shielding. The attempt has been made, therefore, to outline a few general rules and to give representative examples of typical measuring circuits. It will be noted that the examples have been limited largely to the bridge circuit. This is because our experience has shown that this circuit is the most flexible and accurate over the whole of the frequency range over which precise impedance measurements have been made, and because the problems of shielding it are sufficiently difficult and varied to give satisfactory examples of the solution of rather complicated problems. The principles of shielding given have been found to apply equally well at all frequencies and it has been found that up to the maximum frequency at which precision measurements have been made, the shielding methods developed for use with moderate frequencies require practically no modification as the frequency is increased. Experience with measurements and measuring circuits up to 2,000 kilocycles makes it appear probable that when precision measurements are made at still higher frequencies, the shielded bridge circuit will continue to remain the most satisfactory measuring circuit.

Fatigue Studies of Non-Ferrous Sheet Metals[1]

By JOHN R. TOWNSEND and CHARLES H. GREENALL

The paper describes the development of a fatigue test machine for sheet metals and gives results of fatigue tests on five alloys of alpha brass, one alloy of nickel silver, one alloy of phosphor bronze and Everdur.

The results indicate that cold work raises the endurance limit but not proportionally to the increase in tensile strength produced by the same cause.

Micrographs are shown indicating that fatigue failure of the metals investigated is transcrystalline.

Dispersion hardening of alpha brass by nickel silicide increases the endurance limit.

The ratio of endurance limit to ultimate tensile strength of these alloys varies from .12 to .36 depending on composition, heat treatment, and cold work. These ratios are much lower than similar ratios for steel.

THE materials referred to in this paper are those non-ferrous sheet metals that are employed in electromechanical devices such as telephone apparatus and includes a wide variety of equipment, such as switches, relays, jacks, contact springs,[2] etc. These metals are employed principally in springs used for electrical contacting purposes. In many cases these springs have precious metal contacts welded to them and in other cases the metal itself is used for the contact. Many of these springs are subjected to millions of cycles of stress and it is important, therefore, that the endurance limit of these materials be known in order that apparatus may be designed which will endure for its required service life. Very little precedent has been established in the design of fatigue machines for the testing of sheet metals and it was necessary, therefore, to develop a form of fatigue machine especially suitable for these materials.

Another joint paper by one of the authors describes these non-ferrous metals and explains various methods of test and the commercial limits developed for specification purposes.[3]

Sheet Metal Fatigue Specimen

The specimen shown by Fig. 1 was designed to simulate in its major dimensions the normal size of the springs used in telephone apparatus. It will be noted that the design of the specimen provides a section of uniform stress for $\frac{3}{8}$ inch at approximately $\frac{1}{2}$ inch from the clamped end of the specimen. This is accomplished by

[1] Presented before A. S. T. M. Convention, June 24–28, 1929.

[2] "Telephone Apparatus Springs," by John R. Townsend: *Proceedings A. S. M. E.*, 1928; *Bell System Technical Journal*, April, 1929.

[3] "Mechanical Properties and Methods of Test for Sheet Non-Ferrous Metals," by J. R. Townsend, W. A. Straw and C. H. Davis, presented before A. S. T. M. Convention, June 24–28, 1929.

designing a cantilever beam that will have a uniform bending moment for part of its length. The dashed lines indicate the shape of a beam of uniform bending moment; the heavy lines show the final shape of the specimen and how the portion of uniform bending moment has been connected by means of fillets. This is essential in order to eliminate the possibility of the clamping stresses affecting the applied stress as would be the case if a uniform rectangular cantilever beam specimen had been employed. The specimen is deflected at a point ½ inch from the operated end or, in other words, where the hypothetical beam of uniform bending moment terminates.

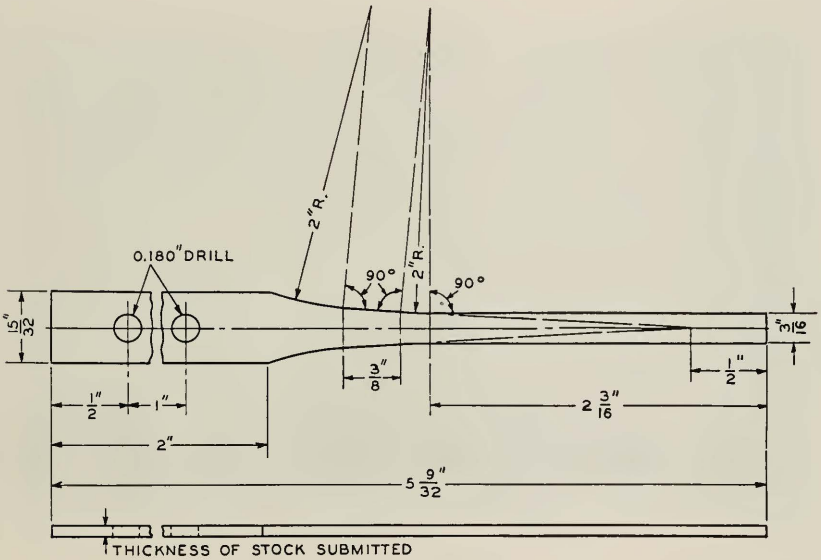


Fig. 1—Sheet Metal Fatigue Specimen.

The specimens are prepared by blanking rectangular samples which are clamped together and then cross milled with a form milling cutter in a manner similar to that described elsewhere for the preparation of tensile specimens.[4] The specimens were cut with the direction of rolling parallel with their length.

Sheet Metal Fatigue Machine

Referring now to Fig. 2, it is seen that the specimens (*S*) are clamped between phenol fiber blocks (*B*). This is done in order that reasonably rigid material will be provided and at the same time a material

[4] "Methods for Determining the Tensile Properties of Thin Sheet Metals," by R. L. Templin. *Proc. A. S. T. M.*, Part II, Vol. 27 1927.

sufficiently dissimilar to the metal to accomplish good clamping without scoring the surface of the clamped portion. Furthermore, the use of the phenol fiber blocks provides a means of automatically recording the breaking of the specimen since the specimen is insulated from the machine and may be employed to break an electrical monitoring circuit. The deflected end of this specimen is held between two fingers ($F_{1.2}$) which have a vertical cylindrical half section. The cylindrical portion is in contact with the specimen. This is necessary in order to compensate for the angular movement of the reciprocating arm (A) in relation to the specimen. One of these fingers (F_1) is fixed and the other is movable (F_2) but bears against the specimen with



Fig. 2—Sheet Metal Fatigue Machine.

a tension provided by compression spring (S_p). This permits some slight movement of the specimen in relation to the fingers which is necessary by reason of the change in free length of the specimen as the reciprocating arm moves backward and forward. The deflection of the specimen is determined in two ways, first by measuring the movement of the reciprocating bar and also by observing the specimen in operation by means of a stroboscope. In this way the static and dynamic deflection of the specimen may be measured and for all practical purposes, these have been found to be the same within the range of deflection and speed used in this investigation.

The speed of the machine is approximately 1,500 r. p. m. It is necessary to adjust the speed of the machine to the material under test since if this is not done the machine may be operated near the

TABLE I

AVERAGE CHEMICAL ANALYSIS OF MATERIALS

Material *	Alloy	Copper	Lead	Iron	Zinc	Nickel	Manganese	Tin	Phosphorus	Silicon	Combined Carbon
[1] High Brass		65.09	0.02	0.03	Balance			0.00			
[1] Alloy "G" Brass		71.73	0.02	0.03	"	0.01					0.018
[1] Nickel Silver		55.23	0.005	0.06	"	18.38	0.11				
[1] Phosphor Bronze		91.84	0.02	0.03	0.00	0.00		8.08	0.03		
[1] Everdur		96.00							3.00	1.00	
Hardened Brass	33	Balance			9.89	2.32				.57	
	34	"			19.89	2.37				.57	
	35	"			30.12	2.36				.66	

* No graphite was present in any of the materials.

[1] Material and Analysis furnished by C. H. Davis, American Brass Co.

TABLE II

PHYSICAL PROPERTIES

Material	Heat Treatment	B & S Nos. Hard	Tensile Strength Psi	P Limit Psi	Mod. of Elasticity Psi × 10^6	% El. in 2″	Rockwell Hardness * 1/16″ Dia. Ball Red Figures	Endurance Limit Psi	Ratio Endurance Limit Ult. Tensile Strength
High Brass.....	600° C. Anneal	0	46,600	13,000	14.5	56	16	12,000	25.7
		4	77,200	32,000		6	79	13,500	17.5
		10	95,000	30,000		2	87	15,000	15.7
Alloy "G" Brass..	600° C. Anneal	0	46,300			61	16	12,000	25.9
		4	81,600			6	84	18,000	22.0
		10	97,800			2	92	20,000	20.5
Nickel Silver.....		0	66,900			42	19	14,000	36.0
		4	98,700			2	69	18,500	17.4
		10	116,200	60,000	20	1.5	79	22,000	18.9
Phosphor Bronze..		0	59,700			67	11	21,000	35.2
		4	95,500			14	71	22,000	23.0
		10	124,800	55,000	15	2	84	24,500	19.6
Everdur.....	"Spring Temper"		80,000	26,000	12.4	22	91	24,000	30.0
Hardened Brass Alloy No. 33....	Quenched 800° C. aged 1 hour at 500° C.		90,000	44,500	19.8	14	86	14,000	15.5
Alloy No. 34....	Quenched 850° C. aged 1 hour at 500° C.		85,800	37,200	17.2	21.5	85	12,500	14.6
Alloy No. 35....	Quenched 800° C. aged 1 hour at 400° C.		85,400	38,000	16.5	28.0	79	16,000	17.3

* 100 kg. load for brass alloys and 150 kg. load used for nickel silver, phosphor bronze and Everdur.

natural frequency of vibration of the specimen and this may superimpose additional stresses upon it. This is determined by observing the operation of this specimen by means of a stroboscope and accurately setting the speed of the machine to a point where the vibratory motion of the specimen is uniform.

The machine has a capacity of forty specimens. Twenty specimens are tested on each end of the motor drive. The machine is statically balanced and is smooth in operation. It is customary to test at least five specimens of each material at each stress. The machine therefore, has a capacity of four alloys of five specimens each at two deflections. The practice of using five specimens for each deflection was adopted because it was seen from the results of previous experimenters in fatigue testing that a more accurate result might be provided by doing so. The capacity of the machine is sufficiently large to permit this being done.

Materials

The materials investigated consisted of five alloys of alpha brass, three of which had been hardened with nickel silicide and one alloy each of nickel silver, phosphor bronze and Everdur. The three alloys of hardened brass have been described previously.[5]

The chemical composition of these alloys is given in Table I and the tensile strength, proportional limit, modulus of elasticity, per cent elongation in 2 inches, Rockwell hardness and endurance limit are given in Table II. The heat treatments and amount of cold work expressed by number of B. & S. gauge reductions from standard anneal are shown also on Table II.

Fatigue Endurance Test

Method of Determining Stress in Specimen

The method of determining the stress in the specimen consists in clamping a specimen in the same manner as on the fatigue machine. The clamped specimen is mounted on the table of a Société Genevoise Star Comparator so that its large surfaces are parallel with the vertical plane, the axis being parallel to the table. By mounting the specimen in this manner its weight has no appreciable effect on its deflection. The stress in the specimen is determined from the load deflection curve obtained by applying dead weights ½ inch from the end of the specimen and measuring the amount of deflection for various units of load by means of a microscope mounted on the comparator in such a

[5] "Heat Treatment and Mechanical Properties of some Copper-Zinc and Copper-Tin Alloys Containing Nickel and Silicon," by W. C. Ellis and Earle E. Schumacher. *Proc. A. I. M. M. E.*, 1929.

manner that the rectangular coordinates of the deflected specimen may be read on the micrometer heads. Readings were taken to 1 micron. The stress per unit deflection is then calculated from the formula

$$S = \frac{6Pl}{bd^2},$$

where S = stress in pounds per square inch,
P = load in pounds,
l = length in inches,
b = width in inches,
d = thickness in inches.

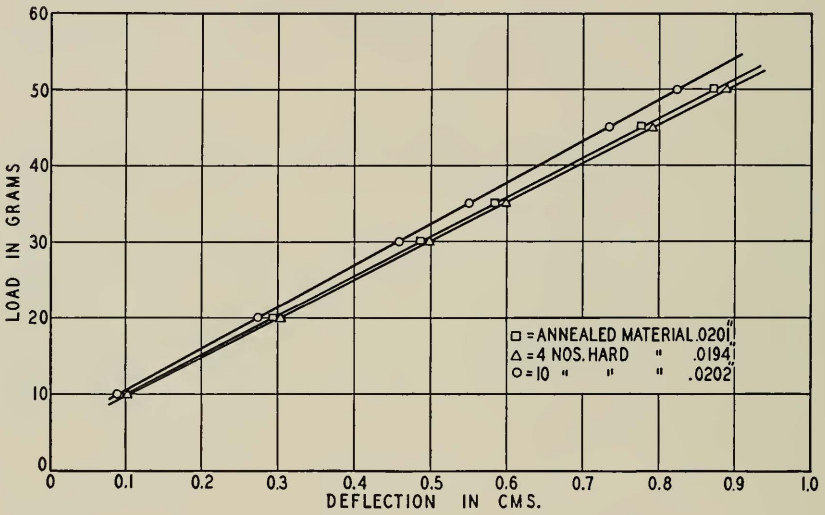


Fig. 3.—Relation of Load to Deflection, Alloy G Brass, No. 24 B. & S. Gauge.

The curve shown by Fig. 3 for alloy "G" brass sheet gives the load for uniform deflection of the specimen upon the fatigue machine. The various stresses are then obtained by varying the amount of deflection of the end of the specimen by adjustment of the roller bearing that operates the reciprocating bar.

Fatigue Endurance Results

The curves shown on Fig. 4 are for high brass sheet annealed and rolled four and ten numbers hard. Figs. 5, 6, 7 and 8 give similar results for alloy "G" brass, nickel silver, phosphor bronze and Everdur respectively for annealed material and rolled four and ten B. & S. gauges numbers hard. Fig. 9 gives the fatigue results for the alpha

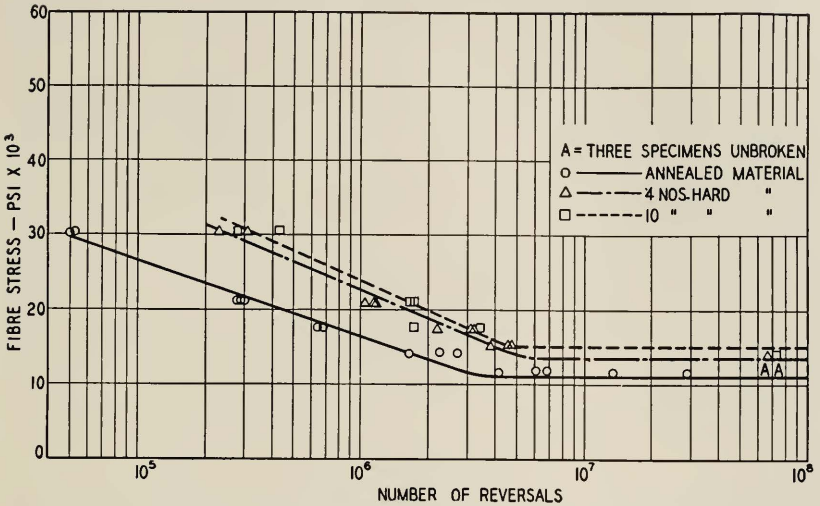


Fig. 4—Relation of Fibre Stress to Reversals, High Brass Sheet, No. 24 B. & S. Gauge.

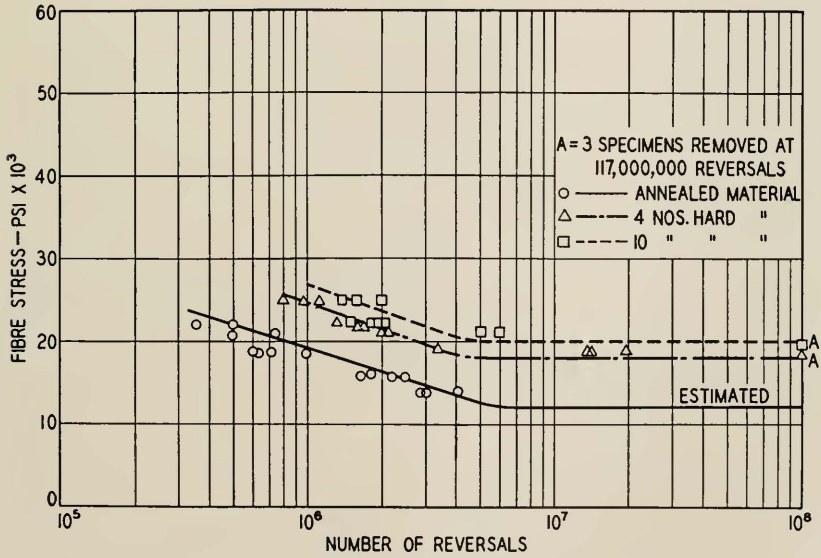


Fig. 5—Relation of Fibre Stress to Reversals, Alloy G Brass Sheet, No. 24 B. & S. Gauge.

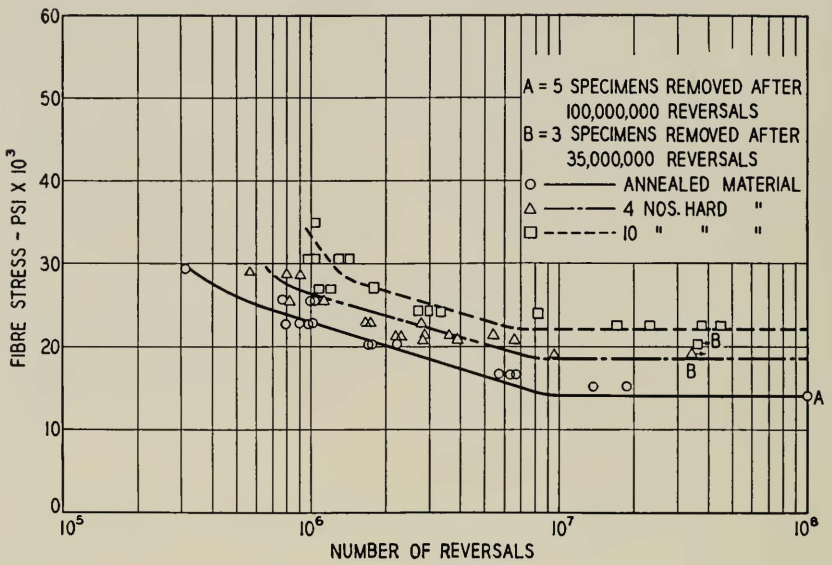


Fig. 6—Relation of Fibre Stress to Reversals, Alloy B Nickel Silver, No. 24 B. & S. Gauge.

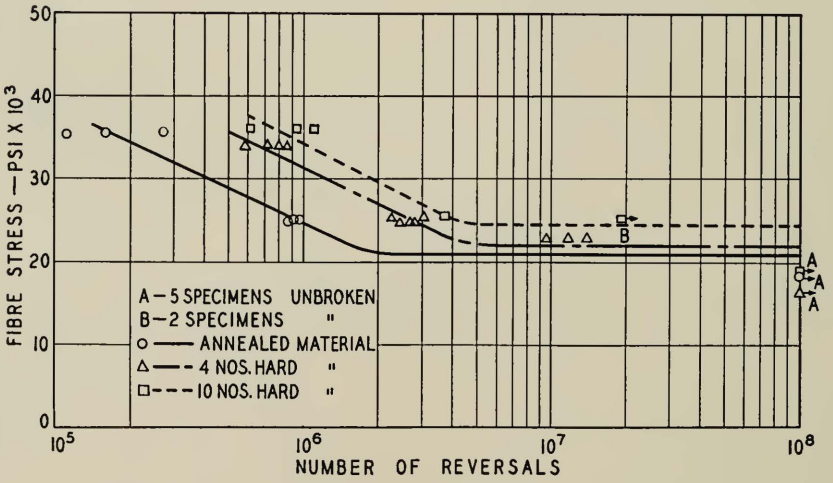


Fig. 7—Relation of Fibre Stress to Reversals, Alloy C Phosphor Bronze, No. 24 B. & S. Gauge.

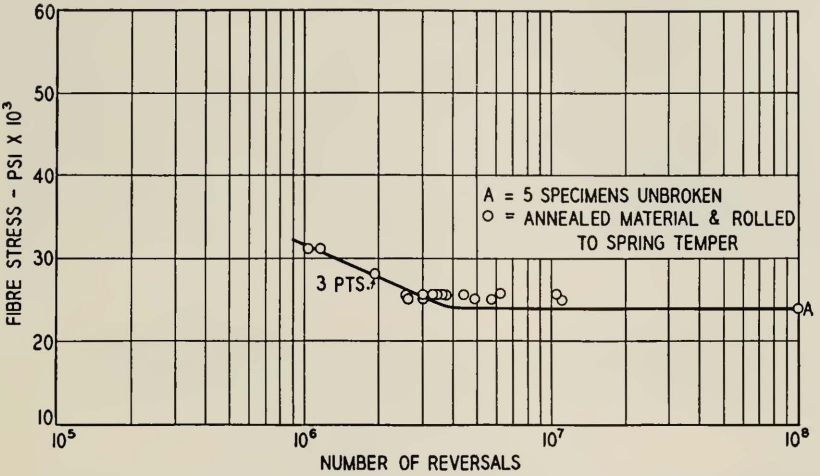


Fig. 8—Relation of Fibre Stress to Reversals, Everdur Alloy Spring Temper, No. 24 B. & S. Gauge.

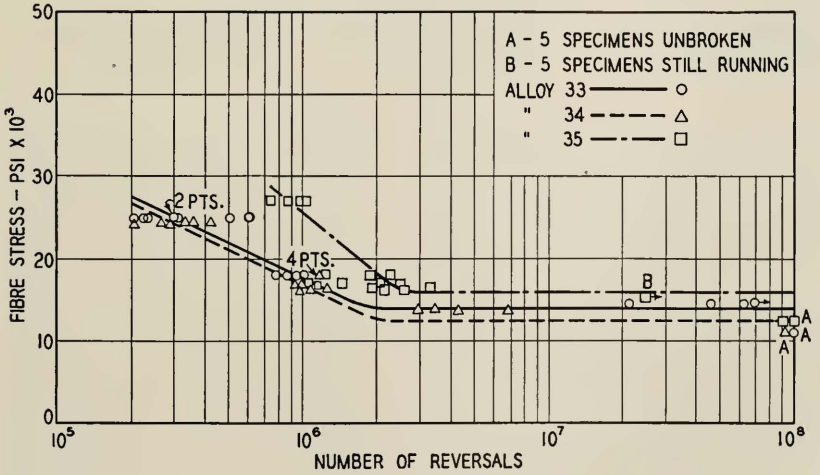


Fig. 9—Relation of Fibre Stress to Reversals, Dispersion Hardened Brasses, Alloys Nos. 33, 34 and 35, No. 24 B. & S. Gauge.

brasses hardened by nickel and silicon otherwise mentioned as alloys Nos. 33, 34 and 35 respectively.

Microstructure

Fig. 10 shows a photograph of a number of broken specimens. The regularity of the break is shown and in every case occurs within the uniformly stressed area. Photomicrographs shown by Fig. 11 are typical of the various alloys. In these cases incipient cracks are revealed within the uniformly stressed area. It is seen that these cracks are, without exception, transcrystalline and there

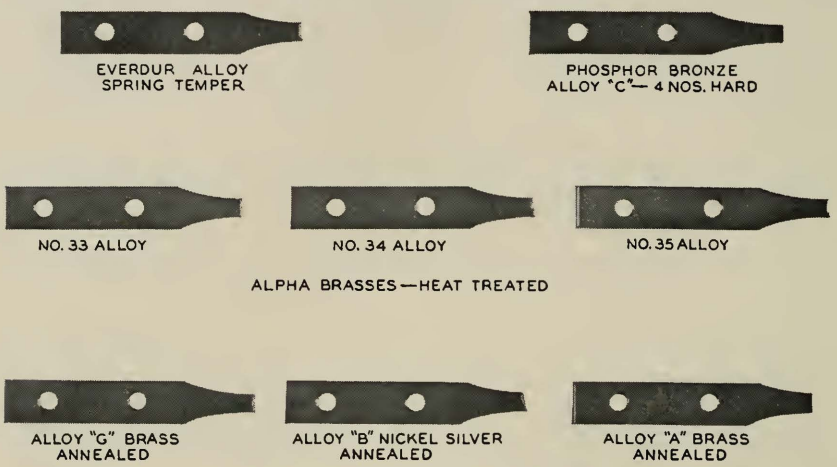


Fig. 10—Typical Broken Fatigue Specimens.

appears to be no distortion of the metal adjacent to the fractures. With regard to the typical structure of the hardened brass alloys reference is made to the previous paper.[6]

Discussion

Examination of the results shown by Table II reveals that the ratio of endurance limit to tensile strength for sheet non-ferrous metals is much lower than that reported for steel rod.[7] These ratios reported for plain carbon and alloy steels in all heat treatments vary from .35 to .67 averaging about .40 whereas for these sheet non-ferrous metals these ratios vary from .14 to .36.

[6] "Heat Treatment and Mechanical Properties of some Copper-Zinc and Copper-Tin Alloys containing Nickel and Silicon," by W. C. Ellis and Earle E. Schumacher. *Proc. A. I. M. M. E.*, 1929.

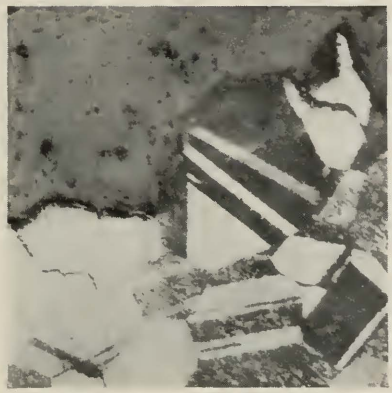
[7] "The Fatigue of Metals," by H. J. Gough, Scott Greenwood & Sons. Also "The Fatigue of Metals," by H. F. Moore and J. B. Kommers, McGraw Hill & Co.



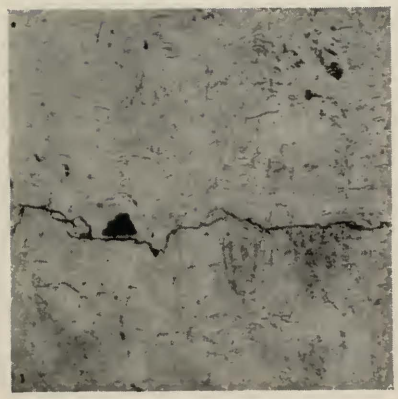
A



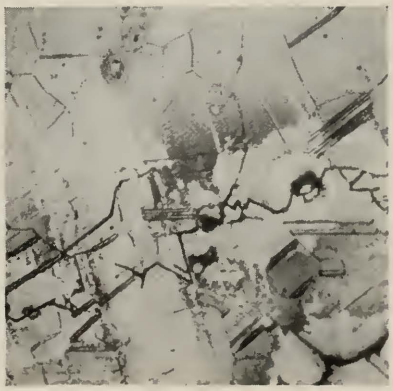
B



C



D



E

Fig. 11—Photomicrographs of Fatigue Specimens.

A. Nickel-Silver, 4 B. & S. Nos. Hard, Stress 28,750 Psi, 800,000 cycles, mag. 200.

B. Phosphor Bronze, 4 B. & S. Nos. Hard, Stress 25,500 Psi, 2,600,000 cycles, mag. 200.

C. Alloy G Brass, Annealed, Stress 15,750 Psi, 2,471,200 Cycles, Mag. 200.

D. Everdur, Spring Hard, Stress 25,500 Psi, 1,186,900 Cycles, Mag. 200.

E. Hardened Brass Alloy No. 33, Stress 18,200 Psi, 850,000 Cycles, Mag. 200.

Prepared by Miss A. K. Marshall

The improvement in the endurance limit due to cold rolling brass, nickel silver and phosphor bronze is not consistent with the increase in tensile strength produced by the same means. This is shown by Table II where in practically every instance material hardened by cold work shows a progressive decrease in the ratio of endurance limit to tensile strength.

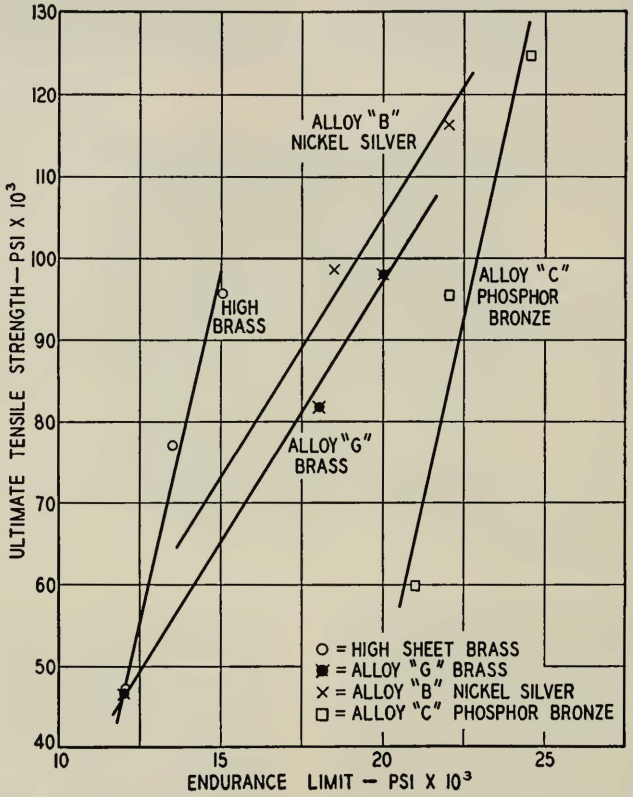


Fig. 12—Relation of Ultimate Tensile Strength to Endurance Limit, No. 24 B. & S. Gauge, Non-Ferrous Sheet.

Previous investigators have shown a close correlation between endurance limit and tensile strength for iron and steel. Fig. 12 fails to reveal such a general correlation for the sheet metals under test. For a particular alloy, however, there appears to be a progressive increase in endurance limit with increase in tensile strength due to cold work but the slope of the curves shown are widely different for the various alloys.

The results for alloy "G" brass and high brass show the effect of change in composition on the endurance limit. The greater improvement in fatigue endurance of alloy "G" brass due to cold work over high brass shows the superiority of this alloy for spring purposes. This was to be expected on the basis of its physical properties.[8, 9]

Alpha brass hardened by nickel and silicon shows considerable improvement in the endurance limit and also an improvement in the ratio of endurance limit to tensile strength. This offers a means whereby some non-ferrous alloys may be improved in this respect.

Attention is called to Fig. 6 which gives the stress cycle graphs for nickel silver in three tempers. It will be noted that the curves tend to turn sharply upward for the higher stresses. This indicates the effect of drastically overstressing the metal. The authors have observed this effect with other metals on the rotating beam machine. It seems that after a limiting stress value that the number of cycles to failure tends to become constant.

Conclusion

From the test results obtained on these non-ferrous metals it is seen that the fatigue endurance limit varies from approximately 12 to 36 per cent of the ultimate tensile strength, whereas the commonly accepted ratio of fatigue endurance to tensile strength of steel is in the neighborhood of 40 per cent of its ultimate tensile strength. In other words, it appears that the low endurance limit of these materials emphasizes the need for their careful selection for use as springs. High tensile strength or proportional limit are not sufficient guarantors that the material will perform satisfactorily in service.

Cold work raises the endurance limit but not in a manner proportional to the increase in tensile strength produced by the same cause. There is no correlation between tensile strength and endurance limit except for cold worked metal of a definite composition. For other compositions the correlation is different.

Precipitation hardening of alpha brass by nickel silicide increases the endurance limit.

The fatigue failure appears to be a result of a fracture across the crystals of the material. Photomicrographs are given showing incipient cracks that were developed in the uniformly stressed areas of the specimen.

[8] "Physical Characteristics of Copper and Zinc Alloys," Bassett and Davis. *Proc. Inst. Metals Div. A. I. M. M. E.*, 1928.

[9] The authors are indebted to Mr. L. E. Abbott for his assistance in obtaining laboratory data.

The curves given for fatigue endurance show the results for each specimen tested. The shape of the fatigue endurance curve for these metals is similar to those published previously for other metals.

A special form of fatigue machine has been developed to test sheet metals. This machine will accommodate forty specimens. Capacity is thereby provided so that as many as five specimens may be employed for each of four materials at two deflections under test.

An Application of Electron Diffraction to the Study of Gas Adsorption

By L. H. GERMER[1]

Under appropriate experimental conditions, electron scattering by a single crystal of nickel can give rise to diffraction patterns of four quite distinct types. We attribute one of these patterns to the space lattice of the nickel crystal, one to the topmost layer of nickel atoms, one to a monatomic layer of adsorbed gas atoms, and one to a thick layer of gas atoms. From these phenomena some conclusions concerning gas adsorption have been drawn. We have at hand a new and important method of crystal analysis.

IN the paper by Dr. C. J. Davisson and myself entitled "Diffraction of Electrons by a Crystal of Nickel," [2] we published a variety of information concerning the gaseous contamination of the surface of our diffracting crystal. This information was obtained from a study of the modifications produced by adsorbed gas in the electron diffraction pattern. We have subsequently succeeded in obtaining some additional facts concerning adsorbed gas from a further study of our original data. Along with the presentation of these new facts, I am taking this opportunity to publish in greater detail the data upon which our original conclusions were based.

In our Physical Review paper we showed that the interaction of a beam of electrons with a single crystal of nickel gives rise to phenomena which, in their most essential characteristics, are similar to the diffraction phenomena which would be observed if the beam of electrons of adjustable speed were replaced by a beam of X-rays of adjustable wave-length. The diffraction patterns produced by electron scattering are, however, substantially more complicated than would be the corresponding X-ray diffraction pattern. We showed that electron scattering can give rise to diffraction phenomena of four quite distinct types. The diffraction patterns of two of these types arise from the nickel atoms in the crystal lattice, while the diffraction patterns of the other two types have their origins in the layer of gas adsorbed upon the surface. The relative intensities of the diffraction patterns of these four types are determined by the amount of gaseous contamination on the surface of the crystal, and also by the temperature of the surface.

The first publication of our discovery of electron diffraction was contained in a note in "Nature." [3] At the time of this first paper we had already discovered two of these types of electron diffraction, and had distinguished sharply between them. The first type, which we later

[1] Translation from "Zeitschrift für Physik," April 12, 1929, pp. 408–421.
[2] C. J. Davisson and L. H. Germer, *Phys. Rev.*, 30, 705 (1927).
[3] C. J. Davisson and L. H. Germer, *Nature*, 119, 558 (1927).

found to be predominant under all conditions, was the normal diffraction from the space lattice of the nickel crystal. The diffraction pattern of this type is analogous to a Laue X-ray pattern produced by the interaction of a beam of heterogeneous X-rays with a single crystal. The second type of diffraction pattern consisted of a curious but very simple assemblage of twelve electron diffraction beams which, at the time of their discovery, we designated "anomalous beams."

At the time of publication of the note in "Nature" the nickel crystal, with which we were experimenting, had not been heated after the bulb containing it was sealed from the pumps. Subsequently the crystal was heated many times by electron bombardment. The effect of these heatings was to destroy the diffraction pattern of the second type, made up of the so-called anomalous beams, and to intensify greatly the first type of diffraction pattern. Simultaneously patterns of two entirely new types appeared for the first time. These new diffraction patterns were comparatively short-lived. Both of them disappeared completely within a few hours after heating the crystal. After its initial increase in intensity, following such a heating, the diffraction pattern of the first type also became weaker with the passage of time, but at a rate very much less rapid than the rate at which these new patterns changed. After the lapse of some days the diffraction pattern of the second type was again found. At first the beams of this pattern could barely be detected, but they became progressively stronger as the diffraction pattern of the first type became weaker.

Experiments more or less similar to the experiment just described were performed many times. After the initial heating of the crystal the rapidity of occurrence of the changes in the diffraction patterns varied greatly, from one experiment to another. These changes occurred very slowly when the experimental conditions were such as to warrant the belief that the vacuum in the experimental tube was unusually high. The changes took place much more rapidly when the vacuum was known to be comparatively poor. We concluded that the changes in the diffraction patterns which were observed following a heat treatment of the crystal were caused by gas being gradually adsorbed upon its surface. The two new types of diffraction pattern must arise from a clean or nearly clean surface, and their complete disappearance seemed conclusive evidence that the surface had become covered by at least one layer of gas atoms.[4] The diffraction pattern of the first type was weakened only slightly at the time of their disap-

[4] We have not obtained any information regarding the nature of the adsorbed gas. In referring to this gas the word "atom" has been used to mean either atom or molecule.

pearance, and we therefore concluded that at the time when this pattern was itself greatly weakened the surface of the crystal was covered by *many layers* of gas atoms. Furthermore, we were able to conclude that the diffraction pattern of the second type had its origin in the film of adsorbed gas, and that *many layers* of gas were necessary for its formation.

In Figs. 1–4 are exhibited polar plots of typical electron diffraction beams belonging to the diffraction pattern of the first type. The curves of these figures show the effect of gaseous contamination of the crystal

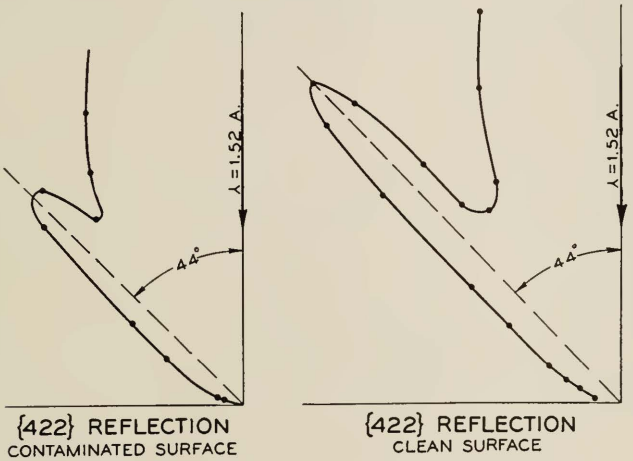


FIG. 1—Showing the effect of the removal of surface gas upon the intensity of the electron diffraction beam whose Miller indices are (422)

surface upon the intensities of diffraction beams arising from the space lattice of the crystal. In these figures the electron diffraction beams have been designated by their Miller indices in accordance with the conventional X-ray nomenclature. Since the refractive index of nickel for electron waves differs from unity by a quite appreciable fraction, in the wave-length range of our experiments, a knowledge of the values of this index was necessary before these Miller indices could be assigned. This knowledge has been supplied by our experiments [5] on electron reflection. (Although Bethe [6] and others deduced refractive indices correctly from the data which we published in the Physical Review (loc. cit.) and assigned the correct Miller indices to the diffraction beams which we observed, these deductions seem to us to have rested upon a rather inadequate experimental basis.)

[5] C. J. Davisson and L. H. Germer, *Proc. Nat. Acad. Sci.*, 14, 619 (1928).
[6] Bethe, *Naturwissenschaften*, 16, 333 (1928).

Fig. 1 consists of two electron scattering curves showing a (422) reflection which occurs at an electron wave-length of 1.52 A. The curve on the left was taken before the crystal had been heated after the sealing of the bulb containing it from the pumps. In the curve on the right is shown this same diffraction beam soon after the crystal was heated. From the reasoning given above we know that at the time the left-hand curve was taken the crystal surface was covered by *many layers* of adsorbed gas, and that when the right-hand curve was taken the surface of the crystal was comparatively clean.

Fig. 2 is similar to Fig. 1. It shows the increase in intensity of a (622) reflection when the crystal surface was cleaned by heating. After such a cleaning the surface became gradually covered again by

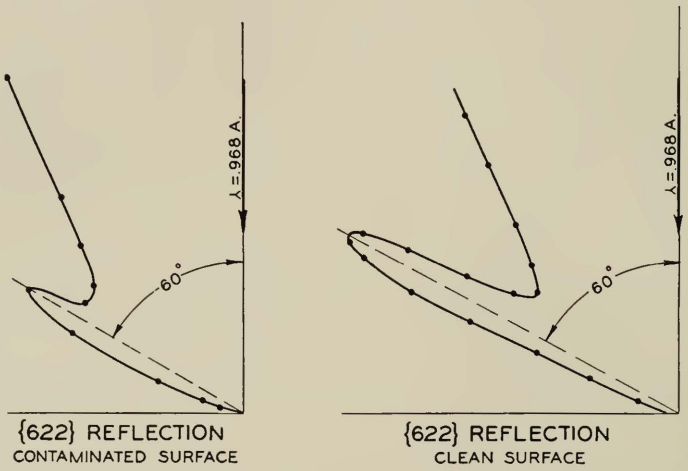


Fig. 2—The effect of the removal of surface gas upon a (622) diffraction beam

adsorbed gas. The intense diffraction beams shown at the right of Figs. 1 and 2 became weaker, finally reaching the intensities shown by the left-hand curves. The time required for this change to occur varied from less than an hour to many weeks, depending upon the vacuum condition of the bulb.

In Fig. 3 is shown the effect of cleaning the crystal surface upon the intensity of a (331) reflection, and in Fig. 4 the effect upon a (551) reflection. The curves either side of the centers in these figures show how selective are the occurrences of these diffraction beams. The disappearance of the diffraction beams as the wave-length is changed from the critical value of 1.67 A. for the (331) reflection and .912 A. for the (551) reflection respectively is much less sudden than the

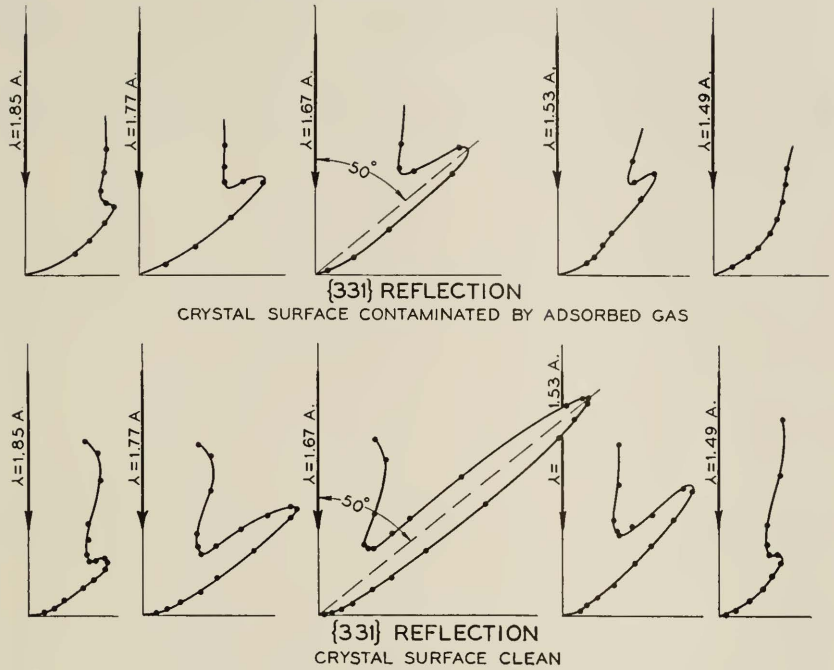


FIG. 3—Development curves of a (331) reflection before the removal of surface gas, and after

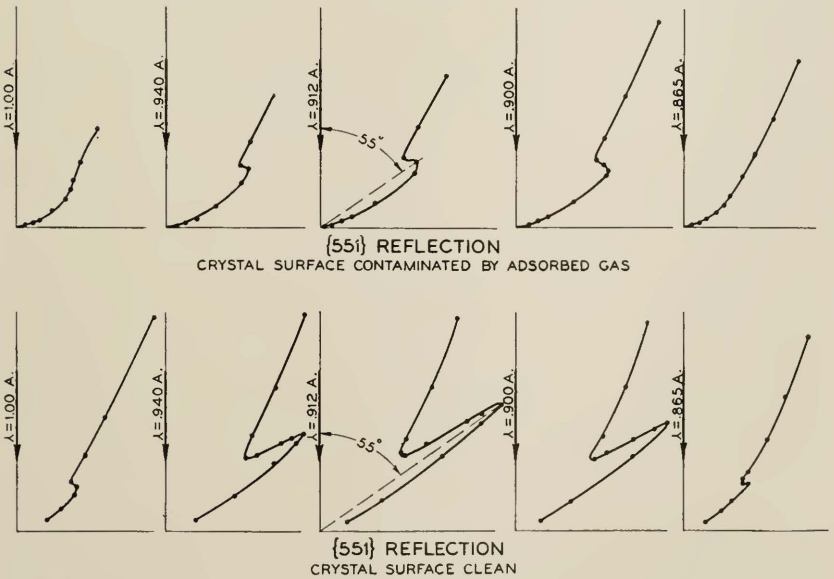


FIG. 4—Development curves of a (551) reflection before and after the removal of surface gas

disappearance of X-ray diffraction beams under similar conditions. In other words, the wave-length resolving power of the crystal is not so good for electron waves as for X-rays of comparable wave-lengths. This is due to the fact that the penetration of electron waves into the metal is very much less than the penetration of X-rays of the same wave-length.[7]

It is, of course, just this slight penetrating power of electron waves which made the diffraction patterns arising from our nickel crystal sensitive to the presence of adsorbed gas, and caused the extraordinary complexity in the observed phenomena. The circumstance, that electron waves are scattered very efficiently by the surface atoms of the crystal and are consequently extinguished on penetrating into the crystal at a very rapid rate, opens up to us the possibility of the use of electron diffraction as a means of studying surfaces.

The first application of this new method of surface study was the analysis (*Phys. Rev.*, loc. cit.) of the two transient diffraction patterns which existed for only a short time after the crystal surface was cleaned by heating. The first of these patterns to appear after the heating was found as soon as the crystal was comparatively cool. This pattern I shall refer to as the electron diffraction pattern of the third type. It consisted of electron beams emerging near to grazing the surface in the principal azimuths of the crystal, occurring in each azimuth just as if the surface of the crystal were a plane diffraction grating. For each diffraction beam the grating constant was equal to the separation between the rows of atoms on the surface of the crystal normal to the azimuth of the beam. Figures showing beams of this type in the principal crystal azimuths were exhibited in our original paper (loc. cit., Figs. 14–16).

The change in intensity with time of a typical beam of this third type after the crystal surface was cleaned by heating is shown by the curve marked "Type-3" in Fig. 5. (The crystal was not cool until nearly twenty minutes after the heating.) From the positions of these "plane grating" (Type-3) beams near to grazing the crystal surface and from their behavior with time after the surface was cleaned, we concluded that, when the beams of this type were most intense, the crystal surface was completely free from adsorbed gas. We concluded also that, until these beams had become quite weak, the surface was not covered by so much as a single layer of gas atoms. The condition,

[7] The lower curves in Fig. 3 represent the data from which we calculated (*Phys. Rev.*, loc. cit.) the rate of extinction of 54 volt electrons ($\lambda = 1.67$ A.) on penetrating into the metal. A similar calculation cannot readily be made from the data of the lower curves in Fig. 4. These matters were considered in detail in our original paper and need not be discussed here.

which we recognize as a completely clean condition of the crystal surface, could be produced by a comparatively mild heating, to a temperature roughtly estimated to be about 900° C. The conclusion that such a mild heating was sufficient to clean the surface completely is of interest, as it appears to be in disagreement with similar evidence concerning gas on surfaces obtained from photoelectric and thermionic measurements.

Transient electron diffraction beams of another type appeared some time after the crystal had become cool following a heating. These beams constitute what I shall call the electron diffraction pattern of the

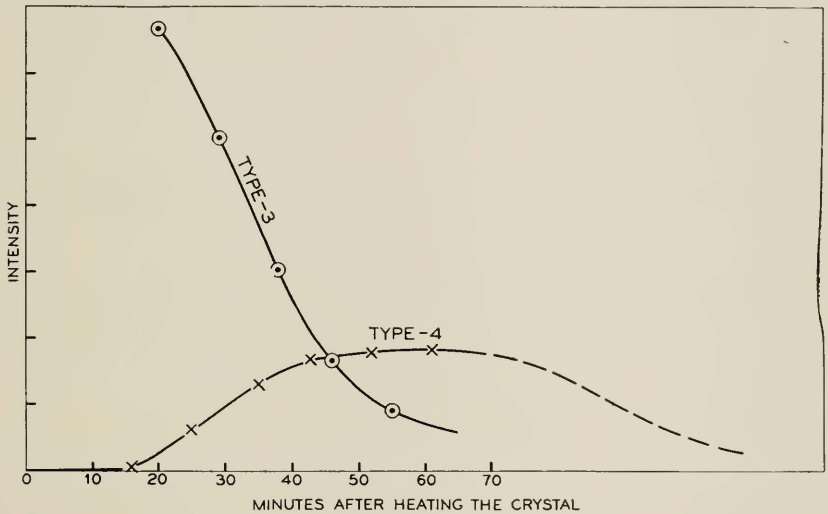


FIG. 5—Change in intensity of a typical "Type-3" beam and of a typical "Type-4" beam as gas settled upon a clean crystal surface. Measurements of the intensities of these two beams were made alternately over a period of an hour.

fourth type. They could not be detected until the diffraction pattern of the third type had become appreciably weakened, and did not attain their maximum intensities until the pattern of the third type had almost disappeared. The curve marked "Type-4" in Fig. 5 shows the life history of a typical beam of this type following a heat treatment of the crystal. In this particular experiment intensity measurements were not made after sixty-one minutes. The dashed continuation of the curve represents its general course as known from other experiments.

The two curves of Fig. 5 represent a single experiment which was carried out at a time when the vacuum condition of our experimental apparatus was intermediate between the best and the worst conditions

which we were able to realize. Under our best vacuum conditions, the intensity changes took place more than twenty times more slowly than the changes shown in Fig. 5, and under our poorest vacuum conditions they took place so rapidly that the diffraction patterns of the third and fourth types could not be found.

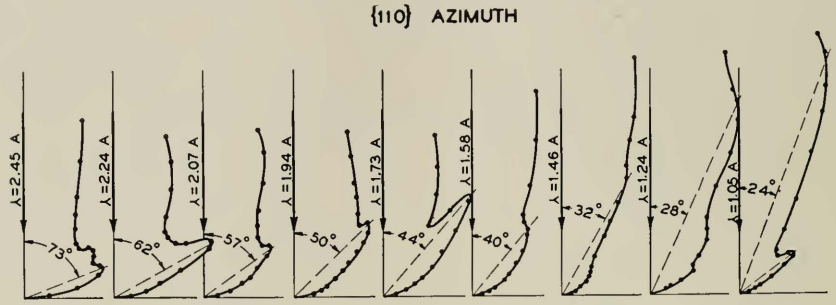


FIG. 6—Scattering curves in a (110) azimuth of the crystal showing the transient electron diffraction beams of the fourth type. On each arrow indicating the primary beam is printed the wave-length of the electrons giving rise to the curve.

The entire diffraction pattern of the fourth type is exhibited by the curves of Figs. 6–8.[8] These figures show a series of electron scattering curves in each of the principal azimuths of the crystal. When these

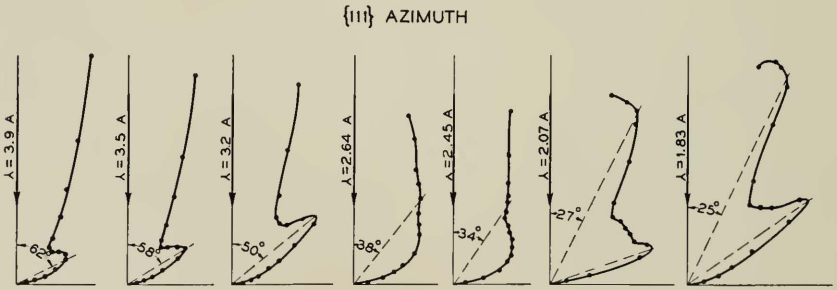


FIG. 7—Scattering curves in a (111) azimuth showing electron diffraction beams of the fourth type

curves were taken, the crystal surface was in the transient gas-covered condition corresponding to the maximum of the "Type-4" curve shown in Fig. 5. The vacuum condition of the apparatus was so good at this

[8] It will be remembered that we were bombarding normally a (111) face of a nickel crystal. Each azimuth of the crystal we designated by the Miller indices of the densest plane of atoms the normal of which lies in the azimuth and in or above the surface of the crystal. Diffraction beams were found only in the three most important azimuths, the designations of which are the (111) azimuth, the (100) azimuth, and the (110) azimuth. (Other azimuths were explored without finding diffraction beams.)

time that the intensities of the beams shown in these figures remained sensibly unchanged for several hours.

The scattering curves near the right-hand ends of these three figures show weakened diffraction beams of the third type, which we recognize readily. The curve second from the right in Fig. 8 shows also a diffraction beam of the first type, a beam arising from the space lattice of the nickel crystal. This is a (311) reflection according to X-ray terminology. The other electron beams shown in Figs. 6–8 can be most easily considered by correlating the wave-length of each beam with the sine of its co-latitude angle. This is the procedure which was followed in Fig. 17 of our original paper in considering the diffraction beams of the *first* type. (Loc. cit. Fig. 17.) Fig. 9 is

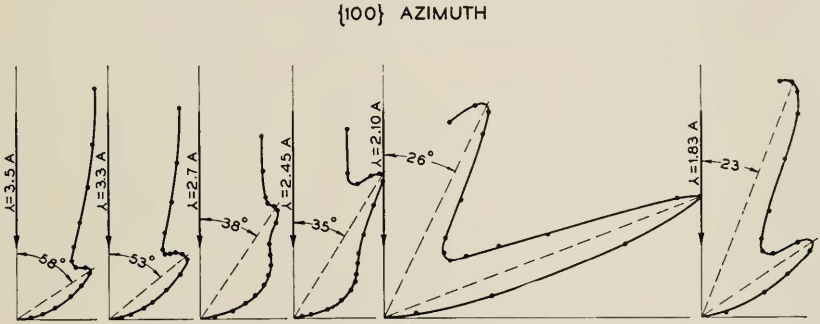


FIG. 8—Electron diffraction beams of the fourth type in a (100) azimuth

similar to Fig. 17 of the original paper. The diagonal lines in Fig. 9 are the plots in the various orders of the plane grating formula $n\lambda = d \sin \theta$, where in each azimuth the grating constant d is equal to the separation between lines of nickel atoms on the surface of the crystal normal to the azimuth. On this figure are plotted as dots all of the originally reported diffraction beams of the first type and as crossed circles the diffraction beams of the fourth type shown in Figs. 6–8.[9]

It is clear from Fig. 9 that the diffraction beams of the fourth type are "plane grating" beams of "one-half order." The obvious interpretation of the occurrence of such beams is that, for these beams, *the value of the grating constant is really 2d instead of d.*

It was known from the consideration of data similar to those shown in Fig. 5 that the diffraction pattern of the fourth type had its origin

[9] In Fig. 9 there is plotted a crossed circle for each of the separate curves shown in Figs. 6–8. In this respect the crossed circles of Fig. 9 are not similar to the dots. Each dot represents a diffraction beam of the first type *at its intensity maximum*, whereas the crossed circles represent all the beams of the fourth type which were found. The intensities of some of these were very small.

in a film of gas on the surface of the crystal, and that this film of gas was thin. The last figures show that this gas film consisted of only a single layer of atoms, and that these atoms were arranged in a crystalline structure similar to a parallel layer of nickel atoms but separated by distances twice as great as the separations of nickel atoms (*Phys. Rev.*, loc. cit., Fig. 20). This arrangement of gas atoms and surface nickel atoms is shown in Fig. 10.

Very interesting information was obtained by taking data similar to those represented in Fig. 5 with the crystal maintained slightly

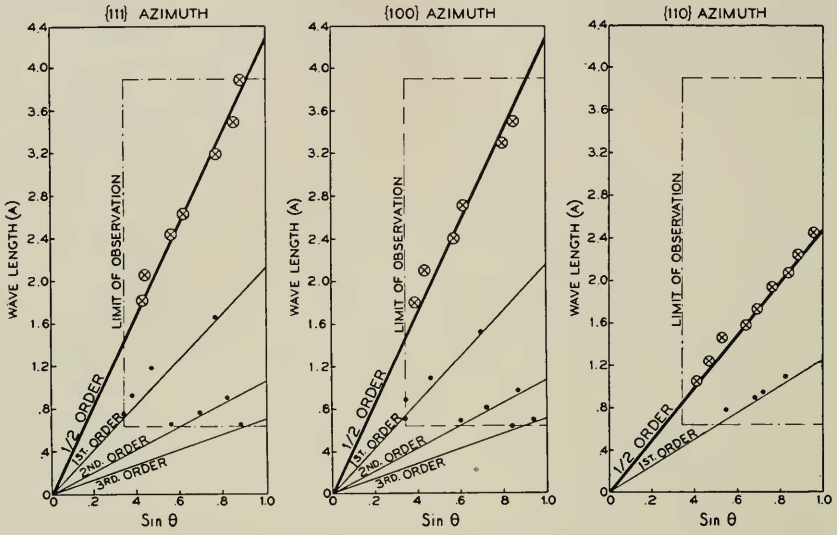


FIG. 9—Plots of the grating formulas in the principal azimuths of the crystal. Positions of electron diffraction beams of the fourth type are indicated by crossed circles and positions of beams of the first type by dots.

warm, at a temperature roughly estimated to be 150° C. For this slightly warm crystal the life history of a typical "Type-3" beam was substantially the same as that shown in Fig. 5. The "Type-4" beams, however, did not develop at all. We concluded that gas atoms settled upon the crystal when warm much the same as when cold, but that upon the warm surface they were unable to take up the regular arrangement which they assumed upon the cold surface. One is led to say that the temperature of 150° C. was above the melting point of the two dimensional "gas crystal." It would have been comparatively easy to have found out whether or not the gas crystal melted abruptly at some critical temperature, but this information was not obtained.

One of the most striking features of Figs. 6–8 is the marked change

in intensity of the "Type-4" beams from one curve to another. In our original publication we pointed out a possible interpretation of these intensity differences. The data concerning these beams have now been more carefully studied, and it has turned out that this suggested interpretation gives quantitative results in agreement with the observations.

What we suggested was that the changes in the intensities of the "Type-4" beams were due to interference between the beams diffracted from the layer of gas atoms and the beams arising from the underlying

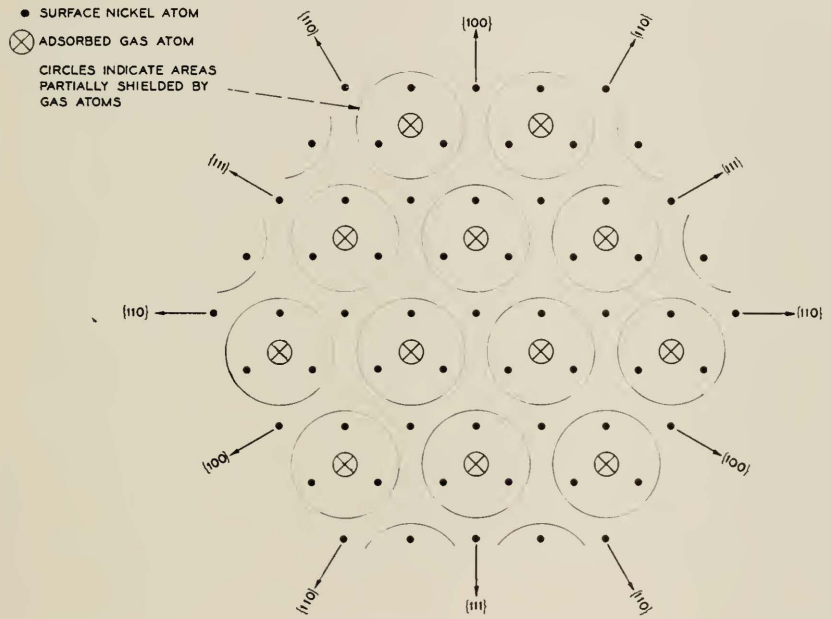


FIG. 10—The arrangement of gas atoms which gives rise to the "Type-4" diffraction beams. The designation of crystal azimuths is the one upon which the calculations of Table I are based.

nickel. That the nickel crystal can produce diffraction beams capable of interfering with the beams from the doubly-spaced layer of gas atoms arises from the fact that the gas atoms divide the surface nickel atoms into two classes. Three-fourths of the surface nickel atoms are adjacent to gas atoms and are presumably partially shielded by these atoms from the incident electron waves (Fig. 10). The other fourth of the nickel atoms on the surface, which are not so shielded, are arranged in a structure similar to that of the gas atoms and having the same scale factor. Thus it is that the surface layer of nickel atoms can give

rise to differential diffraction beams capable of interfering with the "plane grating" beams from the gas atoms.

Believing that the intensity differences of the beams in Figs. 6–8 arise in this manner, one can hope to determine from the intensities of these beams the space relation existing between the layer of gas atoms and the surface layer of nickel atoms. In carrying out the plan of making this determination one naturally assumes a certain relative spacing, and upon this assumption calculates the wave-lengths (and angles) at which the diffraction beams from the gas layer are in phase with the differential beams from the surface nickel layer.[10] These wave-lengths are then compared with the wave-lengths in Figs. 6–8 at which the "Type-4" beams are found to be strong.

Calculations of this kind have been carried out, with the gas atoms placed as shown in Fig. 10, for each of a series of assumed separations between the plane of the gas atoms and the plane of the surface nickel atoms. In comparing the results of these calculations with the observations of Figs. 6–8, the azimuths are to be taken either as shown in Fig. 10 or with the (111) and (100) azimuths interchanged. As far as the first layer of nickel atoms is concerned the (111) and (100) azimuths are identical. When one must distinguish between these azimuths (as in Fig. 10), this distinction is equivalent to a certain definite location of the gas atoms relative to the *deeper layers* of nickel atoms. It turns out that, when the centers of the gas atoms are assumed to lie in a plane separated by 3.0 A. from the plane of the centers of the surface nickel atoms and when the azimuths are chosen as shown in Fig. 10, the calculated wave-lengths of the intensity maxima of the "Type-4" beams agree pretty well with the wave-lengths at which the beams are observed to become strong. These calculated wave-lengths of the intensity maxima are given in Table I.

TABLE I

CALCULATED WAVE-LENGTHS OF THE INTENSITY MAXIMA OF "TYPE-4" BEAMS ASSUMING THE AZIMUTH DESIGNATIONS OF FIG. 10 AND A SEPARATION OF 3.0 A. BETWEEN THE PLANE OF THE GAS ATOMS AND THE PLANE OF THE SURFACE NICKEL ATOMS

(110) Azimuth (compare with Fig. 6)	(111) Azimuth (compare with Fig. 7)	(100) Azimuth (compare with Fig. 8)
2.20 A.	3.07 A.	3.53 A.
1.72	2.10	2.36
1.38	1.57	1.72
1.13		
0.96		

[10] One does not, of course, obtain in this simple manner the intensities of the diffraction beams, but one does obtain the positions and wave-lengths at which these beams reach intensity maxima.

They are to be compared directly with Figs. 6–8. Better agreement could not be expected in view of the fewness of the experimental data. For other assumed separations the calculated wave-lengths of the intensity maxima do not agree so well with the observations.

Designating azimuths as shown in Fig. 10 is tantamount to stating that the gas atoms locate themselves directly over nickel atoms of the third layer. This location appeals to one as inherently probable. The separation of 3.0 A.[11] is greater by about fifty per cent than the separation of layers of nickel atoms parallel to the surface (2.03 A.). This also seems reasonable in view of the fact that the gas atoms are separated laterally by distances twice as great as the separations of nickel atoms.

. . . . . .

Our analyses of the diffraction patterns of the first, third and fourth types seem to us to be fairly satisfactory. The study of the diffraction pattern of the second type has, however, turned out to be futile. This pattern consisted of twelve diffraction beams, one in each of the principal azimuths of the crystal (see Fig. 10). All these beams appeared at the same co-latitude angle of about 58° and for the electron wave-length 1.17 A. We know that this curious pattern had its origin in a heavy layer of adsorbed gas, but we have been unable to determine what were the crystalline regularities in this layer. Our present feeling is that these beams may really have reached their intensity maxima at slightly different angles and wave-lengths in the different crystal azimuths, and that perhaps there may have been other beams belonging to this pattern which we did not detect. We are inclined to believe that we shall not be able to obtain information concerning the crystalline arrangement of a heavy gas layer until further experimental data are available.

. . . . . .

Looking back over the information recorded here concerning gas upon a crystal surface, one should think of it as information which has been obtained by a new method of diffraction analysis. Thinking of it in this way I realize that the observations are of a rather elementary nature. The facts which we have discovered are by no means comparable in complexity with the elaborately detailed information concerning crystal structure which is so readily obtained by X-ray analysis.

Electron diffraction is, however, a new field. At the time when the

[11] Logically this value of 3.0 A. should perhaps be changed by a few per cent to take into account a refractive index slightly different from unity.

experimental data described here were obtained, we could have learned a great deal more concerning gas upon our crystal surface with comparatively little effort. In the above pages there are many suggested continuations which were not followed up. When our measurements were made our interests were quite naturally centered upon the phenomenon that electrons are diffracted, and we neglected many lines of physical investigation which were open to us. The work described here must be regarded as only the very first application of what may possibly develop into a useful method of crystal analysis, occupying a field entirely distinct from the field of X-ray analysis.

Abstracts of Technical Articles From Bell System Sources

Relation of Nitrogen to Blue Heat Phenomena in Iron and Dispersion-Hardening in the System Iron-Nitrogen.[1] R. S. Dean, R. O. Day and J. L. Gregg. It has been generally observed that iron, as an outstanding exception among metals, increases its hardness and strength by low-temperature annealing after cold work, and also by increase of testing temperature to the range of 150° C. to 300° C. This investigation was made with the object of ascertaining if similar phenomena were observed in high purity iron and, if not, to the presence of which impurities these phenomena could be traced. After describing the tests made and giving the results, the authors come to the conclusion that commercial irons owe their property of hardening by reheating after cold work, as well as their increase in tensile strength in the range 100° C. to 300° C., to the solution of small amounts of iron-nitride present.

Heat Treatment and Mechanical Properties of Some Copper-Zinc and Copper-Tin Alloys Containing Nickel and Silicon.[2] W. C. Ellis and Earle E. Schumacher. The addition of nickel and silicon to the copper-zinc and copper-tin systems results in alloys which can be hardened by heat treatment. The heat treatment, in general, consists of a quench from 800° C. followed by hardening at 400° C. to 500° C. The dispersion-hardening effect of nickel and silicon in these alloys opens a considerable field in the manufacture of high strength brasses. The mechanical properties in the rolled condition of the hardened brass containing 30 per cent of zinc and 3 per cent of nickel plus silicon are in general similar to those of high brass sheet in the spring temper. The endurance limit in reversed flexure for this alloy in the hardened condition is, however, approximately 20 per cent higher than that of high brass sheet in the same temper.

A Metallographic Study of Tungsten Carbide Alloys.[3] J. L. Gregg and C. W. Küttner. This paper gives the results of an investigation of the structure of five of the tungsten-carbon alloys by means of microscopic and X-ray methods, the samples studied being small tools or wire-drawing dies. After a general discussion of the constituents of tungsten-carbon alloys, the preparation of the samples is described, and the structures found are shown in twenty-one figures accompanying the text.

[1] *Mining and Metallurgy*, Vol. 10, March, 1929, p. 163 (abstract).
[2] *Mining and Metallurgy*, Vol. 10, March, 1929, p. 162 (abstract).
[3] *Mining and Metallurgy*, Vol. 10, February, 1929, p. 94 (abstract).

Motion Pictures in Relief.[4] HERBERT E. IVES. In this article Dr. Ives describes the method by which stereoscopic motion picture projection can, theoretically at least, be attained. The method is relatively complicated and has severe practical limitations. It appears to be theoretically sound and capable of realization, at least on an experimental scale.

The Absorption of Oxygen by Rubber.[5] G. T. KOHMAN. The work reported in this paper was planned for the purpose of determining the part played by oxygen absorption in the natural aging of rubber. To do this, the effects of a number of factors known to influence natural aging on rates of oxygen absorption were studied. A piece of apparatus, developed for determining these rates, which involves special means for keeping the oxygen pressure surrounding the sample constant is described. The results obtained lead to the conclusion that oxygen absorption is the predominating factor in the natural aging of rubber and that rates of oxygen absorption are of value in predicting the natural life of rubber.

An Electrical Test for Tin Coating on Copper Wire.[6] H. M. LARSEN and C. M. UNDERWOOD. The method described is essentially a deplating process. The wire samples are placed in an acid solution and a current of suitable value applied to effect the deplating. The weight of tin on the wire surface and that alloyed with the copper are determined separately, the measuring means being two graduated tubes containing electrodes (sometimes called voltameters). The gas evolved in these voltameters is proportional to the current and hence to the tin being removed. As soon as the copper surface is exposed, an auxiliary electrode in the deplating bath actuates a relay which brings into operation the second voltameter, permitting determination of the tin alloyed with the copper.

Very simple formulæ permit determining the amount of tin from the volume of gas accumulated in the two voltameters. The method is said to save time and permit the use of relatively unskilled operators as compared with the usual chemical tests applied to tin coatings.

Further Observations on the Microstructure of Martensite.[7] FRANCIS F. LUCAS. This paper is a further contribution by Dr. Lucas on the microstructure of martensite. It describes a number of quenching and

[4] *Journal of the Optical Society of America and Review of Scientific Instruments*, Vol. 18, February, 1929, pp. 118–122.
[5] *Journal of Physical Chemistry*, Vol. 33, February, 1929, pp. 226–243.
[6] *Wire & Wire Products*, Vol. 4, April, 1929, pp. 118–119, 140.
[7] *Transactions of the American Society for Steel Treating*, Vol. 15, February, 1929, pp. 339–364.

tempering experiments in which commercial high quality tool steels were used. Representative structures found in the quenched and various tempered conditions are illustrated and discussed.

Technique of the Talking Movie.[8] DONALD MACKENZIE. In this article the talking movies are described in some detail as to mechanical features, production and exhibition. The author tells some interesting things about producing these pictures and the human reactions that must be considered in preparing a picture with sound record so that it will seem natural and the changes that producers will have to make to satisfy the public.

Some Long Distance Transmission Problems.[9] H. MOURADIAN. This paper discusses the transmission properties of high voltage power transmission lines with incidental reference to telephone transmission. The method of improving the performance of power lines by means of synchronous condensers at the ends and at intermediate points is discussed and compared with a proposed method in which neutralizing networks are neutralized at intervals. Each network consists of a pi whose series and shunt elements neutralize the corresponding elements of the line at the frequency of transmission. It is stated that the synchronous condensers increase the power transfer limits of the line but decrease the transmission efficiency, while the neutralizing networks increase both the power transfer limits and the efficiency. Illustrative numerical examples are given for a 220,000-volt line, 500 miles long. Some possibilities of a transcontinental power transmission line are discussed.

Electrical Conduction in Textiles. Part II—Alternating Current Conduction.[10] E. J. MURPHY. This paper shows the variation of the equivalent parallel capacity and conductance of cotton and silk with relative humidity and frequency (for a small range). It also shows the effect of changes in the amount of electrolytic material in the textile. The main results are: At high humidities the capacity is greatly reduced by a reduction in the amount of electrolytic material in the textile. The a.c. and d.c. conductivities of cotton approach each other as the humidity is increased and become equal at humidities greater than 80–85 per cent (that is, dielectric loss is entirely due to direct current conductivity in this range). At humidities lower than this a large part of the dielectric loss is not due to d.c. conduction, but this

[8] *Journal of the Western Society of Engineers*, Vol. 34, February, 1929, pp. 95–102.
[9] *Journal of the Franklin Institute*, Vol. 207, February, 1929, pp. 165–192.
[10] *Journal of Physical Chemistry*, Vol. 33, February, 1929, pp. 200–215.

part of the dielectric loss is also strongly affected by the amount of electrolytic material in the textile. These characteristics can be explained if the textile is regarded as an electrolytic cell in which the absorbed water and dissolved salts form the electrolyte and the solid constituents of the textile act as a container which determines the volume, geometric form and specific conductance of the electrolyte. The capacity at high humidities is regarded as due chiefly to the electrolytic polarization capacity of this electrolyte.

Electrical Conduction in Textiles. Part III—Anomalous Properties.[11] E. J. MURPHY. This paper deals with the increase of conductivity with increasing applied voltage (the Evershed effect), and with the residual electromotive forces and changes in resistance produced by the passage of current through the textile. The results point to the conclusion that the Evershed effect is due to two factors, a back-e.m.f. due to electrolytic polarization, and an increase, caused by the increase in voltage, in the amount to which the ions adsorbed by the interface between the aqueous solutions and the solid material of the textile contribute to the total conductivity. The characteristics of the residual e.m.f. change with humidity; at high humidities the e.m.f. is apparently caused by chemical changes in the aqueous solutions due to their electrolysis. It was found that the passage of a current through a textile causes its resistance to become non-uniformly distributed, the distribution depending on the nature of the electrode material; this is interpreted as due to changes in the chemical composition of the solutions in different parts of the textile. The anomalous properties can be explained in terms of the electrolytic cell mechanism suggested in the preceding paper by attributing to the solid in which the aqueous conducting paths are contained the properties of adsorbing ions and of hindering the equalization of concentration differences in the solutions by diffusion. Thus, all of the properties of conduction in textiles observed in this investigation can be explained in terms of a single general mechanism which appears to be a probable consequence of the colloidal structure of the materials.

Study of Weller Brittleness Test for Paper.[12] R. L. PEEK, JR. and J. M. FINCH. On the assumption that paper possesses certain basic properties, an expression is theoretically obtained relating the results of the Weller brittleness test to these basic properties, the dimensions of the sample, and the conditions of testing. Experimental data are presented which show that the effect of the sample dimensions and the

[11] *Journal of Physical Chemistry*, Vol. 33, April, 1929, pp. 509–532.
[12] *Paper Trade Journal*, Vol. 88, February 7, 1929, pp. 56–62.

conditions of testing are substantially as indicated by the theoretical expression. The theory is then employed to interpret the results of the test and to indicate the best form in which these may be expressed. The general question of testing for flexibility and brittleness is considered in the light of this study.

Diffusion of Water through Rubber.[13] EARLE E. SCHUMACHER and LAWRENCE FERGUSON. This article gives data on the diffusion of water through thirteen rubbers of different compositions. The mathematical derivation of a simple formula for calculating the rate of diffusion has been given. The diffusion measurements have shown: (*a*) that the rate of diffusion of water through a rubber membrane is inversely proportional to the square of the thickness: (*b*) that the rate of diffusion decreases greatly with increase in hardness; (*c*) that the effect of saturating the rubber with water is to increase the rate of diffusion through it, due probably not only to an increase in the water vapor pressure within the rubber, but also to a decrease in hardness; (*d*) that there is no intimate relationship between rate of diffusion and minor variations in the composition of the rubber.

Effect of Arsenic on Dispersion-Hardenable Lead-Antimony Alloys.[14] K. S. SELJESATER. Arsenic has no solid solubility in lead and is known to form a continuous series of solid solutions with antimony. Therefore, immediately after annealing and quenching the antimony is in solid solution in the lead, and there is a certain amount of eutectic between the lead-antimony solid solution and arsenic. After quenching, the lead-antimony solid solution is supersaturated (the same as if arsenic were not present) and minute crystals of antimony separate. Since arsenic is soluble in antimony, some of the arsenic present will be concentrated in the surface layer of the minute antimony particles, which will then possess surface conditions different from those of pure antimony particles. The condition of the alloy at this stage is analogous to a suspension in a liquid which has been partly stabilized by a third constituent. Agglomeration and precipitation will occur, but at a much slower rate than if the third constituent were not present. Arsenic, therefore, is to be considered as a retardant for the agglomeration of minute antimony particles in the lead matrix. The length of the stabilization time decreases at elevated temperatures. The offered explanation is in agreement with the fact that the increase in hardness is practically independent of the percentage of arsenic within limits investigated. The addition of a third constituent insoluble in the

[13] *Industrial and Engineering Chemistry*, Vol. 21, February 1, 1929, pp. 158–162.
[14] *Mining and Metallurgy*, Vol. 10, February, 1929, p. 94 (abstract).

solvent and forming a continuous series of solid solutions with the solute, might be of advantage to other kinds of age-hardenable binary alloys.

A Precision Regulator for Alternating Voltage.[15] H. M. STOLLER and J. R. POWER. In this paper a recently developed precision voltage regulator for use with alternating current is described. It will maintain its output voltage constant to within 0.03 per cent over an input voltage range of 10 per cent and a load range of from zero to full load.

This regulation is accomplished by means of a small transformer inserted in one of the lines which boosts or bucks the impressed voltage by the required amount. The transformer is controlled by a vacuum tube circuit acting through an inductance bridge.

[15] *Journal of the A. I. E. E.*, February, 1929, Vol. 48, pp. 110–113.

Contributors to this Issue

C. J. Davisson, B.Sc., University of Chicago, 1908; Ph.D., Princeton University, 1911; Instructor in Physics, Carnegie Institute of Technology, 1911–17; Engineering Department of the Western Electric Company, 1917–25; Bell Telephone Laboratories, 1925–. Dr. Davisson's work since coming with the Bell System has related largely to thermionics and electronic physics.

G. W. Elmen, B.Sc., University of Nebraska, 1902; M.A., 1904; Research Laboratories of the General Electric Company, 1904–06; Engineering Department of the Western Electric Company, 1906–25; Bell Telephone Laboratories, 1925–. Mr. Elmen's principal line of work has been magnetic investigations. He is the inventor of the permalloy alloys. For this he was awarded the John Scott Medal in 1926 and the Elliott Cresson Medal in 1928.

J. G. Ferguson, B.Sc., University of California, 1915; M.Sc., 1916; Research Assistant in Physics, 1915–16; Engineering Department, Western Electric Company, 1917–25; Bell Telephone Laboratories, 1925–. Mr. Ferguson's work has been in connection with the development of methods of electrical measurement.

L. H. Germer, A.B., Cornell, 1917; M.A., Columbia, 1922; Ph.D., Columbia, 1927; Engineering Department, Western Electric Company, 1917–1925; United States Army, 1917–1919; Bell Telephone Laboratories, 1925–. Dr. Germer has been engaged chiefly upon work in thermionics and electron scattering.

C. H. G. Gray, B.Sc., Massachusetts Institute of Technology, 1918; Instructor, Electrical Engineering, Massachusetts Institute of Technology, 1918–19; Engineering Department, Western Electric Company, 1919–25; Bell Telephone Laboratories, 1925–. Mr. Gray, in addition to his work on the master reference system, has been engaged in the development and application of machine testing methods for the inspection of transmitters and receivers.

Charles H. Greenall, M.E., Lehigh University, 1922; Engineering Department, Western Electric Company, 1922–25; Bell Telephone Laboratories, 1925–. Mr. Greenall's work has involved investigation of telephone apparatus, and testing methods as applied to specification requirements and development of metallic materials.

F. R. Lack, B.Sc., Harvard, 1925; Engineering Department, Western Electric Company, 1913–22; First Lieutenant, Signal

Corps, A.E.F., 1917–19; Harvard, 1922–25; Bell Telephone Laboratories, 1925–. Mr. Lack has been engaged in experimental work connected with radio communication.

W. A. MARRISON, Royal Flying Corps, later Royal Air Force, Canada, 1917–18; B.Sc., Queens University, Canada, 1920; A.M. in Physics and Mathematics, Harvard University, 1921; Western Electric Company, 1921–25; Bell Telephone Laboratories, 1925–. Mr. Marrison is engaged in the study of picture transmission and methods for the production of constant frequency.

W. H. MARTIN, A.B., Johns Hopkins University, 1909; B.Sc., Massachusetts Institute of Technology, 1911; American Telephone and Telegraph Company, Engineering Department, 1911–19; Department of Development and Research, 1919–. Mr. Martin's work has related particularly to transmission of telephone sets and local exchange circuits, transmission quality and loading.

W. O. PENNELL, B.Sc., Massachusetts Institute of Technology, 1896; Instructor, Lafayette College, 1896–1898; Equipment and Traffic Engineer, Bell Telephone Company of Philadelphia, 1898–1902; Engineering Department, American Telephone and Telegraph Company, 1902–03; Chief Engineer, Missouri and Kansas Telephone Company, 1903–12; Southwestern Bell Telephone Company, 1912–29, and for the last eleven years their Chief Engineer. In addition to his activities as an engineer, Mr. Pennell has from time to time contributed articles of a mathematical nature to various technical publications.

J. R. TOWNSEND, Engineering Department, Western Electric Company, 1919–25; Bell Telephone Laboratories, 1925–. Mr. Townsend has been largely concerned with the testing of telephone apparatus and more recently with the development of requirements of test for metallic materials.

The Bell System Technical Journal

October, 1929

A Method of Sampling Inspection

By H. F. DODGE and H. G. ROMIG

This paper outlines some of the general considerations which must be taken into account in setting up any practical sampling inspection plan. An economical method of inspection is developed in detail for the case where the purpose of the inspection is to determine the acceptability of discrete lots of a product submitted by a producer. By employing probability theory, the method places a definite barrier in the path of material of defective quality and gives this protection to the consumer with a minimum of inspection expense.

ONE of the common questions in every day inspection work is, How Much Inspection? The answer must always be arrived at in the light of economy for only the least amount of inspection which will accomplish the purpose can be justified.

We wish here to consider the problem of setting up an economical inspection plan whose immediate purpose is the elimination of individual lots of product which are unsatisfactory in quality. By a lot of unsatisfactory quality is meant one that contains more than a specified proportion of defective pieces. This proportion is usually small and is based on economic considerations. The interest in inspections made for the elimination of such lots is shared by two parties,—the producer and the consumer. The consumer establishes certain requirements for the quality of delivered product. The producer so arranges his manufacturing processes and provides such an inspection routine as will insure the quality demanded. Our problem recognizes that the consumer runs some risk of receiving lots of unsatisfactory quality, if the quality of each lot is judged by the results of inspecting only a sample. The method of attack presumes the adoption of a risk whose magnitude is agreeable to the consumer and the selection of a particular inspection procedure which will involve a minimum of inspection expense on the part of the producer while guaranteeing the protection agreed upon.

Considering various possible concepts of a risk, the consumer would prefer to have one adopted that was worded as follows:

> "Not more than a specified proportion of the delivered lots shall be unsatisfactory in quality."

In other words, he would like to have the assurance that "the risk

of receiving a lot of unsatisfactory quality shall not exceed some definite figure." This risk involves two probabilities:

(*a*) that an unsatisfactory lot will be submitted for inspection.
(*b*) that the inspector will pass as satisfactory, an unsatisfactory lot submitted for inspection.

Without definite information regarding probable variations in the producer's performance and without an absolute assurance that this performance will remain consistently the same, the use of probability (*a*) in stating the risk to the consumer might be misleading. This probability will therefore receive no further consideration in this paper so far as the definition of the risk to the consumer is involved.

For probability (*b*) a reasonably low value for an upper limit can be given without any knowledge of the producer's performance. This upper limit (as defined below) has been taken as the starting point for the inspection method presented in this paper. The concept of risk which has been adopted gives the consumer unconditional assurance that his "chance of getting any unsatisfactory lot *submitted for inspection* shall not exceed some definite magnitude." This magnitude is the Consumer's Risk used.

Our problem may hence be stated as follows:

> Given a product of a specified type of apparatus or material coming from a producer in discrete lots, what inspection plan will involve a minimum of inspection expense, and at the same time insure that under no conditions will more than a specified proportion of the unsatisfactory lots submitted for inspection be passed for delivery to consumers?

This problem is economic in character and its solution involves the use of probability theory for establishing the height of the barrier to be placed in the path of unsatisfactory material.

Our attention will be directed particularly to the inspection of material which is produced more or less continuously on a quantity basis as distinguished from intermittent production in relatively small amounts. Under these conditions the producer is able to secure a continuing record of performance and to set up an inspection program which takes advantage of current quality trends.

