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 wavelength 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 wavelength 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 I*B*

RESISTANCE OF COTTON AND SILK THREADS

Length ½ in. Temperature 25 deg. cent.

	Resistance Megohms	
	Cotton	Silk (Spun)
Sample	Humidity 77%	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 I*C*

RESISTANCE BETWEEN TWISTED PAIRS OF COTTON AND SILK INSULATED WIRES

Humidity 77 per cent Temperature 25 deg. cent.

	Resistance Megohms	
Sample	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

	Resistance Megohms	
Length Inches	Total	Per Inch
0.5	21,700	43,400
1	41,750	41,750
3	153,000	51,000

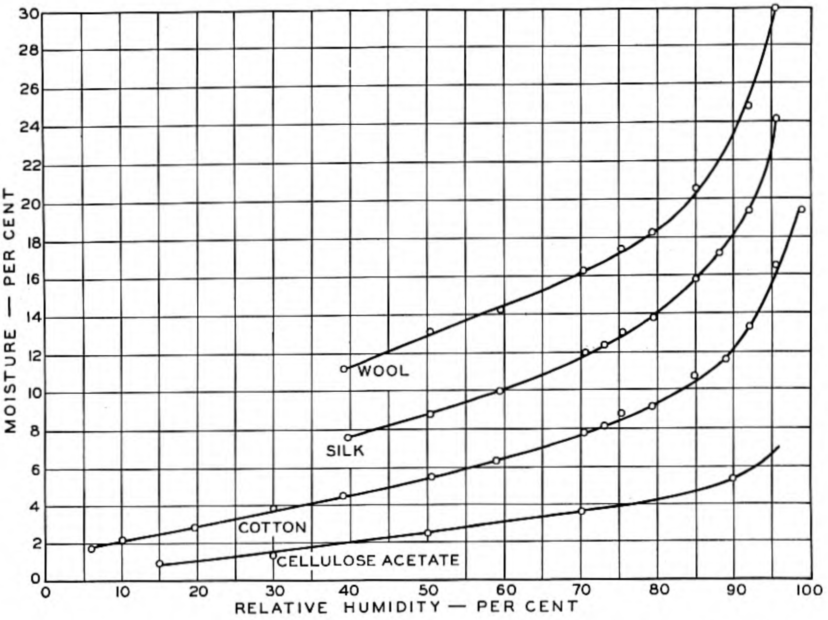


Fig. 1—Dependence of moisture content of textiles upon relative humidity of atmosphere with which they are equilibrated

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.

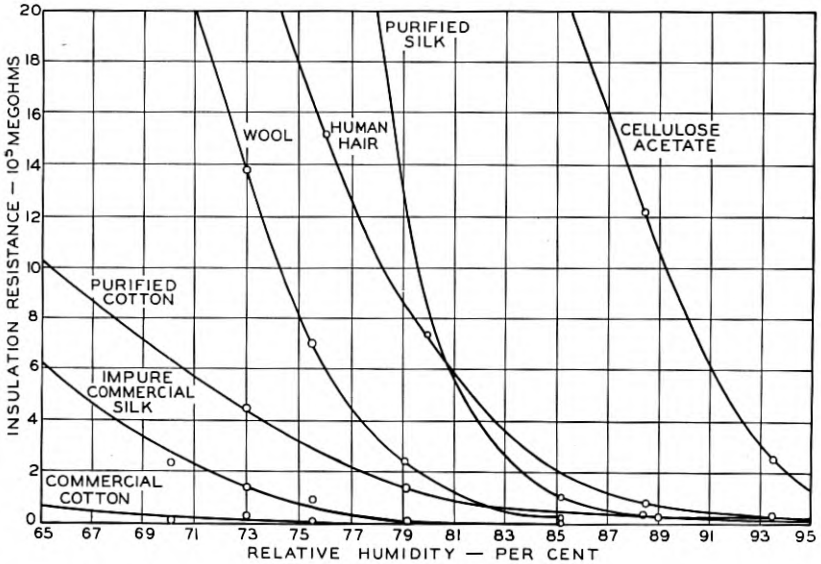


Fig. 2—Insulation resistance of ½ inch lengths of textile threads as affected by relative humidity of atmosphere. The purified cotton and purified silk had been submitted to a washing procedure to remove electrolytes. The impure silk was a commercial specimen representing somewhat more than usual contamination 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

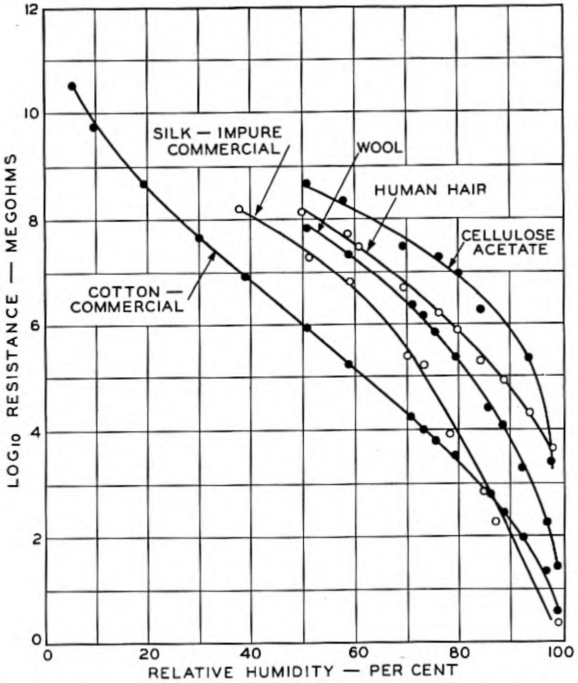


Fig. 3—Log insulation resistance of ½ inch lengths of textile threads *vs*. relative humidity of the atmosphere. For description of samples see Fig. 2. For tabulated data see reference 3*a*

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.[3 c] 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 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.

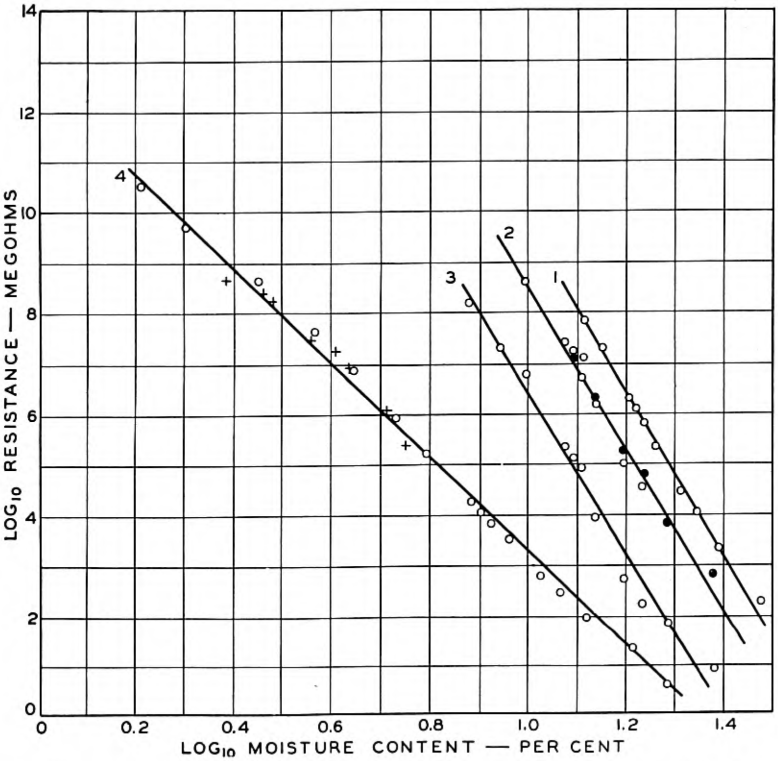


Fig. 4—Insulation resistance as a function of moisture content of textiles
1. Wool yarn
2. Silk threads purified [sample 1 (o); sample 2 (●)]
3. Silk threads impure
4. Cotton threads (o); cellulose acetate threads (+)

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

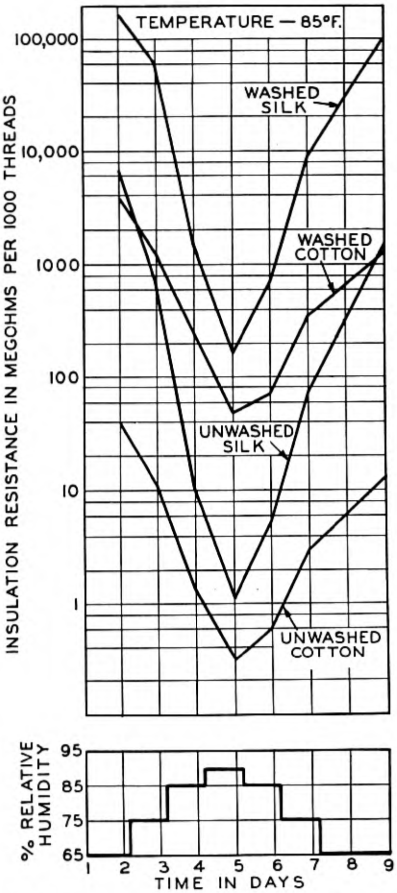


Fig. 1—Typical d.c. resistance characteristics of washed and unwashed silk and cotton threads of equal size.

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 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

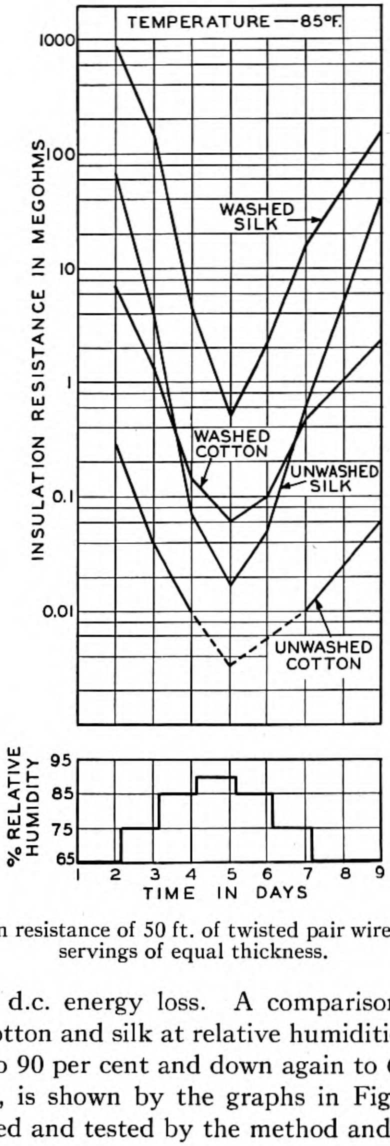


Fig. 2—D.C. insulation resistance of 50 ft. of twisted pair wire insulated with double servings of equal thickness.

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

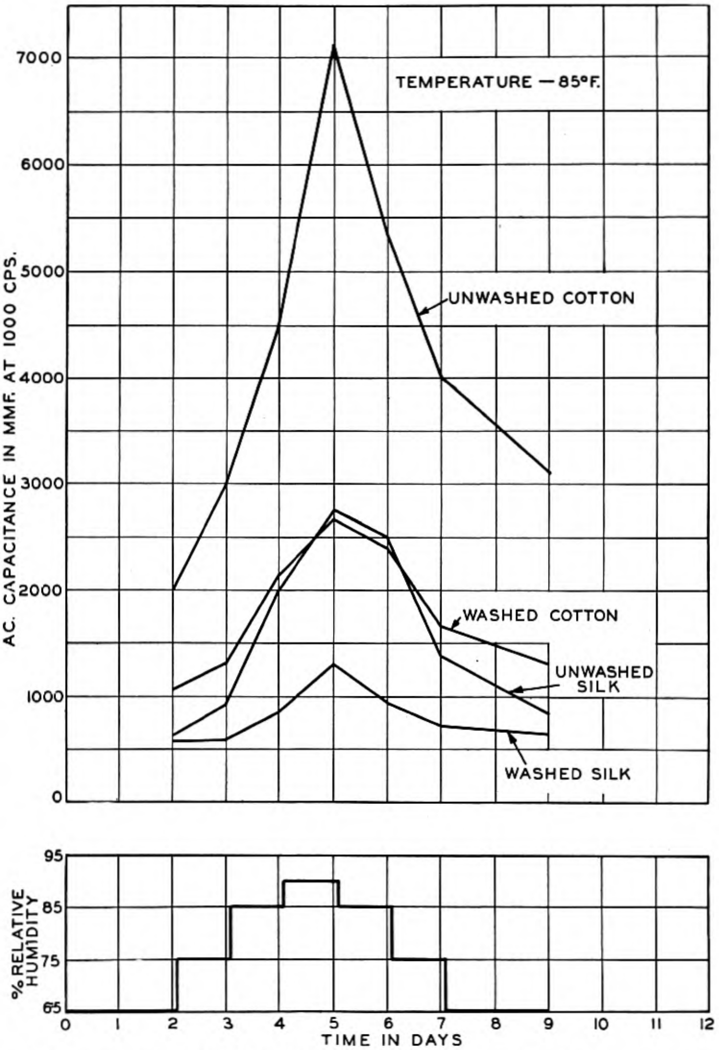


Fig. 3—A.C. capacitance of 50 ft. of twisted pair wire insulated with double serving of equal thickness.

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.

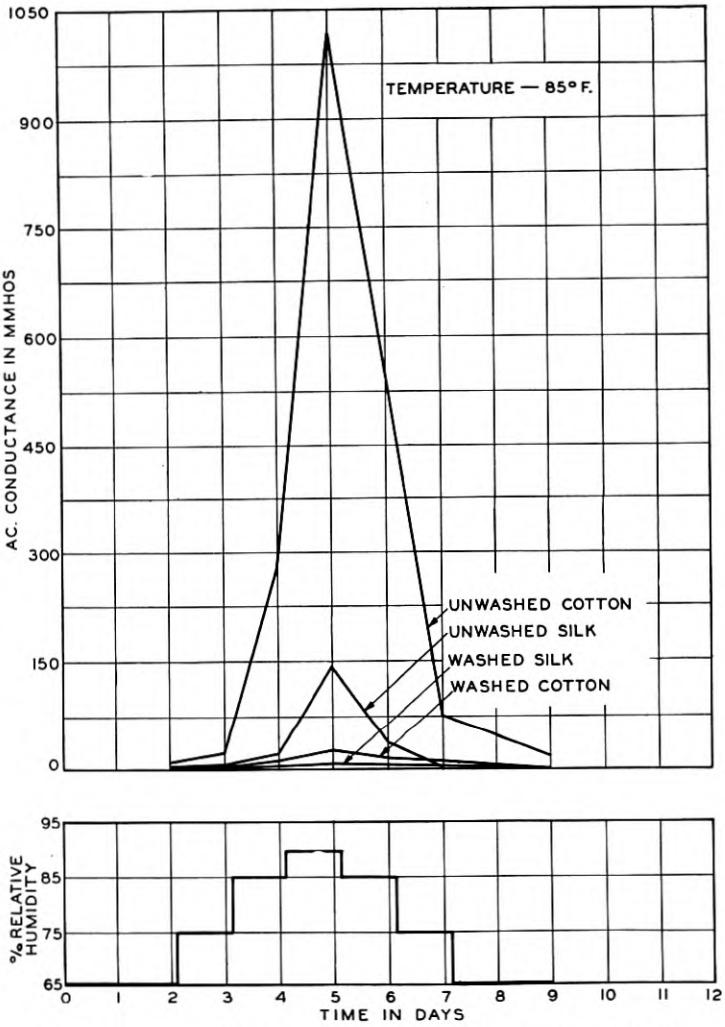


Fig. 4—A.C. conductance of 50 ft. of twisted pair wire insulated with double servings of equal thickness.

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 found that temperature variations within the tank are reduced to the vanishing point for all practical purposes.

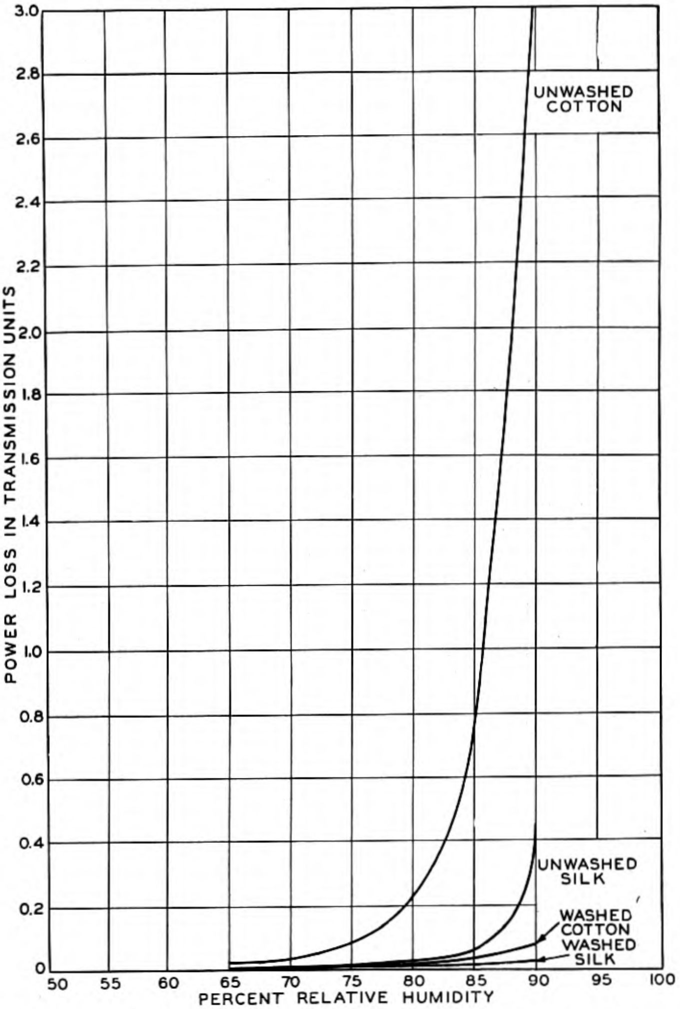


Fig. 5—Transmission loss in 50 ft. of twisted pair wire insulated with double servings of equal thickness.

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. 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.

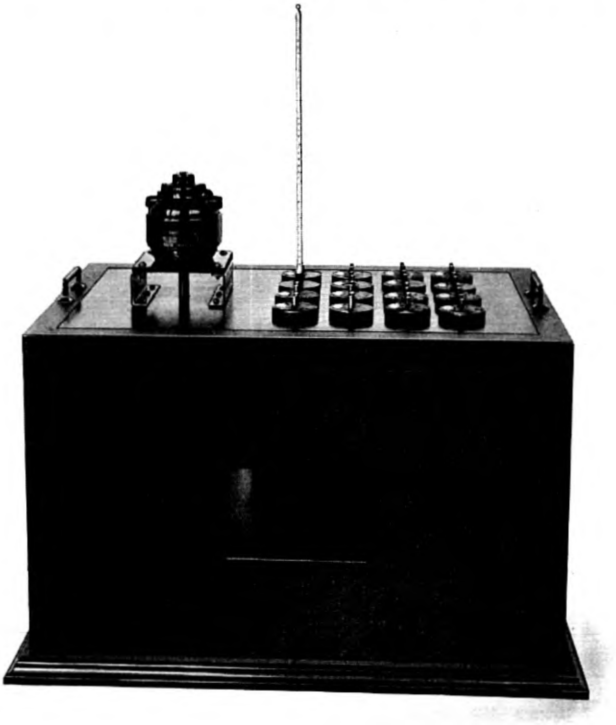


Fig. 6—Humidity cabinet for conditioning samples.

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.

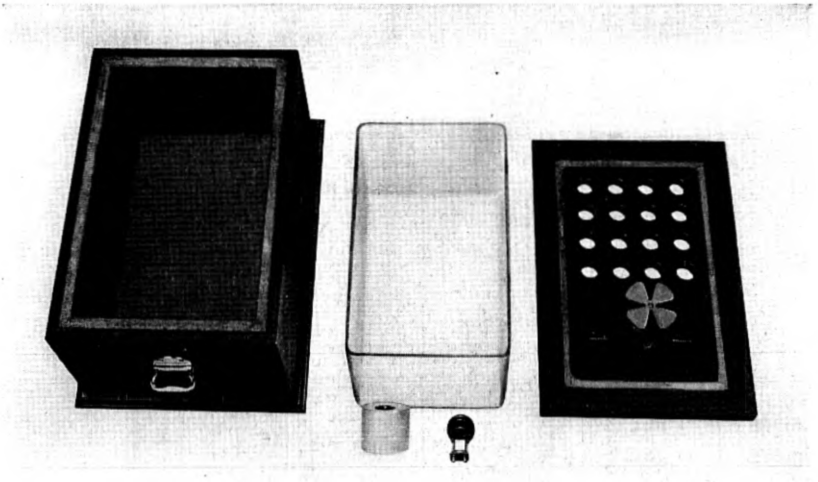


Fig. 7—Humidity cabinet disassembled.

The distance between the electrodes is not particularly critical. That is, it is not important, for example, whether the distance is ⅝ in. or ¾ in. It is important, however, that having decided upon a certain separation, say ¾ 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, ¾ 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.

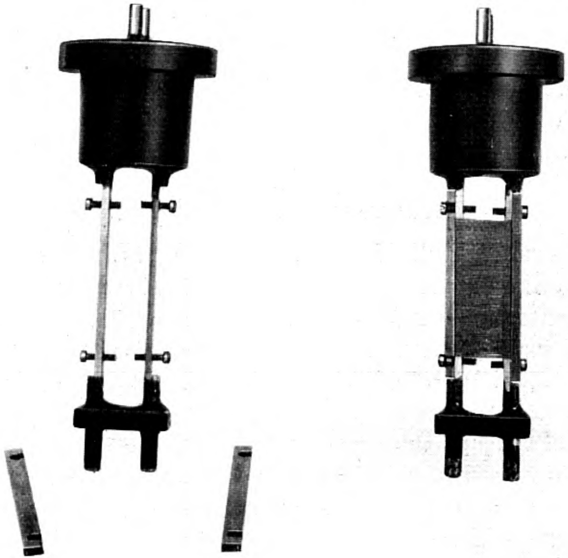


Fig. 8—Electrodes on which samples are wound for test.

