The Bell System Technical Journal

April, 1928

Joint Meeting of the Institution of Electrical Engineers and the American Institute of Electrical Engineers

ON February 16th last, a joint session of the Institution of Electrical Engineers in London and the American Institute of Electrical Engineers in New York was made possible by the transatlantic telephone. The audience in New York numbered over one thousand and that in London was several hundred. The two audiences were called to order at 10:30 A.M. New York time by Mr. Bancroft Gherardi, president of the American Institute of Electrical Engineers, and several papers were read both in New York and London to which the two audiences listened simultaneously. This joint meeting marks such an important milestone in the history of electrical communication that its entire proceedings are reproduced herewith. They are entirely self-explanatory.

Having called the meeting to order, Mr. Gherardi said:

"I will ask Mr. Charlesworth, Chairman of our Meetings and Papers Committee, to say a few words concerning the London meeting, and then to arrange for our joint session."

MR. CHARLESWORTH: Before proceeding with the joint session with our associates in the British Institution of Electrical Engineers, I wish to say just a few words concerning their London meeting in order to help you visualize the nature and significance of our joint session.

Our British associates have assembled in the auditorium of the Institution of Electrical Engineers Building located on the Victoria Embankment. The time is about 3:30 in the afternoon. Their meeting includes their President, Archibald Page, Chief Engineer of Central Electricity Board, their Vice President, Colonel Purves, Engineer and Chief of the British Post Office, the full Council of the Institution, members and invited guests from all parts of Great Britain, men prominent in all branches of the electrical industry.

Through the medium of developments which have been made in electrical communication, we are in effect to wipe out the great distance which separates the meeting places of our two societies, and to come together in a joint session in which our respective Presidents may exchange greetings in our behalf and in which other distinguished

representatives of our two societies may take part. By means of the loud speakers we shall be able to hear these proceedings as though we were all located in one great auditorium.

Colonel Purves has just finished making a statement to his associates concerning our meeting here in New York.

I will now speak to Colonel Lee who is at the telephone in London, and we will then proceed with our joint session.

Good morning, Colonel Lee.

COLONEL LEE: "Good afternoon, Mr. Charlesworth."

MR. CHARLESWORTH: "Are we ready to proceed with our joint session, Colonel Lee?"

COLONEL LEE: "We are, Mr. Charlesworth."

MR. CHARLESWORTH: "I will hand the telephone to Mr. Bancroft Gherardi, President of the American Institute of Electrical Engineers."

COLONEL LEE: "I will also hand the telephone to Mr. Archibald Page, President of the British Institution of Electrical Engineers."

MR. GHERARDI: Good morning, Mr. Page.

MR. PAGE: Good afternoon, Mr. Gherardi.

MR. GHERARDI: Mr. Page, it would give us great pleasure, if as President of the Institution of Electrical Engineers—the senior society, founded in 1871—you would act as chairman of this joint meeting.

CHAIRMAN PAGE: I regard it as a great honour to be asked to take the chair on this historic occasion. It is also a gracious compliment to our institution, and in accepting, which I do gladly, I desire to thank you, Mr. President, and the members of the American Institute of Electrical Engineers most heartily. I welcome all present at the meeting now in session, and venture to predict that the proceedings will prove exceedingly interesting and likely to live not only in our memories, but to be quoted by succeeding generations of electrical engineers as marking an important milestone in the advancement of electrical science. I am sure I interpret the desire of those assembled if I request Mr. Gherardi to address us, which I now do.

MR. GHERARDI: Mr. President and Members of the Institution of Electrical Engineers: On behalf of the American Institute of Electrical Engineers, I extend to you greetings and our best wishes. We are meeting here in New York at our Midwinter Convention. In the auditorium of the Engineering Societies Building in New York City, from which I am speaking, there are assembled about one thousand members of our organization from all parts of the United States, from Canada, and from other parts of the New World. It is with the greatest satisfaction that, as a result of the accumulated work of the scientist, the inventor, and the electrical engineer, it is possible for us

to hold this joint meeting—the first of its kind. It is with feelings of deep appreciation and respect that we think of the men who have exemplified the ideals of your organization—Faraday, Maxwell, Kelvin—and of the many others, past and present, who have contributed to Electrical Engineering and to the scientific foundations upon which it rests. These developments have been notable and have contributed in the greatest degree to the welfare of mankind. One of these developments is the art of electrical communication—the electric telegraph, and the telephone. These have made communication independent of transportation and no longer subject to all of its difficulties and delays. By the telephone, distance has not only been annihilated, but communication by means of the spoken word has become possible. Starting in 1876 with instruments and lines which, with difficulty, permitted communication over distances limited to a few miles, the telephone art has been improved year by year until continents have been spanned and, at last, even the limitations of the Atlantic Ocean have been overcome, and today telephone conversation between the two great capitals of the English-speaking world is a reality. We are gratified that transatlantic communication has made this meeting possible; it has added one more to the many ties existing between our two institutions and has added still another opportunity for friendly communication between us.

CHAIRMAN PAGE: Mr. Gherardi and gentlemen: Please regard me for the time being, not as chairman but rather as representing the thirteen thousand members of the Institution of Electrical Engineers. My first desire is to thank you, sir, for your most kind message of good will to us all. In turn we hail the President and members of the American Institute of Electrical Engineers with feelings of the utmost warmth and of everything included in the term good comradeship. The telephone must rank as one of the greatest inventions of the nineteenth century and it has transformed the daily life of all civilized people. Our indebtedness to Graham Bell for the boon he has conferred upon us increases with the years, and his memory, along with that of Franklin and Henry, will be cherished as becomes such benefactors of mankind. It would indeed be a gigantic task to attempt to exhaust the list of those of your society who have contributed so largely to the progress of electrical science and I must content myself by paying tribute to a great institution which has given proof time and again that engineering is truly international. It cannot be questioned that we are living in a period of extraordinary change due to scientific discovery, and in no field has the advance been more marked than in that of communication engineering. The commercial radio services

thus placed at our disposal assure closer and closer association between the English-speaking races, new spice is added to life and bonds of friendship materially strengthened. I rejoice that our two institutions can combine in the future even more effectively than in the past and that this is the outcome of the splendid work done in one of the branches of our own profession. I will now resume my chairmanship and call upon Dr. Jewett to speak.