General Considerations in Setting Up an Inspection Method

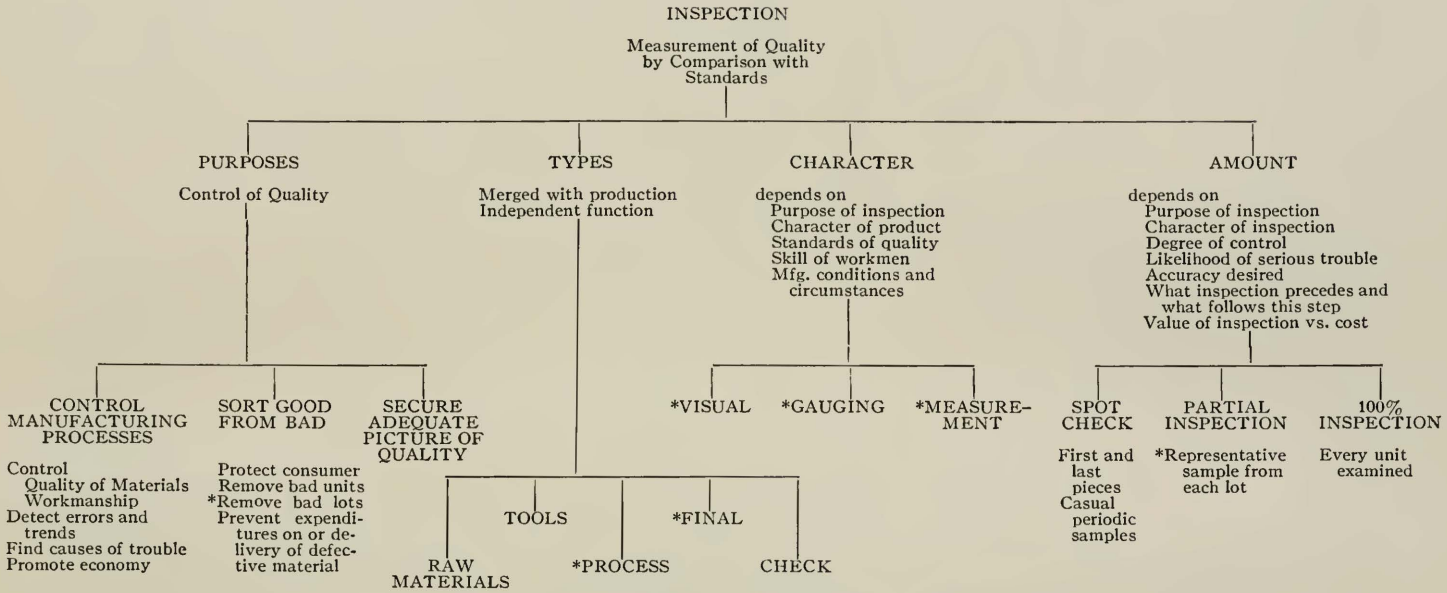
The broad purpose of inspection is to control quality by critical examinations at strategic points in the production process. Raw materials must be inspected. Some of the rough and finished parts must be inspected. In the manufacture of even the simpler kinds of merchandise, inspections dot the chart of progress from raw materials

to the finished product. The distribution of inspection activities throughout any process must be so ordered that the net cost of production will be consistent with the quality demanded by the customer. To determine whether an inspection should be made, or how much should be made at any one of the formative stages, is a major problem involving questions of both quality and economy.

One hundred per cent inspection is often uneconomical at a point in the production process where inspection is clearly warranted, particularly when preceded or followed by other inspections, inasmuch as the cost of more inspection at that point may not be reflected by a corresponding increase in the value of the finished product. In special cases, for example where inspection is destructive, 100 per cent inspection may be totally impracticable. Sampling inspections are often best from the standpoint of both the producer and the consumer when the value of quality and the cost of quality are weighed in the balance.

To arrive at an answer to the question, How Much Inspection, it is first essential to define clearly just what the inspection is intended to accomplish and to weigh all of the important factors both preceding and following the inspection in question which have a direct influence on the quality of the finished product and which as a whole determine how large a part this inspection step must play in controlling quality. Should it serve as an agency for making sure that the product at this stage conforms 100 per cent with the requirements for the features inspected? If so, 100 per cent inspection is required. Or should it serve to make reasonably certain that the quality passing to the next stage is such that no extraordinary effort would have to be expended on defective material? If so, sampling inspection may be employed. Ahead of all else, decisions are needed as to the specific requirements that must be satisfied by the inspection plan itself. This part of the problem is a practical one—one which must be approached in the light of experience, knowledge of conditions and the statistics of past performance. Once the basic requirements of the plan are agreed upon, probability theory can assist in formulating the details which will accomplish the desired results. It is important to hold in mind that statistical methods are aids to engineering judgment and not a substitute for it.

An attempt has been made in Fig. 1 to show schematically some of the outstanding general considerations which must be taken into account in establishing a proper setting for any problem that seeks to determine the economical amount of inspection. Inspections vary widely in purpose, type and character. While their broad purpose is



*Conditions involved in the inspection method of this paper.

Fig. 1—Chart showing some of the factors to be considered in establishing an inspection plan.

to control quality, the immediate objects of individual inspection steps differ. For example, the object of one step may be to secure information which will assist directly in controlling the manufacturing process by detecting errors or trends in performance which would become troublesome if allowed to persist unchecked. In other places the immediate object may be to determine the acceptability of definite quantities of product or to provide a screen for sorting the bad pieces or the bad lots from the good ones. Materials, parts in process and finished units are scrutinized with these objects in view. Depending on circumstances, the character and completeness of inspections vary from visual examination of small samples to careful measurement or testing of each piece.

Conditions Under Which the Present Method Applies

A large amount of industrial inspection work consists in comparing individual pieces with a standard—such as gauging a dimension or measuring an electrical property—to determine whether the pieces do or do not conform with the requirements given in specifications. This is often referred to as inspection by the "method of attributes." Consideration will be directed here to the case where non-destructive sampling inspection of this kind is conducted on discrete lots of product for the purpose of determining their acceptability.

From the standpoint of sampling theory, one of the general requirements is that each lot should be composed of pieces which were produced under the same essential conditions. Practically, this means that an attempt should be made to avoid grouping together batches of material, which, due to manufacturing conditions or methods, are apt to differ in quality. It is presumed, of course, that the sample drawn from any lot will be a random sample so that it may fairly represent the quality of the entire lot.

Summing up the general conditions for which a solution is sought, we assume

(1) The purpose of the inspection is to determine the acceptability of individual lots submitted for inspection, i.e., sorting good lots from bad.
(2) The inspection is made by the "method of attributes" to determine conformance with a particular requirement, i.e., each piece does or does not meet the limits specified.
(3) A lot is homogeneous in quality and the sample from it is a random sample.

The starred items in Fig. 1 indicate the set of conditions involved in our problem.

Protection and Economy Features of the Method

The adoption of sampling inspection at any stage of manufacture carries with it the premise that the product emerging from this point does not have to conform 100 per cent with specification requirements. It is often more economical, all things considered, to allow a small percentage of defective pieces to pass on to subsequent assembly stages or inspections for later rejection than to bear the expense of a 100 per cent inspection. Under these conditions, the status of the inspection can be clarified by establishing a definite tolerance for defects for the lots submitted to the inspector for acceptance. This may be specified as an allowable percentage defective, a figure which may be considered as the border line of distinction between a satisfactory lot and an unsatisfactory one. Thus, if the percentage defective is greater than this "tolerance per cent defective," the lot is unsatisfactory and should be rejected. We say "should be" rejected but this cannot be accomplished with absolute certainty if only a sample is examined. Sampling inspection involves taking chances since the exact quality of a lot is not known when only a part is inspected. According to the laws of chance, a sample will occasionally give favorable indications for bad lots which will result in passing them for delivery to consumers.

The first requirement for the method will therefore be in the form of a definite insurance against passing any unsatisfactory lot that is submitted for inspection.

The second requirement that will be imposed is that the inspection expense be a minimum, subject to the degree of protection afforded by the first requirement.

For the first requirement, there must be specified at the outset a value for the tolerance per cent defective as well as a limit to the chance of accepting any submitted lot of unsatisfactory quality. The latter has, for convenience, been termed the Consumer's Risk and is defined, numerically, as the probability of passing any lot submitted for inspection which contains the tolerance number of defects.

As will be shown further on, the first requirement can be satisfied with a large number of different combinations of sample sizes and acceptance criteria. To satisfy the second requirement, it is necessary then to determine the expected amount of inspection for a variety of inspection plans, determine the cost of examining or testing, add the costs other than those incurred in the simple process of examining samples, and choose among these plans that which involves a minimum of inspection expense.

There are, of course, a number of possible general methods of

inspection procedure, such as single sampling, double sampling, multiple sampling, etc., which allow the examination of only one sample, of two samples, or of more than two samples before a prescribed disposition of the entire lot is made. For each of these general methods, different combinations of sample sizes and acceptance criteria can be found which will satisfy the first requirement. We now prescribe that any lot which fails to pass the sampling requirements shall be completely inspected. Under this condition, one of the above mentioned combinations will give a lesser amount of inspection than the rest. Since a major cost item is that associated with the *amount* of inspection, we will carry through in detail the problem of finding the combination which will result in the *minimum* amount of inspection for one simple general method of inspection.

Single Sampling Method of Inspection

Attention is now directed to what is termed the Single Sampling method of inspection, which involves the following procedure:

(*a*) Inspect a sample.
(*b*) If the acceptance number for the sample is not exceeded, accept the lot.
(*c*) If the acceptance number is exceeded, inspect the remainder of the lot.

The term "Acceptance Number" is introduced to designate the allowable number of defects in the sample.

For this procedure, the first requirement reduces the problem to one which can be solved readily by determining probabilities associated with sampling from a finite lot containing the tolerance number of defects. For any sample size, there is a definite probability of finding no defects, of finding exactly one defect, exactly two defects, etc. If, under the above conditions, the acceptance number were 1, for example, there is one value of sample size such that the probability of finding one or less defects is equal to the value of the Consumer's Risk specified. Since a lot will be accepted if the observed number of defects does not exceed the acceptance number, the probability of finding one or less defects in a sample selected from a lot of tolerance quality is the risk of accepting any lot of tolerance quality submitted to the inspector. It follows that the risk of accepting a lot of worse-than-tolerance quality is less than the Consumer's Risk just defined. If the producer gets into trouble and begins to submit lots of unsatisfactory quality, the consumer has the assurance that his chance of getting them will not exceed this figure. In fact, the worse the quality, the less will be their chance of passing without a detailed

inspection. Thus the amount of inspection is automatically increased as quality degenerates.

For every acceptance number, such as 0, 1, 2, etc., there is an unique size of sample which will satisfy the specified values of tolerance per cent defective and Consumer's Risk. We thus have many pairs of values of sample size and acceptance number from which to choose.

The second requirement dictates which pair shall be chosen. We will select that pair which involves the least amount of inspection for product of *expected* quality. In industry, the quality emerging from any process tends to settle down to some level which may be expected more or less regularly day by day. If this level could be maintained quite constant, if the variations in quality were no larger than the variations that could be attributed to chance, then inspection could often be safely dispensed with. But practically, while such a level may be adhered to most of the time, instances of man failure or machine failure are bound to arise spasmodically and as a result the quality of the output may gradually or suddenly become unsatisfactory. The method of solution takes into consideration this usual or expected quality and requires an estimate of the expected quality under normal conditions. A satisfactory estimate of this can usually be obtained by reviewing data for a past period during which normal conditions existed and by utilizing such other pertinent information as bears on manufacturing performance under present or anticipated conditions. This expected value is defined as the *process average* to be used in the solution. Thus the method under discussion will assure the producer of a minimum of inspection expense so long as he holds to his expected performance. If he gets into trouble and the quality becomes poorer than normally expected, the method automatically increases the inspection by an amount which varies with the degree of quality degeneration. This reacts on the producer directly by increasing his inspection expense and serves as an incentive to the elimination of the causes of trouble. The producer's expected performance is thus made use of in a way that affects the economy of the producer's inspection work but is not used to color or affect the magnitude of the Consumer's Risk.

The amount of inspection, that will be done in the long run for uniform product [1] of process average quality is made up of two parts:

(1) The number of pieces inspected in the samples.
(2) The number of pieces inspected in the remainder of those lots which fail to be accepted when a sample is examined.

[1] By "uniform product" is meant one produced under a constant system of chance causes, giving rise to a quality which is a chance variable. In the present paper, this chance variable is assumed to be the Point Binomial.

But what proportion of the lots will fail to be accepted on the basis of the sampling results? Here is where probability theory comes in again. There will be a definite probability of exceeding the acceptance number in samples drawn from material of process average quality. Since we are interested in the amount of inspection *in the long run*, the sample at this stage of the problem may be regarded as drawn from a very large (mathematically infinite) quantity of homogeneous product whose percentage defective is equal to the process average per cent defective. Thus, for example, with an acceptance number of 1, the average [2] number of pieces inspected *per lot* as a result of *extended* inspections is equal to the number of pieces in the remainder of a lot multiplied by the probability of finding more than one defect in a sample drawn from an infinite quantity of material of this quality. This value plus the number of pieces inspected in the sample gives the average amount of inspection per lot for an acceptance number of 1. Similar results are found for all other acceptance numbers and the desired solution is obtained by choosing that acceptance number for which the average amount of inspection per lot is a minimum.

The plan thus provides the inspector with a definite routine to follow, such that his inspection effort will be a minimum under normal conditions.

Charts for Single Sampling

For any specified value of Consumer's Risk, charts may readily be constructed to give the acceptance number, the sample size, and the average number of pieces inspected per lot for the conditions outlined above. To illustrate the general character of these charts, Figs. 2, 3 and 4 are presented for a Consumer's Risk value of 10 per cent.

In the appendix it is shown that the acceptance number which satisfies the condition of minimum inspection is dependent on two factors, (1), the tolerance number of defects for a lot, and (2), the ratio of the process average to the tolerance for defects. Fig. 2 based on this relationship defines *zones* of acceptance numbers for which the inspection is a minimum.

Fig. 3 gives curves for finding the sample size. The mathematical basis for these curves is likewise given in the appendix. For a given tolerance number of defects and the acceptance number found from Fig. 2, the value of tolerance times sample size is determined. This quantity divided by the tolerance gives the sample size. The curves shown are based on an approximation which is satisfactory for practical use when the tolerance for defects does not exceed 10 per cent and the sample size is not extremely small.

[2] It is to be noted that wherever "average" appears in the paper, "expected" value, in the rigorous probability sense, is meant.

Fig. 4 gives curves which enable one to determine the minimum average number of pieces inspected per lot for uniform product of process average quality. For a given tolerance number of defects and a given ratio of process average to tolerance, a value of tolerance times minimum average number inspected per lot is determined. This when divided by the tolerance gives the desired value of minimum average number inspected per lot.

Illustrative Example

Suppose that lots of 1,000 pieces each are inspected for a characteristic having a specified tolerance of 5 per cent defective and that the process average quality of submitted lots is 1 per cent defective. If it is desired to have a risk of 10 per cent of accepting a 5 per cent defective lot, what single sampling plan should be followed by the inspector to give a minimum amount of inspection, and how much inspection per lot will be required on the average?

Referring to Fig. 2, for "Tolerance Number of Defects" (.05 $\times$ 1,000) = 50 and "Ratio of Process Average to Tolerance" (.01/.05) = .20, we find the acceptance number = 3.

Referring to Fig. 3, for "Tolerance Number of Defects" = 50 and "Acceptance Number" = 3, we find "Tolerance Times Sample Size" = 6.5. Dividing by the tolerance (expressed as a fraction defective) = .05, gives a sample size of 130.

Referring to Fig. 4, for "Tolerance Number of Defects" = 50 and "Ratio of Process Average to Tolerance" = .20, we find by interpolation that the "Tolerance Times Minimum Average Number Inspected" = 8.2. Dividing by the tolerance, .05, gives an average number inspected per lot of 164.

This solution has thus been obtained by the initial specification of first, the tolerance per cent defective for a single lot and a value for the Consumer's Risk, these two factors being combined to give a definite measure of protection against passing faulty material, and second, a minimum amount of inspection for product of process average quality.

These two requirements control the exact details of inspection procedure and must be initially chosen on the basis of practical considerations and circumstances. The specification of these factors lends definiteness to the problem of inspection and provides a rationalized basis of procedure which can be depended on to give the desired degree of protection. Obviously, any value of Consumer's Risk may be chosen according to circumstances. The value which is proper in any case is dependent on the conditions associated with the product

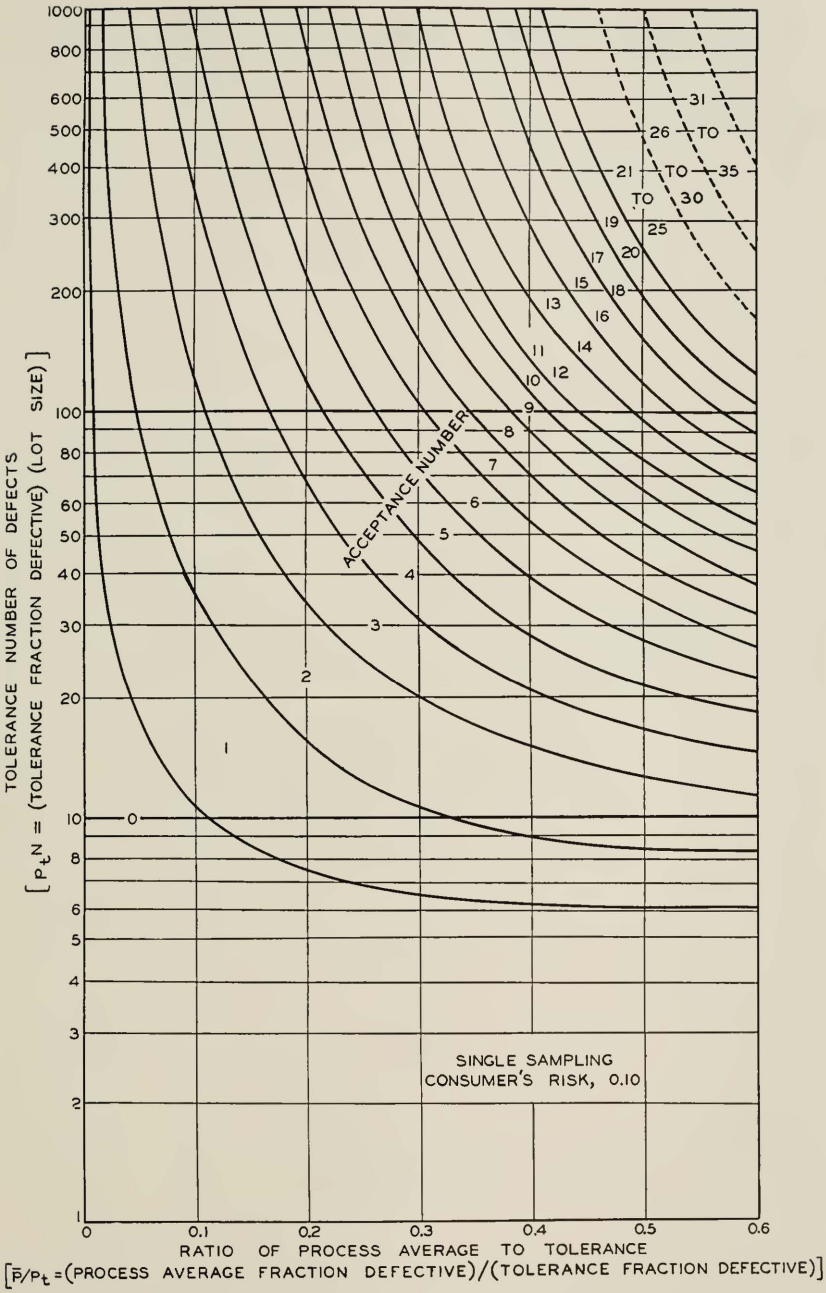


Fig. 2—Chart for finding acceptance number.

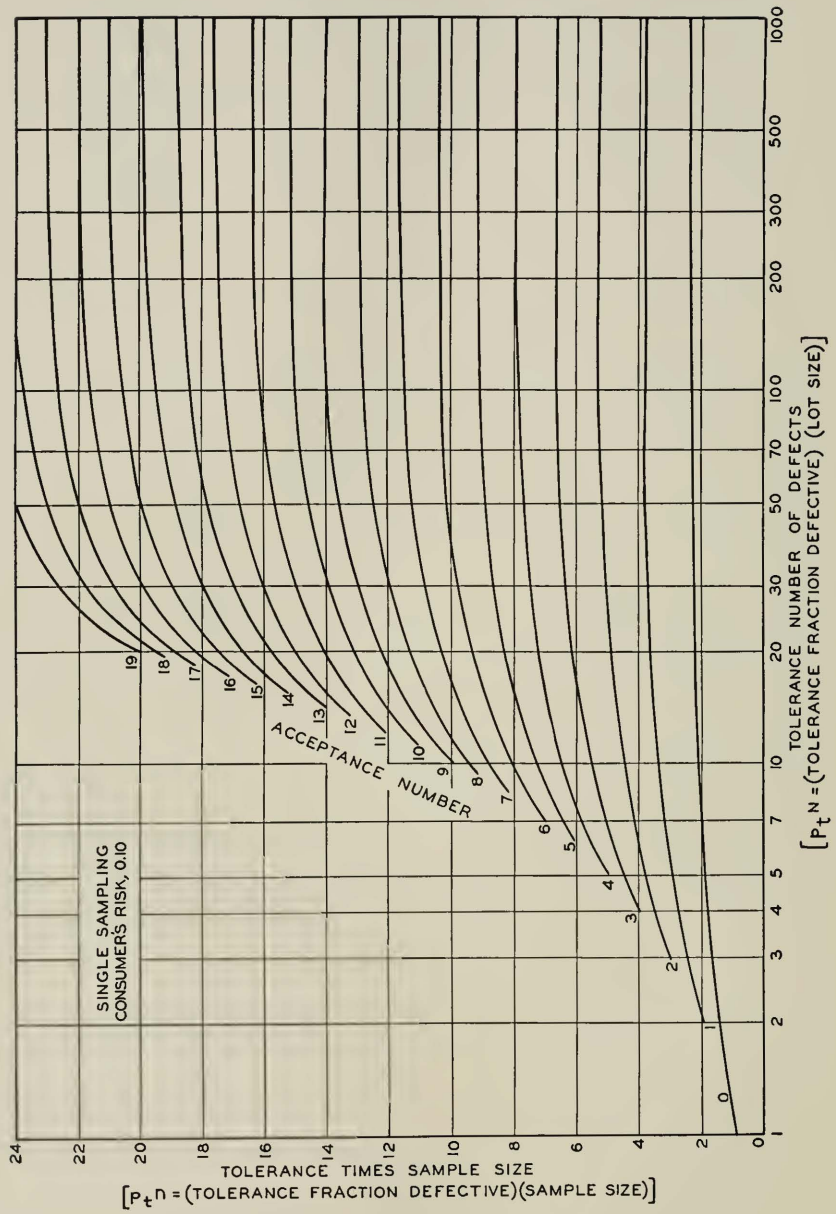


Fig. 3—Curves for finding size of sample.

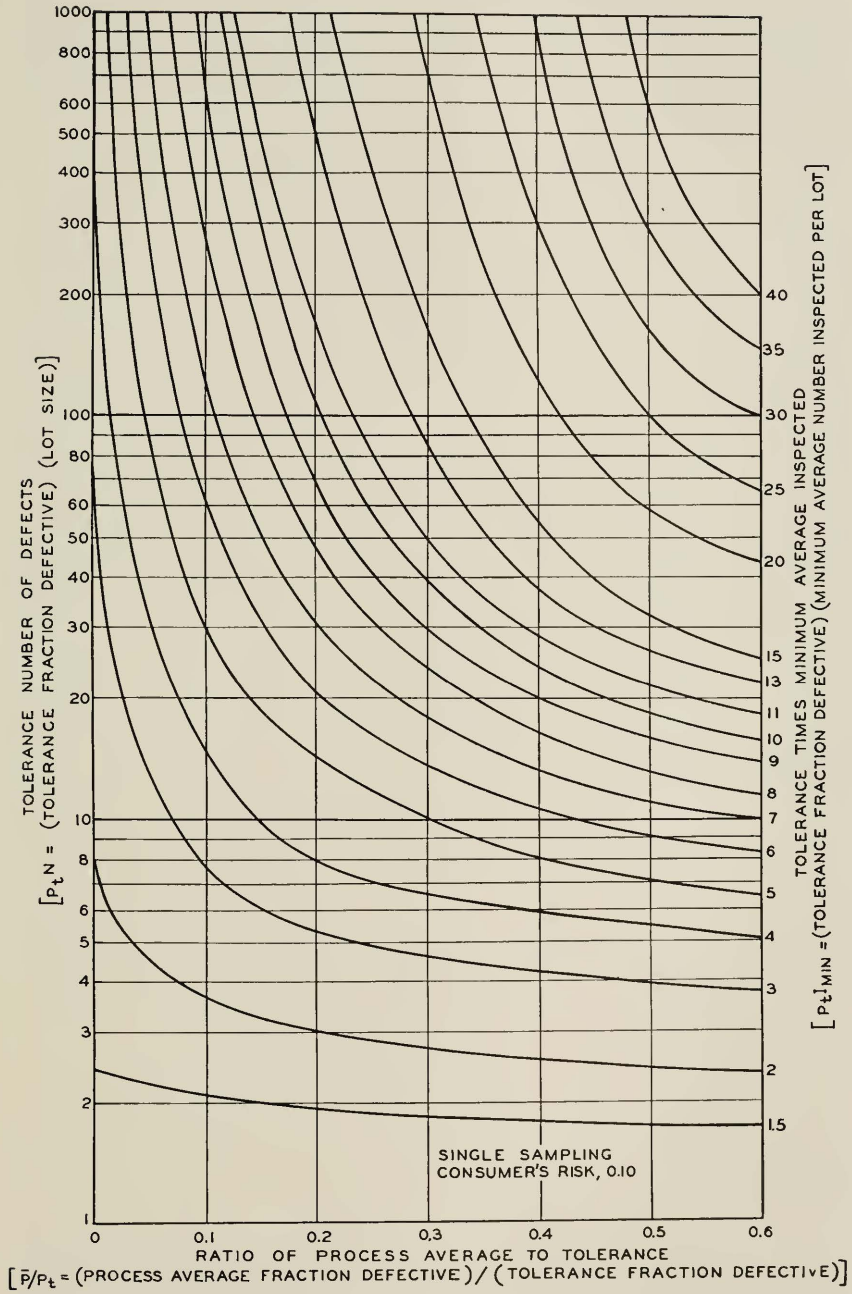


Fig. 4—Curves for finding the minimum amount of inspection per lot.

inspected, such as the degree of control exerted by the producing agencies, or the protective measures which precede or follow the inspection step in question. For any Consumer's Risk value, sets of charts similar to Figs. 2, 3, and 4 may be constructed.

Other Methods of Inspection

The above detailed discussion has been limited to one simple method of inspection, Single Sampling, in order to show how certain concepts and principles may be applied with the aid of sampling theory. The same principles are readily extended to a plan of double sampling or of multiple sampling in which cases a second sample may be examined if the first fails or a third, fourth, etc. examined if the preceding samples fail, before resorting to a detailed inspection of the remainder of a lot. As a matter of fact, for given values of tolerance and risk, the minimum average amount of inspection per lot will be somewhat less for plans which permit the examination of more than one sample before detailing, but when consideration is given to the costs associated with interruption of work, extraction of additional random samples, inconveniences or difficulties in handling the routine called for, etc., it has not been found economical in general to examine more than two samples from any lot.

It may be well to point out that other basically different requirements may be chosen for setting up economical sampling inspection plans. For example a satisfactory method has been devised to meet the following requirements:

(1) A limiting value to the average per cent defective after inspection.
(2) A minimum amount of inspection for product of process average quality.

This method has been found of value in continuous production where the inspection is intended to serve as a partial screen for defective units. It differs from that described above in that it provides a fixed limit to the *average* quality of product after inspection rather than a limit to the quality of each individual lot.

The solution of such problems, which employ probability theory as an aid, always demands a concise statement of the conditions and the specification of numerical requirements which the inspection must satisfy.

MATHEMATICAL APPENDIX

The problem considered is to minimize the average number of pieces inspected per lot in Single Sampling Inspection. The method and equations developed below may be extended to problems involving

2, 3, 4, etc., samples. The method of extension is somewhat complicated although the procedure is identical in nature. For example, in Minimum Double Sampling, first and second acceptance numbers and corresponding first and second sample sizes must be found, and at the same time the total Consumer's Risk must be properly divided between the two samples. We will restrict our attention to the problem of Minimum Single Sampling.

The first seven variables defined below in Table I enter into the two equations needed for the solution of the problem, the first three being fixed by the requirement of the method that a definite protection be provided against accepting faulty material, the fourth being fixed by the requirement that the average amount of inspection shall be a minimum for uniform product of process average quality. Therefore, the three unknown variables are c, n, and I. The five variables N, n, I, $\bar{p}$, and p_t are replaced in the solution by four variables which have been obtained from the original variables by combining p_t with the other four, viz. $M = p_t N$, $a = p_t n$, $z = p_t I$, and $k = \bar{p}/p_t$. Since M, P, and k are specified by the method, the unknown variables are c, a, and z. Two tables showing respectively the notation [3] and the disposition of the variables are presented below.

TABLE I

NOMENCLATURE

N = number of pieces in lot,
P = Consumer's Risk, the probability of accepting a submitted lot of tolerance quality,
p_t = tolerance fraction defective,
$\bar{p}$ = process average (expected) fraction defective,
c = acceptance number, the maximum allowable number of defective pieces in sample,
n = number of pieces in sample,
I = average (expected) number of pieces inspected per lot,
$M = p_t N$ = number of defective pieces in lot of tolerance quality,
$a = p_t n$ = expected number of defective pieces in sample drawn from lot of tolerance quality,
$z = p_t I$ = product of tolerance and average (expected) number of pieces inspected per lot,
$k = \bar{p}/p_t$ = ratio of process average to tolerance,
m = number of defects found in sample,
$C_n^N = \dfrac{N!}{(N-n)!n!}$ = number of combinations of N things taken n at a time.

The solution of the problem requires the consideration of the

[3] The symbols $\bar{p}$ and p_t, as used in this problem, are assumed true parameters of the universe sampled and according to the notation adopted by these Laboratories should be primed, i.e. $\bar{p}'$ and p_t'. For the sake of simplicity here the prime notation has been omitted in the equations. Ref. W. A. Shewhart, "Quality Control," *Bell System Technical Journal*, Vol. VI, p. 723, footnote 3, October, 1927.

following two equations:

$$z = f_1(M, a, c, k), \tag{1}$$

$$P = f_2(M, a, c), \tag{2}$$

where f_1 and f_2 represent symbolic functions which are to be determined later.

We wish to find a pair of values (c, a) which will make z a minimum,

TABLE II

Disposition of Variables

Initial Variables Involved	Initial Fixed Variables	Initial Unknown Variables	Variables Used in Method	Fixed Variables of Method	Unknown Variables of Method
N	N		$M = p_t N$	M	
P	P		P	P	
p_t	p_t		$k = \bar{p}/p_t$	k	
$\bar{p}$	$\bar{p}$		c		c
c		c	$a = p_t n$		a
n		n	$z = p_t I$		z
I		I			

subject to the condition that this pair (c, a) satisfies equation (2). Hence, due to the discreteness of c, pairs (c, a) satisfying (2) are substituted in (1) until a minimum value of z is found. Thus, for $P = .10$ and for given values of M and k, we read c from Fig. 2, a for this value of c from Fig. 3, and the minimum value of z from Fig. 4.

Basis of Fig. 2 Giving Minimum Acceptance Numbers

In determining the function f_1 involved in equation (1), the average number of pieces inspected per lot, I, is treated as the dependent variable. Since a sample is always taken, n pieces will be inspected from every lot submitted. The number of times that the remainder of the lot $(N - n)$ will be inspected on the average is determined from the expression giving the probability that more than c defective pieces will be found in n. The sample is assumed to be drawn from a product of which a fraction, $\bar{p}$, is defective. The probability that c or less defective pieces will be found in n pieces selected at random from a product containing $\bar{p}$ fraction defective pieces is given by the sum of the first $c + 1$ terms of the Point Binomial $[(1 - \bar{p}) + \bar{p}]^n$. Hence the average number of pieces inspected per lot is determined from the relation,

$$I = n + (N - n)\left[1 - \sum_{m=0}^{m=c} C_m^n (1 - \bar{p})^{n-m} \bar{p}^m\right].$$

For the condition, $\bar{p} < .10$, which is usual in practice, it has been found satisfactory to replace the Point Binomial by the Poisson Exponential.[4] By multiplying both sides of the equation by p_t we obtain z in the form,

$$z = M - (M - a) \sum_{m=0}^{m=c} \frac{(ka)^m e^{-ka}}{m!}, \qquad (1')$$

which is the function f_1 desired.

To obtain f_2, we state the probability of finding c or less defects in a sample n taken from a lot N containing $M = p_t N$ defective pieces. This is given by the equation,

$$P = \frac{1}{C_n^N} \sum_{m=0}^{m=c} C_{n-m}^{N-M} C_m^M.$$

But this equation is too difficult to handle in general computations on a large scale. When $p_t < .10$ and n is sufficiently large, a satisfactory approximation to the above equation may be developed from the first $c + 1$ terms of the Point Binomial, that is

$$P = \sum_{m=0}^{m=c} C_m^n \left(1 - \frac{M}{N}\right)^{n-m} \left(\frac{M}{N}\right)^m, \qquad \text{since} \qquad p_t = \frac{M}{N}.$$

An even better approximation is obtained by interchanging[5] n and M in the latter equation giving the expression,

$$P = \sum_{m=0}^{m=c} C_n^N \left(1 - \frac{n}{N}\right)^{M-m} \left(\frac{n}{N}\right)^m.$$

Since $\frac{a}{M} \equiv \frac{n}{N}$, we obtain the final form,

$$P = \sum_{m=0}^{m=c} C_m^M \left(1 - \frac{a}{M}\right)^{M-m} \left(\frac{a}{M}\right)^m, \qquad (2')$$

which is the function f_2 desired.

Now that we have f_1 and f_2 as expressed in equations (1′) and (2′) we must explain how Fig. 2 was obtained. When $P = .10$, for any pair (M, k), a particular pair (c, a) was found which made z a minimum. The acceptance number c may assume only discrete values since any

[4] G. A. Campbell, "Probability Curves Showing Poisson's Exponential Summation," *Bell System Technical Journal*, Vol. II, pp. 95–113, January, 1923.

[5] Paul P. Coggins, "Some General Results of Elementary Sampling Theory for Engineering Use," *Bell System Technical Journal*, Vol. VII, p. 44, Equation (11), January, 1928.

piece must be considered either as defective or non-defective. Hence minimum values of z ($z_{\min.}$) will be found for many pairs (M, k) for the same value of c. From this it is evident that on an M, k plane there exist zones in which the acceptance numbers are identical. To find the boundary lines of these zones it was noted that for certain pairs (M, k) two pairs of (c, a) exist, giving the same minimum value for z. These values of c were found to differ by 1 in all such cases. Designating in general two such adjacent acceptance numbers as c and $c + 1$ and corresponding values of a which satisfy the Consumer's Risk as a_c and a_{c+1}, we may obtain these boundary curves from the equation,

$$(M - a_c) \sum_{m=0}^{m=c} \frac{(ka_c)^m e^{-ka_c}}{m!} = (M - a_{c+1}) \sum_{m=0}^{m=c+1} \frac{(ka_{c+1})^m e^{-ka_{c+1}}}{m!}.$$

In using the above equation to determine these boundary curves for Fig. 2, the following steps were taken:

(1) Assume values for c and $c + 1$.
(2) Determine a_c and a_{c+1} for a given value of P assuming N to be infinite.
(3) For any given value for k, solve the linear equation in M obtained by substituting the assumed values in the above equation.
(4) Using the value of M thus found, determine the exact values of a_c and a_{c+1} from equation (2′) for $P = .10$ (Fig. 3).
(5) Using the same value of k, again solve the linear equation in M substituting the values of a_c and a_{c+1} obtained from step (4).
(6) If the values a_c and a_{c+1} obtained in step (4) satisfy the value of M thus found, the values of M and k define a point on the boundary curve between two adjacent acceptance numbers. If these values of a do not satisfy the value of M thus determined, steps (4) and (5) may be repeated until the limiting conditions are satisfied.

Basis of Fig. 3 Giving n for Any c

For given values of the Consumer's Risk P and acceptance number c, the sample size n may be obtained from equation (2′) since $a = p_t n$. For the case $P = .10$ values of a are presented in Fig. 3 for selected ranges of M and c.

Basis of Fig. 4 Giving the Minimum Average Number of Pieces Inspected Per Lot

The curves in Fig. 4 represent specific values of $z_{\min.}$ on an M, k plane for $P = .10$. Each curve was obtained by substituting given

values of $z_{\text{min.}}$, k, and (c, a) in equation (1′) and solving for M. If, for this value of M thus found, the selected values of c and a coincide with those read respectively from Figs. 2 and 3, a point was established for the given value of $z_{\text{min.}}$ on the M, k plane. If not, sufficient trials were made until the condition given by Figs. 2 and 3 were met. The curves for $z_{\text{min.}}$ were thus determined. To obtain $I_{\text{min.}}$ it was only necessary to use the relation $I_{\text{min.}} = \frac{z_{\text{min.}}}{p_t}$.

The Frequency Distribution of the Unknown Mean of a Sampled Universe

By E. C. MOLINA and R. I. WILKINSON

In drawing conclusions as to the reliability of the mean of a sample it is important that all relevant information be taken into consideration. The mathematical analysis in this paper is based on the Laplacian Bayes Theorem which implicitly comprehends the results of a sample together with the a priori knowledge available concerning the parameters of the universe.

The discussion is limited to a universe assumed to be normal but whose mean and precision constant are unknown. Several simplifying, yet quite reasonable, assumptions regarding the forms and independence of the a priori frequency distribution of the true mean and standard deviation are incorporated in the analysis so that numerical answers may more easily be deduced.

Conclusions, properly drawn, are usually quite definitely dependent upon the a priori assumptions made, and especially so in the case of small samples. A considerable space is, therefore, devoted to the solution of a problem in which the sample is only five, taking up a wide variety of these a priori assumptions. They give, in consequence, a wide range of numerical results, appearing in the form of probable errors in the mean of the sample. Each set of assumptions is briefly discussed indicating how the sampling technician may be able to make a selection consistent with his a priori knowledge of a particular problem.

EVERY observation or series of observations upon the items composing a "universe" or "population" may be regarded as constituting a sample. We may divide sampling into two broad natural classes, (1) Sampling of Attributes, and (2) Sampling of Variables. The theory of the first class concerns itself with some particular characteristic, such as the color red, which each item of the universe definitely does or does not possess, and endeavors to assign, ultimately, a numerical value to the probability that the number or proportion of the items in the universe having this characteristic lies within any given range. The second division comprehends that wide variety of problems in which each item of the universe displays to a greater or less degree the same particular quality, such as length, weight, or resistance. After having drawn a random sample of items, probability theory is called upon to assert with what likelihood certain important descriptive constants or "parameters" of the universe lie within any given ranges.

In either class the problem is legitimately attacked by means of a posteriori probability theory. This theory makes use of the two important distinct kinds of knowledge which, in varying amounts, are always at hand, namely, (1) a priori or preexisting information regarding the universe and the possible values which the unknown

parameters may assume, and (2) the actual observed value of the studied characteristic in each item of the sample. The a priori information may be meager, in some instances hardly more than the limits between which the parameters must lie, and again, from past experience a great deal may be known about the universe, such as its general form of frequency distribution, the most likely value for each of its parameters to take, and a general feeling that they will not, except in rare cases, lie outside of certain well defined ranges closely bordering their believed most likely values. When the a priori knowledge is meager, more weight must be attached to the results of the sample, but when considerable a priori information is at hand relatively less reliance should be placed in the sample; and in some rare cases it is conceivable that so much is known before the drawings are made that a particular sample, especially if small, would justifiably be disregarded entirely.

The Sampling of Attributes on the a posteriori basis for both infinite and finite universes has already been set forth in these pages at considerable length.[1] The theory of Sampling of Variables when the samples are large becomes usually a matter of assuming that some of the parameters of the sample are sufficiently close to those of the universe that no sensible error will be made in assuming them to be equal. In this case the a priori knowledge of the universe, unless far more exact than is normally found in practice, would exercise but a slight effect in the conclusions which might be drawn, and is therefore quite often properly neglected.

When, for one reason or another, some conclusions are demanded after having taken a small sized sample, it cannot safely be assumed that the sample itself adequately describes the universe, and what a priori knowledge we have must, of necessity, play an important rôle in the determination of any legitimate statements as to the constitution of the universe.

The purpose of this paper is to study in strict accordance with the theory of probability the conclusions which may be drawn concerning the true parameters of the unknown universe after a "sample of variables" of any size has been examined.

The paper is divided into the following five sections:

I. The general equation is given for the a posteriori probability

[1] "Deviation of Random Samples from Average Conditions and Significance to Traffic Men," by E. C. Molina and R. P. Crowell, January 1924. "Some General Results of Elementary Sampling Theory for Engineering Use," by P. P. Coggins, January 1928. This second paper is based on another by Mr. E. C. Molina presented before the Statistical Section of the International Mathematical Congress, held at Toronto in August 1924.

that the true mean of a sampled normal universe lies within a given range.

II. Certain mild restrictions are placed on the general equation of (I) to facilitate its use in practice.

III. The selection of a priori frequency functions in practice is discussed.

IV. A typical example is selected and solved for various a priori existence probability distributions with a discussion of the ranges of errors.

V. Conclusions.

I. The General A Posteriori Equation

It is common, unless information is known to the contrary, to assume that the universe from which the sample is to be made is composed of an infinite number of items all having a particular characteristic whose numerical value from item to item follows the normal frequency law. In the remainder of this paper we shall limit ourselves to a discussion involving only this type of universe. The problem may now be precisely stated:

A set of n observations has been made on a variable quantity drawn from a universe wherein the normal law of errors

$$\left(\frac{h}{\pi}\right)^{1/2} e^{-h(x-m)^2}, \qquad h = 1/2\sigma^2$$

is satisfied but the values of the mean and the precision constant, or standard deviation, are unknown; *before* the observations were made the probability in favor of the simultaneous existence of the inequalities

$$m < \text{mean} < m + dm \tag{1}$$

$$h < \text{precision constant} < h + dh \tag{2}$$

was some function of m and h, say $W(m, h)dmdh$; what is the probability that *after* the observations were made the unknown mean satisfies the inequality (1)?

Let $x_1, x_2 \cdots x_n$ be the values for x given by the n observations. Set

$$n\bar{x} = \sum_1^n x_i, \qquad ns^2 = \sum_1^n (x_i - \bar{x})^2.$$

Now if m and h were known the probability that a set of n observations, *not yet* made, would give values $x_1, x_2, \cdots x_n$ would be

$$\left(\frac{h}{\pi}\right)^{(1/2)n} e^{-h\Sigma(x_i-m)^2} dx_1 dx_2 \cdots dx_n. \tag{3}$$

Therefore, by the Laplacian generalization of the Bayes formula, the a posteriori probability that

$$m < \text{mean} < m + dm$$

is (cancelling factors which do not involve m or h)

$$P(m)dm = \frac{dm \int_0^\infty W(m, h)h^{(1/2)n}e^{-h\Sigma(x_i-m)^2}dh}{\int_{-\infty}^\infty dm \int_0^\infty W(m, h)h^{(1/2)n}e^{-h\Sigma(x_i-m)^2}dh}^2$$

$$= Adm \int_0^\infty W(m, h)h^{(1/2)n}e^{-h\Sigma(x_i-m)^2}dh, \tag{4}$$

where A is a constant such that

$$\int_{-\infty}^\infty P(m)dm = 1.$$

II. Introduction of Restrictions on General Equation

We are now confronted by a difficulty inherent to a posteriori probability problems. What do we know as to the form of the a priori existence probability function $W(m, h)$? If in a specific practical problem the form of $W(m, h)$ is unknown, no conclusions can be drawn from the set of observations *unless some assumptions are made* and then the weight assignable to the conclusions drawn is a delicate question depending on the reasonableness of the assumptions.[3]

The analysis and results given below are based on assumptions which the writers believe will be found justifiable in many problems of practical interest.

A first assumption which suggests itself is that m and h are independent a priori so that we may write

$$W(m, h) = W_1(m)W_2(h). \tag{5}$$

On this assumption

$$P(m)dm = AW_1(m)dm \int_0^\infty W_2(h)h^{(1/2)n}e^{-h\Sigma(x_i-m)^2}dh. \tag{6}$$

As a second step toward tentative solutions assume that

$$W_2(h) = Kh^{(1/2)c}e^{-ah}, \tag{7}$$

[2] See Poincare: "Calcul des Probabilites"; 2d edition; articles 178 and 179.

[3] In this connection see italicized paragraph, page 266, "Probability and Its Engineering Uses," T. C. Fry, 1928.

where K, c and a are constants. This, by means of the change of variable

$$y = h[a + \sum(x_i - m)^2]$$

and throwing the definite integral

$$\int_0^\infty y^{(1/2)(n+c)} e^{-y} dy$$

in with the constant A, reduces (6) to

$$P(m)dm = A'W_1(m)[a + \sum(x_i - m)^2]^{-(1/2)(n+2+c)} dm. \qquad (8)$$

We are still confronted with the a priori existence probability function $W_1(m)$.

A plausible form, suggested by the well known "Student" [4] distribution of the ratio $(\bar{x} - m)/s$ for a set of observations *to be* made from a normal universe of *known* mean and standard deviation, is

$$W_1(m) = A_1[1 + B(M - m)^2]^{-(1/2)N}, \qquad (9)$$

where M is the value of m which is a priori most probable, N and B are positive constants while the equation

$$\int_{-\infty}^{\infty} W_1(m)dm = 1$$

gives

$$A_1 = \frac{B^{1/2}\Gamma(\frac{1}{2}N)}{\pi^{1/2}\Gamma[\frac{1}{2}(N-1)]}.$$

With this assumed form and noting that

$$\sum(x_i - m)^2 = ns^2 + n(\bar{x} - m)^2$$

equation (8) gives

$$P(m)dm = A''[1 + B(M - m)^2]^{-(1/2)N} \times \left[1 + \left(\frac{ns^2}{a + ns^2}\right)\left(\frac{\bar{x} - m}{s}\right)^2\right]^{-(1/2)(n+2+c)} dm, \qquad (10)$$

the integral of $P(m)dm$ between plus and minus infinity determining A''.

Recapitulating: formula (10) gives us the a posteriori frequency distribution for m in terms of the observed data and the arbitrary constants a, c, N, B, M which have entered into the problem in

[4] The writers are aware of the fact that the "Student" frequency function has been put forward in more than one place as the solution for an a posteriori problem. But it should be noted that the various *deductions* of this function which have been given by "Student" and others are entirely a priori.

consequence of the three assumptions made regarding the form of the a priori existence probability function $W(m, h)$; the three assumptions being embodied in equations (5), (7) and (9).

III. Practical Selection of A Priori Frequency Distributions

In equation (10) we have first to assign a numerical value to each of the five constants a, c, N, B, M, before the probability $P(m)$ can be evaluated for any desired range of m. Obviously, in actual practise, the selection of their values is extremely important and too much care cannot be exercised in an attempt to satisfy the engineering judgment that all of the a priori information at hand has been nicely comprehended.

In an endeavor to reduce the number of constants to which we must assign values we shall consider first the a priori function

$$W_2(h) = Kh^{(1/2)c}e^{-ah}.$$

Setting

$$h = c/2a \tag{11}$$

makes $W_2(h)$ a maximum. On the other hand, the value of h which would make the observed set of values of x most probable is given by the equation

$$\frac{1}{h} = \frac{2\sum(x_i - m)^2}{n},$$

or, if m be set equal to $\bar{x}$, we obtain the simpler equation

$$h = 1/2s^2. \tag{12}$$

Upon eliminating h from (11) and (12),[5]

$$a = cs^2. \tag{13}$$

In Fig. 3 are shown four frequency curves of $W_2(h)$. Curve I is plotted for $c = 3$ according to equation (13), and to illustrate the wide possibility of forms, curves II and III have been constructed, keeping $c = 3$, after changing equation (13) to

$$a = \frac{cs^2}{1 - .1s^2} \quad \text{and} \quad a = \frac{cs^2}{1 + .1s^2},$$

respectively. Curve IV again satisfies equation (13) but has c increased from 3 to 6.

[5] It should be carefully noted that there is no necessary relation between the *a priori most probable value of h* and the *value of h which would make the observed event most probable.* The elimination of h between (11) and (12) is justified solely by the practical consideration that a tentative relation between a and c will reduce by one the number of arbitrary constants to which numerical values must be assigned.

For every assumption of a and c in the a priori distribution of h there is, of course, a corresponding a priori distribution, $\phi(\sigma)$, of the standard deviation. Here

$$\phi(\sigma)d\sigma = K'\sigma^{-(c+3)}e^{-(1/2)a/\sigma^2}d\sigma$$

and

$$K' = \frac{a^{(1/2)c+1}}{2^{(1/2)c}\Gamma(\frac{1}{2}c + 1)} .$$

The distributions of σ corresponding to each of the four frequency curves of h in Fig. 3 are shown in Fig. 4 with similar designations.

In many cases, too, it is obvious that very little is known concerning the shape and the parameters of the mean's a priori distribution beyond that it is generally unimodal and quite likely to be fairly symmetrical about its most probable value; a mathematical expression of this has been set up in equation (9). In this circumstance we may not introduce serious restrictions if we make two further assumptions which greatly simplify (9).

The first is that we set $M = \bar{x}$ which says that, a priori, the most probable value of the unknown mean was the same as that which was later calculated as the mean in the sample.[6] It is admitted that the chance of exactly fixing on $M = \bar{x}$ from a priori information is very small, yet if our knowledge is so slight that we must introduce some guesswork here, the selection of the value of $\bar{x}$ at least has the advantage of being a *possible* one which M might assume and, except in rare cases, it will not be greatly distant from the true m in that particular lot. The logical difficulty here also may be minimized by selecting a form of $W_1(m)$ of such flatness that over a considerable range of values in the neighborhood of $\bar{x}$ the existence probability does not take on widely differing magnitudes.[7]

The second assumption can more readily be allowed, and consists in empirically defining

$$B = \frac{n}{a + ns^2} .$$

This removes a degree of freedom from the $W_1(m)$ function but, as far as its form is concerned, except in special cases, the one variable, N, may serve quite well in characterizing the pre-existing information. As is clearly shown in Fig. 1, the increase of N indicates a greater

[6] While it does not matter in this particular problem, the authors wish to carefully distinguish, at least in thought, between an "observed" parameter and a parameter calculated from individual observations.

[7] The setting of $M = \bar{x}$, it should be noted, has no effect if all values of the mean are made a priori equally likely by setting $N = 0$ (that is, $W_1(m) = A_1$).

certainty in the investigator's mind that the true value of m lies closer and closer to the assumed most probable figure, M.

With these two assumptions incorporated in equation (10) we may now write

$$P(m)dm = f(t)dt = A'''(1 + t^2)^{-(1/2)T}dt, \tag{10'}$$

in which

$$t^2 = \left(\frac{\bar{x} - m}{s}\right)^2\left(\frac{ns^2}{a + ns^2}\right) = B(\bar{x} - m)^2, \tag{14}$$

$$T = n + 2 + c + N,$$

$$A''' = \frac{1}{\int_{-\infty}^{\infty}(1 + t^2)^{-(1/2)T}dt} = \frac{\Gamma(\frac{1}{2}T)}{\pi^{1/2}\Gamma[\frac{1}{2}(T - 1)]}.$$

The formula (10') is a "Student" [8] frequency form with the arguments n and $\frac{\bar{x} - m}{s}$ replaced by $n + 2 + c + N$ and $\frac{\bar{x} - m}{s(1 + a/ns^2)^{1/2}}$ respectively.

Fig. 2 shows curves plotted for ranges of t such that

$$A'''\int_{-t}^{+t}(1 + t^2)^{-(1/2)T}dt = .50, .80, .90, \text{ and } .9973,$$ [9]

and the errors in the mean corresponding to any of these probabilities, after determining t, may be found by evaluating $\bar{x} - m$ in equation (14).

IV. Solution of a Typical Example

Five samples of retardation coils rated at 47 ohms are taken from a large lot, and careful measurements show them to have resistances of 46.30, 44.40, 47.72, 50.50, and 45.58 ohms respectively. We are asked to determine the probable and 99.73 per cent errors of the average of these resistances, assuming that the samples have been drawn from a normal universe.

The average of these five values is $\bar{x} = 46.90$ ohms and their standard deviation about this average is found to be $s = 2.097$.

From the preceding discussion it is evident that as many answers to this problem may be obtained as there are assumptions made regarding, in general, the a priori distributions of the mean and

[8] Student: "The Probable Error of a Mean," Biometrika, Vol. VI, No. 1, March 1908.

[9] Student: "New Tables for Testing the Significance of Observations," Metron, Vol. V, No. 3, I–XII–1925. Tables I and II, pages 114–118, for values of $n' = 2$ to 21.

precision constant, and in our particular analysis the five constants found in equation (10). In Table I and Fig. 1 we tabulate and portray graphically twenty-one complete solutions of the example based upon as many sets of values given to these constants. A wide

TABLE I

$\bar{x} = 46.90 \quad n = 5 \quad s = 2.097$

No.	Parameters in Existence Probability Distributions							$T = n + 2 + c + N$	A Posteriori Probability Values of Errors in Observed Mean	
	Mean				Precision Constant				Probable or 50% Error	99.73% Error
	M	N	A_1	B	c	a	K			
1	46.9	0	—	.2274	−3	0	—	4	.923	11.16
2	46.9	0	—	.2274	−2	0	—	5	.776	6.94
3	46.9	0	—	.2274	0	0	—	7	.623	4.23
4	46.9	0	—	.1421	3	13.19*	475.5	10	.626	3.61
5	46.9	0	—	.1098	3	23.55†	2,024	10	.712	4.10
6	46.9	0	—	.1605	3	9.163‡	191.2	10	.589	3.39
7	46.9	0	—	.1034	6	26.38*	80,770	13	.625	3.39
8	46.9	1	—	.2274	0	0	—	8	.566	3.59
9	46.9	1	—	.1421	3	13.19*	475.5	11	.587	3.33
10	46.9	2	.1518	.2274	0	0	—	9	.524	3.17
11	46.9	2	.1200	.1421	3	13.19*	475.5	12	.558	3.08
12	46.9	2	.1055	.1098	3	23.55†	2,024	12	.635	3.50
13	46.9	2	.1275	.1605	3	9.163‡	191.2	12	.525	2.90
14	46.9	2	.1023	.1034	6	26.38*	80,770	15	.575	3.01
15	46.9	4	.3036	.2274	0	0	—	11	.464	2.63
16	46.9	4	.2400	.1421	3	13.19*	475.5	14	.511	2.71
17	46.9	4	.2110	.1098	3	23.55†	2,024	14	.581	3.09
18	46.9	4	.2551	.1605	3	9.163‡	191.2	14	.480	2.55
19	46.9	4	.2047	.1034	6	26.38*	80,770	17	.537	2.76
20	41.0	2	.1200	.1421	3	13.19*	475.5	—	.665§	3.87§
21	49.0	2	.1200	.1421	3	13.19*	475.5	—	.635§	3.65§

* $a = cs^2$.

† $a = \dfrac{cs^2}{1 - .1s^2}$.

‡ $a = \dfrac{cs^2}{1 + .1s^2}$.

§ Errors determined by planimeter method.

variety of a priori conditions is assumed giving, in consequence, widely varying probable and 99.73 per cent errors.

The a priori frequency functions, $\phi(\sigma)$, for the standard deviation in the first seven cases are shown in broken lines superposed upon the distributions of precision constants in Fig. 1. The scales of h and σ are not to be confused, the attempt being only to represent the *form* of the $\phi(\sigma)$ frequency curves.

(*a*) If we wish to be very conservative we might select values for the unknown constants which would make all values of m and σ

APRIORI PROBABILITY DISTRIBUTIONS

ERRORS IN OBSERVED MEAN

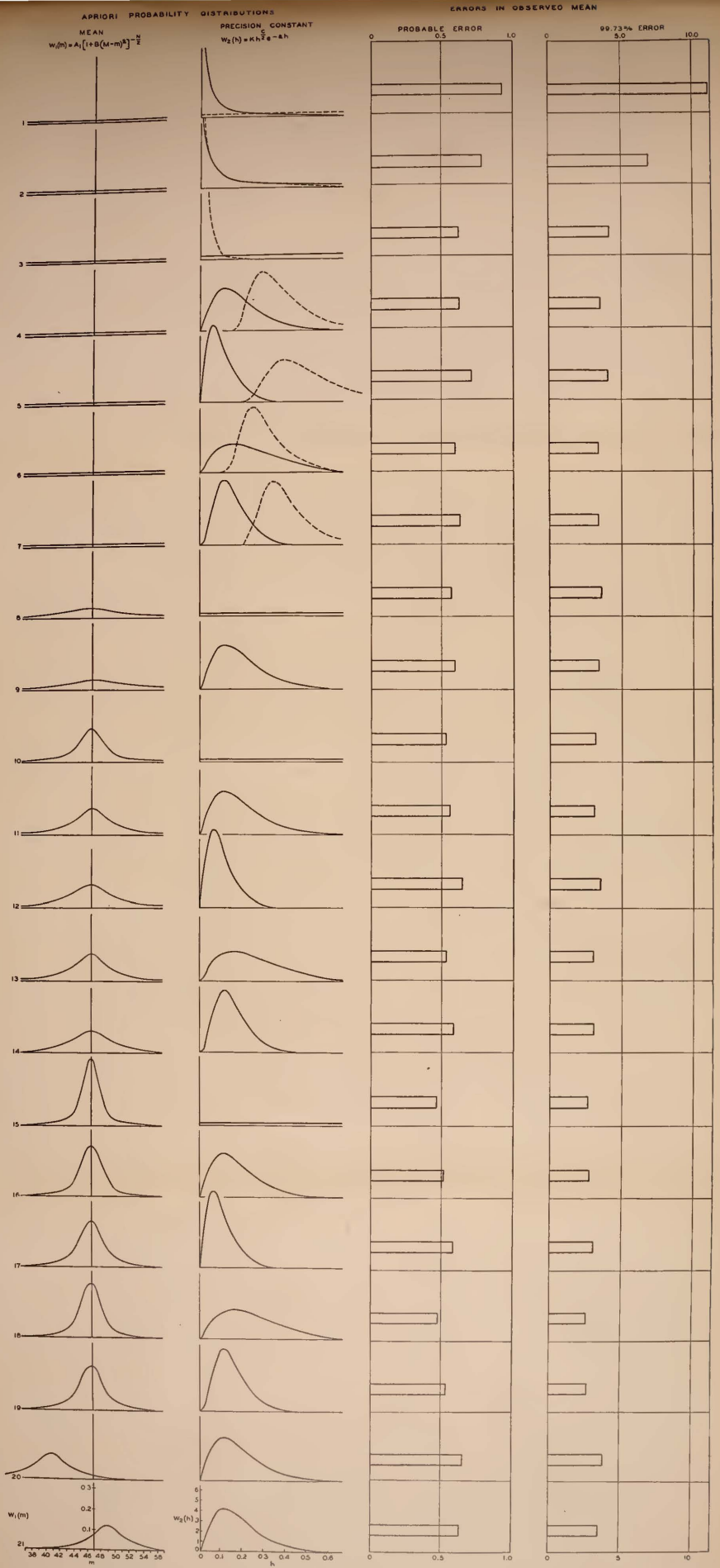

APRIORI PROBABILITY DISTRIBUTIONS
MEAN
$w_1(m) = A_1\left[1+B(M-m)^2\right]^{-\frac{n}{2}}$
PRECISION CONSTANT
$w_2(h) = Kh^{\frac{n}{2}}e^{-Ah}$
ERRORS IN OBSERVED MEAN
PROBABLE ERROR
99.73% ERROR
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
$w_1(m)$
36 38 40 42 44 46 48 50 52 54 56
m
$w_2(h)$
0 0.1 0.2 0.3 0.4 0.5 0.6
h
0 0.5 1.0
0 5.0 10.0
0 0.5 1.0
0 5 10

FIG. 1.

equally likely, that is, $N = 0$, $c = -3$, $a = 0$.[10] Here the precision constant's a priori distribution is decidedly exponential and we might predict the large probable and 99.73 per cent errors in the observed average which actually result.

Case 1 in Table I and Fig. 1 presents the problem in its entirety with the resultant errors tabulated as well as shown graphically.

(*b*) The engineer's knowledge, however, in all probability, is not so limited as in (*a*) above, at least regarding the precision constant

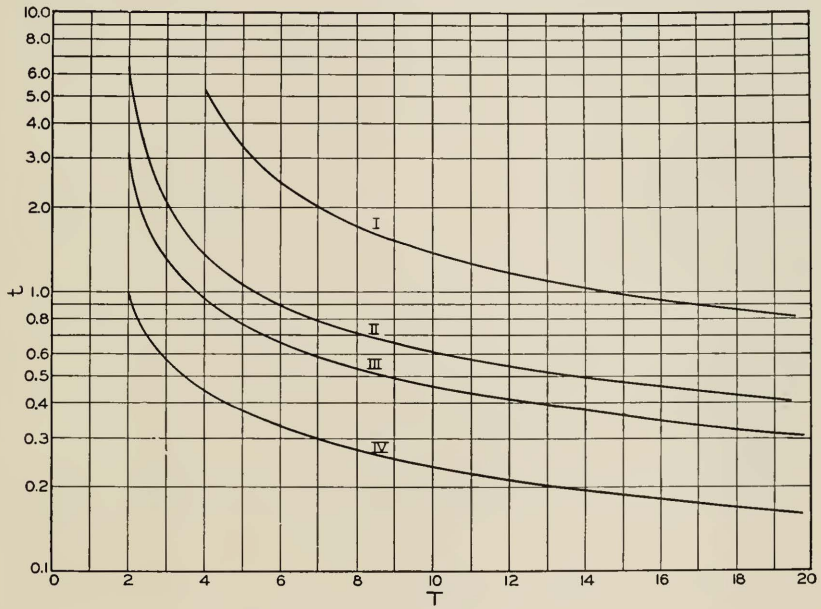


Fig. 2—Errors of averages of samples of size n.
I—99.73 per cent Error.
II—90.00 per cent Error.
III—80.00 per cent Error.
IV—50.00 per cent Error.

Note: Abscissa: $T = n + 2 + c + N$.
Ordinate: t = The Product of the Error of the Average and the Square Root of B.

(or the standard deviation). He knows that extremely small values of the precision constant are less likely than larger ones, and to some extent we picture the transition from (*a*) to this impression in the Cases Nos. 2 and 3 which as before may be found completely portrayed

[10] The formula for $P(m)$ resulting from a substitution of these constants in equation (10′) reproduces the result obtained by Drs. J. Neyman and E. S. Pearson for all values of the a priori function $W'(m, \sigma)$ equally likely: Biometrika, Vol. XX*A*, Parts I and II, July 1928; "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference," page 196, equation XXXV.

equally likely, that is, $N = 0$, $c = -3$, $a = 0$.[10] Here the precision constant's a priori distribution is decidedly exponential and we might predict the large probable and 99.73 per cent errors in the observed average which actually result.

Case 1 in Table I and Fig. 1 presents the problem in its entirety with the resultant errors tabulated as well as shown graphically.

(*b*) The engineer's knowledge, however, in all probability, is not so limited as in (*a*) above, at least regarding the precision constant

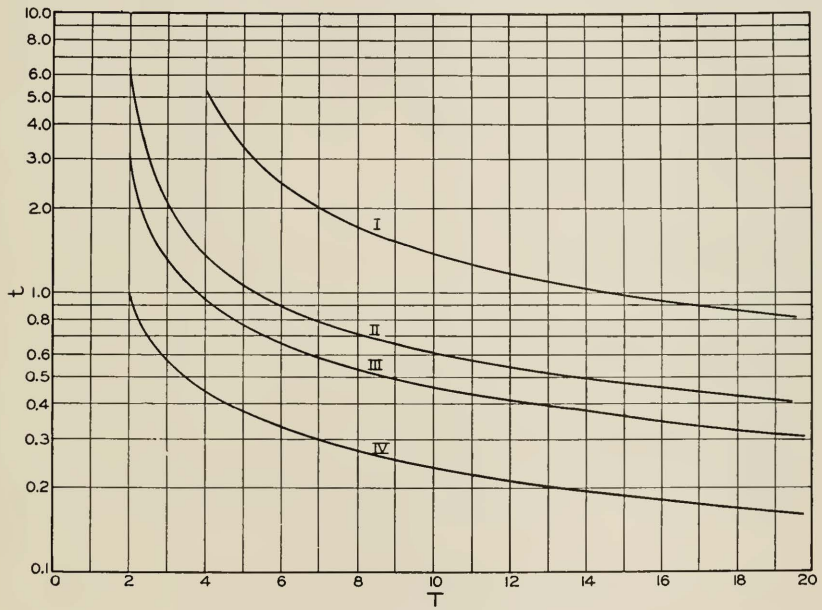


Fig. 2—Errors of averages of samples of size n.
I—99.73 per cent Error.
II—90.00 per cent Error.
III—80.00 per cent Error.
IV—50.00 per cent Error.

Note: Abscissa: $T = n + 2 + c + N$.
Ordinate: t = The Product of the Error of the Average and the Square Root of B.

(or the standard deviation). He knows that extremely small values of the precision constant are less likely than larger ones, and to some extent we picture the transition from (*a*) to this impression in the Cases Nos. 2 and 3 which as before may be found completely portrayed

[10] The formula for $P(m)$ resulting from a substitution of these constants in equation (10′) reproduces the result obtained by Drs. J. Neyman and E. S. Pearson for all values of the a priori function $W'(m, \sigma)$ equally likely: Biometrika, Vol. XX*A*, Parts I and II, July 1928; "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference," page 196, equation XXXV.

in Table I and Fig. 1. Case No. 2, it is interesting to note, is the familiar "Student" formula; Case No. 3's outstanding characteristic is that all values of h are a priori equally likely. The errors in the mean have here been greatly reduced by merely changing the existence probability distribution of the precision constant.

(*c*) Again, the experienced analyst is quite likely to assume willingly that the distribution of the precision constant (and likewise the standard deviation) is of a unimodal form having its maximum value not greatly distant from the figure determined in the sample. Cases Nos. 4 to 7 inclusive typify this kind of assumption while, at the same time, all values of the true mean are held a priori, equally likely.

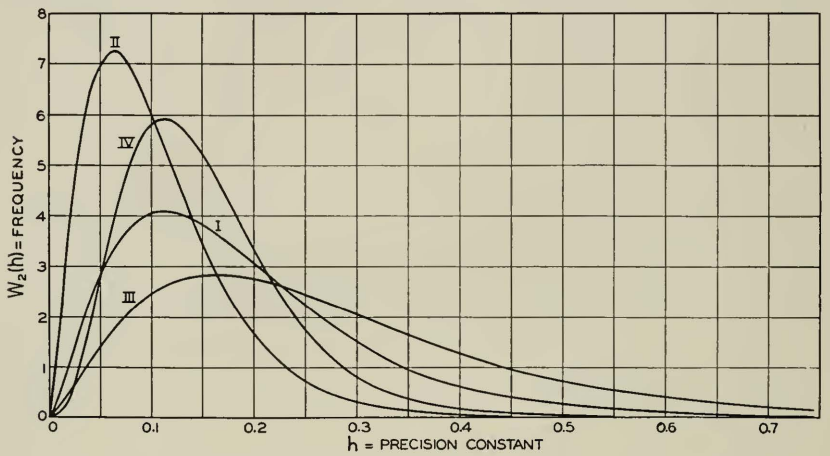


Fig. 3—Typical a priori frequency distributions of the precision constant.

$$W_2(h) = Kh^{(1/2)c}e^{-ah}.$$

I—$c = 3$, $a = cs^2 = 13.1922$.

II—$c = 3$, $a = \dfrac{cs^2}{1 - .1s^2} = 23.5467$.

III—$c = 3$, $a = \dfrac{cs^2}{1 + .1s^2} = 9.1629$.

IV—$c = 6$, $a = cs^2 = 26.3845$.

The constants for Cases Nos. 4 and 7 have been so selected as to bring the modal value of h at that found from the sample, that is, that value of h has been made most likely a priori which will make the probability of occurrence of the particular value $(1/2s^2)$ calculated from the observations, a maximum. Case No. 7 is a considerably more peaked distribution than Case No. 4 indicating more faith in the modal figure selected as being close to the true value. Cases Nos. 5 and 6 illustrate how the mode of the $W_2(h)$ function which always lies at $h = c/2a$ may be shifted either down or up and the extent of modification in the resulting errors which may be expected.

The four frequency distributions of h just considered are the same as those shown in more detail in Fig. 3; the corresponding distributions of the standard deviation are found detailed on Fig. 4.

(*d*) If the interpreter of the data is closely familiar with the sampled product and has been observing similar lots for some time he may have a reasonably good idea as to the value of the general average

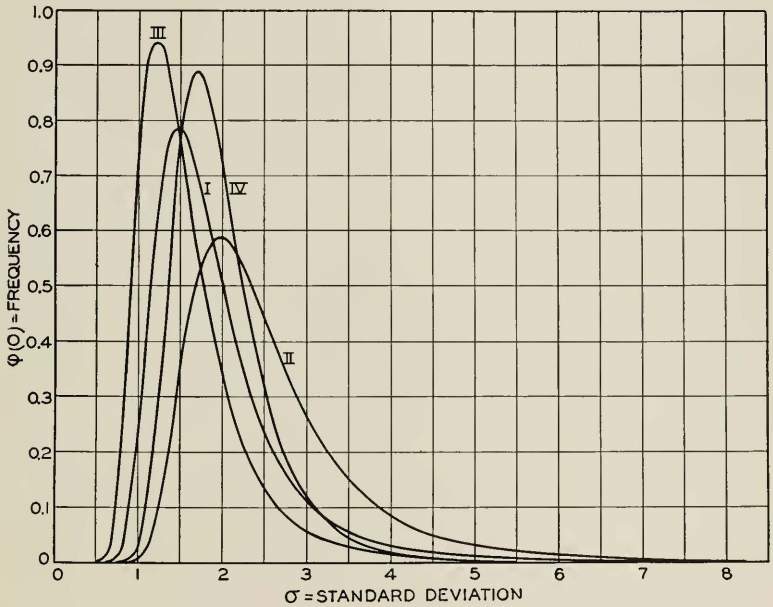


Fig. 4—Typical a priori frequency distributions of the standard deviation.

$$\phi(\sigma) = K'\sigma^{-(c+3)}e^{-(1/2)a/\sigma^2}.$$

$$\text{I}—c = 3,\ a = cs^2 = 13.1922.$$

$$\text{II}—c = 3,\ a = \frac{cs^2}{1 - .1s^2} = 23.5467.$$

$$\text{III}—c = 3,\ a = \frac{cs^2}{1 + .1s^2} = 9.1629.$$

$$\text{IV}—c = 6,\ a = cs^2 = 26.3845.$$

of items produced under these same essential conditions. In Cases Nos. 8 to 19, inclusive, use is made of this knowledge on the assumption that $\bar{x}$, the mean of the sample, turns out to be so nearly equal to M, the most likely a priori value of the true mean m, that we may safely call them identical. Three values of N, regulating the spread of the $W_1(m)$ distribution to conform to the observer's best judgment of the true circumstances have been associated with the same sequence of a priori assumptions regarding the precision constant as were presented in Cases Nos. 3 to 7.

The errors found in Cases Nos. 8 to 19 on the various combinations of a priori frequency curves lie in a fairly narrow band distinctly below those determined from the more conservative assumptions underlying Cases Nos. 1, 2 and 3. This well illustrates the importance of carefully surveying and as far as possible completely utilizing the knowledge available before the sample has been made.

(*e*) Finally, cases are bound to occur in which the engineer can quite definitely say that some value of M other than $\bar{x}$ is a priori most probable; this situation is encountered in Cases Nos. 20 and 21. These are identical with Case No. 11 except that in Case No. 20, M has been reduced about 6 ohms and in Case No. 21 raised about

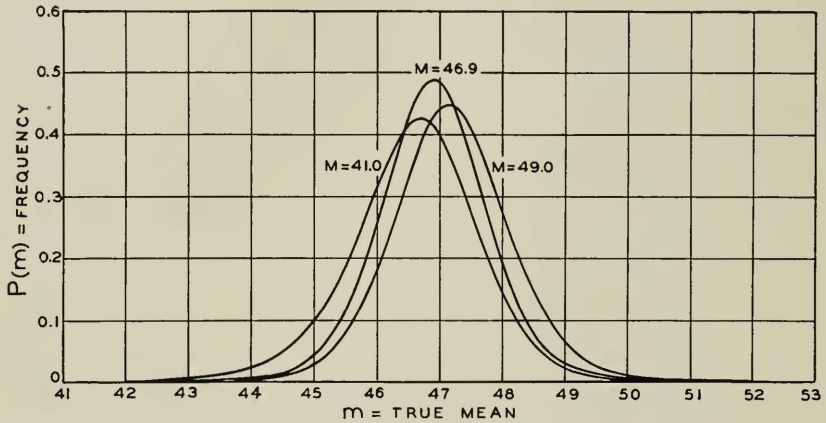


Fig. 5—Typical a posteriori frequency distributions of the unknown mean.
$P(m) = A''[1 + B(M - m)^2]^{-(1/2)N}\left[1 + \left(\frac{ns^2}{a + ns^2}\right)\left(\frac{\bar{x} - m}{s}\right)^2\right]^{-(1/2)[n+2+c]}$.

2 ohms. The errors are somewhat increased by these changes in M, as, of course, we should have predicted. Comparisons such as this should help the investigator to decide whether or not his previously selected figure for M is sufficiently close to $\bar{x}$ that they may safely be equated.

In the event that it is decided that M may not be set equal to $\bar{x}$, in any particular problem, as in Cases Nos. 20 and 21, the symmetrical "Student" form of distribution for $P(m)$, (except when $N = 0$) no longer occurs. This is clear from an inspection of Fig. 5 which shows the three cases plotted on the same scale.

It is suggested then, since the integral of $P(m)dm$ here may become difficult to handle, that recourse be had to the use of a planimeter on the distribution plotted from equation (10) on rectangular co-ordinate paper. In this way may be determined within what range,

equidistant above and below $\bar{x}$, lies the proportion of the total area corresponding to the desired probability.[11]

V. Conclusions

We have presented a general equation for the probability that the true mean of a sampled normal universe lies within a given range, incorporating the kind of knowledge the investigator may be expected to have before the sample was made as well as the information directly presented by the individual observations themselves. It cannot be overemphasized that the problem by its very nature is indefinite since it would be a rare instance indeed to find a mathematical expression which would completely and exactly summarize the a priori knowledge, impressions and beliefs in the mind of any person confronted with its solution. All that can be found is, at best, an approximate probability based upon certain assumptions we are willing to make in order to arrive at a numerical result. And only by utilizing as far as possible all of the available knowledge will the most nearly correct probability values ascertainable be realized.

[11] On certain test cases of "Student" distributions, the error in planimeter readings averaged about one-half of one per cent, and in no case exceeded one per cent.

Speech Power and Its Measurement [1]

By L. J. SIVIAN

The paper is chiefly concerned with the important speech power quantities—frequency spectra, distributions of instantaneous, average, syllabic and peak amplitudes, etc.—as they obtain in actual speech for a large range of voices, talking levels, and subject matters. The analysis is not nearly so complete nor so fine-grained as that which, in principle, can be derived from oscillographic records of individual speech sounds. Its advantage is in the speed with which data can be secured, under widely varying conditions and on a scale which warrants statistical conclusions. Some of the methods in use for measurements of this type are described. A "level analyzer" has been developed, primarily for the measurement of average and peak pressure amplitudes in speech and music, both as to magnitude and as to position in the frequency spectrum. Illustrative results are given for samples of speech, music and noise.

SPEECH sounds are so variable from one to another, from one individual to another, from one conversation to another, as to make it necessary to deal with speech power in a statistical manner. This is particularly true of engineering applications, as distinguished from studies in phonetics and voice dynamics. It is largely from the viewpoint of the former that the subject is here treated.

Speech power occurs in several states, e.g. acoustic, electric, magnetic, mechanical, optical, thermal, etc., but its measurement largely refers to speech in the acoustic or in the electric form. Acoustically, the simplest quantity to define is the instantaneous power transmitted through unit area tangent to the wave front. That power is $L = P \cdot U$, where P is the pressure and U the air particle velocity. With P expressed in bars (dynes per cm.^{2}), and U in cm./sec., L is given in ergs/cm.$^2 \times$ sec. We do not directly measure the product $P \cdot U$. No suitable means for doing that has been developed. In a progressive plane wave, or with good approximation in any progressive wave at sufficient distance from the source, P and U are in phase, and the expression for L simplifies down to $L = P^2/\rho c. = U^2 \cdot \rho c$, where ρc is a constant equal to 41.5 mechanical ohms—the sound radiation resistance of air. Hence for the type of waves mentioned the wattmeter type of measurement (pressure $\times$ velocity, corresponding to voltage $\times$ current in the electrical case) may be replaced with the simpler measurement of pressure (voltage) or velocity (current) alone.

What are the acoustic voltmeters and ammeters that are capable

[1] Presented before Acoustical Society of America, May 11, 1929.

of measuring the instantaneous pressures and velocities respectively? First, as to acoustic ammeters: perhaps the two best known forms are the Rayleigh disk and the hot-wire microphone. The disk gives absolute values of U but its response is so slow that it can be used only to measure the effective values in comparatively sustained sound waves lasting, say one second or longer, or for the ballistic integration of shorter pulses. Hence it does not give a measure of the instantaneous velocities for even the slowest audio-frequency vibration. Furthermore, its use when exposed to the speaker in an open room, is rendered difficult by its susceptibility to spurious air currents. The only application of the disk to vocal power measurements which has come to my notice, is one by Prof. Zernov,[2] published in 1908. For sustained loud singing and shouting he found energy densities ranging from 0.3 to 2.0×10^{-4} ergs/cm.^{3}, at 2 meters distance from the singer. Assuming uniform distribution over a hemisphere of 2 m. radius, this gives a total power output of the voice of the order of 50,000 microwatts. At 2 m. distance reflections from the walls of the room materially raise the energy density as compared with that in a progressive wave.

Another acoustic ammeter is the hot-wire microphone. The resistance variations of the wire tend to follow the instantaneous values of the air particle velocity in the sound wave, but the sensitivity is a complicated function of the frequency, rapidly decreasing as the latter increases. Hence it does not give true oscillograms of speech sounds which in general include a frequency range of six or seven octaves.

Still another device in this general class is the glow-discharge microphone. Its response is determined largely by the amplitude of the air particle motion (E. Meyer, E.N.T., v. 6, 17-21, 1929). Hence, for constant sound intensity the response is inversely proportional to the frequency, roughly. No method for its absolute calibration has been proposed other than comparison with a microphone of known performance. The frequency response and the somewhat erratic behavior of the device in its present status, render it rather unsuitable for speech power studies.

Nearly all our information concerning speech power and the waveform of speech sounds has been obtained with acoustic voltmeters, i.e., with devices responding to pressure in the sound wave. To a first approximation, the ear belongs to this class. It includes the resonators which Helmholtz used in his vowel studies. Generally, the vital element in these devices is a diaphragm which vibrates with

[2] *Ann. der Phys.*, *26*, 94, 1908.

the alternating sound pressure and whose motion is converted (indirectly, as a rule) into a visual record. These acoustic voltmeters have gradually evolved from Scott's phonautograph and Koenig's manometric capsules and indicating flames of some 70 years ago to the present day technique. Their extensive use accounts for the fact that sound measurements frequently are expressed in terms of pressure rather than in terms of power. In many cases, when the relation between pressure and velocity is not known or too complex a function of frequency to be manageable,—the description of the sound wave must be confined to pressure values alone.

The first requisite for absolute measurements of speech pressures is a microphone which admits of an absolute electroacoustic calibration over the range of audio frequencies, and which has a substantially uniform sensitivity over the most important part of that range. The development of the condenser microphone and of the thermophone method for calibrating it, supplied this need. This fundamental contribution to the subject is due to E. C. Wente.[3] Using such a calibrated condenser microphone in conjunction with a vacuum tube amplifier and an oscillograph which were uniformly sensitive up to about 6000 p.p.s., I. B. Crandall and C. F. Sacia,[4] and I. B. Crandall [4] obtained oscillograms of the fundamental speech sounds in the English language. Essentially, these oscillograms give a picture of the instantaneous pressures throughout the duration of the sound, impressed on the microphone diaphragm at 2.5 cm. from the speaker's lips. A certain amount of similar work on German speech sounds has been published by Trendelenburg.[5] From the standpoint of phonetics these pressure amplitude oscillograms of individual sounds are a most comprehensive source of information. Their chief use has been in determining the frequency spectra of the fundamental speech sounds. Sacia [6] and Sacia and Beck [7] have used them to compute the mean and the peak powers of those sounds.

From the standpoint of engineering applications, speech power has a twofold interest: (*a*), there is the question of designing microphones, receivers, circuits, room acoustics, of controlling noise levels, etc., (*b*), the control of apparatus (chiefly adjustment of amplification) while it is handling speech from persons who acoustically are not controllable.

[3] *Phys. Rev.*, July 1917, April 1922, June 1922.
[4] *Bell Sys. Tech. Jour.*, April 1924 and Oct. 1925, resp.
[5] Wiss. Veroff. a. d. Siemens-Konzern, 1924 and 1925.
[6] *Bell Sys. Tech. Jour.*, Oct. 1925.
[7] *Bell Sys. Tech. Jour.*, July 1926.

In connection with (*a*) we can use data obtained under "laboratory" conditions, i.e. with the speakers, their voice levels, the talking speed, etc., suitably selected. Perhaps the simplest statistical characteristic

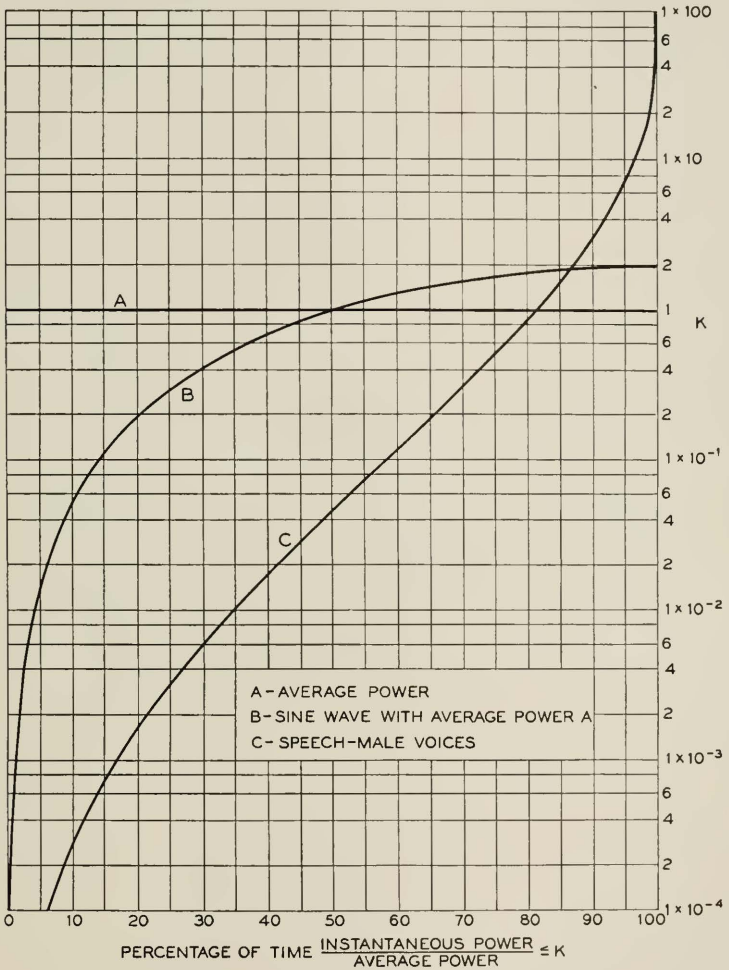


Fig. 1—Time distribution of speech power.

of speech is the average power, i.e. the ratio of the total energy in a large number of speech sounds divided by the total speech time. The quantity immediately measured is the pressure on the transmitter diaphragm. The corresponding power flow through unit area is computed. In doing so it is roughly assumed that the pressure on the diaphragm is twice as great as it would be in free air, which is

assuming total reflection by the transmitter. It is further assumed that the power flow is uniform over a hemisphere whose radius is the distance from the speaker's lips to the diaphragm. For this average power flow with "normally modulated" voices, Crandall and Mackenzie [8] found the value of 12.5 microwatts. More recently Sacia [6] gave 7.4 microwatts, including pauses between speech sounds, which

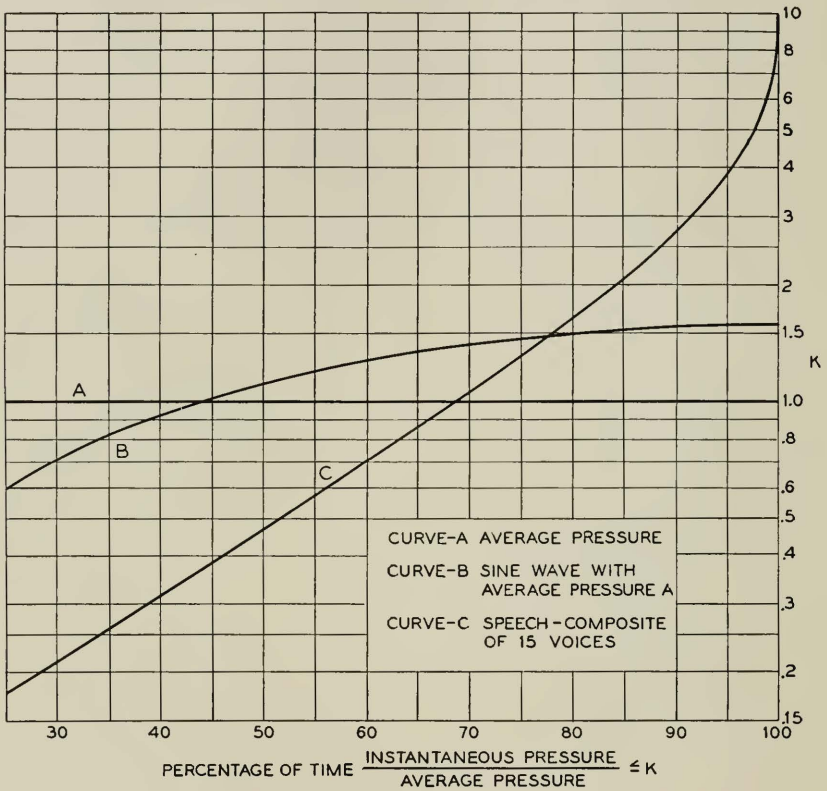


Fig. 2—Time distribution of speech pressures.

occupied about one-third of the total time. Some measurements made since then have led to substantially similar values, and so the value of 10 microwatts may be taken as fairly representative of a normal level. The corresponding average pressure on the diaphragm, at 5 cm. distance from the lips, was found to be about 5 bars. The averaging, of course, is for the absolute values of the pressures. Some notion of the level involved may be gained by noting th it is at

[8] *Phys. Rev.*, March 1922.

decidedly fatiguing to maintain a speech level 20 db higher for more than a few seconds.