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.

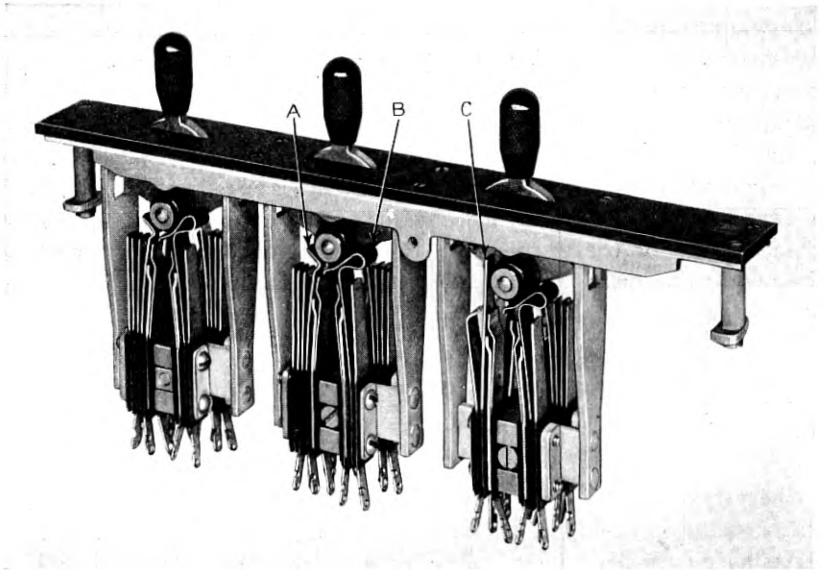


Fig 1—Common type of switchboard key illustrating use of sheet-metal springs: *A*, plunger spring with crimp; *B*, crook spring; *C*, straight plunger spring.

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 ×10⁶
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 ×10⁶
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 ×10⁶
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.

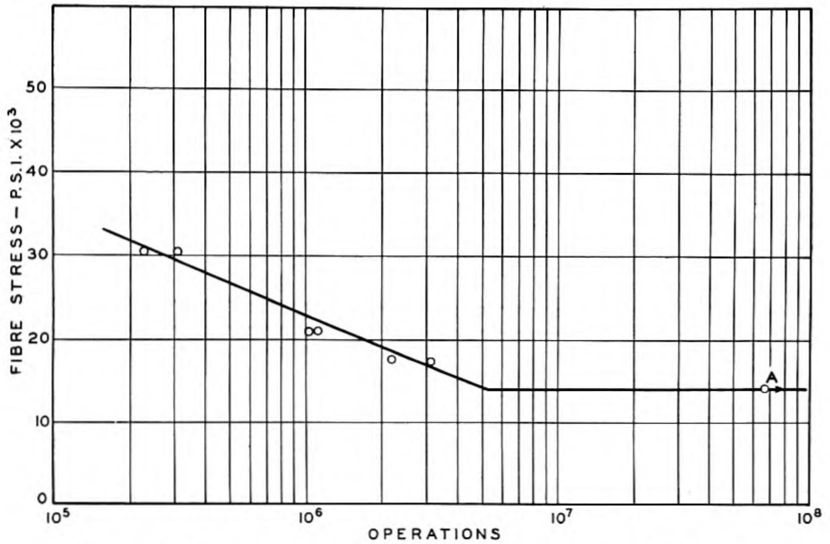


Fig. 2—Fiber stress vs. fatigue—brass sheet—24 gage—4 Nos. hard: *A*, average of three specimens taken from test without failure.

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-

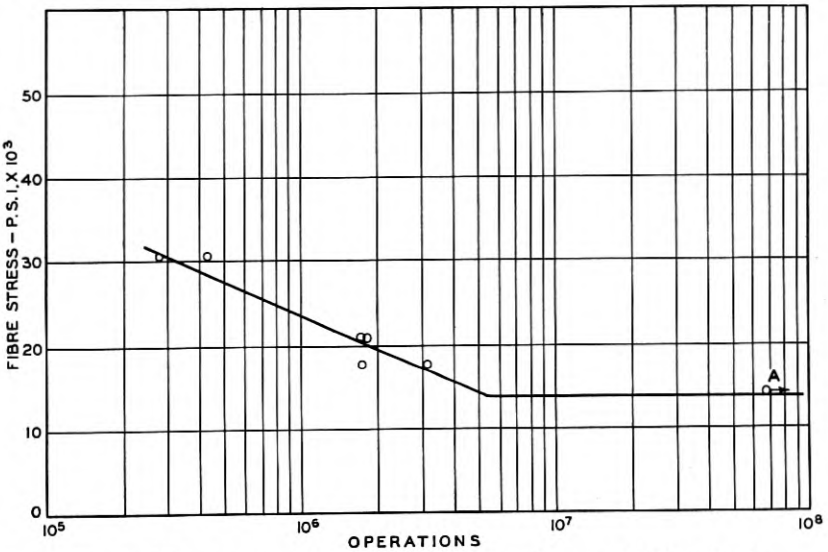


Fig. 3—Fiber stress vs. fatigue—brass sheet—24 gage—10 Nos. hard: *A*, average of three specimens taken from test without failure.

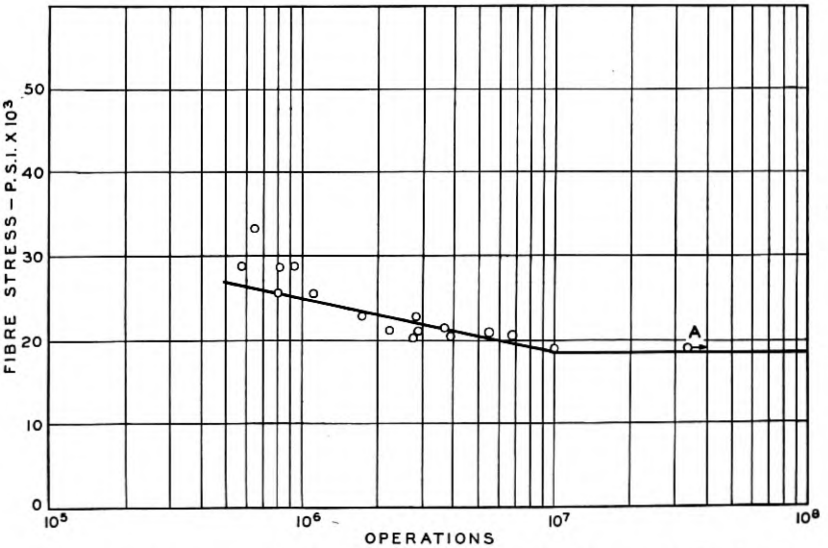


Fig. 4—Fiber stress vs. fatigue—nickel silver sheet—24 gage—4 Nos. hard: *A*, average of three specimens taken from test without failure.

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.

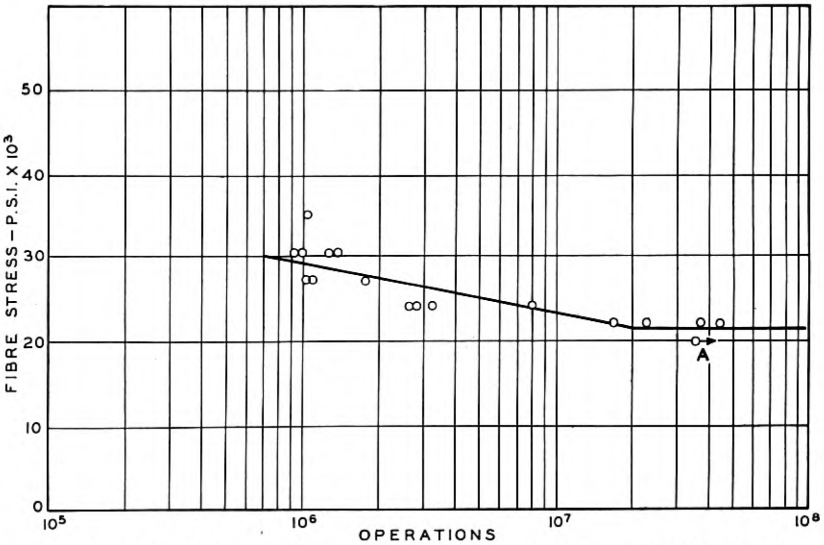


Fig. 5—Fiber stress vs. fatigue—nickel silver sheet—24 gage—10 Nos. hard: *A*, average of three specimens taken from test without failure.

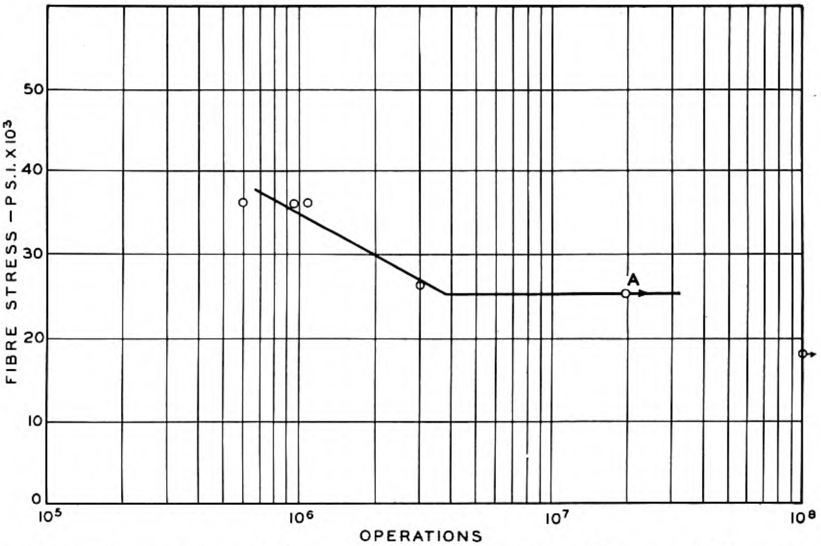


Fig. 6—Fiber stress vs. fatigue—phosphor bronze sheet—24 gage—10 Nos. hard: *A*, average of two specimens taken from test without failure.

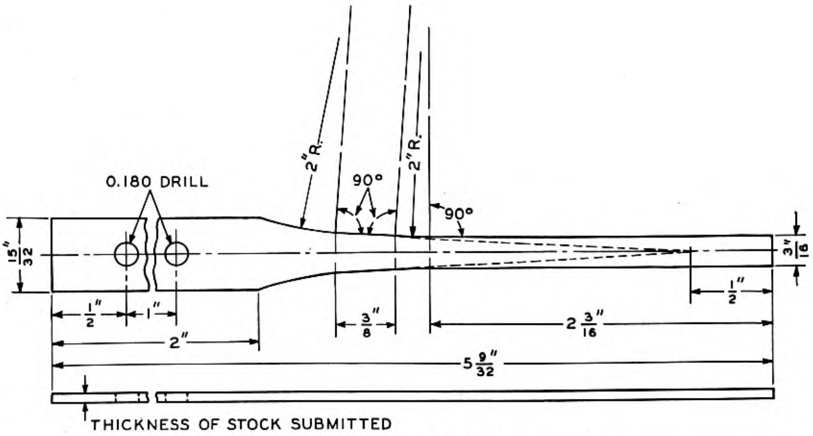


Fig. 7—Test specimen for fatigue study.
(The piece must be free from scriber and tool marks. The 2 in. radius fillets must join the tapered portion without an abrupt break. The intersection of the tapered sides produced must be ½ in. from the end.)

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.

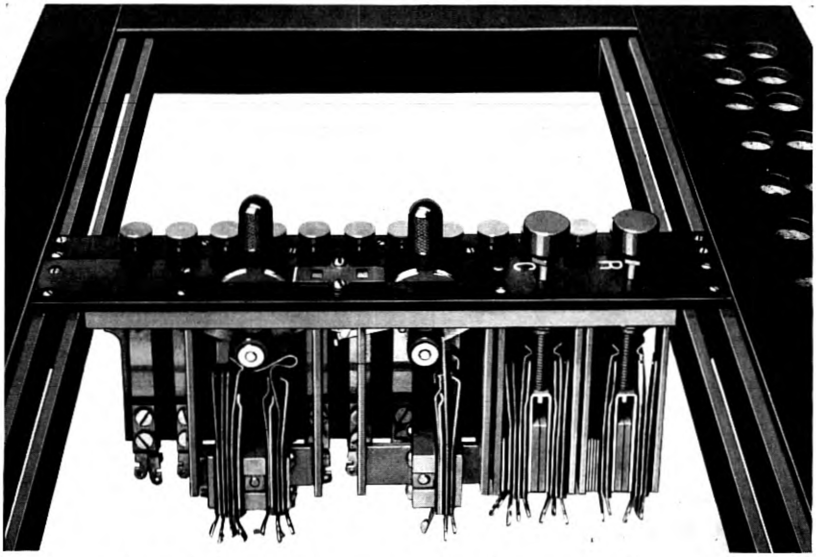


Fig. 8—Switchboard keys employing sheet metal springs and helical compression springs

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.

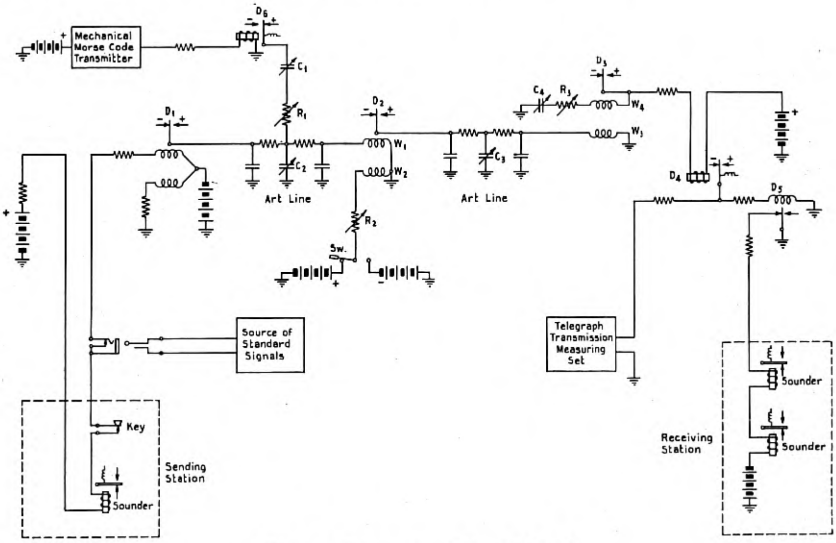


Fig. 1—Diagram of Test Circuit.

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 transmitting 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 unsatisfactory. 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.

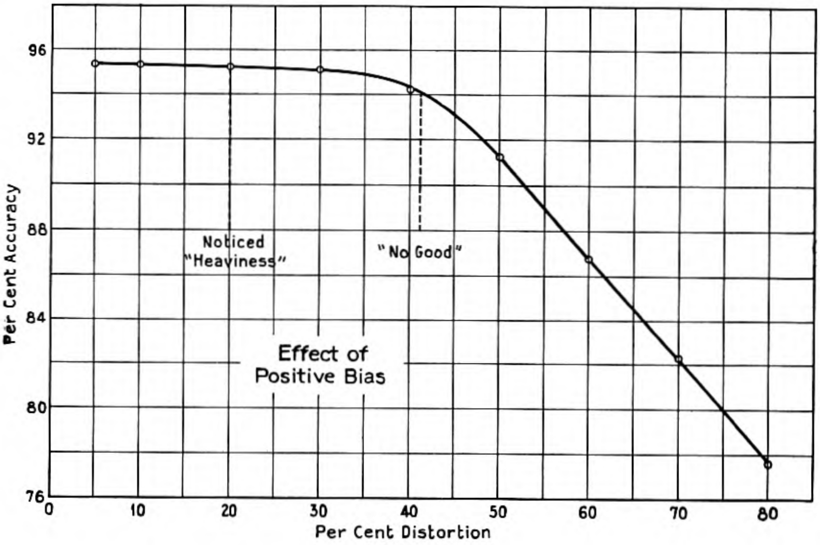


Fig. 2.

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.

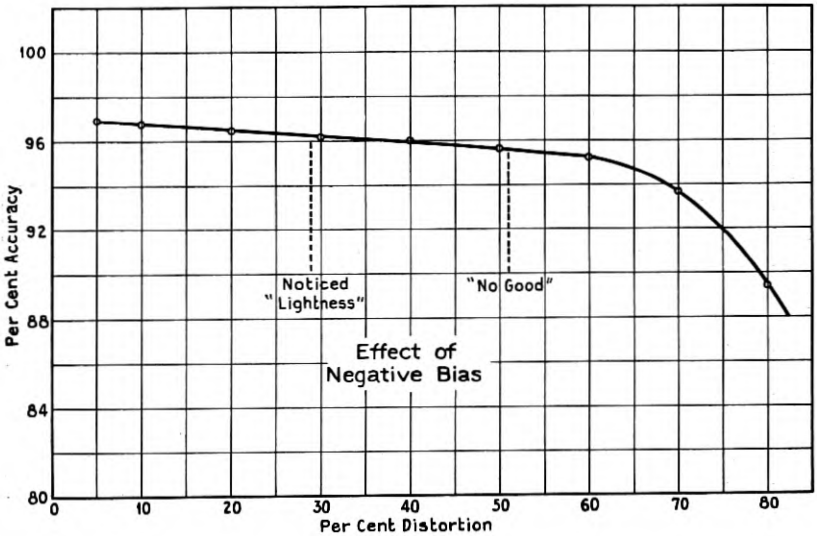


Fig. 3.

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.

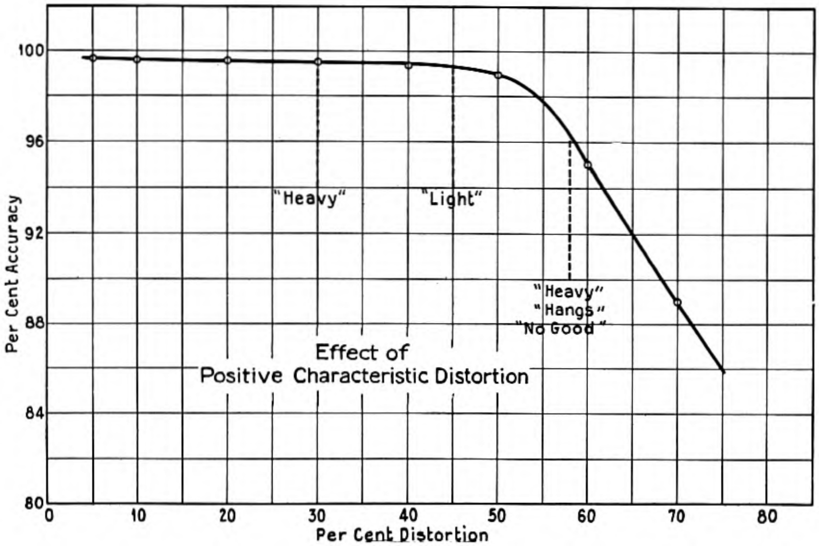


Fig. 4.

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.

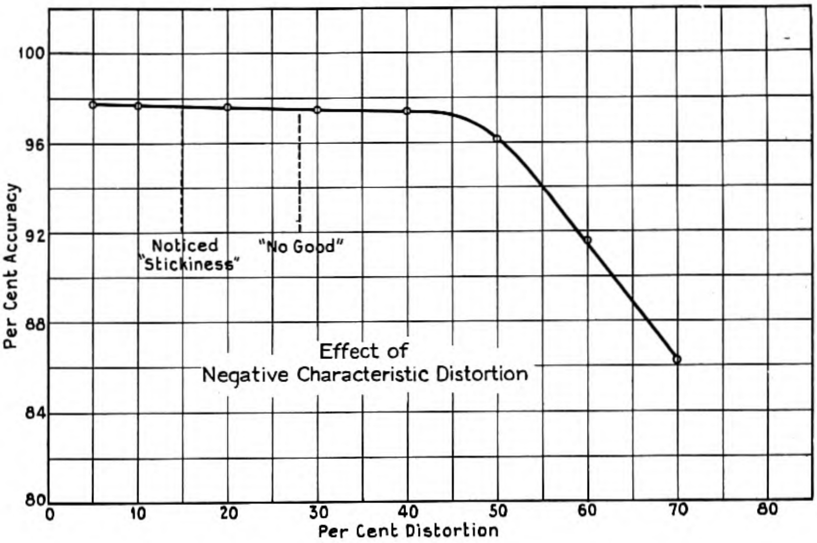


Fig. 5.

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.

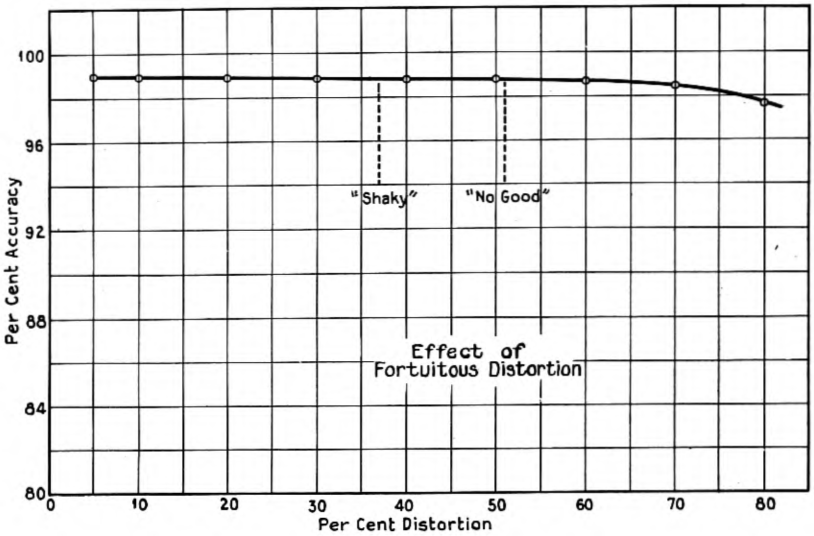


Fig. 6.