DR. JEWETT: Mr. Chairman, Mr. Gherardi, and fellow members of the Institution of Electrical Engineers and of the American Institute of Electrical Engineers: The opportunity which this occasion offers of addressing jointly two widely separated groups of engineers whom, in times past, I have addressed vis-a-vis in London and New York, affords me the liveliest satisfaction.

I am gratified to participate in an event which marks both a notable advance in electrical communication and a pioneer demonstration of a wider use for electrical communication.

I am frankly pleased that, in common with numerous associates on both sides of the Atlantic, it has been my good fortune to play a part in the development work which has made this occasion possible.

Col. Purves and Mr. Gherardi will remember, and the rest of you will be interested to know, that in London more than a year ago, when we were engaged in final considerations preliminary to the opening of commercial transatlantic telephony, we discussed the details of just such a meeting as this. That our discussion should have been serious and not a pleasant mental diversion at a time when the channels of communications were not in operation is a striking evidence of the sound basis which underlies present-day electrical engineering. The fact that we saw and appraised the many obstacles to be overcome did not in the least diminish the assurance with which we talked of and planned for a distant event.

While therefore the present occasion is highly gratifying to the engineers whose work has made it possible, it is in no sense a surprise.

The success of this occasion is significant also in that it is the tangible evidence of a cooperation both intimate and full between men so situated as to make cooperation difficult. On behalf of my associates in America, I salute our associates in England.

CHAIRMAN PAGE: It is now Colonel Purves' turn to speak.

COLONEL PURVES: Mr. President, Mr. Gherardi, Dr. Jewett and gentlemen: It is an honour and a privilege to be associated with this notable event, which, one can justly feel, is breaking new ground in the advance of nations towards closer relationship. It is a great thing that two large assemblies, separated by wide expanses of ocean, can

join themselves together as we are doing now and interchange their thoughts and ideas by the simple and natural medium of direct speech to a combined audience. It opens up the prospects of results which thrill the imagination, and which are bound to be beneficent, and to conduce, by the way of clearer and mutual understanding, to the good of mankind. On this first occasion it is inevitable that the many professional interests which our two institutions share and which we should dearly like to talk over with each other should be pushed a little into the background, and that we should find ourselves preoccupied mainly with the wonder of the thing itself.

The radio art has given us its essential basic principles and the high power amplifying tubes, which over here we call valves. Long distance telephony has contributed a host of new devices which are equally essential. Specialized broadcast has given us the loud speaking receiver. As we sit and talk to each other our speech is launched into the air by the radio transmitting stations at Rugby and at Rocky Point with an electromagnetic wave energy of more than two hundred horsepower, and, I may add, the combined effect of the various refinements and special devices included in the transmitting and receiving systems is to make the speech efficiency of each unit of this power many thousands of times greater than that of an equivalent amount of power radiated by an ordinary broadcasting station. Many further improvements are being studied.

I should like to express the feelings of great personal pleasure with which I am listening to the voices of my old and valued friends of the American Telephone and Telegraph Company, Mr. Gherardi, Dr. Jewett and General Carty, and to assure them and their colleagues, both on my own behalf and on behalf of the engineering staff of the British Post Office, that the increased opportunities of cooperation with them which the development of the transatlantic telephone system has afforded us, are appreciated in a very high degree. We have to thank them for much helpful counsel in this and in many other matters and we look forward with great pleasure to a continuance of our close association with them on the long road forward, over which we still have to travel together.

CHAIRMAN PAGE: We are delighted to have with us in New York General John J. Carty, Vice President of the American Telephone and Telegraph Company and Past President of the American Institute of Electrical Engineers. It gives me great pleasure indeed to have this opportunity to congratulate General Carty on the presentation which he received last evening of the John Fritz Medal. This was presented to him by the National Engineers Societies of the United States for

his outstanding achievements in the engineering field. General Carty is widely regarded as the *doyen* or, to be more correct, the dean of the telephone engineering profession, and we shall be glad if he will say a few words and propose a resolution on the subject of our joint meeting.

General Carty spoke for a moment and then offered the following resolution.

WHEREAS on this 16th day of February, 1928, the members of the Institution of Electrical Engineers assembled in London, and the members of the American Institute of Electrical Engineers assembled in New York, have held, through the instrumentality of the trans-atlantic telephone, a joint meeting at which those in attendance in both cities were able to participate in the proceedings and hear all that was said, although the two gatherings were separated by the Atlantic Ocean; and as this meeting, the first of its kind, has been rendered possible by engineering developments in the application of electricity to communication by telephone; therefore,

Be it resolved that this meeting wishes to express its feelings of deep satisfaction that, by the electrical transmission of the spoken word, these two national societies have been brought together in this new form of international assembly, which should prove to be a powerful agency in the increase of good will and understanding among the nations; and

Be it further resolved that a record of this epoch-making event be inscribed in the minutes of each society.

CHAIRMAN PAGE: Sir Oliver Lodge, who needs no introduction, is sitting beside me and I have asked him to second the motion.

SIR OLIVER LODGE: Mr. Chairman, I think it very kind of you and the Council to allow me to take part in this important occasion, to send greetings to our many American friends. It is surely right and fitting that a record of the transmission of human speech across the Atlantic be placed upon the minutes of those societies whose members have been most instrumental in making such an achievement possible, and I second the proposal that has just been made from America. All those who in any degree have contributed to such a result from Maxwell and Hertz downwards, including all past members of the old British society of telegraph engineers, will rejoice at this further development of the power of long distance communication. Many causes have contributed to make it possible; that speech is transmissible at all is due to the invention of the telephone. That speech can be transmitted by ether waves is due to the invention of the valve and the harnessing of electrons for that purpose. That ether waves are constrained by the atmosphere to follow the curvature of the earth's surface is an unexpected bonus on the part of Providence, such as is sometimes vouchsafed in furtherance of human effort.

The actual achievement of today, at which we rejoice and which posterity will utilize, must be credited to the enthusiastic cooperation owing to the scientific and engineering skill of many workers in the background whose names are not familiar to the public as well as to those who are well known. The union and permanent friendliness of all branches of the English-speaking race, now let us hope more firmly established than ever, is an asset of incalculable value to the whole of humanity. Let no words of hostility be ever spoken.