It should be noted that the distributions of the instantaneous power and instantaneous pressure values in respect to their average values are quite different for speech from what they are in a sinusoidal wave. Fig. 1 and Fig. 2 show the extent of the difference for powers and pressures respectively. Clearly, speech as a whole is decidedly more "peaked" than a sinusoidal wave. Values of peak factors for

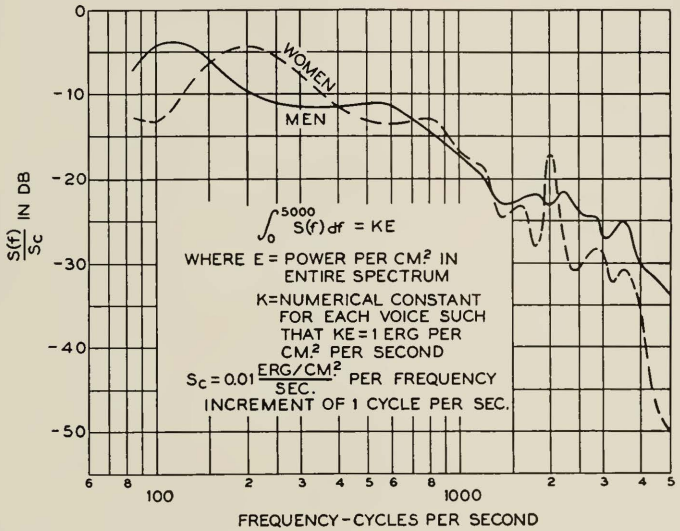


Fig. 3—Energy spectrum of connected speech-composite curves for 4 men and for 2 women.

individual speech sounds, i.e. the ratio of peak amplitude to mean effective amplitude, are given in Sacia's paper.[5]

A paramount characteristic of speech is the distribution of power and pressure in the frequency spectrum. That is so because in general the systems responding to or carrying speech (including the ear) have pronounced frequency characteristics. The same is true of noise sources interfering with speech. Here, as for total powers and pressures, the simplest quantity is the average spectrum for a large number of speech sounds. Crandall and Mackenzie [8] obtained the energy spectrum shown in Fig. 3, based on speech from six voices. They used a condenser microphone whose output was analyzed by a series of resonant circuits covering the range from 75 to 5000 p.p.s. Fig. 4 shows a recent determination in which the average pressure is given as a function of frequency. The apparatus

used to obtain the data will be described in detail elsewhere. Briefly, a series of band-pass filters are used to separate the frequency components. The output of any one filter goes into a rectilinear vacuum

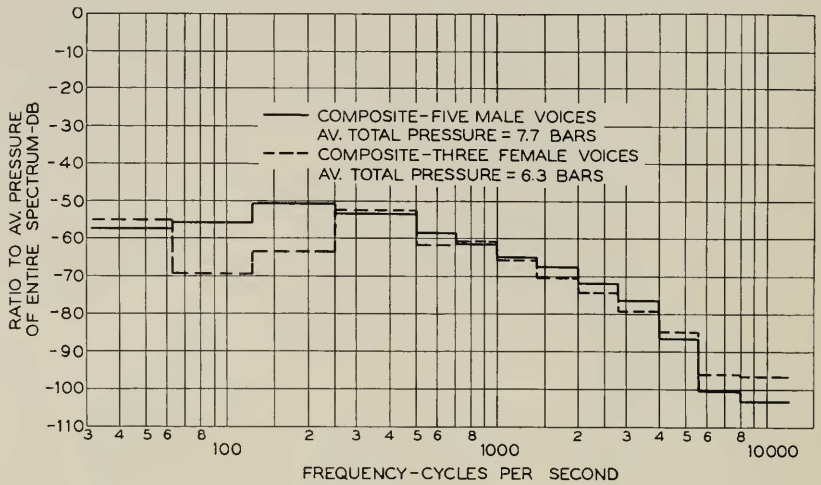


Fig. 4—Average speech pressures per frequency interval of 1 cycle per second—normal conversational voice. Distance 2″.

tube rectifier, and a fluxmeter integrates the rectified current over the duration of the speech. Simultaneously with this measurement the total spectrum also is rectified and integrated. The integration periods used were 15 seconds long. The pressure in any one band is

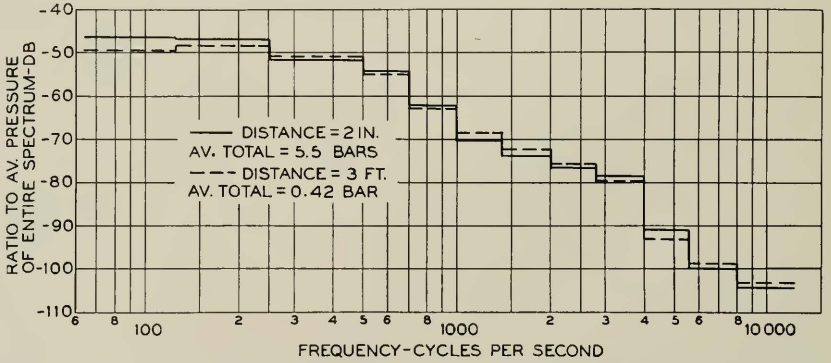


Fig. 5—Effect of distance on voice spectrum—average pressure per frequency interval of 1 cycle per second—normal speech-composite, three male voices.

given by the product: ($10^{0.05\alpha}$ × band width in cycles × total pressure), where α is the ordinate expressed in decibels (db). In using average spectra of this sort it is well to remember that they are

determined not only by the amplitudes in the individual speech sounds but equally by frequency of occurrence. Thus the fact that nearly all vowels for male voices at normal levels have a fundamental frequency between 80 and 150 p.p.s. tends to accentuate that region even though the fundamental amplitudes in individual speech sounds not be outstandingly large.

The data in Fig. 4 are for close talking conditions, 5 cm. from the lips to the diaphragm. Fig. 5 shows the effect of distance on the spectral distribution. The two distances are 5 cm. and 90 cm. respectively, in both cases the transmitter being set into a large fiber-board wall. The shapes of the two spectra are almost identical

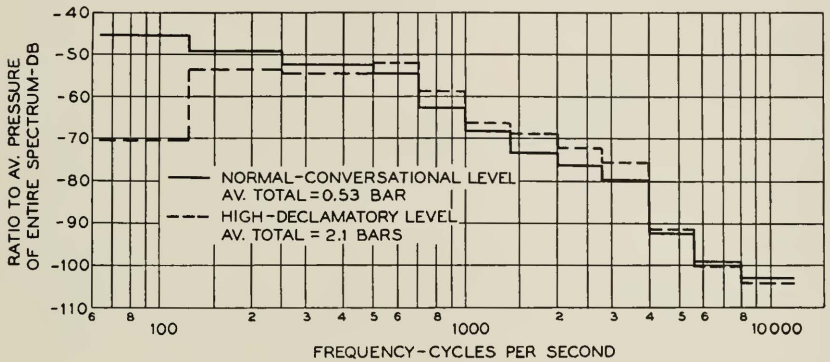


Fig. 6—Effect of level on voice spectrum—average pressures per frequency interval of 1 cycle per second—composite, three male voices, distance 3 ft.

indicating that, on the average, even at 5 cm. the condenser microphone does not greatly affect the voice as a sound generator. The largest difference between the two is in the lowest band, from 62 to 125 p.p.s., perhaps owing to the relatively low radiation efficiency of the voice at those frequencies.

The ratio of the distances is 18 : 1, that of the average pressures 14 : 1. The average pressure is nearly inversely proportional to the distance, part of the difference probably being chargeable to the more nearly total reflection for the distant condition. It is implied, of course, that even for the latter condition the direct sound is large compared with that reaching the microphone by reflections.

So far the spectra discussed were those of normally modulated voices. Fig. 6 shows what happens to the average spectrum when a high, rather declamatory level is adopted. The higher level is relatively poorer in frequencies below 500 p.p.s., relatively richer between 500 and 4000 p.p.s., and above 4000 p.p.s. its spectrum is nearly the

same as for normal levels. The decrease at low frequencies is most pronounced in the band from 62 to 125 p.p.s., i.e. in the region of the voice fundamental frequency at normal levels. This leaves two possibilities. Either at the high speech level the fundamental frequency is the same as for normal, but its amplitude is relatively small; and the spacing of the overtones is the same as at normal levels. Or else—and more probably as indicated by the auditory pitch sense —passage to the high level is attended by an actual upward shifting of the fundamental frequency, and a correspondingly larger spacing

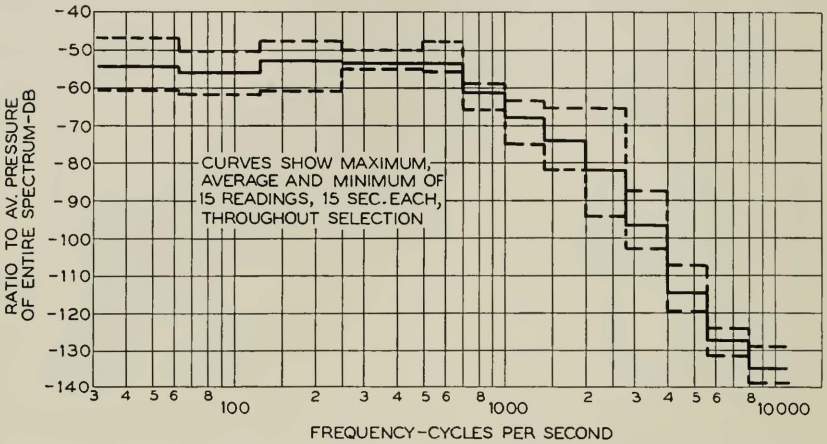


Fig. 7—Average pressures per frequency interval of 1 cycle per second—piano selection—Liszt's "Hungarian Rhapsody No. 2"—average total pressure = 3.5 bars.

of the overtones. In addition the amplitude of this new fundamental would be relatively lower than for the normal level. The average spectrum method is incapable of deciding between the two alternatives. Some oscillograms were made at normal and high speech levels, which clearly indicated that the second alternative is the correct one, or at least the prevalent one. For most vowel sounds, the fundamental frequency was found shifted from about 100 p.p.s. at normal to about 200 p.p.s. at the higher level. This result has an intimate bearing on the loss of naturalness encountered when speech originally picked up at normal voice levels is subsequently reproduced at much higher levels. A corresponding change, though in the opposite direction, probably takes place in going from normal to subnormal levels. Some evidence of this will be seen in the section on peak amplitudes.

The method of average pressure-frequency spectra is equally

applicable to sounds other than speech, provided they are sustained or can be repeated. Two illustrations are given. Fig. 7 is for a piano composition. Characteristic in comparison with speech is the relative smallness of high frequencies. Spectra of radically different types of compositions were found to be rather similar. The instrument, the room acoustics and the position of the microphone largely determine the picture. In this case, the microphone was about 7 ft. from the keyboard, and the reverberation time between 1.2 seconds at 100 p.p.s. and 1.0 second at 4000 p.p.s.

Fig. 8 gives the spectrum of street noise entering a laboratory

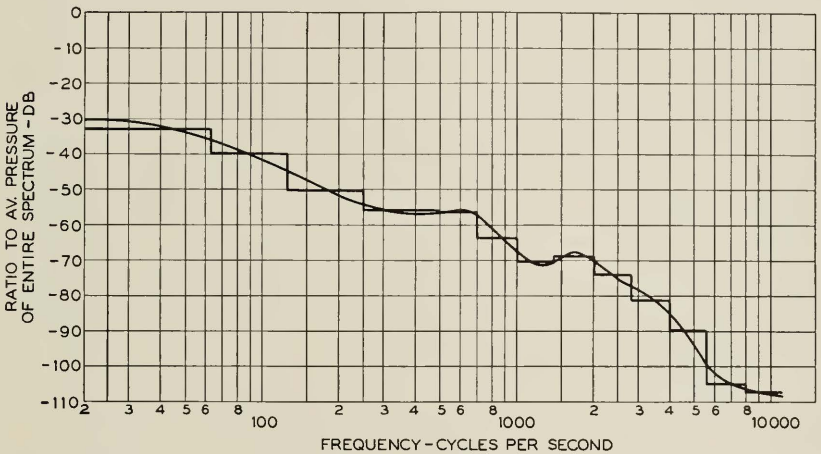


Fig. 8—Average pressures per frequency interval of 1 cycle per second—street noise—10th floor Bell Telephone Laboratories facing east—2.50–3.20 P.M., Oct. 10, 1928—average total pressure = 0.75 bar.

window, ten floors above the street. Street traffic and an elevated railway are the chief contributors, the measurements being taken during traffic peak hours. The relative poverty in high frequencies is even more pronounced than for the piano.

We shall now consider the type of apparatus intended primarily for measurements when the speakers are not acoustically controllable. Instead of averaging over a large number of words the measurement is essentially that of mean power in syllables. Usually it covers the whole spectrum rather than a particular frequency band. It has been widely used for controlling amplification in radio broadcasting, in phonograph and film recording of speech, and for rapid measurement and electrical control of speech levels in telephone conversations. A typical device of this sort is the "volume indicator" [9] shown in

[9] Essential features of this apparatus are shown by E. L. Nelson, U. S. Patent No. 1523827, filed 8/31/22.

Fig. 9, developed about ten years ago. Essentially it is a vacuum tube rectifier with a rapid action d.c. meter in the plate circuit. It is operated on a part of the characteristic such that the rectified plate current is roughly proportional to the square of the speech voltage. The rectifier is preceded by an amplifier of adjustable gain. For the speech level under measurement the gain is adjusted to such a value that the fluctuating meter deflections attain a prescribed

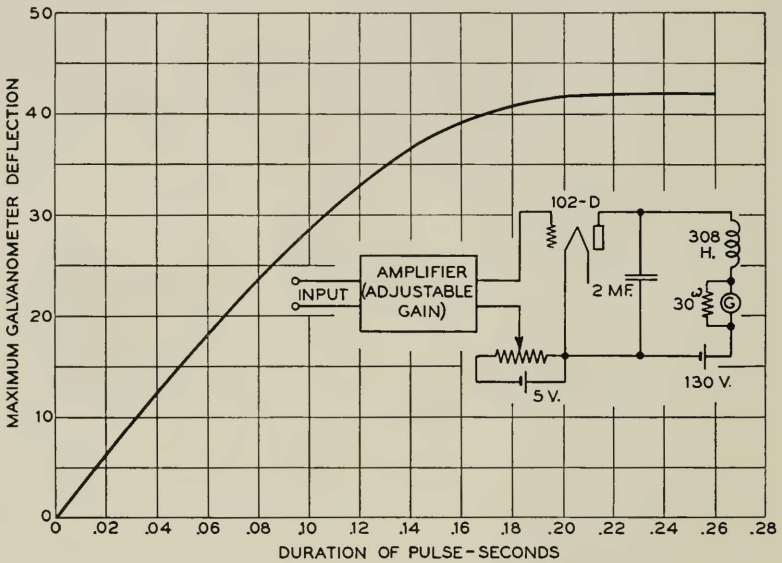


Fig. 9—Volume indicator—deflection as a function of a.c. pulse duration.

maximum value on the average of once in about three seconds. That value of gain, expressed in decibels with respect to a certain normal value of the gain, gives the volume indicator measure of the speech level. The meter, combined with the electric circuit, has a dynamic characteristic as shown in the curve, which gives the maximum deflection as a function of the duration of the a.-c. input. For inputs lasting more than about 0.18 second the maximum deflection remains the same. Since the average syllable duration is of the order of 0.2 second, it follows that the maximum deflection of the "volume indicator" meter is approximately proportioned to the mean power in the syllable. By comparison with oscillograms, or by equivalent methods, the volume indicator readings can be correlated by absolute quantities. Fig. 10 is an example showing the relation between speech level as measured with the volume indicator and the average

instantaneous amplitude of the speech waves on a certain laboratory telephone circuit of the commercial type.

Fig. 11 illustrates the type of data which can be expeditiously secured by means of the volume indicator. Here the levels for a

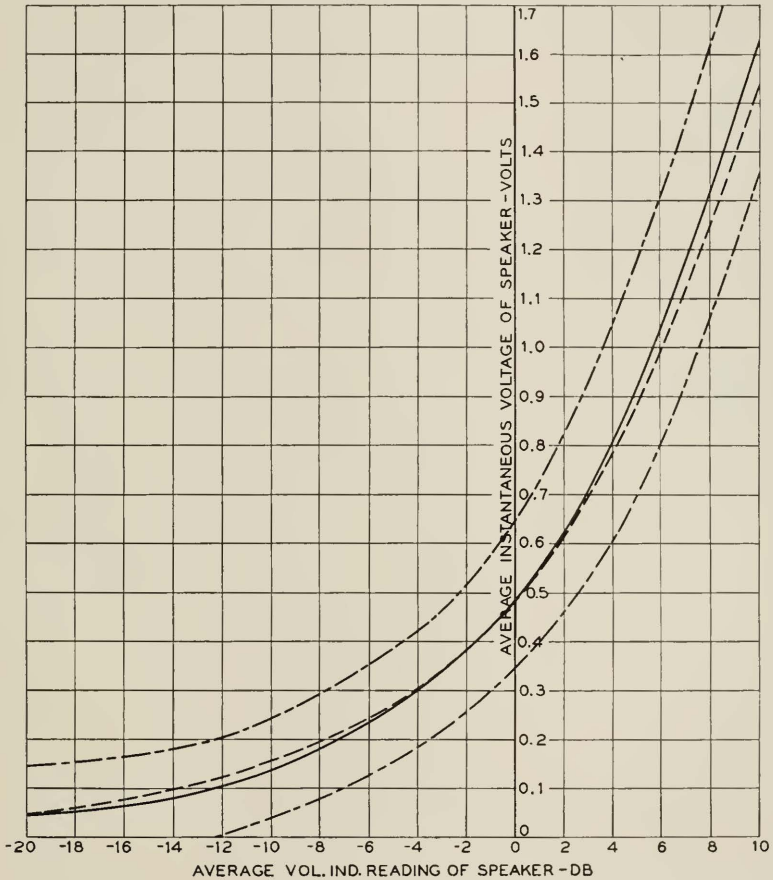


Fig. 10—Instantaneous voltage V.S. Volume indicator reading 85% of observed points included between outer dashed curves—ordinates of middle dashed curve are inversely proportional to gain of volume indicator.

large number of speakers and conversations (over a laboratory telephone circuit of the commercial type) are determined. To cover the same ground by means of oscillographic measurements would have been a decidedly formidable task.

Another device for the rapid measurement of speech levels is the "impulse meter," [10] shown in Fig. 12. This is essentially a peak

[10] D. Thierbach, *Zs. F. Techn. Physik*, No. 11, 1928.

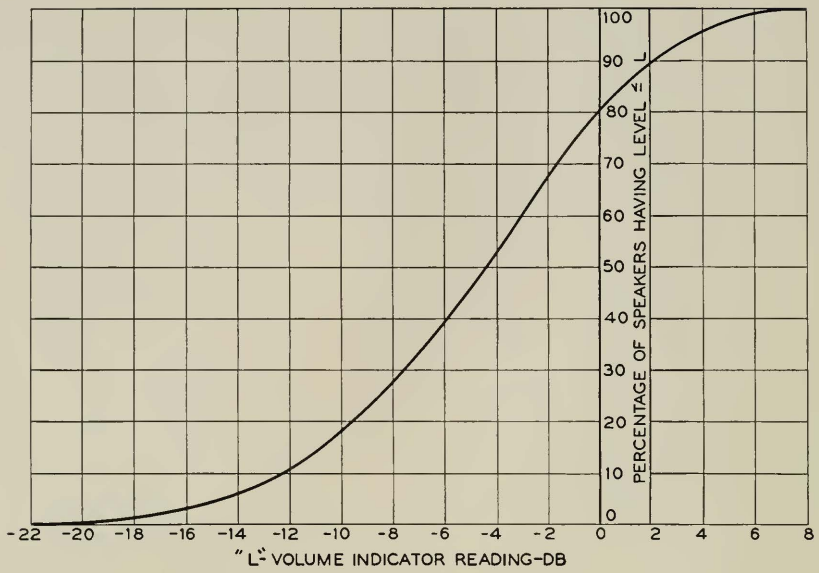


Fig. 11—Distribution of speaker's levels—87 men and 59 women.

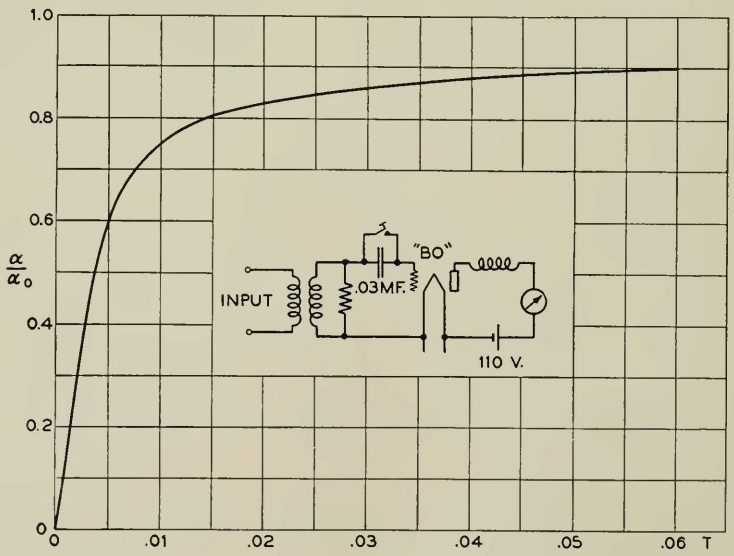


Fig. 12—Impulse meter.

T = Duration of a.c. impulse
α = Corresponding ultimate plate current
α_0 = Ultimate plate current for steadily applied a.c.

voltmeter. The curve shows the time rate at which the potential on the blocking condenser builds up. The time required for the galvanometer to reach its maximum deflection is determined by the dynamic characteristic of the meter and associated plate circuit, as well as by the time constant of the condenser charging circuit.

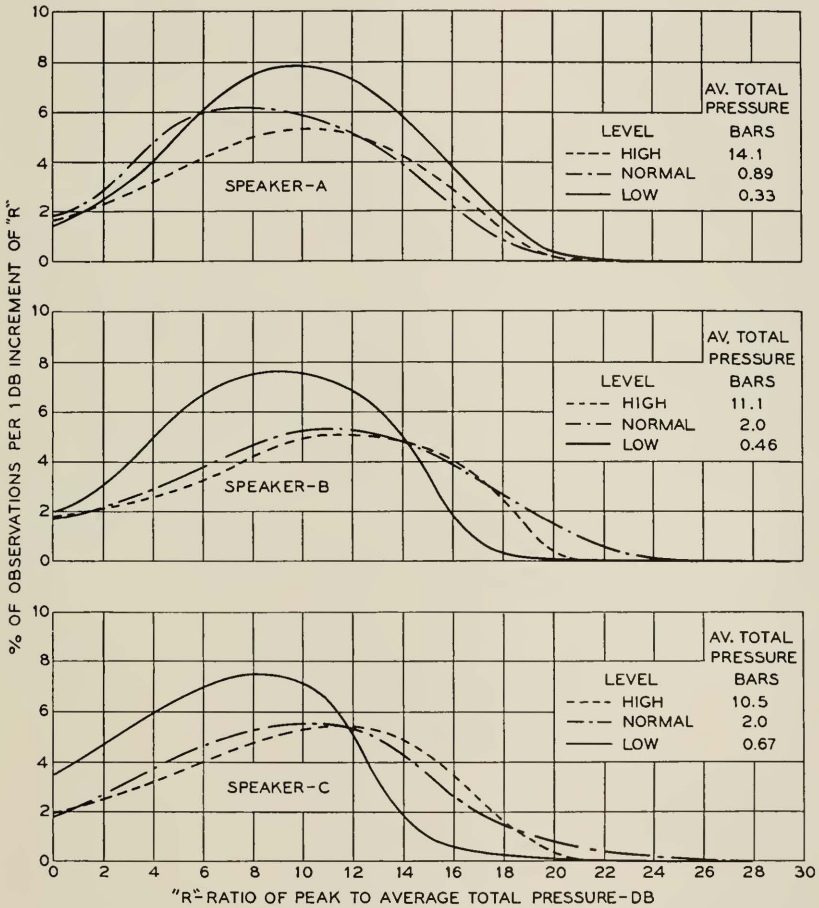


Fig. 13—Distribution of peak pressures for entire speech spectrum—three male voices—each peak is the maximum instantaneous pressure during 1/8 sec.

In using the volume indicator type of instrument in the telephone plant, limits are observed which have been determined by trial to give satisfactory operating results. The fact that it is possible to set such limits indicates a correlation between mean syllabic power and peak pressures. The overload capacity of apparatus is so important that

laboratory investigations of quantities affecting it merit more fundamental study than can be made with instruments of the volume indicator type. The quantities of interest are the instantaneous amplitudes of the wave peaks, the frequency of their occurrence and their distribution in the frequency spectrum. The apparatus used to obtain the data described below will be described in detail elsewhere. It is sufficient to state here that it is capable of directly (and automatically) registering the magnitudes and frequency of occurrence of the in-

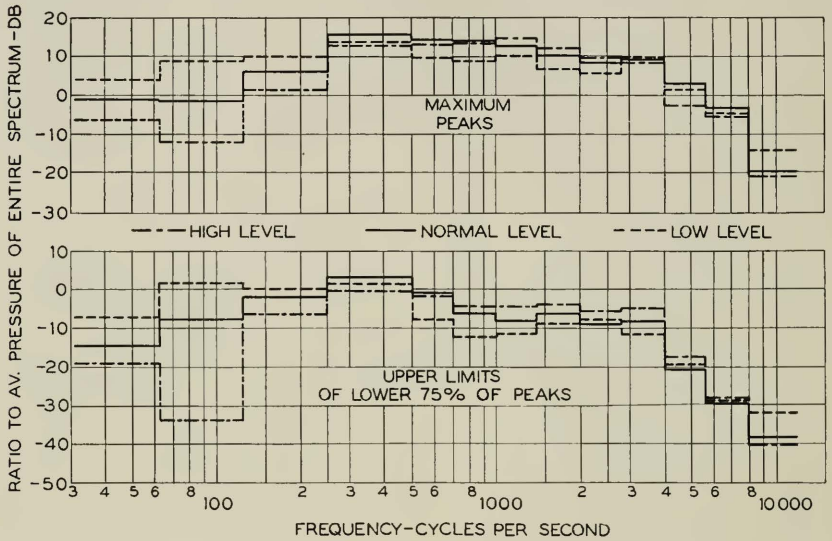


Fig. 14—Peak pressures of speech—composite, three male voices—each peak is the maximum instantaneous pressure during 1/8 sec.—average total pressure:
High (declamatory) level—8.7 bars;
Normal (conversational) level—1.6 bars;
Low (confidential) level—0.5 bar.

stantaneous peaks over a range of 60 db, corresponding to a power ratio of 1,000,000 : 1. The peak amplitudes were measured for the whole spectrum and also for restricted frequency bands selected by filters.

Fig. 13 and Fig. 14 show the results of some measurements with undistorted speech. Each individual observation gives the magnitude of the peak pressure in a 1/8 second interval. This is about as short an interval as one can use and still retain a high degree of probability that the individual measurements will give the maximum peak amplitudes in syllables. Otherwise, from the standpoint of the apparatus, the individual observations could be confined to much shorter time intervals, resulting in many more measurements for a

given length of speech. The ordinates give the ratios of the peak pressures to the average total speech pressure. In Fig. 13 the peaks are those of speech as a whole; in Fig. 14 the ordinates give the peak pressures in the several frequency bands. This is done for three widely different average levels spread over a range of 30 db. From

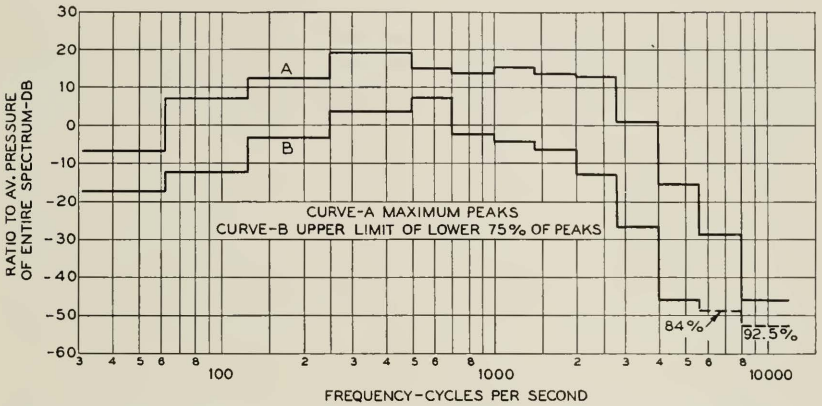


Fig. 15—Peak pressures in piano music—Liszt's "Hungarian Rhapsody No. 2" —each peak is the maximum instantaneous pressure during 1/8 sec.—average total pressure 4.0 bars.

the standpoint of providing apparatus overload capacity it is sometimes important to know not only the maximum values of all peaks (which might be uneconomical to provide for) but the upper limit of a certain percentage of all the peaks. The lower half of Fig. 14 illustrates this method of analyzing the peak data. It is interesting to note how much larger the transmission capacity must be to take care of the highest 25 per cent of the peaks.

Fig. 15 gives a similar picture of the peak amplitude distribution in a piano composition, for which the average amplitudes are given in Fig. 7. The microphone was about 7 ft. laterally from the center of the keyboard, and the measurements were made in a room having an average reverberation of about 1 second.

My thanks are due to Dr. H. K. Dunn and Mr. S. D. White of Bell Telephone Laboratories, who have obtained most of the data discussed in this paper.

Asymptotic Dipole Radiation Formulas

By W. HOWARD WISE

THE analysis of the radiation from dipoles as given by Sommerfeld and by von Hoerschelmann is deficient in one respect: it does not give the true [1] asymptotic expressions for the radiation leaving at a considerable angle from the horizontal. The correct asymptotic formulas have already been easily supplied by an appeal to the Reciprocal Theorem; lately M. J. O. Strutt [2] has got them directly from the boundary conditions and H. Weyl [3] has derived the correct asymptotic formula for a vertical dipole at the surface of the earth by a method quite different from Sommerfeld's. In the present paper it is shown how they can be got by merely improving the rigor of Sommerfeld's analysis.

The present analysis begins with the formulas of von Hoerschelmann for the wave potentials of vertical and horizontal dipoles at a finite distance above the surface of the earth and generally follows Sommerfeld. The derivation of an asymptotic approximation for the wave potential of a vertical dipole is considerably different from Sommerfeld's and results in the simpler and more precise formulas deduced from the reciprocal theorem.

Most of the analysis is somewhat simplified by taking the permeability of the earth to be unity.

The notation used is chiefly that of Bateman.[4]

τ = variable of integration, throughout the paper.

$$k^2 = \epsilon\mu\omega^2 + 4\pi\sigma\mu i\omega$$

$$l = \sqrt{\tau^2 - k_1^2}, \qquad m = \sqrt{\tau^2 - k_2^2}.$$

The subscripts 1 and 2 refer to air and ground respectively.

R_1, R_2, a, ρ, φ, w, x, y and z are adequately defined by Fig. 1.

$$\cos\theta_x = x/R, \qquad \cos\theta_y = y/R, \qquad \cos\theta_z = z/R, \qquad R^2 = x^2 + y^2 + z^2.$$

The wave potential of a horizontal dipole is [5, 4]

[1] See paragraph following equation (8).

[2] M. J. O. Strutt, *Ann. d. Phys.*, Bd. 1, p. 721, 1929.

[3] H. Weyl, *Ann. d. Phys.*, Bd. 60, p. 481, 1919.

[4] "Electrical and Optical Wave Motion," pp. 73–75.

[5] H. v. Hoerschelmann, *Jahrb. d. draht. Teleg.*, Bd. 5, 1912, pp. 14–188.

$$\Pi^h = i \times \left(- \frac{e^{ik_1R_1}}{R_1} + \frac{e^{ik_1R_2}}{R_2} - \int_0^\infty \frac{2}{l+m} J_0(\tau\rho) e^{-wl} \tau d\tau \right) + j \times 0$$

$$+ k \times 2 \cos \varphi \int_0^\infty \frac{k_2^2 - k_1^2}{(l+m)(k_2^2 l + k_1^2 m)} J_1(\tau\rho) e^{-wl} \tau^2 d\tau. \quad (1)$$

$- \frac{e^{ik_1R_1}}{R_1} + \frac{e^{ik_1R_1}}{R_2} = \Pi_x^\infty$ is the wave potential of a dipole placed parallel to a perfectly conducting plane, k_2 infinite. So, writing

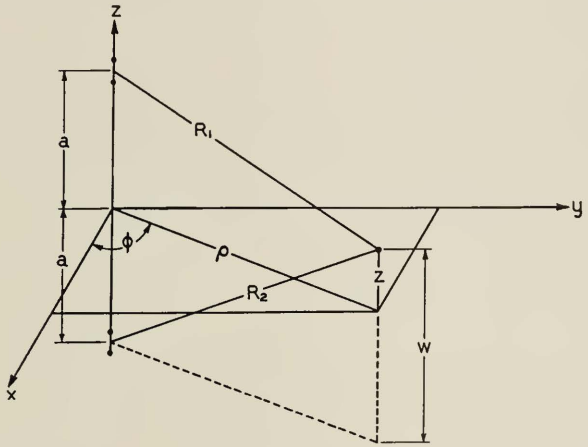


Fig. 1

$\Pi^h = \Pi_x^\infty + \Pi^\Delta$ and correspondingly $E^h = E^\infty + E^\Delta$

$$E_x^\Delta = - i\omega \left[\Pi_z^\Delta + k_1^{-2} \frac{\partial}{\partial z} \left(\frac{\partial}{\partial x} \Pi_x^\Delta + \frac{\partial}{\partial y} \Pi_y^\Delta + \frac{\partial}{\partial z} \Pi_z^\Delta \right) \right]$$

$$= - i\omega 2 \cos \varphi \frac{1}{k_1^2} \int_0^\infty J_1(\tau\rho) \frac{e^{-wl}}{l+m} \left(\frac{(k_2^2 - k_1^2)\tau^2}{k_2^2 l + k_1^2 m} - l \right) \tau^2 d\tau$$

$$= - i\omega 2 \cos \varphi \int_0^\infty J_1(\tau\rho) e^{-wl} \frac{- m\tau^2}{k_2^2 l + k_1^2 m} d\tau,$$

$$= - i\omega \frac{2}{k_1^2} \frac{\partial}{\partial x} \int_0^\infty J_0(\tau\rho) e^{-wl} \left(\frac{- k_2^2 l}{k_2^2 l + k_1^2 m} + 1 \right) \tau d\tau$$

$$= - i\omega \frac{1}{k_1^2} \frac{\partial^2}{\partial x \partial w} \left(V - 2 \frac{e^{ik_1R_2}}{R_2} \right), \quad (2)$$

where

$$V = \int_0^\infty \frac{2k_2^2}{k_2^2 l + k_1^2 m} J_0(\tau\rho) e^{-wl} \tau d\tau. \quad (3)$$

When $k_2 = \infty$, $V = 2R_2^{-1} \exp ik_1R_2$ and E_z^{Δ} is zero. When $k_2 = k_1$, $V = 1R_2^{-1} \exp ik_1R_2$ and E_z^{Δ} just cancels the field of the image dipole in E^{∞}.

The wave potential of a vertical dipole is [5, 4]

$$\Pi_z^v = \frac{e^{ik_1R_1}}{R_1} - \frac{e^{ik_1R_2}}{R_2} + V. \tag{4}$$

V is the function Π_0 analysed by Sommerfeld in Riemann-Webers Differentialgleichungen der Physik. Sommerfeld transforms the integration from 0 to ∞ into an integration from $-\infty$ to $+\infty$ by replacing the Bessel function by its equivalent Hankel functions and then wraps the real axis path of integration around the zero of $k_2^2l + k_1^2m$ and the two branch cuts from k_1 and k_2 to $+i\infty$. He thus gets $V = P + Q_1 + Q_2$ where P is the integral around the zero of $k_2^2l + k_1^2m$, Q_1 is the integral around the branch cut from k_1, Q_2 is the integral around the branch cut from k_2 and the function integrated is

$$\frac{k_2^2}{k_2^2l + k_1^2m} H_0^1(\tau\rho)e^{-wl}\tau d\tau. \tag{5}$$

l and m are pure imaginaries on the branch cuts, which are rectangular hyperbolas, from k_1 and k_2 respectively. As one carries τ counterclockwise around the branch cut from k_1 to $+i\infty$, l travels up the right hand side of the imaginary axis from $-i\infty + \epsilon$ to $+i\infty + \epsilon$. A similar statement holds for m.

Multiplying numerator and denominator of equation (5) by $k_2^2l - k_1^2m$ and then taking out a factor $(k_2^4 - k_1^4)^{-1}$ our integrand becomes

$$\frac{k_2^2}{k_2^4 - k_1^4} \frac{k_2^2\sqrt{\tau^2 - k_1^2} - k_1^2\sqrt{\tau^2 - k_2^2}}{(\tau - s)(\tau + s)} H_0^1(\tau\rho)e^{-w\sqrt{\tau^2 - k_1^2}}\tau d\tau,$$

where $s = +k_1k_2 \div \sqrt{k_2^2 + k_1^2}$. Integrating around the pole at $\tau = s$, we get

$$\begin{aligned} P &= \frac{k_2^2}{k_2^4 - k_1^4} 2\pi i \frac{k_2^2\sqrt{s^2 - k_1^2} - k_1^2\sqrt{s^2 - k_2^2}}{2s} H_0^1(s\rho)e^{-w\sqrt{s^2 - k_1^2}}s \\ &= 0 \quad \text{if} \quad k_2^2\sqrt{s^2 - k_1^2} - k_1^2\sqrt{s^2 - k_2^2} = 0 \\ &= -2\pi \frac{k_2^3k_1s}{k_2^4 - k_1^4} H_0^1(s\rho)e^{-iwsk_1/k_2}, \end{aligned} \tag{6}$$

if

$$k_2^2\sqrt{s^2 - k_1^2} - k_1^2\sqrt{s^2 - k_2^2} = 2k_2^2\sqrt{s^2 - k_1^2} = 2ik_1k_2s.$$

l and m must have their real parts positive and so in taking the square roots of $s^2 - k_1^2 = -k_1^4 \div (k_2^2 + k_1^2)$ and $s^2 - k_2^2 = -k_2^4 \div (k_2^2 + k_1^2)$ one halves the smallest angle with the positive real axis. If they both lie on the same side of the real axis P is zero. In order that they may lie on opposite sides of the real axis it is necessary that

$$\arg k_1^4 < \arg (k_1^2 + k_2^2) < \arg k_2^4.$$

Writing $k_1^2 = \alpha + i\beta$ and $k_2^2 = x + iy$ this means that

$$y > \frac{2\alpha\beta}{\alpha^2 - \beta^2} x + \frac{\alpha^2 + \beta^2}{\alpha^2 - \beta^2} \beta.$$

The goal of the paper being asymptotic formulas for the sky waves of vertical and horizontal dipoles the ground wave, P, will hereafter be ignored. This is possible because at the high frequencies for which dipoles are useful the ground wave is very highly damped.

Sommerfeld gets an asymptotic expression for Q_1 by noting that if we are at a great distance from the source most of the value of the integral comes from that portion of the path of integration very close to k_1. The solution he arrives at is

$$-2(\Omega + \Omega^2 + \Omega^3 + + \cdots) \frac{e^{ik_1R}}{R} \quad \text{where} \quad \Omega = \frac{k_2^2}{k_1^2\sqrt{k_1^2 - k_2^2}} \frac{\partial}{\partial z}.^{6} \qquad (7)$$

Neglecting higher powers of $1/R$ than the first, equation (7) sums up into

$$Q_1 \sim \frac{2k_2^2 \cos \theta_z}{k_2^2 \cos \theta_z + k_1\sqrt{k_2^2 - k_1^2}} \frac{e^{ik_1R}}{R}. \qquad (8)$$

But in getting equation (7) Sommerfeld has replaced $\sqrt{\tau^2 - k_2^2}$ by $\sqrt{k_1^2 - k_2^2}$. This is a needless approximation which ruins the symmetry, damages the utility and tends to hide the physical meaning of the final result. To get the true asymptotic formula for Q_1 it is necessary to confine the approximations to the purely operational variety, i.e. make no approximations of substitution before integrating but let the approximation reside wholly in the manner of integrating, as follows

$$Q_1 = \int \frac{k_2^2}{k_2^2 l + k_1^2 m} H_0^1(\tau\rho) e^{-wl} \tau d\tau$$

$$= -\frac{\partial}{\partial w} \int H_0^1(\tau\rho) \frac{e^{-wl}}{l} (C_0 + C_1 l + C_2 l^2 + \cdots) \tau d\tau$$

[6] *Annalen der Physik*, Band 28, 1909, page 705.

where

$$C_0 + C_1 l + C_2 l^2 + \cdots = k_2^2 \div (k_2^2 l + k_1^2 \sqrt{l^2 + k_1^2 - k_2^2}).$$

To the extent that most of the value of the integral comes from that portion of the path of integration very close to $\tau^2 = k_1^2$ the expansion in powers of l is valid. Replacing each l by $-(\partial/\partial w)$ we have then

$$Q_1 \sim -2\frac{\partial}{\partial w}\left(C_0 - C_1\frac{\partial}{\partial w} + C_2\frac{\partial^2}{\partial w^2} - + \cdots\right)\frac{e^{ik_1R_2}}{R_2}. \tag{9}$$

Now

$$-\frac{\partial}{\partial w}\frac{e^{ik_1R_2}}{R_2} = (-ik_1\cos\theta_z)\frac{e^{ik_1R_2}}{R_2} + \frac{\cos\theta_z}{R_2}\frac{e^{ik_1R_2}}{R_2}$$

$$= \left[\gamma + \frac{\cos\theta_z}{R_2}\right]\frac{e^{ik_1R_2}}{R_2} \quad \text{where} \quad \gamma = -ik_1\cos\theta_z,$$

$$\frac{\partial^2}{\partial w^2}\frac{e^{ik_1R_2}}{R_2} = \left(\gamma^2 + 2\gamma^1\frac{\cos\theta_z}{R_2} + 1\cdot 2\gamma^0\frac{ik_1\sin^2\theta_z}{2R_2} + \cdots\right)\frac{e^{ik_1R_2}}{R_2},$$

$$-\frac{\partial^3}{\partial w^3}\frac{e^{ik_1R_2}}{R_2} = \left(\gamma^3 + 3\gamma^2\frac{\cos\theta_z}{R_2} + 2\cdot 3\gamma^1\frac{ik_1\sin^2\theta_z}{2R_2} + \cdots\right)\frac{e^{ik_1R_2}}{R_2},$$

. . . . . .

etc., and so

$$Q_1 \sim 2(\gamma^1 C_0 + \gamma^2 C_1 + \gamma^3 C_2 + + \cdots)\frac{e^{ik_1R_2}}{R_2}$$

$$+ 2\cos\theta_z(1\gamma^0 C_0 + 2\gamma^1 C_1 + 3\gamma^2 C_2 + + \cdots)\frac{e^{ik_2R_2}}{R_2^2}$$

$$+ ik_1\sin^2\theta_z(1\cdot 2\gamma^0 C_1 + 2\cdot 3\gamma^1 C_2 + 3\cdot 4\gamma^2 C_3 + + \cdots)\frac{e^{ik_1R_2}}{R_2^2}$$

$$+ \text{higher order terms}$$

$$= \frac{2k_2^2\gamma}{k_2^2\gamma + k_1^2\sqrt{\gamma^2 + k_1^2 - k_2^2}}\frac{e^{ik_1R_2}}{R_2}$$

$$+ \left(\cos\theta_z\frac{\partial}{\partial\gamma} + \frac{ik_1\sin^2\theta_z}{2}\frac{\partial^2}{\partial\gamma^2}\right)\frac{2k_2^2\gamma}{k_2^2\gamma + k_1^2\sqrt{\gamma^2 + k_1^2 - k_2^2}}\frac{e^{ik_1R_2}}{R_2^2}$$

$$+ \text{higher order terms.} \tag{10}$$

The second order term can be neglected if $k_1R_2 >> 1$, say if $R > 20\lambda$. It will hereafter be supposed that this is the case.

Multiplying numerator and denominator of the first order term by i and canceling a k_1 we have finally

$$Q_1 \sim \frac{2k_2^2 \cos\theta_z}{k_2^2 \cos\theta_z + k_1\sqrt{k_2^2 - k_1^2 \sin^2\theta_z}} \frac{e^{ik_1R_2}}{R_2}. \tag{11}$$

This Q_1 behaves as one would rightfully expect a true asymptotic formula to behave. It is $2(\exp ik_1R_2)/R_2$ at $k_2 = \infty$ and $1(\exp ik_1R_2)/R_2$ at $k_2 = k_1$.

In so far as k_2 is considerably larger than k_1 and the expansion in powers of l is valid Q_2 is negligible in comparison with Q_1.[7] Perhaps the easiest way of seeing this is to note that equation (10) might just as well have been obtained directly from V instead of from Q_1.

Substituting equation (11) in equation (4) we get

$$\Pi_z^v \sim \left(1 + \frac{k_2^2 \cos\theta_z - k_1\sqrt{k_2^2 - k_1^2 \sin^2\theta_z}}{k_2^2 \cos\theta_z + k_1\sqrt{k_2^2 - k_1^2 \sin^2\theta_z}} e^{ik_1(R_2-R_1)}\right) \frac{e^{ik_1R_1}}{R_1}, \tag{12}$$

whence

$$E_z^v = -i\omega\left(\Pi_z^v + k_1^{-2}\frac{\partial^2}{\partial z^2}\Pi_z^v\right)$$

$$\sim -i\omega \sin^2\theta_z(1 + R_1e^{ik_12a\cos\theta_z})\frac{e^{ik_1R}}{R}, \tag{13}$$

where

$$R_1 = \frac{k_2^2 \cos\theta_z - k_1\sqrt{k_2^2 - k_1^2 \sin^2\theta_z}}{k_2^2 \cos\theta_z + k_1\sqrt{k_2^2 - k_1^2 \sin^2\theta_z}}.$$

Substituting equation (11) in equation (2) and adding E_z^∞ we get

$$E_z^h \sim -i\omega \cos\theta_x \cos\theta_z(1 - R_1e^{ik_12a\cos\theta_z})\frac{e^{ik_1R}}{R}. \tag{14}$$

R_1 is the coefficient of reflection for a plane wave polarized in the plane of incidence.

The horizontal fields of a horizontal dipole are

$$E_y^h = -i\omega\left[\Pi_y^h + k_1^{-2}\frac{\partial}{\partial y}\left(\frac{\partial}{\partial x}\Pi_x^h + \frac{\partial}{\partial y}\Pi_y^h + \frac{\partial}{\partial z}\Pi_z^h\right)\right]$$

$$= -i\omega k_1^{-2}\left[\frac{\partial^2}{\partial y\partial x}\Pi_x^\infty + \frac{\partial}{\partial y}\left(\frac{\partial}{\partial x}\Pi_x^\Delta + \frac{\partial}{\partial z}\Pi_z^\Delta\right)\right] \tag{15}$$

and

$$E_x^h = -i\omega\left[\Pi_x^h + k_1^{-2}\frac{\partial}{\partial x}\left(\frac{\partial}{\partial x}\Pi_x^h + \frac{\partial}{\partial y}\Pi_y^h + \frac{\partial}{\partial z}\Pi_z^h\right)\right]$$

$$= -i\omega\left[\Pi_x^\infty + k_1^{-2}\frac{\partial^2}{\partial x^2}\Pi_x^\infty\right]$$

[7] Riemann-Weber's "Differentialgleichungen der Physik," p. 556.

$$- i\omega \left[\Pi_x^{\Delta} + k_1^{-2} \frac{\partial}{\partial x} \left(\frac{\partial}{\partial x} \Pi_x^{\Delta} + \frac{\partial}{\partial z} \Pi_z^{\Delta} \right) \right]$$

$$\sim - i\omega(\Pi_x^{\infty} + \Pi_x^{\Delta}) + \frac{\cos \theta_x}{\cos \theta_y} E_y^{h}$$

$$= - i\omega(\Pi_x^{\infty} + \Pi_x^{\Delta}) + \cot \varphi E_y^{h}; \tag{16}$$

$$\therefore \quad E_{\varphi}^{h} = E_x^{h} \sin \varphi - E_y^{h} \cos \varphi \sim - i\omega \sin \varphi(\Pi_x^{\infty} + \Pi_x^{\Delta}). \tag{17}$$

Evidently the procedure which yielded the true asymptotic expres-

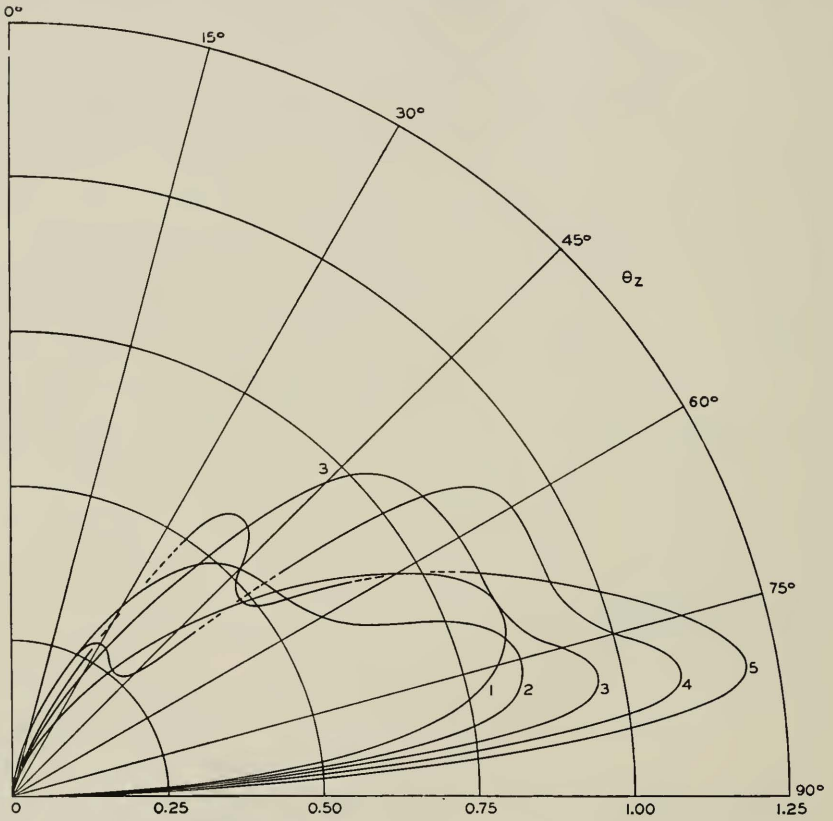


Fig. 2—Vertical dipole polar diagrams computed for $\lambda = 6$ meters, $\epsilon = 9$, $\mu = 1$, and $\sigma = 10^{-13}$

sion for V will do the same for Π_x^{Δ}. The details are not interesting. The result is

$$\Pi_x^{\Delta} \sim - \frac{2k_1 \cos \theta_z}{k_1 \cos \theta_z + \sqrt{k_2^2 - k_1^2 \sin^2 \theta_z}} \frac{e^{ik_1 R_2}}{R_2}. \tag{18}$$

Substituting equation (18) in equation (17) we get

$$E_{\varphi}{}^{h} \sim + i\omega \sin \varphi(1 - R_2 e^{ik_1 2a \cos \theta_z}) \frac{e^{ik_1 R}}{R}, \tag{19}$$

where

$$R_2 = \frac{\sqrt{k_2{}^2 - k_1{}^2 \sin^2 \theta_z} - k_1 \cos \theta_z}{\sqrt{k_2{}^2 - k_1{}^2 \sin^2 \theta_z} + k_1 \cos \theta_z}.$$

R_2 is the coefficient of reflection for a plane wave polarized perpendicular to the plane of incidence.

Formulas (13), (14) and (19) are just what one would get by applying

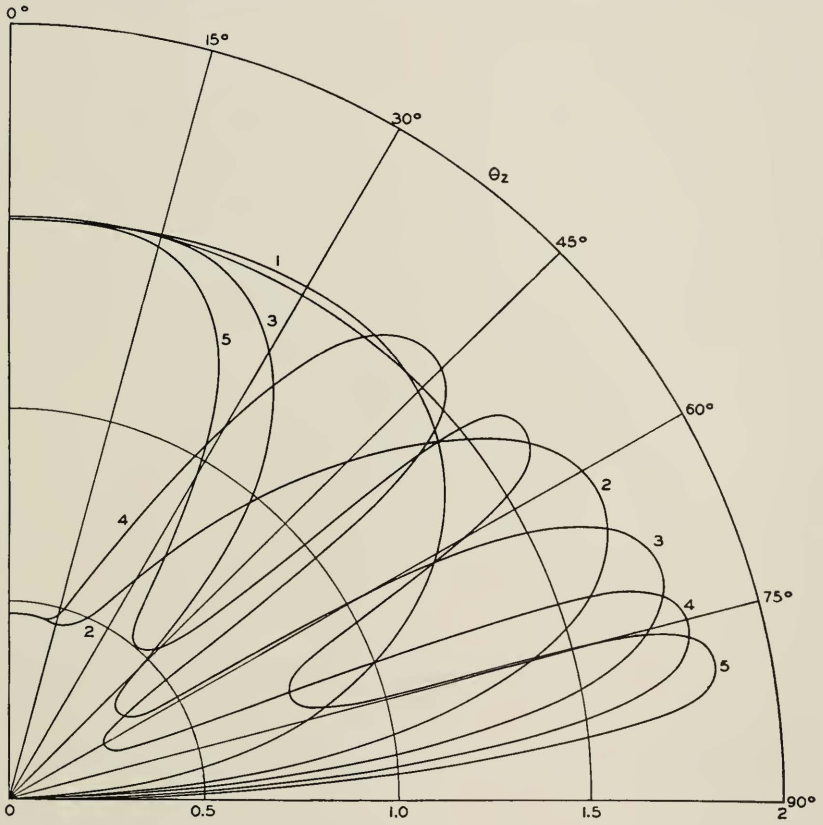


Fig. 3—Polar diagrams of the horizontal fields of horizontal dipoles computed for $\lambda = 6$ meters, $\epsilon = 9$, $\mu = 1$ and $\sigma = 10^{-13}$

the coefficient of reflection from an imperfect reflector to the reflected waves of the corresponding electrostatic formulas.

Returning to equation (15)

$$E_y{}^h = -i\omega k_1{}^{-2}\frac{\partial^2}{\partial y\partial x}\left[\Pi_x{}^\infty - \int_0^\infty \frac{2}{m+l}\left(1-\frac{(k_2{}^2-k_1{}^2)l}{k_2{}^2 l+k_1{}^2 m}\right)J_0(\tau\rho)e^{-wl}\tau d\tau\right]$$

$$= -i\omega\frac{\partial^2}{\partial y\partial x}(k_1{}^{-2}\Pi_x{}^\infty - k_2{}^{-2}V)$$

$$\sim -i\omega\cos\theta_x\cos\theta_y\left(-1+\left[1-\frac{k_1{}^2}{k_2{}^2}(1+R_1)\right]e^{ik_1 2a\cos\theta_z}\right)\frac{e^{ik_1 l}}{R}\cdot \quad (20)$$

Usually one cares only for $E_\varphi{}^h$ at $\varphi = \pi/2$ and $E_x{}^h$ and $E_y{}^h$ are of no particular interest. Figs. 2, 3 and 4 are polar diagrams in the

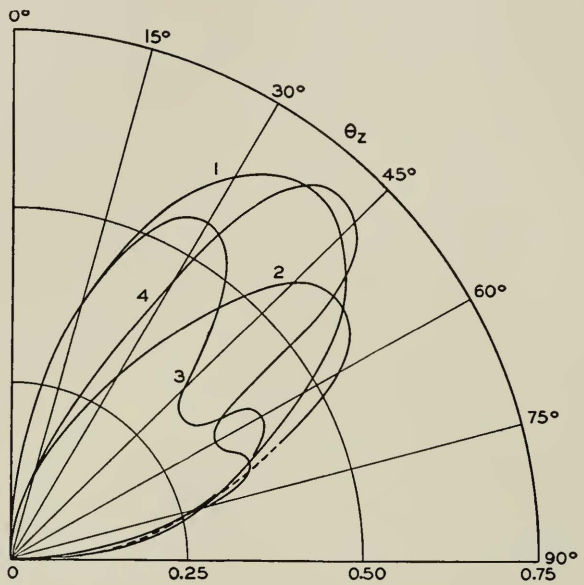


Fig. 4—Polar diagrams of the vertical fields of horizontal dipoles computed for $\lambda = 6$ meters, $\epsilon = 9$, $\mu = 1$ and $\sigma = 10^{-13}$

vertical plane of equations (13), (14) and (19). Assuming the conductivity of the air to be zero and the dielectric constant to be unity

$$k_2{}^2 = k_1{}^2(\epsilon + i2c\lambda\sigma), \qquad k_1 = 2\pi/\lambda,$$

where ϵ is the dielectric constant of the ground referred to air as unity and σ is the conductivity of the ground in electromagnetic units. The values of ϵ and σ used in computing the polar diagrams are generally supposed to be somewhere in the neighborhood of their

average values for earth but they vary so much with the locality that the diagrams can scarcely be regarded as giving more than a general idea as to what may be expected of the formulas.

The number attached to each curve is the height of its dipole in quarter wave-lengths.

Formulas (13), (14) and (19) are just what one would get by applying the Reciprocal Theorem to two dipoles, one near the earth and the other far away. The electric field acting on the one near the earth is composed of a direct field and a reflected field which is R_1 or R_2, as the case may be, times the direct field.

When the depth to groundwater, bedrock, an orebody or any marked discontinuity in the electrical properties of the ground is known and is not too great the effect of this discontinuity on the polar diagram ought not to be ignored. The asymptotic formulas for any stratified ground are got by putting the coefficients of reflection for a plane wave reflected from the surface of that ground in place of the corresponding coefficients in equations (13), (14) and (19). For a number of rather obvious reasons it would usually be out of the question to deal with more than one plane of discontinuity; one is bad enough. The coefficients for a single plane of discontinuity at a depth Δ are

$$R_1^1 = \frac{k_2^2 \cos\theta_z - \eta_1 k_1 \sqrt{k_2^2 - k_1^2 \sin^2\theta_z}}{k_2^2 \cos\theta_z + \eta_1 k_1 \sqrt{k_2^2 - k_1^2 \sin^2\theta_z}}$$

and

$$R_2^1 = \frac{\sqrt{k_2^2 - k_1^2 \sin^2\theta_z} - \eta_2 k_1 \cos\theta_z}{\sqrt{k_2^2 - k_1^2 \sin^2\theta_z} - \eta_2 k_1 \cos\theta_z},$$

where

$$\eta_1 = \frac{\mu_2}{\mu_1}\frac{1+\delta_1}{1-\delta_1}, \qquad \eta_2 = \frac{\mu_2}{\mu_1}\frac{1+\delta_2}{1-\delta_2},$$

$$\delta_1 = \frac{k_2^2\mu_3\sqrt{k_3^2 - k_1^2\sin^2\theta_z} - k_3^2\mu_2\sqrt{k_2^2 - k_1^2\sin^2\theta_z}}{k_2^2\mu_3\sqrt{k_3^2 - k_1^2\sin^2\theta_z} - k_3^2\mu_2\sqrt{k_2^2 - k_1^2\sin^2\theta_z}} e^{i2\Delta\sqrt{k_2^2 - k_1^2\sin^2\theta_z}},$$

and

$$\delta_2 = \frac{\mu_3\sqrt{k_2^2 - k_1^2\sin^2\theta_z} - \mu_2\sqrt{k_3^2 - k_1^2\sin^2\theta_z}}{\mu_3\sqrt{k_2^2 - k_1^2\sin^2\theta_z} + \mu_2\sqrt{k_3^2 - k_1^2\sin^2\theta_z}} e^{i2\Delta\sqrt{k_2^2 - k_1^2\sin^2\theta_z}}.$$

If Δ is not large and k_3 is considerably different from k_2 then η_1 and η_2 will differ considerably from unity.

Statistical Theories of Matter, Radiation and Electricity [1]

By KARL K. DARROW

The atomic or "kinetic" theory of gases, with its interpretations of such qualities as temperature, pressure, viscosity and conductivity, has ranked for more than half a century as a very important part of theoretical physics. A corresponding theory for radiation and for negative electricity is much to be desired, since it is known that in many ways each of these entities behaves as though it were atomic. There are, however, differences among the three, and only within the last five years have these been formulated suitably. This article is devoted to the resulting statistical theories.

THE major subjects of this article are two extensions of what formerly was called *atomic theory*—that is to say, the attempt to explain as many as possible of the properties of pieces of matter large enough to be visible and tangible and ponderable, by visualizing these as swarms of tiny particles each endowed with only a very few and simple qualities. Among the properties of gases, for example, are pressure and viscosity and entropy and temperature. Conceivably one might invest the ultimate atoms with all four. The atomic theory of gases as it stands today, however, is the outcome of a very different procedure. It is the achievement of an effort to interpret these four properties and several more as features of a hypothetical assemblage of very many corpuscles all alike, and not possessing them nor any others except position and velocity and mass (and moment of inertia, sometimes) and the liability to make elastic impacts with each other. On the whole the effort has been remarkably successful. Therefore viscosity and temperature and entropy are not attributed to single atoms, but pictures and expressions for them are derived as qualities of the assemblage. The theory which leads to these results is called *statistical;* it is based on certain assumptions which, in the form in which they were originally made, constitute the *classical statistics*. The successes of the classical statistics are a part of the evidence that matter is corpuscular. Once they were nearly the whole of the evidence, for they antedated the striking demonstrations of individual atoms which now spring to the mind whenever one is asked to state the reasons for accepting the atomic theory.

Radiation resembles a gas in some respects. Entropy and temperature and pressure, for example, are properties displayed by radiation when enclosed in a space surrounded by a wall of even temperature, just as they are by a gas in a like situation. It seems quite natural that one should try to interpret them in the same way as for a gas they are interpreted by the atomic theory: imagining the radiation

[1] *Physical Review Supplement*, Vol. 1, July 1929, pp. 90–155.

as an assemblage of innumerable particles, a swarm of photons or corpuscles of light. Nowadays at least the idea seems quite natural; but of course, in the years when no one as yet had broken away from the tradition that light is altogether wavelike, it would doubtless have been thought a very wild one. Even after Einstein had ventured such a breach with the past, nearly a score of years elapsed before there was developed out of the theory of quanta an adequate conception of the "radiation-gas." The historical sequence in the growth of the atomic theory of matter was here inverted: there was abundant evidence for the corpuscular theory of light, in phenomena such as the Compton effect and the photoelectric effect showing the work of individual photons, before the statistical theory of these corpuscles was perfected. We now see that the trouble was, that even when one accepts the notion of corpuscles of light without reserve, and even when one knows the proper values of energy and momentum to be assigned to these corpuscles, it still is not correct to apply to them the same statistics as gives such good results when applied to the atoms of matter. Bose discovered how to remodel the statistics, in order to construct a competent atomic theory of the radiation in thermal equilibrium in an enclosure.

The other of the new extensions of atomic theory is partly a revival —the resurrection of the theory, first proposed some thirty years ago, that part at least of the negative electricity within a metal acts like a swarm of freely-flying corpuscles which collide now and again not with each other but with the atoms. It was of course the classical statistics which was always used in developing this theory. Moribund because of several incurable discordances with fact, the theory was resuscitated by Pauli and by Sommerfeld with a revision of the statistics. It was not quite the same revision as enabled Bose to set up an atomic theory of radiation, but a very similar one, invented first by Fermi and later independently by Dirac. One cannot say that the so-renovated "electron-gas theory" is a perfect explanation of all the multifarious phenomena of the flow of electricity and heat inside of metals and outward through the boundaries of metals. Its initial successes, however, are so auspicious as to suggest that the hope of further progress lies not in renouncing it (as seemed to be almost inevitable before the alterations) but in amending it in its details.

Is the atomic theory of material gases to remain untouched by these novel ideas? Apparently all three forms of statistics, the classical and the two recent types, lead to very nearly the same conclusions when applied to material gases. Only at remarkably low temperatures and remarkably high densities do their predictions diverge; and under

these conditions the experimental data are not easy to interpret for that purpose. Suppose, however, that eventually the data are proved to decide for one of the new forms of statistics against the old: what then? Probably we shall merely remove one of the theoretical foundation-stones of the kinetic theory of gases and insert another to take its place, meanwhile leaving practically intact the great superstructure of formulæ and equations whereby the kinetic theory makes contact with experience. Happily this is an easier process in theoretical physics than in architecture.

Custom has lately changed the meaning of the term *atomic theory*, making it almost synonymous with *theory of the structure of the atom;* but the province which this latter has taken for its own is one to which its forerunner disclaimed all right of entry. It was never supposed that all of the properties of a gas can be interpreted as statistical features of a swarm of corpuscles. The earlier atomic theory conceded some of them to the individual atoms, thus in effect renouncing the ambition to explain them; and among these were the spectra. Where the statistical theory left off, the builders of atom models took up the work. Bohr, for example, designed a model for the individual hydrogen atom, competent—at least to a great extent—to explain the Balmer series and the rest of the line-spectrum of "atomic hydrogen." This model he constructed, following Rutherford, out of a pair of corpuscles. What he and his successors thus developed was in a way an *atomic theory of the atom*—a degree deeper, or further, or higher perhaps, then the atomic theory of matter which had provided him with the notion and the scale of the atom to begin with.

What then distinguishes this new "atomic theory of the atom" from its ancestor? Well, the major differences in method and in aim are traceable to the fact, that in the later theory the number of elementary particles which constitute the system is quite manageably small, while in the earlier, it is inconceivably tremendous.

Bohr constructed his model for the hydrogen atom with only a pair of corpuscles, and those for all the other atoms out of not more than a few dozen each. Now with a model consisting only of two particles, one can specify positions and velocities for these with the utmost of precision, and go merrily ahead predicting and describing orbits with as much exactness as one cares to lavish. Even with dozens of electrons and a nucleus one can attain at least a specious accuracy of detail; remember the portraits of the electron-orbits of massive atoms which six or seven years ago were so profuse. Perhaps it is not wise to make such definite assertions; but it is feasible. Not so, however, with the subjects of the older theory.

The model proposed for a cubic centimetre of gas under the ordinary conditions of temperature and pressure consists of something like 10^{20} particles. Merely the mention of so extravagant a figure is sufficient to persuade that it is vain to dream of making any progress by postulating a definite position and a definite velocity for each of these. The life of the human race would not be long enough to write down even the postulates, to say nothing of the inferences.

This seems a fearful handicap; but it is not so at all. Adopting the statistical method, one does not even begin upon the hopeless task of fixing place and motion for every particle. We content ourselves with writing down a function, which states how many among the multitude of particles we assume to be situated in each small (but not too small) element of volume; and how many we assume to have momenta which lie in each small (but not too small) range of momentum. These are specifications much more modest and vague; but they are ample. For the things which we wish to interpret—entropy and temperature, viscosity and conduction and diffusion—the atomic picture need not be made one whit more definite.

In saying this I am understating the case. If the atomic picture could be made more definite, say by stating the locations and the velocities of all the atoms with absolute precision, the meanings which we shall presently attach to entropy and temperature would be dissolved. Our theory of these entities depends upon the vagueness of the picture. Seemingly they appeal to us as physical realities because our senses and our instruments are too obtuse to perceive the atoms. Our minds must feign a somewhat similar obtuseness, pretending not to fix the particles of the imagined swarm too sharply; therefore it does not matter that they are so numerous that the pretence becomes sincere. Exact knowledge of the individual atoms is unattainable; but it is useless, is not desirable even. One remembers Æsop's tale of the fox and the inaccessible grapes; in this case it is probable that the grapes really *are* sour.

For that matter, perhaps they do not even exist. One of the most striking of the very recent ideas in theoretical physics is the thought, that even for atom-models with but a few particles, even in thinking of an isolated particle, it may be altogether pointless to assign exact positions and velocities. In dealing with a swarm of particles by the statistical way, we do in effect fix the position of each corpuscle, but with a certain latitude; we fix the velocity of each, but again with a certain latitude. Perhaps this latitude, this indefiniteness, is something inherent in nature. Insisting as I am upon the contrast between the theory of the structure of the atom and the statistical

theory of matter and radiation, I may in effect be insisting on the contrast between a faulty way of visualizing some phenomena, and a correct way of visualizing all.

A function of the sort which I just mentioned, a so-called *distribution-function*, is the goal of every statistical theory. I have said that it states how many among the multitude of particles we *assume* to be located in each small element of space and to have momentum comprised in each small range of values of momentum. So it does; but the purpose of a statistical theory is, to derive it from assumptions still more fundamental, in preference to assuming it outright. Of course one might say instead, that the reason for deriving a distribution-function is to put the fundamental assumptions to their test. Whichever viewpoint one prefers, it is the distribution-function which is tested by experiment: indirectly, in that it supplies numerical values for such things as conductivity, viscosity, specific heat; and directly, for there are now immediate ways of observing it in certain cases.

A distribution-function commonly appears in an equation of this form:

$$dN = f(x, y, z, p_x, p_y, p_z) \cdot dxdydzdp_xdp_ydp_z. \qquad (1)$$

Such an equation will as a rule refer to some particular assemblage of particles, say N altogether, occupying some definite region of space: a gas in a tube, radiation in a cavity, electrons in a wire. It is to be read as follows: "dN, equal to $f \cdot dxdydzdp_xdp_ydp_z$, stands for the number of particles having coordinates in dx at x, in dy at y, in dz at z, and components of momentum in dp_x at p_x, in dp_y at p_y, in dp_z at p_z." The phrasing "in dx at x" is a succinct alternative for "between x and $x + dx$."

The function f is the distribution-function in the variables in question—here the *coordinates* of the particles referred to some Cartesian frame in the ordinary or "coordinate" space, and the components of momentum resolved along the axes of that frame, the *momenta*. Heretofore it has been customary to use the components of velocity rather than the momenta, but these are much to be preferred: partly because it is they which figure in the canonical equations, but chiefly because we shall find when we pass over to the study of assemblages of photons that the momenta play the same role in these as they do in assemblages of atoms, while the speeds of all photons are the same. There is a well-known formula for translating a distribution-function from one set of variables to another set dependent

on the first, which we shall use in special cases.[2] It is also well known that to obtain the distribution-function in *some* of the independent variables from the distribution-function of *all* of them, it is necessary to integrate the latter over the entire range of all the *other* variables: in such a case as is symbolized by equation (1), the distribution in p_z would be obtained by integrating f with respect to the first five variables over the entire range of each.

The product $dxdydz$ is an element of volume in ordinary or *coordinate-space;* the product $dp_x dp_y dp_z$ is an element of volume in *momentum-space,* in which each particle is represented by a point having for its coordinates in a Cartesian frame the values of its momenta; the product $dxdydzdp_x dp_y dp_z$ is an element of volume in *phase-space.* The function f describes the distribution of the assemblage in this phase-space of six dimensions. In some cases—for instance, that of electrons in a metal not at an even temperature, and that of oscillators —we shall have to think continually of this six-dimensional space. In others—whenever we deal with photons, and whenever we consider atoms or electrons in a region where neither temperature nor potential varies from place to place—we shall be able to assume that the distribution in the coordinate-space is uniform (that f is independent of x, y, z) and to dismiss it from mind, and to derive the distribution in the three-dimensional momentum-space quite separately as if there were no other. Even in these simplest cases it would no doubt be more consistent to operate always in the phase-space. Unhappily the human mind is so constructed, that no matter how much it may ratiocinate about space of six dimensions or six trillion, it always visualizes in space of three.

In an equation such as (1), the differential element or the product of such elements which terminates the right-hand member must be neither too large nor too small. If it is so large that f varies considerably from one point in it to another, then its multiplier, which is by definition the *mean value* of f in the said element, must be computed by the methods of integral calculus. If on the other hand it is so small that it contains only a few of the corpuscles, then the product of f into its size may be many times as great or many times as small as the number which it does contain. This is easily perceived by proceeding to the absurd limit of dividing the space into say ten times as many elements as there are corpuscles, so that in at least

[2] Let $u_1, u_2, \cdots$ represent the variables of the first set, $v_1, v_2, \cdots$ those of the second; let $f(u_1, u_2, \cdots)$ and $F(v_1, v_2, \cdots)$ stand for the distribution-functions in the two sets; then

$$F(v_1, v_2, \cdots) = f(v_1, v_2, \cdots, \frac{\partial(u_1, u_2, \cdots)}{\partial(v_1, v_2, \cdots)} .$$

nine-tenths of the elements the number of participles is zero while f is greater than zero, and in the others the number is generally much greater than f times the size of the element. To subdivide the space so finely would be to make the atomic picture too definite, and ruin it for the purposes for which we now require it.

It is not too early in this paper for me to say emphatically that the differential elements which figure in equations such as (1) *must not be identified* with the elementary compartments of the phase-space, which we shall presently encounter, and which are so important in the new statistics and in the old alike. It takes a great many of these latter to make up an element large enough to be employed in an equation like (1). Otherwise expressed: the subdivision of the phase-space into the elementary cells or compartments of the forthcoming theory is much too fine to be used in connection with the distribution-function. Much confusion may arise from failing to realize this.[3]

In speaking of the distribution-function, I have been tacitly assuming that there is such a thing as a stable, self-sustaining, changeless distribution of the atoms of a gas, the photons in a cavity, the electrons in a wire. This assumption must now be examined. It is scarcely self-evident; one might guess at first that the more numerous the particles, the more abruptly would the distribution vary from one moment to the next, and that an assemblage of 10^{20} particles would be in such unceasing turmoil that it would be senseless to imagine one single distribution for it.

Experience however shows the reverse. The gas in a tube remains uniformly dense and stationary, it does not surge forever to and fro nor huddle in a corner nor become spontaneously hot at one end and cold at the other. In the radiation in a cavity with heated walls the intensity comprised within any portion of the spectral range remains unchanged so long as the temperature of the walls is constant. The distribution-in-velocity of the electrons streaming from a heated filament does not appear to change. Moreover, when by artifice the gas in a tube is forced to assume uneven density, non-uniform temperature, or any sort of flow or turbulence, it settles down very quickly into a stagnant uniformity as soon as it is left to itself.

Now we know that while a gas is passing from an unstable state—a state of non-uniform temperature, for instance—to its stable and permanent condition, a property which we call its *entropy* and denote

[3] I am thinking particularly of the fact that in most expositions of the classical statistics one is adjured that there must be many particles in each compartment, and then in taking up the Fermi statistics one is told that there must be not more than one in each compartment; yet the two lead to formulæ which in the limiting case are the same.

by S is increasing; in certain simple cases we can evaluate this rate of change of entropy. We know that when a gas is in its stable condition, its entropy is at a maximum; we know how to compute the entropy (except perhaps for an additive constant) of a given quantity of a gas in this condition, as a function of its temperature and others of its measurable properties. And when we have evaluated both the entropy S and the energy E of a gas under any specific conditions, we know that its absolute temperature is determined by the following equation,

$$dS/dE = 1/T, \tag{2}$$

which is the definition of absolute temperature.

If we had obtained by some independent way an adequate atomic picture of entropy, so that whenever a distribution-function was suggested we could compute the value of S: then necessarily the stable distribution would be the one for which S has the greatest value compatible with the given number of particles and the given amount of energy. We do not have an independent way. But if instead we adopt some *tentative* atomic picture of entropy, some function S of which we can compute the value for any given distribution: then the test of our picture will be, whether the distribution for which this tentative S has its greatest value is verified by experiment to be the stable one. It will be found that this distribution "of maximum S" involves the derivative dS/dE, and therefore the absolute temperature; so the temperature enters into the postulated distribution-function in the course of its derivation, not by separate assumption or by an afterthought.

This method is the very notable one invented by Boltzmann, and continued by Planck. One choice of the function S which is to be identified with entropy leads to the classical or Maxwell-Boltzmann distribution-law; another leads either to the Bose or to the Fermi distribution, the difference between these two entering in at another point.

Each of these suggested functions is logarithmic; it is proportional to the logarithm of a function which is called *probability*. In theoretical physics it is a fairly general rule, that when a theorist introduces the word *probability* he is abandoning all hope of explaining by cause-and-effect the phenomena of which he is discoursing. This is the disadvantage of Boltzmann's method. The "distribution of maximum S" is baptized "the most probable distribution"; there is even a numerical estimate of its "probability," and in general it turns out to have so much greater a probability than all the others put together

that one accepts without demur the conclusions that in practice it will be stable. But there is no proof that the "most probable" distribution is always or even usually followed by another exactly like it, nor that an "improbable" distribution is always or even usually followed by another of greater probability; there is no study of the way in which one distribution is transformed into another, there are no assumptions about the collisions or encounters which presumably offer to the particles their means of interchanging speed and energy, and to the assemblage its means of approaching the stable distribution. There are other statistical methods in which account is taken of these things, and we shall have a glimpse of one of them in the last section of this paper; but the notion of causality is absent from the method which will be followed in deriving the distribution-laws of Maxwell and Boltzmann, of Bose, and of Fermi and Dirac.

These three distribution-laws will be applied to freely-flying particles in regions which are either field-free, or else pervaded by a field (electrostatic or gravitational) derivable from a potential. It may surprise the reader to hear so little about oscillators, considering that the statistics which Planck applied to these objects was the first of all the modifications of the classical statistics, was the source of the entire quantum-theory, and therefore the most important advance of the physics of the last quarter-century. The history of this period is very curious; but I cannot mention more than a couple of the salient points.

The Planckian oscillators served two purposes: they enabled Planck to derive the law of distribution of radiant energy at uniform temperature in a cavity, by supposing the radiation to be entirely wavelike and to be in equilibrium with myriads of oscillators in the walls of the cavity; and they enabled various savants to develop, step by step, a progressively improving theory of the specific heat of solids. The Bose statistics made them quite superfluous for the first purpose: by applying this statistics to the radiation supposed to consist of corpuscles, we can derive the same law of distribution without invoking the oscillators at all. As for the second: as early as 1912 (which seems remarkable, now) Debye had replaced the concept of a solid as a latticework of vibrating atoms by the concept of a solid as a system of stationary waves agitating a continuum. I do not mean to imply, of course, that the existence of the atoms was denied; I mean no more than to say, that in these statistical reasonings the individual vibrating atom was replaced by an individual pattern of stationary waves. Today we are becoming familiar with the idea

that in certain reasonings, a freely-flying electron or quantum or even an atom in a region bounded by walls may be replaced by a pattern of stationary waves filling the whole of this region. Thus it seems that the free particle, the oscillator, the stationary wave-pattern, are in close affinity with one another; they may simply represent different ways of looking at the same thing. Though on almost every page of this article I shall write in the language of the strictest corpuscular theory, it is probable that every one of the results could be translated into the language of oscillators or the language of waves.

There are still assumptions to be made about the individual particles. They are to have position, momentum and energy. Momentum may be separated into mass and velocity; often it is better left as an elementary concept. It will turn out that one of the essential differences between photons on the one hand, electrons and atoms on the other—that is to say, between the particles out of which we shall try to build a picture of radiation, and those of which we shall build models of gases and of electricity in metals—lies in the relation between momentum and energy.

Experience with matter in bulk leads to the well-known equations connecting kinetic energy K and momentum p with mass m and speed v:

$$K = \tfrac{1}{2}mv^2, \qquad p = mv \tag{3}$$

and these are supposed to hold for the ultimate particles of matter and of electricity.

During the years in which the corpuscular theory of light was struggling into existence—for, it will be remembered, light was still considered to be entirely wavelike even after Planck had founded the quantum-theory by his statistics of oscillators—Einstein proposed at two different times (1905 and 1917) the following formulæ for the energy and the momentum of photons in terms of their wavelength:

$$E = hc/\lambda, \qquad p = h/\lambda. \tag{4}$$

Historically it is interesting that he proposed the latter formula because of certain statistical studies which he had made of the equilibrium between photons and atoms. The verification of the latter by the Compton effect, of the former by the photoelectric effect and many other phenomena, is too familiar to require comment.

Now from equations (3) we deduce, for particles of matter and of electricity:

$$E = \frac{1}{2m}p^2 = \frac{1}{2m}(p_x^2 + p_y^2 + p_z^2) \tag{5}$$

and from equations (4) we deduce, for particles of light:

$$E = pc. \tag{6}$$

The difference between these two relations is responsible for some of the contrasts between radiation-gas on the one hand, electron-gas and material gases on the other; but by no means for the major part. The major difference lies in the statistical theory, as we shall now find out.

The Classical Statistics

We are going to represent three kinds of objects—ordinary or material gases, radiation in enclosures, negative electricity in metals—as assemblages of particles possessing location and momentum. We may visualize such an assemblage first as a swarm of points in ordinary space, with a coordinate-frame along the axes of which the coordinates x, y, z of the particles are measured; then as a swarm of points in momentum-space with a frame along the axes of which the momenta p_x, p_y, p_z are measured.

I will first illustrate the method of classical statistics by using it to ascertain the most likely distribution of particles in ordinary space, a case where seemingly the result may be foreseen. For it seems a truth of intuition that inside a box of ordinary space, with nothing (*e.g.* no variations of potential) to distinguish one region from another, the particles must tend to distribute themselves uniformly. This is a conclusion to which the statistical method *must* lead. The uniform distribution *must* be the most probable. How then should we define the "probability" of a distribution so that it shall be greatest for the uniform one?

But in the first place, what *is* a uniform distribution? We must divide the space—mentally, of course—into conpartments of equal volume. The distribution will then be called uniform, if the numbers of particles in the various compartments are about the same. But this clearly requires that these subdivisions be of a certain size. Their linear dimensions cannot for example be smaller than the average distance between particles, as then a "uniform distribution" would be impossible. To partition the space too finely would be like studying a painting with a microscope. The quality which we wish to define evades too sharp a scrutiny. The compartments should contain large numbers of particles, both for the stated reason and for the convenience of a certain mathematical approximation which is made.

Denote then by N the total number of particles, by m the number

of compartments into which the volume V is divided, by N_i the number of particles in the ith compartment. A distribution is described by stating all the numbers $N_1, N_2 \cdots, N_i, \cdots N_m$.