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.

	Limiting Values	
Kind of Distortion	From Opinion of Operators	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.

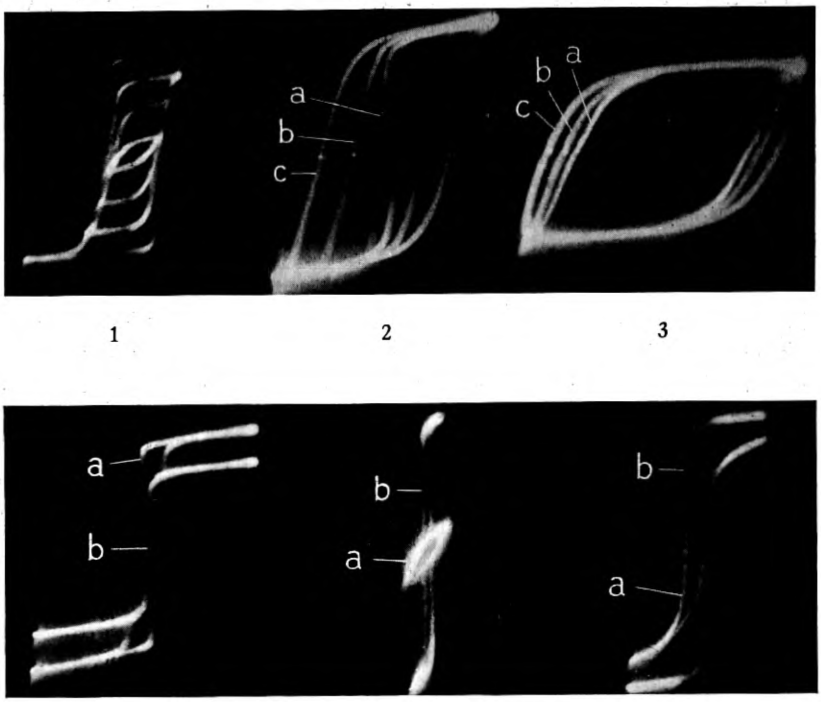
The hysteresis curves of Plate I (1) are those of a ribbon of Armco iron. The curves were made with 25 cycle magnetization and at four different magnetizing fields, the highest field being sufficient to saturate the sample. The effects of frequency and of permeability on the eddy current loss in materials 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.

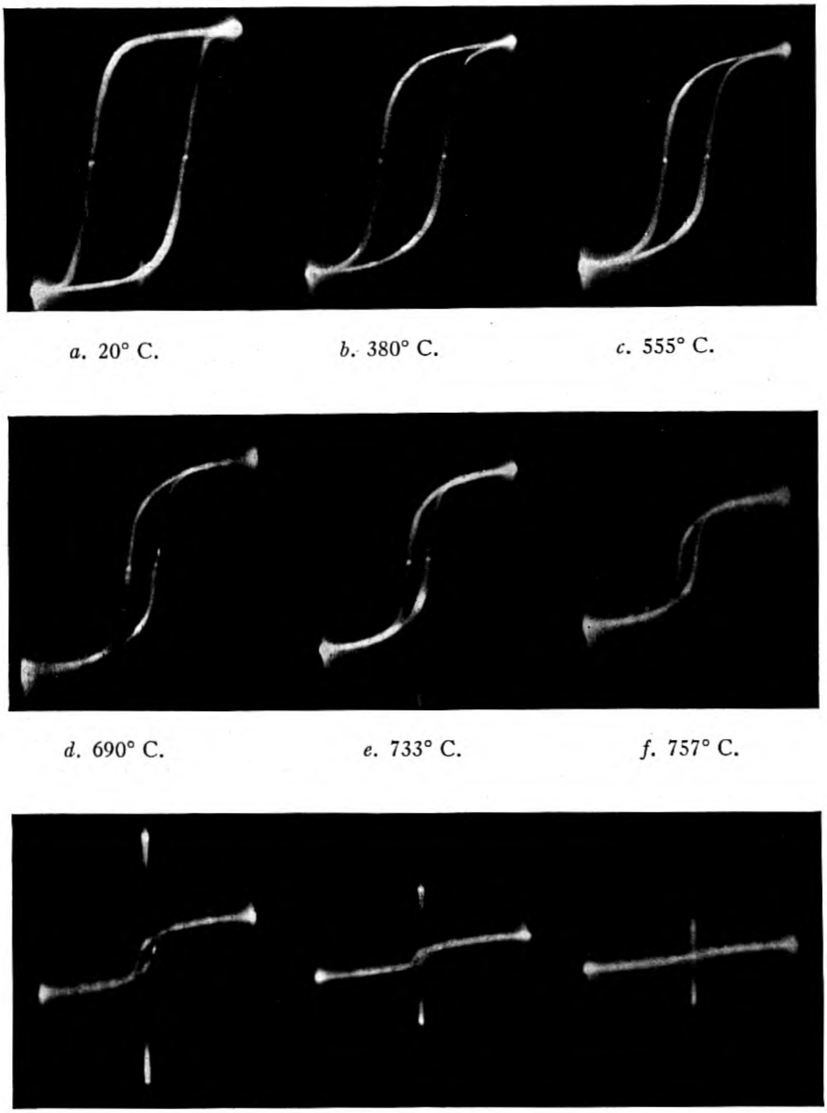


1 2 3

4 5 6

1. Armco iron at various fields.
2. Permalloy at (*a*) $25^\sim$, (*b*) $200^\sim$, (*c*) $1{,}000^\sim$; high field.
3. Armco iron at (*a*) $25^\sim$, (*b*) $200^\sim$, (*c*) $1{,}000^\sim$; 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 magnetized 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.*

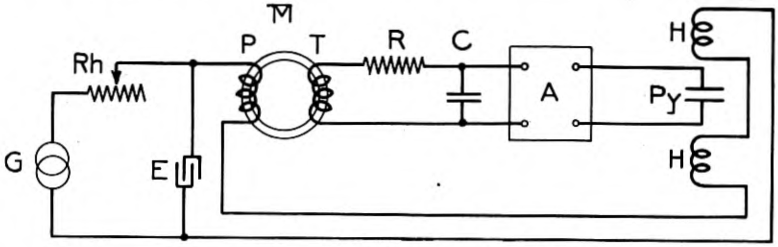


Fig. 1—Elementary diagram of the circuit.

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

$$B = \frac{b}{d}\,\frac{RC}{NS} \times 10^8 \text{ gauss.} \tag{2}$$

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.

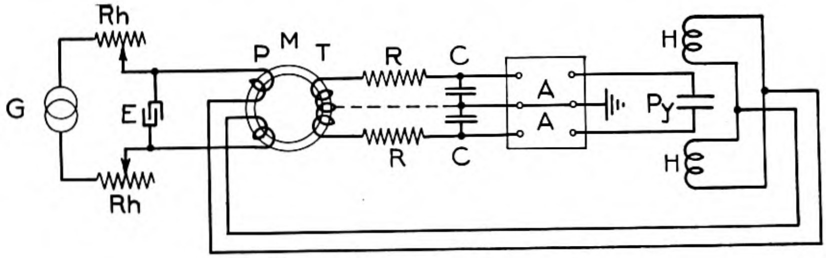


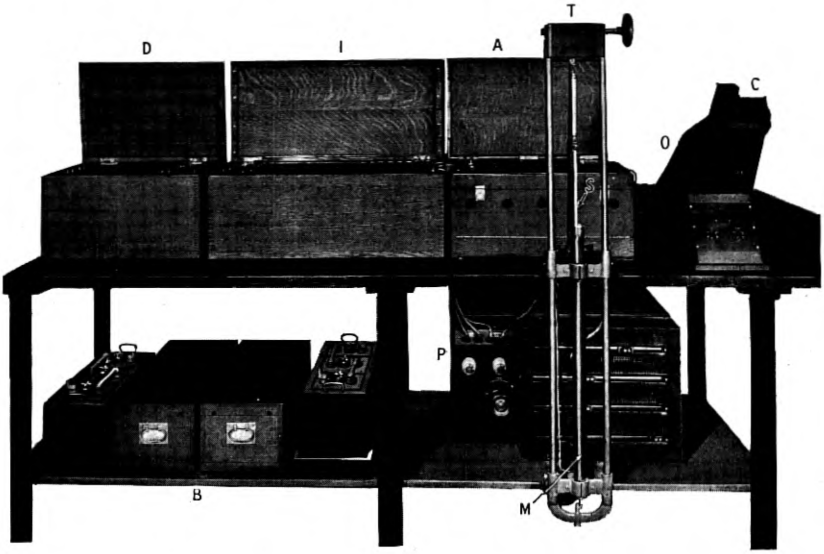
Fig. 2—Diagram of the symmetrical circuit.

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 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

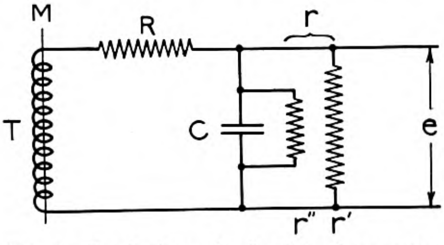


Fig. 3—Equivalent circuit of the integrator.

$$G = \frac{\omega C}{Z_1} + \frac{\omega^2 C}{Z_2} + \cdots. \tag{3}$$

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 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

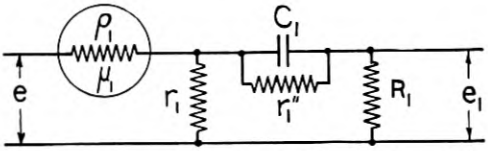


Fig 4—Equivalent circuit of one stage of the amplifier.

$$\left.\begin{array}{c} e_1 = \dfrac{NS}{RC}\dfrac{B\mu_1R_1r_1}{(R_1+r_1)(\rho_1+r_1)-r_1^2} \\ \left[\cos\omega t-\left(\dfrac{1}{r'C\omega}+\dfrac{1}{RC\omega}+\dfrac{1}{R_1C_1\omega}+\dfrac{2}{Z}\right)\sin\omega t\right]\end{array}\right\}. \tag{8}$$

Similarly, the output from s stages of the amplifier is

$$e_s=\frac{NS}{RC}B\left[\prod_1^s M_s\right]\left[\cos\omega t-\left(\frac{1}{r'C\omega}+\sum_0^s\frac{1}{R_sC_s\omega}+\frac{s+1}{Z}\right)\sin\omega t\right], \tag{9}$$

where

$$M_s=\frac{\mu_sR_sr_s}{(R_s+r_s)(\rho_s+r_s)-r_s^2}\doteq\mu_s\frac{r_s}{\rho_s+r_s}.$$

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

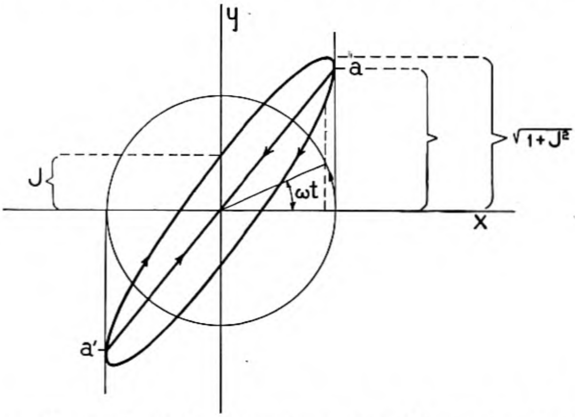


Fig. 5—Distortion produced by integrator and amplifier at low flux densities.

the maximum amplitude in the B direction may be taken as unity. The ellipse is the curve

$$\left.\begin{aligned} x &= \cos \omega t, \\ y &= \cos\cdot\omega t - \sum_0^s J_s \sin \omega t, \end{aligned}\right\} \qquad (10)$$

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

$$\begin{aligned} e_s &= \frac{NS}{RC} B_m \frac{4}{\pi} \left[\prod_{1}^{s} M_s \right] \sum_{1}^{n} \left[(-1)^{n-1} \left\{ \frac{\cos m\,\omega t}{m} \right.\right. \\ &\quad \left.\left. - \left(\frac{1}{r'C\omega} + \sum_{0}^{s} \frac{1}{R_s C_s \omega} \right) \frac{\sin m\,\omega t}{m^2} - \frac{s+1}{Z} \frac{\sin m\,\omega t}{m} \right\} \right] \\ &= \frac{NS}{RC} B_m \left[\prod_{1}^{s} M_s \right] \left[\alpha - \left(\frac{1}{r'C\omega} \quad \sum_{0} \frac{1}{R_s C_s \omega} \beta - \frac{s+1}{Z} \gamma \right) \right], \end{aligned} \tag{12}$$

where

$$\alpha = 1 \text{ from } \omega t = -\frac{\pi}{2} \text{ to } +\frac{\pi}{2};$$

$$\alpha = -1 \text{ from } \omega t = \frac{\pi}{2} \text{ to } 3\frac{\pi}{2}; \text{ (Fig. 6)}$$

$$\beta = \omega t \text{ from } \omega t = -\frac{\pi}{2} \text{ to } +\frac{\pi}{2};$$

$$\beta = (\pi - \omega t) \text{ from } \omega t = \frac{\pi}{2} \text{ to } 3\frac{\pi}{2}; \text{ (Fig. 7)}$$

$$\gamma = \frac{4}{\pi} \sum_{1}^{n} (-1)^{n-1} \frac{\sin m\,\omega t}{m}; \text{ (Fig. 8)}$$

$$m = 2n - 1.$$

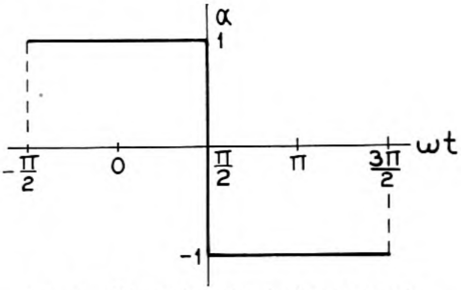


Fig. 6—The value of α during one cycle.

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).

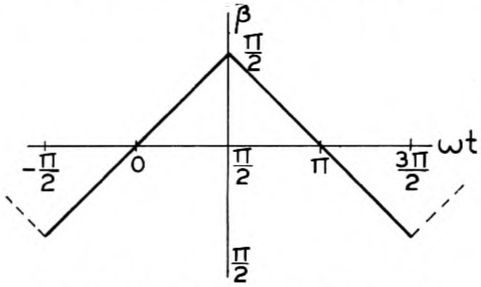


Fig. 7—The value of β during one cycle.

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 clockwise direction. When the apparatus is used for an actual material having positive hysteresis, the loop is made narrower at the *H* axis by this distortion, but the amount of the decrease in width

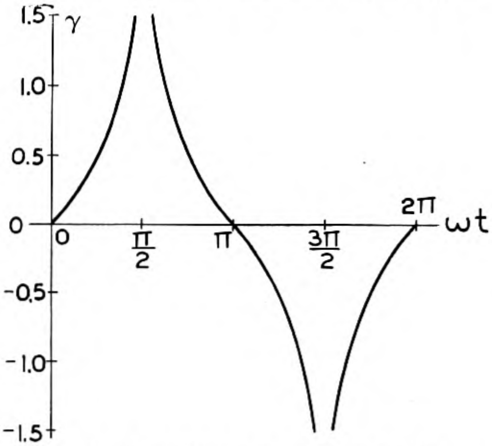


Fig. 8—The value of γ during one cycle.

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 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.

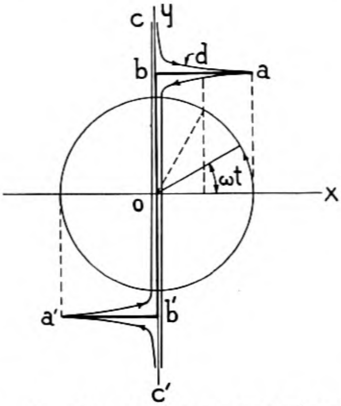


Fig. 9—Distortion produced by the integrator and amplifier for an ideally saturated material.

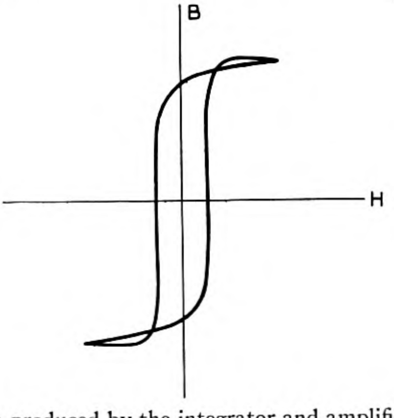


Fig. 10—Distortion produced by the integrator and amplifier with a normally saturated material.

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

$$B = \mu H = \mu H_1\left(\cos \omega t + \frac{4\pi}{10}\times 10^{-8}\frac{N^2S}{lR}\mu\omega \sin \omega t\right), \tag{14}$$

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

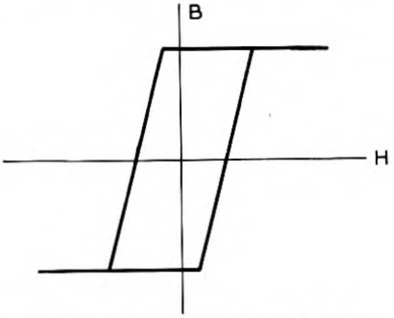


Fig. 11—Simplified hysteresis loop.

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

$$\Delta H = \frac{4\pi}{10} 10^{-8} \frac{N^2S}{lR} \frac{dB}{dt} = \frac{4\pi}{10} 10^{-8} \frac{N^2S}{lR} \frac{dB}{dH} \frac{dH}{dt} \cdot$$

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

$$\Delta H = \frac{4\pi}{10} 10^{-8} \frac{N^2S}{lR} \frac{dB}{dH} \omega H_1. \tag{15}$$

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

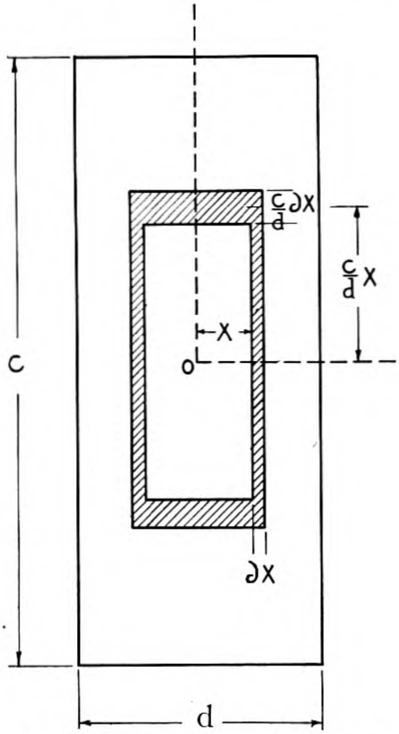


Fig. 12—Cross-section of specimen.

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

$$i_x = \frac{e}{r} = 10^{-8}\,\frac{d\varphi_{cx}}{dt}\,\frac{1}{\rho}\,\frac{cd}{4(c^2+d^2)}\,\frac{1}{x}\cdot \qquad (16)$$

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

$$\varphi_{cx} = \int_0^x d\varphi_x = \int_0^x 8\,\frac{c}{d}\,xdx\,\frac{dB}{dH}\left[H - H_c - \frac{4\pi}{10}\int_x^{d/2} i_x d_x \right]. \quad (17)$$

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,

$$\varphi = 8\,\frac{c}{d}\,\frac{dB}{dH}\,(H - H_c)\int_0^{d/2} xdx - 8\,\frac{c}{d}\,\frac{4\pi}{10}\,\frac{dB}{dH}\int_0^{d/2} xdx \int_x^{d/2} i_x dx \quad (18)$$

$$= cd\,\frac{dB}{dH}\,(H - H_c) - \varphi_2. \quad (19)$$

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

$$\Delta H_c = H_a - H_c = \frac{\pi}{4}\,\frac{10^{-9}}{\rho}\,d^2\,\frac{dB}{dH}\,\frac{dH}{dt}\cdot \quad (21)$$

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.

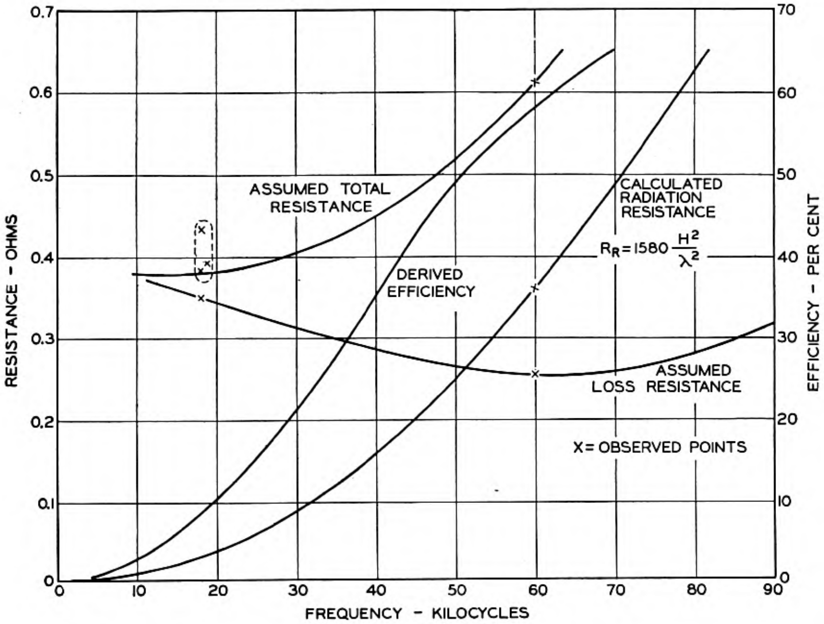


Fig. 1—Assumed resistance and efficiency of Rocky Point antenna. (Effective height 75 meters.)