CHAIRMAN PAGE: Gentlemen, you have heard the motion proposed by General Carty and seconded by Sir Oliver Lodge. I now put it to the joint meeting. Those in favor. Contrary. Carried unanimously. I suggest, Mr. Gherardi, that we now adjourn the meeting. I feel that it has been eminently successful and that we should regard it as the forerunner of many more to come.

PRESIDENT GHERARDI: Mr. Page, before we adjourn, I should like to take this opportunity to thank you for the gracious manner in which you have acted as chairman of this meeting, the first of its kind that ever has been held. We on this side send you our goodbye greetings and consent to the adjournment of the meeting.

CHAIRMAN PAGE: That is all the business, gentlemen. The meeting is adjourned. Goodbye.

Transatlantic Telephony—the Technical Problem

By O. B. BLACKWELL

SYNOPSIS: This paper, which, as it was read, was prefatory to the joint meeting, describes in rather non-technical terms the engineering problems involved in developing the transatlantic radio trunk by means of which the American telephone system of some 18,000,000 stations can communicate with the English telephone system of about 1,500,000 telephones, and also with the telephone systems of other European countries.

WE wish to give you a picture, necessarily very briefly sketched, of the physical makeup of the transoceanic telephone circuit, why it has been given its present form, and what further improvements are expected as the result of development work now under way.

The problem in brief is suggested by Fig. 1. A telephone system in America of some 18,000,000 stations, and distances of upwards of 3,000

Fig. 1—Map showing U. S. and European telephone systems separated by ocean

miles. A telephone system in England of about 1,500,000 telephones and the possibilities already partly realized of wire extensions to the other European nations. Three thousand miles of ocean between these two systems.

The establishment of a connection across the ocean presented two problems. First, the problem of setting up the radio circuit between the United States and England and second, the problem of making this radio circuit function as a link between these two widely extended telephone systems.

Fig. 2 shows the geographical layout of the long wave transoceanic circuit. The course followed by the currents in a connection is as follows: voice currents originating at any substation in America are

transmitted to New York City over the wire circuits in the usual way and thence also by wire to the sending station at Rocky Point, Long Island, where they are radiated into space. These waves are picked

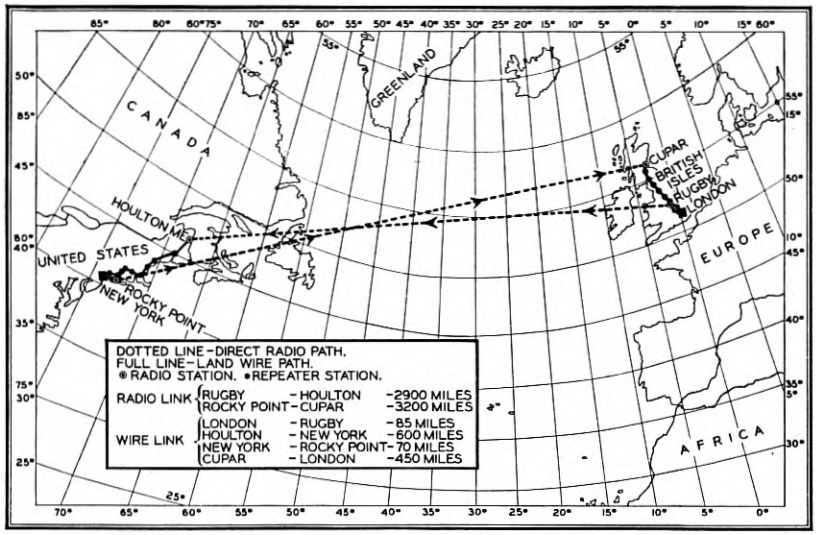


Fig. 2—Showing long wave routes

up at Cupar, Scotland, and transmitted by wires to London from which point they go by the usual wire connections to the subscriber in England or on the continent.

The answering voice waves are transmitted from the European subscriber by wire to London and thence by wire to Rugby, England, at which point they are radiated into space. The waves are picked up at Houlton in the northern part of Maine and transmitted by wires to New York City and thence by wires to the American subscriber.

You will note that the east and westbound radio systems are entirely separate from each other. Please note also that the receiving points in both countries are carried as far north as convenient—to Houlton, Maine, in this country, and to Cupar, Scotland, in the British Isles.

The radio and wire plant in Great Britain is owned and operated by the Post Office Department of the British Government.

As a supplement to the long waves there is a short wave circuit being formed which is so far only partially in use. In Fig. 3 the heavy lines show the long wave and the lighter lines the short wave routing. The short wave circuit from the United States to London was employed as an emergency routing during the severe static season last

summer extending from Deal Beach, N. J., to New Southgate, London. The British Post Office in January started sending on a return short

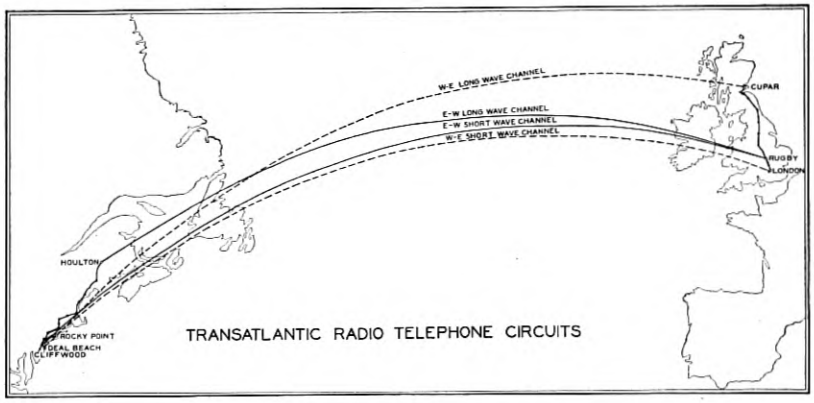


Fig. 3—Showing both long and short wave routes

wave circuit from Rugby, England, which is now being received at Cliffwood, N. J., but is not yet ready for service.