The basis of the classical statistics is the fact that if the particles have identities—if each of them is labelled by a distinctive letter, for instance—there are different ways of arranging them in the same distribution. One starts with any arrangement compatible with the prescribed "populations" $N_1, N_2, \cdots N_m$, and obtains all the other arrangements by interchanging particles *ad libitum* among the compartments, respecting only the condition that each of these shall always have as many as it had at first. The total number of distinct arrangements, the number of *permutations of the combination* $N_1, N_2, \cdots N_m$, is by a well-known theorem:[4]

$$W = \frac{N!}{N_1!N_2!\cdots N_m!}. \tag{7}$$

This number has its minimum value of unity for a distribution in which all the particles are crowded into one compartment, which would be the most non-uniform conceivable; and its maximum value for the uniform distribution, as I now proceed to show.[5]

Let us use the logarithm of W instead of W itself. If W has a maximum for any distribution so also will its logarithm, which is easier to handle, and will presently be chosen as the representation of entropy. We have:

$$\log W = \log N! - \sum \log N_i!. \tag{8}$$

Now we introduce Stirling's approximation for the factorial of a large number—by far the greatest and the most frequently invoked

[4] Imagine yourself stationed beside a set of m baskets and an urn filled with N lettered but otherwise indistinguishable balls, which are to be lifted out at random and dropped into the baskets under the following rules of the game: the first N_1 which come to your hand are to be dropped into basket 1, the next N_2 to come to your hand are to go into basket 2, and so on to the end. Having acted accordingly, you note down the assortments of balls in the various baskets, and repeat the process *ad infinitum*. Now there are $N!$ different orders in which the balls may come out of the urn. When the inspection of the baskets after two drawings reveals different results, the orders must certainly have been different. But two different orders need not reveal two different results to the inspection. Take any order, to start with; then there are $(Q - 1) = (N_1!N_2! \cdots N_m! - 1)$ others which yield the same result. For there are $N_1!$ orders in which the earliest N_1 balls emerge might come out, without any of them losing its place among the first N_1; there are $N_2!$ orders in which the next N_2 might come, without any losing its place in the second basket; and so forth. Each of the $N!$ orders then is but one among Q altogether which lead to the same result; so that there are only $N!/Q$ different results.

[5] What will actually be shown is that for the uniform distribution the function W is stationary; that it is maximum (not minimum) seems fairly obvious from the physics of the case, and can be proved.

of the mathematical aids in statistical theory. It is:

$$x! = (2\pi x)^{1/2}(x/e)^x, \qquad (9)$$
$$\log x! = x \log x - x + \tfrac{1}{2} \log (2\pi x).$$

The first two terms of this latter expression form an approximation singularly good even when x is no greater than ten or thereabouts. Using it we have:

$$\log W = \text{const.} - \sum N_i \log N_i. \qquad (10)$$

Denote by W^0 the value of W for some particular distribution N_1^0, N_2^0, $\cdots N_m^0$ and by $W = W^0 + \delta W$ its value for some other only slightly different distribution $N_1^0 + \delta N_1$, $N_2^0 + \delta N_2$, $\cdots N_m^0 + \delta N_m$. The difference between the values of $\log W$ for these two distributions is to first order of approximation:

$$\delta \log W = \delta W/W = -\sum_i (1 + \log N_i^0)\delta N_i. \qquad (11)$$

If W^0 is a maximum for the distribution $N_1^0 \cdots N_m^0$, then the difference between $\log W^0$ and the value of $\log W$ for any other slightly different or "slightly varied" distribution must vanish to first approximation. The quantity on the right of (11), the "first variation" of $\log W$, must be zero for any permitted set of values of $\delta N_1, \cdots \delta N_m$; meaning by "permitted" any set of integer values adding up to zero, for we consider an assemblage of an invariable number of particles.

Now one sees immediately that the right-hand side of (11) does vanish, if all the populations N_i^0 have the same value, say α; for then

$$\log W = -\sum_i (1 + \log \alpha)\delta N_i = \text{const.} \sum_i \delta N_i \qquad (12)$$

and the permitted variations are precisely those, for which the summation $\sum N_i$ is zero.

We do therefore reach the result which was desired. Failing it, this mode of "counting the ways in which a distribution may be realized" would have been unprofitable. As it is, the quantities W and $\log W$ are greatest for the uniform distribution which seems intuitively the most probable and is the rule for gases, and least for the utterly non-uniform one which seems the least probable. Tentatively the former is adopted as measure of the "probability" of a distribution.

I point out in passing that while the foregoing result is mathematically valid for any value of the constant α, the total number of particles prescribed for the assemblage determines the value of α which is physically permissible: viz. N/m.

We proceed to apply this method to the swarm of points in momentum-space representing the assemblage.

Like the coordinate-space, the momentum-space is to be divided into equal compartments large enough to contain each a multitude of particles. We are to define a distribution by specifying how many particles are in each compartment, and calculate as before the number W which is to measure the "probability" of the distribution. The values for W, for $\log W$ and for the variation of $\log W$ are obtained just as before. There is however an important novelty. Since the energy of a particle depends on its position in momentum-space, different distributions usually entail different values for the total energy of the assemblage. If we compute the variation of $\log W$ due to a slight change in distribution, we shall usually be computing a variation in $\log W$ correlated with a certain variation of the total energy U of the assemblage.

We now take the very great step of *identifying the quantity log W with entropy.*

More precisely, we assume that the entropy S is proportional to the logarithm of W:

$$S = k \log W, \tag{13}$$

introducing a constant factor k, and relying on subsequent experiments to teach us its numerical value.

Now when a gas being initially in thermal equilibrium at temperature T receives an infinitesimal amount of energy dE, and regains thermal equilibrium with its augmented energy, its entropy ascends by the amount of dS given by the equation:

$$dS/dE = 1/T. \tag{14}$$

If then the foregoing model of the gas and the foregoing picture of entropy are justified, the variation of $\log W$ in passing from the most probable distribution consonant with a total energy E to the most probable distribution consonant with a total energy $E + dE$ (the total number of particles remaining the same) must be equal to $(1/kT)dE$.

If we start from the most probable distribution for energy E and make *any* slight change in it involving an energy-change dE, the new distribution will presumably differ but little from the most probable distribution for $E + dE$. We therefore say: the most probable distribution for energy E is the one of which the first variation is dE/kT. This expression vanishes, if we are comparing distributions for which E is the same; which is as it should be.

It is now easily shown that such a distribution is the following

$$N_i = \alpha \exp(-\epsilon_i/kT), \tag{15}$$

in which α stands for any constant and ϵ_i for the average energy of particles in the ith compartment, which is related to the average momenta of these particles by the equation

$$\epsilon = \frac{1}{2m}(p_x^2 + p_y^2 + p_z^2)' \tag{16}$$

for we have only to write down the expression for δS as furnished by equation (11), and introduce into it the value of $\log N_i$ as supplied by equation (15):

$$\begin{aligned}\delta S &= -k\sum(1 + \log N_i)\delta N_i \\ &= -k(1 + \log \alpha)\sum\delta N_i + \sum\epsilon_i\delta N_i/T = \delta E/T,\end{aligned} \tag{17}$$

the result which was desired.

The value to be chosen for the constant α will be determined as before by the total number of particles. Denote this number by N, and conceive the compartments as tiny cubes of volume H, so that there are $1/H$ of them per unit volume of the momentum-space. The density ρ of the particles in momentum-space, which is no other than the *distribution-function in the momenta*, is given anywhere by the value of N_i/H computed for the value of energy there prevailing:

$$\begin{aligned}\rho = N_i/H &= \frac{\alpha}{H}\exp(-\epsilon/kT) \\ &= \frac{\alpha}{H}e^{-p_x^2/2mkT}e^{-p_y^2/2mkT}e^{-p_z^2/2mkT}\end{aligned} \tag{18}$$

and it is the integral of this expression over the whole of momentum-space which is equal to N:

$$N = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\rho dp_x dp_y dp_z. \tag{19}$$

The integration is easily affected; the triple integral is the product of three identical single integrals, and we have:

$$N = \frac{\alpha}{H}\left[(2mkT)^{1/2}\int_0^{\infty}e^{-w^2}dw\right]^3, \tag{20}$$

w being a symbol for each of the three momenta in turn; so that

$$\alpha = \frac{NH}{(2\pi mkT)^{3/2}}. \tag{21}$$

The expression for the number of particles in any compartment thus becomes:

$$N_i = \frac{NH}{(2\pi mkT)^{3/2}} \exp\,(-\epsilon_i/kT), \tag{22}$$

involving the four constants m, N, k and H. The first three are determinable by experiment, the third is the universal constant known in Boltzmann's honor by his name, though he himself never evaluated it. The fourth, the volume H assigned to the compartments, drops out of the distribution-functions—out of the function ρ, out of the distribution-in-energy soon to be deduced, out of the fundamental distribution-function f in the coordinates and momenta defined by equation (1), and which I now set down in place of ρ:

$$f = \frac{N}{V(2\pi mkT)^{3/2}} \exp\left(-\frac{p_x^2 + p_y^2 + p_z^2}{2mkT}\right), \tag{23}$$

V standing for the volume in ordinary space of the enclosure which contains the assemblage. This evasion of H is very deceptive; for it suggests not merely that the exact volume of the compartments is of no importance, but that the compartments themselves were invented only as a momentary stepping-stone to the distribution-functions, and should be allowed to shrink to zero like the infinitesimals of the calculus. This however is precisely what is not allowed. It is of the essence of the argument that there are compartments of finite size. As will presently transpire, I suspect that the division of momentum-space into compartments should be regarded as a quantum postulate, even in this case of the derivation of the Maxwell-Boltzmann law which seems to be at the opposite extreme from all the notions of quantum-theory.

The next step is the derivation of the *distribution-in-energy*. In preface I point out that the distribution which we are considering is, in respect to the directions of motion of the particles in ordinary space, *isotropic*. Mathematically, this occurs because p_x, p_y and p_z enter symmetrically into all the distribution functions; physically it occurs because we have made no assumption leading to a preference of any direction over any other. Later on we may establish a preferred direction by introducing a field of force, and then the impending steps may have to be reconsidered. Until then the distribution which we shall study will be described completely by saying that they are isotropic and giving the distribution-function-in-energy. This may be obtained from the distribution-function-in-the-momenta

by transforming to a polar coordinate frame in the momentum-space.[6]

I follow practically the same route.

Divide up the momentum-space into spherical "shells" by means of a sequence of spheres all centered at the origin. Each sphere corresponds to a value of ϵ, each shell to a range $d\epsilon$ of values of ϵ. Take one of the latter at random; call it shell s, denote by ϵ_s and by ϵ_{s+1} or $\epsilon_s + d\epsilon$ the energy-values at its boundary spheres, by r_s and $r_s + dr$ the radii of these, by dV the volume of the shell. Then:

$$r_s = (2m\epsilon_s)^{1/2}, \qquad dr = \left(\frac{m}{2\epsilon_s}\right)^{1/2} d\epsilon,$$

$$dV = 4\pi r_s^2 dr = \frac{2}{(\pi)^{1/2}} (2\pi m)^{3/2} (\epsilon_s)^{1/2} d\epsilon. \tag{24}$$

Suppose to begin with that each shell is large enough to contain very many compartments. The number Q_s of compartments in shell s will then be:

$$Q_s = dV/H = \frac{2}{H(\pi)^{1/2}} (2\pi m)^{3/2} \epsilon_s^{1/2} d\epsilon \tag{25}$$

and the average number of particles per compartment in shell s, call it N_s, will be:

$$N_s = \alpha \exp(-\epsilon_s/kT) \tag{26}$$

and the total number M_s of particles in the shell will be:

$$M_s = Q_s N_s = \frac{2N\pi}{(\pi kT)^{3/2}} \epsilon_s^{1/2} e^{-\epsilon_s/kT} d\epsilon = F(\epsilon_s) d\epsilon. \tag{27}$$

This is the number of particles having energy-values in $d\epsilon$ at ϵ_s. Hence the distribution-function-in-energy F is the factor multiplying $d\epsilon$ (it would be well to discard the subscript s in writing it). I have copied the value of α from (21), but it could have been derived by integrating F from $\epsilon = 0$ to $\epsilon = \infty$ and equating the integral to N.

The separation of M_s or $F(\epsilon)d\epsilon$ into two factors—Q_s the number of compartments in the shell s, N_s the average number of particles per compartment—is highly advantageous in searching for the distinctions

[6] Denote by p the quantity $(p_x^2 + p_y^2 + p_z^2)^{1/2}$ which is the magnitude of the momentum; and by θ and ϕ the angles which with p constitute a spherical coordinate system. We have

$$\rho dp_x dp_y dp_z = \rho p^2 \sin\theta d\theta d\phi dp = \frac{\alpha}{H} e^{-p^2/2mkT} p^2 \sin\theta d\theta d\phi dp$$

and the distribution-in-momentum is obtained by integrating over all values of θ and ϕ, the distribution-in-energy from it by means of the relation (5).

between the various proposed statistical laws. We shall see that in passing from one to another sometimes one of the factors is changed, sometimes the other, sometimes both.

In particular, we may pass from the Maxwell-Boltzmann law to a distribution like that which Planck derived for oscillators, simply by changing the factor Q_s. We have been dividing the momentum-space into compartments of equal volume, so that the number comprised in a shell s between spheres ϵ_s and $\epsilon_s + d\epsilon_s$ is proportional to $\epsilon_s^{1/2} d\epsilon_s$. Let us instead divide it into compartments of which the volumes increase steadily from the origin outward, at such a rate that the number in a shell s is proportional to $d\epsilon_s$ without the factor $\epsilon_s^{1/2}$.

This is, of course, not the way in which Planck's postulate is habitually stated, though it is substantially the way in which Planck stated it himself. Usually it is said, that Planck restricted the energy of the particles of the assemblage to a set of "permitted values" spaced at equal intervals: say the values a, $a + b$, $a + 2b$, $a + 3b$, $\cdots$ where a and b stand for constants. Each of these permitted values corresponds to a sphere in the momentum-space. In the shell s there are approximately $d\epsilon_s/b$ of these "permitted spheres"; the approximation being closer, the larger ϵ_s and $d\epsilon_s$ are in comparison to b. Now whether we conceive that the $d\epsilon_s/b$ sets of particles in the shell s are located on the surfaces of as many permitted spheres, or alternatively that they are scattered through as many compartments, is for the statistical results of no importance. There may be other reasons for preferring one picture to the other; but the predictions of the statistical theory are the same, whichever is adopted. I will therefore alternate between the two pictures, retaining for the moment that of a subdivision of the momentum-space into compartments; but now it will be expedient to think of these as thin spherical films, centered at the origin and increasing in volume from the innermost outward at the specified rate.

If the shell s is large enough to contain many of these compartments or permitted spheres, we may use the first approximation for the number which it contains:

$$Q_s = d\epsilon/b, \tag{28}$$

and putting the expression (26) for the number of particles per compartment, we get:

$$M_s = Q_s N_s = \frac{\alpha}{b} \exp\left(-\epsilon_s/kT\right) d\epsilon_s = F(\epsilon_s) d\epsilon_s \tag{29}$$

for the number of particles having energy-values between ϵ_s and

$\epsilon_s + d\epsilon_s$. The value of the constant is fixed as heretofore by the condition that the integral of F over the entire range of energy from 0 to ∞ shall be equal to N:

$$\int_0^\infty F(\epsilon)d\epsilon = \frac{\alpha kT}{b}\int_0^\infty e^{-w}dw = N, \tag{30}$$

so that we arrive at the following distribution-in-energy function:

$$F(\epsilon) = \frac{N}{kT}\exp(-\epsilon/kT). \tag{31}$$

This function certainly does not display any feature which suggests the achievements of Planck! It looks as smooth and continuous as the Maxwell-Boltzmann function itself, and the constant b, the step or interval between the successive permitted energy-values or the boundaries of successive compartments, is nowhere to be seen. The constant b however has slipped out for the same reason as the constant H from the function (23), and the apparent continuity is due in both cases to the same cause. In preparing and effecting the integration (30) in order to obtain a value for the constant α, we assumed that the various permitted energy-values within the range $d\epsilon_s$ are all sufficiently nearly equal to be identified with the single value ϵ_s. That is to say, we smoothed over the discontinuities which had previously been brought in by the assumption of separate compartments. No wonder that there is not a sign of them in the function (31), even as there is not a sign of them in the Maxwell-Boltzmann law!

We might however avoid this smoothing-over, if we could attain the value of α by an actual summation over the various compartments instead of by integration. Now with Planck's postulate this is mathematically feasible and indeed easy. For the number of particles in the ith compartment being

$$N_i = \alpha \exp(-\epsilon_i/kT) \tag{32}$$

the total number of particles is computed thus:

$$\begin{aligned} N = \sum N_i &= \alpha e^{-a/kT}\sum_{i=0}^{\infty} e^{-ib/kT} \\ &= \frac{\alpha e^{-a/kT}}{1 - e^{-b/kT}} \end{aligned} \tag{33}$$

by virtue of the very convenient consequence of the binomial theorem that $(1 + x + x^2 + \cdots) = (1 - x)^{-1}$; so that for α we obtain the

exact value:

$$\alpha = Ne^{(a-b)/kT}(e^{b/kT} - 1) \tag{34}$$

and for the populations of the various compartments, the formula:

$$N_i = Ne^{-b/kT}e^{-ib/kT}(e^{b/kT} - 1). \tag{35}$$

Here the discontinuity implied in the classical picture of a momentum-space divided into compartments is admitted and accepted, as it never was in the process of deriving the Maxwell-Boltzmann law. Planck did not put discontinuity into the classical statistics; it was there already; he refrained from disregarding it. Instead of confining his studies to the circumstances in which it can safely be ignored, he extended them to ranges where it had to be taken account of, and he took account of it.

As I intimated, the distribution (35) was proposed by Planck not for freely-moving particles, but for oscillators. The "Planckian oscillator" may be visualized as a particle which executes simple-harmonic vibrations back and forth in a straight line across a position of equilibrium, to which it is attracted by a force proportional to its displacement. It is like a free particle, in that its state at any moment is described by giving the values of its position q and momentum p, q being measured from its point of equilibrium; but it is unlike a free particle in that its energy depends not on p alone but on both p and q, being a function of the form $(Ap^2 + Bq^2)$. Therefore we must envisage not the momentum-space alone but the phase-space of the variable p and q. In principle it would have been better, had we envisaged the phase-space all along; but since for an assemblage of free particles that space has six dimensions, it was impractical to visualize more than the momentum-space, and since the energy depended only on the momenta that compromise was not detrimental except for one feature which I can later introduce. Here the compromise would be ruinous, but it is unnecessary since the phase-space has only two dimensions.

Visualize then this two-dimensional phase-space as a plane with p and q axes at right angles to each other. Suppose all the oscillators to have the same mass and the same natural frequency, which is to say, the same values of the constants A and B in the above-mentioned formula for their energy; but let them differ in amplitude. The point representing any oscillator in the phase-space runs round and round in an elliptical orbit centered at the origin. Different amplitudes correspond to different ellipses. The energy of an oscillator depends on its amplitude; therefore different energy-values correspond

to different ellipses, and reversely. If we divide the phase-space into compartments by a succession of ellipses centered at the origin, each of these compartments corresponds to a specific range of energy-values. If the dividing ellipses are so spaced that these compartments are of *equal* area (equal volume of the phase-space), they correspond to *equal* ranges of energy-values—an important difference between this case and the one which was previously treated.

If the dividing ellipses are spaced to form equal compartments, they themselves correspond to energy-values forming a linear sequence: call these a, $a + b$, $a + 2b$, $\cdots a + ib \cdots$ as before. Whether we call these the "permitted" energy-values and allow the oscillators only the choice among them, or whether we sprinkle the oscillators uniformly through the compartments, makes only a secondary difference. In this case, in fact, we can easily see exactly what difference it makes. If the oscillators are sprinkled uniformly in each compartment, then by applying the classical statistics we get just the same distribution (35) as when we assume them restricted to the energy-values $(a + ib)$. But when we undertake to evaluate the average energy of all the oscillators, then in the one case we must put down the mean energy of those in the ith compartment as the arithmetic mean of the values $a + ib$ and $a + (i + 1)b$, while in the other case we must put down the energy of those at the ith permitted ellipse as $a + ib$. Hence to change over from the picture of permitted energy-values to the picture of compartments is the same thing as to replace the original sequence of permitted energy-values by another sequence of values located midway between them. I mention this chiefly in order to emphasize that the subdivision of phase-space into compartments is *ipso facto* quantum-theory.

As every reader knows, Planck postulated that the quantity b—the interval between the permitted energy-values, or the energy-range within a compartment, whichever picture is chosen—is the product of a universal constant (h) and the frequency of the oscillators (ν). The area of the equal compartments is then equal to the universal constant [7] whatever the frequency of the oscillators. From this latter statement the general principle is derived: To state it one must first adopt a symbol (say n) and a name (say *number of degrees of*

[7] The point in the phase-space representing an oscillator of mass m, frequency ν, and amplitude C describes an ellipse having semi-axes C and $2\pi m\nu C$ and area $2\pi^2 m\nu C^2$; its energy is $U = 2\pi^2 m\nu^2 C^2$; hence the relation between energy U and area F is

$$U = \nu F$$

and the area between two ellipses is equal to h if the energy-difference between them is equal to $h\nu$.

freedom) for the number of distinct coordinates q required to describe the individual member of whatever assemblage one may be considering; this is also the number of distinct *momenta* p, there being one p for each q. Then the principle generalized out of Planck's postulate for oscillators is this: *For an assemblage of individuals with n degrees of freedom the phase-space is to be divided into compartments of volume* h^n.

We will now see what the classical statistics, supplemented by this principle, proposes for an assemblage of particles for which the relation between energy and momentum is $\epsilon = cp$ as it is for corpuscles of light, instead of $\epsilon = p^2/2m$ as it is for corpuscles of matter.

Different energy-values correspond as before to different spheres all centred at the origin of the momentum-space, but the numerical relations are changed. Instead of equations (24), we have:

$$\begin{aligned} \epsilon_s &= cr_s, \qquad d\epsilon_s = cdr_s, \\ dV &= 4\pi r_s^2 dr_s = (4\pi/c^3)\epsilon_s^2 d\epsilon_s, \end{aligned} \tag{36}$$

dV standing for the volume of the shell s covering the energy-range between ϵ_s and $\epsilon_s + d\epsilon_s$. Divide the momentum-space into compartments of equal volume H. We derive the "smoothed-over" distribution-function for the case in which N_i varies so little from one compartment to the next that even when the shell s is thick enough to comprise very many compartments the values of N_i for all of them may be equated to a mean value N_s. Under these conditions we may write for the number of compartments in the shell s,

$$Q_s = dV/H = (4\pi/c^3H)\epsilon_s^2 d\epsilon_s. \tag{37}$$

Putting down the classical value (15) for the number of particles in any of these compartments, remembering that N_i is identified with N_s, we obtain for M_s the number of particles in the shell s:

$$M_s = Q_s N_s = \alpha(4\pi/c^3H)\epsilon_s^2 e^{-\epsilon_s/kT} d\epsilon_s \equiv F(\epsilon_s)d\epsilon_s \tag{38}$$

and evaluate α by the same procedure as before. The result is:

$$F(\epsilon) = \frac{N}{3k^3T^3}\frac{\epsilon^2}{e^{\epsilon/kT}}. \tag{39}$$

This is the smoothed-over distribution-in-energy predicted for the radiation-gas by the classical statistics, it being assumed that the momentum-space is to be divided into compartments of equal volume. Experiment however supplies a quite different distribution-in-energy, to wit:

$$F(\epsilon) = \left(\frac{8\pi V}{c^3h^3}\right)\frac{\epsilon^2}{e^{\epsilon/kT} - 1}. \tag{40}$$

It looks as if (39) might be the limiting form of (40)—as if the actual distribution-law might be obtained by avoiding the approximations whereby we came to the formula (39), as Planck's law of distribution for oscillators was obtained by refraining from approximation. Such however is not the case. True, the second factor in (39) is evidently the limiting form, for very high temperatures, of the second factor in (40). But the first factor in (40) contains nothing but the volume of the gas and some universal constants, while the first factor in (39) contains the temperature and an apparently disposable constant standing for the number of particles in the assemblage. The former is not the limit of the latter. It will be noted also that although I said that the volume of the elements of phase-space was to be set equal to h^3, this assumption in no wise enters into the function (39). Bose in fact found it necessary to upset the basis of the classical statistics, in order to arrive at (40) instead of (39).

The Bose Statistics

The momentum-space of the photons is to be divided as heretofore into equal compartments, and various distributions of the particles among these are to be compared, in order that we may elect one of them as "the most probable" and make a picture of the entropy of the assemblage. But the manner of defining a distribution, the manner of "counting the ways" in which it may be realized and computing its "probability," is to be changed, and changed in a most thoroughgoing and fundamental way.

Start with any distribution of the particles, defined as heretofore: defined that is, by saying that there are N_0 of the particles in the compartment 0, N_1 in the compartment 1, and in general N_i in the compartment i.

Count the number of compartments containing no particle; call it Z_0. Count the number of compartments containing one particle apiece; call it Z_1. In general, let Z_i stand for the number of compartments containing i particles apiece. Put down the values of all the numbers Z_i.

Now change the terminology. Elect some neutral word, "arrangement" say, to denote what we have heretofore denoted as a "distribution," and use the latter word in the following new sense: a *distribution* shall henceforth be described by stating the values of the numbers $Z_0, Z_1, Z_2, \cdots, Z_i, \cdots$ and the total energy of the assemblage.[8]

[8] It is of course confusing thus to change the meanings of words, but in the long run less confusing (I think) than to use some other word than *distribution* for the concept always called by that name in the new statistics.

This means that each distribution in the new sense comprises a number of distinct distributions in the old sense. This we shall regard as the number of different ways in which the distribution in the new sense may be realized. Going over entirely to the new terminology: we shall now identify the probability W^* of a distribution with the number of arrangements which are included in it. Previously we identified the probability W of an arrangement with the number of permutations included in it, according to equation (7). This we now must forget; we must proceed as if the probability of each arrangement were the same.

The number W^* is now to be evaluated. In doing this we must remember that we have to count, not the total number of arrangements yielding the prescribed set of values of the quantities Z_i, but the portion of these which give the prescribed value to the total energy of the assemblage.

As before, we superpose upon the partitioning of the momentum-space into small compartments of equal volume H, another partitioning into spherical shells each of which is sufficiently large to contain many of the compartments, yet sufficiently small so that the same value of ϵ may be assigned to all the compartments within it. The final result is thus to be a "smoothed-over" formula. It is rather singular that whereas Planck introduced the quantum into physics by avoiding the smoothing-over which had been customary in the classical statistics the quantum-formula for radiation is now derived by a method in which it is accepted.

Consider then any shell at random, say the "shell s." Denote by Q_s or by Z the total number of compartments in it; by Z_{is} the number of these compartments which contain i particles apiece; by M_s the number of particles in the shell. According to the scheme now being tried out, the number of ways of attaining the particular distribution characterized by the numbers Z_{is} is given by the formula:

$$W_s^* = \frac{Q_s!}{Z_{0s}!\,Z_{1s}!\,Z_{2s}!\cdots}, \qquad Q_s = \sum_i Z_{is}. \tag{41}$$

As the energy-values for all the compartments in the shell are (by hypothesis) approximately the same, these various ways of attaining the distribution Z_{is} all corresponding to approximately the same total energy, as well as the same total number of compartments and the same total number of particles.

Suppose this process repeated for every one of the shells s. The total number of ways of attaining the actual distribution, compatible

with the conditions of constancy of total energy, total number of particles and total number of compartments, is then the product of all the quantities W_s^*: call it W^*:

$$W^* = \prod_s W_s^*. \tag{42}$$

As before, and with the same end in view, we form the expression for $\log W^*$, and employ Stirling's formula (assuming thus in effect that none of the quantities Z_{is} is smaller than ten or so):

$$\log W^* = \sum_s \log W_s^* = \sum_s [Q_s \log Q_s - \sum_i Z_{is} \log Z_{is}]. \tag{43}$$

We now take the very great step of *identifying* not $\log W$ of equation (10), but $\log W^*$, *multiplied by a constant* k, *with the entropy of the assemblage.*

$$S = k \log W^*. \tag{44}$$

Then, when the numbers Z_{is} are changed by small amounts δZ_{is}, the ensuing change δE in the total energy E of the assemblage must be linked to the ensuing change in $(k \log W^*)$ by the equation:

$$\delta S = \delta(k \log W^*) = \delta E/T. \tag{45}$$

The first variation of $(k \log W^*)$ is given thus:

$$\delta(k \log W^*) = -k \sum_s \sum_i (1 + \log Z_{is}) \delta Z_{is}. \tag{45a}$$

Let us try the distribution:

$$Z_{is} = \alpha_s e^{-i\epsilon_s/kT}. \tag{46}$$

Substituting this expression into (45a), we get:

$$\begin{aligned} \delta(k \log W^*) &= -k \sum_s \sum_i (1 + \log \alpha_s - i\epsilon_s/kT) \delta Z_{is} \\ &= -k \sum_s (1 + \log \alpha_s) \sum_i \delta Z_{is} + \frac{1}{T} \sum_s \epsilon_s \sum_i i \delta Z_{is} \\ &= \delta E/T, \end{aligned} \tag{47}$$

for the summation $\sum_i \delta Z_{is}$ vanishes because the number of compartments in each shell is invariable, while the quantity $\epsilon_s \sum i Z_{is}$ is equal to the total energy of the particles in the shell s.

The new statistics, in proposing a new conception of entropy as embodied in equation (44), therefore leads to a new distribution for thermal equilibrium. This distribution is expressed by equation (46) in the new-fashioned way, by stating the number of compartments

in each shell which contain each of the permissible quotas of particles. We must translate it into a distribution-function-in-energy such as we used to express the results of the old statistics.

Before undertaking this translation, we compute the values of the constants α_s by summing the numbers Z_{is} over all values of i for each shell separately, and equating the sum to the total number Q_s of compartments in the shell. We obtain:

$$Q_s = \sum_i Z_{is} = \alpha_s \sum_i e^{-i\epsilon_s/kT} = \alpha_s(1 - e^{-\epsilon_s/kT})^{-1}. \tag{48}$$

On substituting these values of α_s into (46) we get something which begins to look familiar.

Next for the number M_s of the particles in the shell s, we compute:

$$M_s = \sum_i iZ_{is} = \alpha_s e^{-\epsilon_s/kT}(1 - e^{-\epsilon_s/kT})^{-2}$$

$$= Q_s \frac{1}{e^{\epsilon_s/kT} - 1}, \tag{49}$$

which begins to look very familiar indeed.

Now for Q_s, the number of compartments in the shell s, we put the value already stated in equation (37), derived from the assumption that the compartments are all of the same volume H:

$$M_s = \frac{4\pi}{c^3 H} \frac{\epsilon_s^2}{e^{\epsilon_s/kT} - 1} d\epsilon_s \equiv F(\epsilon_s)d\epsilon_s. \tag{50}$$

Here the function $F(\epsilon_s)$ is the "smoothed-over" distribution-in-energy in which the new statistics culminates. Unlike those which we earlier derived from the old statistics, it involves the volume of the elementary compartments directly. Whereas from the old statistics we obtained formulæ involving the quantum only by avoiding the approximation, here we obtain a quantum-formula even when we admit the approximation—a contrast on which, I think, it is worth while to insist.

Let us then, in preparing for the final assumption, accept the principle generalized from Planck's assumption about oscillators: let the elementary cell of phase-space be given the volume h^3. This is not yet an assumption about the compartment of momentum-space. Supplement it, then, by supposing that the compartment of phase-space h^3 is the product of the compartment H of momentum-space and the entire volume V occupied by the radiation-gas. Then:

$$H = h^3/V. \tag{51}$$

This assumption—let me remark in passing—takes a very elegant form if we replace the compartments by the permitted energy-values, and then the corpuscular picture by the wave picture; for then we have a series of permitted wave-lengths, which are precisely those which can form stationary waves in a cube of volume V.

So the new statistics leads to the distribution-law:

$$F(\epsilon)d\epsilon = \frac{4\pi V}{c^3h^3} \frac{\epsilon^2}{e^{\epsilon/kT} - 1} d\epsilon. \tag{52}$$

Dividing out the factor V, we get the number of particles per unit volume having energy-values between ϵ and $\epsilon + d\epsilon$, in an assemblage having the most probable distribution at the temperature T. Multiplying this by ϵ, we get the total energy per unit volume in the possession of such particles. Identifying these particles with photons, we observe that they have wave-lengths between ch/ϵ and $ch/(\epsilon + d\epsilon)$, frequencies between ϵ/h and $(\epsilon + d\epsilon)/h$. Transforming then from the variable ϵ to the new variables λ and ν, we obtain distribution-functions which give the *density of radiant energy* as functions of wave-length and frequency. It turns out that these agree absolutely with the observed distributions, except that they lack a factor 2. This factor is at once imported, and is ascribed to the fact that light is polarizable. So we arrive at the black body radiation-formula:

$$\rho(\nu)d\nu = \frac{8\pi h}{c^3} \frac{\nu^3 d\nu}{e^{h\nu/kT} - 1} \tag{53}$$

and the new statistics is justified by its success.

It will be observed that the new statistics leads to a precise value for the number of photons per unit volume, at any prescribed temperature; whereas the old statistics led to nothing of the sort, but to a formula which contained the number of atoms per unit volume as a disposable constant. This corresponds to a profound physical difference between radiation-gas and material gases. When I state the temperature and the volume of a box containing helium, I am not giving data enough to fix the quantity of helium inside the box; on the contrary, the quantity and the density of the helium in the box can be varied *ad libitum* while the temperature and the volume are held constant. But when I state the temperature and the volume of an enclosure containing radiation, I am giving data sufficient to fix the amount of radiant energy and the number of quanta in the enclosure absolutely. This is a fact of experience, and the new statistics is evidently in accord with it. But if one were tempted to

try out the new statistics upon a material gas, would there be any way of avoiding the inadmissible conclusion that the number of atoms in such a gas is also absolutely fixed by temperature and volume?

There is such a way. One might replace the distribution proposed in equation (46) by a more general one involving a disposable constant B, as follows:

$$Z_{is} = \alpha_s e^{-iB - i\epsilon_s/kT}. \tag{54}$$

On substituting this into the expression for $(k \log W)$ we get instead of (47) the equation:

$$\delta(k \log W^*) = -k\sum_s (1 + \log \alpha_s)\sum_i \delta Z_{is} + kB\sum_s\sum_i i\delta Z_{is} + \frac{1}{T}\sum_s \epsilon_s \sum_i i\delta Z_{is}. \tag{55}$$

The right-hand member must as before reduce to $\delta E/T$ if the distribution (54) is acceptable; and this it will do, provided that not only $\sum_i \delta Z_{is}$ but also $\sum_s \sum_i i\delta Z_{is}$ is zero. Now the second of these quantities is zero for all variations in which the total number of particles remains the same. The distribution (54) enjoys a greater entropy than any other which is compatible with the same total energy and the same total number of particles. The distribution (46) was still more exalted; it enjoyed a greater entropy than any other compatible with the same total energy, even including those for which the total number of particles was somewhat different. But the distribution (54) is sufficiently distinguished to be qualified as the most probable distribution for a material gas. It seems rather singular that the distribution (46) is required for radiation-gas. Here is evidently one of the deep differences between matter and radiation.

Following the same routine as before, we arrive at the following expression for the number of particles in the shell s:

$$M_s = Q_s \frac{1}{e^{B+\epsilon_s/kT} - 1}, \tag{56}$$

and in dealing with radiation-gas we have put $B = 0$ and have taken the value of Q_s from equation (37). If in dealing with a material gas we take the value of Q_s from equation (25) instead and put $H = h^3/V$ we obtain:

$$M_s = \frac{2V}{h^3(\pi)^{1/2}}(2\pi m)^{3/2} \frac{(\epsilon_s)^{1/2} d\epsilon_s}{e^{B+\epsilon_s/kT} - 1} \equiv F(\epsilon_s)d\epsilon_s, \tag{57}$$

and now it is obvious that we must evaluate B in terms of the total

number of particles N by the already so familiar way of integrating $F(\epsilon_s)$ over the entire energy-range from 0 to ∞ and setting the integral equal to N.

Einstein proposed this as an alternative to the Maxwell-Boltzmann law derived from the classical statistics. It is not easy to decide which of the two is supported by experiment, as with increasing temperature the formula (57) becomes more and more nearly like the classical one, and it turns out that throughout the convenient ranges of temperature and pressure the two are indistinguishable. It would be very valuable to determine between the two, as then we should know which of the two ways of defining a distribution and estimating the probability thereof, which of the two pictures of entropy, is the proper one for a material gas. The reader may have remarked that if one were to apply Bose's method to the problem of determining the most probable distribution of particles in ordinary space, one would reach a result at variance with that of the classical statistics, and therefore at variance with intuition. One must deal altogether with the six dimensional phase-space, to be perfectly consistent. This is to be regretted.

The Fermi Statistics

The statistics invented by Fermi, and later independently by Dirac, involves the same fundamental assumptions as that of Bose—the same manner of counting the ways in which a distribution may be realized, of defining its probability, of picturing its entropy. But there is an additional assumption, of the nature of a limitation: it is postulated, that a compartment may contain not more than some specific maximum number of particles. In particular for a gas to which no external field is applied, it is postulated that each compartment must either be empty, or else contain one particle only.

The "exclusion-principle of Pauli" gave the hint from which the Fermi-Dirac theory sprang. This principle may be paraphrased as follows. In Bohr's "atomic theory of the atom" the electrons belonging to an atom are forbidden to revolve in any except certain specific orbits, set apart from the rest as the "permitted" orbits, and labelled by specific "quantum-numbers." In later versions of the theory the "permitted orbits" are less conspicuous, the "permitted quantum-numbers" more so; but the picture is acceptable at all events as a beginning. Upon this prohibition, then, Pauli superposed another; not more than one electron is allowed in each orbit or to each set of quantum-numbers. Perhaps it would be better to say "not more than some definite small number of electrons . . ."

instead of "not more than one." The affinity of this to one of Fermi's assumptions will soon be manifest. It would take much too long to give an idea of the successes of the Pauli principle; they are however so great as to increase the inherent plausibility of Fermi's idea very much—or perhaps I should say, so great as to render the idea plausible, which otherwise it might not seem.

The reasoning follows exactly the same course as when we were deriving the distribution-law (56), except that all the summations over the variable i are now summations of two terms only, the term for $i = 0$ and the term for $i = 1$. For each of the shells there are only two numbers Z_{is} required to describe the distribution: viz. Z_{0s} the number of empty compartments and Z_{1s} the number of compartments containing one particle apiece. We try the distribution (54):

$$Z_{0s} = \alpha_s; \qquad Z_{1s} = \alpha_s e^{B-\epsilon_s/kT}, \tag{58}$$

and easily find that it is the distribution of maximum probability, by comparison with all the others compatible with the same total number of particles and the same total energy. We arrive then at the following expression for the number of particles in the shell s:

$$M_s = Q_s \frac{1}{e^{B+\epsilon_s/kT} + 1}. \tag{59}$$

There is no point in putting for Q_s the value appropriate to radiation-gas, since the Bose formula has already proved adequate for that case. Fermi put the value appropriate to material gases, and obtained:

$$M_s = \frac{2V}{h^3(\pi)^{1/2}} (2\pi m)^{3/2} \frac{(\epsilon_s)^{1/2} d\epsilon_s}{e^{B+\epsilon_s/kT} + 1} = F(\epsilon_s) d\epsilon_s. \tag{60}$$

This formula is the point of departure for the theory of the electron-gas in metals revived and remodelled by Pauli and Sommerfeld, to the experimental tests of which most of the rest of this article will be devoted.

Application of the Fermi Statistics to the Electrons in Metals

We are asked to conceive of a piece of metal as a region populated with "free" electrons, and surrounded by a wall; the electrons being distributed according to the formula of Fermi.

The Fermi distribution-function involves the total number N of the electrons, which is a disposable constant. It also involves the volume V which this assemblage of N particles pervades. For this

we set the volume of the piece of metal—a decision which is tantamount to ignoring the atoms, to supposing the metal a vacuum inhabited by free electrons only. So remarkable an assumption, even though it be made only in approximation, requires some excuse. Its strangeness may be mitigated by recalling, first, that slow electrons may go through atoms (at least through certain kinds of atoms) as imperturbably as if the atoms were not there; and second, that a wave-train may go without being scattered at all through a crowd of particles individually quite able to scatter it, provided that the particles are arranged in a regular lattice having a spacing smaller than the wave-length of the waves. The speeds attributed to electrons in metals are so low and their wave-lengths are so great, that perhaps they do behave in such a way.

The "wall" is the agency which prevents the electrons from escaping; it is commonly imagined as a sharp and sudden gradation of potential at the surface of the metal. Any electron moving towards it from within, with a velocity of which the component normal to the bounding surface may be denoted by u, is supposed to be driven back into the body of the metal if the corresponding "component of kinetic energy" $\frac{1}{2}mu^2$ is less than a certain constant W_a; while if $\frac{1}{2}mu^2 > W_a$ the electron escapes, but with its kinetic energy diminished by W_a. According to newer ideas electrons may sometimes escape even when their values of $\frac{1}{2}mu^2$ are smaller than W_a, and may sometimes fail to escape in the contrary case; but the earlier and simpler conception remains approximately valid, and I will abide by it for a time. The constant W_a may be named the *work-function*.

Like the constant N, the work-function figures as a disposable constant in the theory. It is an ambition of physicists to explain as many as possible of the differences between different metals, by varying only the values of these two constants. Later we shall find it necessary to introduce others, beginning with the one which in the older theories appeared as the mean free path of the electrons; but there are several results of value which can be obtained with no other but these two.

I repeat now from (60) the Fermi formula for the distribution-in-energy of an assemblage of N particles in volume V at temperature T, with two changes made to bring the notation into harmony with that of Sommerfeld:

$$F(\epsilon) = G\frac{2}{(\pi)^{1/2}}\frac{V}{h^3}(2\pi m)^{3/2}(\epsilon)^{1/2}\frac{1}{A^{-1}e^{\epsilon/kT}+1}. \qquad (61)$$

Here the symbol $1/A$ replaces e^B, and a factor G to which we shall

assign the value 2 is introduced for a reason which will be stated later. The corresponding distribution-function in the coordinates and momenta is this:

$$f(x, y, z, p_x, p_y, p_z) = \frac{G}{h^3} \frac{1}{A^{-1}e^{\epsilon/kT} + 1}. \tag{62}$$

The first step now is the same as in the classical statistics: to determine the constant A in terms of N by integrating $F(\epsilon)$ over the whole range of energy-values from 0 to ∞, and equating the integral to N:

$$\int_0^\infty F(\epsilon) d\epsilon = N. \tag{63}$$

This was an easy step in the classical statistics, but here it is very hard. The integral of $F(\epsilon)$ is not one of the common well-known functions to be found in mathematical tables, nor a combination of such; and we do not get a simple equation to be solved for A in terms of N. Sommerfeld indeed found it necessary to compromise by deducing two series-expansions for the integral, one being available for values of A smaller than unity, the other for the opposite extreme. By a stroke of luck which seems almost too good to be true, the first one or two terms of one or the other of these expansions form an approximation amply good enough for all the cases where as yet theory and experiment can be compared.

I consider first the approximation which is of *no* importance in the theory of electrons in metals—the one for values of A so very small that the second term in the denominator of $F(\epsilon)$ is negligible by comparison with the first. Then the distribution-law approaches that of Maxwell and Boltzmann, and of necessity the constant A must possess the value which in the limit makes $F(\epsilon)$ identical with the classical expression written in equation (27): to wit, the value:

$$A = \frac{Nh^3}{GV}(2\pi mkT)^{-3/2}. \tag{64}$$

It would however be a great error to suppose that this value of A can be substituted into the function F under all circumstances. This value is acceptable only if A is very small relatively to unity, which is to say, if the quantity to which A is here equated is very small. So the question arises: in any physical case, is the combination on the right-hand side of equation (64) a small fraction of unity, or is it not?

Now for any material gas under any conditions usual in the labora-

tory, A is indeed very small. The new statistics leads to a result indistinguishable from that of the old statistics. To discriminate between the two by experiments on material gases, one would have to work with temperatures so low and densities so high that the gases would probably either be liquefied already, or at least would be in a condition very different from that "ideal" state to which the statistics is tacitly supposed to apply. Perhaps though it is not impossible to make the test with helium or hydrogen.

A this point apparently Fermi stopped. But it occurred to Pauli that if the new statistics were applied to an electron-gas as dense as that which Riecke and Drude had supposed to pervade the interiors of metals, the deviations from the classical distribution would be much more pronounced. For, in the first place, the mass m of the individual electron is smaller by several orders of magnitude than the mass of the atoms or molecules of any material gas. And, in the second place, if the number N of free electrons in a piece of metal is as great as or greater than the number of atoms, then it is thousands of times as great as the number of particles in an equal volume of a material gas. Now the quantity equated to A in equation (64) contains N in the numerator and $m^{3/2}$ in the denominator, and for the hypothetical electron-gas within the metals it is no longer small. The expression (64) for A is then no longer acceptable.

While the statement just made about m is based on a fact of experience, the statement about N is not so firmly grounded. We have no direct knowledge of the number of free electrons in a given volume, say the number n $(= N/V)$ in unit volume, of a metal. This as I said above is a disposable constant of the theory. One of the tests of the theory is whether one can obtain correct numerical values of half-a-dozen properties of a metal by choosing a single value of n for that metal. So long as the classical statistics was applied to the electron-gas, this was impossible. If the value of n was put as high as the number of atoms in unit volume, the predicted value of specific heat (and we may now add, the predicted value of susceptibility) turned out to be too large; if n was lowered sufficiently to avoid this particular discordance, other predictions were impaired. It was however the general impression, that one should put n equal to the number of atoms or a small multiple thereof. I suspect that this decision was largely due to a feeling that since the free electrons are detached from atoms, and since all the atoms are alike, any atom should supply as many free electrons as any other. However that may be, it was natural though not inevitable for Pauli and for Sommerfeld to link the Fermi statistics with the postulate that there are as

many free electrons as there are atoms, and test the combination of these two assumptions.

On putting for m the mass of the electron, for N/V the number of atoms per unit volume of any metal, for T any temperature from zero absolute up to several thousand degrees, and for G any small integer, one finds that the quantity equated to A in (64) is very large. Thus with $N/V = 5.9 \cdot 10^{22}$ (the number of atoms in a cc. of silver), $T = 300°$ K., $G = 2$, Sommerfeld computed:

$$(nh^3/G)(2\pi mkT)^{-3/2} = \text{about } 2400 \tag{65}$$

a result which invalidates equation (64).

We turn then to the other series-expansion of the integral $\int F(\epsilon)d\epsilon$, the one which Sommerfeld proved applicable for large values of A. The first two terms of this expansion are as follows:

$$\int_0^\infty F(\epsilon)d\epsilon = N = \frac{GV}{h^3}\frac{4\pi}{3}(2mkT \log A)^{3/2}\left(1+\frac{\pi^2}{8}(\log A)^{-2}+\cdots\right). \tag{66}$$

Taking the first term only of this expansion and putting the aforesaid values of N/V and T and solving for $\log A$, one finds a very large value indeed ($\log_e A = 325$). Assuredly then we may use the first two terms of this expansion by themselves when we are dealing with the electron-gas in a metal, and indeed the first term will for some purposes be amply sufficient.

We have thus the following first—and second—approximation formulæ for A in terms of n or N/V:

$$\begin{aligned} 2mkT \log A &= h^2(3n/4\pi G)^{2/3} && \text{first approx.} \\ 2mkT \log A &= h^2(3n/4\pi G)^{2/3}\left[1-\frac{(2\pi mkT)^2}{12h^4}\left(\frac{3n}{4\pi G}\right)^{-4/3}\right] && \text{second approx.} \end{aligned} \tag{67}$$

(the second approximation being computed by putting the first-approximation value of $\log A$ into the second term of the series expansion).

On substituting one or the other of these into the distribution-functions (61) and (62), we have the postulated distribution of the free electrons expressed to as high a degree of approximation as we require, with no disposable constant except n; and we are ready for the applications.

The Specific Heat

As it was the notorious difficulty with the specific heat which spoiled the old electron-gas theory in which the classical statistics

was coupled with the assumption that there are as many free electrons as atoms, let us first of all find out whether this difficulty remains.

Any distribution-law for an assemblage leads immediately to a formula for the total energy E thereof as function of the temperature, which is:

$$E = \int_0^\infty \epsilon F(\epsilon) d\epsilon. \tag{68}$$

Putting the distribution-in-energy (27) derived from the classical statistics, we find:

$$E = \frac{3}{2} NkT \tag{69}$$

and putting the one just derived from the Fermi statistics with the first-approximation value of A, we find:

$$\begin{aligned} E &= E_0 + \tfrac{1}{2}\gamma V T^2, \\ E_0 &= \frac{2\pi}{5} \frac{VGh^2}{2m} \left(\frac{3n}{4\pi G} \right)^{2/3}, \\ \gamma &= \frac{\pi}{3} mG \frac{(2\pi k)^2}{h^2} \left(\frac{3n}{4\pi G} \right)^{1/3}, \end{aligned} \tag{70}$$

two exceedingly different formulæ.

The derivatives dE/dT of these expressions are the formulæ supplied by the two statistics for the specific heat of the electron-gas. The classical theory predicts for the specific heat a constant value, while the Fermi statistics makes it proportional to the temperature—being thus in harmony with Nernst's heat theorem, while the other is not—and gives it even at room temperatures but a small fraction of the classical value.

Experimentally the specific heat of the electron-gas cannot be measured separately from that of the lattice of atoms, which constitutes the metal—an admission which seems to condemn as vain all hope of testing these formulæ. Nevertheless one can conclude with fair certainty that the classical expression is inadmissible; for the specific heat of an ordinary metal agrees so well with the value attributed by statistical theories both old and new *to the atoms alone*, that there is simply none left over for the electrons—no such great excess, that is to say, as the amount $3nk/2$ which the classical theory requires. The device of reducing n to so low a value that $3nk/2$ would be inappreciable makes trouble in other directions, as I have intimated. But with the new statistics the theoretical value γVT is inappreciable even when n is made as great as the number of the atoms and T as

great as many hundreds of degrees. Probably no one who did not often lament the defeat of the old and so very desirable electron-gas theory by that hard fact about the specific heat will ever quite realize the rejoicing caused by this victory of the new, which by this achievement succeeded *a se faire pardonner* many deficiencies in other fields.

Features of the Fermi Distribution

I will now mention some of the features of the Fermi distribution which has thus justified itself by passing its first test.

The most startling of these may be inferred from the distribution-function (62) or (61), by inserting the first-approximation formula for A presented in equation (67), and a new symbol W_i:

$$f = \frac{G}{h^3} \frac{1}{e^{(\epsilon - W_i)/kT} + 1}; \qquad W_i = \frac{h^2}{2m} \left(\frac{3n}{4\pi G} \right)^{2/3}. \tag{71}$$

At the absolute zero the exponential term is either infinity or zero, according as the variable ϵ is greater or less than W_i. Therefore the density of the electrons in phase-space is constant and equal to G/h^3 for all energy-values less than W_i, zero for all values of energy greater than W_i.

This striking result can easily be deduced from Fermi's basic assumption, without any statistics at all. Absolute zero is by definition the temperature of the state, being in which the assemblage can give away no energy whatever. If not more than one electron may occupy any compartment of the phase-space, absolute zero is attained when there is an electron in every compartment from the origin outwards to a sphere which is centered at the origin, and which has just the volume needful to contain as many compartments as there are electrons. The number of electrons in unit volume of the phase-space is and remains equal to the number of compartments in unit volume, *i.e.* to the reciprocal of the volume of the elementary compartment, from the origin outward to this sphere; there it suddenly sinks to zero, and so continues. Cooling-down of an assemblage is settling-down of the particles into this the most condensed of all permissible arrangements; it is like crystallization upon a lattice, only the lattice is in the phase-space.

The foregoing statements may all be repeated, with the words *phase-space* replaced by *momentum-space*. In the momentum-space, a sphere of radius p, consequently of volume $4\pi p^3/3$, contains $(4\pi p^3/3)/(h^3/V)$ of the elementary compartments. If we set this number equal to the total number of electrons N, and solve the resulting equation for p, we get the radius of the sphere which would just contain all the

electrons if there were one in each compartment. If we set the volume of the sphere equal to Nh^3/VG and solve for p, we get the radius p_m of the sphere which would just contain all the electrons if there were G of them in each compartment. But this is the maximum value of the momentum of the electrons, it is the momentum of the fastest of the electrons. The corresponding speed v_m of the fastest of the electrons is p_m/m, therefore is given by the expression:

$$v_m = \frac{h}{m}\left(\frac{3n}{4\pi}\right)^{1/3} \tag{71a}$$

and the corresponding kinetic energy $\frac{1}{2}mv_m^2$ is the same as W_i.

It is expedient to set down for future reference the *mean values* of speed v and of several integer powers of v, for a gas distributed according to the Fermi law at the absolute zero. The general formula for the mean value of any power v^s of v is this:

$$\overline{v^s} = \frac{1}{n}\int v^s f(v)dv = \frac{1}{n}(4\pi G/m^3h^3)\int_0^{v_m} v^{s+2}dv$$

and in particular:

$$\overline{v^{-1}} = 3/2v_m; \qquad \bar{v} = 3v_m/4; \qquad \overline{v^2} = 3v_m^2/5; \qquad \overline{v^3} = \tfrac{1}{2}v_m^3. \tag{72a}$$

The corresponding values for the Maxwell distribution are these:

$$\begin{aligned} \overline{v^{-1}} &= 2(m/2\pi kT)^{1/2}; \qquad v^2 = (2kT/m); \\ \bar{v} &= 2(2kT/\pi m)^{1/2}; \qquad \overline{v^3} = 4\pi(2kT/\pi m)^{3/2}. \end{aligned} \tag{72b}$$

Plotted as functions of ϵ, the distribution-function f in the coordinates and momenta starts out as an horizontal straight line at a distance G/h^3 from the axis of abscissæ, while the distribution-function F in the energy starts out as a concave-upward parabolic arc; these continue as far as the abscissa $\epsilon = W_i$, and from then on the curves coincide with the axis of abscissæ.

The foregoing statements are valid for absolute zero; what happens as the temperature rises? Sommerfeld has proved that the sharp angles in the distribution-curve are *very gradually and slowly* rounded off, the curve always traversing the midpoint of the vertical arc BC (Fig. 1). The far end of the curve sinks down to the axis of abscissæ in the fashion of the Maxwell law. Even at room-temperature and even far above, however, the distribution departs so little from the absolute-zero form that many phenomena may be interpreted in a qualitative way, simply by imagining the absolute-zero distribution—the completely degenerate distribution, it is called—to persist all

through the observable range of temperatures. Indeed, in calculating electrical resistance and certain other properties of metals, one may use the mean values of the various powers of v which are tabulated in (72*a*). There are however other properties of metals, thermal conductivity for instance, for the estimation of which it is not sufficient to assume that the mean values of the powers of v are always the same as at absolute zero, and one must derive more nearly approximate values for them; for these however I refer the reader to Sommerfeld.

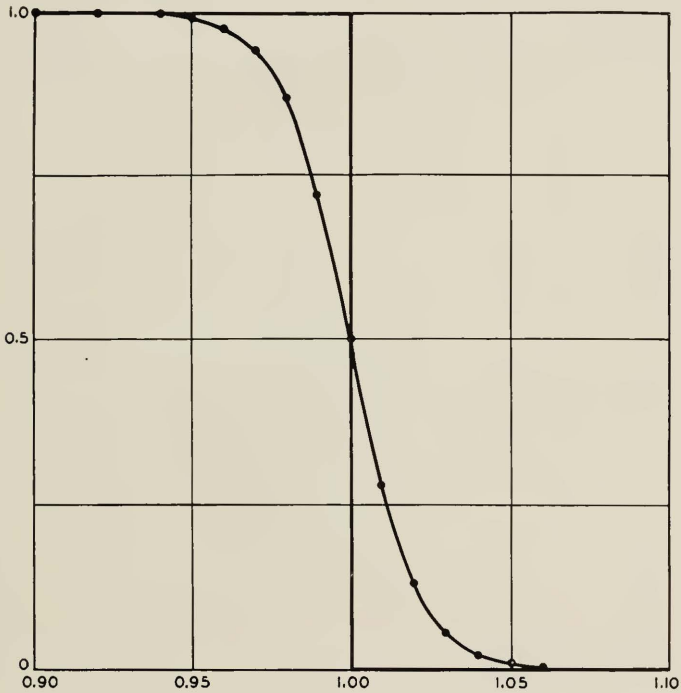


Fig. 1—Graphs of the Fermi distribution-function f plotted against ϵ/W_i as independent variable, for an electron-gas having $6.5 \cdot 10^{22}$ particles per cc., at temperatures zero (rectilinear curve) and 1500° K. (rounded curve). The value of W_i is 6 equivalent volts.

In practice, the values of W_i are rather astonishingly great; no less, for example, than 5.6 equivalent volts for silver, 5.7 for tungsten, 6.0 for platinum. Obviously they depend on the compactness of the lattice, being greater the more closely-packed the atoms are. In potassium and sodium the atoms are relatively widely spaced, and the corresponding values of W_i are about 2.1 and 3.2 in equivalent volts.

The contrast between this and the classical situation is evidently enormous. Where formerly we were asked to think of the electrons in a metal at usual temperatures as being distributed Maxwell-wise about a very modest mean energy, say about 0.02 of an equivalent volt, we are now invited to conceive them as distributed all through a range of energies extending from zero up to as much as half a dozen equivalent volts, and more abundantly the nearer one approaches to the top of this range, abruptly though the distribution ceases when the very top is reached. This is "zero-point energy" with a vengeance!

The pressure of the electron-gas is related to the energy-per-unit-volume by the equation valid also in the classical theory:

$$p = \frac{2}{3}\frac{E}{V},$$

and therefore varies like the total energy—starting from a value absurdly high at the absolute zero (hundreds of thousands of atmospheres) and increasing therefrom very slowly at first, though according to a T^2 law, as the temperature rises. I do not know of any manometer for measuring internal electron-pressures, but if anybody should invent one he had better make it strong.

There is manifest ground for doubting these remarkable proposals: thermionic data seem to show that the work-function which opposes the egress of the electrons from a metal is itself less than half-a-dozen volts (in the usual measure), for some metals less than two—what then keeps these fast electrons confined within the metal? It turns out, however, that in augmenting the *vis viva* of the electrons the new theory also raises the top of the wall which they must overleap. Here indeed we meet with the first of the new experiments which tend to confirm the new theory.

Thermionic Emission

The simplest theory of the thermionic current is, that it consists of all the electrons belonging to the interior electron-gas which fly against the boundary-surface of the metal with velocities such that the component u thereof perpendicular to the boundary-surface is great enough to make the "energy-component" $\frac{1}{2}mu^2$ greater than a constant W_a—the said constant being interpreted as the work-function or the retarding potential-drop at the edge of the metal. The thermionic electrons are those which swim up to the surface with an outward-bound velocity-component so large, that by means of the kinetic energy of their outward motion they can climb over the wall.

Evidently any thermionic emission must distort the distribution of the electron-gas inside the metal, as it is an unbalanced outflow of electrons. The situation in which the efflux is balanced by a corresponding influx from an electron-gas outside the metal is much regarded in thermodynamic theory, but one cannot measure currents in that situation any more than one can measure heat-flow between two bodies at equal temperature. In assuming, then, that the distribution of the internal electrons is that of Fermi or that of Maxwell, we shall probably be invalidating our conclusions except for the limiting case of an infinitesimal emission. It seems probable, however, that with the thermionic currents of practice the approximation is good enough.

The simplest theory of the thermionic current, then, consists entirely of the equation:

$$i = e\left(\frac{m}{h}\right)^3 G \int_{w=u_0}^{\infty} du \int_{-\infty}^{\infty} dv \int_{-\infty}^{\infty} dw \cdot u \frac{1}{A^{-1}e^{\epsilon/kT} + 1}. \qquad (73)$$

This is a restatement of the first sentence of this section, *plus* the assertion that even when electrons are leaking out through the wall of the metal the distribution within remains practically that of Fermi. The factor e stands for the electron-charge; the symbol i thus for the thermionic current-density in electrostatic units. The factor m^3 enters because, in conformity with usage, I have translated from the momenta into the velocity-components u, v, w as independent variables. The quantities u_0 and W_a are related by the equation:

$$W_a = \tfrac{1}{2}mu_0^2. \qquad (74)$$

In the integrand we are of course to put $\frac{1}{2}m(u^2 + v^2 + w^2)$ for ϵ.

Setting for A the first-approximation value from (107) with the symbol W_i defined in (71), we obtain:

$$i = e\left(\frac{m}{h}\right)^3 \int_{u_0}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u \frac{1}{e^{(\epsilon - W_i)/kT} + 1} dudvdw. \qquad (75)$$

The integration is perfectly straightforward if the second term in the denominator may be neglected relatively to the first. This seems unnatural, for we have just been noticing that over part of the energy-range, from $\epsilon = W_i$ downwards, the second term is larger than the first. But if W_a is considerably larger than W_i—and by this I mean, if $(W_a - W_i)/kT$ is positive and considerably larger than unity—then the electrons which escape are those which belong to the extreme

upper part of the energy-range, where the first term is much the larger. Writing then $u \cdot \exp\left[-(\epsilon - W_i)/kT\right]$ for the integrand, we integrate with ease by well-known formulæ, and get:

$$i = \frac{2\pi emG}{h^3}(kT)^2 e^{-(W_a - W_i)/kT}. \tag{76}$$

The experimental test consists in plotting $(\log i - 2 \log T)$ against $1/T$; we should get a straight line provided that W_a does not vary with temperature. The experiments do lead to precisely this result. The slope of the line varies from metal to metal, and depends on the state of the metal surface; identifying it with $(W_a - W_i)/k$, one finds that W_a exceeds W_i by amounts ranging from one equivalent volt upward to five or six; so the approximation just mentioned is abundantly justified, even up to temperatures of incandescence.

The contrast with the predictions of the classical theory is peculiarly interesting. Assuming the Maxwell distribution for the interior electrons, one arrives easily (the reader can do it by substituting the value of A from (64) into (73)) at the formula:

$$i = \frac{en}{(2\pi m)^{1/2}}(kT)^{1/2} e^{-W_a/kT}. \tag{77}$$

On testing this formula by plotting $(\log i - \frac{1}{2} \log T)$ against $1/T$, it is found that the experiments yield lines as beautifully straight as those obtained by plotting $(\log i - 2 \log T)$. Indeed—as everyone knows who has dabbled in thermionics—a function of the type $\exp(-c/T)$ varies so exceedingly rapidly with $1/T$ that it makes no perceptible difference to the graph whether or not the function is multiplied by a constant or by any modest power of T. One cannot then use the graphs to distinguish between the theories, even if one could be sure that W_a is not a function of T. But the classical theory proposes that we identify the slope of the aforesaid line with W_a/k; and it has been the custom so to do.

Now *if the Fermi distribution-function is the right one, physicists have been underestimating the work-function all along.* They have plotted experimental curves which agreed in shape with (76) and (77) and from these they have evaluated the constant figuring in the exponent, a constant which they have denoted usually by $-b/k$; and then they have equated b to W_a, whereas if it is right to apply the new statistics they should have added W_i to b and then equated the sum to W_a. Nor is the alteration slight: for if there are as many free electrons in the metal as there are atoms, then W_i is six volts

or thereabouts, and the quantity to be added to the observed constant b is larger than b itself.

Is there then any other way of determining the work-function than out of this apparently ambiguous current-vs-temperature curve? If there is a direct and independent way, it may serve not only to decide between the two statistics, but also—if it favors the new by yielding a value for W_a greater than the thermionic b—to give an *experimental* value for $(W_a - b) = W_i$ and hence for the disposable constant n which is the one uncertain quantity in the theoretical formula for W_i. It may, that is to say, serve to determine the number of electrons per unit volume of the electron-gas.

Now it seems that the diffraction of electrons by crystals provides an independent and direct way of determining the work-function. For in the phenomena in which negative electricity behaves as a wave-motion, the work-function figures in the index of refraction; and the index of refraction of a metal may be determined from the diffraction-patterns which it forms when irradiated with slow electrons. Ample data concerning one metal—nickel—have already been acquired by Davisson and Germer, and from these it transpires that the work-function is much in excess of the thermionic constant b—so much indeed, that the corresponding value of W_i implies that there are twice as many free electrons as atoms in the metal, or even more.[9] The values deduced for the work-function from the refractive index vary however with the speed of the electrons; and it is evident that much remains to be understood.

The factor which multiples $T^2 \exp\left[-(W_a - W_i)/kT\right]$ in the right-hand member of (76) involves universal constants only (supposing that G is such) and is therefore the same for all metals—a principle derived by Richardson from the first and second laws of thermodynamics twenty years ago, without any assumptions at all about the distribution of the electrons. Its actual value $2\pi k^2 meG/h^3$ differs only by the factor G from the value derived by Dushman, which is numerically equal—in the customary units—to 60.2 amperes per cm.^{2} per degree squared. There are several metals for which the experimental value of this quantity—commonly known as A, a symbol which in this article is monopolized by another meaning—agrees well with 60.2. One might infer that G must be unity, a choice which would demolish the theory of paramagnetism; but there is another recourse; one may suppose that half the electrons which

[9] L. Rosenfeld, E. E. Witmer (*l. c. infra*). From Rupp they cite values of refractive index for six other metals (Al, Cr, Cu, Ag, Au, Pb) and compute values of n. In considering these, however, the reader should assess G. P. Thomson's criticism of Rupp's values.

come up to the bounding surface from within with energy sufficient to escape are nevertheless reflected. A factor $(1 - r)$—r being called the coefficient of reflection, and being put equal to $\frac{1}{2}$—then enters into the formula, and balances out the factor 2 introduced by giving the preferred value to G. Moreover one may explain values of the constant still smaller than 60.2, or between 60.2 and 120.4, by adjusting r accordingly. But there are also recorded values enormously greater than 120; so evidently something remains to be understood.[10]

The methods of wave-mechanics have been applied by Fowler and Nordheim to the problem of evaluating this coefficient of reflection. They have attained some notable results in the fields of thermionics, cold discharge, and photoelectric effect. These however are consequences not of the new statistics only, but of a combination of the new statistics with the new way of considering the transmission and reflection of electron-waves at surfaces. There is not space for me to deal with the latter, beyond indicating its point of departure and its chief results.

Thus far I have been speaking of a metal as an equipotential region surrounded by a surface at which there is a sharp potential-drop, and beyond which there is the equipotential region of outer space. Fowler and Nordheim however, like Schottky and others before them, conceive a metal as an equipotential region surrounded by a surface, beyond which lies a region in which there is a field (or at all events an image-field) the strength of which is a function of the distance from the surface. The quantity W_a appears as the integral of this field-strength, from the surface to infinity. Electrons of a given kinetic energy being supposed to fall against the surface from within, the fraction which fails to pass completely through the region of the field depends upon the kinetic energy of the electrons and upon the *shape* (not solely upon the integral) of the field-strength-vs-distance curve. The average value of this fraction for all the electrons of all speeds coming up to the surface from within—the average being taken with due regard to the relative proportions of the electrons of various speeds, that is to the distribution-in-velocity—is the coefficient r aforesaid. For certain simple shapes of the field-vs-distance curve it may be calculated. One thus arrives at the be-

[10] If the empirical equation is of the form $i = aT^n \exp(-b/T)$, then whatever may be the values of the constants a and b, one can always claim that it agrees with the foregoing theory provided one assumes that W_i or r or both vary with temperature in just the proper way. But, as Fowler puts it: "the variety of possible uncontrollable hypotheses (if such assumptions are to be admitted) becomes too large for profitable discussion."

ginnings of a theory of the effects of surface-conditions on thermionic-emission, and of the cold discharge.

The Cold Discharge

Suppose that a potential-difference is now applied between the metal and a neighboring electrode, such that near the metal the resulting field-strength is very great. Does it penetrate the region which I have just been describing? Assume that it does penetrate as far as the "surface" just defined. Then, everywhere beyond the surface, the actual field is the resultant of this "applied" and the previously-mentioned "intrinsic" field. The shape of the field-strength-vs-distance curve is thus changed, in a way which is calculable if we have postulated some particular original shape; and in certain simple cases it is possible to calculate the consequent change of the coefficient r, and therefore the electron-current—or the additional electron-current—which the applied field causes to emerge from the metal. This is the current known as the "cold discharge."

Nordheim adopted for the field-strength-vs-distance curve, in the absence of applied P. D., the shape which is most commonly proposed —the inverse-square curve, the law of the "image-force" which a charge in the vicinity of a conductor experiences because of its "electrical image" in the conductor. For the current i of the cold discharge he then derived this approximate formula:

$$i = c'SF^2 \exp(-c''/F), \tag{78}$$

in which F stands for the applied field, S for the surface-area of the metal exposed to it, c' and c'' for constants which can be calculated when W_i is known.

There are certain experiments (for an account of which I refer to Nordheim's paper in the Physikalische Zeitschrift) which indicate that the actual relation between the current and the field-strength agrees in form with (78), but that the predicted values of c' and c'' are too large—too large by a factor of the order of ten in the latter case, by several orders of magnitude in the former. However it is possible to explain away these contradictions. One may assume, for instance, that the actual surface concerned in the discharge is a collection of small spots which altogether have but a small fraction of the total area of the metal surface; and that over these small spots, the field-strength is much higher than it would be if the metal were everywhere uniform and smooth. The ratio of the observed to the predicted value of c' then gives the fraction of the total surface which is covered by these "active" spots, and the ratio of the observed to

the predicted value of c'' gives the reciprocal of the factor by which the field-strength must be multiplied.

This explanation has the disadvantage of being not only plausible but much too easy. So long as there is not any independent evidence about the area of the effective spots or the field-strength prevailing over them, the theory simply delivers an equation with two disposable constants, which is not very valuable for testing the underlying assumptions. It appears from the data examined by Nordheim, however, that the ratio of the predicted to the observed value of c'' is always between 10 and 20, and the ratio of the observed to the predicted value of c' is always about 10^{-10} (at least for tungsten). This uniformity of the two quantities which figure in the theory as the disposable constants may be taken as a confirmation of some weight.

Photoelectric Effect

According to the former theory, the elementary process of the photoelectric effect runs thus: a quantum dives into a metal, and gives its whole energy (say E_0) to an electron initially at rest, which then may escape from the metal after suffering a reduction of kinetic energy equal at least to W_a and possibly more (more, that is to say, if the electron loses kinetic energy on its way to the surface). Even if we suppose that the electron originally belonged to an electron-gas conforming to the classical statistics, its initial energy would almost always be quite negligible compared to that which the quantum gives it. But if the electron-gas obeys the new statistics, it comprises electrons with energy-values ranging up to W_i. However if the new statistics is valid, then the reduction of kinetic energy at the boundary is also greater by W_i than we have hitherto supposed. The net result is, that by the new statistics as by the old we derive Einstein's equation for the maximum kinetic energy of the electrons expelled by quanta of frequency ν:

$$E_{\max} = E_0 + W_i - W_a = h\nu + \text{const.} \tag{79}$$

only the additive constant is now $(W_i - W_a)$ instead of $(-W_a)$. This additive constant should as before agree with the thermionic constant b.

The new theory has one marked distinction, probably an advantage, over the old: it implies a sharply definite maximum kinetic energy—that is to say, the distribution-in-energy function of the escaping electrons should jump suddenly from zero to some definitely higher value at the energy-value $E = E_{\max}$ prescribed by (79); the slope of

the curve representing this function should make an acute angle with the axis of E where they intersect at E_{max}. The Maxwellian distribution predicted by the classical statistics for the interior electrons suggests however that the curve in question should approach the axis asymptotically. The new statistics leads also to certain inferences about the shape of the curve for values of E less than E_{max}. A great quantity of data bearing on this subject has been obtained by Ives and his collaborators; but the interpretation is made difficult by the presumption that some allowance must be made for the energy-losses suffered by electrons after they absorb quanta but before they reach the surface, and will require much study.

Paramagnetism of the Electron-Gas

The susceptibility of the electron-gas was calculated by Pauli even before the specific heat was evaluated by Sommerfeld, but as it involves an extra complication I have inverted the historical order.

The complication is due of course to that assumption which is made in order to explain why the electron-gas should be magnetic at all—the assumption that electrons are magnets. Perhaps I am too cautious in referring to it as an assumption, it being so well authenticated by the gyromagnetic effect and by the general usefulness of the "spinning electron" in the explanation of spectra. These phenomena impose a special value on the magnetic moment μ_0 of the electron, to wit, the value,

$$\mu_0 = eh/8\pi m_0 c, \tag{80}$$

m_0 standing for the rest-mass of the electron. Further they require that when the electron is floating in a magnetic field, its moment (considered as a vector) shall be either parallel or anti-parallel to the field. Denote by θ the angle between the moment of the electron and the magnetic field: then θ must be either 0 or π.[11] Now when a magnet of moment M is inclined at an angle θ to a magnetic field H, its "extra magnetic energy" is $-MH\cos\theta$.[12] In dealing with the electron-gas, then, we are in effect assuming that when a field H is applied to it the energy of every electron is either increased or decreased by the amount:

$$\Delta = ehH/8\pi m_0 c. \tag{81}$$

[11] It comes to the same thing, and may on other grounds be preferable, to assign μ_0 twice the value given in (80), and to θ the values 60° and 120°.

[12] The magnetic energy, or energy due to the "interaction between the magnet and the field" is put equal to zero when the magnet is transverse to the field, which is consistent with the picture that the field alters the energy of the magnet by speeding up or slowing down the revolving electricity. The formula here given is a first approximation.

As I mentioned earlier, the idea of compartments in the momentum space may be replaced by the idea of "permitted energy-values," at least in some situations. Let us for convenience return to the latter conception. Then we may say that when a magnetic field is applied to an electron-gas of which the particles are magnets, each of the permitted energy-values is split into two. To any previously-permitted value ϵ correspond a pair of new ones, $\epsilon + \Delta$ and $\epsilon - \Delta$; or let me say $\epsilon + m\Delta$, using m as a symbol which may have only the values $+1$ and -1.

Pauli assumed that the most probable distribution of the electrons among this doubled set of energy-values is to be determined by the new statistics, including Fermi's postulate so modified as to state that not more than one electron may possess any one of the permitted values.

Previously I used the symbol Z_{is} to denote the number of compartments or permitted energy-values which lie in the shell s and are occupied by i electrons apiece; and the symbol Q_s to denote the total number of permitted energy-values in the shell s. Now there are Q_s permitted values which are shifted upward by Δ from the original ones, and Q_s more which are shifted downward by Δ from the original ones. Let Z_{is+1} stand for the number in this upward-shifted group which are occupied by i electrons apiece, and Z_{is-1} for the corresponding number in the downward-shifted groups; Z_{ism} shall be the general symbol for the two.

Now consider the distribution:

$$Z_{ism} = \alpha_{sm} e^{-iB - i(\epsilon_s + m\Delta)/kT}, \qquad \begin{matrix} m = +1, -1, \\ i = \quad 0, \quad 1. \end{matrix} \tag{82}$$

This answers the standard requirements for thermal equilibrium. For if we define W^* in Bose's way, and then say that the entropy is $k \log W^*$, we find that when the energy of the assemblage is varied by δE—the total number of cells and the total number of particles remaining constant—the first variation of the entropy is $\delta E/T$, as it should be. I leave it to the reader to prove this statement as corresponding statements for other distributions (*e.g.* (47)) were earlier proved, and to determine the values of the quantities α_{sm}; on doing which, and substituting the results into this distribution, he should obtain

$$Z_{1s,+1} = \frac{Q_s}{e^{B+(\epsilon_s+\Delta)/kT} + 1}, \qquad Z_{1s,-1} = \frac{Q_s}{e^{B+(\epsilon_s-\Delta)/kT} + 1}, \qquad Q_s = \frac{2}{(\pi)^{1/2}} \frac{V}{h^3} (2\pi m)^{3/2} (\epsilon_s)^{1/2}. \tag{83}$$

It is best, perhaps, to regard these formulæ as the descriptions of two electron-gases, one composed entirely of magnets parallel to the field, the other of anti-parallel magnets; the actual electron-gas is a mixture of the two. The quantity $Z_{1s,\,-1}$, for example, is the number of parallel magnets having energy-values shifted downwards through Δ from the Q_s originally-permitted energy-values which lay in the shell s. True, these magnets are no longer themselves in the shell s, owing to the shift; their energies lie between $(\epsilon_s - \Delta)$ and $(\epsilon_s - \Delta + d\epsilon_s)$, not between ϵ_s and $\epsilon_s + d\epsilon_s$; but for ease of integration it is better to think of them as being associated with the original unshifted energy-values. The total number of magnets comprised in the parallel gas is then given by the equation:

$$N_1 = \frac{2}{(\pi)^{1/2}} \frac{V}{h^3} (2\pi m)^{3/2} \int_0^\infty \frac{(\epsilon)^{1/2} d\epsilon}{e^{B+(\epsilon-\Delta)/kT} + 1} \cdot \tag{84}$$

For brevity denote by L the constant before the integral; and use the symbol ϕ for the function defined as follows:

$$\phi(u) = \int_{\epsilon=0}^\infty \frac{(\epsilon)^{1/2} d\epsilon}{e^{\mu+\epsilon/kT} + 1} \cdot \tag{85}$$

Then expanding N_1 as a power-series in the variable Δ/kT which is $\mu_0 H/kT$, we find:

$$N_1 = L \left[\phi(B) - \frac{\mu_0 H}{kT} \phi'(B) + \text{terms of higher order in } H \right] \cdot \tag{86}$$

Similarly one obtains, for the total number N_2 of electron-magnets in the "anti-parallel gas," a formula:

$$N_2 = L \left[\phi(B) + \frac{\mu_0 H}{kT} \phi'(B) + \text{terms of higher order in } H \right] \cdot \tag{87}$$

Now the total magnetic moment of the "parallel gas" is $N_1\mu_0$ in the same sense as the field, and the total magnetic moment of the "anti-parallel gas" is $N_2\mu_0$ in the sense opposed to the field; so that the net magnetic moment of the entire assemblage of electrons is $(N_1 - N_2)\mu_0$. We will carry out the computations only for values of H so low that we may ignore all terms beyond the second in the expansions for N_1 and N_2. The net magnetic moment is then approximately proportional to H; its quotient by H, the *susceptibility* χ of the electron-gas, is constant. It is a fact of experience that with nearly all paramagnetic substances the susceptibility *is* independent of H up to the highest attainable values of this variable. The limi-

tation which we are here accepting will probably therefore not prove serious. Our approximative formula for χ is then as follows:

$$\chi = (N_1 - N_2)\mu_0/H = -\frac{2L\mu_0^2}{kT}\phi'(B) = -\frac{N\mu_0^2}{kT}\frac{\phi'(B)}{\phi(B)}, \quad (88)$$

and all that remains is to make the step made in every previous case —to determine the last remaining unspecified constant, B, in terms of the total number N of the particles of the assemblage.

This number N is the sum of N_1 and N_2. Ignoring the terms of higher order in H, we have:

$$N = 2L\phi(B) \quad (89)$$

and this is substantially the equation which was used to determine Sommerfeld's constant A in terms of N; for e^B and $1/A$ are one and the same. To make this equation identical with (63), or rather to make (63) identical with this one, we must there put $G = 2$, as we did—this is the reason for having introduced that factor G.

The procedure is then as follows: put $-B$ for $\log A$ in the right-hand member of equation (66)—differentiate it with respect to B—insert into the derivative the value of B obtained by equating to N the right-hand member of (66), *i.e.*, the value given in (67)—and substitute into (88). The resulting value for χ is this:

$$\chi = 12\left(\frac{\pi}{3}\right)^{2/3}\mu_0^2 n^{1/3} m_0 h^{-2}. \quad (90)$$

To pass now to the experiments: is it permissible to suppose that the susceptibility of any metal is due entirely to the electron-gas within it? This is the same sort of uncertainty as confuses the question of the specific heat. Here we have every reason to expect that the magnetization of an ordinary paramagnetic metal is a threefold effect, involving not only the orientation of the electrons but also the orientation of the atoms, and finally that alteration of the electron-orbits in the atoms which gives rise to diamagnetism. To disentangle these three contributions to the net magnetic moment seems almost beyond the powers of any theory. With the alkali metals, however there is strong evidence that the second may be absent. Spectroscopic data show quite definitely that the magnetic moment of the alkali-metal ion—the atom minus its valence electron—is zero. If every atom in an alkali metal has surrendered its valence electron to the electron-gas, then there will be no orientation of the ions by the magnetic field, and the number of electrons forming the electron-gas will be

equal to the number of atoms. The values of χ computed with this last assumption should then be not less than the actual susceptibility; they may be somewhat greater, because of the diamagnetic effect which is opposed in sign to the paramagnetic effect and therefore neutralizes it in part. Of this diamagnetic effect we can predict the order of magnitude, and we may expect that it will be greater, the higher the atomic number of the metal.

The value of χ proposed above is the value for absolute zero; for higher temperatures a closer approximation can be obtained by using two terms of the expansion in (66), instead of the first term only. It appears, however, that the alteration is slight. Like the average energy and the pressure, the susceptibility of the electron-gas should be very nearly the same at all temperatures from absolute zero up through room-temperature and far beyond. Now it is a fact that the susceptibility of the alkali-metals is independent of temperature—a fact so surprising, that the desire to explain it seems to have been Pauli's principal incentive in undertaking this research. For if the electron-gas were governed by the classical statistics, and the electrons were as many as the atoms, the susceptibility of a metal would increase as the temperature diminished and attain enormous values near the absolute zero.

When Pauli published the theory to which this section is devoted, the experimental data indicated that the susceptibilities of potassium and sodium were somewhat lower, those of rubidium and cæsium markedly lower than the predicted values—divergences which might be charged to the diamagnetic effect or to faults in the theory. Recent Canadian work, coming out very much *a propos*, has improved the situation remarkably. This tabulation (taken from E. S. Bieler, to whose article I refer for the sources) shows the comparison:

	Na	K	Rb	Cs
Theoretical (Pauli)	0.66	0.52	0.49	0.45
Experimental:				
McLennan et al.	0.61	0.42	0.31	0.42
Lane	0.65	0.54		

(All numerical values to be multiplied by 10^{-6})

Singularly enough, the agreements are too good! one would expect the diamagnetic effect to be more considerable than the very slight discrepancies between the experimental and the theoretical values for sodium, potassium, and cæsium. Perhaps further work on the theory of the diamagnetic effect would now be desirable.

Returning once more to the meaning of G, one sees that the placing of the value 2 for G in equation (61) and all of its descendants amounts to the making of the assumption that the electron-gas is really a mixture of two equally numerous and entirely independent assemblages of particles, each for itself obeying the Fermi statistics. This seems a rather odd idea, but inevitable.

Theory of Conduction

The new theory of conduction developed by Houston and Bloch is based upon the wave-theory of negative electricity, in which the interior of a metal is conceived to be filled not with darting corpuscles, but with stationary waves—as many distinct patterns of loops and nodes, it may be, as in the corpuscle-picture there are free electrons. It is not a consequence of the Fermi statistics alone, but of the Fermi statistics plus the wave-theory. Of course, if we come to decide that the Fermi statistics implies the wave-theory and *vice versa*, this warning will seem superfluous; but it is not superfluous, so long as the new statistics is used with reference to corpuscles. Now the corpuscle-picture of negative electricity is not only familiar, but seems likely to survive as the most convenient for describing most of the phenomena in which electrons figure. I will therefore express as much as possible of the new theory of conduction in the language of corpuscles, although eventually I shall be forced to make an assumption which will come to the same result as converting the corpuscles into waves.

To realize the things to be explained, conceive a slab of metal, having a thickness d measured along the x-axis; suppose a potential-difference V to exist between its faces, so that a field $E = V/d$ directed along the axis of x pervades it.

If the electrons in the metal moved perfectly freely, then any which was introduced without kinetic energy at the negative side of the slab would fall forthwith to the positive side, arriving there with the full kinetic energy eV and the full corresponding velocity of magnitude $(2eV/m)^{1/2}$ directed along the axis of x. Certainly nothing of the sort occurs. When a potential-gradient exists along a wire, for instance, heat is developed uniformly everywhere and there is nothing to suggest that the electrons are moving more rapidly at the positive than at the negative end.

We must then suppose that the free flight of the electron is interrupted at frequent intervals, and that at every interruption it loses the kinetic energy and the component of velocity up the potential gradient which it has acquired from the field since the last one previous. Or at least, the average loss of kinetic energy and of "drift-speed"

at interruptions must be balanced by the average gain between interruptions.