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 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.)

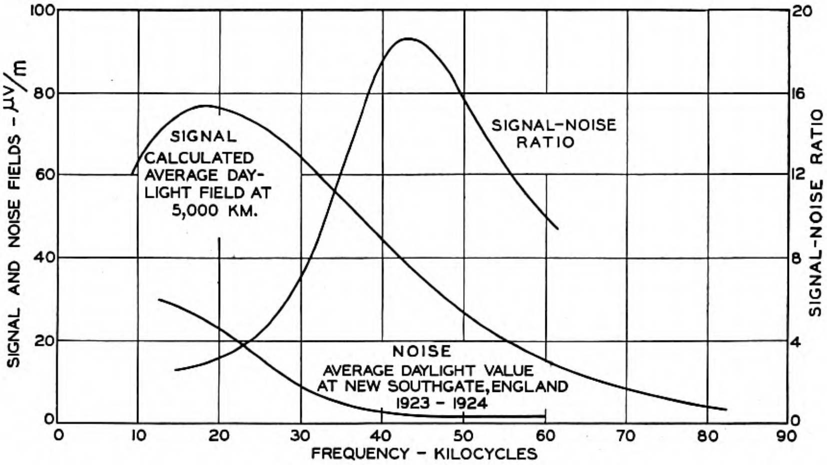


Fig. 2—Variation of signal, noise, and signal-noise ratio with frequency. Transmission from U. S. A. to England. 85.9 kw. supplied to antenna of Rocky Point characteristics.

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.

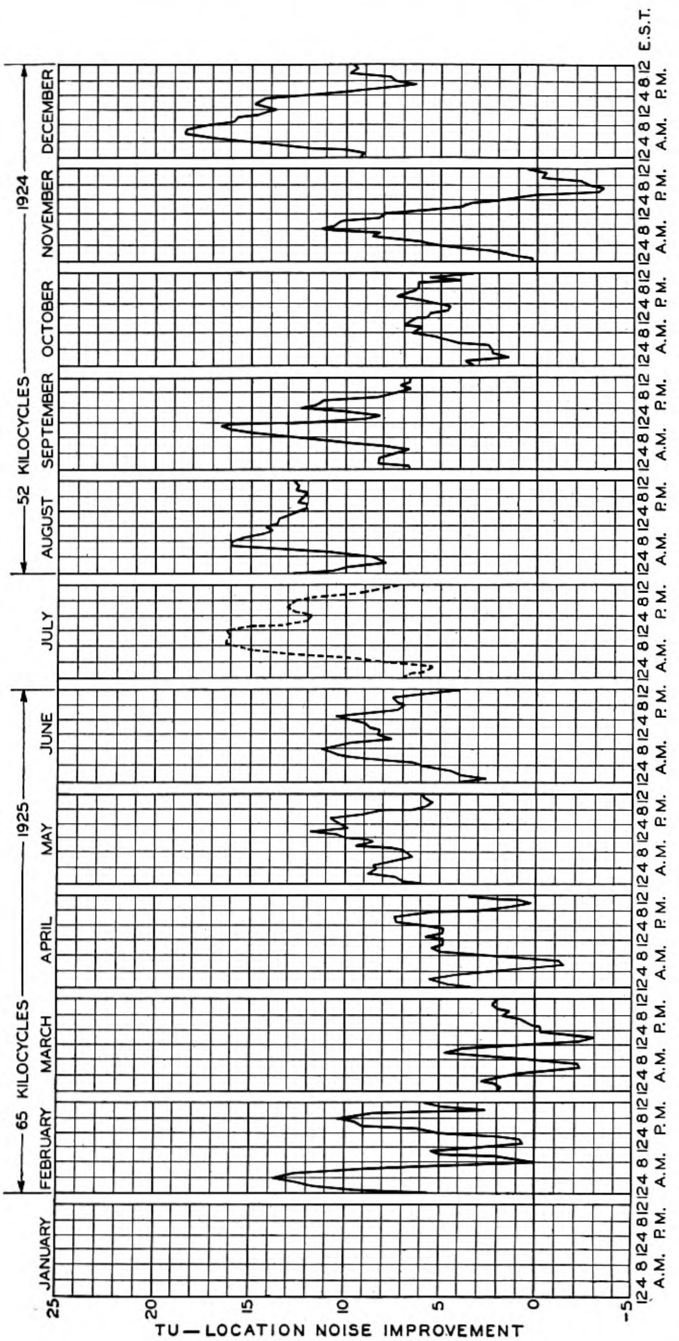


Fig. 3—Transatlantic radio noise measurements. Diurnal variations of location noise improvement (in transmission units) of Belfast, Maine, over Riverhead, New York. Three-hour moving averages of simultaneous observations.

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 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

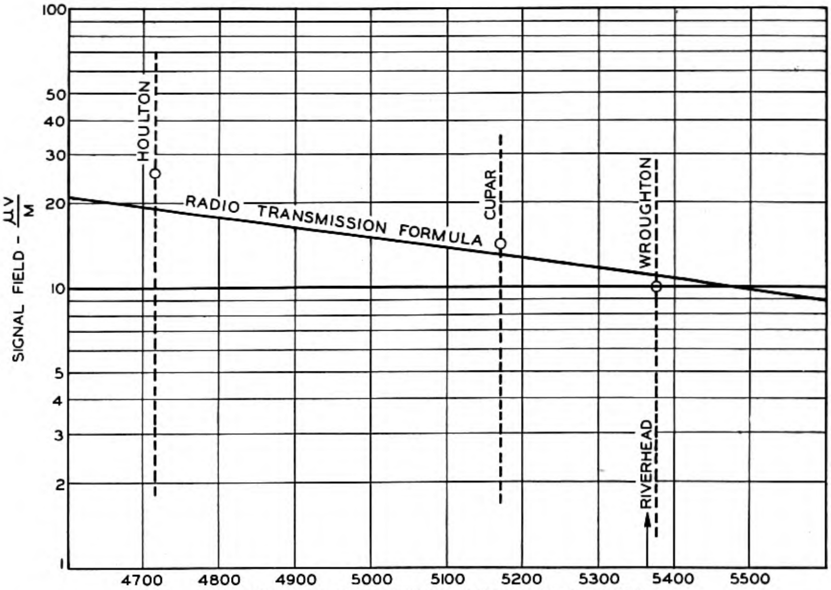


Fig. 4—Transatlantic radio daylight field strength. Average of hourly observations during 1927. Corrected to 50 kw. radiated power—frequency 60 kilocycles

[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.

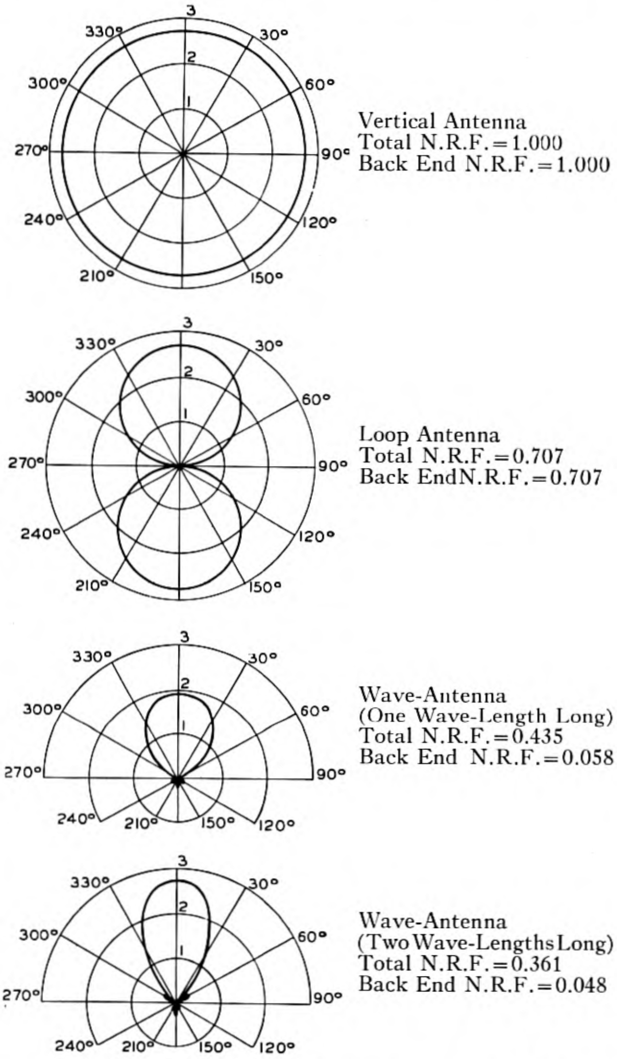
A measure of the directional discrimination of the various antenna types is the Noise Reception Factor (abbreviated *NRF*) which is defined as the ratio of the total noise current received from the antenna in question to that received from a vertical antenna under the conditions of continuous, constant 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 efficiently, it is desirable that they be tuned, introducing the frequency discrimination of a tuned circuit. If these types of antennas be used



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.)

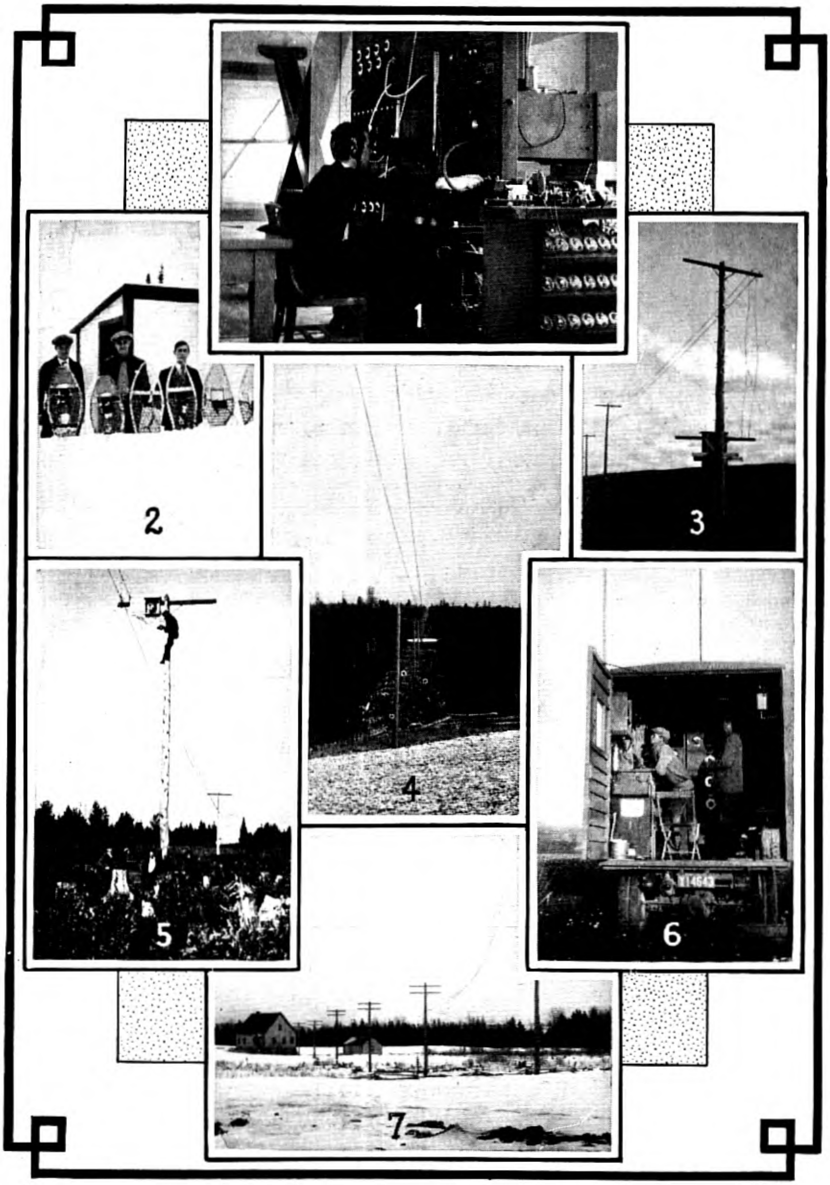
it is necessary, therefore, that the resonance characteristic be studied and means provided to eliminate excessive frequency discrimination within the desired band. On the other hand, the wave-antenna is an aperiodic structure and, in consequence, its transmission-frequency characteristic is so flat 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.

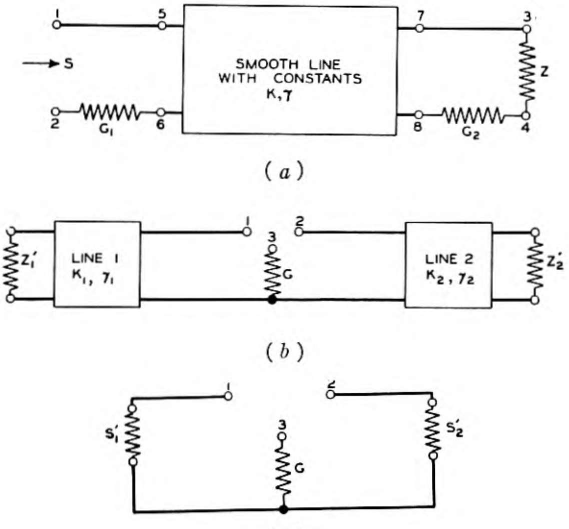
The Wave-Antenna

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 available 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:



(*a*)

(*b*)

(*c*)

Fig. 6.

$$Z' = Z + G_2 \tag{1}$$

$$S' = S - G_1 \tag{2}$$

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' fiom 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 deteımination 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:

$$K = \sqrt{\frac{(S_1 - G_1)(S_2 - G_1)(Z_1 - Z_2) + (Z_1 + G_2)(Z_2 + G_2)(S_2 - S_1)}{(S_2 - S_1) + (Z_1 - Z_2)}} \qquad (10)$$

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 of 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 cuirent 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 ot 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

$$|I_\theta| = H_\theta \epsilon^{-\alpha S\lambda'} \left|\frac{E'}{2K}\right|. \tag{14}$$

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:

$$(a + jb) = \frac{h}{S\lambda'}\left(1 - \epsilon^{-[\alpha S\lambda' + j2\pi S(m - \cos\theta)]}\right)\epsilon^{-j2\pi S\cos\theta} \tag{17}$$

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

$$g = \frac{H_\theta}{S\lambda'}\epsilon^{-\alpha S\lambda'}. \tag{20}$$

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.

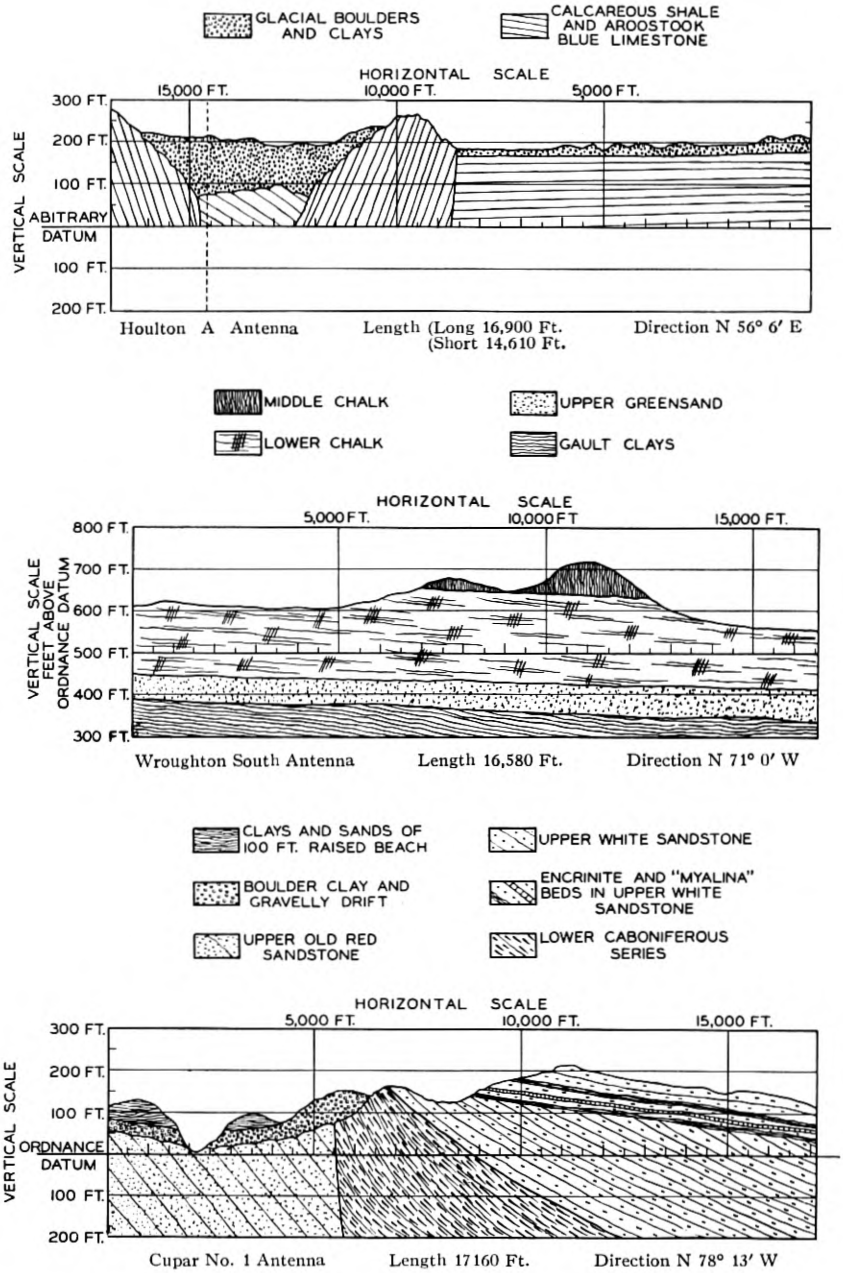
The probable geological formations under individual antennas at each of the three antenna sites mentioned in the preceding paragraph are shown in Fig. 7. The data for the British locations were compiled from the published reports of geological surveys conducted by the British Government, and the data 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-



Houlton A Antenna Length (Long 16,900 Ft. (Short 14,610 Ft. Direction N 56° 6′ E

Wroughton South Antenna Length 16,580 Ft. Direction N 71° 0′ W

Cupar No. 1 Antenna Length 17160 Ft. Direction N 78° 13′ W

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

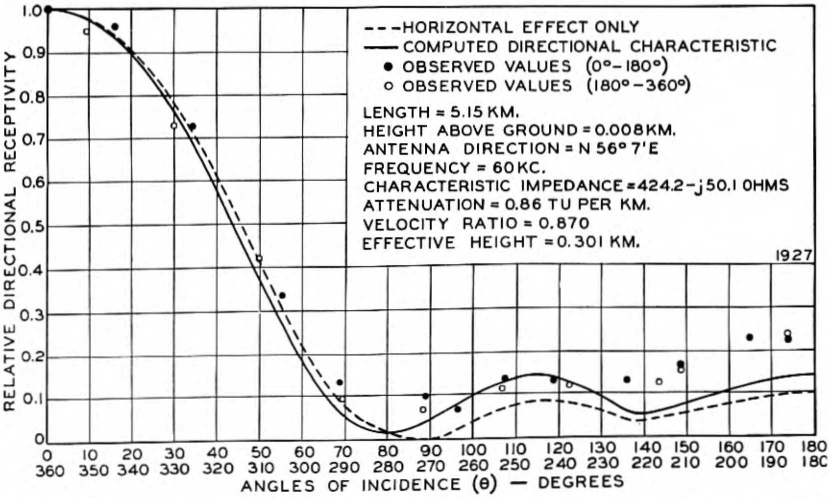


Fig. 8*a*—Relative Directional Receptivity of Houlton Antenna "A" Uncompensated (Long)

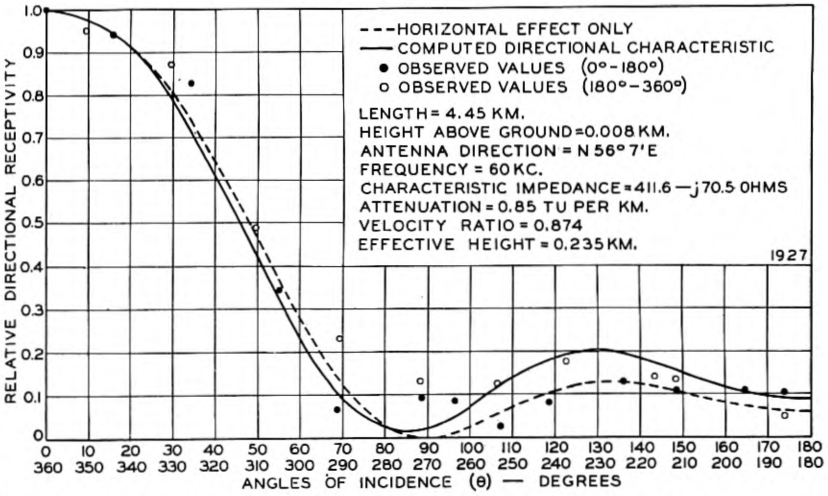


Fig. 8*b*—Relative Directional Receptivity of Houlton Antenna "A" Uncompensated (Short)

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.