The right-hand drawing of Fig. 4 shows, necessarily on a logarithmic scale, approximately the frequency ranges covered by radio as we now know it. At the lower end are the long waves used in long distance

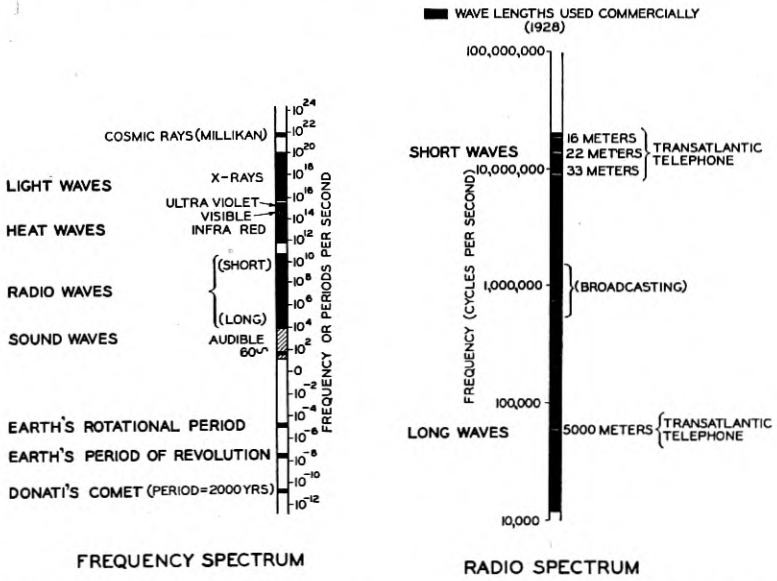


Fig. 4—Showing two plots on a logarithmic scale, one the radio frequency range, the other a general plot of frequencies

telegraphy extending down to nearly 10,000 cycles. At the upper end are the short waves already more or less exploited extending to about 10 meters, that is, 30,000,000 cycles.

It is interesting to note that one frequency range around 60,000 lying near the lower end of the scale and another frequency range extending from about 10,000,000 to 20,000,000 near the upper end of the present scale appear to be the most suitable for transoceanic transmission.

The left-hand figure has no bearing on our present subject. It is, however, rather interesting. It shows the whole gamut of frequencies with which we are familiar. This plot has near its lower end a frequency of one cycle per 2,000 years which is supposed to characterize a particular comet. From this it proceeds through the frequencies corresponding to solar periodicities, through the frequencies used in commercial power systems, through the voice frequencies, the wire carrier, the radio frequencies, the longer heat waves, the visual light rays, ultra-violet rays, X-rays and to the very hard rays sometimes called cosmic rays with which Dr. Millikin's name is here associated because of the investigations which he has carried out regarding them. This whole matter of frequency range and the relation of each part to human needs is of the greatest significance and interest.

In considering now how these long and short radio waves are handled in forming the transoceanic circuit, we will look first at the transmitting stations and antennæ, next at what happens to these waves in space and then at the receiving antennæ and stations.

At the transmitting end Fig. 5 shows a picture of the long wave antenna at Rocky Point, which well suggests the characteristics of these long waves. A frequency around 60,000 cycles corresponds to a wave length of about three miles and needs these physically large structures to effectively radiate the power into space. This antenna has six towers each 400 feet high. These long waves are in a frequency range which is much used and relatively narrow so that it is essential that the frequencies be employed the most economical way possible. This has resulted in the employment for the long waves of what is known as a single side band carrier suppression method of transmission, a refinement of transmission which, so far as I know, is not employed anywhere else in radio services although it is employed to a large extent in carrier over our wire circuits. With ordinary radio telephone transmission such as is used in broadcasting there is a constant steady frequency emitted even when no speech is being sent out and somewhat over 2/3 of the total energy radiated is in this steady carrier frequency which, of course, transmits no message. In the

system here employed, however, this steady frequency is practically eliminated so that when there is no speech to be transmitted practically no energy whatever leaves this antenna. Furthermore, in ordinary

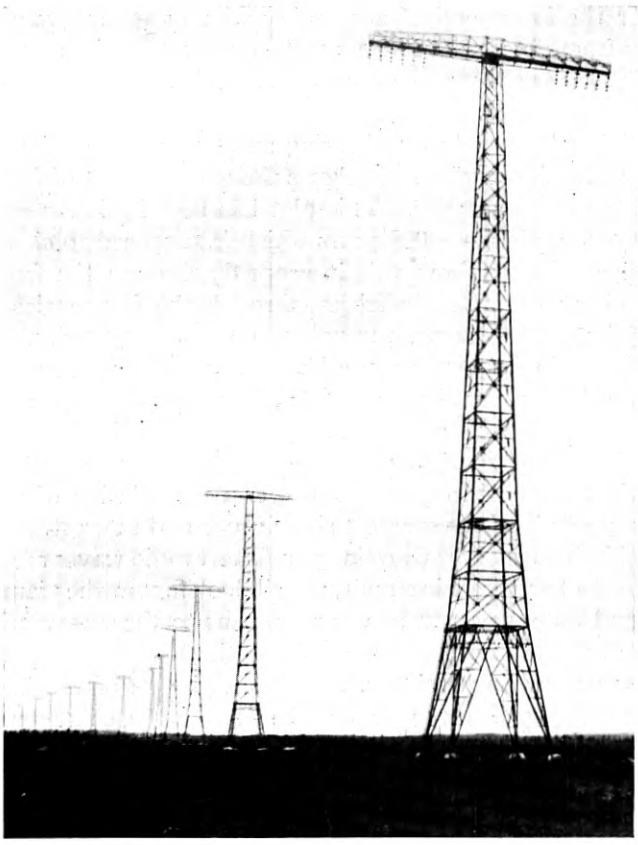


Fig. 5—A picture of Rocky Point sending antenna

transmission when there is, say, a 1,000-cycle tone to the voice, this appears in the transmission as two frequencies 1,000 cycles above and 1,000 cycles below the carrier frequency. This evidently is an ineffective use of the available frequencies so that one of the so-called frequency side bands in this system is eliminated. In this way a single tone in the voice is represented by a single frequency in the transmission from the station. A further frequency economy by the use of the same frequency range for talking in two directions is brought out below.