In the corpuscle-theory these interruptions are pictured as actual impacts or collisions of the electrons with the atoms. Evidently, if we could assume that whenever an electron hits an atom it rebounds in some direction perfectly transverse to the field, then we should have a mechanism in which the drift-speed of the electron up the potential-gradient is annulled at every impact. This would be much too artificial. But if we think of both the electrons and the atoms as elastic spheres, the latter being so massive that they never budge when struck, the result is in effect the same. For then, the angle between the direction along which an electron approaches an atom and the direction along which it flies away after collision is *on the average* 90°. The rebound is as likely to be backward as forward; the rebounding sphere retains on the average no memory of its former direction of flight. This I will prove later.

There is a difficulty, which I must not leave unmentioned, although in this place I can do nothing to clear it away. In the development of these ideas we shall in effect assume that at the end of each free path the electron loses not only the forward drift-speed but the whole of the kinetic energy which it acquired while traversing that free path under the influence of the field. But if it collides with infinitely massive spheres it does not lose kinetic energy at all. If it collides with spheres of the mass of an atom, it loses kinetic energy, but does not completely lose its drift-speed. The theory of this latter case has been developed by Compton and Hertz for use in the study of conduction in gases, and might be applied to the problem presented by metals, but probably fits them no better than does the other hypothesis.

With this elastic-sphere model, then, the average interval between impacts is the average interval during which the electron is piling up drift-speed, only to lose it all at the end of the interval and be forced to start afresh from scratch. Denote by t_0 the length of this average interval. Since the acceleration of the electron is eE/m, its drift-speed at the end of the period t_0 is (eEt_0/m), its average drift-speed is half as great. Now I must dispel the impression that the drift-speed is the whole of the speed which the electron has. On the contrary, the mean speed of the thermal agitation—let me call it $\bar{v}$—is immensely larger than the small contribution which any ordinary field (indeed, any not very extraordinary field) can impart to an electron over a distance comparable with the distance between atoms. The field must not be supposed to do more than bend very slightly the rectilinear

paths of the electrons from impact to impact. This statement is true with the classical statistics, *a fortiori* with the new. Denote by l the average distance traversed by an electron between impacts. Then t_0 is $l/\bar{v}$, and the average drift-speed is $\frac{1}{2}(eEl/m\bar{v})$.[13] The corresponding current-density is the product of this by the numer of electrons in unit volume multiplied by the charge of each. So, for the current-density produced by unit-field-strength, which is by definition the *conductivity* σ, we obtain the formula:

$$\sigma = \tfrac{1}{2}ne^2l/m\bar{v}. \tag{91}$$

The constant l, the *mean free path*, is the third disposable constant of the theory of electrons in metals.

I fear that the foregoing passage sounds very old-fashioned; but nevertheless it expresses the corpuscle-theory of conduction. The notion of elastic spheres is only accessory—an image which may or may not be the best to represent the central idea, the idea that the life-history of a corpuscle in a metal pervaded by a field is an alternation of gradual gain and sudden loss. The mean free path is the average distance of uninterrupted gain.

The common test of the formula (91) is the test by the temperature-variation. The result of this, incidentally, was regarded as quite as grave a demerit of the old electron-gas theory as the difficulty with the specific heat.

It is a fact of experience that the resistivity $\rho = 1/\sigma$ of any metal varies rapidly with temperature. For many metals it varies directly as T over quite a wide range; at low temperatures even more swiftly, not to speak of the strange phenomenon of supraconductivity. Now in equation (151) we have ρ set equal to a combination of two universal constants with three quantities $\bar{v}$, n, l between which last the responsibility for these great variations must be divided.

According to the classical statistics $\bar{v}$ is proportional to $T^{1/2}$. This is a variation in the right sense, but not fast enough. To make ρ vary as T we must then make nl vary as $1/T^{1/2}$. With the Fermi statistics the requirement is harder. The mean speed of thermal agitation is almost independent of temperature, and the burden of the whole responsibility for making ρ proportional to T must be loaded upon nl. The first step with the new statistics is a step backward.

Can we reasonably assume n to be the cause of the variation?

[13] It is the mean of the reciprocal of the speed, not the reciprocal of the mean speed, which should figure here; but with so rough a formula the distinction is scarcely worth making.

If so, then it must diminish with rising temperature. It would seem reasonable enough for n to *increase* with rise of temperature, for presumably the free electrons arise from ionization of the atoms, and ionization is promoted by heat; but for n to *decrease* would seem very odd, notwithstanding Waterman's successes in accounting for some of the data by such a theory.

Part of the burden, then, must be cast on the mean free path—indeed the whole of it, if we adopt the new statistics so that $\bar{v}$ is held constant, and suppose in addition that n does not vary with temperature. But the elastic-sphere conception cannot stand the strain. It gives for the mean free path a value independent of temperature,[14] except insofar as the metal dilates with increase of heat. This is pretty nearly checkmate.

If however we might suppose that an electron may sometimes go clear through an atom without being reflected or deflected, and that the chance of such a piercing is relatively smaller and the chance of a rebounding relatively greater, the more violently the atom is vibrating—then by this theory the mean free path would diminish as the temperature rises, which is what is desired. This is an idea proposed long since by Wien.

The new idea is in result the same. The probability of the rebounding, or let me say of the *scattering* of an electron by an atom, is supposed to increase with the vigor of the vibrations of the latter. But for this a new reason is advanced: the reason, that while vibrating the atom is most of the time away from its equilibrium-place in the crystal lattice, and its relations with its companions are distorted. The probability of scattering is made to depend not only on the presence of the atom somewhere along the path down which the electron is rushing, but also on the relative positions of all the other atoms in the crystal.

To make such an assumption is, in effect, to compromise between the corpuscle-theory and the wave-theory. For what is the evidence from which it is inferred that a beam of light falling upon a grating, say, or a beam of X-rays falling upon a crystal, are undulatory? Essentially this: the way in which the beam is scattered or diffracted by the regular array of rulings on the grating or atom-groups in the crystal is different—greatly and strikingly different—from the way in which we know that it would be scattered by a single ruling, or suspect with good reason that it would be scattered by a single atom-

[14] The value $1/N\pi R^2$ familiar in the kinetic theory of gases, N standing for the number of fixed spheres per unit volume, R for the sum of the radii of a fixed and a moving sphere.

group. For instance, there are directions in which no light at all is sent by the regular array, though assuredly light would be scattered in those directions by any member of the array if it were solitary. These facts are explained by invoking interference of waves. The wavelets expanding outwards from the various rulings or scattering particles are supposed to arrive in opposite phases at the "dark fringes" of the diffraction-pattern, so that they cancel each other. But one might also say that the beam of light is a stream of corpuscles which are deflected or scattered by the atom-groups or rulings which they happen to strike, and that the law of scattering of the individual atom-group is altered by the marshalling of the scattering elements into a regular pattern, so that in particular the probability of a deflection towards one of the dark fringes is reduced to zero.

I am not prepared to say that such a compromise is a full alternative for the wave-theory, though modern theoretical physics seems to be tending in that direction. But if we wish to describe with the language of the corpuscle-theory the phenomena of diffraction by a crystal, whether of waves of light or waves of negative electricity: then we must certainly adopt the idea of a probability-of-scattering, of a mean-free-path, which varies with the irregularity of the placing of the atoms.

The principle is especially simple and especially startling, if we deal with a beam of which the wave-length—considering it as a beam of waves—exceeds the spacing of the lattice. Waves of such a magnitude would not be diffracted at all by scattering particles placed exactly at the points of the lattice. Though any particle singly would scatter them, they flow through the lattice intact. If then we wish to interpret the beam as a stream of corpuscles, the probability of deflection of any corpuscle by any atom must sink to zero when the arrangement is made perfect; the mean free path must then be considered infinite.

The resistance of a perfect crystal of an element should then be zero when all the atoms are stationary in their places on the lattice—if they ever are, which apparently they are not; and should increase steadily with increasing temperature, in a way which can be computed if we know two things: the way in which the scattering of waves by particles on a lattice varies with the amplitude of the quiverings of the particles about their lattice-points, and the way in which the amplitudes of the particles vary with the temperature. The second of these questions is the subject of the theory of specific heats of solids, developed principally by Debye. The first has been profoundly studied by Debye and by several other physicists interested chiefly in the scattering of X-rays by crystals. Transferring their results

into the theory of the diffraction of electron-waves, Houston demonstrated that over a wide range of temperatures the resistance of a perfect crystal should vary as the absolute temperature.

To determine not only the law of variation of resistance with temperature but the actual value of the resistivity for any metal, it would be necessary to evaluate the mean free path of the electrons. By the thoroughgoing corpuscular theory, this depends on the size of the atoms from which the corpuscles rebound; by the wave-theory, it depends on the scattering-power of the individual atom, which thus takes the place of the "size of the atom." The problem of computing the scattering-power of an atom for electron-waves belongs to the new mechanics. Houston was able to obtain good numerical agreements for several metals.

Another way of introducing irregularity into a crystal of an element consists in replacing a small fraction of the atoms, chosen at random here and there on the lattice, by atoms of another element. Certain alloys, known as "solid solutions," are of this type; and it is not only known that the resistance of such an alloy is greater than that of the element which is most abundant in it, but it has been shown by Nordheim that the dependence of resistance on percentage of substituted atoms follows the rule to be expected from the diffraction theory of resistance.[15]

Since then the conception of mean-free-path can be reinterpreted in terms of the wave-theory, and since it appears to be possible to deduce from the wave-theory a law of variation of mean-free-path with temperature which can be incorporated intact into the corpuscle-theory it is permissible to return to the corpuscle-picture to set up a theory of conduction of heat and of electricity, and of the thermo-electric effects in crystals.

We shall apply what I may call the method of the *perturbed distribution-function,* developed by Lorentz. The idea is, to begin by deriving a distribution suiting the actual case. The functions which we have hitherto employed, that of Maxwell and that of Fermi, are isotropic; it is only in the combination $(\xi^2 + \eta^2 + \zeta^2)$, hereafter to be called v^2, that the velocity-components ξ, η, ζ appear in them.[16] These "standard" functions may be appropriate to a uniform metal in which the temperature and the potential are uniform. Evidently they are not appropriate to a metal in which there is an electric field,

[15] The idea that the free paths of electrons extend from one irregularity of the crystal to another was propounded before the advent of the wave-theory of negative electricity.

[16] I shall use the velocity-components hereafter in lieu of the momentum components, to conform with the custom.

or a temperature-gradient, or which varies in its chemical nature from place to place, as might an alloy. If in such a case we orient the axis of x parallel to the gradient of the variable quantity—be it electric potential, temperature, or whatever else—we must expect ξ to enter differently into the distribution-function from η and ξ.

Various arguments show that as a rule the departures from the standard function must be rather small. Lorentz therefore postulated that in the presence of a gradient directed parallel to the x-axis, the actual distribution should differ from the standard function $f_0\,(v)$ —this he of course assumed to be Maxwell's—by virtue of a small additive term, a new function of v multiplied by the velocity-component ξ:

$$f = f_0(v) + \xi g(v), \tag{92}$$

and he proceeded to determine the new function g by the condition that f should remain constant in time despite the collisions of the electrons with the atoms. More precisely, he found for each of the three cases with which we shall be concerned a function g, such that the distribution-function obtained by adding ξg to f_0 conforms to that condition. This justifies the procedure.

Much the simplest case of the three is that of a uniform metal at a uniform temperature, subject to an electric field; for here the distribution-function need not vary from place to place. It will be well to go through the reasoning in this instance, though the formula for conductivity in which it leads differs but little from (91).

It is required, to find a function g of the combination $(\xi^2 + \eta^2 + \zeta^2)^{1/2}$ such that if at any moment the distribution $(f_0 + \xi g)$ prevails, it continues unchanged throughout time—the number of particles in any compartment or cell of the velocity-space (the momentum-space of the earlier pages, with its unit of length altered in the ratio $m : 1$) stays constant. Choose a compartment enclosed between planes ξ and $\xi + d\xi$, η and $\eta + d\eta$, ζ and $\zeta + d\zeta$. Call it the compartment C. Its volume is $d\xi d\eta d\zeta$, which to save a profusion of Greek letters I will usually denote by $d\tau$. The number of particles in it is $f \cdot d\tau$. This number must remain unchanged, though individual particles are constantly moving into and out of C in either of two ways—by "drift" and by "collision."

Owing to the field E, all of the particles have a steady acceleration $\alpha = eE/m$, because of which they are continually and continuously drifting from cell to cell. One easily sees that the number which thus drift out of C per unit time (it is best to think of "unit time" not as one second, but as a period small compared with the mean

time between impacts) is equal to first approximation to $\alpha f(\xi + d\xi, \eta, \zeta)d\eta d\zeta$.[17] This loss is partly balanced by an inward drift of particles which are accelerated into the compartment from the one lying beyond the plane ξ; the balance is not perfect, for the number drifting in per unit time is equal to $\alpha f(\xi, \eta, \zeta)d\eta d\xi$ and there is a difference $\alpha(df/d\xi)d\tau$ outstanding.[18] This difference must be balanced by the entrances and exits of particles which undergo collisions.

Let adt represent the number of electrons which, being initially in the compartment C, suffer impacts during unit time and are thus suddenly bumped out of it; and bdt the number which, being initially in other compartments, suffer such impacts during unit time that they suddenly turn up in C. The function g must be so chosen, that the lack of balance between the electrons drifting out and the electrons drifting in is just compensated by the lack of balance between those bumped into the compartment and those which are bumped out:

$$\frac{eE}{m}\frac{df}{d\xi} = b - a. \tag{93}$$

We must therefore evaluate $(b - a)$ in terms of the distribution-function.

We already have a formula ready-made for the number of impacts experienced per unit time by the particles of speed v; it is v/l for each particle, so that:

$$a = (v/l)fd\xi d\eta d\zeta. \tag{94}$$

Since however we have also to compute b, we shall find it expedient to classify these impacts according to the destinations of the particles, so to speak—according to the compartments of velocity-space into which they are bounced. A particle of speed v is located, in the velocity-space, on a sphere of radius v centered at the origin. Collision with an immovable atom changes the direction, but not the magnitude of the velocity; in the velocity-space, the particle suddenly moves to some other point on the same sphere. When electrons jump out of the compartment C because of impacts, they land in the other compartments which with C form a spherical layer around the origin. When electrons jump into C because of impacts, they come from the other compartments of that same layer. We shall derive an expression

[17] During a time dt so short that αdt is small by comparison with $d\xi$, the particles which initially lie between the plane $(\xi + d\xi - \alpha dt)$ and the side $(\xi + d\xi)$ of the compartment move out of it, while the particles which initially lie between the side ξ of the compartment and the plane $(\xi - \alpha dt)$ move into it; these two classes of particles number $f(\xi + d\xi, \eta, \zeta)\alpha dt \cdot d\eta d\zeta$ and $f(\xi, \eta, \zeta)\alpha dt \cdot d\eta d\zeta$ respectively.

[18] This expression figures in the equations as a net loss, but in fact has a negative sign (since $df/d\xi < 0$) and therefore is actually a gain.

for the number of particles leaping from C into any other cell C' of the layer, and an expression for the number leaping reversely. The difference or lack of balance between these numbers, integrated over all the cells C', will be the required quantity $(b - a)$.

We begin by inquiring how many particles make such impacts that their paths (in the coordinate-space, of course—not the velocity-space) are deflected through angles between say θ and $\theta + d\theta$. To be deflected through an angle θ, an electron must strike an atom at a point where its surface is so oriented, that the normal (which is the line of centres at the instant of collision) is inclined to the line of approach of the electron at the angle $\psi = \frac{1}{2}(\pi - \theta)$. Denote by R the radius of the atom, and suppose that the radius of the electron is negligibly small.[19] Think of all the $f \cdot d\xi d\eta d\zeta$ electrons which at some particular moment of time are in unit volume of the metal, and belong to the compartment C of the velocity-space. Imagine each of these to be the centre, in the coordinate-space, of a pair of circles lying in the plane perpendicular to its path, and having radii $R \sin \psi$ and $R(\sin \psi + d \sin \psi) = R(\sin \psi + \cos \psi d\psi)$. As time goes on, let these circles travel in the direction normal to their plane with the speed v. During unit time each pair of circles traces out a pair of cylinders of length v, containing between them a cylindrical sheath of volume $v \cdot 2\pi R^2 \sin \psi \cos \psi d\psi$. Multiplying this by the number $f \cdot d\xi d\eta d\zeta$ of the electrons, we get the total volume included in all of these sheaths. Multiplying this by the number N of atoms in unit volume, we get the number of atoms located with their centres in these sheaths—which is the number of atoms so placed that in unit time, electrons of the stated cell impinge on them at angles between ψ and $\psi + d\psi$—which is the number of impacts per unit time in which electrons are deflected through angles between θ and $\theta + d\theta$, which accordingly is this:

$$N \cdot f d\xi d\eta d\zeta \cdot 2\pi v R^2 \sin \psi \cos \psi d\psi = f d\xi d\eta d\zeta \cdot N \pi R^2 v \cdot \tfrac{1}{2} \sin \theta d\theta. \quad (95)$$

It will be convenient to express this as a fraction of the total number of impacts—call it $Z d\tau$—experienced per unit time by all the electrons in question, which by integrating (95) is found to be:

$$Z d\tau = \pi N R^2 v \cdot f \cdot d\tau, \quad (96)$$

substituting which into (95) we get:

$$Z \cdot 2 \sin \psi \cos \psi d\psi = Z \cdot \tfrac{1}{2} \sin \theta d\theta. \quad (97)$$

[19] The formulæ remain valid even if the diameter of the electron is supposed not negligibly small, provided that we interpret R as the sum of the radii of atom and electron; but the generalization is not, so far as I know, of any practical value.

It will be observed that deflections smaller than 90° are equally numerous with deflections greater than 90°, so that on the average the electrons after impact have no reminiscence of their prior direction of motion, as I mentioned earlier. Also, comparing (96) with (94), one derives the expression for mean free path,

$$l = 1/N\pi R^2, \tag{98}$$

cited already in a footnote. To appreciate the most important feature of the expression (97) we must however return to the velocity-space.

In the velocity-space, the electrons of which the paths in coordinate-space are deflected through angles between 2θ and $2\theta + d2\theta$ execute leaps from the compartment C into other compartments of the spherical layer aforesaid, located on a certain region thereof. These occupy a belt or collar on the sphere, intercepted between two cones drawn with their common apex at the centre, their common axis pointing towards C and their apical semi-angles equal to θ and $\theta + d\theta$ respectively. Now the area of this belt is itself proportional to $\sin \theta d\theta$. This is very important: for it means that the electrons which are bounced out of C by collisions are sprinkled uniformly over all the rest of the sphere. More yet: it means that the electrons which are bounced out of *any* cell of the spherical layer are sprinkled uniformly through all the rest of the layer.

Consider then the interchange of electrons between two cells of the layer, say C at (ξ, η, ζ) with volume $d\tau$, and C' at (ξ', η', ζ') with volume $d\tau'$. The number leaping from C to C' is equal to the total number of impacts occurring in C multiplied by the ratio which the volume of C' bears to the volume V of the layer. The number leaping from C' to C is equal to the total number of impacts occurring in C', multiplied by the ratio which the volume of C bears to the volume of the layer. The excess of the latter over the former is then:

$$Z(\xi', \eta', \zeta')d\tau'(d\tau/V) - Z(\xi, \eta, \zeta)d\tau(d\tau'/V), \tag{99}$$

which with the aid of (96) and (98) may be written thus:

$$\frac{vd\tau}{l}[f(\xi', \eta', \zeta') - f(\xi, \eta, \zeta)]d\tau'. \tag{100}$$

This is the quantity of which the integral with respect to ξ', η', ζ', extended over the spherical layer, is equal to $(b - a)d\tau$—the net rate at which compartment C gains particles through impacts.

Making Lorentz' postulate about the form of the distribution-

function f, and remembering that throughout the spherical layer the combination $(\xi'^2 + \eta'^2 + \zeta'^2)^{1/2}$ is confined within a narrow range of values around v we obtain:

$$b - a = (v/l) \cdot \mathfrak{F}(v) \cdot \int (\xi' - \xi) d\tau'/V. \tag{101}$$

To effect the integration it is expedient to change over to polar coordinates in the velocity-space. Leaving the origin where it was, and directing the polar axis towards C, we make the radial coordinates and the colatitude-angle identical with our v and θ, and for the meridian-angle we use the symbol ϕ. Dividing up the layer into compartments by latitude circles and meridians, we have for any one of them:

$$d\tau'/V = (1/4\pi) \sin \theta d\theta d\phi. \tag{102}$$

Consequently we obtain:[20]

$$b - a = (v/4\pi l) g(v) \cdot \int_0^{\pi} d\theta \sin \theta \int_0^{2\pi} d\phi (\xi' - \xi). \tag{103}$$

Everything therefore hinges on the evaluation of $(\xi' - \xi)$—the change in the x-component of velocity which the electron incurs when it leaps from C' to C—as a function of θ and ϕ. Now it may be shown without much difficulty,[21] that:

$$\xi' - \xi = -2v \cos \psi \cos \omega = -2v \sin \tfrac{1}{2}\theta \cos \omega \tag{104}$$

wherein ψ stands as before for the angle between the line of approach of the electron and the line of centres at the instant of impact, and ω stands for the angle between the line of centres and the axis of x. There is also a standard formula[22] relating $\cos \omega$ to ψ, ϕ and the

[20] Lest someone be disconcerted by the apparent difference between this equation and that given by Lorentz, I remark that I am using θ to designate an angle twice as great as the one which he denoted by θ.

[21] Let $\boldsymbol{v}$, $\boldsymbol{v}'$ represent the vector velocities of the electron before and after impact, $\boldsymbol{c}_1$ the unit vector along the line of centres at the moment of impact. The components of $\boldsymbol{v}$ and $\boldsymbol{v}'$ along the line of centres are equal in magnitude and opposite in direction; the components perpendicular to the line of centres are equal in magnitude and direction. Writing these statements down in vector notation:

$$\boldsymbol{v} \cdot \boldsymbol{c}_1 = \boldsymbol{v}' \cdot \boldsymbol{c}_1; \; \boldsymbol{v} - (\boldsymbol{v} \cdot \boldsymbol{c}_1)\boldsymbol{c}_1 = \boldsymbol{v}' - (\boldsymbol{v}' \cdot \boldsymbol{c}_1)\boldsymbol{c}_1 = \boldsymbol{v}' + (\boldsymbol{v} \cdot \boldsymbol{c}_1)\boldsymbol{c}_1, \text{ hence,}$$
$$\boldsymbol{v}' - \boldsymbol{v} = -2(\boldsymbol{v} \cdot \boldsymbol{c}_1)\boldsymbol{c}_1 = -2v \cos \psi \cdot \boldsymbol{c}_1; \; \xi' - \xi = -2v \cos \psi \cos \omega.$$

[22] Imagine two planes P_1 and P_2 intersecting along a vertical axis, the dihedral angle between them being ϕ. Through a point O on the axis draw a horizontal plane N, and from O draw two lines of unit length OR_1 and OR_2, the former in plane P_1 and inclined at β to the vertical axis, the latter in plane P_2 and inclined at ψ to the vertical axis; ω is the angle between them. The points R_1 and R_2 are at heights $\cos \beta$ and $\cos \psi$ above the plane H. On the vertical line through R_2 put a point R_3 at distance $\cos \beta$ above H. Expressions for the sides of the right-angled triangle $R_1R_2R_3$ are easily obtained, and (214) is derived by applying the theorem of Pythagoras to them.

angle $\beta = \arccos(\xi/v)$ between the initial path of the electron and the axis of x, as follows:

$$\cos\omega = \cos\beta\cos\psi + \sin\beta\sin\psi\cos\phi. \tag{105}$$

We now have everything necessary to do the integration in equation (103), and we find:

$$b - a = (v\xi/l)g(v). \tag{106}$$

Substituting this into equation (93), the condition that the distribution-function $(f_0 + \xi g)$ shall be stable by virtue of perfect balance between the rates at which electrons are shifted from compartment to compartment by the impacts and by the accelerating field, we get:

$$(eE/m)\frac{d}{d\xi}(f_0 + \xi g) = (v\xi/l)g. \tag{107}$$

If the term $\xi g(v)$ is truly small by comparison with the term $f_0(v)$, we may neglect the second term on the left; and since $(df_0/dv) = (df_0/dv)(dv/d\xi) = (\xi/v)(df_0/dv)$, the culmination of all the argument is in the formula:

$$\xi g(v) = \frac{\xi l}{v^2}\frac{eE}{m}\frac{df_0}{dv} \tag{108}$$

for the alteration which the applied electric field imposes on the distribution. Notice that g involves ξ and η and ζ only in the combination v; this justifies the procedure of Lorentz.

Now each electron which during unit time crosses any surface imagined in the metal contributes an amount e to the current through that surface; but the contributions made by electrons crossing in opposite senses are opposite in sign—what we perceive as current is *net* current, the excess of the flow of charge one way over the flow the other. Conceive a plane surface-element of area da, normal to the field, therefore normal to the axis of x. We must classify the electrons which traverse it according to their values of ξ. Let $H(\xi)d\xi da$ represent the number passing through in unit time, and having at the moment of passage x-components of velocity in the range $d\xi$ at ξ. This is equal to the number which at any instant have their x-components of velocity in this range, and are situated in the right prism having da for its base and extending a distance ξ down the direction of x:[23]

$$H(\xi)d\xi da = \xi da d\xi \int_{-\infty}^{\infty} d\eta \int_{-\infty}^{\infty} d\zeta f(\xi, \eta, \zeta). \tag{109}$$

[23] This would be immediately obvious, if all the electrons were moving parallel to the x-axis and made no impacts. Electrons having y and z components of velocity in addition to the x-component will drift obliquely out of the prism, and electrons making impacts will be thrown out of the range $d\xi$; but each electron thus lost will be balanced by another coming in from outside.

Since for electrons crossing in opposite senses ξ is of opposite signs, the integral of this expression over all values of ξ, multiplied by e, gives the net current through da when the proper distribution-function is inserted. Evidently the integral will vanish when f is isotropic; there are enormous flows of charge both ways through da, but they are balanced. Unbalance is brought about by the non-isotropic perturbation-term in the distribution-function. Making the postulate of Lorentz, we obtain for the net current through unit area perpendicular to the field, the current-density J, the expression:

$$J = e \int H(\xi) d\xi = e \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} d\xi d\eta d\zeta \cdot \xi^2 g(v). \tag{110}$$

Set for $g(v)$ the expression in (217), and to effect the integration transform into polar coordinates in the velocity-space, orienting the polar axis along the axis of ξ; integrating over the angles, one obtains:

$$J = \frac{le^2E}{m} \frac{4\pi}{3} \int_0^{\infty} v^2 \frac{df_0}{dv} dv. \tag{111}$$

The final integration is easy if one chooses for f_0 the Maxwell function, not quite so easy if one chooses that of Fermi. A further step will be of some advantage. Integrating by parts, and noticing that v vanishes at the lower and f_0 at the upper limit, we find:

$$J = -(8\pi le^2E/3m) \int_0^{\infty} f_0 v dv \tag{112}$$

and the integral remaining, divided by n the number of electrons per unit volume, is seen to be $1/4\pi$ times the mean value of v^{-1},—the average of the reciprocal of the speed of the electrons, in the absence of the field. Denoting this by $\overline{v^{-1}}$, we may write as the general formula for conductivity:

$$\sigma = \frac{J}{E} = \frac{2}{3} \frac{le^2n}{m} \overline{v^{-1}}. \tag{113}$$

The analogy with (91) is obvious, but we must not be misled into identifying the average of the reciprocal speed with the reciprocal of the average speed; they are not quite equal.

The actual final formulæ obtained by the old and the new statistics —substituting, that is to say, the appropriate values of $\overline{v^{-1}}$ from (72b) and (72a), and putting $G = 2$—are as follows:

$$\sigma = \frac{4}{3} \frac{e^2ln}{(2\pi mkT)^{1/2}} \quad \text{(old)}, \tag{114a}$$

$$\sigma = \frac{8\pi}{3} \frac{e^2l}{h} \left(\frac{3n}{8\pi} \right)^{2/3} \quad \text{(new)}. \tag{114b}$$

As Sommerfeld has shown, all of the reasoning by which (108) was reached remains intact even when it is supposed that l is a function of v; in that case, l remains under the integral sign in (111), and the integral itself is equal to $n/4\pi$ times the mean value of $v^{-2}d(lv^2)/dv$. This generalization may be of some value.

Uniform Metal with a Temperature-Gradient; Thermal Conduction

We have now to find a function g such that the distribution $(f_0 + \xi g)$ is stable in a metal in which there is a constant gradient of temperature along the axis of x. When we find it, we shall be able to evaluate the integral

$$W = \tfrac{1}{2}m \int\int\int d\xi d\eta d\zeta \cdot g(v) v^2 \xi^2, \tag{115}$$

which is like the integral in (219) except for the differently-chosen form of g and the substitution of $\frac{1}{2}mv^2$ for e, and therefore represents the net rate of flow of kinetic energy borne by electrons through unit area perpendicular to the gradient—the contribution of the electrons to the flow of heat, under the circumstances stated.

The standard distribution-function f_0, involving as it does the temperature T, is in this case itself a function of x. One might expect that this dependence of f_0 on x would be sufficient to fulfil the requirements. An isotropic distribution which varies from point to point cannot however be stable; the particles conforming to such a one at any given moment would proceed to drift off down the gradient. A stable distribution cannot be isotropic. We must repeat the process of balancing the rates at which particles enter and leave each compartment of the phase-space through collisions and through drift. I say the phase-space now, instead of the velocity-space; for this case is made more complicated than the previous one by reason of the fact, that we now must make the balance separately for the electrons contained in each of the six-dimensional cells $d\xi d\eta d\zeta dx dy dz$, whereas previously we could make it *en bloc* for all of the particles in the entire metal comprised within any velocity-cell $d\xi d\eta d\zeta$.

Consider then the six-dimensional cell $d\xi d\eta d\zeta dx dy dz = ds$, and the $f(\xi, \eta, \zeta, x, y, z)ds$ electrons in it. The first three factors in ds denote the range of velocity, the last three the range in position, within which an electron must lie if it is to belong to ds. Electrons in the proper range of position are continually entering or leaving the proper range of velocity, because of impacts. The net rate at which ds gains electrons in this way is given, as before, by (106). Electrons in the proper range of velocity are continually drifting into the proper

range of position, coming into $dxdydz$ from the region adjacent to it on the side towards smaller or greater values of x, according as ξ is positive or negative. By the same sort of reasoning as led to the term $\alpha df/d\xi$ in (93), one sees that the net rate at which ds loses electrons in this way is $\xi df/dx$. Equating the two, we have:

$$\xi df/dx = b - a = (\xi v/l)g(v), \tag{116}$$

and when we put f_0 for f as before, on the ground that the term (ξg) in the full expression for f makes but a small contribution to the left-hand member, we have all that is required for computing g from (116) and W from (115) for whatever standard function we elect.

At this point, however, there arises a difficulty. If having adopted this way of determining g we proceed to compute the electric current J in the metal by formula (110) we find that it is not zero. The reasoning has led to the conclusion that wherever there is a net current of heat in a metal, there is also a net current of electricity. This conclusion is not in accord with experiment. Yet there is apparently no other way to circumvent it, than to suppose that when a gradient of temperature is maintained in a metal there arises a spontaneous internal electric field, of just such a magnitude as to counteract the electric current which would otherwise persist. The gradient of temperature calls forth a gradient of potential; the actual distribution-function is the one which is stable under both these gradients combined. In the bookkeeping of the compartment ds, the net gain from impacts $(b - a)$ is balanced against the sum of the net loss through drift in the coordinate-space $(\xi df/dx)$ and the net loss through drift in the velocity-space $(\alpha df/d\xi)$. Putting these statements into the form of equations, and denoting by E the hypothetical electric field and by $\alpha(= eE/m)$ the acceleration which it imparts to each electron, we have:

$$\alpha(df/d\xi) + \xi(df/dx) = b - a = (\xi v/l)g \tag{117}$$

$$J/e = \int\int\int d\xi d\eta d\zeta \cdot \xi^2 g = 0, \tag{118}$$

a pair of equations for determining E and the function g.

Lorentz, adopting the Maxwell-Boltzmann function for f_0, solved the equations, and obtained:

$$W = \frac{8}{3}\frac{nlk^2T}{(2\pi mkT)^{1/2}}\left(\frac{dT}{dx}\right) \qquad \alpha = \frac{1}{2}\frac{k}{m}\frac{dT}{dx}. \tag{119}$$

Sommerfeld adopted the Fermi function, and obtained for the degenerate case:[24]

$$W = \frac{8\pi^3}{9}\frac{lk^2T}{h}\left(\frac{3n}{8\pi}\right)^{2/3}\left(\frac{dT}{dx}\right). \qquad (120)$$

The coefficient of dT/dx in these expressions for W is by definition the *thermal conductivity*, usually denoted by κ. One notices that these expressions for κ like those for σ, involve the more or less disposable constants n and l. This however is not true of the ratio of the conductivities.

The Wiedemann-Franz Ratio

For the ratio of thermal to electric conductivity, the old statistics and the new supply expressions involving nothing but T and the ratio of the universal constants k and e, and differing only by a slight numerical factor:

$$\kappa/\sigma = 2(k/e)^2T(= 4.2 \times 10^{-11} \text{ at } T = 291^\circ K) \qquad (121)$$

by the old statistics, and

$$\kappa/\sigma = \tfrac{1}{3}\pi^2(k/e)^2T(= 7.1 \times 10^{-11} \text{ at } T = 291^\circ K) \qquad (122)$$

by the new.

This "Wiedemann-Franz ratio" seems to have been predestined to encourage the devotees of the electron-gas theory. Every other formula offered by the theory contained either n or l or both, and therefore could not serve as an ultimate critical test; for any discrepancy with the data could be removed by adjusting these constants. True, the ensemble of the formulæ provided by the classical theory ran counter to the data in so many different ways, that the net result was quite unfavourable; but one could not point out any single prediction which was certainly wrong. If however the Wiedemann-Franz ratio had departed by an order of magnitude or more from the value of $2(k/e)^2T$, the electron-gas theory could hardly have survived the blow. But in this one case where disagreement would have been fatal, there was agreement; not perfect, but rather too good to be discarded as fortuitous. For many of the familiar metals the ratio, when measured at room-temperature, turned out to be around 6 or $7 \cdot 10^{-11}$. This more than any other one fact was what kept alive the feeling, that in spite of all its difficulties the electron-gas theory must be fundamentally right.

[24] To derive this formula it was necessary to proceed to the second-approximation expression for the Fermi distribution-function; the first approximation merely yielded zero for W.

For the twelve metals Al, Cu, Ag, Au, Ni, Zn, Cd, Pb, Sn, Pt, Pd and Fe, the average of the values of κ/σ at 291° K is $7.11 \cdot 10^{-11}$. The agreement with the prediction of the new statistics is more than good. It is so very good, that it must be partly accidental, especially as the individual values from which the average is formed depart from it by varying amounts. One may still doubt whether it is to be admitted as one of the items which compel the adoption of the new statistics. Drude, be it recalled, obtained the value $6.3 \cdot 10^{-11}$ out of the crude assumption that all of the electrons in any volume-element have the same speed.[25] It used to be regarded as rather amusing that the elaborate calculations of Lorentz merely impaired the agreement which Drude had attained in a naively simple way.

Both theories require that the ratio be proportional to T; this is fairly well satisfied over wide ranges of temperature, but at extreme degrees of cold there is marked divergence, which is inconvenient. It may be desirable to invoke other mechanisms of conduction to supplement the free electrons—as for instance the passing-along of electrons from atom directly to atom to assist in the conduction of electricity, or the transmission of elastic vibrations to aid in the transfer of heat. Indeed, when one reflects that insulators though they lack free electrons yet have some device for the transmission of heat, one wonders why this device should not be available to metals also, and exalt their values of κ and of κ/σ above the predictions of the electron-gas theory.

Intrinsic Potential Difference

We have seen that in a metal where there is no electric current and yet there is a current of heat, an internal electric field must be imagined. We shall now see that in a metal where there is no electric current, but the number of electrons per unit volume varies from point to point, there must also be an internal electric field. This sounds plausible to intuition, for one would expect the electrons to diffuse from regions of higher to regions of lower density unless they were impeded by some force. The equations (117) and (118) enable us to evaluate this force.

Returning to these equations, introduce polar coordinates v, θ, ϕ in the velocity-space as we have formerly done; then $\xi = v \cos \theta$. Multiply both sides of equation (117) by $\cos \theta$; the right-hand member of the new equation is then proportional to $\xi^2 g$. Integrate both

[25] Drude of course could evaluate the ratio k/e without knowing either k or e accurately or at all, since it is the same as the ratio $N_0 k/N_0 e = N_0$ standing for the number of molecules in a gramme-molecule, the Loschmidt number—and $N_0 k$ is the gas-constant R while $N_0 e$ is the Faraday constant of electrolysis.

members of this new equation over the entire velocity-space. The integral on the right then vanishes by reason of (118), and for the integrals on the left we have:

$$\alpha \int d\tau (df_0/d\xi) \cos\theta + \int d\tau \xi (df_0/dx) \cos\theta = 0. \tag{123}$$

By obvious transformations we get:

$$\alpha \int d\tau \frac{df_0}{dv} \cos^2\theta + \frac{d}{dx} \int d\tau f_0 v \cos^2\theta = 0. \tag{124}$$

Integrating over the angles:

$$\frac{4\pi}{3} \alpha \int \frac{df_0}{dv} v^2 dv^3 + \frac{4\pi}{3} \frac{d}{dx} \int f_0 v^3 dv = 0. \tag{125}$$

Leaving the second term as it is, but integrating the first by parts, we find that as f_0 vanishes (whichever statistics we use) at one limit and v at the other limit of integration, we get:

$$-\frac{2}{3} \alpha \int 4\pi f_0 v^{-1} \cdot v^2 dv + \frac{1}{3} \frac{d}{dx} \int 4\pi f_0 v \cdot v^2 dv = 0. \tag{126}$$

The integrals are written in this curious fashion, to bring out the feature that they are proportional to the mean values of functions—the functions v^{-1} and v, respectively—averaged over the electrons in question; which is to say, all the electrons contained in the compartment $dxdydz$ of coordinate-space, to which equation (117) has reference. They are in fact equal to the products of these mean values, to wit the *mean reciprocal speed* and the *mean speed*, by the number $ndxdydz$ of the electrons in the compartment $dxdydz$. Rewriting (125) accordingly, with overlinings to signify averages, and dividing out the factors $dxdydz$ and $1/3$, we get:

$$-2\alpha n \overline{v^{-1}} + d(n\bar{v})/dx = 0, \tag{127}$$

and this is the equation for the acceleration α or the accelerating field $E = m\alpha/e$, required to counteract the electric current which otherwise would be produced in the presence of the gradient $d(n\bar{v})/dx$ of the quantity $n\bar{v}$. This is the gradient which evokes the hypothetical electric field; gradients of temperature or of concentration act indirectly, by making $n\bar{v}$ vary.

With the classical statistics the development is extremely simple, for $\bar{v}$ depends on temperature only, while n may be varied at will.

Heretofore we have tacitly assumed that n remains the same while T and therefore $\bar{v}$ vary along the axis of x, so that:

$$d(n\bar{v})/dx = nd\bar{v}/dx = n(d\bar{v}/dT)(dT/dx), \tag{128}$$

and the reader can verify the expressions for α and W given in (119) by starting from this point. But now we will assume that T and $\bar{v}$ remain the same while n varies along the x-direction with a gradient dn/dx. Then:

$$d(n\bar{v})/dx = \bar{v} \cdot dn/dx. \tag{129}$$

Putting this into (127), and recalling that in the Maxwell distribution the mean values of v and v^{-1} are thus related,

$$\bar{v} = (2kT/m)\overline{v^{-1}}, \tag{130}$$

one perceives that $\bar{v}$ disappears by division from the two sides of the equation, leaving this:

$$\alpha n = (eE/m)n = (kT/m)dn/dx, \tag{131}$$

the desired equation for the necessary electric field. Integrating it, we obtain another of very familiar aspect:

$$n = n_0 \exp (eE/kT)(x - x_0) = n_0 \exp (- [V - V_0]/kT). \tag{132}$$

This is the celebrated equation of Boltzmann embodying the statement that if in an assemblage of particles at uniform temperature there are variations in the number-per-unit-volume from place to place, then there must also be a field of force against which work must be done to move a particle from place to place—and *vice versa*. Specifically: if at any two points P and O the number-per-unit-volume of the particles has values n and n_0, there must be a field of force such that when a particle is moved from O to P its potential energy is increased by $- kT \cdot log(n/n_0)$. If the particles are electrons and the field of force is electric and derived from a potential having values V at P and V_0 at O, then of course this change in potential energy is expressed by $e(V - V_0)$.

Boltzmann's equation is so deeply rooted in modern physics, that it seems strange and suspicious that the new statistics should substitute another but it does. The reason for the innovation stands out very clearly in (127) when the absolute-zero extreme of the Fermi distribution is applied. Owing to the dependence of $\bar{v}$ on n, *owing to the interrelation between average speed and number per-unit-volume* which

distinguishes a system conforming to the new statistics, the second term in (127) is no longer proportional to dn/dx. Instead, we have:

$$n\bar{v} = \frac{3}{4} n v_m = \frac{3hn}{4m} \left(\frac{3n}{4\pi G} \right)^{2/3}; \qquad \overline{v^{-1}} = 3/2 v_m, \tag{133}$$

substituting which values into (237), we obtain:

$$eE/m = \alpha = \frac{h^2}{3m^2} \left(\frac{3}{4\pi G} \right)^{2/3} n^{-1/3} dn/dx, \tag{134}$$

and integrating:

$$-E(x - x_0) = +(V - V_0) = \frac{h^2}{2me} \left[\left(\frac{3n}{4\pi G} \right)^{2/3} - \left(\frac{3n_0}{4\pi G} \right)^{2/3} \right]. \tag{135}$$

This is the formula which supplants Boltzmann's equation.

Consulting (71), we see that (135) may be rewritten thus:

$$e(V_1 - V_0) = (W_i)_1 - (W_i)_0, \tag{136}$$

which is to say: if there is equilibrium between two samples of electron-gas, both being at absolute zero and distributed according to the Fermi law, and the fastest electrons of the two having values of kinetic energy W_{i1} and W_{i0} respectively—then there is a potential-difference between the two, such that if the fastest electron of either group were to cross over to the other, its kinetic energy on arrival would be equal to that of the fastest electron of the group which it joins. So stated, the proposition is easy to remember, and one might even come to think it obvious.

Consider now a pair of pieces of different metals, in contact with one another. One may conceive that they are welded together by an alloy in which the proportion of either varies continuously from zero to one hundred per cent, if one feels the need for a mathematical continuity. If the two pieces were separate, the number of electrons per unit volume would probably not be the same for the two; certainly it is not the same if the number of electrons is equal to the number of atoms per unit volume, for this varies from metal to metal. If the process of welding the metals together does not alter the concentration of the electrons in either at points remote from the junction, then a potential-difference given by (132) or (135)—according as the old or the new statistics is the proper one—must arise between the metals. Taking Sommerfeld's example of potassium and silver: if in unit volume of each of these metals there are as many electrons as atoms, and if this state of affairs continued when the two are welded

together, then between the interiors of the metals across the weld there must be a potential-difference of 4.2 volts,[26] potassium being negative. This figure is calculated by the formula (135); the classical formula gives a value considerably lower, about 0.04 volt. This contrast is characteristic. Both the new and the old statistics associate an internal or intrinsic potential-difference with a difference in electron-concentration, and *vice versa;* but the amount of the P. D. associated with a given pair of concentrations is by no means the same by the two theories; and in actual cases, the new statistics gives much the larger amount.

Though it is not actually possible to measure the potential in the interior of a metal, there are phenomena which indicate that between

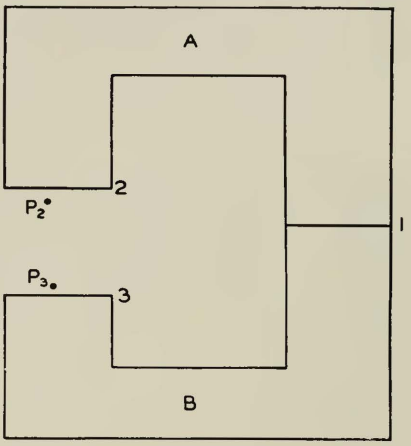


Fig. 2.

two metals touching one another, or between two parts of a metal maintained at different temperatures, there is a difference of potential. These are the *thermoelectric* phenomena—Peltier effect, Thomson effect, thermal electromotive force. The internal potential-gradient reveals itself through the fact that when an electric current is sent through the region where it exists, the rate of generation of heat departs from that which is calculated by Joule's law. We must therefore apply the statistics—this will be the last application which I shall consider—to the problem of evaluating the transport and the generation of heat in an electron-gas, in which the distribution-function is perturbed by an electric field and simultaneously by either of the two other influences—varying temperature, varying concentration of electrons—which we have heretofore considered separately.

[26] Sommerfeld originally computed 5.7 volts, having put $G = 1$; the value 4.2 corresponds to $G = 2$.

Before undertaking this I had better dispel any notion that the "contact" or "Volta" potential-difference between a pair of metals is the measure of the P. D. between their interiors for which we have just been deriving theoretical expressions. It is in fact a measure of something else, as one sees by examining an arrangement like that of Figure 2, where A and B signify pieces of two metals which are in contact at 1, and face one another across a gap between 2 and 3. Consider an electron anywhere inside A, and estimate the potential-barriers which it must cross in order to arrive at the point P_2 just outside of the boundary 2, and also those which it must cross in order to pass through the metal B and reach the point P_3 just outside of the boundary 3. Recalling the symbols and the relations introduced in the section on thermionics, one sees that there is a potential-difference between P_2 and P_3 given by the expression

$$(W_{aA} - W_{aB}) - (W_{iA} - W_{iB}) = b_A - b_B. \qquad (137)$$

This is the contact potential difference; and we see that if the new statistics is correct, it is equal (at the absolute-zero limit) to the difference between the values, for the two metals in question, of that quantity b which appears in Richardson's equation and used to be regarded as the surface work-function. By the old statistics, it differs from $(b_A - b_B)$ by the amount of the internal potential-difference between the metals across the junction 1. Perhaps this difference between the consequences of the two theories could be tested by experiment.

Theory of the Thermoelectric Phenomena

We turn now to the problem of evaluating the rate of generation of heat in a metal through which an electric current is flowing, and in which (according to these theories) there is an intrinsic electric potential-gradient due to a temperature-gradient, or to a gradient of electron-concentration, or both together. The process leads to formulæ which can be tested by experiment, furnishing thus some additional ways of finding out whether these ideas of the new statistics, of the perturbed distribution-function and of the internal electric field are justifiable.

The expression for the rate of generation of heat per unit volume in a conductor traversed by currents of electricity and heat flowing along the axis of x and having current-densities J and W respectively, is $(JE - dW/dx)$. Here E stands for the electric field—not in general for the applied electric field alone, but for the sum of this and the hypothetical internal field. I denote the corresponding acceleration

by α, as before, and the rate at which heat is generated per unit volume by r; then:

$$-r = (m\alpha/e)J - dW/dx. \tag{138}$$

The current-density of heat is given by the formula (115), which I repeat:

$$W = \tfrac{1}{2}m \int v^4 \cos^2\theta g d\tau. \tag{139}$$

Here g stands as always for the non-isotropic perturbation-term in the distribution-function. This and the acceleration α are to be determined from the two equations,

$$\alpha(df_0/d\xi) + \xi(df_0/dx) = (v\xi/l)g = (v^2 \cos\theta/l)g, \tag{140}$$

$$J/e = \int \xi^2 g d\tau = \int v^2 \cos^2\theta g d\tau, \tag{141}$$

which are the same as (117) and (118) except that the electric current is no longer set equal to zero.

Multiply both sides of (140) by $\frac{1}{2}mv^2 \cos\theta$, and integrate over the entire velocity-space. The integral of the right-hand member is W/l; developing the integral of the left-hand member, we find:

$$W = \tfrac{1}{6}ml(-4\alpha n\bar{v} + d(\overline{nv^3})/dx). \tag{142}$$

Multiply both sides of (140) by $\cos\theta$, and integrate over the entire velocity-space; the integral of the right-hand member is J/le; developing the integral of the left-hand member, we get the equation of which (127) was a special case, to wit:

$$-2\alpha\overline{nv^{-1}} + d(n\bar{v})/dx = 3J/el. \tag{143}$$

Evidently these equations suffice to translate (138) into an expression for r in terms of the mean free path, the universal constants, and the averages of various powers of v.

Postulating the Maxwell-Boltzmann distribution-function for f_0; working out the expressions for α and for dW/dx, and importing the formulæ for σ (equation 114a) and κ (equation 119), one finds:

$$\begin{aligned} J(m\alpha/e) &= J \cdot \frac{1}{2}\frac{k}{e}\frac{dT}{dx} - \frac{J^2}{\sigma} \\ dW/dx &= \frac{d}{dx}\left(\kappa \frac{dT}{dx}\right) + 2J\frac{k}{e}\frac{dT}{dx}, \end{aligned} \tag{144}$$

so the value predicted for the rate of generation of heat per unit volume amounts to this:

$$r = +\frac{J^2}{\sigma} + \frac{d}{dx}\left(\kappa \frac{dT}{dx}\right) + \frac{3}{2}\frac{k}{e}\frac{dT}{dx} J. \tag{145}$$

The first term is obviously the Joule heat; the second is not directly a consequence of this current-flow, as it would occur whatever the agency which set up the temperature-distribution in question. It is the third term which concerns us; this is a "reversible heat," proportional to the first power of the current, so that when the current flows in one sense heat is absorbed and when it is reversed heat is evolved. The sign is such, that heat is absorbed when the electrons are flowing towards the hotter part of the metal; the magnitude is such, that as the electrons move onward they acquire just enough energy to raise their temperature to that of the regions which they enter. The coefficient of this term therefore represents the specific heat of the electron-gas, which is the same as that of any other monatomic gas when referred to equal numbers of particles.

Adopting instead the Fermi distribution, and inserting into (142) and (143) the values of $\bar{v}$ and $\overline{v^{-1}}$ and $\overline{v^3}$ prevailing at absolute zero, we find on making the substitutions in the expression (138) for r that the terms containing the first power of J balance one another out. This might have been expected; for we have just seen that these terms form a sum which is proportional to the specific heat of the electron-gas; and if this result may be extended to an electron-gas conforming to the Fermi distribution, then since the specific heat vanishes at zero so also must this "reversible heat." Working through the second approximation, Sommerfeld found that the net coefficient of the term in J in the expression for r is in fact proportional to the specific heat of the electron-gas, being therefore proportional to T, and given by the formula:

$$\frac{2\pi^2}{3}\frac{mk^2}{eh^2}\left(\frac{2\pi G}{3n}\right)^{2/3} T. \tag{146}$$

Now it is a fact of experience that when an electric current flows along a uniform wire of uneven temperature, heat is generated at a rate which involves a term proportional to the current and which changes sign when the current changes sense. This "Thomson heat," like the maximum value which experiments allow us to admit for the specific heat of the electron-gas, has always been much smaller than the value which the classical statistics requires provided that the free

electrons are about as numerous as the atoms. For the Thomson heat as for the specific heat, the new statistics sharply reduces the amounts demanded—to about one per cent of those on which the classical theory insists, at room-temperature that is to say and assuming always that the free electrons are equal in number to the atoms. Agreement in order of magnitude is now attained, and for some metals the advantage is possibly greater; there are indications, too, that the Thomson heat is proportional to T over wide ranges of temperature.

Finally we consider the "Peltier heat"—a term proportional to the current and changing sign when the current changes sense, observed when there is a flow of electricity across a weld or area of contact between two metals. This is clearly to be interpreted as a term in the first power of J, occurring when into the combination $J(m\alpha/e)$—all that remains of the expression (138) for r, when the gradient of temperature is annulled—we substitute the value of α derived from (143) with the assumption that n varies continuously across the weld from the value appropriate to the one metal to the value appropriate to the other. Using the classical statistics, we find that there is such a term; denoting by n and n_0 the electron-concentrations in the two metals, we find for its value:

$$(kT/e) \log (n/n_0)J. \tag{147}$$

Its value for unit current is obviously the intrinsic potential-difference between the metals. Using instead the new statistics, we find that at the absolute zero there is no such term; we must proceed to the next approximation, doing which, Sommerfeld obtained the expression:

$$\frac{2\pi^2}{3}\frac{m(kT)^2}{eh^2}\left[\left(\frac{4\pi G}{3n}\right)^{2/3} - \left(\frac{4\pi G}{3n_0}\right)^{2/3}\right]. \tag{148}$$

Putting the current equal to unity, we find a value very considerably smaller than the intrinsic potential-difference between the metals—a fraction of a millivolt. This is the order of magnitude of the Peltier heat as it is actually observed in many cases. Curiously enough, this fact by itself is in accord with both the theories. By the classical statistics, the intrinsic potential-difference between two metals is generally small, and the Peltier heat for unit current gives its value directly; by the Fermi statistics, the intrinsic potential-difference is generally large, but the Peltier heat for unit current is only a small fraction of it.

Omissions

Among the subjects omitted from this article there are several of much interest, which the reader may trace from the annexed bibliography; in particular:

Sommerfeld's theory of the Hall effect;

Houston's extension of Sommerfeld's theory of intrinsic potential difference, including especially an explanation of the Peltier heat arising when current flows between two differently-oriented crystals of a single substance;

Bloch's use of the new methods of quantum mechanics to make allowance for the influence of the atoms on the conduction-electrons;

Fermi's application of statistical methods to the problem of determining the distribution of electrons in the individual atom;

Fuerth's work on the fluctuations in the new statistics.

The bibliography will indicate other interesting advances in a variety of problems.

Bibliography

J. M. Adams: Value of W_a for silver. Zeits. f. Physik **52**, p. 882 (1929).
W. Anderson: Zeits. f. Physik **50**, p. 874–877 (1928). Astrophysical applications.
H. M. Barlow: Phil. Mag. (7) **7**, pp. 458–470 (1929). A criticism of the electron theory of metals.
R. S. Bartlett: Proc. Roy. Soc. **A121**, pp. 456–464 (1928). Thermionic and cold discharge.
E. S. Bieler: Journ. Franklin Inst. **206**, pp. 65–82 (1928). General account of new statistics of electron-gas.
H. F. Biggs: Nature **121**, pp. 503–506 (1928). General.
F. Bloch: Zeits. f. Physik **52**, pp. 555–559 (1928). Quantum mechanics of electrons in crystal lattices.
F. Bloch: Susceptibility and change of resistance in magnetic field. Zeits. f. Physik **53**, pp. 216–227 (1929).
Bose: Zeits. f. Physik **26**, pp. 178–181 (1924). First paper on the Bose statistics.
L. Brillouin: C. R. **184**, pp. 589–591 (1927); **188**, pp. 242–244 (1929).
P. A. M. Dirac: Proc. Roy. Soc. **A112**, pp. 661–677 (1926).
J. W. M. DuMond: Proc. Nat. Acad. Sci. **14**, pp. 875–878 (1928); Phys. Rev. (2) **33**, pp. 643–658 (1929). Scattering of x-rays by the free electrons.
C. Eckart: Zeits. f. Physik **47**, pp. 38–42 (1928). Contact P. D.
A. Einstein: Berliner Sitzungsberichte, 1924, pp. 261–267; 1925, pp. 3–17. Application of Bose statistics with modification to material gases.
E. Fermi: Zeits. f. Physik **36**, pp. 902–912 (1926). Fermi statistics.
E. Fermi: Zeits. f. Physik **48**, pp. 73–79 (1928). Application of statistics to electrons within atoms.
E. Fermi: Zeits. f. Physik **49**, pp. 550–554 (1928). Statistical evaluation of the *s*-levels.
E. Fermi: *Quantentheorie und Chemie* (Hirzel, 1928), pp. 95–111. Application of statistics to problems of atomic structure.
V. Fock: Zeits. f. Physik **49**, pp. 339–357 (1928). Dirac's statistical equation.
R. H. Fowler: Proc. Roy. Soc. **A117**, pp. 549–552 (1928). Thermionics.
R. H. Fowler, L. Nordheim: Proc. Roy. Soc. **A119**, pp. 173–181 (1928).
R. H. Fowler: Proc. Roy. Soc. **A120**, pp. 229–232 (1928). Photoeffect.
R. H. Fowler: Proc. Roy. Soc. **A122**, pp. 36–49 (1929). Thermionic constant A.
R. H. Fowler: Monthly Notices Royal Astron. Soc. **87**, pp. 114–122 (1926). Astrophysical application.
J. Frenkel: Zeits. f. Physik **49**, pp. 31–45 (1928). Theory of magnetic and electric properties of metals at absolute zero.
J. Frenkel, N. Mirolubow: Zeits. f. Physik **49**, pp. 885–893 (1928). Theory of metallic conduction by wave mechanics.
J. Frenkel: Zeits. f. Physik **50**, pp. 234–248 (1928). Cohesion.
R. Fuerth: Zeits. f. Physik **48**, pp. 323–339 (1928), and **50**, pp. 310–318 (1929). Fluctuations in the new statistics.

G. E. Gibson, W. Heitler: Zeits. f. Physik **49**, pp. 465–472 (1928). Chemical constant in the new statistics.
E. H. Hall: Proc. Nat. Acad. Sci. **14**, pp. 366–380, 802–811 (1928). Arguments against the electron-gas theory.
W. V. Houston: Zeits. f. Physik **47**, pp. 33–37 (1928). Cold discharge.
W. V. Houston: Zeits. f. Physik **48**, pp. 449–468 (1928). Conduction and thermoelectric effects in metals.
W. V. Houston: Phys. Rev. (2) **33**, pp. 361–363 (1929). Cold discharge.
O. Klein: Zeits. f. Physik **53**, pp. 157–165 (1929). Reflection of electrons at a potential-discontinuity by Dirac's dynamics.
E. Kretschmann: Zeits. f. Physik **48**, pp. 739–744 (1928). Continuation of Sommerfeld's work.
J. E. Lennard-Jones: Proc. Phys. Soc. Lond. **40**, pp. 320–537 (1928). General account.
J. E. Lennard-Jones, H. J. Woods: Proc. Roy. Soc. **A120**, pp. 727–735 (1928). Electron-gas theory with modifications to allow for atoms in metals.
A. C. G. Mitchell: Zeits. f. Physik **50**, pp. 570–576 (1928). Entropy of the electron-gas.
L. Nordheim: Phys. Zeits. **30**, pp. 177–196 (1929). General report on new theory of thermionic emission, cold discharge and photoeffect. Earlier papers: Zeits. f. Physik **46**, pp. 833–855 (1928); Proc. Roy. Soc. **A121**, pp. 626–639 (1928); **A119**, pp. 173–181 (1928) (with R. H. Fowler).
L. Nordheim: Proc. Roy. Soc. **A119**, pp. 689–698 (1928). Alternative way of deriving the distribution-laws.
L. Nordheim: Naturwiss. **16**, pp. 1042–1043 (1928). Conductivity of alloys.
J. R. Oppenheimer: Proc. Nat. Acad. Sci. **14**, pp. 363–365 (1928). Cold discharge.
W. Pauli: Zeits. f. Physik **41**, 81–102 (1927). Paramagnetism.
W. Pauli: *Probleme der modernen Physik* (Hirzel, 1928), pp. 30–45. H-theorem.
F. Rasetti: Zeits. f. Physik **49**, 536–549 (1928). Statistical evaluation of the M-levels.
O. K. Rice: Phys. Rev. (2) **31**, pp. 1051–1059 (1928). Electrocapillarity.
O. W. Richardson: Proc. Roy. Soc. **A117**, pp. 719–730 (1928). Cold discharge.
L. Rosenfeld, E. E. Witmer: Zeits. f. Physik **47**, pp. 517–521 (1928). The radiation-gas.
L. Rosenfeld, E. E. Witmer: Zeits. f. Physik **49**, pp. 534–541 (1928). Refractive indices of metals for electron-waves, values of W_a.
R. Ruedy: Phys. Rev. (2) **32**, pp. 974–978 (1928). Dispersion of metals.
E. Schroedinger: Phys. Zeits. **27**, pp. 95–101 (1927).
A. Sommerfeld: Zeits. f. Physik **47**, pp. 1–32, 43–60 (1928). The fundamental article on the application of the Fermi statistics to electrons in metals. Preliminary note in Naturwiss. **15**, pp. 825–832 (1927).
A. Sommerfeld: Naturwiss. **16**, pp. 374–381 (1928). Extension of the article just listed.
A. T. Waterman: Phil. Mag. (7) **6**, pp. 965–970 (1928). Dependence of conductivity on pressure.
A. T. Waterman: Proc. Roy. Soc. **A121**, pp. 28–41 (1928). Cold discharge.
G. Wentzel: *Probleme der modernen Physik* (Hirzel, 1928), pp. 79–87. Photoeffect.

Physical Properties and Methods of Test For Some Sheet Non-Ferrous Metals

By J. R. TOWNSEND and W. A. STRAW

This paper covers an investigation which was undertaken to secure a simple and reliable method of test for sheet non-ferrous metals. An account of the early development work leading to the adoption of the Rockwell hardness tester for a preliminary inspection of sheet metals and the tensile test as the basic test to be referred to in case the Rockwell test results were near to or outside the established Rockwell limits for a given lot of material was published in 1927.[1] The continuation of this work including establishment of test limits for four grades of brass and for two grades each of nickel silver and phosphor bronze will be published this year.[2]

This present paper describes the development work reported by these two papers.

Considerable attention has been given to the Rockwell tester, which, as a result of this work, has been found satisfactory for use as a specification instrument for brasses 0.020 in. and thicker, when used under standardized methods of test and calibrated with standard test blocks. Other tests such as the bend test, ductility test and other hardness tests have been studied but further development work is necessary.

The Rockwell hardness and tensile strength limits are given for four alloys of brass, and two alloys each of nickel silver and phosphor bronze. The physical properties of the rolling series upon which these limits were based are presented as well as experience data obtained on shipments of commercial material. The limits for brass alloys A, B and E are considered final, but the limits for the other metals are tentative until more complete experience is available.

Grain size limits are given for annealed brass and nickel silver sheet. For inspection purposes the grain size is estimated by comparison with the standard photomicrographs reproduced in the 1929 report of Committee E4 of the A. S. T. M. on Metallography.

Refinements in the calibration of the Rockwell tester are given, as well as the development of testing technique. An experimental model of a motor-driven bend test machine is described.

THE investigation which this paper covers was undertaken because of the need for a simple method of test which would afford a more definite determination of the properties of thin-sheet non-ferrous metals than any which has yet been developed. Large quantities of

[1] "Physical Properties and Methods of Test for Sheet Brass," by H. N. Van Deusen, L. I. Shaw, and C. H. Davis, *Proceedings of the American Society for Testing Materials*, 1927.

In addition to the authors the following took an important part in the investigational work. Messrs. H. N. Van Deusen C. H. Greenall and W. S. Hayford, of the Bell Telephone Laboratories, Messrs. W. H. Bassett, R. M. Tree, Alden Merrill and G. S. Mallett of the American Brass Co., and Messrs. L. I. Shaw, M. D. Helfrick H. F. Culver and G. R. Brown of the Western Electric Co. Messrs. N. E. Newton and W. H. Eastlake of the Northern Electric Co., Montreal, participated in the conferences which were held as the work progressed.

[2] "Physical Properties and Methods of Test for Sheet Non-Ferrous Metals," by J. R. Townsend, W. A. Straw and C. H. Davis, presented before A. S. T. M., June 1929.

brass, nickel silver and phosphor bronze are used in the manufacture of telephone apparatus as structural members, springs, and bearings. Because of space limitations, the parts are necessarily small; many are formed into irregular shapes; spring parts must maintain accurate adjustments and have long fatigue life; certain other parts must resist wear. All requirements are steadily becoming more exacting because of the increasing complexity of telephone systems and it, therefore, became necessary to insure more closely the uniformity of the material used in the apparatus.

The methods of test developed to check the uniformity of material are of equal concern to both consumer and supplier. Both are interested in a test rapid enough for use in inspection work and also so simple that specially trained men are not necessary for the actual testing. Due to the large amount of material that must be inspected it is necessary that any tests used require little time to apply and be adaptable to modern production methods. Furthermore, close agreement in test results is necessary. Since no test method was available fulfilling these requirements, it became necessary either to develop new methods of test or to modify existing methods. In addition, limits were desired that would represent the best quality of material obtainable consistent with commercial mill practices.

As a result of this need a cooperative program of tests was laid out which would lead to the drafting of requirements on thin-sheet metals. Those cooperating in this program were associated in the production and use of such material, and each member had already done preliminary work which evidenced his interest in the problem. The group was limited to one producer and one consumer in order that the work might be expedited, and the results that have been obtained justify the plan.

The results of the investigation are given herein for such value as they may have to the industry. It is hoped that this work may act as a stimulus to others interested in this problem so that ultimately definite standards may result, including the refinements demanded by modern industry.

The results of the study of the methods of test suitable for high and clock brass sheet of the order of 0.020 in. and thicker are presented and it is shown that the Rockwell tester gives the most reliable information of the several hardness testing machines available at the present time. Methods of operation of the Rockwell hardness tester were worked out by the cooperating laboratories and a method of carrying on tension tests was developed. The limitations in the use of the Rockwell hardness tester for sheet metal are pointed out and

the need for the development of a tester capable of higher precision in the harder materials and also capable of making hardness tests on material less than 0.020 in. thick is emphasized.

The tensile strength limits are determined by the following procedure. The tension test results of the rolling series are plotted against the actual percentage of reduction by cold rolling. Two limiting curves are drawn in giving the minimum and maximum tensile strength for all reductions covering the range of commercial anneals. These two curves are based on years of experience with these metals and also on the rolling series. The tensile strength limits were taken from these curves for the theoretically correct reduction for each temper. Rockwell hardness tests are made on the grip ends of the tension test specimens, thus establishing the Rockwell hardness-tensile strength relationship for the material. Having established the tensile strength-reduction relationship, the corresponding hardness values are obtained from the Rockwell hardness-tensile strength curve. These limits are then subjected to trial on a large number of shipments of material.

The Rockwell test is considered a preliminary inspection test and is mainly useful because of its economy of time and material. The tension test, on the other hand, is considered the test upon which the acceptance or rejection of the material is based. In practice, material within the Rockwell hardness limits is accepted unless the hardness reading is near or outside the hardness limits, in which case a tension test is made.

This paper covers high and clock brass sheet and four other alloys of brass. One of these contains less lead than clock brass and is designed for use where a material combining moderate drawing and cutting properties is desired. Another has a nominal composition of 72 per cent of copper and 28 per cent of zinc. Grain size requirements are given for two others mainly used for drawing purposes. One of these consists of nominally 85 per cent of copper and 15 per cent of zinc and the other of 75 per cent of copper and 25 per cent of zinc. This paper also covers two nickel-silver alloys and two phosphor-bronze alloys.

The Rockwell hardness and tensile strength limits given for all of these alloys other than high, clock and alloy E brass may be considered tentative in view of the limited experience had with commercial shipments of materials purchased in accordance with these limits up to the present time. The requirements for high, clock and alloy E brass are considered final. Alloy E brass has a lead content midway between high and clock-brass.

Method of Handling Work

The round-robin tests on the Rockwell hardness tester, standard test blocks and calibration of the machines were made by the American Brass Co. (Waterbury, Buffalo and Kenosha Branches), Bell Telephone Laboratories, and the Western Electric Co. The tension tests were also made by these laboratories. All of the rolling series were manufactured by the American Brass Co. Experience data on shipments of non-ferrous metals were obtained jointly by the American Brass Co. and the Western Electric Co.

Methods of Test

Tension Test:

As stated previously, the tension test is considered the basic mechanical test for cold worked materials. Complete data on tensile properties were therefore obtained on the rolling series. All tests were made in accordance with the standard testing procedure developed for use in the Bell System and given in Appendix II. This method has been in use over 4 years and has been found satisfactory to all concerned. Three tension test specimens were made from each sample. The tensile strength, proportional limit, percentage of elongation in 2 in. and modulus of elasticity were determined for each sample of the high-brass rolling series. The specimen 14 in. in length was used which allowed the use of an 8-in. gage length Anderson [3] extensometer. The shorter specimen with the 2-in. gage length was used for the clock brass and all later work. The tests were made on an Amsler tension testing machine. The machine was carefully calibrated, the load indications being correct to less than $\frac{1}{2}$ of 1 per cent.

The tensile strength and percentage of elongation on the high-brass sheet rolling series are plotted in Fig. 8 against per cent reduction by rolling. Nearly a straight line relation exists between the tensile strength and percentage reduction by rolling and the values are grouped closely about this line. The curve is plotted to show the average result, but separate curves could readily be drawn for each thickness. This test also has little sensitivity in the harder tempers where the curve becomes asymptotic. The proportional limit is difficult to measure with any degree of accuracy on account of the personal factor involved in interpreting curves. The results in this case showed no definite relation to any of the other values.

In order to evaluate the test results, it is necessary to establish control limits to determine whether any variations in the data

[3] H. A. Anderson, "Tension Tests of Thin Gage Metals and Ligh Alloys," *Proc. A. S. T. M.*, Vol. 24, Part II, p. 990 (1924).

are significant. By significant variations are meant variations that can be assigned to definite sources, such as measuring errors, defects in the metal tested, testing errors, etc. For the case under consideration, a method of analysis was employed which allows for small sample numbers. The method used is described elsewhere.[4]

Figure 1 shows an engineering analysis of the tensile strength

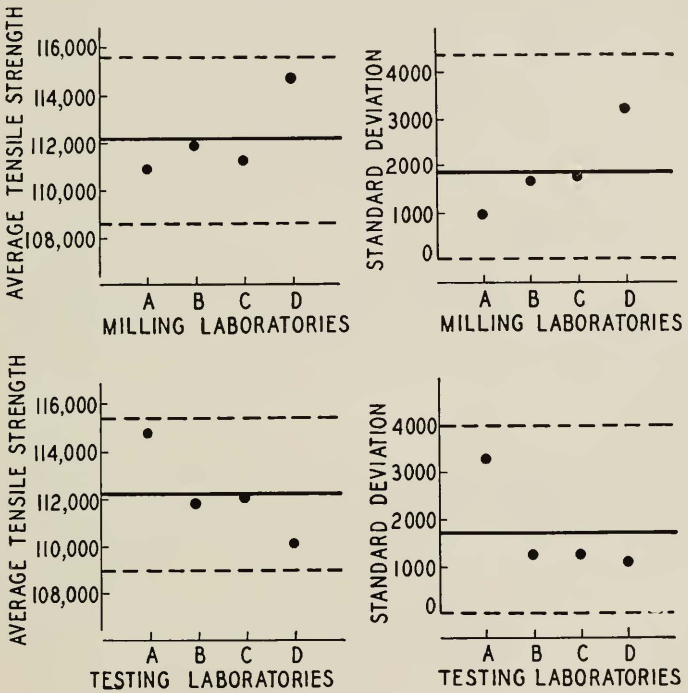


Fig. 1—Statistical Analysis of Round-Robin Tension Tests on Alloy B Nickel Silver.
(*a*) Points are average of all specimens milled at laboratories indicated, and tested by the four laboratories.
(*b*) Points are average of all specimens tested at laboratories indicated, and milled at the several laboratories.

results for alloy B nickel silver plotted in two ways.[5] The data in Fig. 1 (*a*) were obtained by computing, for each of the laboratories, the average tensile strength and the standard deviation [5] of all the specimens milled by that laboratory and tested by each of the four laboratories *A*, *B*, *C* and *D*. The data in Fig. 1 (*b*) were obtained

[4] See Modified Criterion No. 1 in Appendix I to the Report of Sub-Committee XV on Die-Cast Metals and Alloys appended to the report of Committee B-2 on Non-Ferrous Metals and Alloys, presented before the American Society for Testing Materials, June 1929.

[5] The methods used are those described by W. A. Shewhart in a paper on "Quality Control," *The Bell System Technical Journal*, Vol. VI, October 1927.

by computing, for each of the laboratories, the average tensile strength and the standard deviation of the specimens milled in the four laboratories and tested in each of the respective laboratories. In each instance the specimens tested or milled by an individual laboratory are grouped together. The dotted lines on the diagrams represent the control limits within which the data should fall without leaving anything to chance, which means that points falling without these control limits indicate variations due to assignable causes, such as errors in measurements, defects in the material, etc. These diagrams show, therefore, that the tensile strength results between laboratories do not reveal significant or assignable difficulties other than those which could be attributed to chance. In other words, the analysis gave no indication of the presence of assignable variations between the testing laboratories.

These specimens were prepared by cross milling the gage length with a milling cutter shaped to conform to the final shape of the specimen desired. This method, which has been described elsewhere,[6] results in the saving of time and produces specimens of a uniform character.

Scleroscope Hardness Tests:

Previous to this investigation Bell System specifications on nonferrous materials were written in terms of scleroscope hardness, but considerable trouble was encountered between the suppliers and users of metal due to difficulty in checking each other's readings. Because results could not be duplicated on two instruments, it was necessary to allow rather wide limits in each temper of material. This resulted in considerable overlapping in scleroscope limits of the tempers of materials accepted under these specifications. Experience had shown it to be impossible to make correction curves for any two instruments which would hold for any reasonable length of time, and if any replacements such as new hammers were necessary the calibration was changed.

In order to obtain more definite information as to what could be expected from the scleroscope, comparisons were made between four type "C" sclerscopes located in four laboratories using the same samples of materials. Great care was taken in preparing the samples so that they were flat and free from any dirt and roughness. The instruments were in good commercial adjustment and were employed in the usual manner, the magnifier hammer being used. A single

[6] R. L. Templin, "Methods for Determining the Tensile Properties of Thin Sheet Metals," *Proc. A. S. T. M.*, Vol. 27, 1927.

thickness of material was used in each case. The results of one of these comparison tests are shown in Fig. 2. Two other such comparisons were made incidental to the work being done to establish operating technique on the Rockwell tester, but the results shown here are typical. At least five readings were made on each sample

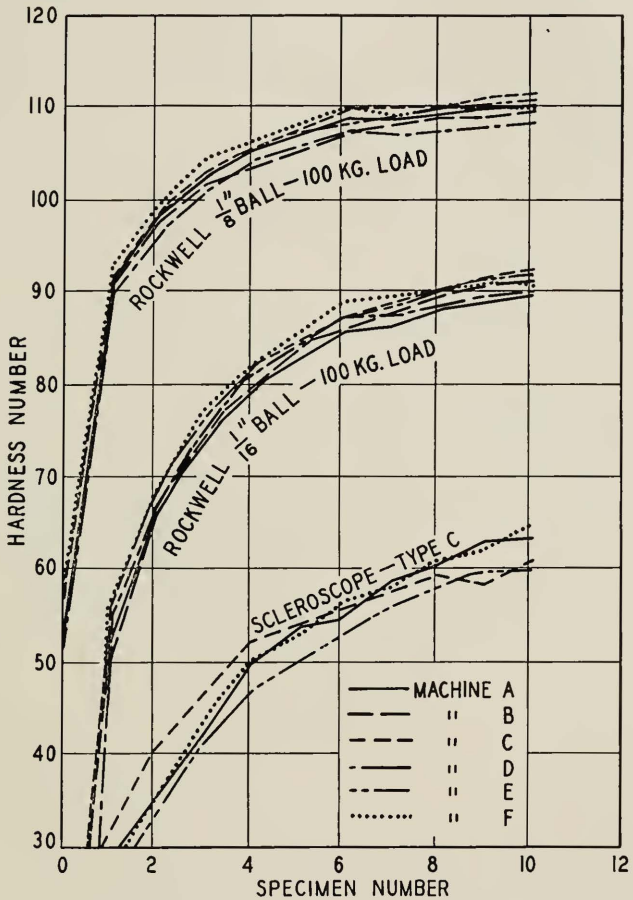


Fig. 2—Brass "Round-Robin" Series. Comparison of Rockwell and Scleroscope Testers.

and the average was used in preparing these curves. It will be noted that the readings on the different scleroscopes varied as much as eight points from each other on cold-rolled materials. The variations in readings between the machines follow no apparent law, one machine reading higher at one part of the scale and lower at another. It will also be noted from the curves that the range of the scleroscope

is much smaller than that of the Rockwell on the same samples. In other words, the sensitivity of the scleroscope is considerably less than that of the Rockwell.

An investigation was made of the type "D" recording scleroscope equipped with a universal hammer to obtain a comparison with the type "C" instruments. The conclusion reached was that there was little difference between the two instruments as regards precision. It was not considered worth while, therefore, to make elaborate comparison tests. Hence, this type of instrument was not used in the round-robin tests.

In view of the difficulties encountered in using the scleroscope, which have also been brought out by previous investigators, no further effort was made to adapt it for use as a specification instrument.

Rockwell Hardness Tests:

While the Rockwell hardness tester has been used quite extensively and with considerable success in testing steels it has been used comparatively little for testing non-ferrous metals. Our experience previous to this investigation indicated that it might prove satisfactory as a specification instrument for non-ferrous metals. Considerable work had already been done by various laboratories to determine the limitations of this machine. Furthermore, the routine testing with the Rockwell of all incoming shipments of material had been instituted for the purpose of accumulating specification data. While each individual machine seemed to give satisfactory results to the user there was little information available as to what agreement could be obtained with machines in other laboratories. Figure 3 gives results of tests on identical samples with five different machines in the laboratories of one of the participating companies before any attempt was made to eliminate mechanical irregularities in the machines. While there was considerable difference between machines, there was a probability that these differences could be reduced by carefully going over the machines and establishing technique of test.

Before making any further comparative tests, careful study was made of each machine and various mechanical irregularities were eliminated. In order to get close comparative results the ball penetrators must be in good condition, the load must be applied without impact and at approximately the same rate in different machines, the value of the load must be the same, and the method of supporting the specimen across the anvil is important. Due to the lever system of applying load and measuring the amount of penetration of the material under this load, a slight amount of friction in the bearings

may result in variations in readings which are not apparent without comparisons with standard blocks. A set of directions for calibrating the Rockwell tester was drawn up which called for a thorough check of these various features of the machine. (See Appendix I.)

Other questions had to be settled, such as the size of ball penetrator to be used and the amount of the load. The possibility of varying the load and size of ball renders the Rockwell machine capable of broad application for testing purposes. Preliminary tests made to determine the most suitable combination of ball and load to be used indicated that the standard "B" scale of the instrument ($\frac{1}{16}$-in. ball and 100-kg. load) would be the most satisfactory for brass. However, it was decided to try also a $\frac{1}{8}$-in. ball with the same load. In our

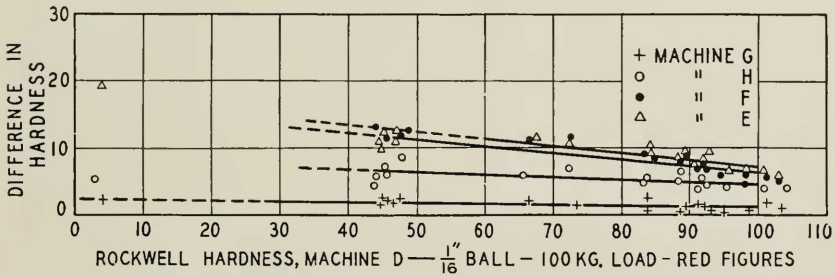


Fig. 3—Preliminary Comparison of Rockwell Testers.

later work on nickel silver and phosphor bronze it was found that the 150-kg. load gave more satisfactory results with those materials. A larger ball is desirable to get readings on annealed brass and copper but with harder materials sensitivity of reading is increased by increasing the load and retaining the $\frac{1}{16}$-in. ball.

The question of the number of thicknesses of material to be tested was also given consideration. On very thin materials there is a distinct anvil effect which can be reduced or eliminated by superposing several samples on each other. However, this method of testing soon proved to be objectionable. The Rockwell tester does not distinguish between actual penetration of the material and descent of the penetrator from any other cause, and since one hardness number of the Rockwell machine corresponds to an 0.00008-in. movement of the penetrator, great care must be taken to insure that movement of the penetrator is due only to actual penetration of the material. Unless the spaces under the penetrator between two or more layers are completely closed by the 10-kg. minor load, low readings will be obtained. With spring materials, which are frequently curved due

to coiling, this condition is often encountered. Furthermore, in very thin materials there is a side flow so that the upper sheets fail to take their share of the load and tend to curl up around the penetrator in such a manner as to prevent the proper measurement of depth. There is also a possibility of including small particles of dirt between the specimens which would result in lower hardness readings. These matters must be very closely watched. An indentation in the anvil just barely visible on a polished surface in strongly reflected light will cause erroneous readings. It is necessary, therefore, that the anvils be sufficiently hard for the purpose. During the past year, harder anvils have been furnished and greatly improved results have been obtained in the testing of thin materials.

A comparison of the methods used by the laboratories in making tests indicated that personal factors might account for some of the differences in readings. One of these personal factors which appeared to be of major importance was the amount of time allowed for the drift of the penetrator after the major load is completely applied. It was the custom of some of the participating laboratories to allow this drift to continue until it had practically ceased while others removed the major load and read the hardness as soon as possible after the major load had been completely applied. This drift is very noticeable in the softer tempers of non-ferrous metals and may amount to as much as 10 Rockwell numbers. In general, it is more noticeable with non-ferrous metals than with steels. In order to eliminate this personal factor, the practice of operating the reset mechanism which removes the major load, immediately upon the full application the major load was adopted.

Three separate rolling series were used in the round-robin tests which were made after the various machines had been checked up. In each case tests were made with both $\frac{1}{16}$ and $\frac{1}{8}$-in. balls using the 100-kg. load. All tests were made on each of six machines located in the six different laboratories. Typical results of one of these series are shown in Fig. 2 which gives the results of a brass rolling series using both the $\frac{1}{8}$ and $\frac{1}{16}$-in. ball. It will be noted that the greatest sensitivity is obtained by using the $\frac{1}{16}$-in. ball, and this was later adopted as a standard ball for testing brass. Much better agreement between Rockwell testers than between scleroscope testers is also apparent.

Rockwell data were obtained on the high-brass and clock-brass rolling series in all thicknesses and tempers using one thickness of material and the standard "B" scale. The plot of the results of these tests on high brass is shown in Fig. 6. These curves appear normal for thicknesses of 0.012 in. and above, but below this thickness the

hardness readings bear no relation to tensile strength due to the anvil effect.

The Rockwell hardness test is carried on in the same manner as given in Appendix I of this paper. All of the Rockwell hardness tests were made on the two grip ends of the tension test specimens. This was done in order to have Rockwell hardness tests and tension tests made on as nearly identical material as possible. Commercial experience with the Rockwell machine in the routine testing of sheet non-ferrous metals has indicated the importance of the following precautions.

Ball Penetrator.—Slight variations in the size and sphericity of the ball penetrators are to be expected and must be guarded against. The results of several hundred measurements made at various points on the "B" scale show that a deviation of less than 0.00002 in. in the diameter of the $\frac{1}{16}$-in. ball has no effect on the hardness readings. Balls showing a greater variation give noticeable error in hardness readings. Penetrator balls have been held within this limit.

Anvil Surface.—It is desirable that the penetrator be perpendicular to the testing surface of the anvil. It is assumed that the penetrator is operating perpendicular to the seating surface and in line with the capstan head. Any lack of perpendicularity would then be due to lack of parallelism of the seating and testing surfaces of the anvil. This is checked by a fixture designed to hold the anvil so that the specimen supporting surface might be checked. A hardened flat, 2 in. square and $\frac{7}{8}$ in. thick, having a $\frac{13}{16}$-in. hole through its center, was lapped so that its two surfaces were flat and parallel to within 0.00005 in. By placing the fixture holding the anvil on the table of an optimeter and sliding the supporting surface of the anvil under a ball-pointed feeler gage, errors in flatness and parallelism may be measured to 0.00005 in. Anvils whose surfaces were found to be plane to within 0.0001 in. were considered satisfactory.

Machine Errors.—Observations on four Rockwell machines at the Western Electric Co., one of which handles approximately 10,000 tests per month and all of which were overhauled and calibrated for their entire range once a week, showed that changes in calibration occurring in this interval were less than one hardness number. Before beginning a series of readings the correction to be applied was determined by taking readings on a standard block. The average change on all three scales was about one hardness number. Variations from calibration were rarely greater than 1.5 numbers unless the penetrator was damaged or the instrument was out of adjustment due to misuse.