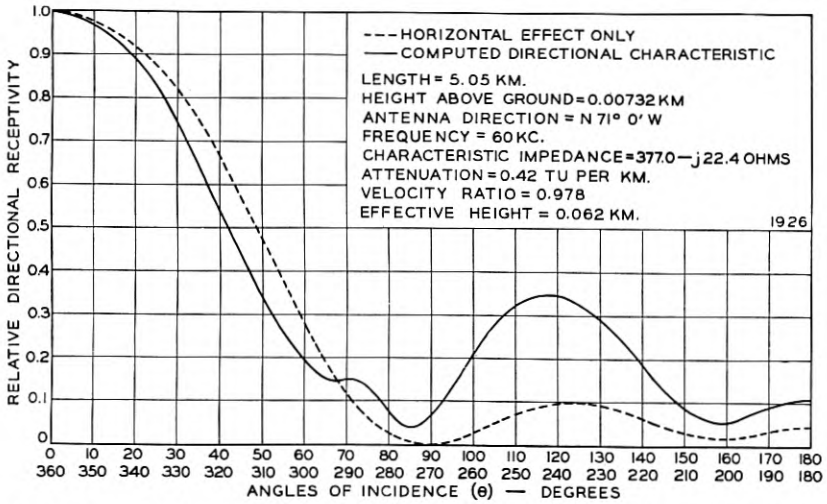


Fig. 8*c*—Relative Directional Receptivity of Wroughton, England—South Antenna

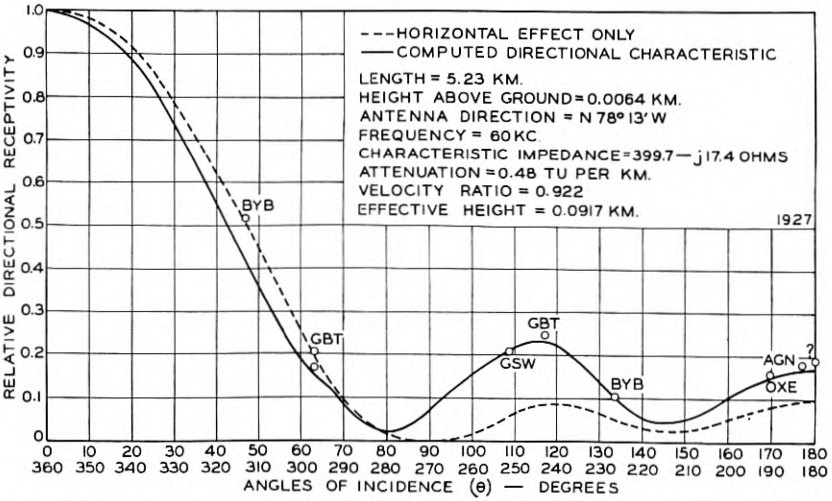


Fig. 8*d*—Relative Directional Receptivity of Cupar, Scotland—Antenna No. 1

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.

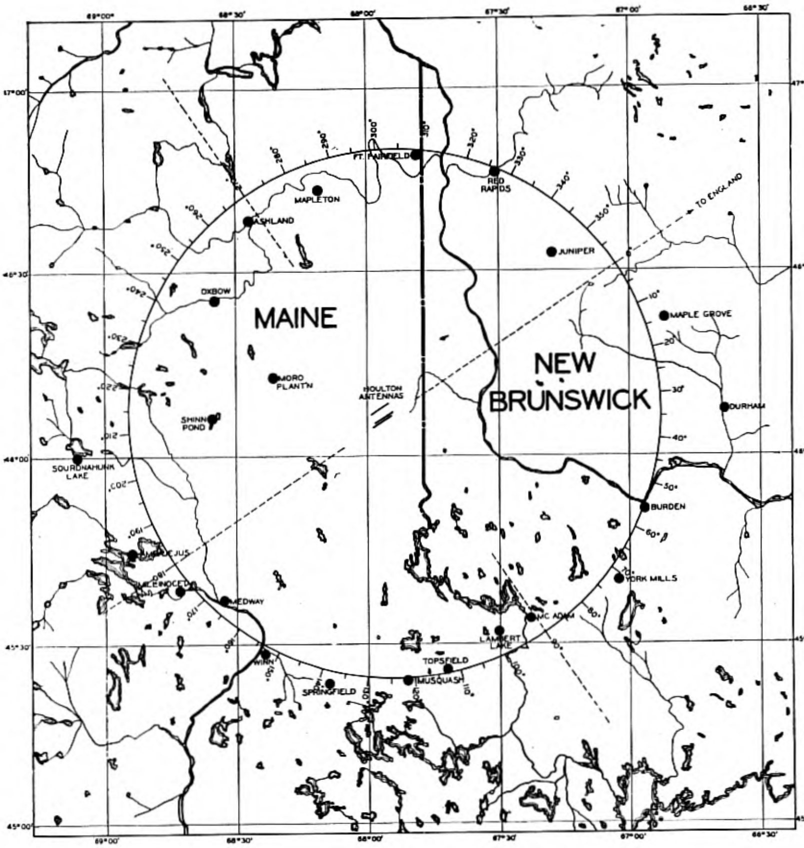


Fig. 9.

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.

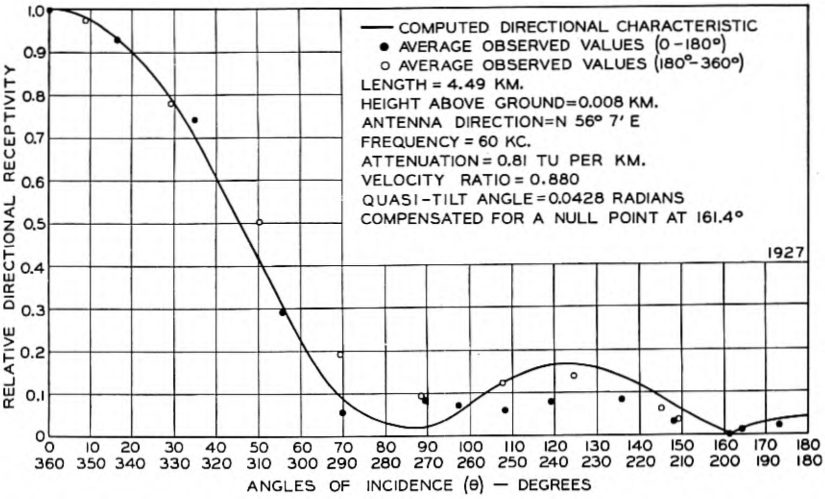


Fig. 10—Wave-antenna directional characteristic. Relative directional receptivity of compensated average Houlton antenna. (Short.)

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-

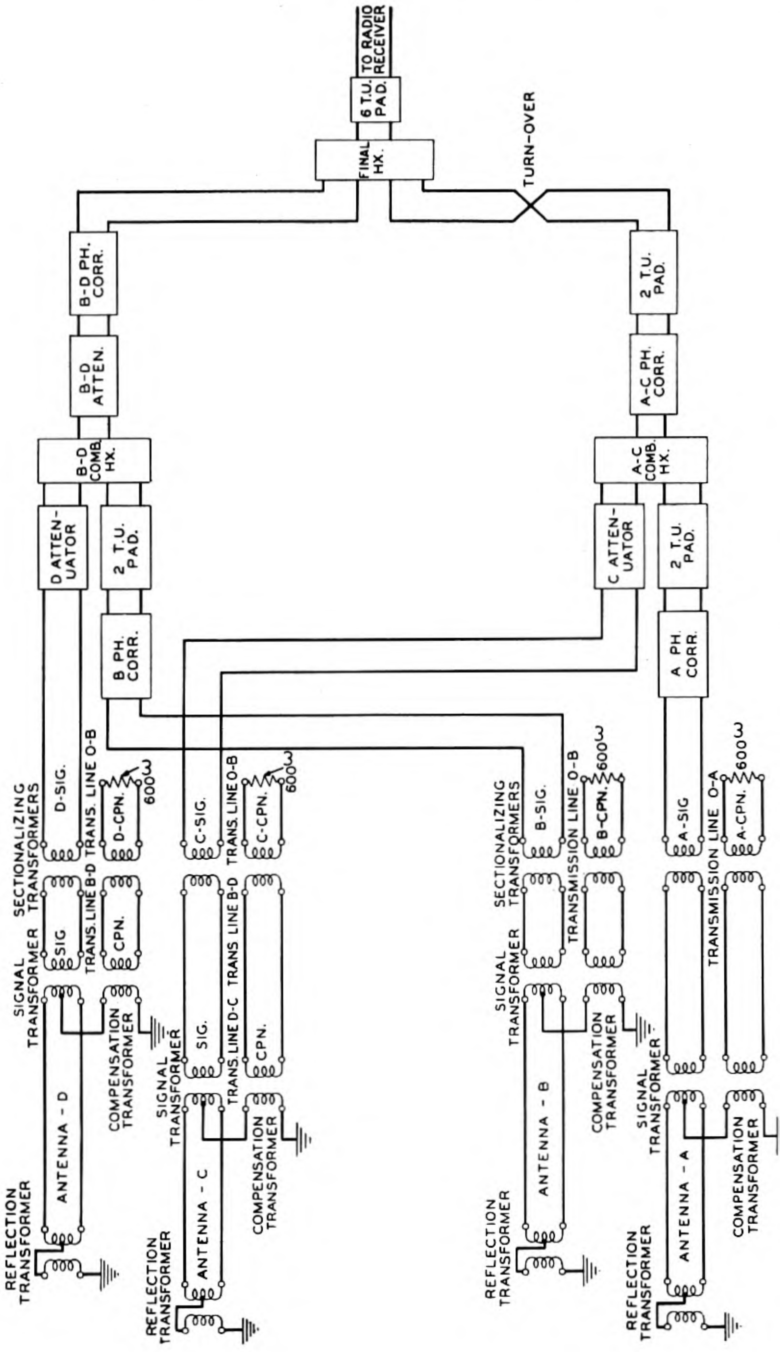


Fig. 11—Houlton antenna system.

antennas are used in either a lateral array, a longitudinal array, or a combination of the two.

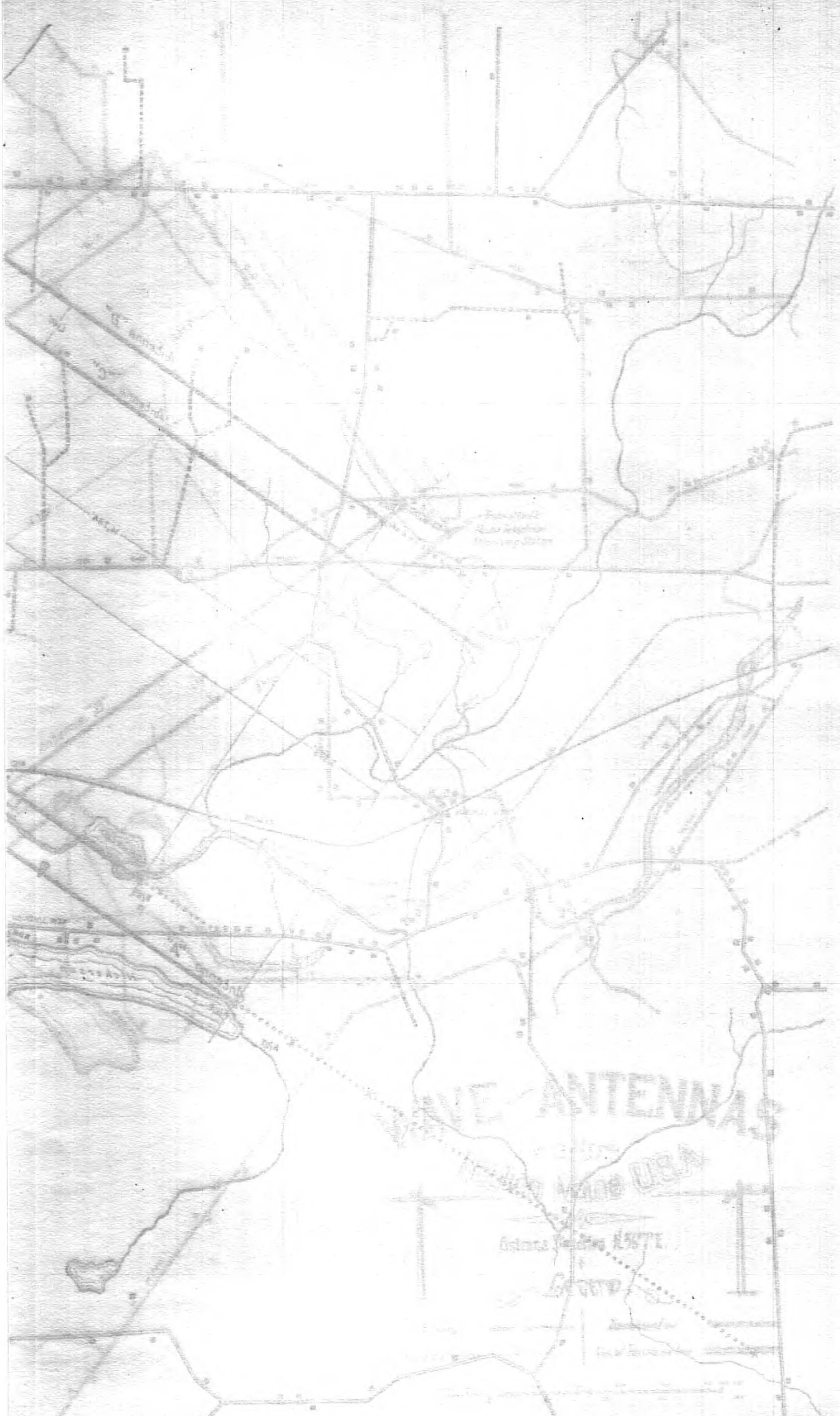
In the lateral array, the initial ends of the wave-antennas are spaced in the direction perpendicular to the axes of the antennas. Since it extends over space in the lateral direction, unless there be undue sacrifice in desired signal, the lateral array can only reduce the width of the directional diagram.

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


ANTENNAS
USA

quite evidently combines the narrowing of the directional diagram due to the lateral array and the reduction of the back end area of the

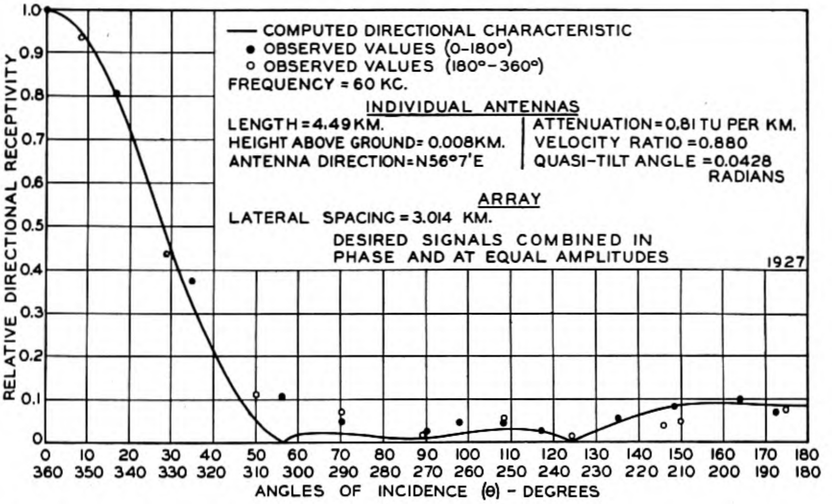


Fig. 13—Wave-antenna array directional characteristic. Relative directional receptivity of lateral array of two Houlton antennas. (Short)

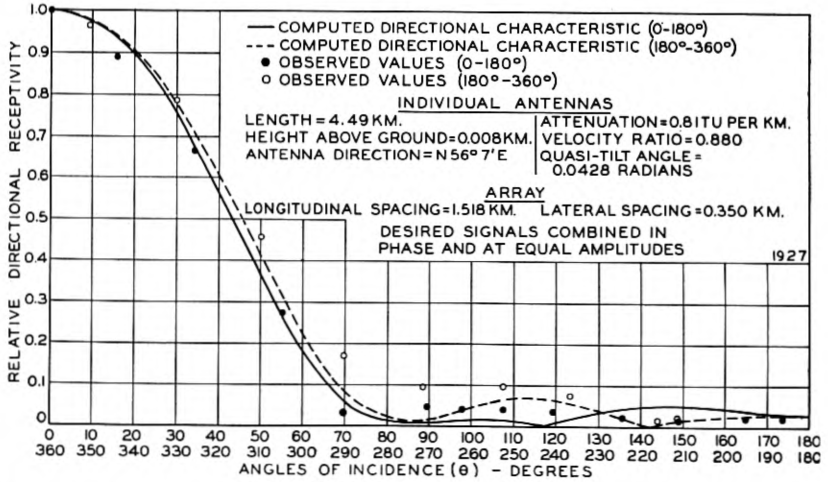


Fig. 14—Wave-antenna array directional characteristic. Relative directional receptivity of modified longitudinal array of two Houlton antennas. (Short)

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 aiive 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*, 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.

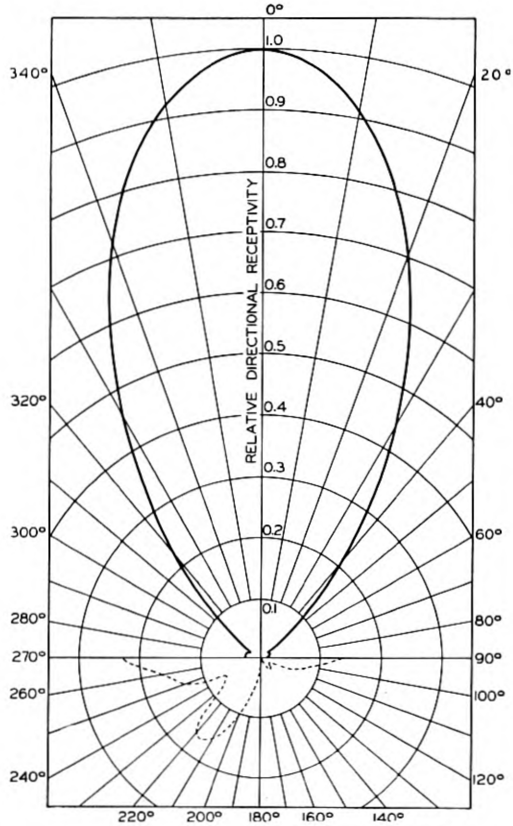


Fig. 15—Wave-antenna array directional characteristic. Calculated relative directional receptivity of array of Houlton antennas *A*, *B*, *C*, and *D*. (From average measured unit antenna characteristic.) Dotted curve—magnified × 10.

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.

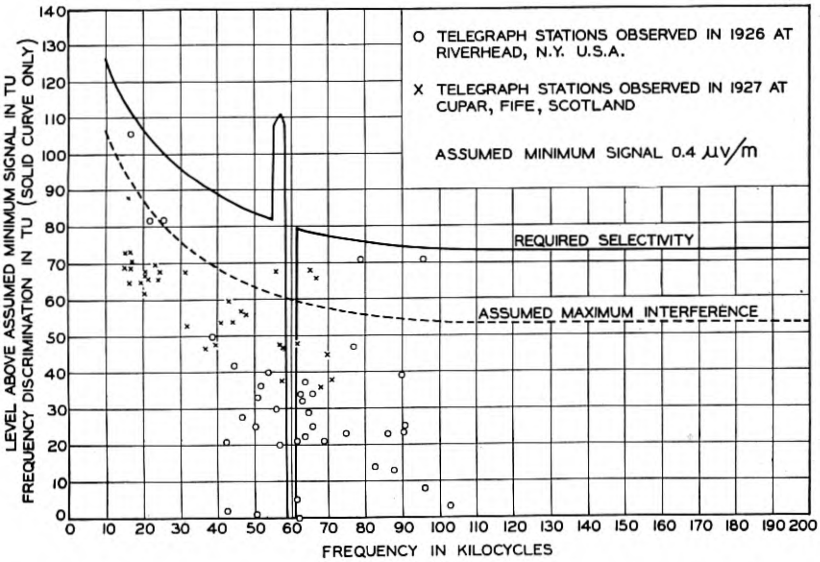


Fig. 16—Selectivity requirements for the long wave transatlantic radio-telephone receiving system.

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):

$$\text{Relative desired signal} = (0) + (\text{Beating oscillator voltage}),$$
$$\text{Relative interference} = (+\,60) + (-\,20) = +\,40.$$

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 il 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 ± 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

$$\begin{aligned} \phi'' &= \phi' + \phi, & f_w'' &= f_w' + f_w, \\ V'' &= V' + V, & f_g'' &= f_g' + f_g, \end{aligned} \tag{101}$$

[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.

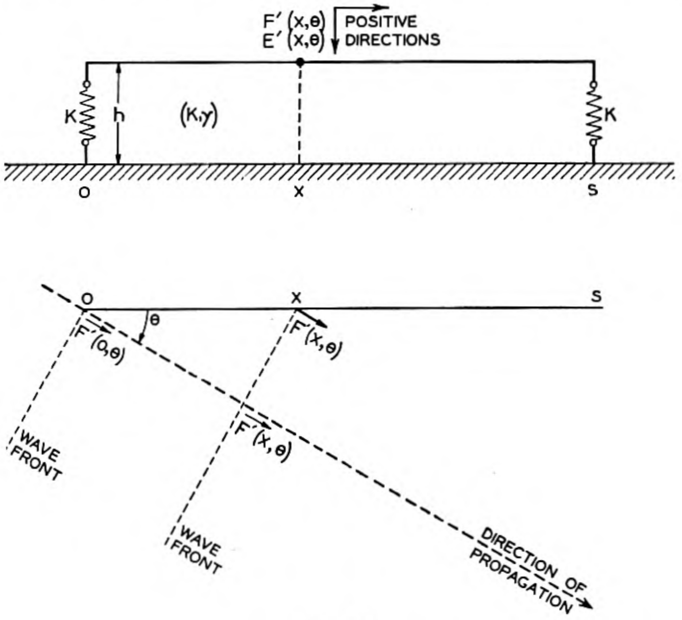


Fig. 17.