Fig. 6 shows the large vacuum tubes used in the last stage of the Rocky Point long wave radio transmitter. As many as 35 such tubes are employed in parallel capable of putting into the antenna something

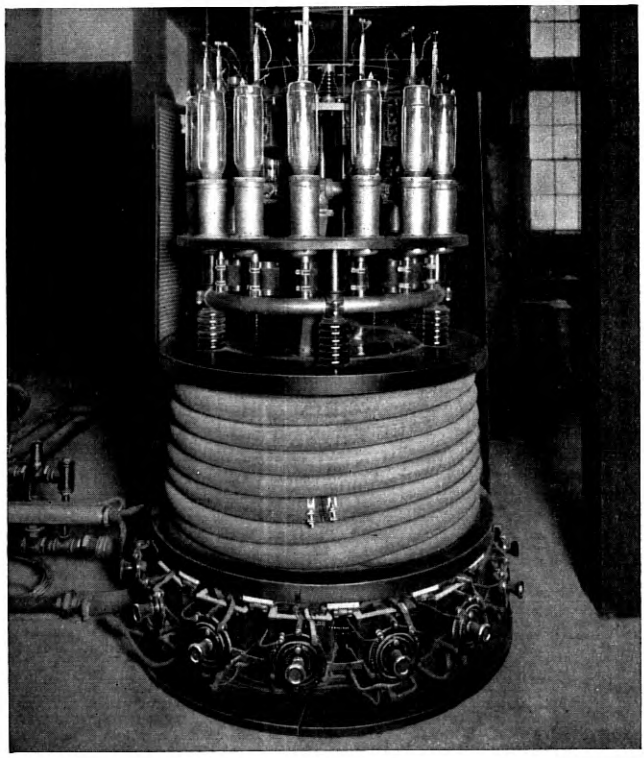


Fig. 6—Showing last power stage

over 200 killowatts which, as already noted, is worth something over six times this amount in the form of ordinary radio transmission.

This is the only picture I shall show of any of the apparatus involved in this work although such apparatus represents, of course, a tremendous amount of fundamental investigation and development and design. In a picture of apparatus, however, you can see little but assembly of cases and wiring and occasional vacuum tubes, and this gives you no adequate idea of what is going on electrically inside of the devices. An interesting feature of the transmitting apparatus in this system is that in the process of suppressing the carrier and stripping off one of the side bands, a double frequency transformation is required. For example, a 1,000-cycle tone coming into the station appears first

as 32,200 cycles and next as 59,500 cycles, which is the frequency transmitted.

Fig. 7 gives us in contrast one of the short wave sending antennæ. This particular one is for a wave of 22 meters wave-length, that is,

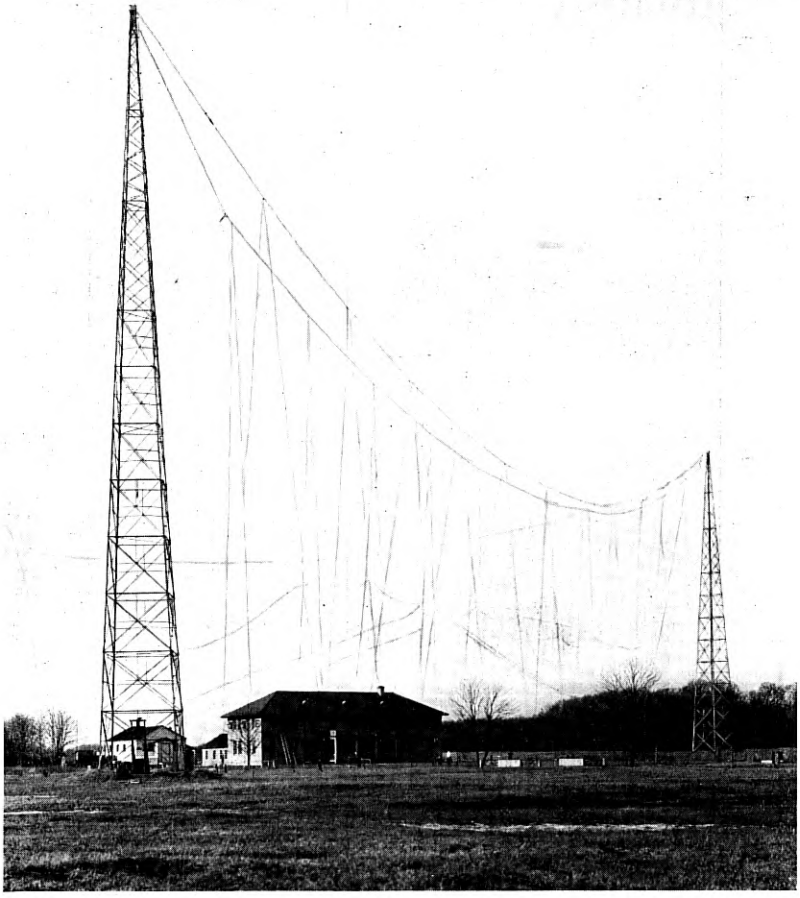


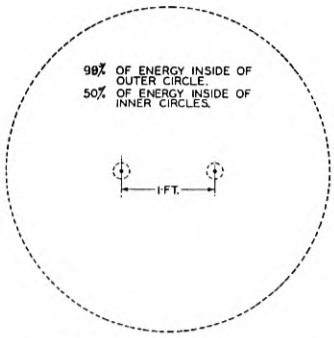
Fig. 7—Deal Beach sending antennæ

about 70 feet, corresponding to a frequency of around 14,000,000 cycles. These shorter wave-lengths lend themselves readily to directional sending and the resultant possibilities of conserving power in that way. The form of antenna shown is essentially a group of vertical antennæ with a transmission line connecting at points of equal phase. With this particular antenna, energy received in England is increased by

about ten times as compared to that received from a non-directive system.

Assuming we have put the power into proper frequency form and radiated it into space, the next question is how does it fare in traversing the great distances before it reaches the receiving points. We can hardly state this as a technical problem since there is nothing the engineer can do to control it. He can find out merely what nature does to such waves and try to arrange his transmitting, and particularly his receiving systems to meet the characteristics of the waves.

Section of New York–Chicago cable



Energy distribution with No. 8 B. W. G. open line circuit carrying A. C. currents

Fig. 8—Cross-sections of wire line with energy circles and cross-section of cable

We could think of no slide to show this space transmission unless possibly a picture of the world rotating in space such as we have seen adorning popular articles on radio. We will suggest the matter, however, by contrast with wire transmission.