The effectiveness with which standard test blocks may be used in bringing Rockwell testers into agreement is shown by the following calibration:

	Approximate Hardness, "B" Scale, 1/16-in. Ball, 100-kg. Load (Red Figures)		
	25	60	80
Before overhauling and checking the entire scale, average difference in hardness readings between three testers	6.0	3.7	3.4
After overhauling and checking the entire scale, average difference in hardness readings between three testers	0.9	1.2	0.9

Calibration of Test Blocks.—It was first thought that a Rockwell hardness machine would hold its adjustment over a long period if carefully maintained and could serve as a standard machine to be used for the calibration of test blocks. More complete experience has shown, however, that wear of the operating parts and slight variations in friction caused the "standard machine" to vary slightly in its readings. Naturally, the calibration of standard test blocks under these conditions is unsatisfactory because it might result in the establishment of several sets of standard blocks, leading to confusion in the application of the Rockwell hardness test limits to commercial material.

A new method of calibrating the Rockwell hardness tester was devised. This method consisted in preparing a set of test blocks covering the range of hardness of the high-brass and clock-brass rolling series. These blocks were checked by the various cooperating laboratories as well as the manufacturers of the Rockwell hardness tester, namely, the Wilson-Maeulen Co. These blocks were considered the basic standards for the Rockwell tests. Sub-standard blocks were calibrated in comparison with these standard blocks and placed in use by the various cooperating laboratories. The high-brass rolling series and the clock-brass rolling series were then retested for Rockwell hardness using the new calibration. It was found that the hardness readings differed one to two points from the values previously determined for several tempers. At the same time experience with commercial shipments of material showed that a shift was necessary in the Rockwell hardness-tensile strength relationship. This verified our conclusions with regard to the adoption of the new method of calibration.

After the establishment of the standard brass test blocks it was

seen that blocks would also be required to cover the nickel-silver and the phosphor-bronze alloys. To prepare new series of blocks especially for these materials would have resulted in a wide variety of blocks and would have been found burdensome to the laboratories. Consequently, a single series of test blocks covering the brass, nickel-silver, and phosphor bronze alloys was selected. This series of test blocks started with the original blocks to which were added a few additional blocks made necessary in order to cover the range of the nickel-silver and phosphor-bronze alloys. The data upon which the calibration of these blocks was based have been submitted to the Section on Indentation Hardness of the A. S. T. M. Committee E-1 on Methods of Testing.

Hardness of Thin Sheet.—The Rockwell hardness test has certain limitations in its application to the testing of thin sheet stock. Material less than 0.020 in. thick gives hardness readings different from thicker sheet of the same temper. Referring to the curves shown hereinafter of Rockwell hardness plotted against tensile strength for materials thinner than No. 24 B. & S. gage the points fall below the curve. This apparently is due to lack of support of the metal about the penetrator and consequently a low reading is given. For still thinner material the penetrator passes nearly through the metal and the hardness reading recorded is inaccurate due to the effect of the supporting anvil in addition to lack of support of the metal.

Cleaning Anvil and Specimen.—Inasmuch as the Rockwell hardness tester measures hardness in terms of penetration, any movement of the penetrator affects the hardness reading. In other words, if the metal has a roughened surface or if the anvil is not polished smooth the metal will flow under the high unit pressure involved and this will cause the Rockwell hardness reading to be lower than its true value. It is considered necessary therefore that the anvil should be polished flat and the material tested should be reasonably free from surface imperfections and oxide film. Table III shows the effect on the hardness readings of polishing the anvil.

In addition to the need of polishing the anvils, if close agreement is to be had, some refinement is also needed in the use of the standard test blocks. It is difficult to obtain test blocks that will not show a variation in hardness. It has been customary, therefore, in calibrating test blocks to take five readings, one in each corner of the test block and one in the center. These readings are made on a machine which has been calibrated with the standard test blocks both before and after calibration of the secondary blocks. Experience in calibrating the Rockwell hardness machine with standard test blocks has empha-

TABLE III—ROCKWELL HARDNESS TESTS ON ALLOY G BRASS SHEET SHOWING EFFECT OF USING UNPOLISHED AND POLISHED ANVIL

Sample		Tensile Strength, lb. per sq. in.	Unpolished Anvil, "B" Scale, 1/16-in. Ball, 100-kg. Load (Red Figures)				Polished Anvil, "B" Scale, 1/16-in. Ball, 100-kg. Load (Red Figures)			
B. & S. Gage	Temper, B. & S. Numbers Hard		Average of 15 Readings	Maximum	Minimum	Range	Average of 15 Readings	Maximum	Minimum	Range
No. 20	2	64,000	73.3	74.0	72.8	1.2	72.9	73.6	71.8	1.8
	4	77,800	83.5	83.8	82.9	0.9	83.5	83.7	83.0	0.7
	6	85,400	87.0	87.6	86.2	1.4	86.7	86.7	86.0	0.7
	8	93,800	90.3	90.0	90.0	0.8	90.7	90.8	90.3	0.5
	10	99,000	92.1	92.7	91.3	1.4	92.7	92.8	92.6	0.2
No. 22	2	65,100	74.2	75.0	73.6	1.4	72.6	73.5	71.3	2.2
	4	79,100	83.4	83.8	83.2	0.5	83.8	84.2	83.0	1.2
	6	88,200	88.3	88.9	88.0	0.8	88.0	88.3	87.8	0.5
	8	99,300	90.0	90.2	89.7	0.5	90.1	90.2	89.8	0.4
	10	98,200	91.8	92.2	91.7	0.5	92.2	92.4	91.8	0.6
No. 24	2	61,000	67.6	68.8	67.1	1.7	63.9	65.3	61.6	3.7
	4	81,600	84.3	84.7	83.7	1.0	82.7	83.0	82.2	0.8
	6	91,400	89.0	89.7	88.7	1.0	87.8	88.0	87.6	0.4
	8	96,600	91.0	91.2	90.7	0.5	89.7	89.9	89.3	0.6
	10	97,800	91.9	92.7	91.2	1.5	91.2	91.5	90.9	0.6
No. 26	2	58,100	64.4	65.9	63.0	2.9	58.4	59.4	56.8	2.6
	4	71,800	77.3	77.8	76.9	0.9	75.1	75.5	74.4	1.1
	6	88,600	87.4	88.0	86.7	1.3	86.2	86.5	85.8	0.7
	8	95,600	90.2	90.7	89.9	0.8	89.5	89.9	89.2	0.7
	10	99,100	91.2	91.7	90.8	0.9	91.0	91.3	90.6	0.7
No. 28	2	61,100	65.4	66.9	64.2	2.7	55.7	57.6	52.3	5.3
	4	71,400	77.0	77.9	76.0	1.7	72.1	73.3	71.1	2.2
	6	83,600	83.2	83.8	82.1	1.7	81.2	81.5	80.7	0.8
	8	94,700	89.6	90.3	89.1	1.2	88.2	88.5	88.0	0.5
	10	100,100	91.5	92.0	90.5	1.5	90.4	90.8	90.1	0.7

sized the need for keeping these blocks clean and free from oxide, grease or other accumulated matter. This applies with equal force to the test specimens. Various methods of cleaning the test blocks have been tried. A successful and convenient method is to rub the surfaces of the test block by hand with chiffon velvet or other lap material dipped in tripoli and water. Test blocks cleaned in this way do not show a greater variation in readings than that of the original test for uniformity.

A variation in hardness must be permitted in these five readings and on the basis of experience with a large number of blocks it is considered that test blocks should have no greater variation than as follows: When tested using the "B" scale, $\frac{1}{16}$-in. ball, 100-kg. load (red figures); for blocks under 40, 3 points variation is allowed; for

blocks from 40 to 60, $2\frac{1}{2}$ points; and for blocks over 60, $1\frac{1}{2}$ points. Acceptance or rejection of blocks is based on these limits, except where a minus reading results and in this case the 60-kg. load is used. Blocks of this uniformity are calibrated for use with the 60, 100 and 150-kg. loads.

Rockwell Scales Employed.—It has been found that the Rockwell "B" scale employing the 100-kg. load, the $\frac{1}{16}$-in. diameter ball and reading the red figures, is satisfactory for rolled brass sheet; but in the case of the nickel-silver and phosphor-bronze alloys it was found that the Rockwell hardness-tensile strength curve became asymptotic, showing very little change in hardness for a large increase in tensile strength in the harder tempers. A load of 150 kg. was substituted for the 100-kg. load and resulted in an improvement since by using the larger load, a greater depth of penetration was obtained and consequently better sensitivity in the higher tempers. The curve, shown by Fig. 13, gives the Rockwell hardness-tensile strength relationship employing the 100 and 150-kg. loads on the Rockwell tester. The 150-kg. load has been adopted for nickel-silver and

TABLE IV

DIAMETER[a] OF AVERAGE GRAIN FOR ANNEALED BRASS AND NICKEL SILVER

Material	Anneal	Average Tensile Strength, lb. per sq. in.	Rockwell Hardness, "B" Scale, $\frac{1}{16}$-in. Ball, 60-kg. Load, Maximum	Diameter of Average Grain m.m.	
				Minimum	Maximum
High-Brass	Light	50,000	80	0.010	0.035
	Drawing	47,000	65	0.035	0.090
	Soft Drawing	43,000	55	0.090	0.200
Alloy C Brass	Light	42,000	76	0.010	0.030
	Drawing	40,000	61	0.030	0.070
	Soft Drawing	39,000	55	0.070	0.150
Alloy D Brass	Light	53,000	82	0.010	0.030
	Drawing	49,000	72	0.030	0.075
	Soft Drawing	45,000	59	0.075	0.200
Alloy E Brass	Light	50,000	80	0.010	0.035
	Drawing	46,000	67	0.035	0.080
	Soft Drawing	44,000	57	0.080	0.160
Alloy A Nickel Silver	Light	55,000	87	0.010	0.025
	Drawing	54,000	80	0.025	0.050
	Soft Drawing	52,500	77	0.040	0.100

[a] The method for determining the grain size of the material is the Jeffries method as described in the note under Section 9 of the Society's Standard Rules Governing the Preparation of Micrographs of Metals and Alloys (E 2–27), see 1927 Book of A. S. T. M. Standards, Part I, p. 778.

phosphor-bronze alloys and the Rockwell hardnesses reported hereinafter for these materials are the values obtained using a 150-kg. load, a $\frac{1}{16}$-in. diameter ball and reading on the "B" scale or red figures. The 60-kg. load is used for testing annealed material. (See Table IV.)

Fatigue Tests:

The endurance limit of each of the metals covered by this paper has been determined for the annealed condition and as rolled four and ten numbers hard. The results of these tests are covered by another paper.[7]

Bend Test Machine:

In connection with the Amsler bend test machine, an improvement on this machine has been developed employing a motor drive and accurately prepared and aligned jaws. Figure 4 shows a model of

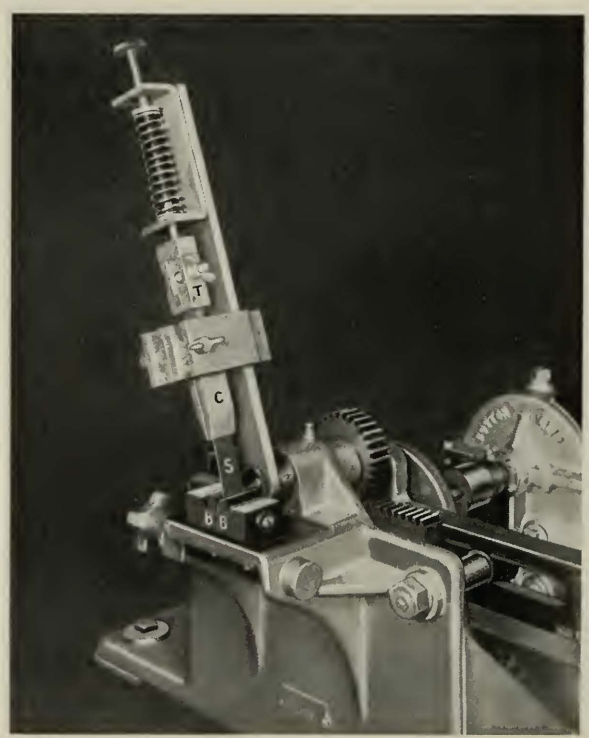


Fig. 4—Experimental Bend Testing Machine.

T is the tensioning grip. *S* is the specimen.
C is the guide clamp. *B* is the radius block.

[7] J. R. Townsend and C. H. Greenall, "Fatigue Studies of Non-Ferrous Sheet Metals," presented before the American Society for Testing Materials in June, 1929

this machine built for experimental purposes. A strip of metal *S*, $\frac{1}{2}$ in. wide, is clamped between a pair of radius blocks, *B*, selected at random. The upper portion of the specimen is held by a tensioning grip, *T*, and guided by two clamps, *C*. The specimen is bent back and forth over the pair of radius blocks and each bend of 90 deg.

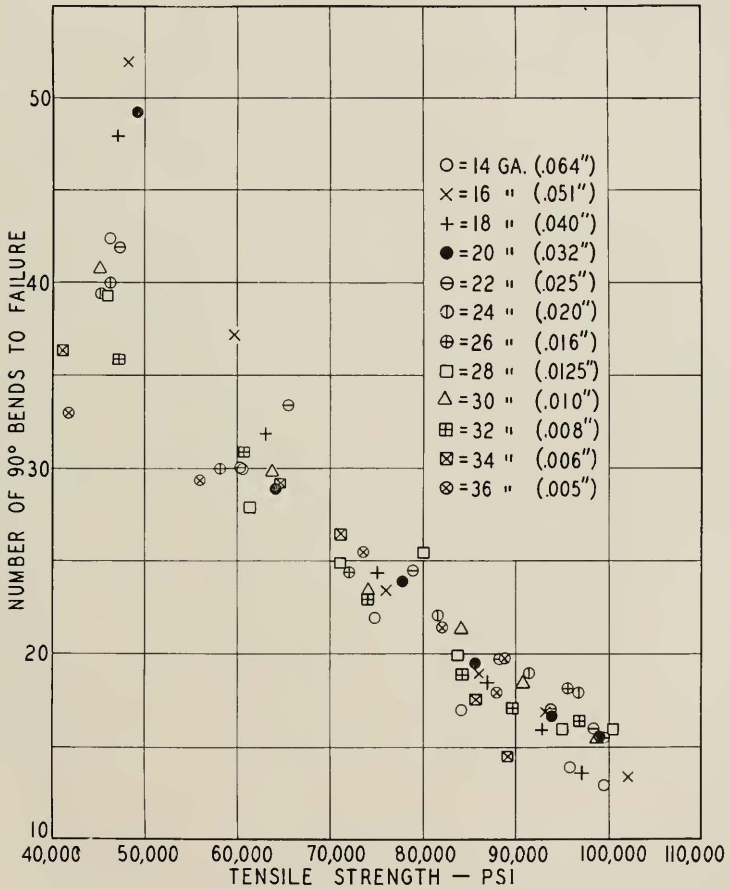


Fig. 5—Relation of Number of 90-deg. Bends to Failure to Tensile Strength of Alloy G Brass. Ratio of Mandrel Radius to Thickness, 8 to 1.

and return is recorded as one bend by a counter. Using another strip of the same metal but a pair of radius blocks of different radii another result is obtained. If the number of bends is plotted on log paper against the ratio of the thickness of the metal to the radius of the blocks, a straight line results. It is possible therefore to select any standard ratio of block radius to thickness of metal and

by two tests employing two different pairs of radius blocks to determine the number of bends for this predetermined ratio by interpolation. The ratio finally selected as most representative was 8 to 1. Figure 5 gives the bend test results for alloy G brass.

The bend test reveals the ductility and toughness of the metal under test. It should show a close correlation to the forming and drawing properties of sheet metal. Further studies will be required to work out such relationships.

Grain Count:

A series of photomicrographs of annealed brass has been adopted covering the grain sizes 0.010 to 0.200 mm. diameter of average grain.[8] Inasmuch as the alpha phase grain structure of nickel silver is similar to that of brass, one set of standards may be employed for these metals. Those standards were carefully counted by the Jeffries method as described in the note to Section 9 of the A. S. T. M. Standard Rules Governing the Preparation of Micrographs of Metals and Alloys (E 2–27) [9] and were photographed at a magnification of 75 diameters. The grain size of a specimen under examination is estimated by comparing the specimen with the standard photomicrographs. For alloy A nickel silver, a magnification of 150 is employed because of the smaller grain size of this material compared with brass. The limits of diameter of average grain of the alloys used for drawing purposes are given in Table IV, namely for high brass sheet, alloys C, D, and E brass, and alloy A nickel silver. (See Table VI for chemical composition.) The grain size limits are based on the range of commercial annealing practice. Inasmuch as these annealed materials are employed only for drawing purposes, no tensile strength limits are necessary. Maximum Rockwell hardness limits are given to exclude cold-worked metal.

Determination of Limits

High and Clock Brass:

Three separate rolling series were used for the round-robin hardness tests. The first of these consisted of a group of materials so arranged as to give a complete range of the hardnesses desired. The second was high brass with a nominal composition of 65 per cent copper and 35 per cent zinc. The third was nickel silver with a nominal composition of 64 per cent copper, 18 per cent zinc, and 18

[8] Report of Committee E-4 on Metallography, presented before the American Society for Testing Materials in June, 1929.

[9] 1927 Book of A. S. T. M. Standards, Part I, p. 778.

per cent nickel. Each of these series included eleven tempers ranging from annealed material to the hardness resulting from 10 B. & S. gage numbers reduction. The samples were approximately 6 in. square and $\frac{1}{8}$ in. thick.

Complete series of samples comprising a fairly wide range of thicknesses were obtained for high-brass sheet in order to thoroughly evaluate the methods of tests. These series were prepared in the following B. & S. gages: Nos. 14, 16, 20, 24, 28, 32 and 36. In each gage the samples had the following hardnesses: Annealed and 2, 4, 6, 8 and 10 B. & S. gage numbers reduction. These rolling series, therefore, included forty-two samples covering the usual commercial range of hardness and the range of thickness of greatest importance in telephone apparatus. These series were made under regular mill conditions with careful supervision to insure representative rolling practice and with complete records of the operations, special attention being given to accurate records of the percentage of reduction. Care was taken to have the metal given approximately a 50 per cent reduction before annealing to insure a uniform grain growth. The temperature of annealing was equivalent to about 600° C. for one-half hour, giving material of about average properties. These series were rolled from four bars of the following compositions:

	Bar No. 1	Bar No. 2	Bar No. 3	Bar No. 4
Copper, per cent	65.19	64.94	65.10	65.13
Zinc, " "	34.77	35.02	34.85	34.82
Lead, " "	0.01	0.02	0.02	0.02
Iron, " "	0.03	0.02	0.03	0.03

The samples were flat strips 10 ft. long and 6 in. wide.

The rolling series on clock brass were made in the following gages and tempers: Nos. 12, 14, 16 and 18 B. & S. gages and 2, 4, 6 and 8 B. & S. gage numbers reduction. The composition was: Copper 61.63 per cent, zinc 36.75 per cent, lead 1.57 per cent, iron 0.05 per cent. This material is so nearly like high-brass sheet in its properties that the data collected when added to that already available were considered sufficient for the preparation of a set of requirements. The tensile strength-percentage of reduction curve for high brass is shown by Figure 10. The physical properties of high and clock brass are shown by Tables I and II.

As mentioned above, Rockwell hardness limits given have been revised to agree with the more recently established Rockwell standards. These limits have been verified by experience data collected on

TABLE I

PHYSICAL TESTS ON HIGH-BRASS SHEET ROLLING SERIES

B. & S. Gage	Thickness, in.	Numbers Hard	Percentage Reduction by Rolling	Proportional Limit, lb. per sq. in.	Tensile Strength, lbs per sq. in.	Elongation in 2 in. per cent	Scleroscope Hardness, Diamond Magnifier Hammer, One Thickness	Rockwell Hardness, "B" Scale, One Thickness	Meyer's Analysis "A," 1/16-in. Ball. 1 mm. in diameter indentation, kg.	Erichsen Ductility, Large Penetrator, mm.	Erichsen Ductility, Small Penetrator, mm.	Olsen Ductility, Large Penetrator, in.	Olsen Ductility, Small Penetrator, in.
No. 36.....	0.0058	Annealed	0	9,500	44,200	40	15	76	80	12.20	3.17	0.462	0.113
No. 36.....	0.0055	2	12.7	33,000	57,200	17	33	76	102	5.64	2.38	0.202	0.064
No. 36.....	0.0058	4	29.25	32,000	67,600	3	41	72	110	4.48	2.17	0.149	0.062
No. 36.....	0.0054	6	46.0	38,000	81,000	1.5	50	76	137	3.27	1.93	0.128	0.062
No. 36.....	0.0057	8	53.6	39,000	86,000	2	54	74	143	3.35	1.88	0.122	0.048
No. 36.....	0.0060	10	62.5	35,500	90,400	1	58	72	151	2.86	1.86	0.107	0.046
No. 32.....	0.0087	Annealed	0	7,900	44,400	53.5	13	48	83	13.24	3.37	0.488	0.116
No. 32.....	0.0078	2	22.0	31,600	60,300	16	36	54	107	5.63	2.41	0.209	0.060
No. 32.....	0.0090	6	43.75	29,200	81,200	2	47	59	134	4.30	2.37	0.153	0.061
No. 32.....	0.0085	8	60.8	40,800	91,800	1.5	53	64	160	4.00	2.10	0.135	0.057
No. 28.....	0.0127	Annealed	0	8,300	44,500	60	9	13	62	13.69	3.73	0.542	0.129
No. 28.....	0.0144	2	10.0	38,800	58,000	25.5	24	60	104	7.21	2.94	0.262	0.087
No. 28.....	0.0135	4	37.8	30,500	75,900	5	31	78	132	6.11	2.76	0.209	0.074
No. 28.....	0.0127	6	49.2	40,300	84,400	2	38	82	142	5.16	2.52	0.188	0.059
No. 28.....	0.0134	8	59.1	34,300	89,600	2	49	86	150	4.37	2.27	0.144	0.054
No. 28.....	0.0136	10	66.0	44,300	95,900	1.5	51	89	157	3.94	2.32	0.149	0.050
No. 24.....	0.0221	Annealed	0	12,000	46,600	58	10	16	66	13.62	3.79	0.532	0.119
No. 24.....	0.0204	2	18.4	23,600	58,300	29	25	60	101	8.51	3.22	0.308	0.089
No. 24.....	0.0199	4	39.3	32,000	77,200	6	33	79	133	6.37	2.87	0.233	0.071
No. 24.....	0.0202	6	49.5	31,000	85,500	4	40	83	142	5.44	2.65	0.204	0.058
No. 24.....	0.0210	8	59.4	31,300	90,500	2	44	86	150	4.59	2.35	0.178	0.052
No. 24.....	0.0211	10	67.6	30,000	95,600	2	50	87	155	4.07	2.26	0.163	0.046

TABLE I—*Continued*

B. & S. Gage	Thickness, in.	Numbers Hard	Percentage Reduction by Rolling	Proportional Limit, lb. per sq. in.	Tensile Strength, lbs. per sq. in.	Elongation in 2 in. per cent	Scleroscope Hardness, Diamond Magnifier Hammer, One Thickness	Rockwell Hardness, "B" Scale One Thickness	Meyer's Analysis "A," $\frac{1}{16}$-in. Ball. 1 mm. in diameter indentation, kg.	Erichsen Ductility, Large Penetrator, mm.	Erichsen Ductility, Small Penetrator, mm.	Olsen Ductility, Large Penetrator, in.	Olsen Ductility, Small Penetrator, in.
No. 20.....	0.0334	Annealed	0	6,100	46,300	63	9	10	63.5	14.17	4.10	0.559	0.124
No. 20	0.0345	2	13.75	23,200	57,300	32	26	60	103	9.05	3.46	0.348	0.107
No. 20.....	0.0336	4	35.1	23,500	72,100	11	34	77	125	7.52	3.05	0.272	0.076
No. 20.....	0.0333	6	48.9	27,100	83,800	5	48	80	138	6.27	2.48	0.231	0.065
No. 20.....	0.0340	8	58.0	33,300	89,000	4	51	86	147	4.83	1.84	0.175	0.063
No. 20.....	0.0331	10	67.6	33,800	92,700	3.5	56	88	152	3.32	1.90	0.130	0.061
No. 16.....	0.0529	Annealed	0	8,300	45,500	70	10	—4	63	15.00	....	....	....
No. 16.....	0.0524	2	19.6	22,900	59,300	35	36	67	106	10.08	....	....	....
No. 16.....	0.0525	4	35.2	23,100	71,700	9.5	42	79	123	7.50	....	....	....
No. 16.....	0.0523	6	48.7	27,600	80,700	7.5	46	84	137	6.53	....	....	....
No. 16.....	0.0521	8	59.3	30,500	87,900	5.5	51	87	146	4.16	....	....	....
No. 16.....	0.0510	10	68.5	29,600	94,500	5	59	91	155	2.86	....	....	....
No. 14.....	0.0661	Annealed	0	7,900	46,600	63	13	8	66	14.99	....	....	....
No. 14.....	0.0651	2	19.6	23,600	58,900	38	34	68	105	10.05	....	....	....
No. 14.....	0.0662	4	35.1	21,600	71,100	13	36	80	125	8.92	....	....	....
No. 14.....	0.0648	6	49.4	22,000	81,700	8	39	85	137	5.66	....	....	....
No. 14.....	0.0664	8	59.0	30,600	89,300	6.5	47	89	147	2.97	....	....	....
No. 14.....	0.0648	10	67.1	28,300	93,500	6	49	90	154	2.08	....	...	....

TABLE II

PHYSICAL TESTS ON CLOCK-BRASS SHEET ROLLING SERIES

Sample Number	Thickness, in.	Temper	B. & S. Gage Numbers Hard	Reduction by Rolling, per cent	Tensile Strength, lb. per sq. in.	Elongation in 2 in., per cent	Sclero-scope Steel Magnifier Hammer, One Thickness	Rockwell "B" Scale, One Thickness	Meyer's Analysis "A," kg.
502–2	0.081	....	2	20.4	59,250	32	35	68	106
502–4	0.064	....	4	36.6	72,750	9	48	80	125
502–6	0.050	....	6	49.8	83,700	6	54	85	137.5
502–8	0.039	....	8	60.4	90,000	4	58	88	152
340	0.143	Soft	..	....	45,900	62	15	21	67
341	0.081	Quarter Hard	..	....	51,400	40	25	56	86
414	0.040	Half Hard	..	....	61,800	25	35	70	106
412	0.079	"	..	....	64,500	23	40	77	119
342	0.033	"	..	....	70,200	14	39	78	125
190	0.020	"	..	....	64,600	13	..	65	113.5
415	0.040	Hard	..	....	74,300	8	46	82	127
413	0.078	"	..	....	74,800	8.5	45	84	125
191	0.032	"	..	....	74,500	5.5	..	79	123.5

material furnished under the limits by a number of suppliers. Table V shows the Rockwell hardness limits. Figure 6 shows Rockwell hardness plotted against tensile strength for the high-brass rolling

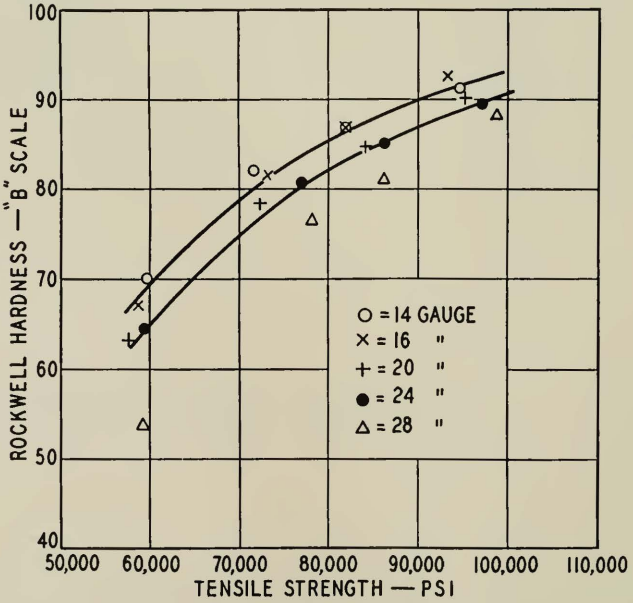


Fig. 6—Relation of High Brass Rolling Series. Rockwell Hardness to Tensile Strength, Using Standard Rockwell Blocks.

TABLE V

REVISED ROCKWELL HARDNESS AND TENSILE STRENGTH LIMITS FOR HIGH, CLOCK AND ALLOY E BRASS SHEET

Thickness	Temper, B. & S. Numbers Hard	Percentage Reduction by Rolling	Tensile Strength,[a] lb. per sq. in.						Rockwell Hardness, "B" Scale, 1/16-in. Ball, 100-kg. Load (Red Figures), for High, Clock and Alloy E Brass Sheet	
			High Brass		Clock Brass		Alloy E Brass			
			Minimum	Maximum	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
0.040 in. and over	Quarter hard. 1	10.9	46,000	56,000	47,000	57,000	47,000	57,000	30	60
0.020 to 0.040 in.		10.9	46,000	56,000	47,000	57,000	47,000	57,000	30	56
0.040 in. and over	Half hard. 2	20.7	53,500	63,500	54,500	64,500	54,000	64,000	50	73
0.020 to 0.040 in.		20.7	53,500	63,500	54,500	64,500	54,000	64,000	49	71
0.040 in. and over	Three-quarters hard. 3	29.5	61,000	71,000	....	....	....	....	70	80
0.020 to 0.040 in.		29.5	61,000	71,000	....	....	....	....	67	77
0.040 in. and over	Hard. 4	37.1	68,000	78,000	68,000	78,000	68,000	78,000	78	85
0.020 to 0.040 in.		37.1	68,000	78,000	68,000	78,000	68,000	78,000	75	83
0.040 in. and over	Extra hard. 6	50.0	79,000	88,500	79,000	88,500	....	....	85	89
0.020 to 0.040 in.		50.0	79,000	88,500	79,000	88,500	....	....	83	87
0.040 in. and over	Spring. 8	60.5	86,000	95,000	85,000	94,500	....	....	88	92
0.020 to 0.040 in.		60.5	86,000	95,000	85,000	94,500	....	....	85	89
0.040 in. and over	Extra spring. 10	68.7	89,500	98,500	88,000	97,500	....	....	89	93
0.020 to 0.040 in.		68.7	89,500	98,500	88,000	97,500	....	....	86	90

[a] Tensile strength values for high and clock brass sheet are the same as in the previous paper by H. N. Van Deusen, L. I. Shaw and C. H. Davis, "Physical Properties and Methods of Test for Sheet Brass," *Proc. A. S. T. M.*, Vol. 27, 1927.

TABLE VI

CHEMICAL COMPOSITION LIMITS FOR NON-FERROUS METAL SHEET

		High Sheet Brass	Clock Brass	Alloy C Brass	Alloy D Brass	Alloy E Brass	Alloy G Brass	Alloy A Nickel Silver	Alloy B Nickel Silver	Alloy A Phosphor Bronze	Alloy C Phosphor Bronze
Copper, per cent	minimum	64.50	61.00	83.00	73.00	64.00	70.50	70.50	53.50	94.40	91.00
	maximum	67.50	64.00	86.00	76.00	67.00	73.50	73.50	56.50	....	....
Lead, per cent	minimum	0.00	1.25	....	....	0.80	....	....	....	....	....
	maximum	0.30	2.00	0.15	0.25	1.10	0.10	....	0.10	0.05	0.02
Iron, per cent	minimum	0.00	0.00	....	....	....	....	....	....	....	....
	maximum	0.05	0.06	0.05	0.05	0.08	0.05	0.35	0.35	0.10	0.10
Zinc, per cent	minimum	Balance	Balance	Balance	Balance	Balance	Balance	8.50	25.50	....	....
	maximum	Balance	Balance	Balance	Balance	Balance	Balance	11.50	28.50	0.30	0.20
Nickel, per cent	minimum	....	....	....	....	....	....	16.50	16.50	....	....
	maximum	....	....	....	....	....	....	19.50	19.50	....	....
Manganese, per cent	minimum	....	....	....	....	....	....	....	....	....	....
	maximum	....	....	....	....	....	....	....	0.50	....	....
Tin, per cent	minimum	....	....	....	....	....	....	....	....	3.80	7.50
	maximum	....	....	....	....	....	....	....	....	4.80	8.50
Phosphorus, per cent	minimum	....	....	....	....	....	....	....	....	0.05	0.05
	maximum	....	....	....	....	....	....	....	....	0.35	0.25
Antimony, per cent	minimum	....	....	....	....	....	....	....	....	....	....
	maximum	....	....	....	....	....	....	....	....	0.01	0.01
Impurities, per cent	minimum	....	....	....	....	....	....	....	....	....	....
	maximum	0.10	0.10	....	....	0.10	0.10	....	....	trace	trace

series retested using standard blocks. There have been no changes in tensile strength limits. Figure 19 shows Rockwell hardness plotted against tensile strength, for high and clock brass, the individual point representing determinations of Rockwell hardness and tensile strength on commercial shipments of material. The lines drawn show the grouping of the data according to thickness.

Alloys C and D Brass Sheet:

The chemical composition limits for alloys C and D brass sheet are given in Table VI.

Alloy E Brass Sheet:

Four rolling series were made from one bar of metal having the analysis shown in Table VII. These series begin at B. & S. gage Nos.

TABLE VII

Chemical Analyses of Rolling Series

	Alloy E Brass	Alloy G Brass	Alloy A Nickel Silver	Alloy B Nickel Silver	Alloy A Phosphor Bronze	Alloy C Phosphor Bronze
Copper, per cent	66.02	71.73	71.31	55.23	95.35	91.84
Lead, per cent	1.08	0.02	....	0.005	0.01	0.02
Iron, per cent	0.03	0.03	0.12	0.06	0.05	0.03
Zinc, per cent	32.87	28.21	10.74	26.27	0.00	0.00
Nickel, per cent	....	0.01	17.62	18.38	0.00	0.00
Manganese, per cent	....	....	0.12	0.11	....	....
Tin, per cent	....	0.00	....	....	4.48	8.08
Phosphorus, per cent	....	....	....	....	0.08	0.03
Graphite, per cent	....	....	0.00	0.00	....	....
Combined Carbon, per cent	....	....	0.013	0.018	....	....

10, 14, 18 and 22 and were rolled 1, 2 and 4 B. & S. gage numbers from commercial anneals in the 500 to 650° C. range. Figure 7 shows Rockwell hardness plotted against tensile strength for alloy E brass.

In a previous paper [10] it was shown that different Rockwell hardness values were required for material 0.040 in. and thicker, and materials less than 0.040 in. thick. This change in limits is dependent upon thickness and is not sharply defined, but it has been found sufficiently accurate for all practical purposes to consider that the division occurs at 0.040 in. In the case of the alloy E brass rolling series results shown by Fig. 7, only one curve was drawn since not sufficient data were available to show this division. The curve practically agrees with the curves given in a previous paper [10] for the

[10] H. N. Van Deusen, L. I. Shaw and C. H. Davis, *loc. cit.*

high-brass rolling series, and also with the data shown on Fig. 19 for commercial shipments of high and clock brass.

The composition of alloy E brass is midway between that of high and clock-brass, and consequently it was to be expected that the Rockwell hardness limits for this material would be the same as for

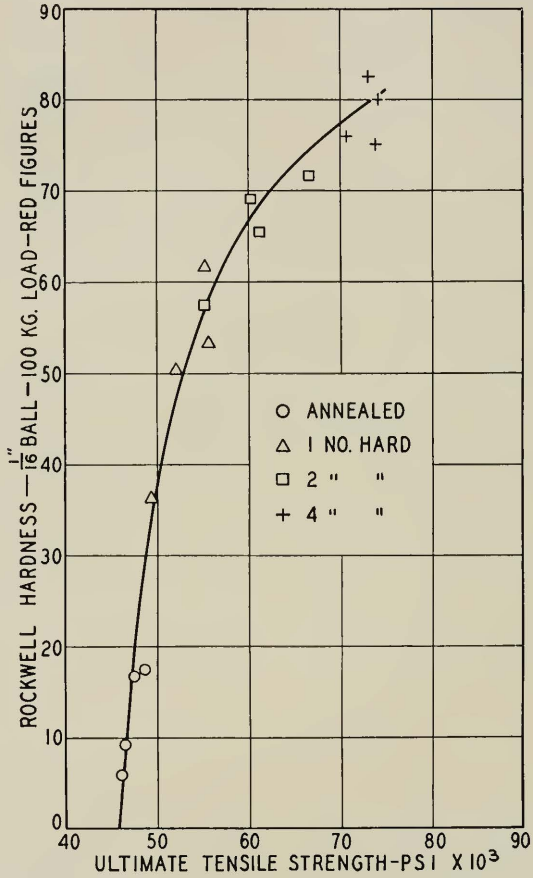


Fig. 7—Relation of Rockwell Hardness to Tensile Strength of Alloy E Brass.

the foregoing. The tension test is more sensitive to such changes in chemical composition and since it is the test upon which acceptance or rejection of material is based, separate tensile strength limits are given. The separate tensile strength and the combined Rockwel hardness limits for these alloys are given in Table V. The physical properties of alloy E brass are given in Table VIII.

TABLE VIII

Physical Properties of Alloy E Brass Sheet

B. & S. Gage	Temper, B. & S. Numbers Hard	Thickness, in.	Percentage Reduction by Rolling		Proportional Limit, lb. per sq. in.	Tensile Strength, lb. per sq. in.	Rockwell Hardness, "B" Scale, 1/16-in. Ball, 100-kg. Load (Red Figures)	Modulus of Elasticity, lb. per sq. in.	Elongation in 2 in., per cent
			At Edges	At Center					
No. 10	0	0.1057	0	0	.....	48,330	17.4	.....	65.8
No. 11	1	0.0903	14.2	15.2	20,050	55,130	61.6	13,740,000	41.0
No. 12	2	0.0817	22.0	23.4	30,417	60,630	71.8	14,600,000	30.7
No. 14	4	0.0657	37.2	38.5	27,230	72,930	82.6	16,540,000	10.8
No. 14	0	0.0656	0	0	.....	46,230	9.1	.....	64.3
No. 15	1	0.0568	10.6	10.6	18,525	52,000	50.3	16,460,000	46.5
No. 16	2	0.0508	23.1	22.8	25,870	60,200	69.2	16,760,000	28.5
No. 18	4	0.0402	38.5	38.5	29,670	74,066	80.1	14,480,000	8.0
No. 18	0	0.0402	0	0	.....	46,130	5.8	.....	67.3
No. 19	1	0.0371	9.2	9.3	18,870	49,600	36.3	.....	52.3
No. 20	2	0.0338	17.4	17.2	20,885	55,230	57.6	15,380,000	37.8
No. 22	4	0.0260	36.7	36.3	23,385	70,530	76.0	17,000,000	11.2
No. 22	0	0.0256	0	0	.....	47,500	16.8	.....	59.0
No. 23	1	0.0225	15.8	14.1	25,410	55,500	53.2	12,340,000	38.7
No. 24	2	0.0205	23.2	21.8	36,460	61,010	65.7	12,300,000	29.0
No. 26	4	0.0169	36.7	35.1	34,090	74,050	75.2	.....	7.5

Note.—Tensile strength and percentage elongation are an average of 3 specimens in each case; Rockwell hardness values are an average of 15 determinations of 5 readings on each of 3 tension test specimens.

It will be noted that the Rockwell hardness and tensile strength limits for alloy E brass are given only for material 1, 2 and 4 numbers hard. This alloy is not normally used in higher tempers as it is especially adapted for use where both forming and machining operations are involved, so that higher leaded or harder materials would not be suitable.

The chemical composition requirements for alloy E brass sheet are given in Table VI.

TABLE IX

PHYSICAL PROPERTIES OF ALLOY G BRASS SHEET

B. & S. Gage	Temper, B. & S. Numbers Hard	Thickness, in.	Actual Percentage Reduction by Rolling	Rockwell Hardness, "B" Scale, 1/16-in. Ball, 100-kg. Load (Red Figures)	Tensile Strength, lb. per sq. in.	Elongation in 2 in., per cent
No. 14	0	0.0655	0	8.5	46,200	68
	2	0.0655	20.6	70.1	60,500	32.5
	4	0.0660	35.0	83.4	74,600	11
	6	0.0658	50.6	88.2	83,800	7.5
	8	0.0655	60.4	92.6	95,600	6.5
	10	0.0655	68.8	93.4	98,600	6.5
No. 16	0	0.0520	0	22.7	48,300	60
	2	0.0522	21.4	69.0	59,700	33
	4	0.0515	37.6	83.7	76,200	9
	6	0.0500	49.9	87.7	86,400	6
	8	0.0503	61.5	91.0	93,000	6.5
	10	0.0512	68.6	93.5	102,000	4.5
No. 18	0	0.0424	0	21.8	48,700	63
	2	0.0409	21.0	73.2	63,400	25.5
	4	0.0412	36.8	81.6	74,900	9.5
	6	0.0408	49.6	87.9	86,800	6
	8	0.0415	59.4	90.1	92,700	4
	10	0.0412	69.6	92.4	96,900	3
No. 20	0	0.0339	0	27.6	48,900	57
	2	0.0333	22.2	73.3	64,000	25
	4	0.0329	37.4	83.5	77,800	7.5
	6	0.0330	49.9	87.0	85,400	5
	8	0.0324	60.7	90.3	93,800	3
	10	0.0320	68.5	92.1	99,000	3
No. 22	0	0.0256	0	15.4	47,100	64
	2	0.0260	23.0	74.2	65,100	22.5
	4	0.0265	39.9	83.4	79,100	7.5
	6	0.0263	50.5	88.3	88,200	2.5
	8	0.0260	60.7	90.0	93,300	3
	10	0.0255	68.8	91.8	98,200	2.5
No. 24	0	0.0201	0	16.0	46,300	61
	2	0.0203	21.1	67.6	61,000	26
	4	0.0204	39.6	84.3	81,600	6
	6	0.0204	53.3	89.0	91,400	3.5
	8	0.0208	59.3	91.0	96,600	2
	10	0.0207	66.6	91.9	97,800	2

TABLE IX—*Continued*

B. & S. Gage	Temper, B. & S. Numbers Hard	Thickness, in.	Actual Percentage Reduction by Rolling	Rockwell Hardness, "B" Scale, 1/16-in. Ball, 100-kg. Load (Red Figures)	Tensile Strength, lb. per sq. in.	Elongation in 2 in., per cent
No. 26	0	0.0169	0	11.0	46,200	58.0
	2	0.0167	17.3	64.4	58,100	26.5
	4	0.0168	34.3	77.3	71,800	9
	6	0.0168	50.9	87.4	88,600	2
	8	0.0165	62.4	90.2	95,600	1.5
	10	0.0172	66.5	91.2	99,100	1.5
No. 28	0	0.0133	0	13.2	45,800	61.5
	2	0.0135	21.1	65.4	61,100	24
	4	0.0134	33.5	77.0	71,400	8
	6	0.0135	49.0	83.2	83,600	2
	8	0.0135	60.1	89.6	94,700	1.5
	10	0.0132	67.7	91.5	100,100	1
No. 30	0	0.0108	0	28.5	45,600	54
	2	0.0103	18.8	55.7	63,600	20
	4	0.0106	36.6	75.7	74,000	4
	6	0.0106	48.7	81.2	84,000	2
	8	0.0108	58.6	86.1	90,600	1.5
	10	0.0108	68.4	90.3	98,600	1
No. 32	1	0.0085	0	....	46,800	51.5
	2	0.0084	17.8	....	60,500	24.5
	4	0.0085	33.6	....	73,900	5
	6	0.0087	46.0	....	84,300	1.5
	8	0.0086	59.4	....	89,700	1
	10	0.0085	63.0	....	96,700	0.5
No. 34	0	0.0071	0	....	41,200	39
	2	0.0065	29.2	....	64,400	20.5
	4	0.0071	39.6	....	70,900	4.5
	6	0.0071	46.9	....	79,800	1.5
	8	0.00705	57.1	....	85,700	1
	10	0.0072	64.5	....	89,200	0.75
No. 36	0	0.0057	0	....	41,600	33
	2	0.0054	16.7	....	55,700	18
	4	0.0056	34.1	....	73,400	4
	6	0.0056	51.5	....	81,900	2
	8	0.0056	60.9	....	87,700	2
	10	0.0052	68.3	....	93,100	1

Note.—Tensile strength and percentage elongation are an average of 3 specimens in each case; Rockwell hardness values are an average of 15 determinations of 5 readings on each of 3 tension test specimens.

Alloy G Brass Sheet:

It has been pointed out in a previous paper [11] that this material, consisting nominally of 72 per cent of copper and 28 per cent of zinc, gives the maximum tensile properties possible with the alpha

[11] W. H. Bassett and C. H. Davis, "Physical Characteristics of Copper and Zinc Alloys," *Proceedings*, Inst. Metals Div., Am. Inst. Mining and Metallurgical Engrs., 1928.

brasses in the cold-rolled condition. A bar of alloy G brass of the chemical composition shown in Table VII was rolled 2, 4, 6, 8 and 10 numbers hard in every even B. & S. gage number from No. 14 to No. 36, inclusive. Table IX gives the physical properties of the

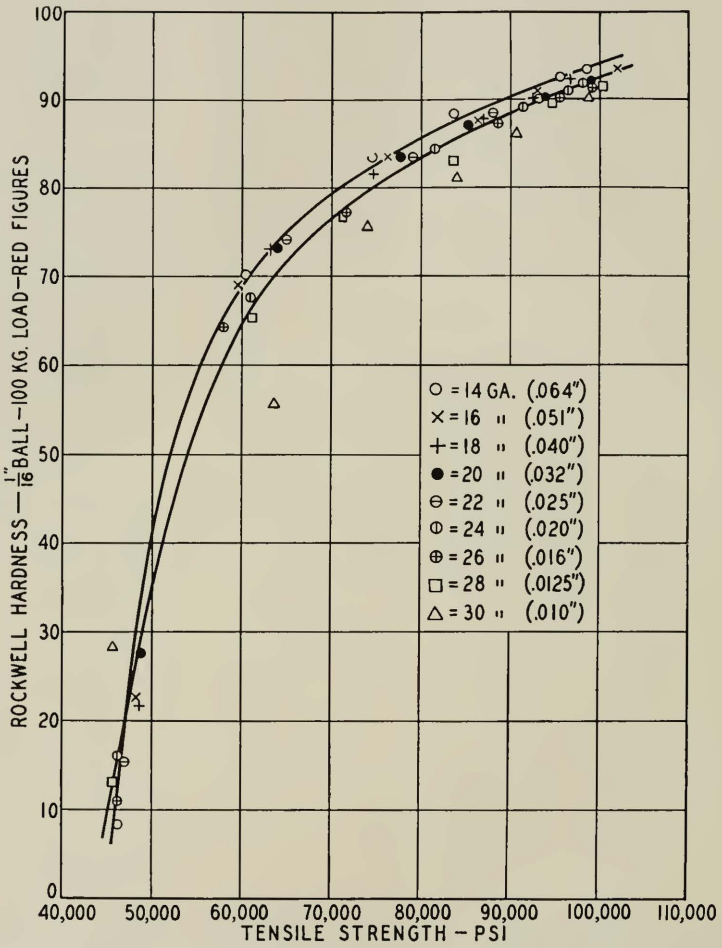


Fig. 8—Relation of Rockwell Hardness to Tensile Strength of Alloy G Brass Rolling Series.

rolling series. The Rockwell hardness-tensile strength curve is shown in Fig. 8 and the tensile strength-reduction curve is shown by Fig. 9. The chemical composition limits are given in Table VI, and the tensile strength and Rockwell hardness limits are given in Table X.

Nickel Silver:

Two nickel-silver alloys were investigated, one containing nominally 72 per cent of copper, 10 per cent of zinc and 18 per cent of nickel, which will be designated "alloy A"; and the other containing 55 per cent of copper, 27 per cent of zinc, and 18 per cent of nickel and designated "alloy B." Alloy A is used mainly for forming and deep drawing purposes, whereas alloy B is used for springs.

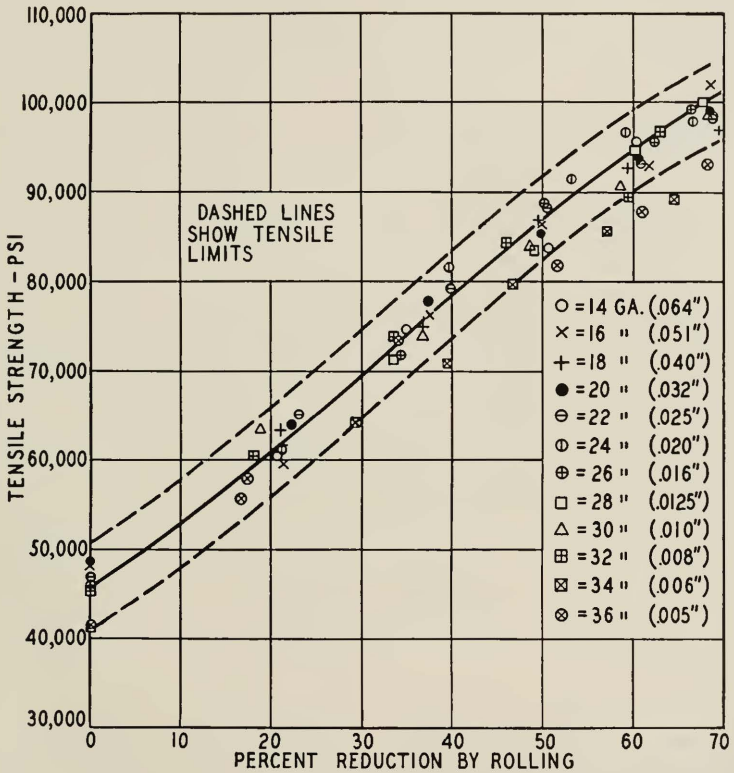


Fig. 9—Relation of Tensile Strength to Percentage of Reduction by Rolling of Alloy G Brass Rolling Series.

Alloy A Nickel-Silver Sheet.—The average composition of three bars of alloy A cast from the same pot of metal are given in Table VII. Table XI shows the physical properties of alloy A. Figure 11 shows the Rockwell hardness-tensile strength relationship and Fig. 12 shows the tensile strength limits plotted against percentage of reduction. The tensile strength and Rockwell hardness limits for this material are given in Table XII and the composition limits in Table VI.

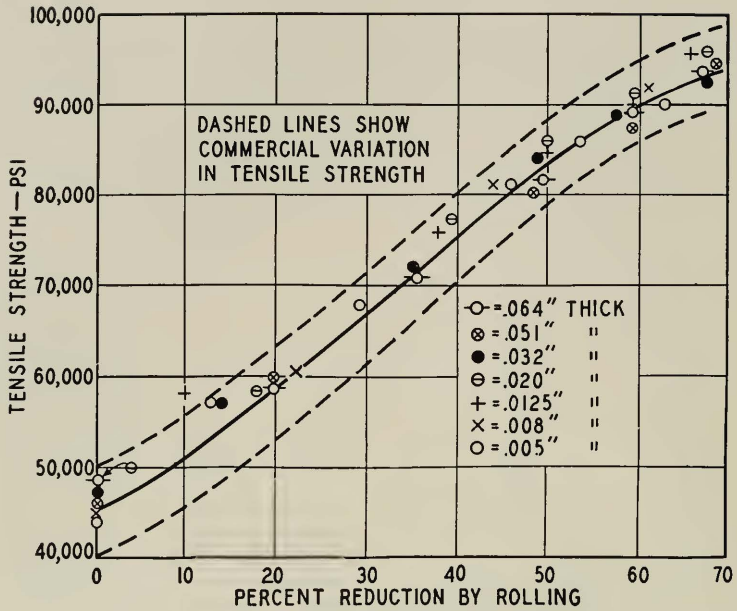


Fig. 10—Tensile Strength *versus* Reduction by Rolling. High-Brass Rolling Series

TABLE X

ROCKWELL HARDNESS AND TENSILE STRENGTH LIMITS FOR ALLOY G BRASS SHEET

Thickness	Temper, B. & S. Numbers Hard	Percentage Reduction by Rolling	Tensile Strength, lb. per sq. in.		Rockwell Hardness, "B" Scale, 1/16-in. Ball, 100-kg. Load (Red Figures)	
			Minimum	Maximum	Minimum	Maximum
0.040 in. and over...	Quarter hard.. 1	11.0	49,000	59,000	35	67
0.020 to 0.040 in....		11.0	49,000	59,000	30	63
0.040 in. and over...	Half hard..... 2	20.7	56,500	66,500	63	76
0.020 to 0.040 in....		20.7	56,500	66,500	57	73
0.040 in. and over...	Hard..... 4	37.1	71,000	81,000	80	86
0.020 to 0.040 in....		37.1	71,000	81,000	77	84
0.040 in. and over...	Extra hard.... 6	50.0	82,500	91,500	87	91
0.020 to 0.040 in....		50.0	82,500	91,500	85	89
0.040 in. and over...	Spring..... 8	60.5	90,500	99,500	90	94
0.020 to 0.040 in....		60.5	90,500	99,500	88	92
0.040 in. and over...	Extra spring... 10	68.7	95,000	104,000	92	96
0.020 to 0.040 in....		68.7	95,000	104,000	90	94

TABLE XI

PHYSICAL PROPERTIES OF ALLOY A NICKEL-SILVER SHEET

B. & S. Gage	Temper, B. & S. Numbers Hard	Thickness, in.	Actual Percentage Reduction by Rolling	Elongation in 2 in., per cent	Rockwell Hardness, "B" Scale, 1/16-in. Ball (Red Figures)		Tensile Strength, lb. per sq. in.
					100-kg. Load	150-kg. Load	
No. 10.....	0	0.1034	0	37.5	37.5	—13.5	54,500
No. 11.....	1	0.0910	7.3	17.5	73.3	37.5	62,200
No. 10.....	2	0.1022	21.1	8.5	80.1	48.7	70,400
No. 10.....	4	0.1037	36.9	5.5	84.0	55.2	79,500
No. 10.....	6	0.1035	49.8	5.0	86.7	59.7	82,900
No. 16.....	0	0.0517	0	36.5	37.3	—16.7	55,100
No. 17.....	1	0.0460	10.6	21.5	69.1	30.1	61,300
No. 16.....	2	0.0516	22.3	7.5	78.7	45.5	69,500
No. 16.....	4	0.0503	39.5	3.5	84.5	56.2	79,700
No. 16.....	6	0.0512	49.8	4.0	86.3	59.0	82,800
No. 22.....	0	0.0257	0	32.0	36.4	....	54,400
No. 23.....	1	0.0301	7.8	20.0	66.6	26.1	60,600
No. 22.....	2	0.0266	17.2	7.0	75.5	41.1	69,100
No. 22.....	4	0.0263	36.1	2.5	81.4	51.0	79,600
No. 22.....	6	0.0263	48.7	2.0	83.3	55.0	83,700
No. 28.....	0	0.0128	0	35.5	23.0	....	54,500
No. 27.....	1	0.0145	11.5	16.0	65.6	....	62,800
No. 28.....	2	0.0132	21.2	4.5	72.2	....	69,200
No. 28.....	4	0.0137	35.6	1.5	77.1	....	80,700
No. 28.....	6	0.0130	47.4	1.0	81.1	....	84,900
No. 30.....	0	0.0105	0	30.5	....	....	54,600
No. 32.....	0	0.0085	0	32.0	....	....	55,100
No. 34.....	0	0.0059	0	28.5	....	....	55,600
No. 34.....	1	0.0077	12.5 17.6	6.0	....	....	65,200
No. 34.....	2	0.0058	25.0 29.4	2.5	....	....	74,400
No. 34.....	4	0.0062	41.2	1.0	....	....	78,800
No. 34.....	6	0.0060	50.8	1.0	....	....	87,900

Note.—Tensile strength and percentage elongation are an average of 3 specimens in each case; Rockwell hardness values are an average of 15 determinations of 5 readings on each of 3 tension test specimens.

Alloy B Nickel-Silver Sheet.—The five bars of alloy B had the average analysis given in Table VII. The rolling series made from these five bars consisted of the tempers 2, 4, 6, 8 and 10 numbers hard for each of the even B. & S. gage numbers from No. 14 to No. 36, inclusive. The physical properties of this material are given in Table XIII. The Rockwell hardness-tensile strength relationship is shown by Fig. 13 and the tensile strength limits plotted against

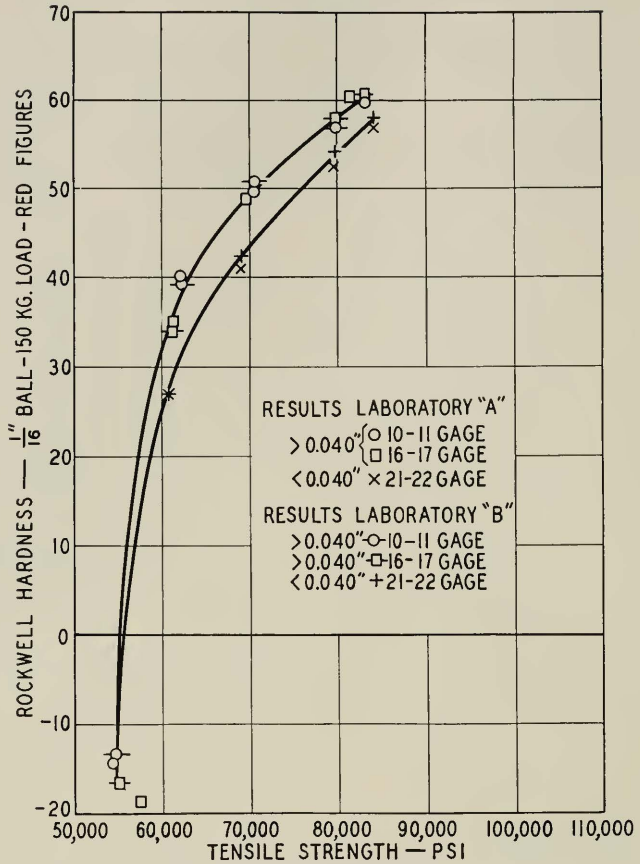


Fig. 11—Relation of Rockwell Hardness to Tensile Strength of Alloy A Nickel Silver Rolling Series.

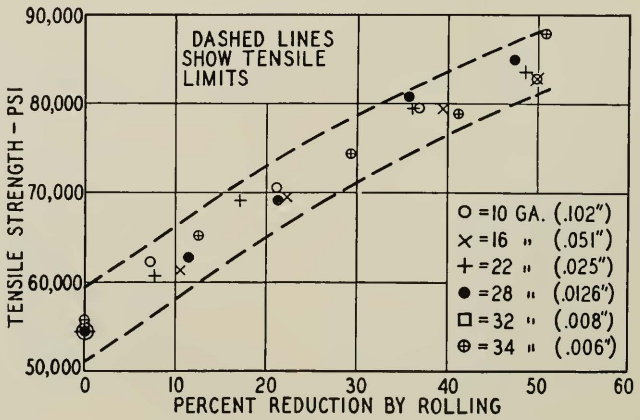


Fig. 12—Relation of Tensile Strength to Percentage of Reduction of Alloy A Nickel Silver Rolling Series.

TABLE XII

ROCKWELL HARDNESS AND TENSILE STRENGTH LIMITS FOR ALLOYS A AND B NICKEL-SILVER SHEET

Thickness	Temper, B. & S. Numbers Hard	Nominal Percentage Reduction by Rolling	Alloy A				Alloy B			
			Tensile Strength, lb. per sq. in.		Rockwell Hardness, "B" Scale, 1/16-in. Ball, 150-kg. Load (Red Figures)		Tensile Strength, lb. per sq. in.		Rockwell Hardness, "B" Scale, 1/16-in. Ball, 150-kg. Load (Red Figures)	
			Minimum	Maximum	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
0.040 in. and over	Quarter hard 1	11.0	58,500	67,000	26	47	.....	.....	..	..
0.020 to 0.040 in.		11.0	58,500	67,000	19	40	.....	.....	..	..
0.040 in. and over	Half hard 2	20.7	65,500	73,500	44	54	78,000	93,000	54	70
0.020 to 0.040 in.		20.7	65,500	73,500	38	48	78,000	93,000	48	67
0.040 in. and over	Hard 4	37.1	75,000	82,000	54	61	92,000	106,500	69	77
0.020 to 0.040 in.		37.1	75,000	82,000	49	55	92,000	106,500	66	75
0.040 in. and over	Extra hard 6	50.0	81,000	88,000	59	64	102,000	115,000	75	82
0.020 to 0.040 in.		50.0	81,000	88,000	54	60	102,000	115,000	72	79
0.040 in. and over	Spring 8	60.5	.....	.....	..	..	108,000	120,000	78	84
0.020 to 0.040 in.		60.5	.....	.....	..	..	108,000	120,000	75	81
0.040 in. and over	Extra spring 10	68.7	.....	.....	..	..	111,000	123,000	80	85
0.020 to 0.040 in.		68.7	.....	.....	..	..	111,000	123,000	77	82

TABLE XIII

Physical Properties of Alloy B Nickel-Silver Sheet

B. & S. Gage	Temper, B. & S. Numbers Hard	Thickness, in.	Actual Percentage Reduction by Rolling	Tensile Strength, lb. per sq. in.	Elongation in 2 in., per cent	Rockwell Hardness, "B" Scale, 1/16-in. Ball (Red Figures)	
						100-kg. Load	150-kg. Load
No. 14.....	0	0.0666	0	66,100	45.5	53.6	9.5
No. 14.....	2	0.0653	22.6	81,300	13.0	87.6	62.2
No. 14.....	4	0.0664	37.2	93,400	5.0	92.8	70.5
No. 14.....	6	0.0670	47.8	103,400	3.0	96.0	76.2
No. 14.....	8	0.0659	58.8	108,600	3.5	97.9	69.0
No. 14.....	10	0.0654	68.3	114,300	3.8	99.1	82.3
No. 16.....	0	0.0513	0	65,100	47.0	51.1	5.3
No. 16.....	2	0.0522	21.6	85,600	12.0	89.6	65.4
No. 16.....	4	0.0506	40.1	97,400	3.0	93.8	70.6
No. 16.....	6	0.0513	51.3	104,500	3.0	96.6	75.3
No. 16.....	8	0.0500	60.8	111,400	3.0	98.9	79.3
No. 16.....	10	0.0510	67.6	114,500	3.0	99.7	80.9
No. 18.....	0	0.0418	0	65,300	45.5	50.6	7.4
No. 18.....	2	0.0413	19.4	81,100	16.0	85.3	59.0
No. 18.....	4	0.0415	37.6	97,400	4.0	93.4	70.7
No. 18.....	6	0.0415	50.9	103,600	2.5	94.9	73.4
No. 18.....	8	0.0398	62.1	110,400	2.0	96.8	77.2
No. 18.....	10	0.0419	66.8	115,000	2.5	98.9	79.1
No. 20.....	0	0.0326	0	67,600	44.0	56.2	14.7
No. 20.....	2	0.0335	19.9	80,700	16.5	85.0	59.6
No. 20.....	4	0.0330	35.4	96,300	3.0	90.9	69.9
No. 20.....	6	0.0325	51.4	107,200	2.0	94.6	74.9
No. 20.....	8	0.0328	61.4	109,500	2.0	95.3	76.3
No. 20.....	10	0.0339	67.8	113,600	2.0	96.7	77.8
No. 22.....	0	0.0286	0	74,600	35.5	67.6	34.5
No. 22.....	2	0.0271	16.5	80,600	19.5	83.4	57.2
No. 22.....	4	0.0259	37.8	98,700	2.0	91.3	70.1
No. 22.....	6	0.0260	49.6	107,000	1.5	94.0	74.2
No. 22.....	8	0.0262	60.6	112,200	1.5	95.5	76.6
No. 22.....	10	0.0264	68.7	113,500	2.0	96.0	77.3
No. 24.....	0	0.0209	0	66,900	42.0	60.6	19.0
No. 24.....	2	0.0210	19.4	94,900	8.5	90.6	68.8
No. 24.....	4	0.0211	34.9	98,700	2.0	90.9	69.0
No. 24.....	6	0.0208	49.1	107,000	1.0	93.4	73.2
No. 24.....	8	0.0211	58.8	112,400	1.5	95.5	76.1
No. 24.....	10	0.0209	68.9	116,200	1.5	96.6	78.5

TABLE XIII—*Continued*

B. & S. Gage	Temper, B. & S. Numbers Hard	Thickness, in.	Actual Percentage Reduction by Rolling	Tensile Strength, lb. per sq. in.	Elongation in 2 in., per cent	Rockwell Hardness, "B" Scale, 1/16-in. Ball (Red Figures)	
						100-kg. Load	150-kg. Load
No. 26....	0	0.0177	0	65,200	41.0	57.9	2.7
No. 26....	2	0.0166	21.9	83,000	11.5	86.3	58.7
No. 26....	4	0.0167	35.8	108,200	2.0	94.2	72.6
No. 26....	6	0.0172	40.3	106,600	1.5	93.3	71.9
No. 26....	8	0.0166	60.2	112,600	1.0	95.4	74.8
No. 26....	10	0.0170	67.4	114,700	1.0	96.2	77.4
No. 28....	0	0.0139	0	64,500	34.0	59.8	3.3
No. 28....	2	0.0136	20.3	83,300	12.5	84.9	50.3
No. 28....	4	0.0126	38.5	102,800	1.5	92.4	63.8
No. 28....	6	0.0128	51.4	116,300	1.5	96.3	72.7
No. 28....	8	0.0133	58.4	111,500	1.0	95.4	73.3
No. 28....	10	0.0133	68.0	116,400	1.0	96.3	74.7
No. 30....	0	0.0106	0	65,300	39.0	56.2	....
No. 30....	2	0.0111	18.4	77,500	14.5	80.6	....
No. 30....	4	0.0112	35.9	95,600	2.5	89.2	....
No. 30....	6	0.0105	46.4	107,500	1.0	94.0	....
No. 30....	8	0.0107	59.1	119,500	1.0	97.1	....
No. 30....	10	0.0106	66.7	115,100	1.0	96.7	....
No. 32....	0	0.0092	0	67,000	38.0	55.1	....
No. 32....	2	0.0087	17.0	76,800	16.0	71.5	....
No. 32....	4	0.0089	34.6	91,200	2.0	83.0	....
No. 32....	6	0.0092	49.1	102,400	1.0	90.4	....
No. 32....	8	0.0084	60.5	112,600	1.0	93.4	....
No. 32....	10	0.0086	66.9	122,200	1.0	97.6	....
No. 34....	0	0.0070	0	66,200	33.0	66.4	....
No. 34....	2	0.0068	23.5	86,200	6.5	75.7	....
No. 34....	4	0.0066	40.0	96,300	2.0	79.1	....
No. 34....	6	0.0069	50.0	100,500	1.0	81.0	....
No. 34....	8	0.0068	61.1	109,200	1.0	83.3	....
No. 34....	10	0.0067	68.3	115,900	1.0	87.2	....
No. 36....	0	0.0055	0	69,400	34.0	78.9	....
No. 36....	2	0.0063	23.0	83,200	8.5	81.5	....
No. 36....	4	0.0058	35.3	96,500	1.5	81.6	....
No. 36....	6	0.0055	45.0	102,700	1.0	83.1	....
No. 36....	8	0.0059	57.7	104,100	1.0	82.9	....
No. 36....	10	0.0057	67.1	111,100	1.0	84.9	....

Note.—Tensile strength values are an average of 3 specimens in each case; Rockwell hardness values are an average of 15 determinations of 5 readings on each of 3 tension test specimens.

percentage reduction are shown in Fig. 14. Table XII shows the Rockwell hardness and tensile strength limits for alloy B. The chemical composition limits are given in Table VI.

Referring to the curve plotted in Fig. 14, showing the tensile strength limits plotted against percentage of reduction, it is seen that one series lies close to the upper curve whereas the remaining rolling

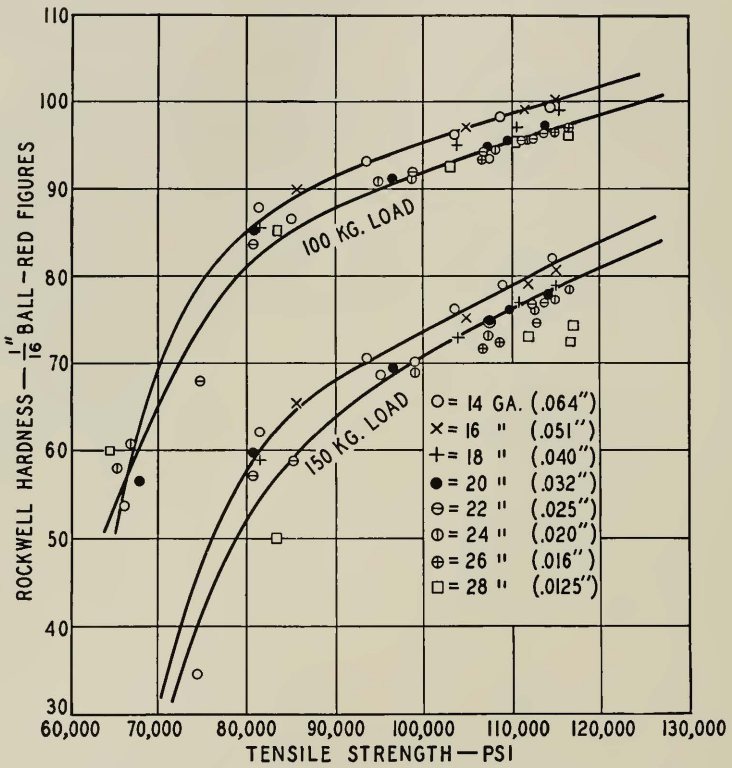


Fig. 13—Relation of Rockwell Hardness to Tensile Strength of Alloy B Nickel-Silver Rolling Series Using 100 and 150-kg. Loads.

series are nearer to the lower curve. This is due to the fact that the bar rolled to No. 22 B. & S. gage sheet had a lighter anneal in the ready-to-finish condition than the other bars, but it was felt that this anneal was close to the commercial range of annealing practice and consequently this series was given some weight when the commercial limits were drawn up.

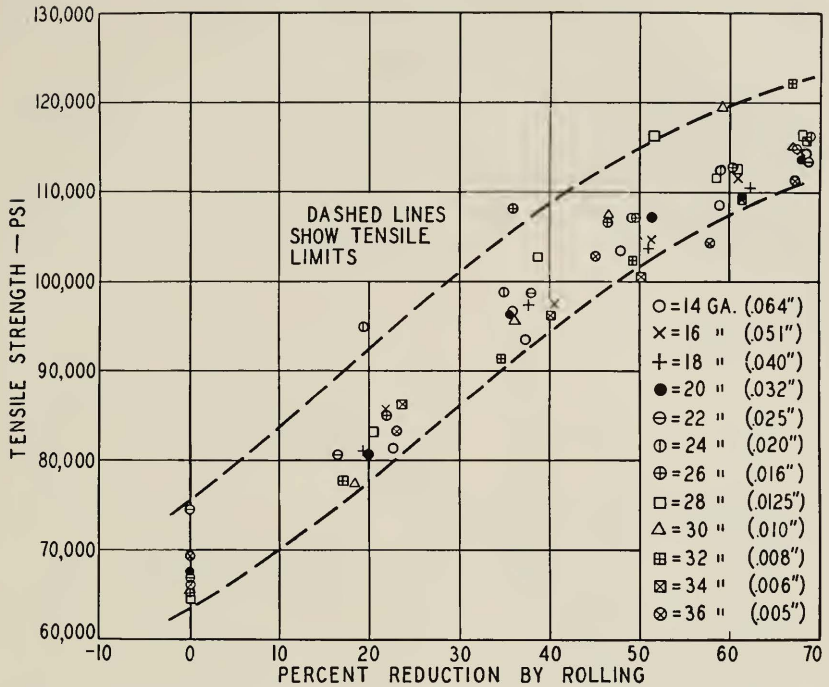


Fig. 14—Relation of Tensile Strength to Percentage of Reduction of Alloy B Nickel-Silver Rolling Series.

TABLE XIV

PHYSICAL PROPERTIES OF ALLOY A PHOSPHOR-BRONZE SHEET

Temper, B. & S. Numbers Hard	B. & S. Gage, in.	Actual Percentage Reduction by Rolling	Tensile Strength, lb. per sq. in.	Elongation in 2 in., per cent	Rockwell Hardness, "B" Scale, 1/16-in. Ball (Red Figures)	
					100-kg. Load	150-kg. Load
Soft	0.0636	0	46,900	53	23.8	—6.9
2	0.0501	21.2	61,200	21.8	75.9	43.3
4	0.0395	37.8	75,150	8.0	84.8	58.3
6	0.0321	49.5	85,700	4.0	88.9	65.2
8	0.0248	61.0	93,900	2.5	91.4	68.8
10	0.0203	68.1	97,450	2.0	92.2	69.7
Soft	0.0324	0	47,000	51	26.6	—7.6
2	0.0242	25.2	62,500	15.5	75.3	43.2
4	0.0199	38.6	73,600	6	80.6	48.0
6	0.0166	48.8	83,950	2.5	84.4	55.0
8	0.0126	61.0	91,900	1.5	88.1	....
10	0.0105	67.6	97,350	1.5	89.3	....
Soft	0.0162	0	46,350	38.5	....	....
2	0.0129	20.4	61,250	15.5	....	....
4	0.0107	33.3	72,800	4.5	....	....
6	0.0082	49.2	84,900	1.5	....	....
8	0.0062	61.7	94,450	1.3	....	....
10	0.0050	68.8	96,400	1	....	....

Note.—Tensile strength and percentage elongation are an average of 3 specimens in each case; Rockwell hardness values are an average of 15 determinations of 5 readings on each of 3 tension test specimens.

Phosphor Bronze:

Two grades of phosphor bronze have been investigated, designated alloys A and C, containing 4 and 8 per cent of tin, respectively.

Alloy A Phosphor-Bronze Sheet.—Three rolling series representative of the entire number investigated were made from the bar of metal of the analysis shown in Table VII. These three series were rolled to all even numbers hard from 2 to 10, B. & S. gage for Nos. 14,

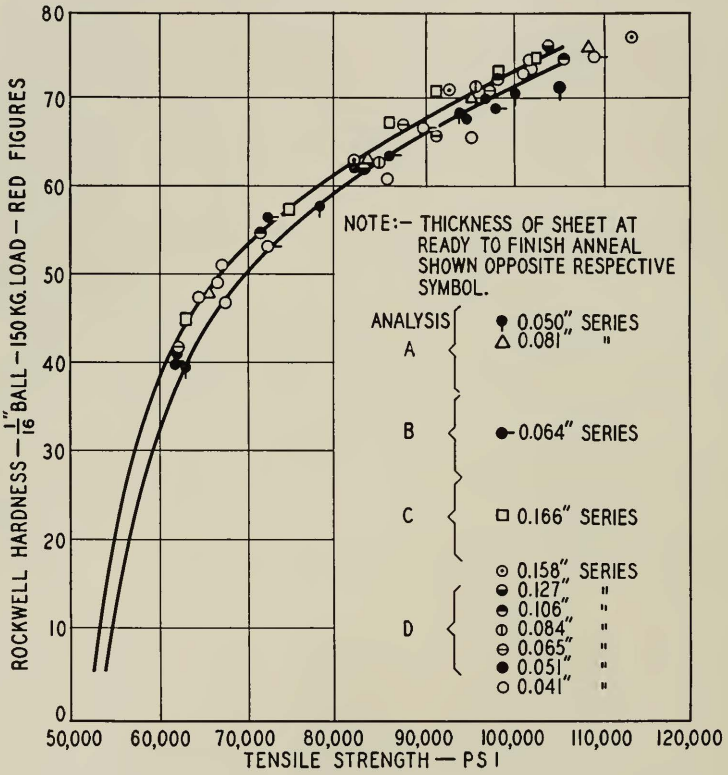


Fig. 15—Relation of Rockwell Hardness to Tensile Strength of Alloy A Phosphor-Bronze Rolling Series.

20 and 28. The physical properties are given in Table XIV. The Rockwell hardness-tensile strength relationship for alloy A is given in Fig. 15 and the tensile strength-reduction relationship is shown in Fig. 16. The Rockwell hardness and tensile strength limits for alloy A are given in Table XV and the chemical composition requirements are given in Table VI.

Alloy C Phosphor-Bronze Sheet.—The alloy C rolling series were made from five bars of metal having the average composition shown in Table VII. This material was rolled to all even hardness numbers from 2 to 10 numbers hard, for all even B. & S. gage numbers from

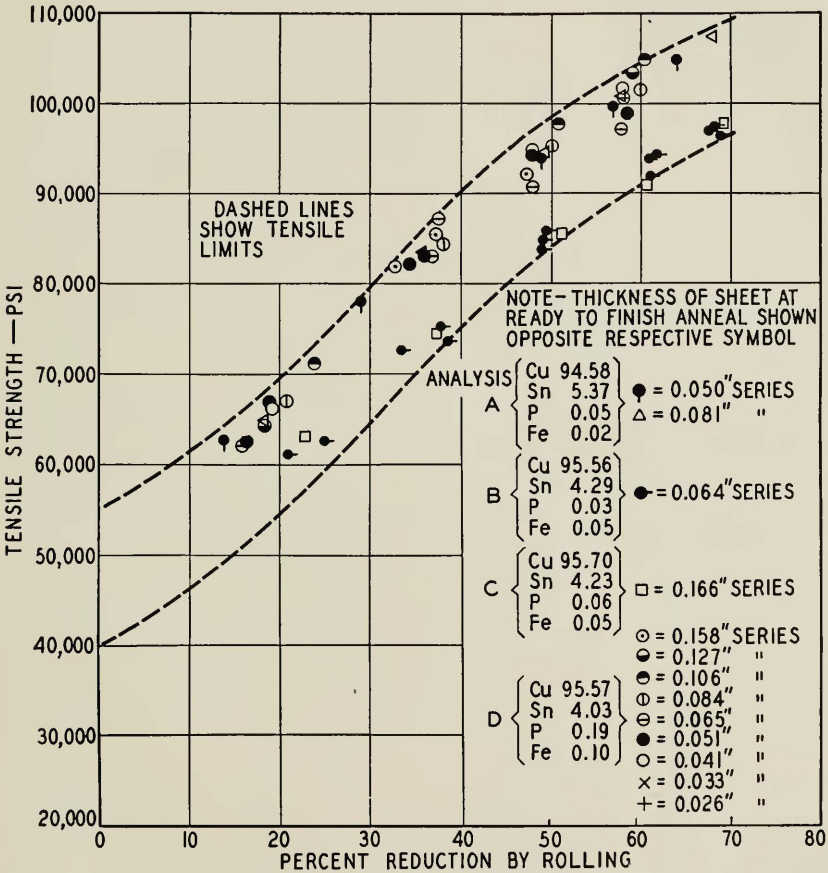


Fig. 16—Relation of Tensile Strength to Percentage of Reduction by Rolling of Alloy A Phosphor-Bronze Rolling Series.

No. 14 to No. 36, inclusive. The physical properties of this material are given in Table XVI. The Rockwell hardness-tensile strength relationship is shown in Fig. 17 and the tensile strength-reduction relationship is shown in Fig. 18. The physical requirements for alloy C are given in Table XV and the chemical requirements in Table VI.

TABLE XV

ROCKWELL HARDNESS AND TENSILE STRENGTH LIMITS FOR ALLOYS A AND C, PHOSPHOR-BRONZE SHEET

Thickness	Temper, B. & S. Numbers Hard	Nominal Percentage Reduction by Rolling	Alloy A Phosphor Bronze				Alloy C Phosphor Bronze			
			Tensile Strength, lb. per sq. in.		Rockwell Hardness, "B" Scale, 1/16-in. Ball, 150-kg. Load (Red Figures)		Tensile Strength, lb. per sq. in.		Rockwell Hardness, "B" Scale, 1/16-in. Ball, 150-kg. Load (Red Figures)	
			Minimum	Maximum	Minimum	Maximum	Minimum	Maximum	Minimum	Maximum
0.040 in. and over	Half hard 8	20.7	55,000	70,000	20	53	69,000	84,000	47	68
0.020 to 0.040 in.	Half hard 8	20.7	55,000	70,000	15	51	69,000	84,000	40	64
0.040 in. and over	Hard 4	37.1	72,000	87,000	55	66	85,000	100,000	69	77
0.020 to 0.040 in.	Hard 4	37.1	72,000	87,000	53	64	85,000	100,000	65	74
0.040 in. and over	Extra hard 6	50.0	84,000	98,500	64	73	97,000	111,500	76	82
0.020 to 0.040 in.	Extra hard 6	50.0	84,000	98,500	62	71	97,000	111,500	73	79
0.040 in. and over	Spring 8	60.5	91,000	105,000	69	75	105,000	118,500	79	85
0.020 to 0.040 in.	Spring 8	60.5	91,000	105,000	67	73	105,000	118,500	76	82
0.040 in. and over	Extra spring 10	68.7	96,000	109,000	72	77	109,500	122,000	81	86
0.020 to 0.040 in.	Extra spring 10	68.7	96,000	109,000	70	75	109,500	122,000	78	83

TABLE XVI

Physical Properties of Alloy C Phosphor-Bronze Sheet

B. & S. Gage	Temper, B. & S. Numbers Hard	Thickness, in.	Actual Percentage Reduction by Rolling	Tensile Strength, lb. per sq. in.	Elongation in 2 in., per cent	Rockwell Hardness, "B" Scale, 1/16-in. Ball (Red Figures)	
						100-kg. Load	150-kg. Load
No. 14.....	0	0.0663	0	62,400	68.0	56.3	14.7
No. 14.....	2	0.0640	22.0	72,800	36.0	83.7	56.0
No. 14.....	4	0.0649	37.5	86,900	17.0	93.0	71.3
No. 14.....	6	0.0659	50.5	100,900	10.0	97.3	78.3
No. 14.....	8	0.0671	59.4	110,300	7.5	99.6	81.6
No. 14.....	10	0.0669	67.7	116,300	6.5	101.4	84.7
No. 16.....	0	0.0515	0	54,700	76.0	42.0	—7.3
No. 16.....	2	0.0536	19.4	80,400	33.0	88.4	63.6
No. 16.....	4	0.0539	35.4	83,800	20.0	91.9	67.7
No. 16.....	6	0.0519	49.8	97,100	9.0	96.4	76.2
No. 16.....	8	0.0535	60.2	107,700	7.0	99.2	80.6
No. 16.....	10	0.0549	67.1	115,700	7.0	101.1	83.2
No. 18.....	0	0.0429	0	60,400	71.0	50.7	7.7
No. 18.....	2	0.0412	20.6	71,200	36.0	82.3	55.0
No. 18.....	4	0.0430	34.9	99,200	11.5	96.0	73.4
No. 18.....	6	0.0431	48.2	95,700	10.0	94.5	73.4
No. 18.....	8	0.0420	59.9	106,500	5.5	97.7	78.4
No. 18.....	10	0.0416	69.3	116,800	4.0	100.2	82.6
No. 20.....	0	0.0336	0	61,200	65.5	50.2	6.7
No. 20.....	2	0.0344	18.6	78,300	34.0	85.4	60.3
No. 20.....	4	0.0332	35.6	87,500	15.0	91.0	69.1
No. 20.....	6	0.0334	49.4	114,200	5.0	98.3	80.4
No. 20.....	8	0.0347	57.9	105,900	5.0	96.1	77.1
No. 20.....	10	0.0332	68.5	113,600	3.0	98.6	80.7
No. 22.....	0	0.0264	0	53,700	74.0	45.8	—0.4
No. 22.....	2	0.0267	22.0	80,200	32.0	85.4	60.8
No. 22.....	4	0.0273	35.5	95,400	12.0	92.9	72.8
No. 22.....	6	0.0263	49.5	102,400	7.0	94.6	75.1
No. 22.....	8	0.0260	61.3	121,000	3.0	99.0	82.9
No. 22.....	10	0.0266	67.7	112,100	3.0	97.3	80.3
No. 24.....	0	0.0221	0	59,700	67.5	52.3	11.0
No. 24.....	2	0.0200	22.4	77,300	26.0	85.8	60.2
No. 24.....	4	0.0219	35.2	95,500	14.0	92.2	71.4
No. 24.....	6	0.0221	48.8	108,500	5.0	95.9	77.1
No. 24.....	8	0.0213	55.1	110,700	3.0	96.9	78.8
No. 24.....	10	0.0216	67.4	124,800	2.0	100.0	84.0

TABLE XVI—*Continued*

B. & S. Gage	Temper, B. & S. Numbers Hard	Thickness, in.	Actual Percentage Reduction by Rolling	Tensile Strength, lb. per sq. in.	Elongation in 2 in., per cent	Rockwell Hardness, "B" Scale, 1/16-in. Ball (Red Figures)	
						100-kg. Load	150-kg. Load
No. 26.....	0	0.0174	0	54,300	69.0	57.3	25.0
No. 26.....	2	0.0176	22.6	73,600	38.0	81.5	52.6
No. 26.....	4	0.0178	32.9	86,800	15.0	89.2	65.9
No. 26.....	6	0.0166	49.9	108,500	4.0	96.1	76.3
No. 26.....	8	0.0173	60.1	116,600	2.0	97.3	78.7
No. 26.....	10	0.0168	65.7	115,700	2.0	97.7	80.3
No. 28.....	0	0.0130	0	57,500	58.0	61.6	26.5
No. 28.....	2	0.0129	23.2	73,400	30.0	82.4	46.4
No. 28.....	4	0.0137	40.1	91,500	11.0	90.8	66.1
No. 28.....	6	0.0128	52.9	105,300	4.0	95.1	69.6
No. 28.....	8	0.0150	55.6	107,000	3.0	95.1	74.1
No. 28.....	10	0.0138	66.4	117,600	1.0	99.0	81.1
No. 30.....	0	0.0110	0	55,400	61.0	65.3	43.1
No. 30.....	2	0.0110	17.6	72,400	27.5	85.2	50.3
No. 30.....	4	0.0111	34.2	86,500	12.5	90.3	54.2
No. 30.....	6	0.0109	53.0	102,200	2.5	94.5	63.1
No. 30.....	8	0.0115	58.8	101,900	1.0	96.1	69.7
No. 30.....	10	0.0114	66.7	111,000	1.5	98.3	74.4
No. 32.....	0	0.0086	0	58,200	52.0	....	....
No. 32.....	2	0.0077	28.6	89,000	8.0	....	....
No. 32.....	4	0.0087	28.6	91,400	9.0	....	....
No. 32.....	6	0.0090	47.6	97,200	3.0	....	....
No. 32.....	8	0.0088	61.3	108,200	1.5	....	....
No. 32.....	10	0.0088	67.0	115,300	1.5	....	....
No. 34.....	0	0.0063	0	63,900	50.0	....	....
No. 34.....	2	0.0063	24.0	81,900	17.5	....	....
No. 34.....	4	0.0061	41.9	99,100	7.0	....	....
No. 34.....	6	0.0063	44.5	111,300	2.5	....	....
No. 34.....	8	0.0064	63.4	110,200	1.5	....	....
No. 34.....	10	0.0066	70.0	118,800	1.0	....	....
No. 36.....	0	0.0054	0	61,300	51.0	....	....
No. 36.....	2	0.0054	16.1	77,900	25.0	....	....
No. 36.....	4	0.0049	35.4	97,900	4.5	....	....
No. 36.....	6	0.0053	52.4	103,000	4.0	....	....
No. 36.....	8	0.0054	58.0	114,800	1.5	....	....
No. 36.....	10	0.0059	68.3	110,500	1.5	....	....