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

$$f_w' - f_g' + \frac{dV'}{dx} = -\frac{d\phi'}{dt}. \qquad (106)$$

Subtracting (106) from (104) we get

$$zI - f_w' + \frac{dV}{dx} = -\frac{d\phi}{dt}. \qquad (107)$$

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

$$zI + L\frac{dI}{dt} + \frac{1}{C}\frac{dQ}{dx} = f_w', \qquad (108)$$

but the line current is decreased by the amount of the charging current and the leakage current

$$-\frac{dI}{dx} = \frac{dQ}{dt} + I_Y, \qquad (109)$$

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

$$I_Y = YV'' = Y(V' + V). \qquad (110)$$

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

$$-\frac{d^2I}{dx^2} = \frac{dQ}{dx}j\omega + Y\frac{dV'}{dx} + \frac{Y}{C}\frac{dQ}{dx}. \qquad (111)$$

By means of (111) we may eliminate Q from (108)

$$(z + jL\omega)I - \frac{1}{Y + jC\omega}\frac{d^2I}{dx^2} = f_w' + \frac{Y}{Y + jC\omega}\frac{dV'}{dx} \qquad (112)$$

and if

$$K = \sqrt{\frac{z + jL\omega}{Y + jC\omega}}, \qquad (113)$$

$$\gamma = \sqrt{(z + jL\omega)(Y + jC\omega)}, \qquad (114)$$

where K is the characteristic impedance and γ the propagation constant of the antenna circuit, equation (112) may be written

$$\frac{K}{\gamma}\left(\gamma^2 - \frac{d^2}{dx^2}\right) I = f_w' + \frac{YK}{\gamma}\frac{dV'}{dx} \cdot \tag{115}$$

When the boundaiy 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)

$$I_{E'\theta} = \frac{hE'(0)}{2K}\left[\epsilon^{(\gamma-\gamma'\cos\theta)s} - 1\right]e^{-\gamma s}. \tag{123}$$

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'	Velocity ratio.	numeric
$m \equiv V'/V = \beta/\beta'$	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. $= \dfrac{I_\theta}{I_0}$	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, a general array of n antennas is indicated, of which only the first and the k'th are portrayed.

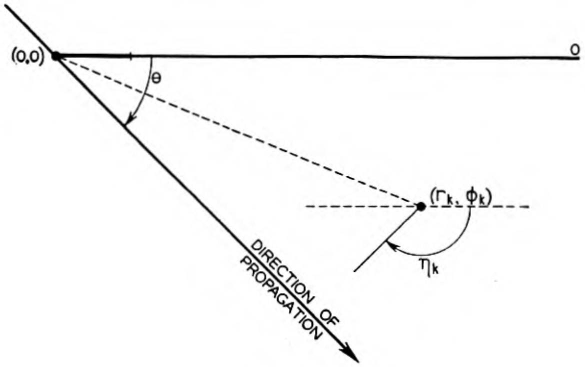


Fig. 18

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 ot 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

$$\beta' = \frac{2\pi V'}{\lambda'} \cdot \tag{204}$$

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.

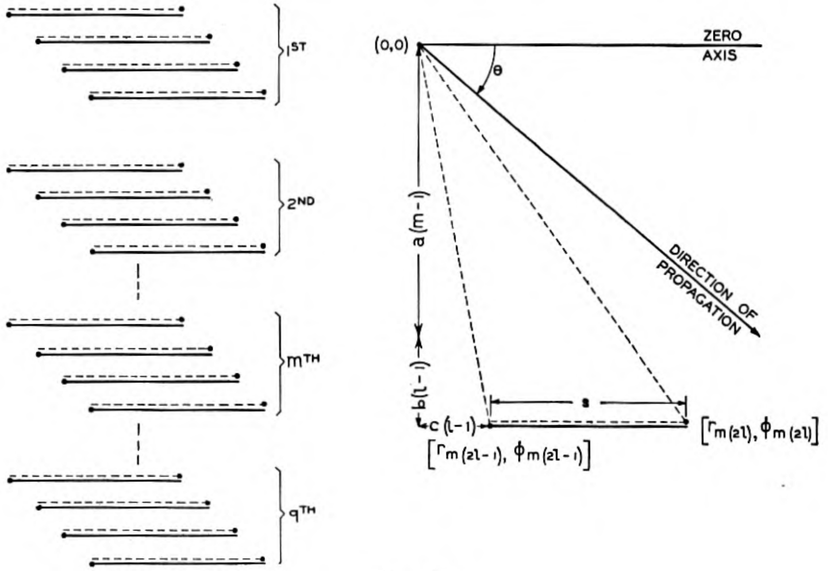


Fig. 19

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

$$-\frac{F'}{E'} = \epsilon^{j\delta}\tan T = \sqrt{\frac{\dfrac{9\times 10^{11}}{\rho_1} + j\dfrac{1}{4\pi}\omega k_1}{\dfrac{9\times 10^{11}}{\rho_2} + j\dfrac{1}{4\pi}\omega k_2}}, \tag{301}$$

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:

$$\begin{aligned} \rho_1 &= \infty && \text{(air)},\\ \rho_2 &= \rho && \text{(earth)},\\ k_1 &= 1 && \text{(air)},\\ k_2 &= k && \text{(earth)}. \end{aligned}$$

Substituting these values into the general equation (301)

$$-\frac{F'}{E'} = \epsilon^{j\delta}\tan T = \frac{1}{\sqrt{k}}\left[\frac{\left(\dfrac{fk\rho}{18\times 10^{11}}\right)^2}{1+\left(\dfrac{fk\rho}{18\times 10^{11}}\right)^2}\right]^{\frac{1}{4}} \epsilon^{j[1/2\tan^{-1}(18\times 10^{11}/fk\rho)]}. \quad (302)$$

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

$$\tan T = \frac{1}{\sqrt{k}}\left[\frac{\left(\dfrac{fk\rho}{18\times 10^{11}}\right)^2}{1+\left(\dfrac{fk\rho}{18\times 10^{11}}\right)^2}\right]^{\frac{1}{4}}, \tag{303}$$

$$\delta = \frac{1}{2}\tan^{-1}\left(\frac{18\times 10^{11}}{fk\rho}\right). \tag{304}$$

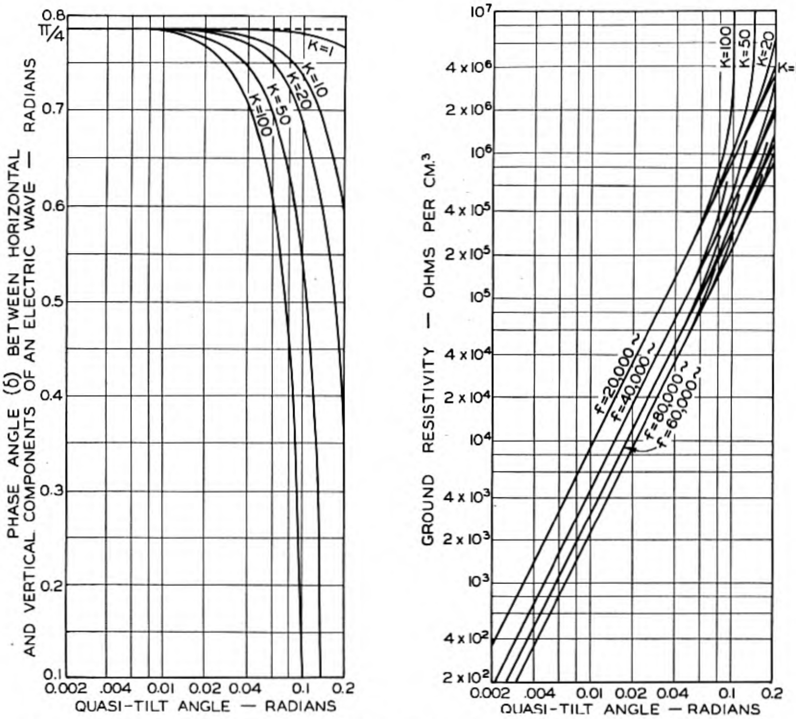


Fig. 20—Relation between quasi-tilt angle, ground resistivity, and phase angle between horizontal and vertical components of an electric wave. (By Zenneck's formula.)

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 probabi'ity 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

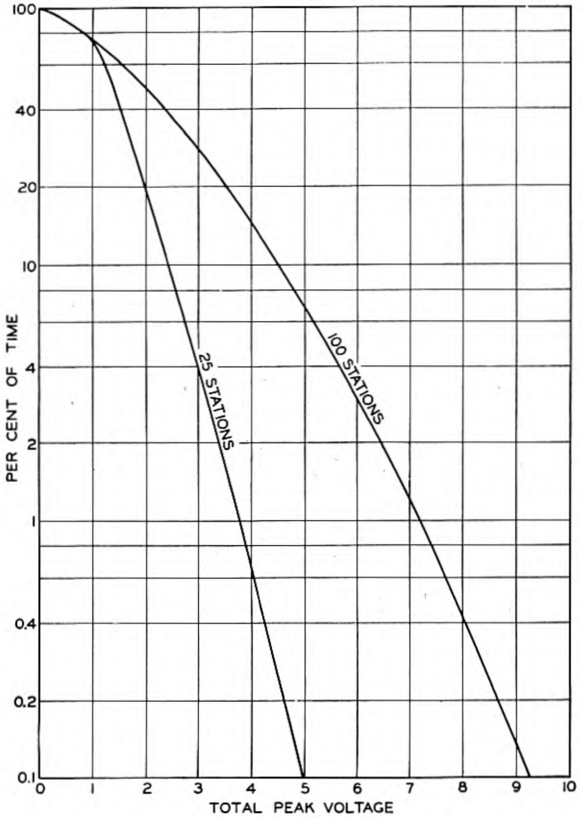


Fig. 21—Voltages resulting from several unit voltages each applied 15 per cent of the time and with random phase and frequency

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 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.

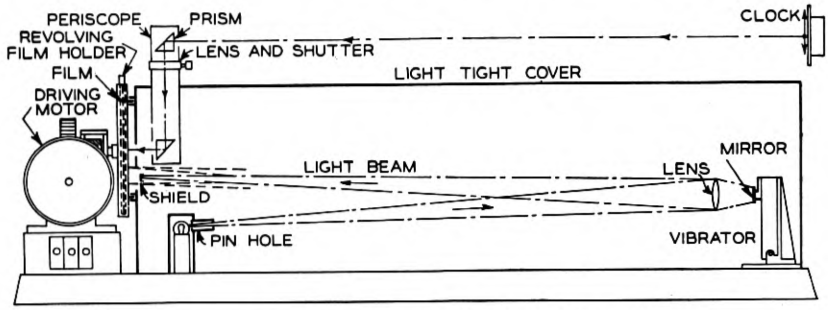


Fig. 1—Essential elements of polar oscillograph.

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.

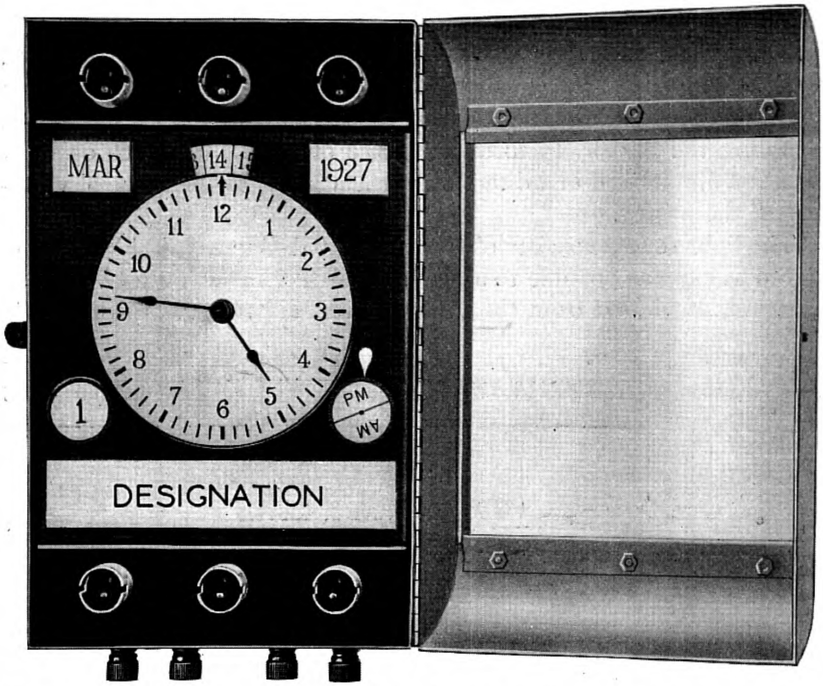


Fig. 2—Calendar clock used with oscillographs. Door is open to show lamp sockets.

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.

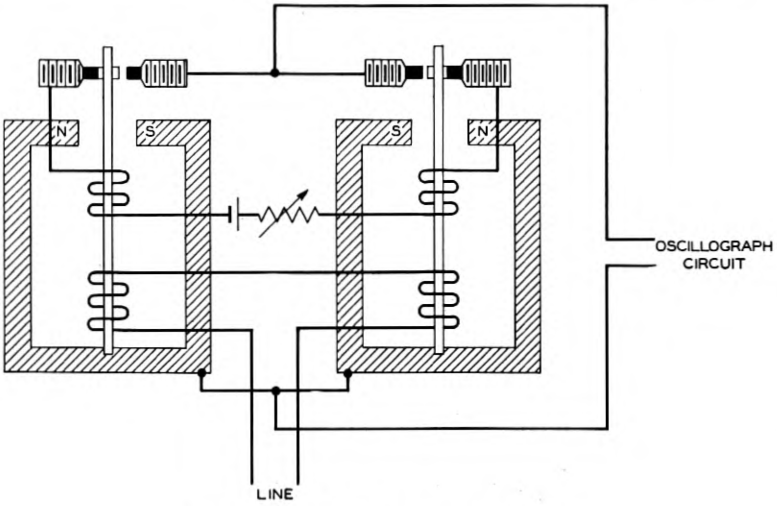


Fig. 3—Circuit of high-speed line-relay.

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

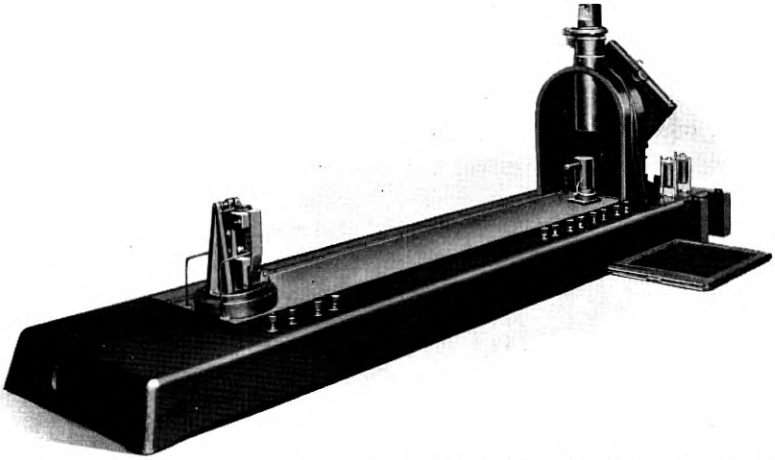


Fig. 4—Polar oscillograph, showing film rotor, periscope, lamp housing, and vibrator.

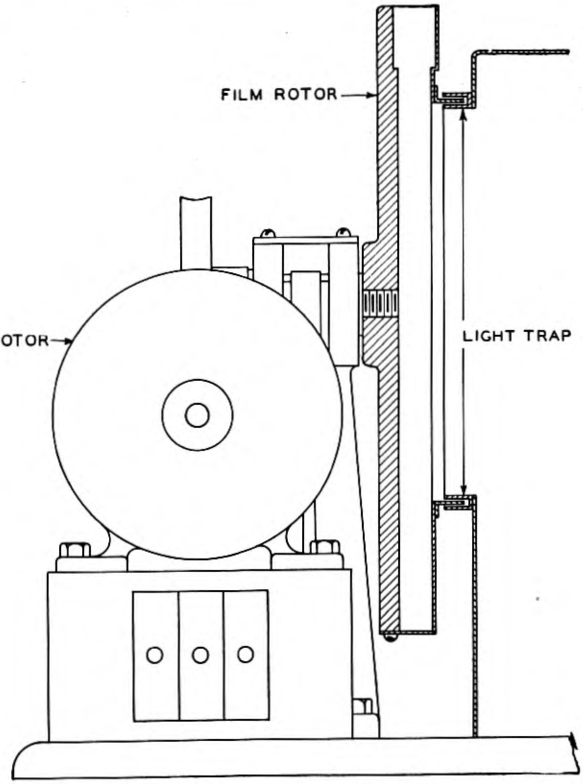
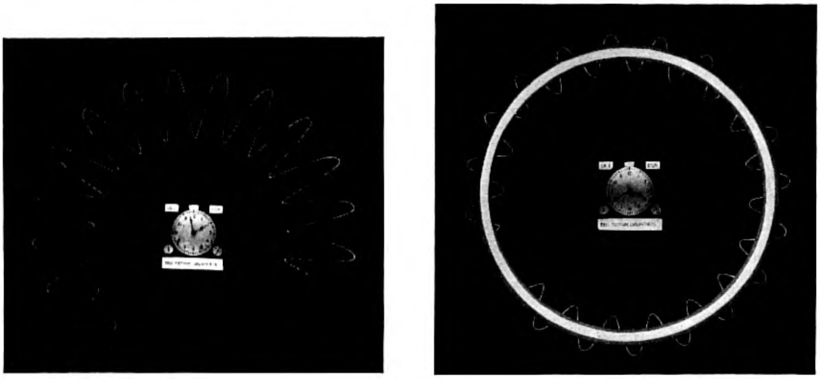


Fig. 5—Scale drawing of rotating light trap.

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, clear space is left where a zero line is usually obtained. This is illustrated in Fig. 6-A.

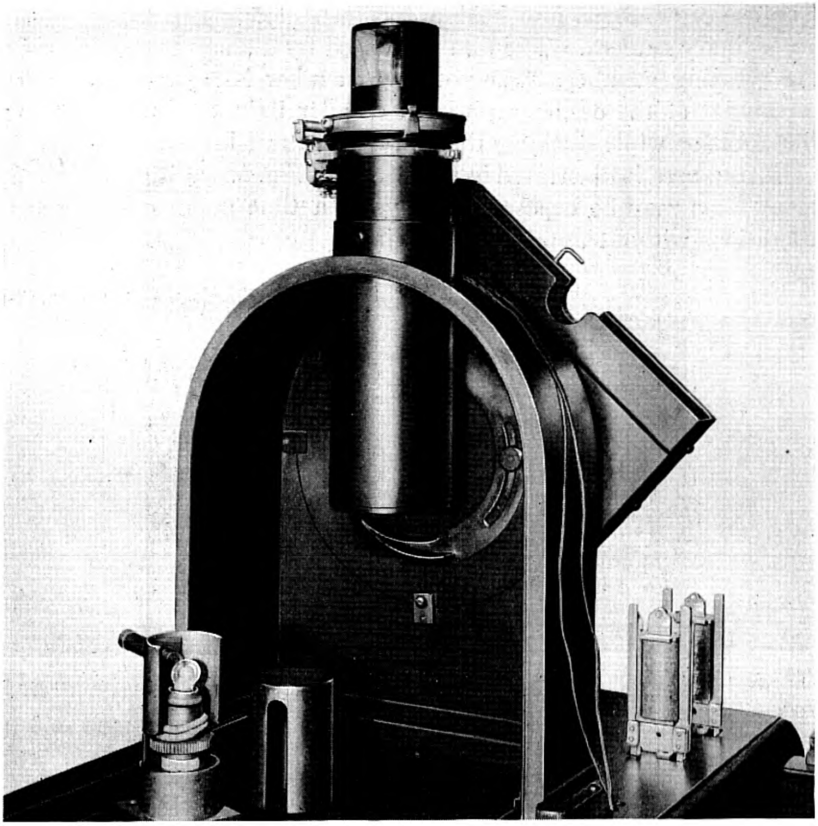


Fig. 7—Shields to prevent fogging on polar oscillograph.

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.

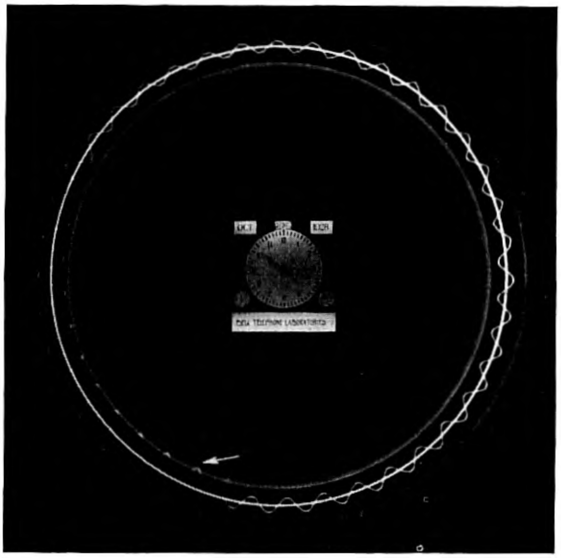


Fig. 8—Oscillogram illustrating the use of removable shield.