The left-hand part of Fig. 8 shows a cross-section of a pair of copper wires spaced a foot apart on a pole line, which is the standard telephone wire arrangement. On such a circuit 98 per cent of the energy is transmitted inside of the outer dotted circle. The right-hand part of the slide shows two views (one a cross-section) of a typical telephone toll cable of somewhat under three inches diameter. Practically all

the energy for about 300 telephone circuits is transmitted inside this sheath.

While both the radio and the wire transmission involve similar electromagnetic waves, there could hardly be a greater contrast in the method of handling waves than that between the radio transmission we are considering in this paper and transmission employing such wire methods and spanning the comparable distance of say San Francisco to New York.

Recently in visiting the short wave receiving station in New Jersey I was shown oscillographs taken on radio telegraph transmission in which each telegraph dot was followed about a tenth of a second later by what appeared to be an echo. The first transmission came some 3,000 miles from England. The second transmission had gone the opposite direction around the world and had travelled some 22,000 miles before reaching the same receiving point. In such long distance radio we may then have a situation in which each individual signal sets up oscillations, perhaps measurable oscillations, in space surrounding practically the whole earth.

Contrast this to the toll cable shown in the picture which, as already noted, contains about 300 circuits. Such cable is used now commercially for distances up to around 1,500 miles and is permissible for 3,000-mile distances such as we are here considering. In such cables each message is practically confined to a strip of space extending between the terminals of the cable and smaller around than a lead pencil. In the radio case we have literally all out of doors but there is little we can do to control it. In the cable case the channel is reduced to the meagerest dimensions but this so reduced space we can pretty nearly call our own, surrounded and shielded as it is by a sheath and containing carefully balanced circuits. Such space is only occasionally penetrated by outside disturbances when some of our power friends set up unusually strong electrical fields in its immediate neighborhood. By loading, by amplifiers, by equalization of various sorts, this meager space is guarded and controlled and rendered efficient and constant.

We are still somewhat in the dark as to how kind nature has been to us regarding short waves and what degree of reliability we can ultimately get from a circuit employing them. There is nothing yet in the picture, however, to suggest a reliability for long or short wave radio approaching that of a cable circuit for similar distances.

A large number of measurements have been made of the strength of the radio fields laid down in England from the long and short wave stations in America and similarly in America from the English stations. Along with such data is taken the amount of noise interference present

at the receiving points. Fig. 9 shows data taken in this way. The curve which reaches the highest point shows for a typical summer's day the field strengths received on the long wave from America. The

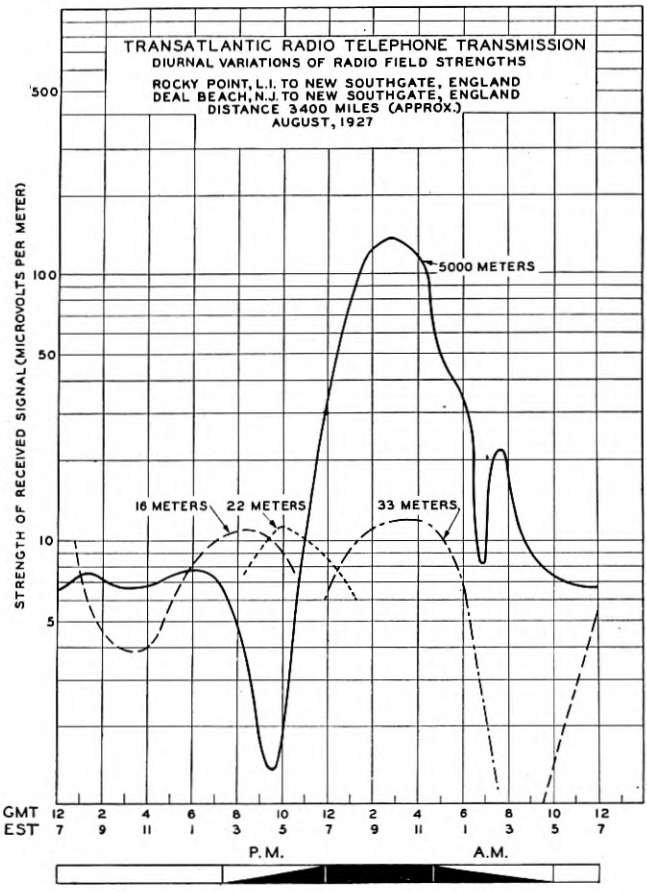


Fig. 9—Long wave curve and corresponding short wave curve of received field strength

other curve shows the field strengths received on short waves employing three wave-lengths, 16, 22 and 38 meters, and taking for each time of day the one of these which was best suited to the conditions.

On this typical day all the wave-lengths were operating well. It will be noted, however, that there are times when the long wave is low and some one of the three short waves is more effective and other times when none of the three short waves are high but the long wave is effective. Furthermore, all of these waves vary a good deal so that

on certain days for hours shown operative on these curves one or more might be entirely out of service. This chart indicates the tremendous advantage in employing a number of separate wave-lengths varying a good deal in their characteristics and choosing at any one time that wave-length which is giving the best performance.

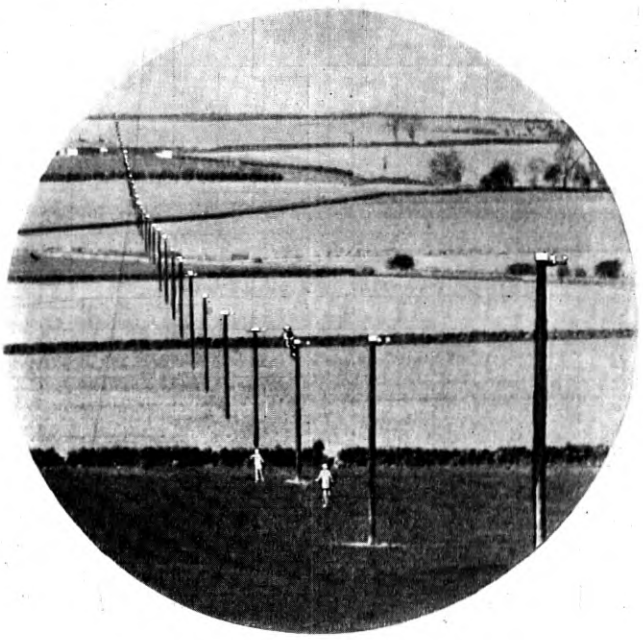


Fig. 10—Showing receiving antenna at Cupar

Having followed the radio transmission as the waves radiate into space and traverse space to the receiving end, we come to the matter of receiving stations. Fig. 10 shows a view of the wave antenna at Cupar, Scotland, used for receiving long waves. This complete antenna arrangement is made up of two pole lines such as shown, each about 3 miles long; a third may be added. These pole lines are placed parallel to each other with separations of about two miles. A pole line joins the two together and connects them to the receiving stations.