Note.—Tensile strength and percentage elongation are an average of 3 specimens in each case; Rockwell hardness values are an average of 15 determinations of 5 readings on each of the 3 tension test specimens.

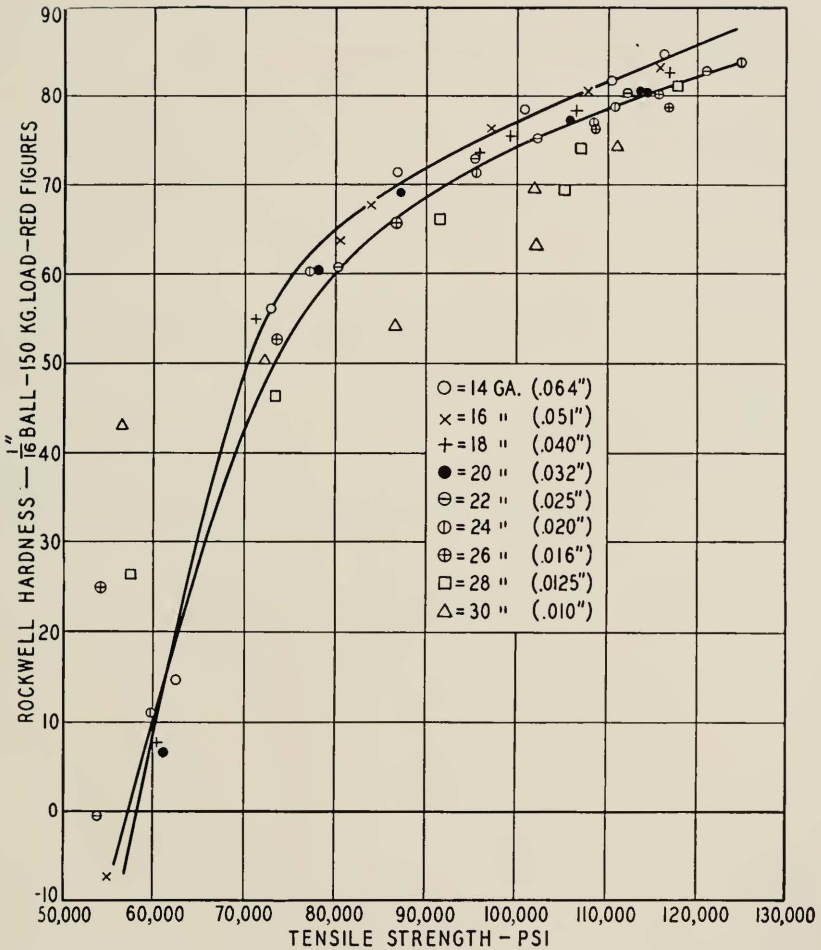


Fig. 17—Relation of Rockwell Hardness to Tensile Strength of Alloy C Phosphor-Bronze Rolling Series.

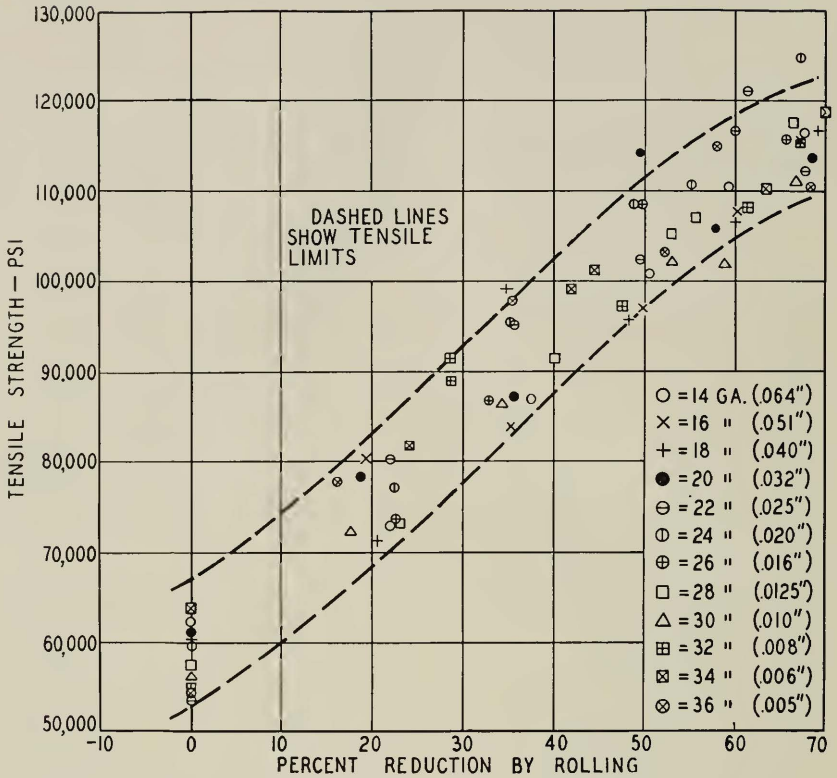


Fig. 18—Relation of Tensile Strength to Percentage of Reduction of Alloy C Phosphor-Bronze Rolling Series.

Experience Data

During and subsequent to the laboratory work covered in this paper, a series of data was obtained on regular commercial shipments of non-ferrous sheet material. About 700 lots of brass, nickel-silver and phosphor-bronze sheet in various tempers and grades were tested on the Rockwell and scleroscope by both the producer and consumer. The lots ranged in size from 100 to about 50,000 lbs. of sheet metal and included material from at least five different mills. The sheet ranged in thickness from 0.010 to 0.500 in. A representative sample about 6 in. long and from 1 to 6 in. wide was taken from each lot. Tests were made on identical samples, five readings being made with each instrument and the average taken. All tests were made by the inspectors who were normally responsible for the control of the material, even though they were not in all cases familiar with the operation of the Rockwell machine. The Rockwell machines were, of

course, put in good mechanical condition and the methods of test outlined in Appendix I were followed. The results obtained constituted an independent verification of the conclusion that the Rockwell machine was more sensitive but much less subject to variation and personal error than the scleroscope.

Figure 19 is representative of the results obtained, showing a

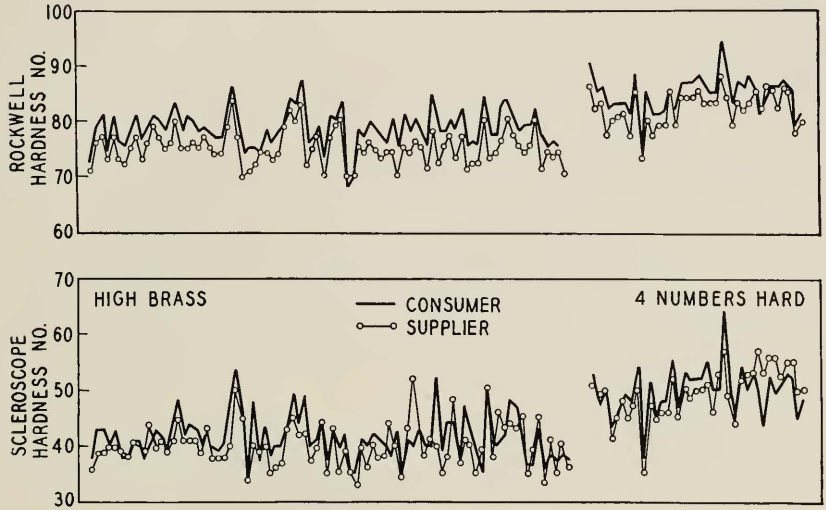


Fig. 19—Typical Comparisons of Rockwell and Scleroscope Machines in Commercial Tests on Brass.

direct comparison of the Rockwell and the scleroscope readings on two tempers of high brass. It illustrates the consistency of the readings of the two Rockwell instruments as compared with the two scleroscopes. Figure 20 represents similar results obtained on nickel

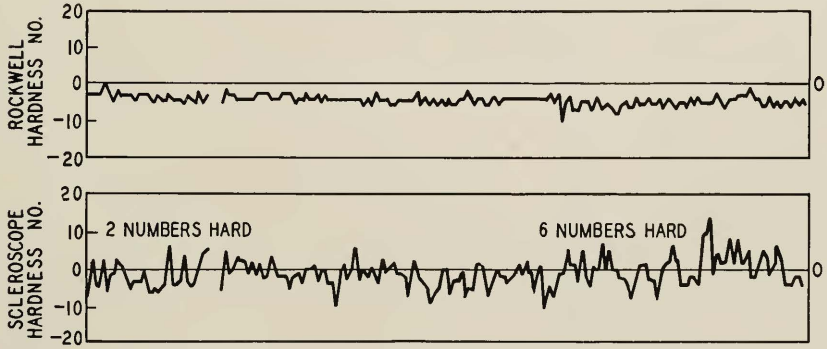


Fig. 20—Typical Results Showing Differences in Rockwell and Scleroscope Machines in Commercial Tests on Nickel Silver.

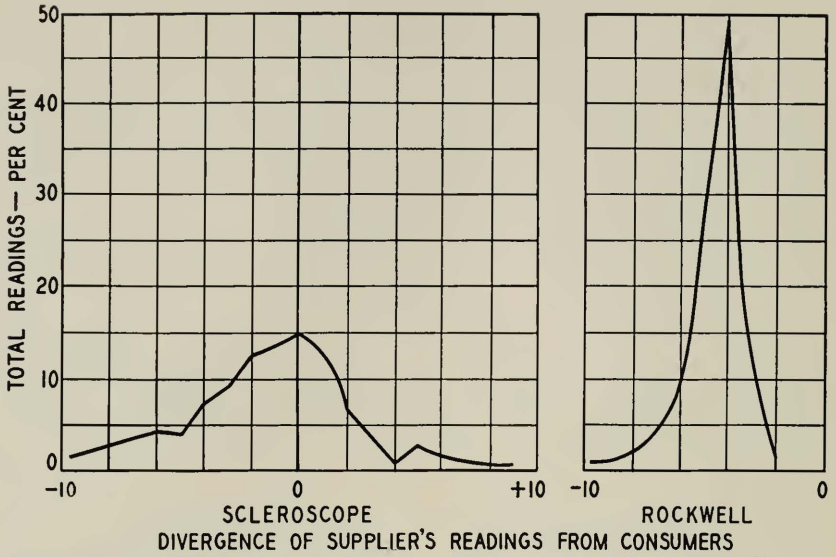


Fig. 21—Typical Results Showing Frequency of Occurrence of Degrees of Difference Between Two Machines.

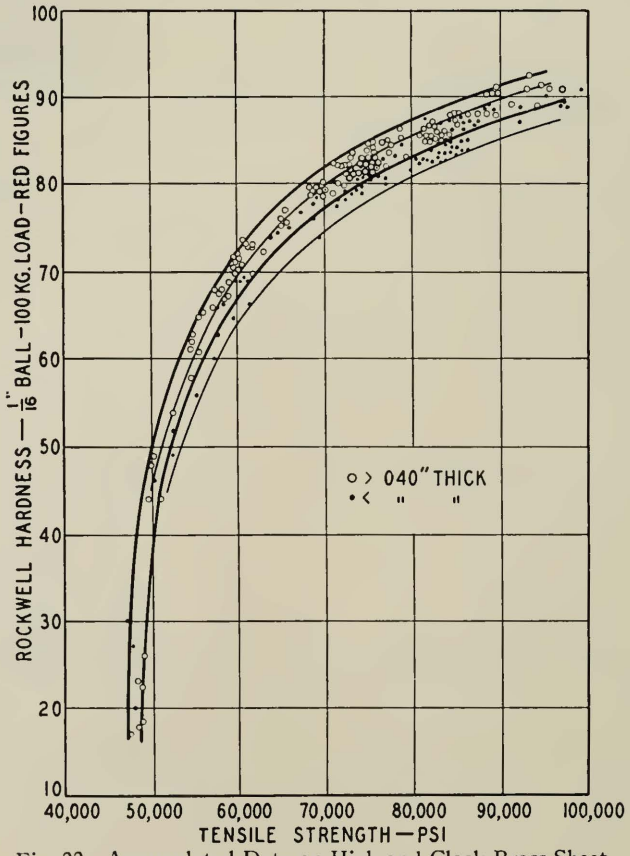


Fig. 22—Accumulated Data on High and Clock-Brass Sheet.

silver plotted to show directly the differences between readings on the two machines of each type, one being assumed as a standard. Figure 21 shows the frequency of occurrence of the various degrees of difference between the two instruments, giving results for both the Rockwell and the scleroscope. One of the instruments was taken as a standard and the variations of the other instrument from it are plotted to left or right as they are negative or positive. The fre-

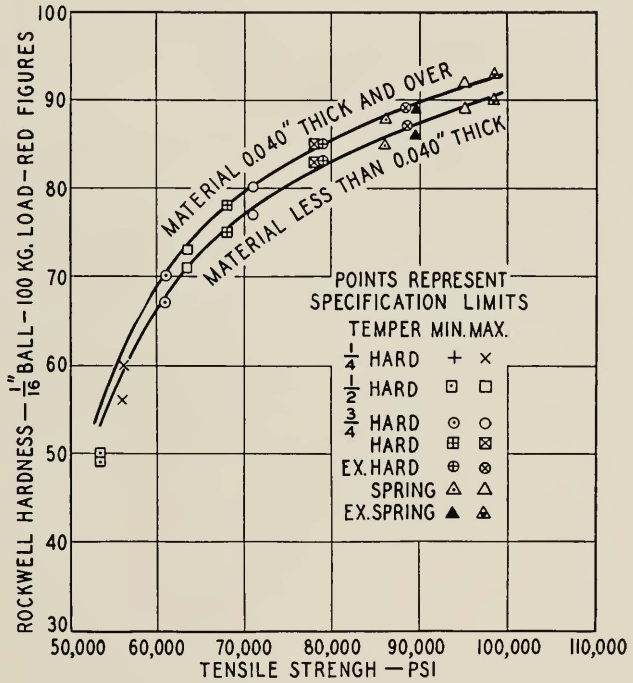


Fig. 23—Comparison of Curves of Results from Experience Data with Limits Adopted, for High-Brass Sheet.

quency of occurrence of differences of the magnitudes represented by the abscissas are plotted as ordinates in per cent of the total readings made on the group of samples for which the curve is drawn. In this typical case it is seen that about 50 per cent of the Rockwell readings from one machine are 4 points below those from the other machine. Application of a conversion factor would, therefore, move the curve to the right 4 points and very close checks between the two machines would result. The scleroscope curve shows that readings

of the two instruments agree on only about 15 per cent of the total, and that in this case the differences are both positive and negative, thus not permitting the application of a conversion factor.

Experience data were also collected for Rockwell hardness and

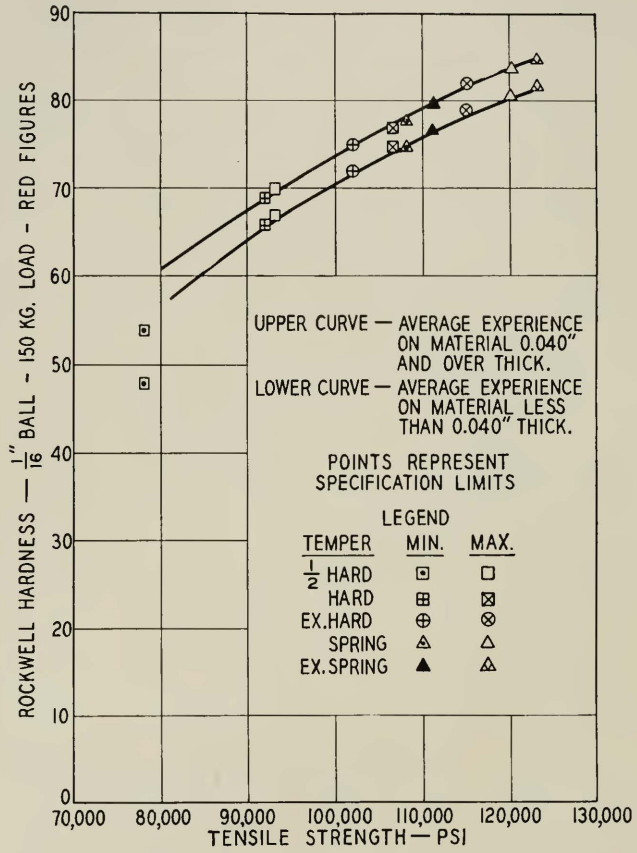


Fig. 24—Comparison of Curves of Results from Experience Data with Limits Adopted, for Alloy B Nickel-Silver Sheet.

tensile strength of the metals, the values being averages of tests on individual shipments.

Figure 22 shows graphically the data obtained on high and clock-brass sheet, and Fig. 23 the experience data which have been averaged and shown as a smooth curve for comparison with the specification

limits shown as points for high-brass sheet. The agreement is close. Figures 24, 25, and 26 show similar smooth curves of experience data for nickel silver and phosphor bronze. In these latter cases the experience data are not sufficiently complete to verify the specification limits in all tempers, but in general the accuracy of the specification limits is verified.

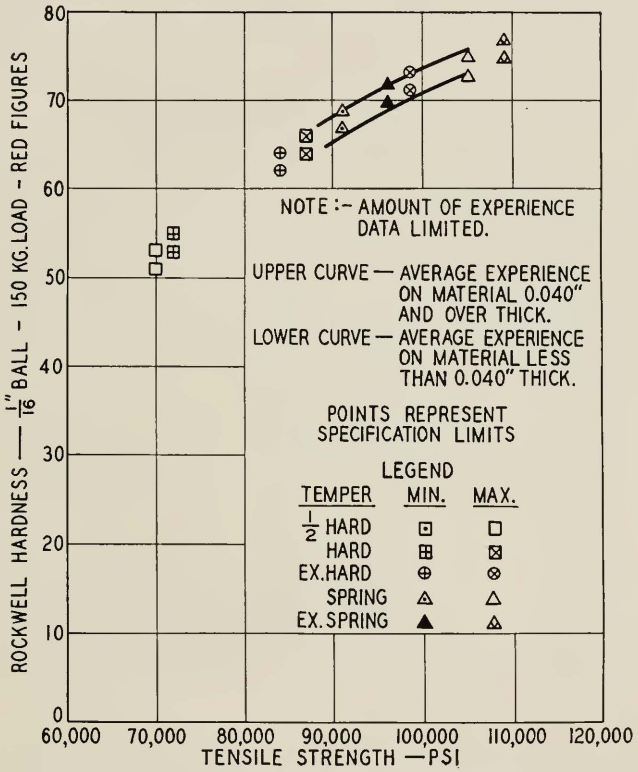


Fig. 25—Comparison of Curves of Results from Experience Data with Limits Adopted, for Alloy A Phosphor-Bronze Sheet.

In addition to the data, the experience of mills supplying material indicates that the limits proposed are satisfactory. Instances where the limits were not met served only to emphasize the advantages of control of the physical properties. Variations from the nominal annealing temperature, composition or percentage reduction in all cases explained the variations from the specification limits.

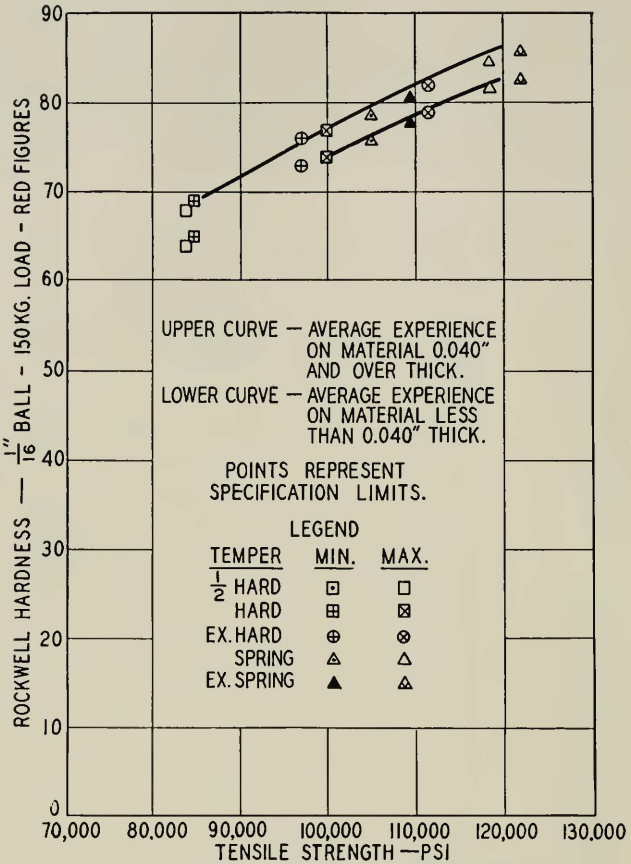


Fig. 26—Comparison of Curves of Results from Experience Data with Limits Adopted, for Alloy C Phosphor-Bronze Sheet.

Conclusion

The purposes stated in this paper have been accomplished, namely, the development of a commercial method of test for sheet non-ferrous metals and limits for use in commercial specifications.

The conclusion reached that the tension test is the best available static test for non-ferrous sheet metal in the hard tempers has been confirmed.

Because of its close correlation to the tension test, the Rockwell hardness test has been adopted for preliminary inspection purposes. Other hardness tests have been studied but this test is the most satisfactory for sheet metals 0.020 in. thick and thicker. It has limited use for material less than 0.020 in. thick. Material near to or outside

of the Rockwell hardness limits is subjected to the tension test. Rejection of material is based solely on the tension test.

Refinements in the application of the Rockwell hardness test as an inspection instrument have been worked out and it is shown that close agreement can be obtained between producer and consumer on commercial shipments of material.

The necessity for the adoption of a series of Rockwell standard blocks covering the range of hardness of certain non-ferrous sheet and the calibration of testing machines to these standards is emphasized.

Grain-size limits are given for annealed brass of four compositions and for one nickel-silver alloy. This is the most satisfactory method of controlling the annealed material.

Tentative tensile strength and Rockwell hardness limits for brass alloy G, nickel-silver and phosphor-bronze alloys have been developed. These limits are considered preliminary until more complete information on actual shipments of material is available. The limits for high and clock brass are considered entirely satisfactory.

A modified bend test has been presented as having considerable value in determining the forming and drawing qualities of sheet metals.

APPENDIX I

STANDARD TEST PROCEDURE FOR ROCKWELL HARDNESS TESTS

Standard Test Blocks

1. Standard test blocks are in the possession of the Bell Laboratories and sub-standard blocks are calibrated from these blocks. These sub-standard blocks are used to calibrate the Rockwell machines. All specifications are written in terms of these standard blocks. Before a Rockwell machine is used for tests on metal supplied to specifications it should be calibrated by the sub-standard block.

Adjustments

2. (*a*) *Dash Pot.*—The dash pot on the Rockwell tester shall be so adjusted that the operating handle completes its travel in from five to ten seconds with no specimen on the machine and with the machine set up to apply a major load of 100 kg.

(*b*) *Index Lever Adjustment.*—As specified in the Rockwell tester instruction book, the following tests (and adjustments, if necessary) should be made.

"Put a piece of material on the anvil and turn the capstan elevating nut to bring the material up against the ball penetrator. Keep turning to elevate the material until the hand feels positive resistance to further turning, which will be felt after the 10-kg. minor load has been picked up and when the major load is encountered. When excessive power would have to be used to raise the work higher, take note of the position of the pointer on the dial. After setting the dial so that C-0 and B-30 are at the top then:

(*1*) If pointer stands between B-50 and B-70 no adjustment is needed.
(*2*) If pointer stands between B-45 and B-50 adjustment is advisable.
(*3*) If it stands anywhere else, adjustment is imperative.

"As the pointer revolves several times when the work is being elevated it is pointed out here that the readings mentioned apply to that revolution of the pointer

which occurs as the reference mark on the gage stem disappears into the sleeve. The object of the adjustment is to see that the elevation of the specimen to pick up the minor load shall not be carried so far as to cause even a partial application of the major load, which to make a proper test, must be applied only through the release mechanism."

"*To Make the Index Lever Adjustment, if Necessary.*—When the test piece is elevated until it starts to pick up the major load, loosen the lock nut of the screw through the index lever which carries at its lower end a small steel plate engaging the ball on the penetrator shaft while using a small screw-driver to firmly hold the screw until after the nut is loose. Then turn the screw very slightly and note result on position of gage pointer which should be at B-60 to B-70 before reclamping the screw with its lock nut."

(*c*) *Protection Against Vibration.*—If the bench or table on which the Rockwell tester is mounted is subject to vibration, such as felt in the vicinity of other machines, the tester should be mounted on a metal plate on sponge rubber at least 1 in. thick or on any type of mounting that will effectually eliminate vibration from the machine. Otherwise the pointer will penetrate farther into the material than when such vibrations are absent.

Cushioning of Latch.—If, when the operating handle is being returned to its normal position, the latch operates with such a snap as to noticeably change the position of the dial pointer, felt or rubber washers should be placed under the trip button in order to cushion this blow. If this snap is severe, difference in reading of several hardness numbers may result.

Testing Methods

3. The anvil used shall have a polished bearing surface for the material of about $\frac{3}{16}$ in. in diameter. It shall be hard enough so that no visible indentation is made when the thinnest material is tested.

Before using the machine it shall be operated several times on a piece of scrap material in order to firmly settle the penetrator, anvil and moving parts of the machine. This should be done every day before the machine is used.

Make a reading on at least one of the standard test blocks. This need not be done every day if it has been found that the machine does not change.

The Rockwell ball should be replaced by a new one occasionally and the operator should be on the lookout for any permanent deformation of the ball which will ordinarily be indicated by high readings on the standard test blocks.

When applying the minor load, the capstan screw should be turned up so that the pointer stops at "O" with a maximum variation of ± five divisions. The last movement of this screw must always be in such a direction as to elevate the specimen.

In applying the major load the operating handle shall be allowed to revolve without interference until the major load is completely applied. This may be observed in two ways, (1) when the pointer suddenly slows down, or (2) watching the weight arm to see that it is completely free from the control of the dash pot. When the major load has thus been completely applied, the operating handle shall be immediately brought back to the latched position without waiting for it to complete its revolution. This should be accomplished in less than two seconds after the major load is completely applied.

The test specimen shall be well supported to prevent errors due to the overhang of the specimen.

All tests shall be made upon one thickness of the material.

APPENDIX II

STANDARD TEST PROCEDURE FOR TENSION TESTING OF THIN SHEET METALS

General

1. (*a*) These specifications cover general methods for testing the tensile strength of thin sheet, strip and flat wire materials for use in the manufacture of telephone apparatus. They shall form part of individual specifications covering sheet metals for specific purposes whenever so provided in the individual specifications.

NOTE.—By thin sheet material is meant material 0.05 in. or less, in thickness. This thickness lies in the range where it becomes necessary to abandon the use of the ordinary micrometer and to use the barrel type micrometer or a still more sensitive device in order to keep the probable error of measurement as low as possible.

(*b*) In case of any difference in the provisions of this and the individual specification, the provisions of the individual specification shall govern.

SPECIMENS FOR TENSION TESTS

2. *Selection of Specimen.*—The test specimens shall be chosen in such numbers and from such locations in each lot of material as to be representative of the quality of this material. The method of selecting specimens will be specified in the individual specifications when found necessary.

3. *Dimensions of Specimens for Tension Test.*—

(*a*) *Materials ½ in. or Less in Width.*—For strip or flat materials less than ½ in. wide, the specimen shall consist of the appropriate length cut from the material as furnished. For gage lengths of 2 and 8 in., the length of the test specimen shall be not less than 8 and 14 in., respectively.

(*b*) *Materials More Than ½ in. in Width.*—For materials over ½ in. wide, two alternative specimens are allowed.

(*1*) *Standard Specimen:*

The specimen which shall be regarded as standard shall have dimensions and be machined as shown in Fig. 27.

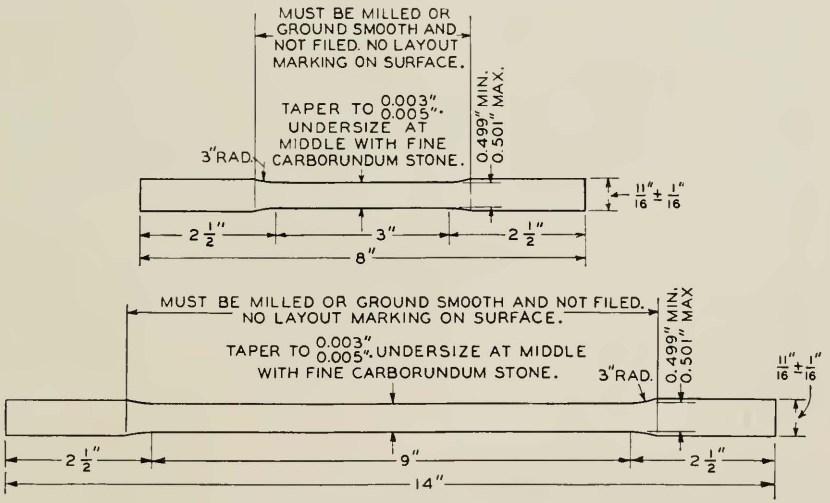


Fig. 27—Tension Test Specimens for Sheet Material Over ½ in. in Width.

(*a*) Specimen for Use in Determining Tensile Strength and Elongation in 2 in.

(*b*) Specimen for Use in Determining Proportional Limit and Modulus of Elasticity.

(*2*) *Alternative Specimen:*

This specimen shall be in the form of a strip up to 1½ in. in width, having parallel sides and having about 0.001 to 0.002 in. of the material removed from the edges of the specimen near the center by grinding with a fine-grained stone. This specimen shall have the lengths given above for materials less than ½ in. wide. It may only be used so long as good breaks (which occur at the point of reduced section and without tearing) are obtained and where the strength of the material does not

approximate the acceptance limits. The milled specimens (see Standard Specimen above) are to be regarded as standard and are to be used where good breaks are not obtained with these wider specimens and as referee test specimens where the strength of the material approximates the acceptance limits.

4. *Cleaning of Specimens.*—Before being measured, each test specimen shall be thoroughly cleaned with carbon tetrachloride applied by rubbing with a soft cloth. When being measured, each specimen shall be wiped clear of dust immediately before being inserted in the measuring instrument.

5. *Measurement of Specimens.*—The dimensions of the specimen shall be measured between flat parallel surfaces of hardened steel, or other appropriate material, approximately $\frac{3}{16}$ in. in diameter which shall be applied in such a manner as to exert on the specimen, while the measurement is being taken, a normal pressure of approximately 16 oz.

6. *Calibration of Measuring Instrument.*—The measuring instrument shall be suitably adjusted and calibrated to indicate the distance between the measuring surfaces of the mandrels when exerting a normal pressure of approximately 16 oz. or 35 lb. per sq. in. of mandrel end area with an error not greater than $\pm$ 0.00005 in., *i.e.*, to the nearest 0.0001 in.

Notes

1. When measuring very thin materials, the micrometer or measuring instrument should be adjusted and calibrated with the greatest practicable accuracy.

2. The measuring instrument should be calibrated preferably by means of a hardened steel gage block having flat parallel surfaces and of a thickness approximately equal to that of the material to be measured.

3. In order to obtain an accurate check, it is essential that the mandrel surfaces and faces of the gage be freed from dust immediately before being brought together.

4. If a barrel type micrometer indicating 0.0001 in. per scale division is used, care should be taken to have the instrument in good operating condition. The instrument can then be adjusted to give the first click of the ratchet at a mandrel pressure of 16 oz. $\pm$ 4 oz. by suitable adjustment of the ratchet spring, measuring the mandrel pressure by means of a small spring balance, having flat parallel contact surfaces, inserted between the mandrels. The micrometer screw and nut shall be so adjusted that no end play or shake of the spindle may be felt when the spindle is gently pushed back and forth in the direction of the axis of the screw.

5. In calibrating the micrometer against a standard gage block, or in measuring a specimen, the spindle shall be rotated slowly into contact with the work by means of the ratchet until the ratchet clicks once, be retracted about 0.002 in. and again slowly rotated into contact until the ratchet clicks once before reading is taken. In order to avoid exceeding the normal pressure of the mandrels in operating the micrometer, the mandrel shall be brought against the work at a rate not exceeding 0.001 in. per second at both first and second contacts given above. This rate is approximately the maximum rate at which the operator is able to count the half thousandths marks on the barrel as they pass the zero line on the frame.

7. *Method of Computing Cross-Section.*—

(*a*) *Material $\frac{1}{2}$ in. or Less in Width.*—For strip material of flat wire $\frac{1}{2}$ in. or less in width, the width and thickness shall be measured along the major axis of the specimen at the middle point and at distances of approximately 2 and 4 in., respectively, at each side of the middle point. The cross-section of strip and flat wire material shall be assumed to be a rectangle. The area of cross-section to be used in computing the tensile strength of the material shall be the product of the average of the five thickness readings and the average of the five width readings.

(*b*) *Cross-Section of Standard Test Specimen for Material Over $\frac{1}{2}$ in. in Width.*—For the standard test specimen described in Section 3 (*b*), the width shall be measured at the point of minimum width as determined by trial measurements. The thickness shall be measured in a plane normal to the major axis of the specimen at the point of minimum width, at the middle and at each edge of the specimen. When measuring the thickness at the edge of the specimen, the outer edges of the contact surfaces of the measuring instrument shall be approximately $\frac{1}{32}$ in. inside of the edge of the material. The area of cross-section shall be taken as the product of the minimum width and the average of the three thickness readings above specified.

NOTE.—In some cases, it may be found that the thickness of the material near the fillets is so much less than the thickness in the plane of minimum width that the minimum area of cross-section occurs near the fillet. In measuring specimens, the operator should be on the lookout for this condition and, in case it is found, the width of the specimen at the point of minimum width should be further reduced by grinding with a stone until the minimum cross-section is brought to the middle of the specimen.

(*c*) *Cross-Section of Test Specimen Over* $\frac{1}{2}$ *in. in Width.*—For specimens over $\frac{1}{2}$ in. wide as described in Section 3 (*b*), the measurements shall be made as on the standard specimen, as described in Paragraph (*b*) above.

METHOD OF TESTING

8. *Testing Machine.*—

(*a*) *Design.*—Testing machines shall be of a design in which the applied load is balanced by increasing the lever arm of a movable weight or by some other means employing weights without the use of springs.

(*b*) *Method of Mounting.*—The testing machine shall be installed on a solid foundation of such a nature that the machine is maintained in a level position.

(*c*) *Range.*—It is desirable that the testing machine should be capable of applying and measuring a load about 50 per cent greater than the normal anticipated breaking strength of the test specimen.

(*d*) *Sensitivity.*—All bearings and knife edges of the testing machine shall be so proportioned and adjusted that the position of the beam, when balanced at the normal breaking load of the specimen, shall be perceptibly changed by a change of load of 0.1 per cent.

(*e*) *Accuracy.*—The machine shall be calibrated with dead weights so that a load equal to the breaking load of the material tested shall be indicated with an error of less than $\pm$ 0.1 per cent.

(*f*) *Method of Applying Load.*—The mechanism for applying load to the test specimen shall be such as to advance the pulling head of the machine at a uniform rate, such as may be obtained by a motor drive operating from a well regulated source of power. The machine shall be capable of being operated at different uniform rates of speed, including the rate of approximately 0.05 in. per minute of the moving head. Unless otherwise specified in individual specifications, the maximum speed of the moving head in making a test shall be 0.025 in. per inch free length of specimen per minute.

(*g*) *Jaws.*—The jaws or grips of the testing machine shall be designed so as to produce and maintain axial alignment of the specimen without producing tearing stresses. The edges of the jaws shall be rounded so as not to noticeably deform the specimen near the borders of the region where the specimen comes in contact with the jaws.

Articulation Testing Methods

By H. FLETCHER and J. C. STEINBERG

This paper is chiefly concerned with the technique of making articulation tests. The construction of a syllabic testing list, the selection of a testing crew, the methods of comparing articulation data for various crews, and the significance of the test as a measure of the speech capabilities of a system are discussed. Various types of lists for different uses are also discussed.

THE transference of thought by means of speech is a very complicated, although common, process. So long as the process runs smoothly, its complications are forgotten. When an auditor fails to understand the speaker, however, inquiry into the reasons for the difficulty begins.

The production, the transmission, and the reception of speech constitute the three important elements of the process. To determine defects in any one of these, it is necessary to have a quantitative means of measuring the recognizability of the speech sounds that the auditor hears. The term "recognizability" as used here refers to correctness with which an auditor identifies a sound as being one, or some combination, of the fundamental speech sounds, when no meaning is associated with the sounds.

During the past few years methods of measuring the recognizability of speech sounds have come into greater and greater use both in this country and abroad. In order to compare the results obtained by various crews in various languages, it is desirable to standardize the methods of test and to set up reference circuits for purposes of calibration. It is the aim of this paper to discuss the methods that have been found the most useful, not only in determining defects in transmission, but defects in the production and reception of speech as well.

One needs only to tabulate the various devices that are used for transmitting speech to realize the importance of a quantitative method of rating their performance. There may be mentioned, for example, the various telephone and radio systems, the phonograph, sound pictures, rooms and auditoriums with various types of acoustic treatment, audiphone sets for the deafened, speaking tubes, etc.

Methods of measuring the recognizability of speech sounds have not been used so extensively for determining the ability of persons to speak properly. Such methods should be of value in the training

of public speakers, actors, students of foreign languages or pupils in deaf schools who are learning to speak.

The rating of auditors by measuring the recognizability of speech sounds which they hear has been used to some extent. For example, such methods have been used to determine the ability of students to interpret a spoken foreign language. Also, the deafness of a person can be determined by such methods. In this case, however, the specialists have usually tried to vary transmission systems between the speaker and listener so as to compensate for the loss of hearing, the amount of such compensation being determined by measuring the recognizability of speech sounds.

The best method of determining the recognizability of the speech sounds naturally depends upon which of the things just enumerated is to be rated. In principle, the method in each case consists in the pronunciation of "selected speech sounds" by a speaker, the transmission of these sounds to an observer's ears, and the recording by the observer of the sounds which he recognizes. Such methods applied to telephone systems have been frequently referred to as articulation tests. The term "articulation" would be more logically used if it were applied only to cases where the speaking abilities of persons are being determined. However, it has been used so frequently in connection with rating transmission systems that it seems convenient to retain it.

The "selected speech sounds" which are ordinarily used in articulation tests are meaningless monosyllables. The percentage of the total number of spoken syllables which are correctly observed is called the syllable articulation. This percentage has frequently been called simply "the articulation."

A syllable is considered to be incorrectly observed, if one or more of the fundamental speech sounds which it contains are mistaken. It is frequently desirable to analyze these mistakes and determine the articulation of the speech sounds. The percentage of the total number of spoken sounds which are correctly observed is called the sound articulation. When the attention is directed toward a specific fundamental sound, such as "b" or "t" or "ā," etc., then the term "individual sound articulation" is used. For example, the individual sound articulation for "b" is the percentage of the number of times that "b" was called that it was observed correctly. Similarly, the terms "consonant articulation" or "vowel articulation" refer to the percentages of the total number of spoken consonant or vowel sounds which are correctly observed.

The articulation values as defined above are taken as the measures

of the recognizabilities of the various phonetic units. English words and short sentences have also been used for testing purposes. When material of this kind is used, a new element enters, namely, the thought or meaning associated with the sentence or word. The criterion for the correct observation of words or sentences is also different from that used in the case of articulation tests. If the thought or meaning of a word or sentence is correctly understood, it is considered to be correctly received, even though the observer may not have correctly recognized each sound that was spoken. The terms "word articulation" or "sentence articulation," therefore, seem inappropriate when referring to the results of such tests. The term "intelligibility" has frequently been used in this sense. Since it has also been used in a more general sense, the terms "discrete word intelligibility" and "discrete sentence intelligibility" will be used when referring to the results obtained by using disconnected words or sentences for the testing material. They are defined as the percentage of the total number of spoken words and sentences, respectively, that are correctly interpreted according to the criterion given above.

Very early in the work of developing the telephone, words and sentences which were chosen in a haphazard way were used for testing purposes. Word lists of various sorts have been worked out and used with some success. Even in very recent years some of these word lists have been used to good advantage. The main objections which have developed, to the continuous use of such lists are: (*a*) it is hard to make the lists equally difficult without resorting to very long lists of words, (*b*) a very large number of lists are necessary in order to avoid memory effects.

Dr. G. A. Campbell [1] was one of the first to propose a system of syllabic speech sounds for testing the transmission characteristics of the telephone system. These syllables had no meaning and were constructed by combining the various initial consonants with the vowel "ee," such as bee, fee, etc. With these lists the consonant articulation was taken as a measure of the system.

Later Dr. I. B. Crandall [2] worked out a system which used both simple and compound consonant forms in a vowel-consonant and consonant-vowel type of syllable. All of the common vowels were used, and the combinations were formed in ways which are usually found in written speech. The sounds occurred with the same frequency as they occur in ordinary written material. As in the Campbell lists, the articulation was based on the consonant sounds alone.

[1] "Telephonic Intelligibility," G. A. Campbell, *Phil. Mag.*, Jan. 1910.
[2] "Composition of Speech," I. B. Crandall, *Phys. Rev.*, 10, p. 74, July 1917.

Several other lists which have not been published were proposed and used, the differences being in the choice of the fundamental speech sounds, in their arrangement into syllables, and in the relative frequency of occurrence, both of the different syllable forms and of the speech sounds in each form. There was a distinct effort to make the lists as nearly like speech as possible by using the syllable forms, and by using the particular combinations of fundamental sounds that occur frequently in English. Difficulties were encountered in testing, however, when this was carried too far in that enough different syllables could not be obtained for continuous testing. On the other hand, when random combinations of sounds were made, without regard to the particular combinations occurring in English, syllables that were very unusual and difficult to pronounce were obtained, unless the combinations were restricted to the simple syllable forms having only two or three sounds. In other words, testing lists must be selected with two things in mind; namely, they must be representative of speech and they must be suitable for making tests. The experience with these various lists also indicated that the results obtained with one system of lists could be calculated approximately from the results obtained with other systems by properly weighting the individual sound articulation values.

This experience led to the adoption by the Laboratories of a system of lists which has been used during the past ten years in studies of the effects of distortion upon the recognition of speech sounds.[3] These lists which have been referred to in the literature as the standard articulation lists were made up of only the con-vow, vow-con and con-vow-con syllable forms. The various fundamental sounds of English were combined at random into the syllables, such that each sound occurred approximately with uniform frequency.

During the past few years it has become evident that still further simplifications in the syllable forms used in the standard articulation lists might be made. Also our methods of making articulation tests and interpreting the results obtained have undergone considerable changes during this time. It is with these new methods that the present paper is chiefly concerned.

In order to distinguish between the old lists and the ones modified as described below, the prefixes "old" and "new" will be placed before the title "Standard Articulation Lists." When there is no chance for confusion, the new lists will be called simply the standard articulation lists, since they are the principal ones now being used in the work at Bell Telephone Laboratories.

[3] "Nature of Speech and Its Interpretation," H. Fletcher, *Journal Franklin Institute*, June, 1922.

New Standard Articulation Lists

In setting up any testing list it is necessary to classify and select a representative group of speech sounds. The National Phonetic Association uses a basic alphabet of 65 different sounds and also uses numerous modifiers which serve to distinguish slight variations in a given sound. Such a system is too complex for testing purposes. The revised scientific alphabet uses 48 simple sounds of which 24 are consonants, 19 vowels, and 5 diphthongs. Besides these fundamental sounds, connected speech contains certain recurrent combinations of them, such as st, ing, etc.

In speech these fundamental sounds are combined into syllables in a large variety of ways, but as mentioned before, in constructing a testing list it is desirable to adhere to very simple syllable forms. More complex forms which include the compound endings are either too few in number or involve unusual speech sound combinations. In either case they are soon memorized by a testing crew working with such lists. In the new lists, therefore, simplifications are made by omitting the con-vow and vow-con types of syllables, leaving only the con-vow-con type. In order to make syllables of this type it is obviously necessary to have the same number of vowels and consonants, provided that each consonant may be used in both the initial and the final position. Some consonants, however, can be used only in the former while others can be used only in the latter position.

With these facts in mind the sounds that are shown in Table I were adopted for these new lists. It will be noticed that all of the consonants are used in both the initial and final positions in the syllable, except h, w, and y, which are used only in the former, and zh, ng, and st, which are used only in the latter position. As was the case in the old standard lists, it will be seen that, in the new lists the vowel variants have been excluded. They occur infrequently in speech and phoneticians do not universely agree on their pronunciation. For this reason they are not included. Also, the diphthongs ī, ou, oi, and ew, which were used in the old lists, were omitted from the new lists. The last two of these diphthong sounds occur very infrequently in speech. Although the diphthongs, ī and ou, do occur quite frequently, it was felt that their essential properties were embraced by the properties of their constituent vowel sounds. By their omission and also by the introduction of the compound st as a final consonant, it is possible to construct any desired number of syllables of the con-vow-con type, from the speech sounds shown in the table.

TABLE I

SPEECH SOUNDS FOR NEW STANDARD TESTING LISTS *

Initial Consonant	I.P.A.	Key Word	Vowel	I.P.A.	Key Word	Final Consonant	I.P.A.
b			a	[ɑ:]	father	b	
d			a			d	
f			ā	[e:]	fame	f	
g		go	ā			g	
k			a'	[æ]	fat	k	
l			a'			l	
m			e	[ɛ]	get	m	
n			e			n	
r			ē	[i:]	greet	r	
p			ē			p	
s			i	[ɪ]	tin	s	
sh	[ʃ]	ship	i			sh	
th'	[ð]	this	o	[ʌ]	but	th'	
th	[θ]	thin	o			th	
t			ō	[o:]	go	t	
v			ō			v	
ch	[tʃ]	church	u	[ʊ]	full	ch	
z			u			z	
j	[dʒ]	judge	ū	[u:]	rule	j	
h			ū			zh	[ʒ]
w			o'	[ɔ:]	haul	ng	[ŋ]
y	[j]	yawl	o'			st	

Note: Final r and ng are used in the list only when they occur in combination with the following vowels:

a'r (as in carry, paragraph)
ar (as in are, far, tar)
er (as in bury, ferry, verify)
ir (as in spirit)
or (as in her, utter, fir)
o'r (as in for, lore)
ūr (as in your, sure)

a'ng (as in bang, sang)
eng (as in geng, e as in ten)
ing (as in sing, wring)
ong (as in sung, hung)
ung (as in gung, u as in took)
o'ng (as in long, wrong)

* The symbols for the sounds are those of the International Phonetic Association's alphabet. See Pronunciation of Standard English in America, Krapp, Oxford University Press, 1919. See also Revised Scientific Alphabet, Funk and Wagnall's Dictionary.

The testing syllables are formed in the following way. Cards upon which the initial consonants are written are placed in one box; others upon which the vowel sounds are written are placed in a second box; and those upon which the final consonant sounds are written are placed in a third box. A card from each box is drawn at random, thus forming the con-vow-con syllable. By drawing all of the sounds, a list of 22 syllables is formed. This process is repeated three times to obtain a list of 66 syllables which is a unit that has been found convenient to use. A list of syllables of about this length can be used without giving callers and observers a rest period. In such a list each initial consonant occurs three times, each vowel six times, and each final consonant three times.

In forming such syllables only those combinations involving final r and ng that are shown in Table I are included. Much confusion exists concerning the pronunciation of other combinations of these sounds. Syllables that represent slang in English are also omitted. These omissions are made by returning the card upon which the sound in question is written to its box and drawing another card. By combining the sounds at random in this manner any desired number of lists may be made which for practical purposes are all of equal difficulty.

In addition to containing a certain speech sound content, connected speech is characterized by inflection, accent, a rate of utterance, etc. In the earlier articulation studies the test syllables were called singly at intervals of about three seconds. When considered with respect to connected speech this procedure seems somewhat artificial. Comparative tests were made in which the syllables were called singly and as parts of introductory sentences. The tests showed the syllable articulation to be somewhat larger when the introductory sentences were used. The increase was due largely to the greater ease in interpreting the initial consonants of the syllables, when they were inserted in the introductory sentences. The effect was most noticeable for the stop and fricative consonants which have relatively short durations. In order to make the technique more nearly like connected speech the syllables are called in the short introductory sentences. The use of such introductory sentences also helps to insure that any element in the transmission system being tested, whose performance depends particularly upon their immediate past history, will be in the condition in which we are interested for determining speech transmission capabilities.

A list of sentences which is used for this purpose together with a sample record of articulation data is shown in the articulation test record of Table II. For calling purposes, the syllables on the cards are written in the spaces under the columns marked "called" of the test record. These sentences are called uniformly at the rate of 15 per minute. When the syllables in the first column are called, the sentences are repeated using the syllables in the second column and then those in the third column.

The observers are provided with blank articulation test record sheets. They write the sounds which they hear in the corresponding "observed" columns. When the test is completed the observed and called sheets are compared and the various articulations obtained.

For good results it has been found advisable to use a testing crew of ten people—5 men and 5 women. Eight people are ordinarily

TABLE II

Articulation Test Record

DATE 3-16-28 SYLLABLE ARTICULATION 51.5%

TITLE OF TEST PRACTICE TESTS CONDITION TESTED 1500~ LOW PASS FILTER

TEST NO. 10 OBSERVER W.H.S. JR.

LIST NOS. 5-9-37 CALLER E.B.

NO.		OBSERVED	CALLED	OBSERVED	CALLED	OBSERVED	CALLED
1	THE FIRST GROUP IS	máv	náv	póz	poth	Kōb ✓	Kōb
2	CAN YOU HEAR	pōch ✓	pōch	nēz	nēzh	shēth	siz
3	I WILL NOW SAY	seng ✓	seng	jóch ✓	jóch	fūch ✓	fūch
4	AS THE FOURTH WRITE	chūd ✓	chūd	thám ✓	thám	thōl ✓	thōl
5	WRITE DOWN	run ✓	run	hab ✓	hab	poth ✓	poth
6	DID YOU UNDERSTAND	chiz	kiz	def	doth	wám ✓	wám
7	I CONTINUE WITH	foz	fozh	chech	chej	gūm	gūn
8	THESE SOUNDS ARE	lól ✓	lól	lun	lon	nāsh	nāth
9	TRY THE COMBINATION	jās	zhāth	shāl ✓	shāl	vóg ✓	vóg
10	PLEASE RECORD	tháth	thásh	muz ✓	muz	lung	long
11	WRITE THE FOLLOWING	wūr ✓	wūr	lēd	bēd	diz	dizh
12	NOW TRY	yāp ✓	yāp	wif ✓	wif	KaK	taK
13	THIRTEEN WILL BE	mad	maj	gōst ✓	gōst	thár	zhár
14	YOU SHOULD OBSERVE	bēch	bēK	thāv	sāv	must ✓	must
15	WRITE CLEARLY	gēm	dēm	Kōf ✓	Kōf	yód ✓	yód
16	NUMBER 16 IS	theb	veb	róg ✓	róg	jet ✓	jet
17	YOU MAT PERCEIVE	jok	jost	thip ✓	thip	rēp	rēj
18	I AM ABOUT TO SAY	gaf ✓	gaf	yar ✓	yar	thēp	hēp
19	TRY TO HEAR	hus ✓	hus	zhūt ✓	zhūt	—	chuv
20	PLEASE WRITE	hiv	thith	Kūk	tūk	thef	thesh
21	LISTEN CAREFULLY TO	tōg ✓	tōg	fung ✓	fung	bās ✓	bās
22	THE LAST GROUP IS	shót ✓	shót	thev	vesh	thóf	shaf

c-1242 (6-28)

employed in a test, the remaining two being held for emergencies in order to keep the work going. One member of the crew calls at a time, and the remaining members act as observers. Ordinarily, eight callers are used with four observers recording simultaneously for each caller, although as many as eight or nine observers may be used. The order is arranged such that the various members are equally represented in the test.

Each observer's sheet (Table II) furnishes a value of syllable articulation (the percentage correctly observed), corresponding to a particular caller-observer pair and to 66 syllables or 198 speech sounds called. These values of syllable articulation are recorded in the form shown in Table III. The average of each column gives the average articulation for each observer. The averages of the rows give the callers' articulation.

TABLE III

ARTICULATION TEST RESULT RECORD

FROM 2/29/28 TITLE PRACTICE TESTS

TO 3/21/28 CONDITION 1500 CYCLE LOW PASS FILTER

REFERENCE MM-2186-8/27/28 REMARKS

CONDITION

OBS. / CALLER	E.B.	W.H.S.	F.S.	C.M.	H.C.	A.H.	E.L.F.	M.W.	P.H.		Ave.
E.B.		51.5	51.5	50.0	48.5	57.5	59.0		45.5		51.9
W.H.S.	72.5		62.0	69.5	58.0	66.5	60.5		56.0		63.7
F.S.	47.0	54.5		56.0	48.5	57.5	50.0		42.5		50.9
C.M.	51.5	62.0	53.0		42.5	65.0	57.5		38.0		52.8
H.C.	48.5	56.0	48.5	36.5		50.0	54.5	30.5			46.4
A.H.	41.0	53.0	50.0	53.0	35.0		47.0	42.5			45.9
E.L.F.	51.5	51.5	51.5	45.5	53.0	57.5		41.0			50.2
M.W.	50.0	48.5	50.0	36.5	51.5	58.0	63.5				51.3
Ave.	51.7	53.9	52.4	47.6	48.3	57.0	56.0	38.0	45.5		51.6

When the articulation values are near 100 per cent or 0 per cent, then a group of values will not distribute itself symmetrically about the arithmetic mean or average value. For the high values, this is due to the fact that one cannot get a higher value than 100 per cent. To some observers, the 100 per cent mark may be obtained very easily and to others it may be obtained only with considerable effort. This difference in difficulty cannot be registered in the percentages obtained. A similar reason exists for the unsymmetrical grouping for values near zero. Our experiments have shown that this grouping is symmetrical in the range from 20 per cent to 80 per cent. For the range from 80 per cent to 100 per cent, the average value is less than the most frequent value, and for the range from 0 per cent to 20 per cent, the average value is greater than the most frequent value. These differences are of the order of 1 or 2 per cent. From an extensive series of tests, the averaging factor curve of Fig. 1 was constructed which enables the data to be averaged, in a way, such that the average value is approximately equal to the value which would be most frequently observed in a large number of tests. To do this, each observed articulation, based on 66 syllables, is converted into an averaging factor by means of the above curve. These factors are then averaged and the average value reconverted into average

articulation by means of the above curve. The value so obtained is taken as the average syllable articulation for the test.

The articulations for the various sounds are recorded on the Articulation Test Analysis Record of Table IV. In making the analysis the total errors for each caller are counted. The occurrences per caller are the products of the number of times the sounds are spoken by the caller, and the number of observers. One analysis sheet contains the results for a complete test.

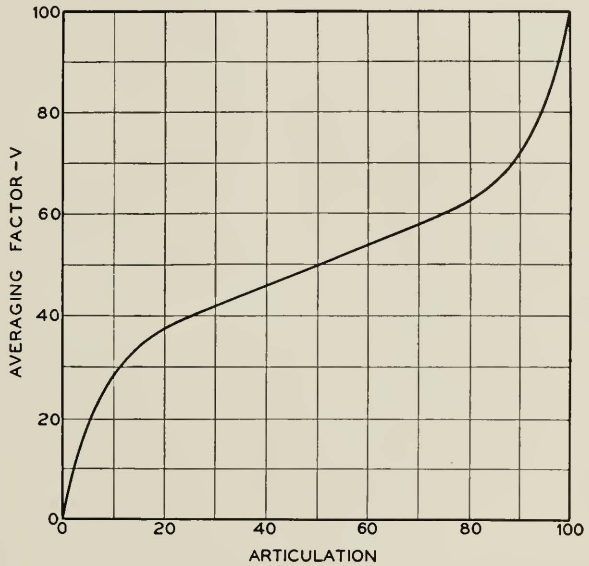


Fig. 1—Averaging factor curve

In dealing with the syllable articulation the unit is 66 called syllables, or the result of one caller and one observer. Hence a number of syllable values are obtained for one test which are averaged by means of the averaging factor curve. When we deal with the articulation of the individual speech sounds, it is advisable to use a larger unit since each sound occurs only six times in a 66 syllable unit. In Table IV the errors are shown for each caller as a unit, and since there were 7 observers per caller, each sound occurs 42 times. This is a unit of sufficient size to qualitatively compare various voices. However, in drawing conclusions as regards the effects of the circuit upon the transmission of the various sounds, it is best to use eight callers as a unit, so that each sound, depending upon the number of observers, occurs of an order of 200 to 300 times. In this case the articulation of an individual sound has a precision that is comparable

with the precision of the syllable articulation when based on 66 called syllables, as will be discussed in a later paragraph.

TABLE IV

ARTICULATION TEST ANALYSIS RECORD

DATE 2/29/28 To 3/21/28 TITLE PRACTICE TESTS

REFERENCE MM-2186 8/27/28 CONDITION 1500~LOW PASS FILTER

OBSERVERS EB, WHS, FS, CM, HC, AH, ELF, MW, PH SOUND ARTICULATION 79.3 SYLLABLE ARTICULATION 51.6

LETTER	OCCUR. PER CALLER	ERRORS PER CALLER																		IND. SOUND ART.	
		EB		WHS		FS		CM		HC		AH		ELF		MW		TOTAL ERRORS		TOTAL OCCUR.	
a	42	3		2		10		1		15		2		2		3		38		336	
ā	42	1		1		4		0		7		2		0		2		17		336	
a'	42	2		0		8		7		20		22		27		11		97		336	
e	42	2		4		4		4		22		19		20		15		90		336	
ē	42	0		0		2		0		1		4		0		1		8		336	
i	42	2		0		2		0		9		8		9		3		33		336	
o	42	12		1		6		5		12		5		4		10		55		336	
ō	42	1		0		0		0		0		1		0		0		2		336	
o'	42	0		0		5		2		0		0		0		0		7		336	
u	42	5		2		19		1		8		3		4		4		46		336	
ū	42	1		1		1		1		4		4		2		1		15		336	
TOTAL	462	29		11		61		21		98		70		68		50		408		3696	
b	42	7		0		11		10		13		1		4		6		52		336	
ch	42	13		4		11		13		2		8		8		7		66		336	
d	42	9		3		6		8		8		7		10		14		65		336	
f	42	6		11		7		12		25		23		21		22		127		336	
g	42	9		4		8		13		2		3		1		2		42		336	
h	21	2		0		0		0		2		0		1		1		6		168	
j	42	17		4		10		9		3		6		3		2		54		336	
k	42	17		6		6		11		3		12		11		6		72		336	
l	42	1		0		3		0		0		2		3		0		9		336	
m	42	1		2		4		6		2		5		7		0		27		336	
n	42	4		5		0		10		7		6		9		15		56		336	
ng	21	0		0		2		3		1		1		4		5		16		168	
p	42	11		11		4		5		19		21		14		8		93		336	
r	42	3		0		0		2		4		3		4		2		18		336	
s	42	18		5		16		9		23		22		18		17		128		336	
sh	42	25		21		28		26		14		25		17		20		176		336	
st	21	6		0		8		2		7		6		2		1		32		168	
t	42	15		8		15		13		33		24		25		13		146		336	
th	42	22		24		17		22		27		16		22		27		177		336	
th	42	20		24		25		19		17		12		20		15		152		336	
v	42	17		7		12		10		4		9		5		14		78		336	
w	21	0		0		0		0		0		0		0		0		0		168	
y	21	1		0		0		0		1		0		1		0		3		168	
z	21	6		4		11		4		12		9		9		13		68		168	
zh	42	33		24		32		35		27		27		23		26		227		336	
TOTAL	924	263		167		236		242		256		248		242		236		1890		7812	

K-1945 (2-28)

SELECTION OF TESTING PERSONNEL

It is necessary to set up the technique of testing, such that the values of syllable articulation can be reproduced within acceptable limits. The limits depend upon the control of the auditory and vocal

characteristics of the testing crew, the control of numerous haphazard factors, and the control of the practice or experience that the crew acquires in the testing of circuits.

The departure from normal, in acuity of hearing of prospective crew members, should be measured with a good audiometer. In our laboratories the 2-A audiometer is used for this purpose. Only those individuals whose average hearing loss departs from normal in the speech range of frequencies (100 to 8,000 cycles per second) by less than 5 db (decibels) are selected. Although of normal hearing, some observers of a crew usually obtain higher values of syllable articulation than do others. The averages of the columns of Table III are a typical set of results for nine observers who have passed such a hearing test and who have had a year or more of experience in observing.

Observer A. H. obtained the highest percentage, namely 59, and observer M. W. obtained the lowest percentage, namely 38. In general this order would be preserved in a series of tests, although haphazard variations in a single test might change the order. The spread in observations is of an order of 20 per cent. More extended tests have shown that the spread tends to decrease as the observed percentages approach 0 or 100. In order to make a replacement in the observing personnel from time to time without causing a probable change of more than 2 per cent in the average percentage, it is necessary to use an observing crew of 8 to 10 persons. Our experience has shown that men and women show no characteristic difference when acting as observers.

The ability of prospective crew members to enunciate the sounds in a normal way is determined in the following manner. An extensive series of tests on various voices have yielded data which are arbitrarily used as a basis for determining normalcy. These tests were made with a simplified list consisting of common English words which will be described in a later paragraph (see Table XVII). Tests were made under three conditions; namely, direct transmission through the air in a quiet, well damped room, transmission over a circuit which uniformly transmitted the frequency range from 100–4,500 cycles, transmission over a circuit having a carbon transmitter. A diagram of this latter circuit is shown in Fig. 7. The sounds were observed by a crew of experienced observers. Table V gives the results of tests that were made upon 21 male and 23 female voices, the personnel being selected from various departments of the Laboratories. The average articulations of the simple consonant sounds are shown. The data are given separately for men and women.

TABLE V

NORMAL ENUNCIATION

Speech Sound	Air Transmission		Band 100 to 4,500~		Carbon Transmitter Circuit	
	Av. % Articulation of Sounds		Av. % Articulation of Sounds		Av. % Articulation of Sounds	
	Men	Women	Men	Women	Men	Women
b	98.7	98.0	96.2	90.1	95.0	91.5
ch	100.0	99.3	98.4	98.0	98.7	98.4
d	99.4	100.0	98.7	98.3	91.9	88.6
f	97.5	84.8	96.5	87.7	79.6	65.2
g	100.0	100.0	99.4	94.6	70.6	63.0
h	100.0	98.0	99.4	98.0	94.4	97.8
k	100.0	100.0	99.1	99.7	78.1	82.1
l	100.0	100.0	98.1	97.7	96.0	87.5
m	99.2	91.3	97.5	95.5	86.5	86.0
n	99.4	100.0	99.4	99.1	95.5	94.5
ng	100.0	100.0	99.7	99.4	93.7	100.0
p	96.9	95.8	97.2	96.6	80.0	73.5
r	100.0	100.0	100.0	99.7	96.9	75.5
s	97.5	99.1	95.0	68.0	91.2	67.8
sh	100.0	100.0	100.0	99.8	98.3	93.1
th	90.4	80.6	75.2	56.3	60.1	50.2
th'	97.5	87.0	93.3	87.7	74.1	77.2
t	98.3	100.0	99.4	96.6	92.9	72.1
v	93.3	78.9	96.2	83.6	87.1	75.0
w	100.0	99.3	100.0	98.9	97.5	84.8
z	100.0	100.0	95.9	70.8	91.2	65.2
Aver.	98.2	94.6	96.4	90.2	87.5	80.5

For our work, a prospective crew member is required to call such a list of syllables to a crew of experienced observers. If the observed articulations of the sounds are reasonably close to those indicated in Table V for each circuit, and if no obvious irregularities are noticed in the speech, the prospect is considered satisfactory for testing work. Measurements are also made upon the individual's speech power, but it has not been found necessary to use the information in the process of selection.

Aside from the practical application to the methods of testing, the table is interesting in showing characteristic differences between the voices of men and women. In general, woman's speech is more difficult to interpret than man's, particularly, in the case of the sibilant and fricative consonants. This is probably due to the fact that in woman's speech, these sounds are not only fainter, but occupy higher frequency ranges than in man's speech. The frequency range from 6,000 to 8,000 cycles for the former, is approximately equivalent to the range from 4,000 to 6,000 cycles for the latter. In the case of

the voiced sounds, woman's speech has only one half as many components as man's, which also may cause greater difficulty in interpreting the former.

With respect to the vowel sounds, the crew members are instructed in the correct manner of enunciation. Only those vowels which have definite differences have been included in the testing lists, so that, slight differences in enunciation do not seriously affect the observed results.

The object of the selection process is to determine in a broad but definite way the normalcy in speech of prospective members, and to eliminate those individuals who have speech characteristics which are not readily reproducible should it be necessary to change the testing personnel. The row averages of Table III show a typical set of results for 8 callers who were selected in the above way, and have had a year or more of experience in calling.

The spread in results is of an order of 20 per cent, so that, if a crew of 8 to 10 callers is used, a replacement may be made in the calling personnel without causing a change in the average percentage of more than 2 per cent. Owing to inherent differences in the voices of men and women, they are equally represented on the testing crew. Individuals who have the equivalent of a high school education, and whose ages range from 18 to 23 years, are usually selected for this work.

Control of Haphazard Factors

Haphazard factors arise from various sources, some of which can be controlled reasonably well. The observers work in a sound-proof room, so that extraneous noises will not affect the articulation results. The calling is ordinarily done in a sound-proof booth that has been especially treated with sound absorbing material so as to reduce the reverberation time to an order of a few tenths of a second. Ordinarily the crew does not test more than two to four hours during the day, and the schedule is usually arranged so that this is not done continuously.

The intensity level of each caller is also measured during the test, as small variations in intensity level may cause rather large variations in articulation. Ordinarily the various callers are permitted to call at the intensity level most natural to them, although in some tests the callers all attempt, by watching an indicator, to call at the same level. Various instruments have been used for measuring the intensity levels during tests. The volume indicator [4] has proven quite satis-

[4] This instrument depends for its readings, essentially upon the syllabic powers of the vowel and semi-vowel sounds, so that the reading of the instrument is determined largely by the amplitudes in the frequency range from 100 to 2,000 cycles.

factory and is the instrument ordinarily used for this work. It has the advantage over some of the other instruments that were tried, of being in much more general use on speech circuits.

Control of other haphazard factors of a more or less psychological character, may best be obtained by taking enough data so as to average out their effects. This involves the number of syllables that are called by each speaker and the number of caller-observer pairs that are used in the test. The variability of caller-observer pairs for a calling unit of 66 syllables may be seen from Table III. The probable error [5] in percentage articulation of a single observation (ϵ_s) i.e., one caller-observer pair as taken from the data in the table, is $\pm$ 9. The probable error of the average articulation ($\epsilon_{av.}$) of the 56 caller-observer pairs is $\pm$ 1.2.

It has been found from a large number of tests that the probable error of a number of crews, each consisting of one caller and one observer, is of an order of $\pm$ 12 (per cent articulation) for a 66 syllable unit when the syllable articulation is around 50 per cent. This value tends to decrease with increasing experience in testing, and with increasing or decreasing values of syllable articulation. The use of 36 caller observer pairs obviously reduces the probable error to an order of $\pm$ 2 in percentage articulation, which is about the order of magnitude of the errors involved in maintaining the testing personnel over a period of time.

Since as will be shown in a later paragraph, the syllable articulation is equal to the cube of the sound articulation, the probable error in the sound articulation [6] for one caller and one observer, or a unit of 198 sounds, is of an order of $\pm$ 6 when its value is around 80 per cent. Since each individual sound is called only six times, the probable error for each individual sound for a single caller-observer pair is of an order of $\sqrt{\frac{198}{6}} \times 6 = \pm 35$. If a test comprises 4 observers per caller and 8 callers, each sound is called 192 times, which reduces the probable error for the articulation of each sound to $\pm$ 6. Under

[5] $\varepsilon_s = .67 \frac{\Sigma d^2}{n-1}$ and $\varepsilon_{av.} = \frac{\varepsilon_s}{\sqrt{n}}$; where, n = number of caller-observer pairs; d = difference between the articulation of a caller-observer pair and the average articulation of n caller-observer pairs.

[6] $\varepsilon_s = \frac{\partial S}{\partial L} \varepsilon_L = 3L^2 \varepsilon_L$

$\varepsilon_s = \pm 12$, $S = .5$, $L = S^{1/3}$,

$\varepsilon_L = \frac{\varepsilon_s}{3S^{2/3}} = \pm 6.$

ε_s = prob. error in syl. art. for one caller-observer pair.
ε_L = prob. error in sound art. for one caller-observer pair.

the same circumstances, the probable error for the average syllable articulation is ± 2, and that for the sound articulation is ± 1.

Control of Practice Effects

The third factor entering into the reproducibility of articulation results is practice and experience. The practice effect manifests itself in various ways. An increase in articulation takes place as the observers become more familiar with the vocal characteristics of the speakers. Similar effects are observed as they become more accustomed to a given technique, or to a particular type of distortion. In general, these effects become smaller as the testing crew becomes more experienced.

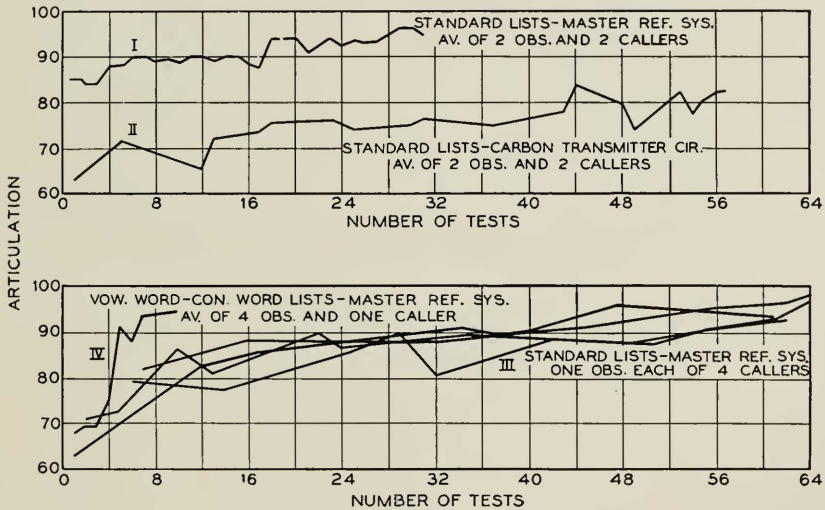


Fig. 2—Typical growth curves

Fig. 2 shows several typical growth curves that were obtained in the process of training new crew members. In this process the new members observe continually on various circuits until the results compare favorably with the results that are obtained by experienced observers. In such tests, experienced speakers are used. The averages for two new observers, of the results that were obtained on a high grade circuit, are shown by Curve I. Two speakers were used in these tests. A limited amount of testing was done by the observers prior to the above tests. Upon the completion of the tests of Curve I about 30 or 40 additional tests were made on various circuits. A

series of tests, in which several speakers were used, were then undertaken on a carbon transmitter circuit. In Curve II the averages for the two observers, of results on two voices, are shown. Three to four weeks' time was spent by the observers in making the various tests mentioned above.

The curves under III show similar data that were taken at a later date by one new observer, for several voices. All of the above tests

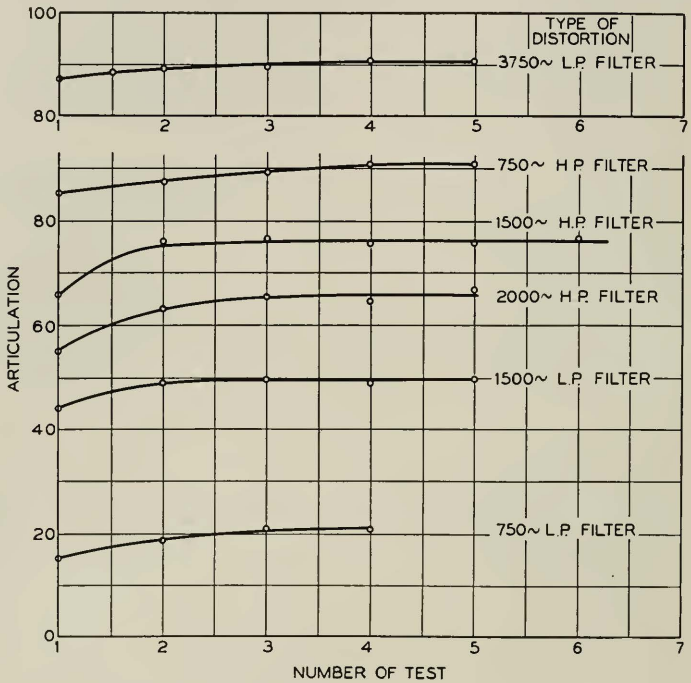


Fig. 3—Practice effects for an experienced crew

were made with the standard lists. In Curve IV, data are shown that were obtained with the vowel and consonant word lists (see Table XVII). In these tests four new observers were used and no preliminary training was given. It is evident that with the word lists, the results reach a state of saturation very quickly.

After a crew has spent several months in testing, its performance becomes largely mechanical. Under such circumstances the practice effects are rather small for types of distorted speech with which it has had experience. When the speech distortion is unusual, however, rather large practice effects may be obtained. Fig. 3 shows such

practice effects for several types of distortion for a crew of eight people. All six of the circuits were tested on each test before going on to the following test. The first three tests were made successively and covered a period of about two months. In each test the types of distortion were interspersed. In other words half of the first test was completed with the filters in one order, and the other half with the filters in the reverse order. The fourth test was made about three

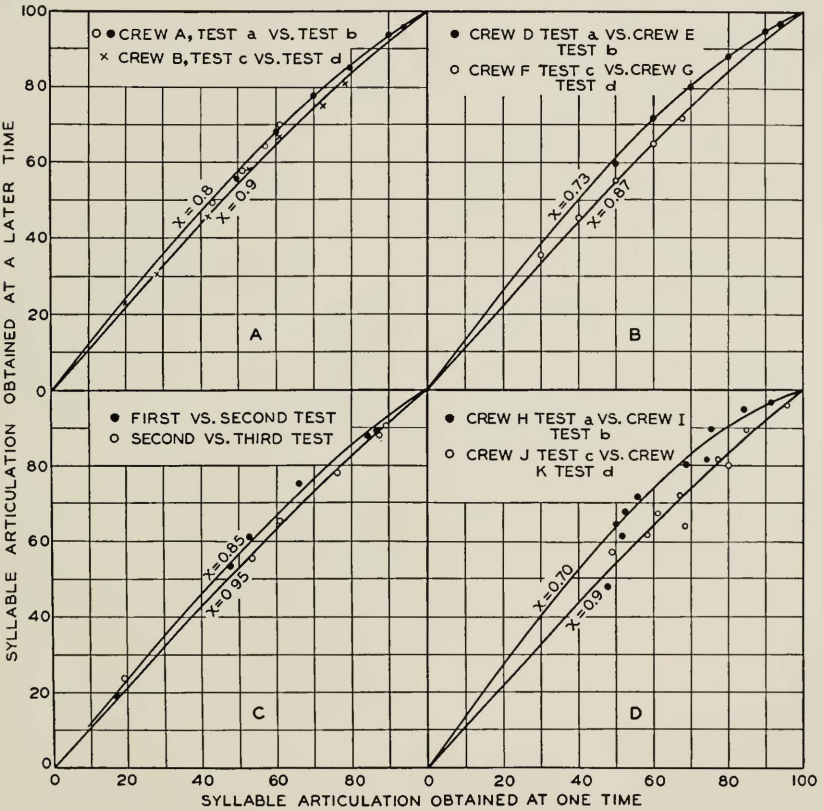


Fig. 4—Practice effects

months later, and the fifth test was made approximately six months later. Although the crew had been testing various circuits for about a year, and were thoroughly accustomed to the routine, these particular circuits had not been previously tested, so that, the crew's experience with these types of speech distortion was small. It is evident that the articulation of an experienced crew reaches saturation very quickly. It is probable that practice effects for a crew of several years' experience with distorted speech would be negligible.

Several procedures are followed in order to correct, in so far as possible, for practice effects. In comparative tests, whenever it is possible, the circuits to be compared are interspersed so as to average out practice effects. If it is desired to compare the articulation of a very new or unusual circuit (from the standpoint of the speech distortion) with one of common experience, several successive tests are made upon the new circuit until no further increase in articulation with practice appears. When it is impossible to intersperse the tests, the data may be corrected to a given state of practice by means of curves which were obtained in the following way. Although as will be seen, this procedure is valid only under certain restrictions, which will be discussed, such a correction will always tend to correct the data to a more comparable basis.

In Fig. 4-*a* a practice curve is shown that was obtained for a crew, from two series of tests that were separated by an interval of three months. The dots represent tests that were made upon a circuit which uniformly transmitted a frequency range from 100 to 5,500 cycles. The circles were obtained from a circuit of the type shown in Fig. 7 involving the carbon transmitter. In both cases the various articulation values correspond to different received speech levels. The crosses represent similar results that were obtained with a different crew on the latter type of circuit.

In Fig. 4-*c* the data of the first three tests in Fig. 3 are shown. In this case the distortion was varied and the received speech level held constant. As previously stated, in so far as was known, the crew had no previous experience with these types of speech distortion so that the practice for the various types of distortion ought to be comparable.

All of the solid curves are graphs of the following equation

$$(1 - S') = (1 - S)^x \cdots, \tag{1}$$

where S' = decimal value of syllable articulation obtained on a given circuit at one stage of a crew's career,

S = the value obtained on the same circuit at a later stage of the crew's career,

x = a number called the practice factor.

The values of the practice factor x that were necessary in order to fit the observed values are shown in the figure. It is impossible to state definitely that a crew has uniform practice with various types of distortion for the reason that experience is cumulative. A crew's experience with one type of distortion may be of aid in the under-

standing of some other type of distorted speech. With this in mind it will be seen that a constant value of x fits the data for the various types of distortion reasonably well. In the case of changing speech levels with a constant type of speech distortion, where the question of uniformity of experience is not so important, the fit is even better.

It is reasonable to suppose that an inexperienced observer must make a greater mental effort than an experienced observer to obtain the same articulation values. In other words the element reflected

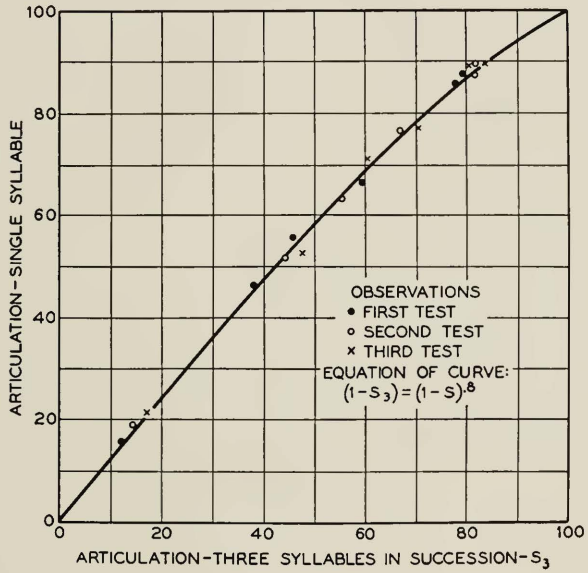


Fig. 5—Relation between techniques

by these curves is closely associated with the burden or strain upon the observer. A somewhat analogous situation obtains when tests are made with two different techniques which differ primarily in the burden imposed on the observer. In making the filter tests described above, two techniques which differed in this respect were used. One was the standard technique, in which one syllable was called with the introductory sentences. In the other, the syllables were called in groups of three (three in succession) with the sentences. The syllables were uttered as nearly in the manner of a three syllable word as was possible.

The results that were obtained with the two techniques are shown in Fig. 5. When the syllables are called in groups of three, the articulation values are smaller than when they are called singly.

It was found, however, that the type of relation shown in Eq. 1 also relates the data obtained with the two techniques. In this case the relation may be expressed as follows:

$$(1 - S_3) = (1 - S)^{\cdot 8 \cdots}, \tag{2}$$

where S_3 = decimal value of syllable articulation when called connectedly,

S = decimal value of syllable articulation when called singly.

The curve of Fig. 5 is a graph of the above equation.

In this case uniformity of experience with the various types of distortion does not enter, as the tests with the two techniques were made simultaneously. The only difference in the techniques was that in the three-syllable case the observer listened to three syllables before writing them down. It seems reasonable to conclude, therefore, that when a crew has the same experience with different types of distortion, then the results obtained by it at one time may be compared with the results obtained by it at some other time by using such a relation. No doubt other types of functions could be found which would also fit the above data. The relation shown here was chosen because it fit both the practice data and the data that were obtained with the two different techniques and is very convenient to use in making such corrections.

It is evident that in order to use the practice curves it is necessary to set up a reference circuit in order to obtain an appropriate value of x. Theoretically, one reference condition should be sufficient, provided that the practice of the crew had the same relative distribution over various types of speech distortion. Since this is usually not the case, it is necessary to use several reference circuits representing various types of speech distortion. When it is desired to correct data for practice effects, the appropriate value of x is determined by making tests upon the reference circuits having types of distortion similar to the circuits for which the corrections are desired. A description of several reference or control circuits which have been found useful with the values of sound and syllable articulation as obtained with the testing crew of five men and five women as previously described, is given below.

(*a*) *Air Transmission. Master Reference System for Telephone Transmission.*—The air transmission tests were made in a quiet, well damped room having a volume of approximately 1,000 cubic feet. The observers faced away and were located at an average distance of 30 inches from the speaker. Sound articulation "L" 99.1 per cent.

Syllable articulation "S" 97.5 per cent. Practically identical results were obtained with the "Master Reference System" [7] with the system set for optimum received speech level, i.e. a sensation level of 70 db, average distance from lips to transmitter 1.5 inches.