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 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

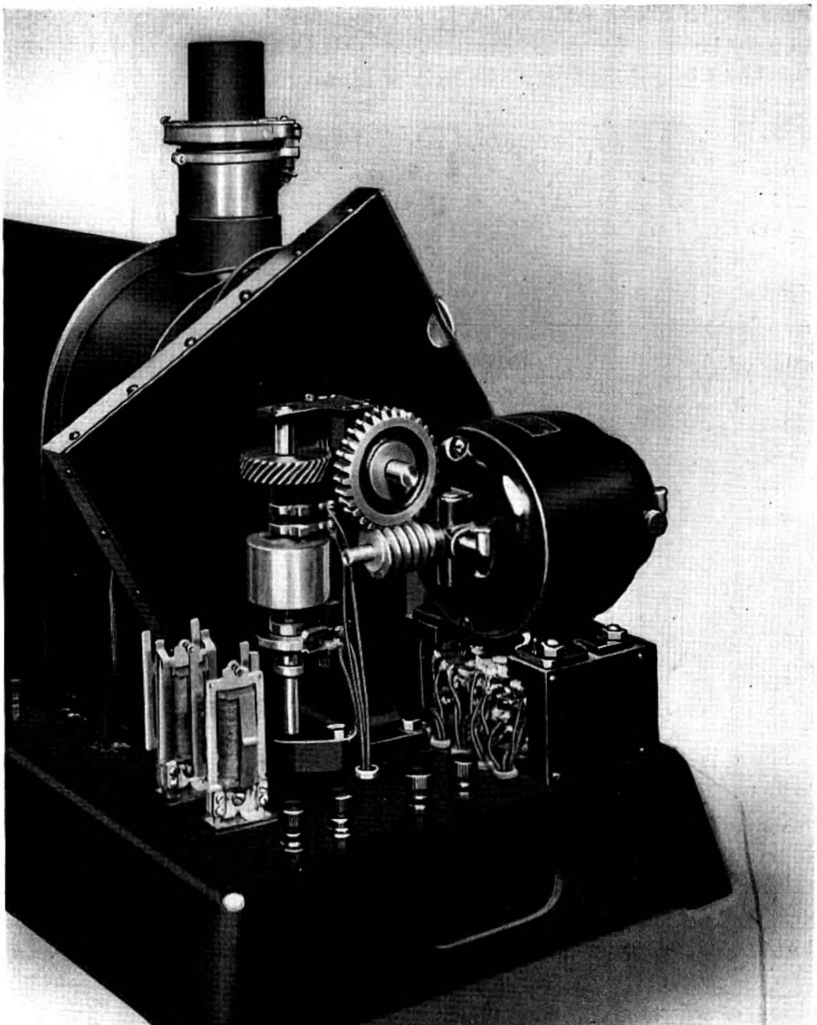


Fig. 9—Device on polar oscillograph to limit recording time.

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 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 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.

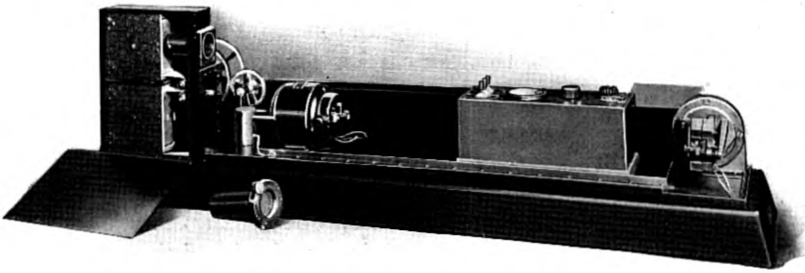


Fig. 10—Continuous-film oscillograph with covers removed.

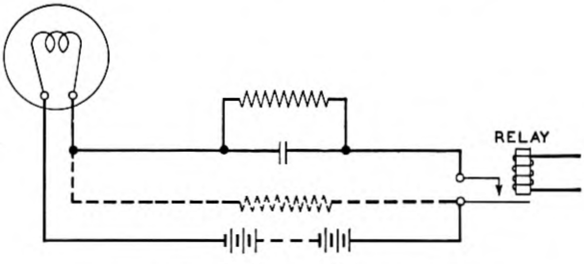


Fig. 11—Circuit for lighting lamp quickly.

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, 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.

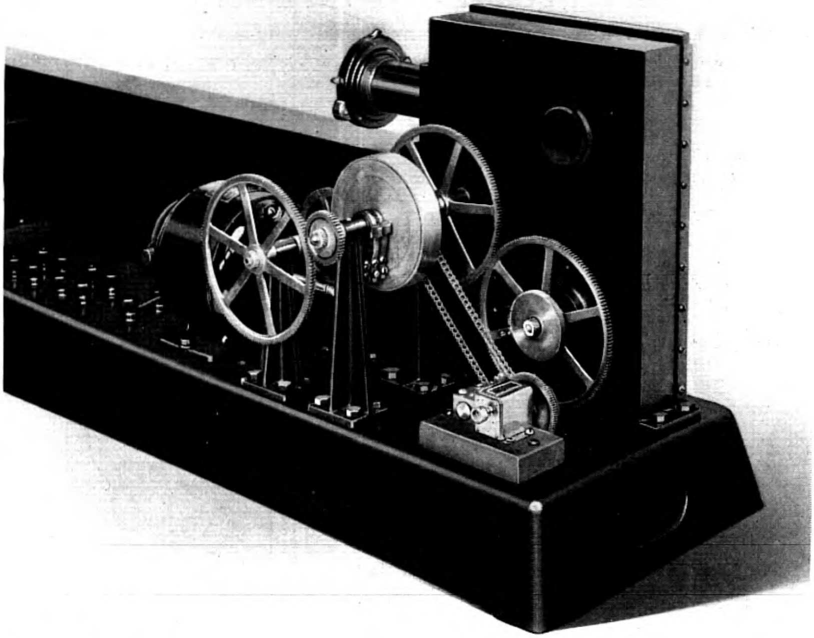


Fig. 12—Quick acting magnetic clutch on continuous-film oscillograph.

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. A resistance of suitable value connected as indicated by dotted lines in Fig. 11 will accomplish this result.

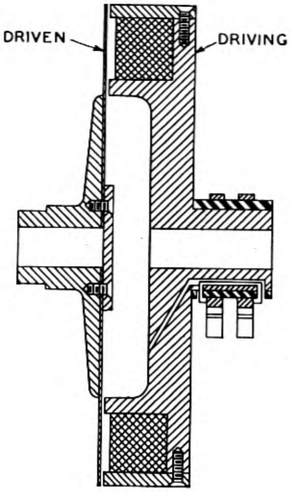


Fig. 13—Sectional scale drawing of magnetic clutch.

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.

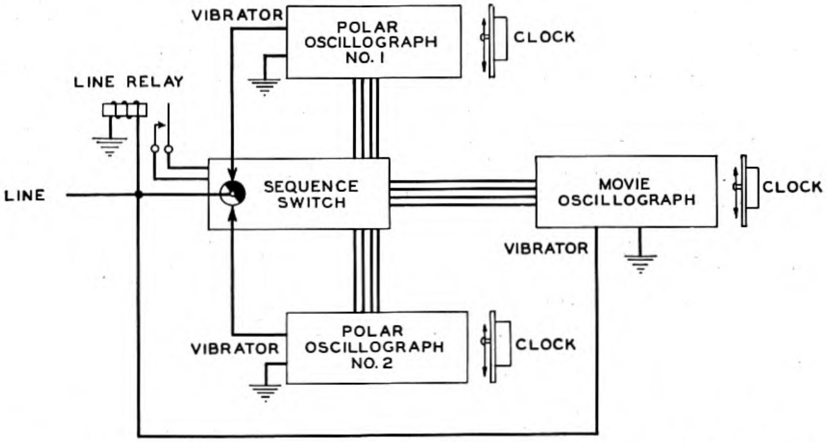


Fig. 14—Arrangement of oscillographs for recording any transient in a line. Two polar oscillographs and one continuous-film oscillograph with control equipment and clocks are arranged so that a transient of any duration occurring at any time will be recorded with a record of the time of occurrence.

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 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.

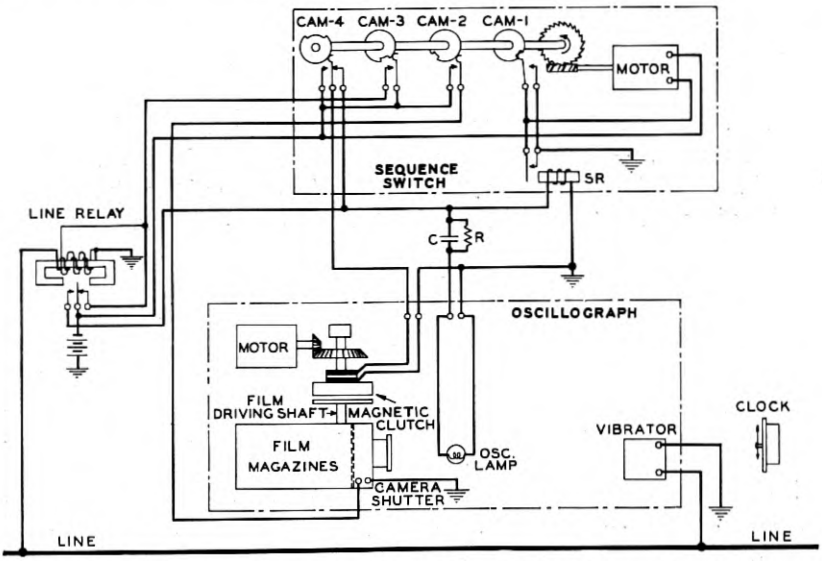


Fig. 15—Schematic arrangement of continuous-film oscillograph with sequence switch, line-relay, and clock, arranged for automatic operation.

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.

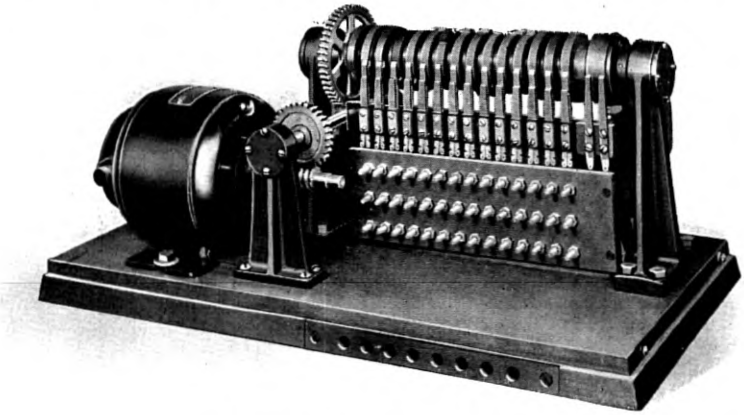


Fig. 16—Sequence switch.

Fig. 16 shcws 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 especially for sampling a wave at regular short intervals instead of making a continuous record or merely a record of unusual disturbances.

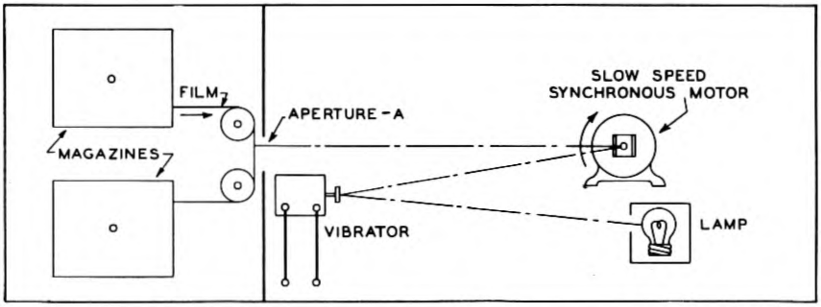


Fig. 17—Schematic drawing of sampling oscillograph. This oscillograph records one wave out of a number at regular intervals, say one cycle in sixty, with considerable resolution in order to record slow variations in wave form.

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.

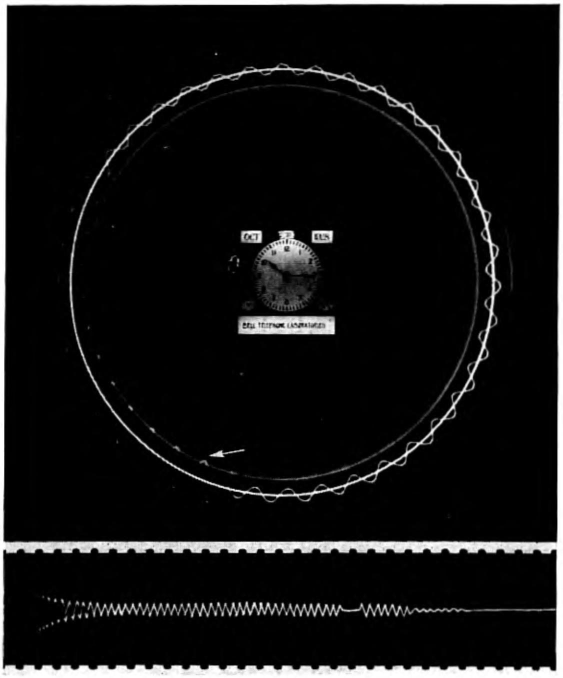
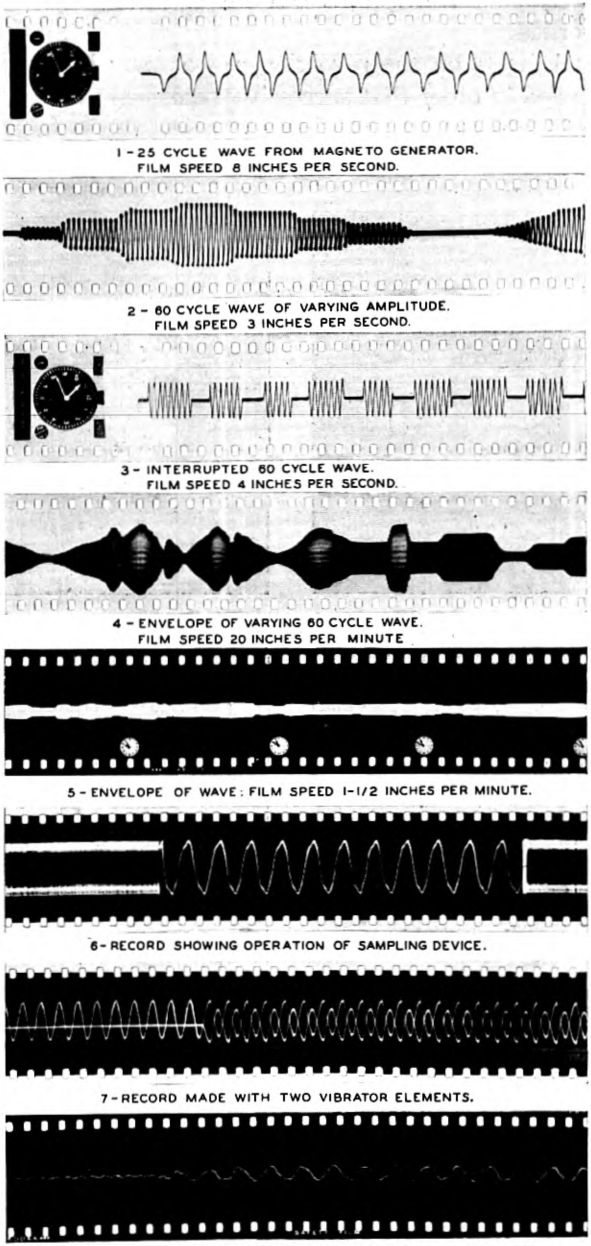


Fig. 18—Sample of records made by polar and continuous-film oscillographs used together.

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



1 - 25 CYCLE WAVE FROM MAGNETO GENERATOR.
FILM SPEED 8 INCHES PER SECOND.

2 - 60 CYCLE WAVE OF VARYING AMPLITUDE.
FILM SPEED 3 INCHES PER SECOND.

3 - INTERRUPTED 60 CYCLE WAVE.
FILM SPEED 4 INCHES PER SECOND.

4 - ENVELOPE OF VARYING 60 CYCLE WAVE.
FILM SPEED 20 INCHES PER MINUTE

5 - ENVELOPE OF WAVE: FILM SPEED 1-1/2 INCHES PER MINUTE.

6 - RECORD SHOWING OPERATION OF SAMPLING DEVICE.

7 - RECORD MADE WITH TWO VIBRATOR ELEMENTS.

8 - WAVE: OF 60 CYCLE AND 25 CYCLE COMPONENTS SUPERPOSED.

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 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.

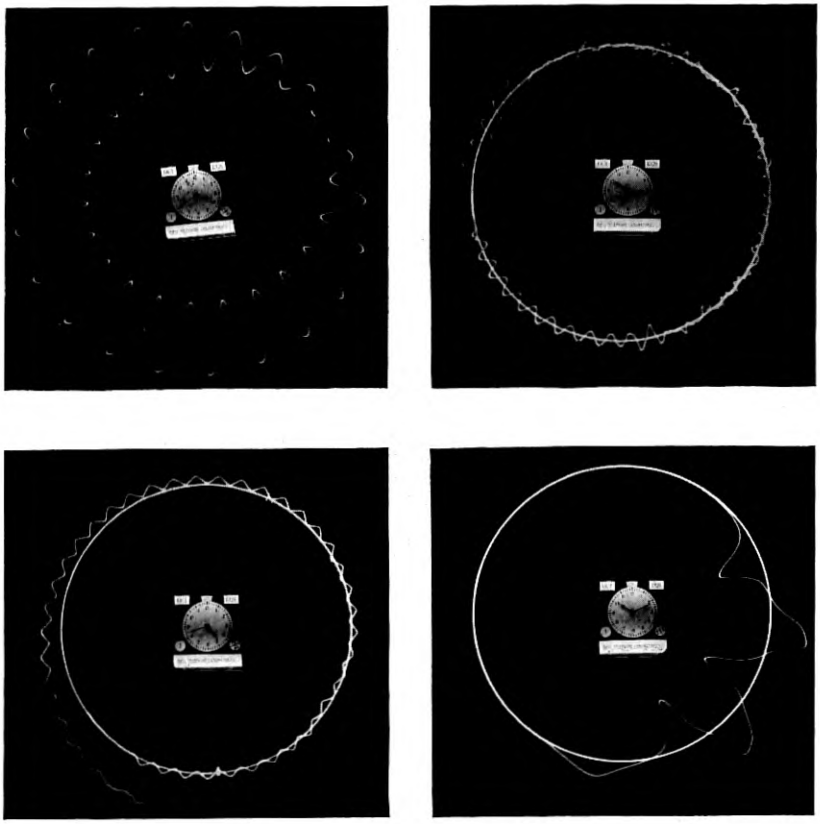


Fig. 20—Samples of polar oscillograms.

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.

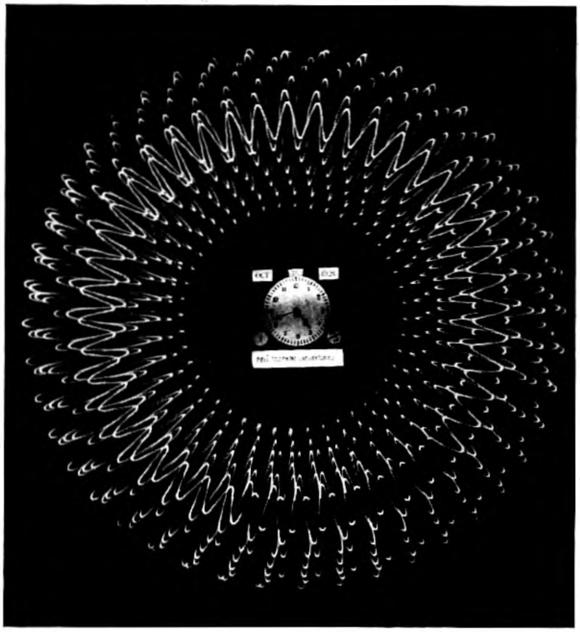


Fig. 21—Amplitude calibration of polar oscillograph.

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" arrangement 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 *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.

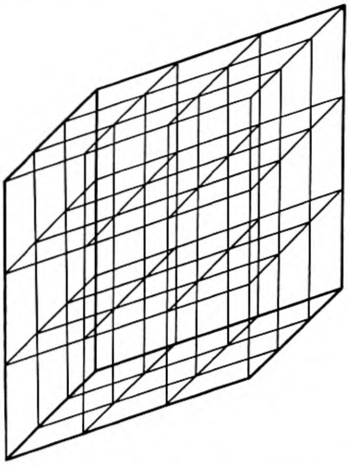


Fig. 1—A space-lattice. Intersections of three lines are lattice-points.

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.

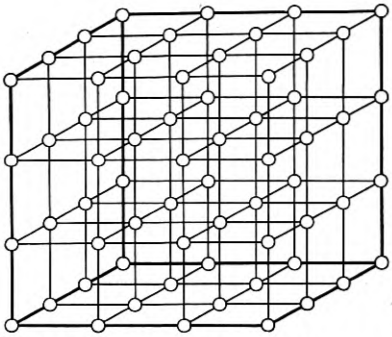


Fig. 2—A cubical space-lattice, with particles of full spherical symmetry indicated at the lattice-points.

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.

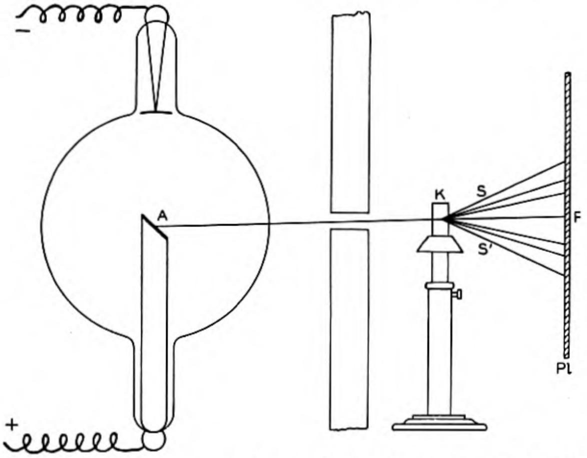


Fig. 3—Illustrating how a beam of waves is split by a crystal (at K) into diffraction beams forming spots on a plane screen (P). (Apparatus of the first experiment on X-rays by Friedrich and Knipping.)

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.

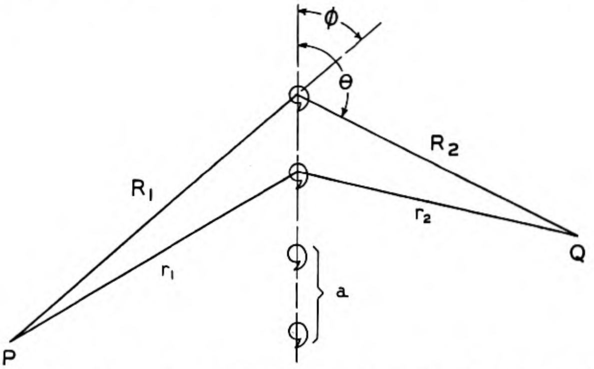


Fig. 4—Illustrating diffraction by a single file of atom-groups.