It is fortunate that in America (and the same is true to a considerable extent in England) the signals come in from a northerly direction and the static tends to come in from almost the directly opposite direction. The directivity brought about by such antennæ has, therefore, a very large effect. It is estimated that under average conditions the present

antenna gives as much improvement as would an increase in power of about 100 times. Furthermore, receiving at Houlton, Maine, rather than in the vicinity of New York, by getting to a more northerly latitude, is equivalent to a power increase of 50 times. The antenna location and directivity, therefore, is equivalent to a transmitted power increase of 5,000 times. The receiving set employed with these antennæ at Houlton is of a double demodulation type, of very high selectivity.

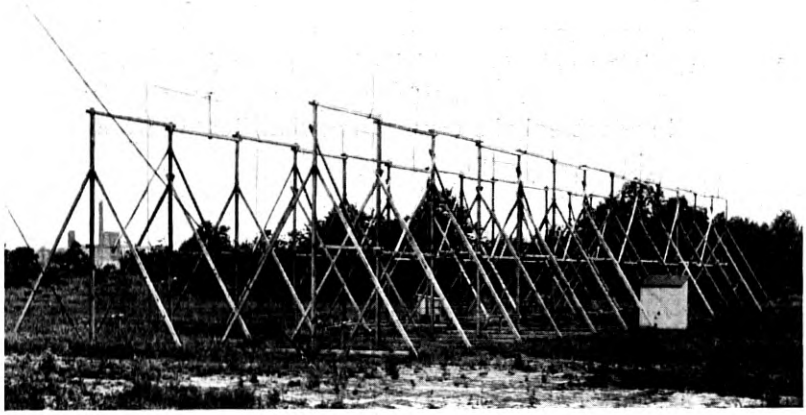


Fig. 11—Picture of receiving antenna in Cliffwood

Fig. 11 shows a receiving antenna at Cliffwood, N. J., of the type used for short waves. It is constructed of a wooden framework on which are held two parallel sets of conductors made up by bolting together lengths of copper tubing. Possibly you can make out the form of the two sets of conductors. Each set consists of vertical elements a quarter of a wave-length high and a quarter of a wave-length between successive elements. The first element is connected to the second by a conductor connecting the upper ends of the elements, the second is connected to the third by a conductor connecting the lower ends and so on.

The whole effect gives a degree of directivity equivalent to an increase in power of about 15 times. This general question of short wave receiving is a fruitful field for investigation and for the ingenuity of the engineers. A large number of arrangements have been devised

and a good many of them tried. Considerable further work is under way.

At the time when the short waves are in trouble apparently either one of two conditions may obtain. In the first of these there may be considerable field received from the distant transmitting station but this field is made up of the result of transmission over several paths so that the waves received over the different paths react on each other, causing a rapidly fluctuating interference pattern somewhat similar to that well known with light waves.

There are other times, however, when either the field which reaches the receiving point is not of sufficient magnitude to be picked up or is below the static noise level at the particular time.

One interesting fact with these very short waves is that the ignition system in automobiles may create large disturbances and it is important that the receiving points should be kept away from roadways frequented by automobiles. For this and other reasons the New Jersey location will undoubtedly be moved somewhat from Cliffwood.

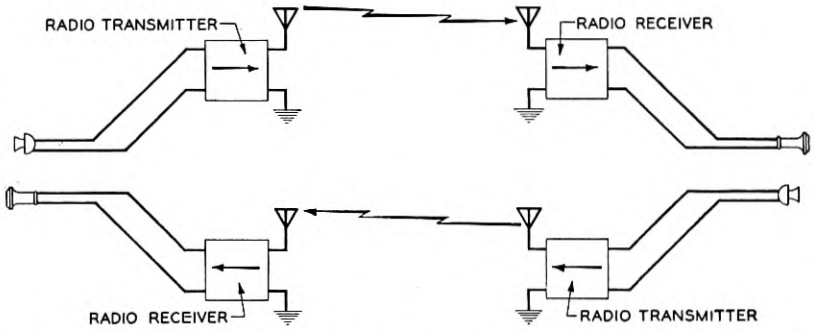


Fig. 12—Indicating diagrammatically east and west transmission on separate channels terminated in ordinary transmitters and receivers

So much then for the question of the radio circuits themselves. The problem remains of making them serve as a link between the two-wire systems on the two sides of the Atlantic. If the problem were merely transmission from one particular subscriber's set to another particular subscriber's set, the very simple arrangement shown in Fig. 12 would be feasible. In this you will note the eastbound and westbound transmission each starting with a telephone transmitter at one end and extending to a telephone receiver at the other are kept entirely separate. Two people could evidently carry on a conversation over this circuit without further complications. In fact this was the way in which the first two-way tests were carried out.

Since eastbound and westbound short wave channels are at entirely different frequencies, there would be no interaction between the east and west-going circuits when used in this way. For the long waves, however, since the east and westbound circuits are at the same frequencies, each transmitting station would send considerable energy into the receiving station on the same side of the ocean, thus giving to each subscriber a heavy side tone of his own speech. This effect, however, could be considerably reduced by separating the transmitting and receiving stations and arranging the directive antennæ of each receiving station so it would receive as little as possible of the corresponding transmitting station. Even with the long waves, therefore, a conversation could be held in this simple way.

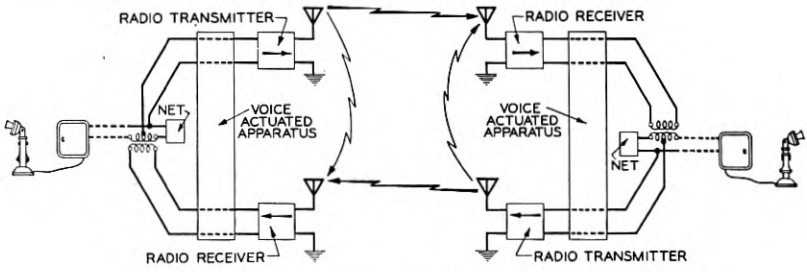


Fig. 13—Transoceanic circuit schematic

When the east and westbound radio circuits are brought together, however, for a connection to a wire circuit at each end, we have introduced some very serious difficulties. Consider first the simpler case of the short waves where the eastbound and westbound transmission are at different frequencies.