(*b*) *Auxiliary Circuit of the Master Reference System.*—The auxiliary circuit of the master reference system consists of networks which are

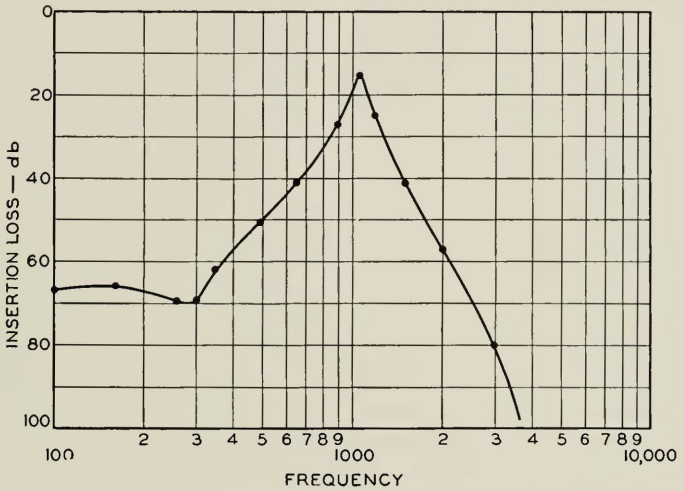


Fig. 6—Insertion loss of auxiliary networks

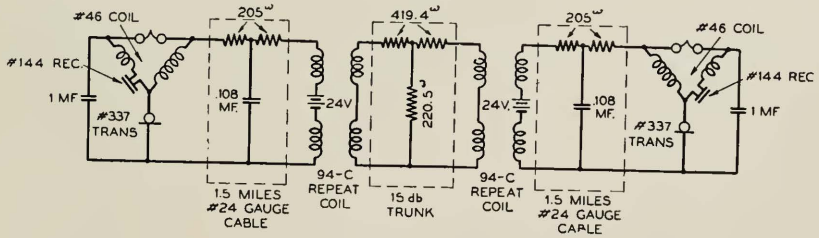


Fig. 7—Carbon transmitter circuit

inserted into the otherwise distortionless reference system, to give it a frequency resonance around 1,100 cycles. The insertion loss of the networks is shown in Fig. 6. This loss is approximately equal to the combined losses of the No. 1 transmitter and receiver distortion networks of the Master reference system. Sensation level 74 db $L = 89.2$ per cent, $S = 72$ per cent.

[7] L. J. Sivian, "A Telephone Transmission Reference System," *Electrical Communication*, 3, Oct. 1924. M. Cohen, "Apparatus Standards of Telephonic Transmission." W. H. Martin and C. H. G. Gray, "Master Reference System," *Bell Tech. Jour.*, July, 1929.

(*c*) *Carbon Transmitter Circuit* (see Fig. 7).—The average values for five transmitters are $L = 93$ per cent, $S = 81$ per cent. In these tests the sensation level of the received speech was 75 db, and the calling level as measured by a volume indicator bridged across the line side of the input repeating coil was − 12.5 db.

(*d*) *Master Reference System Plus Filters.*—System set for a sensation level of 70 db.

3,750~ Low	Pass Filter	$L = 96.7\%$	$S = 91.0\%$
750~ High	" "	$L = 96.7\%$	$S = 91.0\%$
1,500~ Low	" "	$L = 77.7\%$	$S = 49.5\%$
1,500~ High	" "	$L = 91.0\%$	$S = 76.0\%$

The foregoing discussion has been concerned with methods of correcting the articulation results obtained by a given crew at different times to an arbitrary stage of practice or experience. To do this it is necessary to calibrate the crew for types of distortion that are similar to those of the systems for which the corrections are desired. The method has been described in detail because there are times when it is necessary to make such corrections. However, it has been our experience that such practice effects become negligible with a crew that has been set up in accordance with the methods previously described, when the crew's experience with types of distortion is diversified and when unusual circuits are tested successively until no further increase in articulation with practice occurs.

These methods may also be used to correlate the articulation data of various crews and various techniques, provided that the only essential difference between the crews and techniques is in the demand or burden that is placed upon the observer. This means that the crews must have similar vocal characteristics and similar hearing abilities, and that the testing lists must have similar speech sound content. It has been found, for example, that a crew of women callers obtain a considerably higher articulation than men callers on a circuit which eliminates all frequencies below 1,500 cycles and vice versa on a circuit which eliminates all frequencies above 1,500 cycles. It is obvious, therefore, that the methods described above could not be used to correlate the two crews for such circuits. Similarly, the methods could not be used for comparing two crews, if the hearing level of one is 10 db below the other, or to compare two techniques, one of which is made up entirely of vowel sounds and the other entirely of consonant sounds. As shown in Fig. 4-*b*, data have been obtained with various crews on various circuits which can be correlated very

well by means of the above curves. Fig. 4-*d* gives data that were obtained with other crews which show very poor correlation. In neither case are the characteristics of the crew well enough known to satisfactorily account for the observed differences. At the time the work was done the significance of these factors was not so apparent, so that they were not given the attention they now receive. During the past two years a crew of 10 people has been used almost continuously in testing work. During this time numerous changes in personnel have taken place and only five of the original members are now on the crew. The data obtained during this time appear to be strictly comparable. In some cases it is necessary to use the practice curves. In other cases (circuits that are frequently tested), practically identical results are obtained. For this reason, it is believed that if a similar crew of 10 different people were to be selected as previously described, comparable articulation results would be obtained. It seems reasonable to expect that crews testing in various languages should also obtain comparable results provided that the crews were similar in the sense used here and that the lists were phonetically similar. It seems desirable, therefore, to standardize on the factors which affect the comparison of data, such as, the size and type of crew, the type of list, and the type and number of reference circuits. Best results are likely to be obtained when the crews do not differ by amounts which correspond to values of x less than 0.7. Smaller values indicate that the crews have not had sufficient testing experience, or have speech and hearing characteristics which are essentially different, or that the phonetic content of the testing lists are appreciably different. In the latter case the results may be correlated by means of statistical relations that will be given in a later paragraph.

Relation of Articulation to the Transference of Thought by Speech

The foregoing paragraphs have been concerned with the practical problems of setting up a suitable testing technique and correlating the observed articulation results. The procedure that has been discussed enables us to measure the percentages of the various speech sounds which are correctly recognized when they are spoken in a simple con-vow-con syllable. We desire at this point to consider the broader significance of this measure. In other words, how is the articulation result related to the transference of thought by means of speech? This relationship involves many psychological factors which are difficult to evaluate so it must not be expected that a comprehensive answer can be given here, but it is important to understand

as fully as possible those parts of the problem that can be evaluated. Such a relation involves two questions, (*a*) how do the articulations of the sounds as measured with the testing lists compare with their articulations as they are used in speech, (*b*) how should the articulation values be weighted in order to obtain an index of the speech capabilities of a system.

In the first place, certain fundamental sounds of speech were omitted from the above lists. The most important of these are the consonant compounds. The majority of these sounds may be regarded as the product of a very few combining consonants acting as modifiers to the rest of the consonant alphabet. Since the combining consonants or modifiers occur over and over in combination with various consonants, it might be expected that the interpretation of the compounds would depend primarily upon the interpretation of the various consonants, and not upon the modifiers. In other words, the compounds would be interpreted as simple consonant sounds. The tests discussed below show that this is true on the average, although notable exceptions may occur in individual cases.

The testing lists were made up from the sounds shown in Table VI.

TABLE VI

Consonants		*Vowels*
Initial	*Final*	
b, br,	rb, b,	a'
d, dr,	rd, d,	a
g, gr,	rg, g,	e
p, pr,	rp, p,	i
k, kr,	rk, k,	o
t, tr,	rt, t,	
f, fr,	rf, f,	
th, thr,	rth, th,	
s, sl,	nd, d,	
b, bl,	nj, j,	
g, gl,	nz, z,	
p, pl,	nk, k,	
k, kl,	nt, t,	
f, fl,	ns, s,	
r, l,	n, r,	

These sounds were combined at random into syllables of the con-vow-con-con and con-con-vow-con form. Ten lists of 90 syllables each were made, and a crew of 10 callers, with 5 observers per caller, was used. With this number of tests the probable error in the per cent articulation for each sound is approximately 5 per cent. The tests were made on the auxiliary circuit of the master reference system. The sensation level of the received speech was about 80 db. The results are shown in Table VII.

TABLE VII

ARTICULATION OF CONSONANT COMPOUNDS

Initial				Final			
Sound	% Art.	Sound	% Art.	Sound	% Art.	Sound	% Art.
b	89.0	br	85.3	rb	63.3	b	77.0
d	98.0	dr	88.7	rd	60.0	d	91.0
g	96.3	gr	90.0	rg	57.3	g	88.7
p	78.3	pr	52.0	rp	39.3	p	66.7
k	95.3	kr	93.3	rk	52.7	k	90.7
t	79.3	tr	89.3	rt	58.7	t	88.7
f	56.3	fr	60.7	rf	39.3	f	53.3
th	71.3	thr	88.0	rth	42.0	th	52.0
ave.	83.0	ave.	81.0	ave.	51.5	ave.	76.0
s	57.3	sl	72.7	nd	92.7	d	91.0
b	89.0	bl	95.3	nj	91.3	j	96.0
g	96.3	gl	77.3	nz	82.0	z	76.7
p	78.3	pl	68.0	nk	92.7	k	90.7
k	95.3	kl	86.7	nt	84.7	t	88.7
f	56.3	fl	86.7	ns	72.7	s	44.7
ave.	78.7	ave.	81.1	ave.	86.0	ave.	81.3
Init. Ave.	81.1		81.0	Fin. Ave.	66.3		78.3

Ave. Simple Cons. 79.7
" Comp. " . . . . . 73.5 Exclusive of r () 82.7

The articulation of the consonant compounds as a class does not differ appreciably from the articulation of the corresponding simple consonants. The final r compound is seen to be an exception to this general rule. The errors for combinations containing this sound were caused by the large number of omissions of the modifier. For example, if "barb" were called, "bab" would be recorded. When the final r is combined with a consonant, the tendency is to shorten its duration and to stress it less than is done when it occurs as a simple final. Also, as mentioned before, the r sound materially modifies the vowel preceding it and usually in such a way that the vowel and r sounds are spoken as a vowel. For these two reasons, it escapes detection more readily than when used as a simple final. It will be noticed from the table that f is definitely more difficult to recognize than fl, while p is definitely less difficult to recognize than its compounds pl and pr. There is also a large difference between the results for s and ns. Although these differences are large, some tend to increase and others to decrease the average articulation. It is seen from the table that if the r compounds are omitted, the averages for the simple consonant sounds and for their compounds are approximately equal. Since this class of sounds comprises less than 15 per cent of the speech sounds, the results obtained by using a list in which the consonant

compounds are omitted will be very closely the same as those obtained by lists in which such sounds occur. In view of this, and also because their inclusion would greatly extend the time needed for testing, compound consonants have been omitted.

In conversational or written speech some of the sounds are used much more frequently than others, whereas in the testing lists each sound is used the same number of times. Does this procedure lead to essentially different articulation values, for the various sounds, from those obtained by using the sounds in proportion to their frequencies of occurrence in speech?

TABLE VIII

ARTICULATION OF SOUNDS OF EQUAL VS. UNEQUAL OCCURRENCE

Sound	Equal Occ.		Unequal Occ.	
	No. of Occur.	Art.	No. of Occur.	Art.
a	300	91.0	550	88.0
ā	300	98.3	600	96.5
a'	300	88.3	400	82.3
e	300	91.0	1200	87.4
ē	300	100.0	850	98.8
i	300	96.0	1150	94.5
o	300	86.0	500	90.6
ō	300	98.3	500	97.6
o'	300	97.7	450	96.2
u	300	88.3	150	92.7
ū	300	97.3	250	97.6
Ave.		93.8		92.9
b	300	87.0	400	91.0
ch	300	97.3	150	98.0
d	300	91.0	700	89.1
f	300	73.7	400	67.8
g	300	92.0	450	94.2
h	150	100.0	200	100.0
j	300	98.0	150	97.3
k	300	87.3	900	89.8
l	300	98.7	1200	96.2
m	300	95.3	900	81.3
n	300	93.7	1300	95.5
ng	150	93.3	400	90.0
p	300	75.3	500	71.2
r	300	97.0	1250	97.4
s	300	72.3	850	76.3
sh	300	99.3	300	99.7
st	150	96.0	200	96.5
t	300	83.7	1550	86.6
th	300	60.3	200	60.5
th'	300	70.7	150	68.0
v	300	78.7	250	71.6
w	150	96.7	400	97.8
y	150	99.3	100	94.0
z	300	67.0	250	72.0
zh	150	97.3	50	100.0
Ave.		88.0		87.3

Table VIII gives the results of articulation tests that were made with two such types of lists. In both cases the sounds were combined at random into syllables of the con-vow-con type. The tests were made on the auxiliary circuit of the master reference system.

Realizing that the probable error in the articulation value given for each sound is $\pm$ 5, there do not appear to be any outstanding differences in the articulations of the various sounds with the two types of list. The average articulations for the two lists differ by less than the probable error. The test indicate, therefore, that lists having uniform occurrence of sounds give the same individual sound articulation values as lists having the frequencies of occurrence of the sounds proportional to their frequencies of occurrence in speech. At least this is true within the accuracy usually attained in making such tests. The testing advantages of the former type of list have already been pointed out.

It is important to notice that the average sound or the average syllable articulation may not be the same for the two types of lists. even though the articulation for each sound is the same. The averages shown in the table were obtained by assigning equal weights to the articulation for each fundamental sound. If weights which are proportional to frequency of occurrence of the sounds in speech be assigned, the averages obtained will, in general, be slightly different. For the particular circuit corresponding to the data of Table VIII, the averages obtained in the two ways did not differ by more than the observational error. Our data have shown that this is also true for a large class of circuits ordinarily used in telephone work. However, those transmission systems which have a specific effect upon certain consonant or vowel sounds, for example, upon s which occurs 850 times in one list compared to 300 times in the other, would obviously have different values for the sound articulation by using the two methods of obtaining the average.

In speech, certain combinations of sounds occur more frequently than others. In other words, some consonants precede certain vowels more frequently than they do other vowels, and similarly, some consonants follow certain vowels more frequently than others. For example, the combination "es" is used much more frequently than the combination "us" (u as in foot). Since the testing lists are made by random selection, the various con-vow and vow-con combinations occur with uniform frequency. In order to determine how this difference influences the interpretation of the sounds, articulation data on various circuits were examined. Attention was focused first on the final consonant sounds. One hundred errors for each consonant,

or 2,200 consonant errors were selected at random from the articulation data and the number of these 2,200 errors that occurred after each vowel sound ascertained. Similarly, the number of vowel errors, out of a total of 1,100 errors, that occurred after each of the consonant sounds, was determined.

Probability studies indicate that the distribution of these errors as shown in Table IX is of the same order as that to be expected on

TABLE IX

Distribution of Vowel and Consonant Errors

Distribution of Vow. Errors		Distribution of Fin. Con. Errors	
Preceding Con. Sound	No. of Vow. Errors	Preceding Vow. Sound	No. of Fin. Con. Errors
b	58	a	188
ch	73	ā	264
d	52	a'	232
f	49	e	166
g	50	ē	214
h	55	i	172
j	54	o	192
k	60	ō	192
l	45	o'	208
m	49	u	196
n	54	ū	176
p	68		
r	39		
s	50		
sh	53		
th	38		
t	52		
v	40		
w	52		
y	49		
z	60		

the basis that the distribution of errors is due entirely to chance. Since, from the way the lists were constructed, the occurrence distribution is due to chance, it is evident that the errors in recognition of the sounds do not depend upon the particular sounds that they follow. Although the analysis was not made, it would be expected that a similar situation obtains for initial consonant and vowel errors. These data may be interpreted to mean that the consonant articulation, the vowel articulation, the sound articulation, or the syllable articulation, is approximately independent of the particular sound combinations, when a wide variety of combinations are used. The results obtained with these lists, therefore, are as representative of speech as the results that would be obtained with lists employing particular sound combinations in proportion to their frequencies of occurrence

in speech. The analysis was not extensive enough to draw conclusions as to the effects of particular sound combinations upon the articulation of individual speech sounds.

Approximately 40 per cent of the syllables that occur in English are of the con-vow-con type. About 34 per cent are of the con-vow, and vow-con type. The syllables, including the compounds, such as, con-con-vow, vow-con-con, con-vow-con-con, and con-con-vow-con, make up about 16 per cent of the syllables of English. Since, as pointed out above, the interpretation of the consonant compounds depends primarily upon only one of the consonants, the latter syllables may be grouped in the two former classes, which then constitute some 90 per cent of English. Of the remaining syllables, 7 per cent consist of a single vowel, so that the more complex syllable forms constitute only 3 per cent of English.[8] Since 97 per cent of the syllables of English are included in the one, two and three letter forms, there is little reason to include the more complex syllable forms in order to represent speech, when as has been previously stated, they are undesirable from a testing standpoint. As will be shown in a later paragraph, one, two and three-letter syllables all yield equal values of articulation for the various speech sounds. Since the three-letter syllables require a smaller testing time for a given number of called sounds, the other syllable forms were excluded from the testing lists.

Having shown that the standard technique gives, for the various sounds, data that are representative of speech, the question now arises as to the best figure that may be computed from the data obtained with this technique, in order to best represent the speech transmission ability of the system under test. Before discussing this, it is necessary to consider some probability relations existing between the quantities entering into the calculation of such a figure.

Statistical Relations

The syllable articulation S when expressed as the ratio of the number of successes (correct interpretations of the syllables) to the number of trials (syllables called) is the chance of perceiving a syllable correctly. Also, if a similar ratio is used for the sound articulation L, the vowel articulation V, and the consonant articulation C, then these letters represent the probability of perceiving correctly a fundamental sound, a vowel sound or a consonent sound, respectively. If a syllable contains only one fundamental sound, then it is obvious that

$$S = L. \tag{3}$$

[8] These data were obtained from Godfrey Dewey's book "Relative Frequency of English Speech Sounds," Harvard University Press, 1923.

If a syllable has two letter sounds, then the chance of perceiving them both correctly is the same as the chance of perceiving the syllable correctly or

$$S = L^2. \tag{4}$$

Similarly, for a syllable containing m sounds

$$S = L^m. \tag{5}$$

Or if A_1, A_2, A_3, $A_4 \cdots A_m$ give the per cent of syllables in the list containing 1, 2, 3, 4, $\cdots$ m sounds, respectively

$$S = A_1L + A_2L^2 + A_3L^3 + \cdots A_mL^m. \tag{6}$$

Similarly, the chance of perceiving a syllable of the type con-vow or vow-con is VC; of the type con-vow-con, con-con-vow or vow-con-con is VC^2; of the type con-con-vow-con, con-vow-con-con, vow-con-con-con, or con-con-con-vow, is VC^3, etc.

For the old standard articulation lists these formulæ reduced to

$$S = \frac{1}{5} VC + \frac{4}{5} VC^2 = \frac{1}{5} L^2 + \frac{4}{5} L^3. \tag{7}$$

For the new standard articulation lists they reduce to [9]

$$S = VC^2 = L^3. \tag{8}$$

If a list of N syllables is used, then the letter errors and syllable errors will be $3N(1 - L)$ and $N(1 - L^3)$, respectively, or the number of letter errors per mistaken syllable, for the new standard lists, will be

$$m = \frac{3}{1 + L + L^2}. \tag{9}$$

It is seen that m approaches 3 as L becomes small, and unity as L approaches unity. For $L = .30$, $m = 2.06$; for $L = .50$, $m = 1.71$;

[9] When derived from the probability formulæ

$$VC^2 = L^3.$$

However, from the definition of V, C and L,

$$L = (2C + V)/3 \quad \text{so that} \quad VC^2 = L^3.$$

The difference is

$$d = \frac{(V + 8C)(V - C)^2}{27}.$$

Actually V and C are not wholly independent of each other and when values as obtained in tests are substituted in the above equation, the difference turns out to be small.

and for $L = .80$, $m = 1.23$. When observed values of m become consistently greater or less than this theoretical value, it must be concluded that the assumptions underlying this statistical theory are not valid.

All of the above statistical relations are dependent upon the tacit assumption that the chance of perceiving any sound correctly is entirely independent of the other sounds present and also independent of the number of other sounds present. It was shown in the previous section that the articulation of the various sounds is, on the average, independent of the other sounds in the syllables. On the other hand, experiments have indicated that the articulation does depend upon the number of sounds in the syllable. The sound articulation becomes smaller when the number of sounds in the syllables increases beyond three per syllable.

The data from which this conclusion was drawn were taken from three different experiments. In the first, three different transmission systems were tested by using first the standard articulation lists and then the vowel-consonant lists which are described in the last section. When using the vowel list, the vowels only are considered and when using the consonant list the consonants only are considered. These lists together, then may be considered as composed of syllables having only one sound. The syllable and sound articulations are the same when using such lists. The comparison of the results obtained with the two types of lists is shown in Table X. It will be seen that there

TABLE X

ARTICULATION FOR ONE- AND THREE-SOUND SYLLABLES

	Freq. below 1,000~ only			Freq. below 1,950~ only			Freq. above 1,500~ only		
	Vow. Art.	Cons. Art.	Sound Art.	Vow. Art.	Cons. Art.	Sound Art.	Vow. Art.	Cons. Art.	Sound Art.
One-sound Syllables	71.5	62	65	98.5	82.5	88	81	96	91
Three-sound Syllables	69.5	61.5	64	96	83	87	80	96.5	91

is only a slight tendency for the sound articulation to be lower for the three-sound syllable when compared with the one-sound syllable. The differences are within the observational error in testing.

In the second experiment a new list was constructed using the syllable forms con-vow and vow-con. The auxiliary circuit of the master reference system was tested with this list and also with the standard articulation list. The results are shown in Table XI. The

TABLE XI

ARTICULATIONS FOR TWO- AND THREE-SOUND SYLLABLES

	V	*C*	*L*
Two-sound Syllables	95	88	90
Three-sound Syllables	94	87	89

number of syllables used of each type for determining these averages was 1,344. The average sound articulation in each case was determined by giving equal weights to the articulation for each sound. For the three-sound syllables this is done by dividing the number of sounds correctly recognized by the total number of sounds called. For the two-sound syllable the procedure is not so simple. Since each vowel sound occurs twice as often as each consonant sound, it is necessary to obtain an average for the vowels and consonants

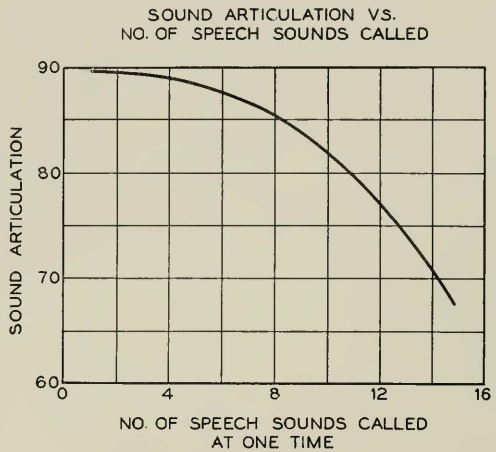


Fig. 8

separately. The final average value for the sound articulation is obtained by assigning weights of 1 and 2 to the vowel and consonant articulations, respectively. It is seen that there is no appreciable difference between the values of *L* obtained by the two types of lists.

In the third experiment the standard lists were used to test the auxiliary circuit but the syllables were called in groups of 1, 2, 3, 4, or 5 at a time. The results of these tests are shown in Table XII.

TABLE XII

Number of Sounds Called at One Time	Sound Articulation
3	89.0
6	87.5
9	84.0
12	77.0
15	67.0

From these three sets of data and from other available data which could be applied to this problem, the curve shown in Fig. 8 was constructed. It gives the sound articulation which would be obtained for a circuit such as the auxiliary circuit of the master reference system, when the number of sounds that are spoken at a time, that is, before the observer starts writing, is represented by the abscissa. It is evident from the shape of this curve that the assumptions underlying the statistical formulæ are valid for syllables having three or less sounds per syllable, and that they will break down for the more complex types of syllables. These assumptions might be expected to break down also, for certain extreme types of distortion.

Definite relations between the vowel, consonant, sound, and syllable articulations for both the old and the new techniques, have been derived by statistical theory. An experimental relationship between these quantities is shown in Figs. 9 and 10. These were obtained by an analysis of the errors of a large number of tests with widely different types of distortion, the data in Fig. 9 being taken with the old and the data in Fig. 10 with the new technique.

In the figures observed values of sound articulation have been plotted against the corresponding observed syllable articulation values. The solid curves in the two cases were calculated from Equations 7 and 8, respectively. The observed values agree reasonably well with the theoretical curves.

There is very little correspondence between the vowel and syllable or consonant and syllable articulation. The table below shows that

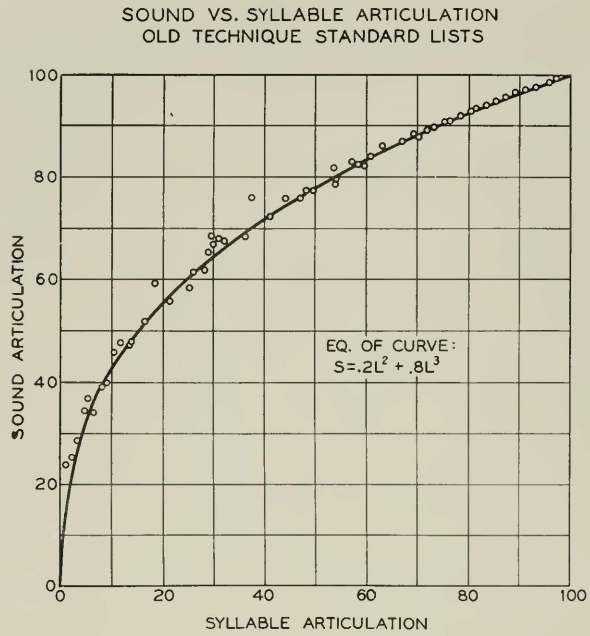
TABLE XIII

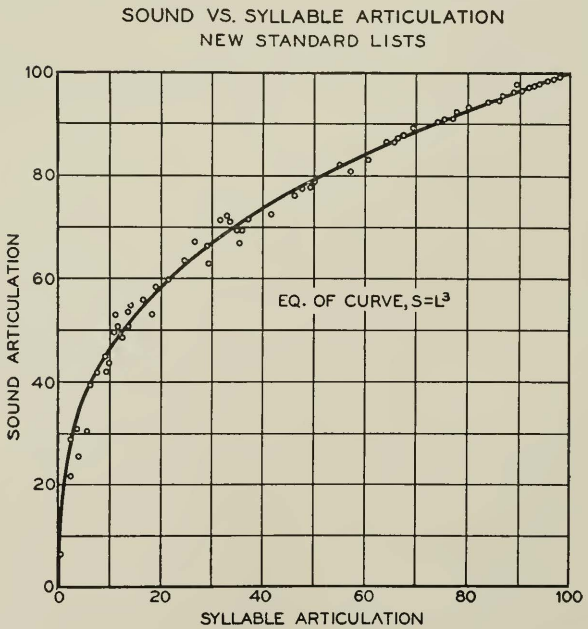
Vowel, Consonant, and Syllable Articulations

	V	C	S	VC^2
3,750 L.P.F.	98.7	95.4	90.8	89.8
750 H.P.F.	93.0	98.6	90.9	90.4
2,850 L.P.F.	98.5	92.9	86.3	85.4
1,000 H.P.F.	89.0	98.4	87.0	86.2

a circuit which discriminates against the vowels may have a syllable articulation equal to another circuit which discriminates against the consonants. However, it is seen that the product VC^2 is equal to S as the statistical theory indicates. If then *the sound, or vowel and consonant articulations are known, it is possible to calculate the syllable articulation*, for the case of two- and three-sound syllables.

We are now in a position to consider the figure which best represents





Figs. 9 and 10—Relation between sound and syllable articulation

the capabilities of systems to transfer thought by means of speech. For giving a complete picture, it is necessary to give the articulation values for each speech sound. Since this involves 36 articulation values, it is difficult to compare various systems. To combine these values into averages raises the question of how such an average shall be taken. At first thought it might seem obvious that the weights assigned to each sound articulation value should be proportional to the frequency of occurrence of that sound in English speech. Many of the most frequently occurring words, however, such as the, of, and, to, in, that, etc., do not carry much of the thought, so that it seems reasonable to exclude the effects of such words in the weighting process. It is evident that many sets of weighting factors could be evolved depending upon how far the exclusion process is carried and depending upon whether written or spoken English is used, in determining the frequencies of occurrence of the sounds. After excluding the twenty or twenty-five most common words, however, further exclusion does not appreciably change the calculated articulation value. The table below gives a set of factors obtained from the frequencies of occurrence used in Table VIII. They are based upon the studies of Messrs. French and Koenig [10] on the frequencies of occurrence of speech sounds in spoken English. The effects of the more common parts of speech, such as, personal pronouns, definite articles, conjunctions, and prepositions have been excluded.

TABLE XIV

Group I	Weight	Group II	Weight	Group III	Weight	Group IV	Weight	Group V	Weight
ā	3.0	i	5.8	r	6.3	d	3.5	z	1.3
ē	4.3	o	2.5	l	6.1	t	7.8	s	4.3
ō	2.5	a'	2.0	ng	2.0	b	2.0	v	1.3
ū	1.3	u	.8	n	6.6	p	2.5	f	2.0
o'	2.3	e	6.0	m	4.5	g	2.3	zh	.3
a	2.8	y	.5			k	4.5	sh	1.5
		w	2.0			j	.8	th'	.8
						ch	.8	th	1.0
						h	1.0	st	1.0
Total Weight	16.2		19.6		25.5		25.2		13.5

It will be noticed that the speech sounds are arranged in five groups. The sounds in each group have very similar characteristics, so instead of dealing with 36 articulation values for a circuit, it is only necessary

[10] "Frequency of Occurrence of Speech Sounds in Spoken English," N. R. French & W. Koenig, *Proc. Acoustical Society of America*, 1929.

to deal with the average value for each of the five groups. The average for the first group is designated V_l, signifying long-vowel index; for the second group V_s, signifying short-vowel index; for the third group C_n, signifying nasalized-consonant index; for the fourth group C_s, signifying stop-consonant index; for the fifth group C_f, signifying fricative-consonant index. If the articulation obtained from any test for each sound be designated by the phonetic symbol for that sound, then,

$$\left.\begin{aligned} V_l &= .19\ \bar{a} + .27\ \bar{e} + .15\ \bar{o} + .08\ \bar{u} + .14\ o' + .17\ a \\ V_s &= .30\ i + .13\ o + .10\ a' + .04\ u + .31\ e + .02\ y + .10\ w \\ C_n &= .25\ r + .24\ l + .08\ ng + .26\ n + .17\ m \\ C_s &= .14\ d + .31\ t + .08\ b + .10\ p + .09\ g + .18\ k + .03\ j \\ &\qquad + .03\ ch + .04\ p \\ C_f &= .10\ z + .32\ s + .10\ v + .15\ f + .02\ zh + .11\ sh \\ &\qquad + .06\ th' + .07\ th + .07\ st. \end{aligned}\right\} \quad (10)$$

The sound index is related to these values by

$$i = .162\ V_l + .196\ V_s + .255\ C_n + .252\ C_s + .135\ C_f. \quad (11)$$

For obtaining the most representative single value for the syllable index I, the equation given below is used.

$$I = .5\ i^2 + .5\ i^3. \quad (12)$$

This equation is based upon the frequency of occurrence of the syllable forms in English speech. As pointed out before, if the compound consonants be considered as simple sounds, then there are less than 10 per cent of syllable forms other than the two- and three-sound type. The frequency of occurrence of these two types is approximately equal.

Similar formulæ to the above may be used to relate articulation results in English to articulation results in a different language. To do this it is necessary to select the fundamental sounds of the different languages that correspond to the 36 fundamental sounds of English, where the correspondence is based on similar phonetic characteristics and similar positions of the vocal organs in producing the sounds. When this is done, the coefficients in Eqs. 10, 11 and 12, must be modified to correspond with the frequencies of occurrence of the sounds and syllables in the language.

Observed values of individual sound articulation are thus reduced

to a single index, or for a more comprehensive picture, to five indices corresponding to the five groups of speech sounds. In order to compare the indices obtained by a given crew with those of a reference crew, it is necessary to correct the data in accordance with Eq. 2 for the effects of practice. To do this, as previously discussed, articulation tests are made upon one or more of the reference circuits by the crew in question. If I' is the syllable index so obtained, the practice factor for the crew is given by the relation

$$x = \frac{\log (1 - I')}{\log (1 - I)} . \tag{13}$$

The practice factors for the other indices may be obtained also, by substituting the appropriate indices for the syllable index in Eq. 13. In Table XV the reference values for the various indices are given for the reference circuits that were previously described.

TABLE XV

REFERENCE VALUES

Circuit	V_l	V_s	C_n	C_s	C_f	i	I
Master Reference System	98.5	98.9	99.6	99.2	98.8	99.3	98.0
Auxiliary Circuit of Master Ref. Sys.	95.0	95.0	96.5	88.5	66.5	90.0	77.0
Carbon Transmitter Circuit	97.0	97.0	96.5	93.5	82.0	94.0	85.5
Master Ref. Sys. plus 3,750- L.P.F.	99.0	99.5	99.5	99.0	86.5	97.6	94.0
" " " " 750- H.P.F.	96.0	92.5	99.0	99.5	98.5	97.1	93.0
" " " " 1,500- L.P.F.	93.5	86.5	90.5	76.0	52.5	80.2	58.0
" " " " 1,500- H.P.F.	85.0	82.5	96.0	97.5	97.0	92.0	81.0

If the values for the sound index be compared with the sound articulation values based on uniform weighting, that were given under the section on practice effects, it will be seen that for these circuits there is very little difference, between the two sets of values. In other words, the average sound articulation is very nearly equal to the average that is obtained when the individual sound articulations are weighted according to the frequencies of occurrence of the sounds in English.

Similar comparisons have been made for a large number of other transmission systems. They showed similar small differences between the weighted and unweighted averages. For this reason we consider it unnecessary to use the weighted average when great accuracy is not required, for example, in a great deal of our routine work where comparisons are being made between circuits which have similar characteristics. This means that when testing an unknown circuit,

having an electrical characteristic similar to one of the reference circuits, the syllable index I can be calculated from the observed syllable articulation S (as obtained with the new standard lists) by means of the equation,

$$I = .5\ S^{2/3} + .5\ S. \tag{14}$$

This value must now be reduced to the reference condition of practice by the methods which have already been described. In such cases it is thus possible to obtain the syllable index from the observed syllable articulation values, and it is unnecessary to analyze the data for the individual sound articulation values.

The weighted average, however, is the more logical way of obtaining a single index and should be used when it is suspected that it might give results which are essentially different from the unweighted average.

It is possible to carry the probability relations a step further and apply them to cases of English words and sentences. In order to do this it is necessary to make assumptions as to how the thought or meaning of the words affects the interpretation of the sounds. These assumptions are not only somewhat uncertain, but owing to psychological factors in testing are difficult to verify experimentally. In general, the meaning associated with words makes them easier to interpret than meaningless words. For single-syllable words, these effects are small. Two-syllable words are easier to interpret than single-syllable words. The interpretation of words containing from three to five syllables, and short sentences, depends almost entirely upon interpreting those parts which are not indicated by the thought or meaning.

Other Testing Methods

For most articulation studies it has been found desirable to use the standard testing technique which has been described, but it is frequently necessary, in special cases, to use other techniques. In such cases it is desirable, if possible to interpret the results in terms of the standard technique. In the course of research work, several different articulation testing methods have been used which give information on the type of correlation between them that may be expected.

The probability relations have been made use of in constructing two other types of lists which are called vowel-consonant and vowel word-consonant word lists. These lists are designed to give the same values of sound articulation as given by the standard lists. The former lists are shown in Table XVI. The various vowels are combined with the same consonant, and the various consonants with the

TABLE XVI

VOWEL LIST

Sound to be Graded	Testing Syllables in the List	
a	at	ta
ā	āt	tā
a'	a't	ta'
e	et	te
ē	ēt	tē
i	it	ti
o	ot	to
ō	ōt	tō
o'	o't	to'
u	ut	tu
ū	ūt	tū

CONSONANT LIST

Sound to be Graded	Testing Syllables in the List					
b	bū	ūb	ba	ab	bē	ēb
d	dū	ūd	da	ad	dē	ēd
f	fū	ūf	fa	af	fē	ēf
g	gū	ūg	ga	ag	gē	ēg
k	kū	ūk	ka	ak	kē	ēk
l	lū	ūl	la	al	lē	ēl
m	mū	ūm	ma	am	mē	ēm
n	nū	ūn	na	an	nē	ēn
r	rū	ūr	ra	ar	rē	ēr
p	pū	ūp	pa	ap	pē	ēp
s	sū	ūs	sa	as	sē	ēs
sh	shū	ūsh	sha	ash	shē	ēsh
th'	th'ū	ūth'	th'a	ath'	th'ē	ēth'
th	thū	ūth	tha	ath	thē	ēth
t	tū	ūt	ta	at	tē	ēt
v	vū	ūv	va	av	vē	ēv
ch	chū	ūch	cha	ach	chē	ēch
z	zū	ūz	za	az	zē	ēz
j	jū	ūj	ja	aj	jē	ēj
h	hū		ha		hē	
w	wū		wa		wē	
y	yū		ya		yē	
zh		ūzh		azh		ēzh
ng		ūng		ang		ēng
st		ūst		ast		ēst

same vowel. The technique of using the list is the same as that previously described, except that the vowel articulation and consonant articulation are measured separately. Only the vowel errors are counted when using the vowel list and only the consonant errors when using the consonant list. These lists have the advantage that they can be used over and over by merely changing the sequence of the syllables.

Table XVII shows two lists similar to the above except that they are made up entirely of common English words. They are designated as vowel word and consonant word lists. This list is used in the same way as the vowel and consonant lists. In using either of these lists

the testing crews should be familiar with the syllables or words in the lists.

TABLE XVII

VOWEL WORD LIST (ENGLISH WORDS)

Sound to be Graded	English Words in the List	
a'	bat	back
ā	bait	bake
e	bet	beck
ē	beat	beak
i	bit	bit
ī	bite	bike
o	but	buck
o'	bought	balk
ō	boat	boat
u	book	book
ū	boot	boot

CONSONANT WORD LIST (ENGLISH WORDS)

Sound to be Graded	English Words in the List	
b	by	by
ch	which	which
d	die	die
f	fie	whiff
g	guy	wig
h	high	high
j		
k	wick	wick
l	lie	will
m	my	whim
n	nigh	win
ng	wing	wing
p	pie	whip
r	wry	wry
s	sigh	sigh
sh	shy	wish
th'	thy	with
th	thigh	thigh
t	tie	wit
v	vie	vie
w	why	why
y		
z	whiz	whiz
st	sty	whist

Note: The h following w is not pronounced in such words as whim, whip, etc.

It usually requires a training period of a month or more for a testing crew to thoroughly master the technique of using the standard lists, that is, to reach a stage where the phonetic symbols are spoken and recorded almost mechanically. The vowel consonant lists require less time, since it is only necessary for the observers to fix their attention on one sound in the syllable. With the word lists this training period is reduced to a minimum. Phonetic symbols are avoided, and attention is given to only one sound in the words. As may be seen from Fig. 2, after a few tests they practically reach a degree of

uniform proficiency. In using the lists, the words are recorded with the English spelling. Only errors in the vowel and consonant sounds of the left-hand column of the above table, are counted. Since only one sound in each syllable is utilized, the above lists require a somewhat greater testing time for a given precision than do the standard lists where all three sounds of the syllables are used.

Table XVIII below, shows data that were obtained with the three types of lists, namely, the standard lists, the vowel consonant lists, and the vowel word consonant word lists. The vowel consonant

TABLE XVIII

ARTICULATION RESULTS WITH VARIOUS LISTS

Circuit	Cons. Vow. List		Cons. Word Vow. Word List		Standard List			Corrected
	C	V	C_w	V_w	S	VC^2	$V_wC_w^2$	$V_wC_w^2$
Air Transmission	99.5	99.0	99.0	99.5	98.0 ± .4	98.0	98.0	97.5
Master Ref. System + 5500 L.P. Fil.	99.0	99.0	99.0	99.5	96.5 ± 1.0	97.0	97.5	97.0
M.R.S. + 3750 L.P. Fil.	95.5	98.5	96.5	99.0	90.5 ± 2.0	90.0	92.0	90.0
M.R.S. + 1950 L.P. Fil.	82.5	98.5	85.5	98.5	67.0 ± 2.0	67.5	72.0	68.0
M.R.S. + 1000 L.P. Fil.	62.0	71.5	67.0	82.0	29.0 ± 3.0	27.5	37.0	33.0
M.R.S. + 1500 H.P. Fil.	96.0	81.0	97.0	86.0	75.0 ± 2.0	75.0	81.0	77.0
Carbon Transmitter Circuit	91.0	96.0	91.5	99.0	81.0 ± 1.5	80.0	83.0	80.0

lists lead to syllable articulations that are essentially the same as obtained with the standard lists. It is evident from the data that the word lists lead to slightly higher values of syllable articulation. The explanation is to be found in the make up of the word list. It was not possible to arrange the lists so that all sounds are equally probable, and still use English words. For instance, an observer would not record "bik" for "beck," nor "wiv" for "with," etc. For this reason, the observers do not make errors which occur frequently in the other two types of lists, where any sound is possible. Hence the observed articulations are somewhat higher with the word lists.

It was found, however, that the word lists may be correlated with the standard lists by means of the relation given by Equation 1. The value of x for this case is 0.9 so that, the word technique may be corrected to the standard technique, by means of the equation.

$$S = 1 - (1 - V_wC_w^2)^{\cdot 9}, \tag{15}$$

where S = syl. art. of standard lists expressed as a ratio,

V_w = vowel art. of vowel word lists expressed as a ratio,

C_w = cons. art. of cons. word list expressed as a ratio.

The corrected values are also shown in the above table.

It is frequently necessary to test very poor systems where the standard lists giving an articulation of a few per cent, are not satisfactory. The vowel consonant lists are somewhat more satisfactory under these circumstances. Lists of sentences have also been found to be very useful for such purposes. The sentences were of the interrogative or imperative form containing a simple idea. They were designed to test the observer's acuteness of perception rather than his intelligence. Tests were made with these sentences and the standard lists on various circuits, involving carbon transmitter circuits and various filter systems. The data are shown in Fig. 11. The

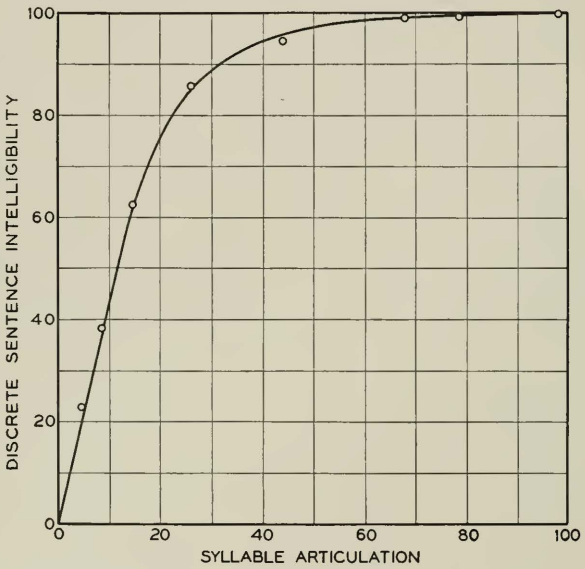


Fig. 11—Discrete sentence intelligibility vs. articulation

sentences were considered to be understood if the observer either recorded the sentence correctly or recorded an intelligent answer. As stated earlier, the percentage correctly observed is called the discrete sentence intelligibility.

It will be seen that for changes in distortion, the changes in the discrete sentence intelligibility will be small for systems having syllable articulations greater than 30 per cent, but very large for systems having syllable articulations below 20 per cent. It is for systems in this latter class that these test sentences are useful. A case in point is the measurement of the degree of secrecy obtained

in sound proofing telephone booths, or in dealing with cross-talk. The sentences have also been found to be useful in making quick qualitative tests of the goodness of an audiphone set for a particular case of deafness.

Because of their general usefulness for these purposes, the complete lists of sentences are given in the appendix. Due to memory effects a set of sentences can be used with the same personnel only a very few times. The psychological factors are also more prominent with sentences than with simple syllable.

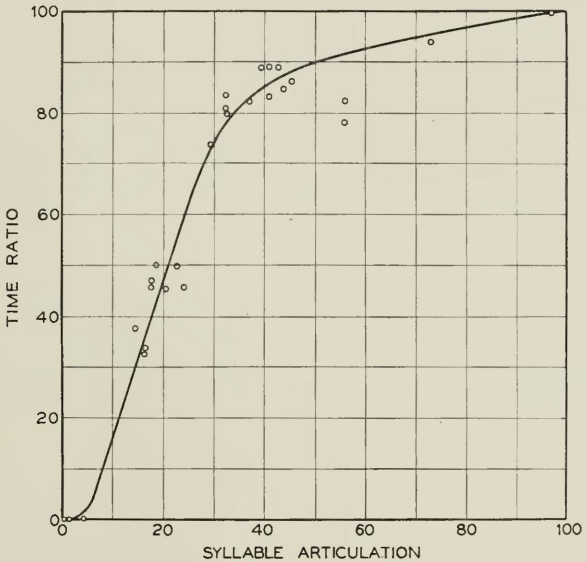


Fig. 12—Articulation vs. time ratio

Sentence lists of the above type have also been used to obtain a notion of how the time taken to transmit an idea correctly over a system depends upon the articulation. To do this, the observer was instructed to reply orally to the question. If the reply indicated that the observer failed to understand, the speaker repeated the question. Both speaker and observer tried to carry out the test in a normal conversational manner. The observer could ask the speaker to repeat, reword or spell out difficult parts of the sentence.

The tests were made on a variety of systems of known syllable articulation. The results that were obtained are shown in Fig. 12. The ordinates of the curve give the ratio of the time required to

transmit correctly one of these test sentences over an ideal system to the time required over the system under test. With the crew used in making these tests, and with an ideal transmission system, it required an average time of 5.2 seconds after the speaker started to pronounce the sentence before the observers grasped the idea. It will be seen from the curve that for systems having approximately 20 per cent articulation, the time required is twice as great. Fig. 11 shows that one out of every four of the sentences is mistaken for this value of articulation. If it is assumed that an observer asks that only sentences which he fails to understand be repeated, it can be shown that this time ratio is equal to the discrete sentence intelligibility.[11]

It is evident from Figs. 11 and 12, that the observed time ratio is appreciably less than the discrete sentence intelligibility. This difference may be taken to indicate that an observer not only asks that sentences which he fails to understand be repeated, but also that sentences about which he is uncertain be repeated. In other words, the time element reflects both factors, the understandability and the uncertainty.

As has been previously mentioned, tests have been made with various types of English word lists. Because of the manner in which the words were selected, and also due to uncertain psychological factors entering into the tests when such words are used repeatedly, it is difficult to compare the results so obtained with syllable articulation results.

However, it was found that if a definite rule were followed in selecting words from a newspaper, consistent results could be obtained with lists containing 500 or more words per list. The method of selection was to take the first word from every third line of a newspaper column. In this selection all proper names and the following six most frequent words of English were excluded, the, of, and, to a, in. When a word was hyphenated from the previous line, the whole word was used. Each of eight callers called a list of 66 words to four observers in the manner of an ordinary standard articulation test. Tests were made with the carbon transmitter circuit and the six circuits indicated in Fig. 3. The data were analyzed to give the discrete word intelligibilities for the one, two, three, four, and five-syllable words occurring in the lists, as well as for the lists as a whole. The lists on the average contained 46.3 per cent one-syllable, 29 per cent two-syllable, 16.8 per cent three-syllable, 6.4 per cent four-

[11] "A Theoretical Study of Articulation and Intelligibility of a Telephone Circuit," John Collard, *Electrical Communication*, 7, page 168, January, 1929.

syllable, and 1.5 per cent five-syllable words, and an average number of two syllables per word. The discrete word intelligibility vs. syllable articulation as obtained with the standard lists is shown in Fig. 13. The dashed curves indicate the relations for the various types of words, and the solid curve for the word lists as a whole. The data for two syllable words practically coincided with the solid curve. Owing to the small amount of data, the curves for the four- and five-syllable words are less reliable than those for the other types. Curves of the above type, both for words and sentences, depend very

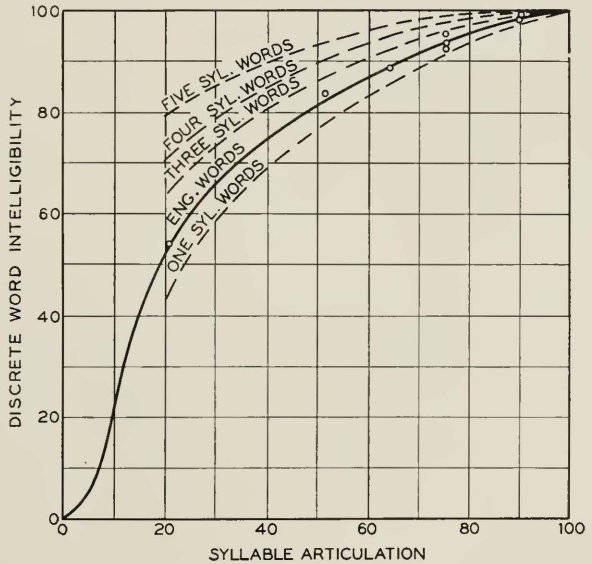


Fig. 13—Discrete word intelligibility vs. syllable articulation

much upon the way the speech material is selected. If, for example, only "different" words had been included in the word lists, appreciably smaller values of discrete word intelligibility would have been obtained.

Tests have also been made with lists made up of the following numbers, 1, 2, 3, 4, 5, 6, 8. These numbers were combined at random into groups of three and called in the manner of an ordinary articulation syllable. The distinguishing characteristic of each of the above numbers is a vowel sound, so that, they are interpreted primarily from recognizing the vowel. Such lists, therefore, do not give a very good picture of the speech capabilities of a system which distorts speech. They are, however, very useful in measuring the deafness

of an observer, for the reason that the number articulation decreases very rapidly as the sounds approach the threshold of hearing. As may be seen from Fig. 14, the number articulation passes from practically 100 per cent to 0 per cent in the short range of 10 or 15 db. It is evident that such lists give a critical measure of the point at which

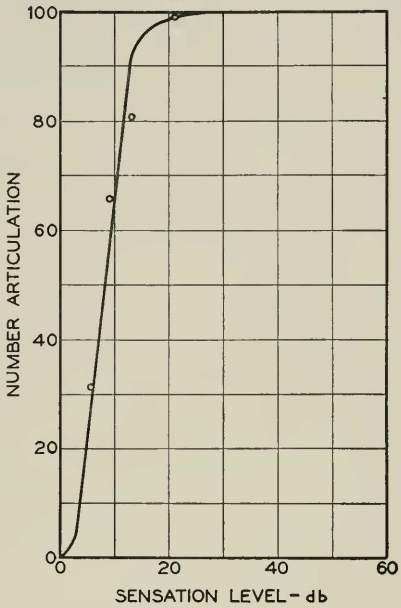


Fig. 14—Number articulation vs. sensation level

an observer fails to hear the sounds. Lists of this type have been used extensively in testing the hearing of school children.

Summary

The standard testing technique is primarily a means of determining the articulation or recognizability of the individual speech sounds when they are spoken in a way that is representative of conversational speech, and in a way which facilitates the carrying out of articulation tests. The articulations of the individual sounds may be converted into an index which indicates the speech capabilities of a system. Other types of lists which yield either the recognizability of speech sounds, or the intelligibility of discrete English words and sentences containing thought, have been described and experimentally correlated with the syllable technique.

It should be emphasized that there may not be a one to one corre-

spondence between all of these measured quantities for all types of speech distortion. The data entering into the sound, vowel, consonant, and syllable articulation curves were very extensive with respect to types of distortion and testing personnel. The theoretical equations relating them seem to rest upon assumptions with few uncertainties. For these reasons, it is felt that these relations can be used with considerable confidence, especially for values of syllable articulation greater than 30 per cent. The curves dealing with English words and sentences are based upon less diversified data and should be regarded as indicative only of the correlation and the type of relation.

During the past few years articulation testing methods have been used more and more, both in this country and abroad. It is felt that in order to compare the results obtained by various crews in various tongues, *it is desirable to use techniques that operate on the same basic principles and to calibrate various crews on similar reference circuits.*

INTELLIGIBILITY LIST

List 1

1. Name a prominent millionaire of the country.
2. How large is the sun compared with the earth?
3. Why are flagpoles surmounted by lightning rods?
4. Give the abbreviation for January and February.
5. Name the tree on which bananas grow.
6. How often does the century plant bloom?
7. What description can you give of the bottom of the ocean?
8. Explain the difference between a hill and a mountain.
9. What is the chief purpose of industrial strikes?
10. Describe the shoes of the native Hollander.
11. Name some uses to which electricity is put.
12. What would cause the air to escape from a bicycle tire?
13. Where is more grain raised, in the East or the West?
14. Tell what is meant by an Indian Reservation.
15. For what invention is Thomas Edison noted?
16. Name a state which has no seacoast.
17. Write the Roman numeral ten.
18. Explain the difference between export and import.
19. Explain why a corked bottle floats.
20. What substance is a good conductor of electricity?
21. Explain why Indians were afraid of firearms.
22. Explain the purpose of fire drills.
23. At what time do ocean waves become dangerous?
24. What medicine would you take to remedy indigestion?
25. What knowledge is covered by the study of astronomy?
26. Name a good restaurant in this vicinity.
27. What is the importance of large windows in stores?
28. Explain why a giraffe eats the foliage of trees.
29. How are the pages of a magazine held together?
30. Explain why the name string-bean is appropriate.
31. Name a nearby city in which there is a shipyard.

32. Name a fruit which grows in bunches.
33. Which of our Presidents went to South Africa?
34. Why are wire springs used in beds?
35. Why are books bound in stiff covers?
36. Why did the home people conserve food during the war?
37. Name an insect that has a hard shell.
38. What symbol on the United States money stands for liberty?
39. What weapons did the Indians use in warfare?
40. In what kind of weather does milk sour?
41. What streets in this city have Dutch names?
42. How does turning a ship's wheel steer the ship?
43. What nation aided us in the Revolutionary War?
44. What are some personal characteristics of the people of Japan?
45. What candy is black and good for colds?
46. Name a famous Indian Tribe.
47. Why is this building lighted by reflected light?
48. Why are most lighthouses situated on rocks?
49. Give some ingredients used in soap.
50. Why is a house built of stone superior to others?

Abstracts of Technical Articles From Bell System Sources

Reciprocal Theorems in Radio Communication.[1] JOHN R. CARSON. Two reciprocal theorems, the generalized Rayleigh theorem and the Sommerfeld-Pfrang theorem, are of great theoretical importance in radio communication. A careful analysis of these theorems and their mathematical derivations shows that they are quite distinct and their practical fields of application different. In particular it shows that the Sommerfeld-Pfrang theorem labors under restrictions, implicit in its mathematical derivation, which seriously limit its field of practical applicability.

Telephone Circuits for Program Transmission.[2] F. A. COWAN. Systems of telephone circuits which are extensively used in the transmission of programs to broadcasting stations are described in this paper. Certain stages in the development of these networks are considered and the general requirements for satisfactory transmission at the present time are enumerated. The arrangements of the networks as well as the procedures used in setting up and maintaining them are discussed.

Correlation of Directional Observations of Atmospherics with Weather Phenomena.[3] S. W. DEAN. This paper analyzes some data on the direction of arrival of static at Houlton, Maine, obtained by means of a recorder and a cathode ray radio direction finder; and points out that in certain cases there is a relation between the direction of static and the location of storm centers. Two cases are discussed in which day by day bearings showed static sources in the direction of moving storm centers.

A Direct-Current Amplifier for Measuring Small Currents.[4] J. M. EGLIN. A direct-current amplifier consisting essentially of a Wheatstone bridge, having the amplifying tube in one arm and a balancing tube in another, has been described by P. I. Wold and by C. E. Wynn-Williams. This circuit has now been developed to give a constant amplification for currents in either direction up to 10,000 times the lowest measurable value. The amplification and the lowest

[1] *Proceedings of the Institute of Radio Engineers*, Vol. 10, June, 1929, pp. 952–956.
[2] *A. I. E. E. Journal*, July, 1929, pp. 538–542 (abridgment).
[3] *Proceedings of the Institute of Radio Engineers*, Vol. 17, July, 1929, pp. 1185–1191.
[4] *Journal of the Optical Society of America and Review of Scientific Instruments*, Vol. 18, May, 1929, pp. 393–402.

measurable current are alterable together by changing the resistance introduced between the grid and filament of the amplifying tube. With tubes of high insulation, the amplification can be made as large as 10^6; and the measurable current as low as 10^{-14} ampere. Some improvements of the circuit are: (1) the insertion of a resistance in series with the tube in one arm of the bridge to "compensate" for variations in plate and grid battery voltages; (2) the suspension of the tubes to protect them from mechanical vibrations; (3) the use of tubes with pure tungsten filaments to avoid changes in contact potentials, and with plates enclosing the filaments completely to lower the effects of wall charges. In a "null" method of using the circuit the values of the grid resistance and an auxiliary potential introduced in the grid-filament circuit are sufficient to determine the measured current.

Meeting Long Distance Telephone Problems.[5] H. R. FRITZ and H. P. LAWTHER, JR. There have been written many papers describing various technical and apparatus developments of value in providing long distance telephone service. Several papers have also appeared covering specific transmission or operating problems, or dealing with the advance planning of the telephone plant. Feeling that it might be of interest, particularly to the young engineering graduates, the writers have prepared this over-all sketch of the general problem of actually providing, year by year, the extensions and additions to a comprehensive network of communication channels necessary to keep pace with a growing public demand for long distance service. Since the writers are most familiar with the area served by the Southwestern Bell Telephone Company, the discussion is restricted to that territory.

Some Measurements on the Directional Distribution of Static.[6] A. E. HARPER. The utility of directional data on static is shown, and two types of apparatus devised for such a directional investigation are compared. It is shown that a method which gives the direction of individual crashes is superior to integrating methods. The distribution of thunderstorms over the world is discussed, and comparisons are drawn between this distribution and the observed directional distribution of static. Probable geographical locations are assigned to the sources, based upon thunderstorm data and directional observations.

[5] *A. I. E. E. Journal*, July, 1929, pp. 547–550 (abridgment).
[6] *Proceedings of the Institute of Radio Engineers*, Vol. 17, July, 1929, pp. 1214–1224.

Maximum Excursion of the Photoelectric Long Wave Limit of the Alkali Metals.[7] HERBERT E. IVES and A. R. OLPIN. Earlier experiments have shown that the long wave limit of photoelectric action in the case of thin films of the alkali metals varies with the thickness of the film. A maximum value is attained greater than that for the metal in bulk, which for the majority of the alkali metals lies in the infra-red. The wave-length of the maximum excursion of the long wave limit was first studied for Na, K, Rb and Cs. In each case it was found to coincide with the first line of the principal series, i.e., the resonance potential. If this relation holds for lithium, its maximum long wave limit should be greater than that of sodium. This was tested and confirmed by experiments in which red-sensitive lithium films were prepared, sensitive to 0.6708 μ. It is suggested that photoelectric emission is caused when sufficient energy is given to the atom, to produce its first stage of excitation. The identity of photoelectric and thermionic work functions suggests that atomic excitation is the initial process in thermionic emission as well.

Magnetic Testing Furnace for Toroidal Cores.[8] G. A. KELSALL. When making magnetic tests at high temperatures trouble is often experienced in maintaining the insulation between turns of the magnetizing and exploring windings and between the windings and the test sample.

This paper describes a magnetic testing furnace for toroidal specimens which eliminates these difficulties. By means of this furnace the test sample may be passed through a definite temperature cycle and the variation in magnetization for a constant magnetizing force determined or the temperature may be held constant while measurements are made for the B-H curve or for a hysteresis loop.

Electrical Wave Analyzers for Power and Telephone Systems.[9] R. G. MCCURDY and P. W. BLYE. This paper describes two types of electrical analyzers which have been developed for the direct measurement of harmonic components of voltage and current on power and telephone systems. These devices are assembled mechanically in a form suitable for use either in the laboratory or in the field. Both instruments, which differ chiefly with respect to sensitivity and input circuit arrangement, employ multistage vacuum tube amplifiers and two duplicate interstage selective circuits.

[7] *Physical Review*, Vol. 34, July 1, 1929, pp. 117–128.
[8] *Journal of the Optical Society of America and Review of Scientific Instruments*, Vol. 19, July, 1929, pp. 47–49.
[9] *A. I. E. E. Journal*, Vol. 48, June, 1929, pp. 461–464 (abridgment).

The power circuit analyzer is designed to measure harmonic voltages in the frequency range from 75 to 3000 cycles and over a voltage range from 0.5 millivolt to 50 volts. The telephone circuit analyzer operates over the same frequency range and measures harmonic currents as low as 0.05 microampere and voltages as small as 0.005 millivolt. Both analyzers are adapted to measure small harmonic voltages and currents in the presence of the fundamental component and other harmonics relatively large in magnitude.

A number of devices are described which have been adopted for eliminating various sources of error. The paper presents in detail the characteristics of both instruments with respect to selectivity, sensitivity, linearity, balance of input with respect to ground, generation of harmonics, and susceptiveness to stray fields.

Solution to a Problem in Diffusion in Employing a Non-Orthogonal Sine Series.[10] R. L. PEEK, JR. In this paper there are developed the equations applicable to diffusion through a membrane between a chamber in which a constant pressure of the diffusing material is maintained and a second closed chamber, initially evacuated, into which the material diffuses. Assuming Fick's law to apply, a solution is obtained in the form of an infinite series of a type similar to those applying to other problems in diffusion. The sine series to which the solution reduces at zero time is non-orthogonal, but it is shown that by a modification of Fourier's method the coefficients of the terms may be directly determined. There is included a proof of the convergence of the series considered.

Telephone Transmission Networks.[11] *Types and Problems of Design.* T. E. SHEA and C. E. LANE. In this paper is given a brief résumé of the nature of telephonic signals showing how the qualities of wave composition which distinguish signals from other electrical waves set the requirements on networks and provide a basis for their design. The principal functions of wave filters, equalizers, telephone transformers, line balancing networks, and artificial lines are outlined. In order that these networks may be used in conjunction with other apparatus in the telephone system they must provide efficient transmission, low distortion, good impedance balance, stoppage of longitudinal currents, stable characteristics with current variations, low external coupling, and low reflection coefficient. In addition to these

[10] *Annals of Math.*, 2d Series, Vol. 30, April, 1929, pp. 265–269.

[11] Presented at the Regional Meeting of the South West District No. 7, of the A. I. E. E., Dallas, Texas, May 7–9, 1929. Abridgment in *A. I. E. E. Journal* of August, 1929, pp. 624–628.

requirements the network must not cross-talk into associated circuits and must have desirable impedance characteristics in the attenuation range of frequencies as well as throughout the transmission range. An illustration of the use of transmission networks in a typical three-channel carrier telephone system is given describing the functions of the line filter sets, the directional filter sets, band filters, and equalizers. Some of the engineering limitations on the design and construction of networks are discussed.

Recent Developments in Telephone Construction Practices.[12] B. S. WAGNER and A. C. BURROWAY. In this paper are described some recent developments in telephone cable installation and maintenance practices. The paper is divided into three sections: (1) Gas pressure testing for detecting and locating sheath defects before they result in failure. (2) Methods for reducing bowing and other movements of cable which in time cause fracture of the sheath. (3) Catenary construction for long spans, such as at river crossings.

[12] *A. I. E. E. Journal*, Vol. 48, May, 1929, pp. 366–369 (abridgment).

Contributors to this Issue

KARL K. DARROW, B.S., 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, 1917–25; Bell Telephone Laboratories, 1925–. Dr. Darrow has been engaged largely in writing studies and analyses of various fields of physics and the allied sciences. Some of his earlier articles on Contemporary Physics form the nucleus of a recently published book entitled "Introduction to Contemporary Physics" (D. Van Nostrand Company).

H. F. DODGE, S.B., Massachusetts Institute of Technology, 1916; Instructor, Electrical Engineering, 1916–17; A.M., Columbia University, 1922; Engineering Department, Western Electric Company, 1917–25; Bell Telephone Laboratories, 1925–. Mr. Dodge was earlier associated with the development of telephone instruments and allied devices, and is now engaged in development work relating to the application of statistical methods to inspection engineering.

HARVEY FLETCHER, B.Sc., Brigham Young, 1907; Ph.D., Chicago, 1911; Instructor of Physics, Brigham Young, 1907–08 and Chicago, 1909–10; Professor, Brigham Young, 1911–16; Engineering Department, Western Electric Company, 1916–25; Bell Telephone Laboratories, 1925–. During recent years, Dr. Fletcher has conducted extensive investigations in the fields of speech and audition.

EDWARD C. MOLINA, Engineering Department of the American Telephone and Telegraph Company, 1901–19; Department of Development and Research, 1919–. Mr. Molina has been closely associated with the application of the mathematical theory of probabilities to trunking problems and has taken out several important patents relating to machine switching.

H. G. ROMIG, A.B., Pacific University, 1921; University of Washington, 1922; A.M., University of California, 1923; Teaching Fellow in Physics, University of California, 1922–24; Instructor in Mathematics and Physics, San Jose State Teachers College, 1924–26; Bell Telephone Laboratories, 1926–. Mr. Romig is engaged in development work relating to the application of statistical methods to inspection engineering.

L. J. SIVIAN, A.B., Cornell University, 1916; Engineering Department, Western Electric Company, 1917–1919 and 1920–1925; Bell Telephone Laboratories, 1925–. Mr. Sivian's work is in acoustics, chiefly in connection with methods of electroacoustic measurements.

J. C. Steinberg, B.Sc., M.Sc., Coe College, 1916, 1917; U. S. Air Service, 1917–19; Ph.D., Iowa, 1922; Engineering Dept., Western Electric Company, 1922–25; Bell Telephone Laboratories, 1925–. Dr. Steinberg's work since coming with the Bell System has related largely to speech and hearing.

R. I. Wilkinson, B.Sc., Iowa State College, 1924; Western Electric Company, 1920–21; American Telephone and Telegraph Company, Department of Development and Research, 1924–. Mr. Wilkinson has studied principally the application to telephone problems of the mathematical theory of probability, including sampling and statistical analysis.

W. Howard Wise, B.S., Montana State College, 1921; M.A., University of Oregon, 1923; Ph.D., California Institute of Technology, 1926; American Telephone and Telegraph Company, Department of Development and Research, 1926–. Since coming to the American Telephone and Telegraph Company, Dr. Wise has been engaged in various theoretical investigations.

Index to Volume VIII

A

Acoustic Considerations Involved in Steady State Loud Speaker Measurements, *L. G. Bostwick*, page 135.

Adsorption, Application of Electron Diffraction to the Study of Gas, *L. H. Germer*, page 591.

Alloys of Iron, Nickel and Cobalt, Magnetic, *G. W. Elmen*, page 435.

Aluminum Electrolytic Condenser, *H. O. Siegmund*, page 41.

Apparatus Springs, Telephone Review of the Principal Types and the Properties Desired of These Springs, *J. R. Townsend*, page 257.

Application of Electron Diffraction to the Study of Gas Adsorption, *L. H. Germer*, page 591.

Application to the Binomial Summation of a Laplacian Method for the Evaluation of Definite Integrals, *E. C. Molina*, page 99.

Articulation Testing Methods, *H. Fletcher* and *J. C. Steinberg*, page 806.

Asymptotic Dipole Radiation Formulas, *W. H. Wise*, page 662.

B

Bailey, Austin, S. W. Dean and *W. T. Wintringham*, Receiving System for Long-Wave Transatlantic Radio Telephony, page 309.

Binomial Summation of a Laplacian Method for the Evaluation of Definite Integrals, Application to the, *E. C. Molina*, page 99.

Bostwick, L. G., Acoustic Considerations Involved in Steady State Loud Speaker Measurements, page 135.

Braun Tube Hysteresigraph, *J. B. Johnson*, page 286.

C

Carson, John R., Ground Return Impedance: Underground Wire with Earth Return, page 94.

Cobalt, Magnetic Alloys of Iron, Nickel, and, *G. W. Elmen*, page 435.

Condenser, Aluminum Electrolytic, *H. O. Siegmund*, page 41.

Contemporary Advances in Physics, XVII. Scattering of Light with Change of Frequency, *Karl K. Darrow*, page 64.

Contemporary Advances in Physics, XVIII. Diffraction of Waves by Crystals, *Karl K. Darrow*, page 391.

Crystals, Diffraction of Waves by, Contemporary Advances in Physics, XVIII, *Karl K. Darrow*, page 391.

D

Darrow, Karl K., Contemporary Advances in Physics, XVII. Scattering of Light with Change of Frequency, page 64.

Darrow, Karl K., Contemporary Advances in Physics, XVIII. Diffraction of Waves by Crystals, page 391.

Darrow, Karl K., Statistical Theories of Matter, Radiation and Electricity, page 672.

Davisson, C. J., Electrons and Quanta, page 217.

Davisson, C. J. and *L. H. Germer*, Test for Polarization of Electron Waves by Reflection, page 466.

Dean, S. W., Austin Bailey and *W. T. Wintringham*, Receiving System for Long-Wave Transatlantic Radio Telephony, page 309.

Decibel: Name for the Transmission Unit, *W. H. Martin*, page 1.

Definite Integrals, Application to the Binomial Summation of a Laplacian Method for the Evaluation of, *E. C. Molina*, page 99.

Diffraction of Waves by Crystals, Contemporary Advances in Physics, XVIII, *Karl K. Darrow*, page 391.

Diffraction to the Study of Gas Adsorption, Application of Electron, *L. H. Germer*, page 591.

Dipole Radiation Formulas, Asympototic, *W. H. Wise*, page 662.

Distortion on Morse Telegraph Transmission Quality, Effect on Signal, *J. Herman*, page 267.

Distribution of the Unknown Mean of a Sampled Universe, Frequency, *E. C. Molina* and *R. I. Wilkinson*, page 632.

Dodge, H. F. and *H. G. Romig*, A Method of Sampling Inspection, page 613.

E

Effect of Signal Distortion on Morse Telegraph Transmission Quality, *J. Herman*, page 267.

Electric Circuits Applied to Communication, Principles of, *H. S. Osborne*, page 3.

Electrical Networks, New Method for Obtaining Transient Solutions of, *W. P. Mason*, page 109.

Electricity, Statistical Theories of Matter, Radiation and, *K. K. Darrow*, page 672.

Electrolytic Condenser, Aluminum, *H. O. Siegmund*, page 41.

Electrolytic Material upon Textiles as Insulators, Predominating Influence of Moisture and, *R. R. Williams* and *E. J. Murphy*, page 225.

Electron Diffraction to the Study of Gas Adsorption, Application of, *L. H. Germer*, page 591.

Electrons and Quanta, *C. J. Davisson*, page 217.

Electron Waves by Reflection, Test for Polarization of, *C. J. Davisson* and *L. H. Germer*, page 466.

Elmen, G. W., Magnetic Alloys of Iron, Nickel and Cobalt, page 435.

Elmen, G. W., Magnetic Properties of Perminvar, page 21.

Evaluation of Definite Integrals, Application to the Binomial Summation of a Laplacian Method for the, *E. C. Molina*, page 99.

Expansion Theorem, Generalization of Heaviside's, *W. O. Pennell, page 482.*

F

Fatigue Studies of Non-Ferrous Sheet Metals, *John R. Townsend* and *Charles H. Greenall*, page 576.

Ferguson, J. G., Shielding in High-Frequency Measurements, page 560.

Fletcher, H., and *J. C. Steinberg*, Articulation Testing Methods, page 806.

Frederick, Halsey A., Recent Advances in Wax Recording, page 159.

Frequency Distribution of the Unknown Mean of a Sampled Universe, *E. C. Molina* and *R. I. Wilkinson*, page 632.

Frequency, High Precision Standard of, *W. A. Marrison*, page 493.

G

Gas Adsorption, Application of Electron Diffraction to the Study of, *L. H. Germer*, page 591.

Generalization of Heaviside's Expansion Theorem, *W. O. Pennell*, page 482.

Germer, L. H., Application of Electron Diffraction to the Study of Gas Adsorption, page 591.

Germer, L. H. and *C. J. Davisson*, Test for Polarization of Electron Waves by Reflection, page 466.

Glenn, H. H. and *E. B. Wood*, Purified Textile Insulation for Telephone Central Office Wiring, page 243.

Gray, C. H. G. and *W. H. Martin*, Master Reference System for Telephone Transmission, page 536

Greenall, Charles H. and *John R. Townsend*, Fatigue Studies of Non-Ferrous Sheet Metals, page 576.

Ground Return Impedance: Underground Wire with Earth Return, *John R. Carson*, page 94.

H

Heaviside's Expansion Theorem, Generalization of, *W. O. Pennell*, page 482.

Herman, J., Effect of Signal Distortion on Morse Telegraph Transmission Quality, page 267.

High-Frequency Measurements, Shielding in, *J. G. Ferguson*, page 560.

High Precision Standard of Frequency, *W. A. Marrison*, page 493.

Hysteresigraph, Braun Tube, *J. B. Johnson*, page 286.

I

Impedance: Ground Return, Underground Wire with Earth Return, *John R. Carson*, page 94.

Inspection, A Method of Sampling, *H. F. Dodge* and *H. G. Romig*, page 613.

Insulation for Telephone Central Office Wiring, Purified Textile, *H. H. Glenn* and *E. B. Wood*, page 243.

Insulators, Predominating Influence of Moisture and Electrolytic Material upon Textiles as, *R. R. Williams* and *E. J. Murphy*, page 225.

Iron, Nickel, and Cobalt, Magnetic Alloys of, *G. W. Elmen*, page 435.

J

Johnson, J. B., Braun Tube Hysteresigraph, page 286.

L

Lack, F. R., Observations on Modes of Vibration and Temperature Coefficients of Quartz Crystal Plates, page 515.

Light Valve, Sound Recording with the, *Donald MacKenzie*, page 173.

Light with Change of Frequency, Scattering of, Contemporary Advances in Physics, XVII, *Karl K. Darrow*, page 64.

Long-Wave Transatlantic Radio Telephony, Receiving System for, *Austin Bailey, S. W. Dean* and *W. T. Wintringham*, page 309.

Loud Speaker Measurements, Acoustic Considerations Involved in Steady State, *L. G. Bostwick*, page 135.

M

MacKenzie, Donald, Sound Recording with the Light Valve, page 173.

Magnetic Alloys of Iron, Nickel, and Cobalt, *G. W. Elmen*, page 435.

Magnetic Properties of Perminvar, *G. W. Elmen*, page 21.

Marrison, W. A., High Precision Standard of Frequency, page 493.

Marrison W. A., Oscillographs for Recording Transient Phenomena, page 368.

Martin, W. H. and *C. H. G. Gray*, Master Reference System for Telephone Transmission, page 536.

Martin, W. H., Decibel: Name for the Transmission Unit, page 1.

Mason, W. P., New Method for Obtaining Transient Solutions of Electrical Networks, page 109.

Master Reference System for Telephone Transmission, *W. H. Martin* and *C. H. G. Gray*, page 536.

Matter, Radiation and Electricity, Statistical Theories of, *K. K. Darrow*, page 672.

Measurements, Acoustic Considerations Involved in Steady State Loud Speaker, *L. G. Bostwick*, page 135.

Measurements, Shielding in High-Frequency, *J. G. Ferguson*, page 560.

Measurement, Speech Power and Its, *L. J. Sivian*, page 646.

Metals, Fatigue Studies of Non-Ferrous Sheet, *John R. Townsend* and *Charles H. Greenall*, page 576.

Metals, Physical Properties and Methods of Test for Some Sheet Non-Ferrous, *J. R. Townsend* and *W. A. Straw*, page 749.

Modes of Vibration and Temperature Coefficients of Quartz Crystal Plates, Observations on, *F. R. Lack*, page 515.

Moisture and Electrolytic Material upon Textiles as Insulators, Predominating Influence of, *R. R. Williams* and *E. J. Murphy*, page 225.

Molina, E. C., Application to the Binomial Summation of a Laplacian Method for the Evaluation of Definite Integrals, page 99.

Molina, E. C. and *R. I. Wilkinson*, Frequency Distribution of the Unknown Mean of a Sampled Universe, page 632.

Motion Picture Theaters, Sound Projector System for Use in, *E. O. Scriven*, page 197.

Murphy, E. J. and *R. R. Williams*, Predominating Influence of Moisture and Electrolytic Material upon Textiles as Insulators, page 225.

N

Networks, New Method for Obtaining Transient Solutions of Electrical, *W. P. Mason*, page 109.

New Method for Obtaining Transient Solutions of Electrical Networks, *W. P. Mason*, page 109.

Nickel, and Cobalt, Magnetic Alloys of Iron, *G. W. Elmen*, page 435.

Non-Ferrous Metals, Physical Properties and Methods of Test for Some Sheet, *J. R. Townsend*, and *W. A. Straw*, page 749.

Non-Ferrous Sheet Metals, Fatigue Studies of, *John R. Townsend* and *Charles H. Greenall*, page 576.

O

Observations on Modes of Vibration and Temperature Coefficients of Quartz Crystal Plates, *F. R. Lack*, page 515.

Osborne, H. S., Principles of Electric Circuits Applied to Communication, page 3.

Oscillographs for Recording Transient Phenomena, *W. A. Marrison*, page 368.

P

Pennell, W. O., Generalization of Heaviside's Expansion Theorem, page 482.

Perminvar, Magnetic Properties of, *G. W. Elmen*, page 21.

Physical Properties and Methods of Test for Some Sheet Non-Ferrous Metals, *J. R. Townsend* and *W. A. Straw*, page 749.

Physics, XVII. Contemporary Advances in, Scattering of Light with Change of Frequency, *Karl K. Darrow*, page 64.

Physics, XVIII, Contemporary Advances in, Diffraction of Waves by Crystals, *Karl K. Darrow*, page 391.

Pictures, Synchronization and Speed Control of Synchronized Sound, *H. M. Stoller*, page 185.

Polarization of Electron Waves by Reflection, Test for, *C. J. Davisson*, and *L. H. Germer*, page 466.

Power and Its Measurement, Speech, *L. J. Sivian*, page 646.

Predominating Influence of Moisture and Electrolytic Material upon Textiles as Insulators, *R. R. Williams* and *E. J. Murphy*, page 225.

Principles of Electric Circuits Applied to Communication, *H. S. Osborne*, page 3.

Properties and Methods of Test for Some Sheet Non-Ferrous Metals, Physical, *J. R. Townsend* and *W. A. Straw*, page 749.

Purified Textile Insulation for Telephone Central Office Wiring, *H. H. Glenn* and *E. B. Wood*, page 243.

Q

Quanta, Electrons and, *C. J. Davisson*, page 217.

Quartz Crystal Plates, Observations on Modes of Vibration and Temperature Coefficient of, *F. R. Lack*, page 515.

R

Radiation and Electricity, Statistical Theories of Matter, *K. K. Darrow*, page 672.

Radiation Formulas, Asymptotic Dipole, *W. H. Wise*, page 662.

Radio Telephony, Receiving System for Long-Wave Transatlantic, *Austin Bailey, S. W. Dean* and *W. T. Wintringham*, page 309.

Receiving System for Long-Wave Transatlantic Radio Telephony, *Austin Bailey, S. W. Dean* and *W. T. Wintringham*, page 309.

Recent Advances in Wax Recording, *Halsey A. Frederick*, page 159.

Recording, Recent Advances in Wax, *Halsey A. Frederick*, page 159.

Recording Transient Phenomena, Oscillographs for, *W. A. Marrison*, page 368.

Recording with the Light Valve, Sound, *Donald MacKenzie*, page 173.

Reference System for Telephone Transmission, Master, *W. H. Martin* and *C. H. G. Gray*, page 536.

Romig, H. G. and *H. F. Dodge*, Method of Sampling Inspection, page 613.

S

Sampled Universe, Frequency Distribution of the Unknown Mean of a, *E. C. Molina* and *R. I. Wilkinson*, page 632.

Sampling Inspection, Method of, *H. F. Dodge* and *H. G. Romig*, page 613.

Scattering of Light with Change of Frequency. Contemporary Advances in Physics, XVII, *Karl K. Darrow*, page 64.

Scriven, E. O., A Sound Projector System for Use in Motion Picture Theaters, page 197.

Sheet Metals, Fatigue Studies of Non-Ferrous, *John R. Townsend* and *Charles H. Greenall*, page 576.

Sheet Non-Ferrous Metals, Physical Properties and Methods of Test for Some, *J. R. Townsend* and *W. A. Straw*, page 749.

Shielding in High-Frequency Measurements, *J. G. Ferguson*, page 560.

Siegmund, H. O., Aluminum Electrolytic Condenser, page 21.

Sivian, L. J., Speech Power and Its Measurement, page 646.

Sound Pictures, Synchronization and Speed Control of Synchronized, *H. M. Stoller*, page 185.

Sound Projector System for Use in Motion Picture Theaters, *E. O. Scriven*, page 197.

Sound Recording with the Light Valve, *Donald MacKenzie*, page 173.

Speech Power and Its Measurement. *L. J. Sivian*, page 646.

Speech Control of Synchronized Sound Pictures, Synchronization and, *H. M. Stoller*, page 185.

Springs, Telephone Apparatus Review of the Principal Types and the Properties Desired of These Springs, *J. R. Townsend*, page 257.

Standard of Frequency, High Precision, *W. A. Marrison*, page 493.

Statistical Theories of Matter, Radiation and Electricity, *K. K. Darrow*, page 672.

Steinberg, J. C. and *H. Fletcher*, Articulation Testing Methods, page 806.

Stoller, H. M., Synchronization and Speed Control of Synchronized Sound Pictures, page 185.

Straw, W. A. and *J. R. Townsend*, Physical Properties and Methods of Test for Some Sheet Non-Ferrous Metals, page 749.

Synchronization and Speed Control of Synchronized Sound Pictures, *H. M. Stoller*, page 185.

Synchronized Sound Pictures, Synchronization and Speed Control of, *H. M. Stoller*, page 185.

T

Telegraph Transmission Quality, Effect of Signal Distortion on Morse, *J. Herman*, page 267.

Telephone Apparatus Springs. Review of the Principal Types and the Properties Desired of These Springs, *J. R. Townsend*, page 257.

Telephone Transmission, Master Reference System for, *W. H. Martin* and *C. H. G. Gray*, page 536.

Temperature Coefficients of Quartz Crystal Plates, Observations on Modes of Vibration and, *F. R. Lack*, page 515.

Test for Polarization of Electron Waves by Reflection, *C. J. Davisson*, and *L. H. Germer*, page 466.

Testing Methods, Articulation, *H. Fletcher* and *J. C. Steinberg*, page 806.

Textile Insulation for Telephone Central Office Wiring, Purified, *H. H. Glenn* and *E. B. Wood*, page 243.

Textiles as Insulators, Predominating Influence of Moisture and Electrolytic Material upon, *R. R. Williams* and *E. J. Murphy*, page 225.

Townsend, John R. and *Charles H. Greenall*, Fatigue Studies of Non-Ferrous Sheet Metals, page 576.

Townsend, J. R., Telephone Apparatus Springs. Review of the Principal Types and the Properties Desired of These Springs, page 257.

Townsend, J. R. and *W. A. Straw*, Physical Properties and Methods of Test for Some Sheet Non-Ferrous Metals, page 749.

Transatlantic Radio Telephony, Receiving System for Long-Wave, *Austin Bailey, S. W. Dean* and *W. T. Wintringham*, page 309.

Transient Phenomena, Oscillographs for Recording, *W. A. Marrison*, page 368.

Transient Solutions of Electrical Networks, New Method for Obtaining, *W. P. Mason*, page 109.

Transmission, Master Reference System for Telephone, *W. H. Martin* and *C. H. G. Gray*, page 536.

Transmission Unit, Name for the Decibel, *W. H. Martin*, page 1.

U

Underground Wire with Earth Return, Ground Return Impedance, *John R. Carson*, page 94.

V

Valve, Sound Recording with the Light, *Donald MacKenzie*, page 173.

Vibration and Temperature Coefficients of Quartz Crystal Plates, Observations on Modes of, *F. R. Lack*, page 515.

W

Wax Recording, Recent Advances in, *Halsey A. Frederick*, page 159.

Wilkinson, R. I. and *E. C. Molina*, Frequency Distribution of the Unknown Mean of a Sampled Universe, page 632.

Williams, R. R. and *Murphy, E J.*, Predominating Influence of Moisture and Electrolytic Material upon Textiles as Insulators, page 225.

Wintringham, W. T., Austin Bailey and *S. W. Dean*, Receiving System for Long-Wave Transatlantic Radio Telephony, page 309.

Wise, W. H., Asymptotic Dipole Radiation Formulas, page 662.

Wood, E. B. and *H. H. Glenn*, Purified Textile Insulation for Telephone Central Office Wiring, page 243.