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]

$$(R_1 + R_2) - (r_1 + r_2) = (R_1 - r_1) + (R_2 - r_2) = n\lambda; \qquad n = 0, \pm 1, \pm 2, \ldots \tag{1}$$

Let P and Q recede to infinity; in the limit:

$$R_1 - r_1 = a \cos \phi, \qquad R_2 - r_2 = -a \cos \theta \tag{2}$$

and the condition for optimum reinforcement at Q is this:

$$a(\cos \theta - \cos \phi) = n\lambda. \tag{3}$$

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?

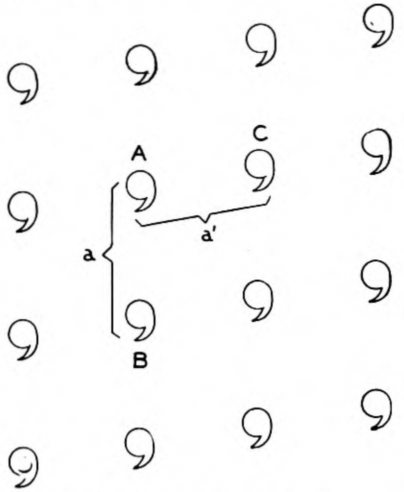


Fig. 5—Illustrating diffraction by a plane of atom-groups in regular array.

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 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.

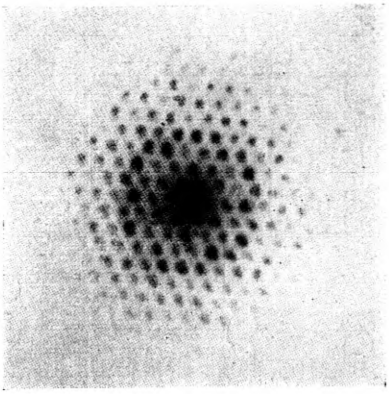


Fig. 6—Diffraction pattern of electron-waves attributed to the array of atom-groups in the superficial plane of a mica crystal. (S. Kikuchi; *Japanese Journal of Physics.*)

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).

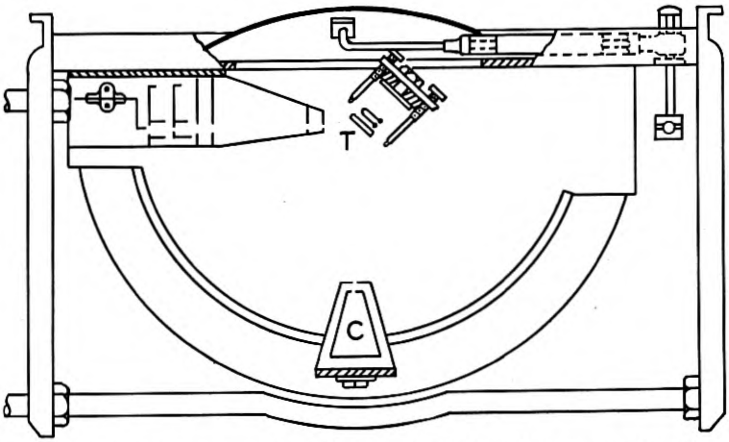


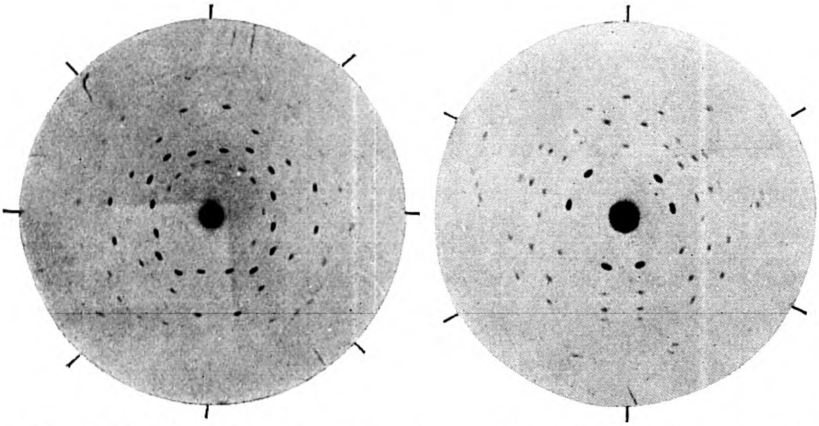
Fig. 7—Illustrating how a crystal (T) is mounted so that it may be oriented in various ways relatively to the oncoming beam of electron-waves (emerging from the electron-gun on the left) while a collector (C) is moved around to catch the diffraction beams. (Davisson and Germer.)

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 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.

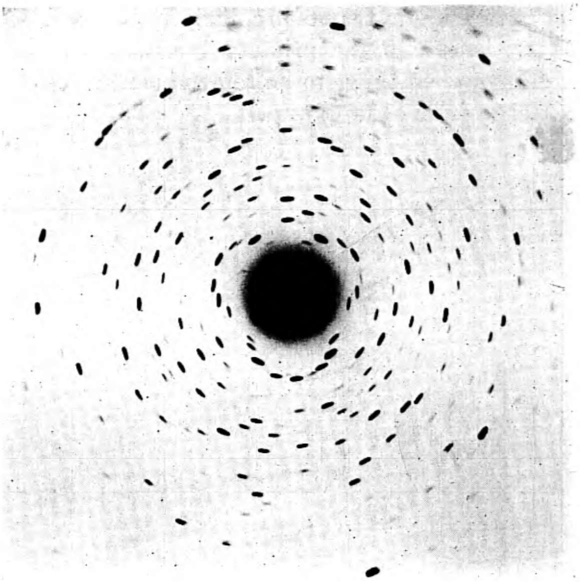


Fig. 10—Laue pattern of a mica crystal. (R. M. Bozorth.)

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?

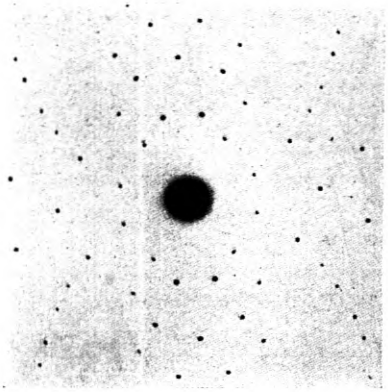


Fig. 11—Laue pattern of an iron crystal, showing that these crystals are built on a cubic lattice, though they are seldom so shaped as to reveal the fact. (G. L. Clark.)

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 cannoτ be correct, for they do not all have 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.

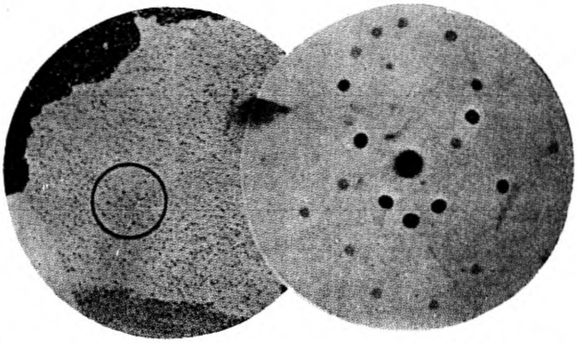


Fig. 12—Laue pattern of an aluminium crystal, introduced to show that if the primary beam does not happen to follow an axis of the crystal lattice there are still diffraction-spots though not so regularly arranged as in the previous cases. The crystal appeared under the microscope as an irregular patch on the metal surface (left-hand figure; the circle shows where the primary beam struck.) (Czochralski.)

(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.

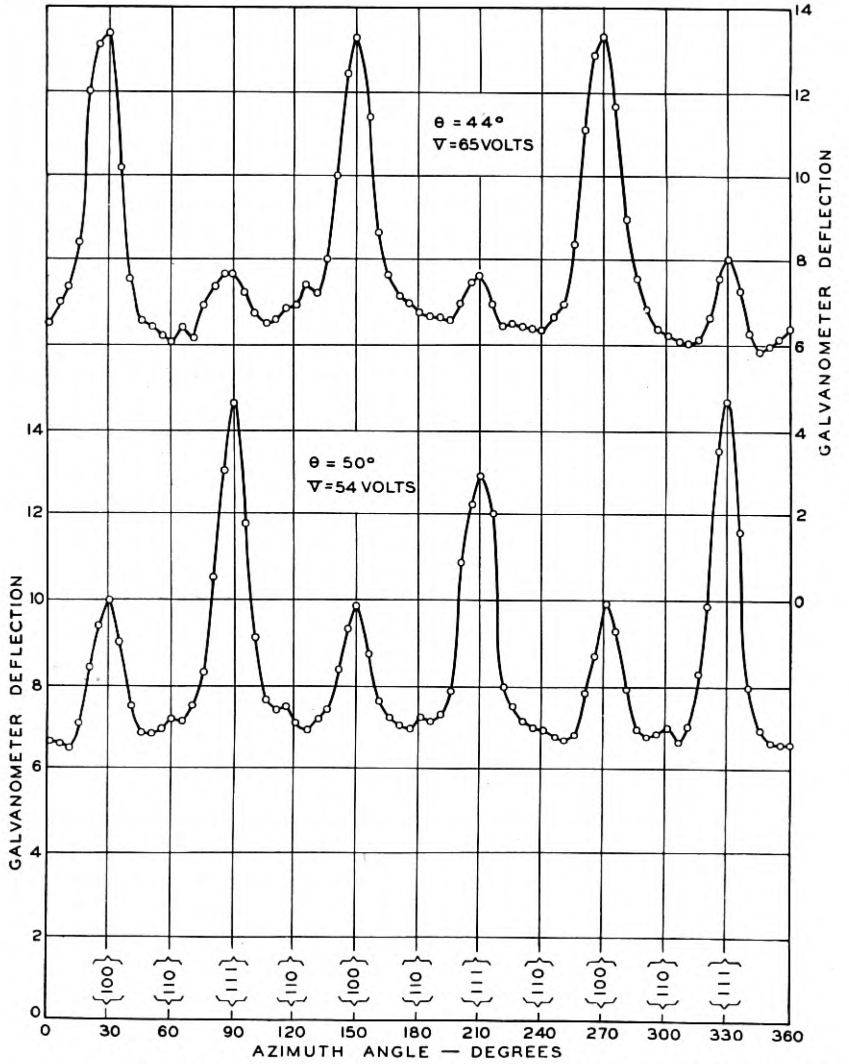


Fig. 13—Diffraction-peaks obtained with a fixed crystal and monochromatic electron waves (the two curves correspond to different wave-lengths chosen because they satisfy the condition for producing diffraction-beams) by moving a collector around so as to capture one beam after another. (Davisson and 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.

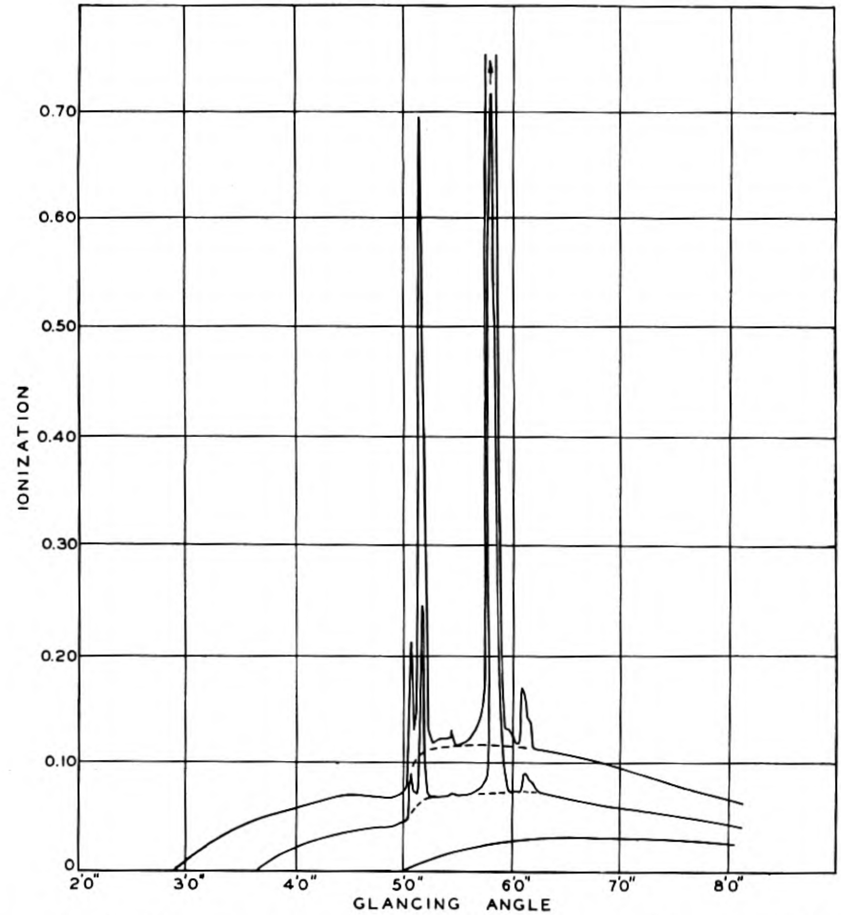


Fig. 14—Diffraction-peaks obtained with X-rays and a revolving crystal. The X-rays contained very intense waves of several different wave-lengths, and the collector was shifted continually so that it would capture a certain strong diffraction-beam produced by each of these in turn. (D. L. Webster.)

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 (6*d* 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 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.

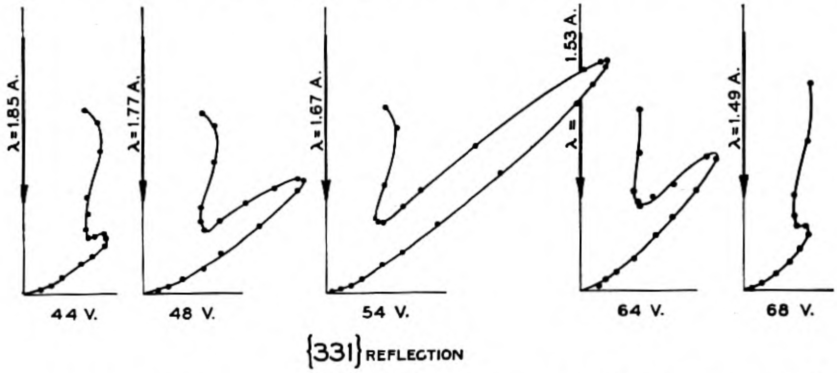


Fig. 15—Gradual appearance and disappearance of a diffraction beam as the mean wave-length of the primary waves passes through one of the values compatible with equations (6). (Davisson and Germer.)

In the foregoing passage it seems as ıf 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 (*6a*, *6b*, *6c*), 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 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:

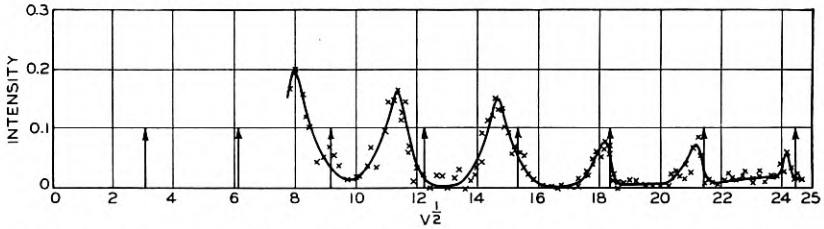


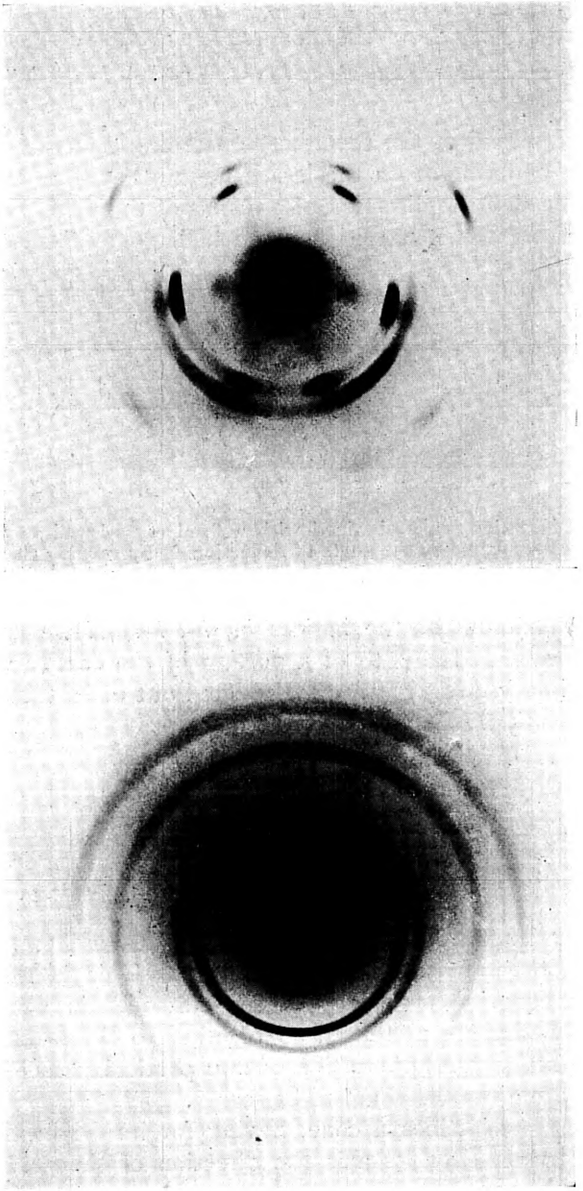
Fig. 16—Diffraction-peaks obtained with a fixed crystal and a fixed collector, *i.e.* with a constant value of the angle of deflection Φ, by varying the wave-length. (Davisson and Germer.)

$$1 - \cos\Phi = \frac{\lambda^2}{2a^2}(h_1^2 + h_2^2 + h_3^2), \qquad (8)$$

$$\sin\frac{1}{2}\Phi = \frac{\lambda}{2a}\sqrt{h_1^2 + h_2^2 + h_3^2}. \qquad (9)$$

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.

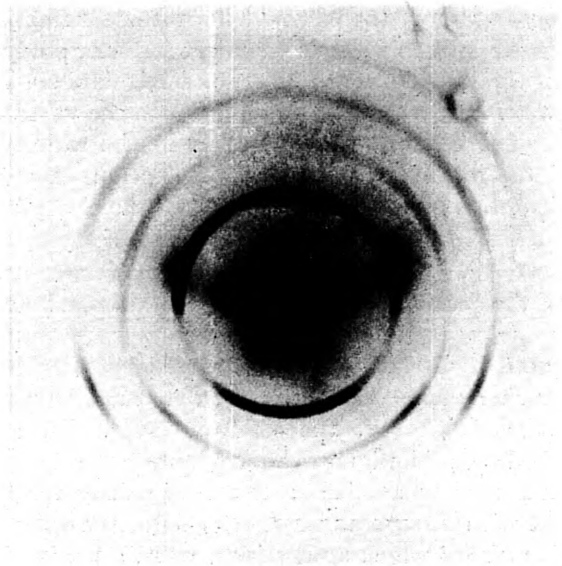


Fig. 19—"Powder method" diffraction-rings obtained with X-rays and a nickel-iron alloy. Like the alloys used in Figs. 17, 18, the crystals of this are built on a cubic lattice but with a differently-shaped atom-group, whence the changed appearance. (R. M. Bozorth.)

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

$$\lambda = (h/m_0v)\sqrt{1 - v^2/c^2}$$

instead of $\lambda = h/m_0v$ (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 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:

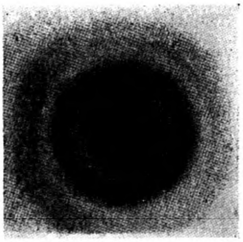


Fig. 20—"Powder method" diffraction-rings obtained with electron-waves and a thin film of gold. (G. P. Thomson, *Proc. Roy. Soc.*)

$$\theta = \tfrac{1}{2}\pi - \tfrac{1}{2}\Phi; \qquad \theta_0 = \tfrac{1}{2}\pi + \tfrac{1}{2}\Phi. \tag{10}$$

We wish to deduce the direction-cosines of the normal—denote them for the moment by the symbols $\gamma_1, \gamma_2, \gamma_3$—from the conditions that it makes the angles θ_0 and θ with the rays having the direction-cosines $\alpha_1, \alpha_2, \alpha_3$ and $\beta_1, \beta_2, \beta_3$ 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:

$$\gamma_1=\frac{h_1}{\sqrt{h_1^2+h_2^2+h_3^2}};\ \gamma_2=\frac{h_2}{\sqrt{h_1^2+h_2^2+h_3^2}};\ \gamma_3=\frac{h_3}{\sqrt{h_1^2+h_2^2+h_3^2}}\cdot \tag{13}$$

[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 $EFGII$. 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.

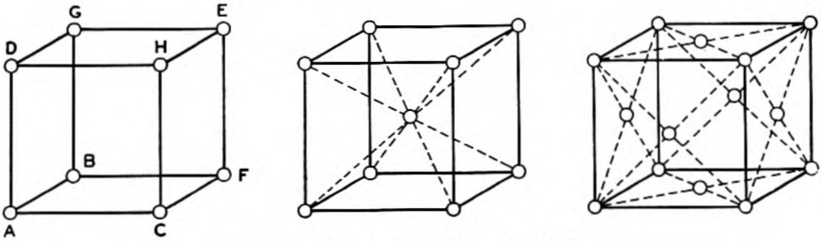


Fig. 21—Illustrating the cubic lattice.

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 *BCGH*. 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 *BCGH*. 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 answer 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 ol 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 $(\gamma_1l_1 + \gamma_2l_2 + \gamma_3l_3)a$; 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.