The voice waves reaching London from an American talker will be reflected in part either at the London office where the east and westbound channels are brought together or at some point before reaching the European subscriber. Unless means are taken to prevent it, this reflected energy can then pass to the English transmitting station and be transmitted back to America. At the American end a similar partial reflection can take place throwing part of the energy back again to England. In this way, according to circuit conditions, it is possible for the whole circuit either to build up and act as a widely flung oscillator or if the damping at the moment makes this impossible, the speaker and the listener can be much interfered with by electrical echoes and distortion.

In the case of long waves there is the added difficulty as already noted that each transmitting station can throw a good deal of power into the

receiving station on its own side of the ocean, thus bringing in the possibility of local oscillations and distortions.

For either the long or short waves then it is very advantageous, in fact practically necessary, to employ switching devices actuated by the voice waves themselves to prevent the effects just stated. In this diagram you will note drawn so as to involve the wire connections both to the transmitting and the receiving station at each end a rectangle which is labeled "voice-actuated apparatus." This is so arranged that when there is no transmission on the circuit the wires connecting to each of the transmitting stations are short-circuited, making both transmitting stations inactive. Speech coming in then at one of the ends from a telephone subscriber operates a relay which opens the wire circuit to the transmitting station at his end and at the same time short-circuits the receiving path at the same end.

One interesting phase of this voice-operated switching mechanism is the employment of a delay circuit. At the New York terminal of the circuit, when the voice currents reach it from some distant subscriber and after these voice currents have actuated the switching

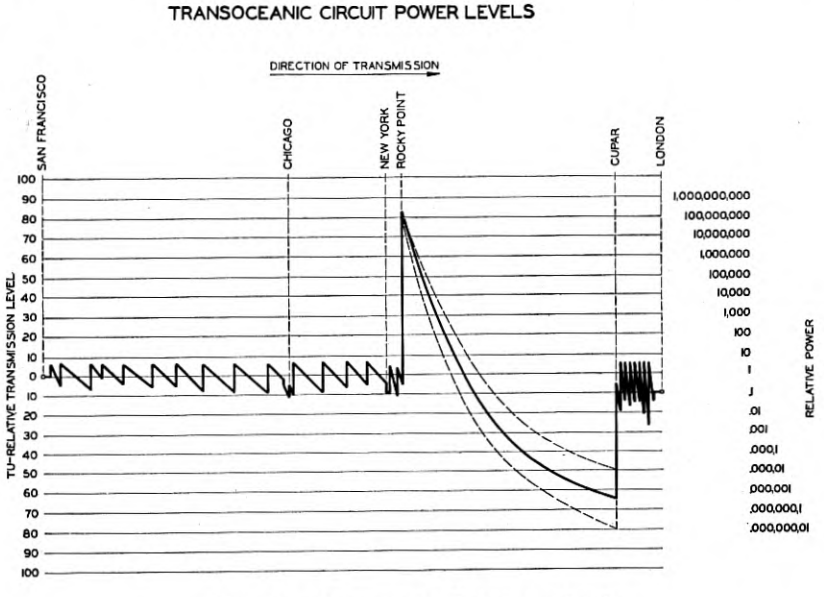


Fig. 14—Diagrams showing energy levels

mechanism noted above, they pass into an artificial line down which they travel, are reflected and travel back to the sending end of the artificial line. In this way the voice waves are allowed to idle away

two one-hundredths of a second during which time the switching mechanisms have performed the operations just noted and have thus put the circuit into shape for the voice waves to go forward.

Fig. 14 shows the shifts in power level in going from one end of the circuit to the other for a connection from San Francisco to London. The zero of the scale corresponds to the power level at which power is given out by an ordinary substation set when actuated by a loud voice. The power ratios compared to this are shown at the right on a logarithmic scale.

The comparatively small ups and downs in the power level corresponding to transmission over the wire lines represent, of course, the line attenuations and successive amplifications by telephone repeaters. The highest power level is naturally at the output of the radio transmitter where it is about one hundred million times the starting level At the English receiving station it has been dropped to about the same ratio below the zero of the scale.

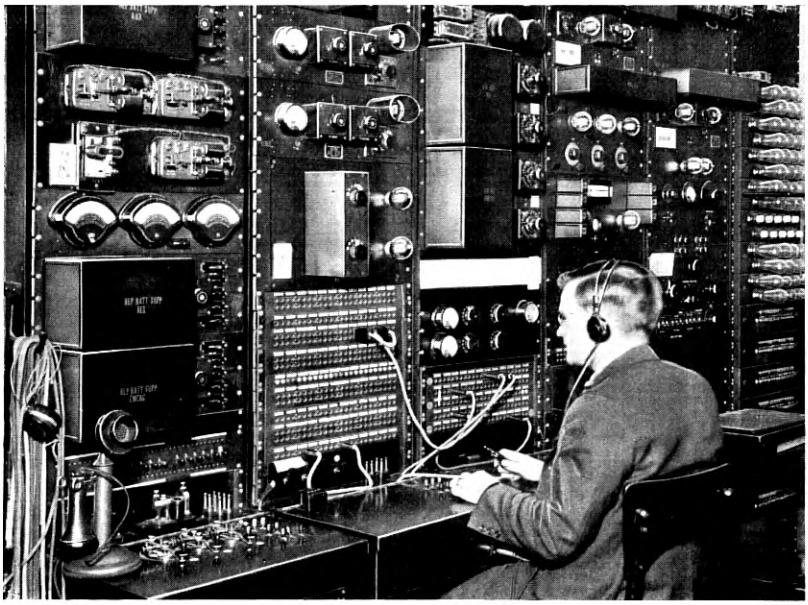


Fig. 15—Technical operators' position

In view of the character of radio, in particular its variable nature, the circuit is under constant supervision during operation by two technical operators, one located in New York and one in London. Fig. 15 shows the special terminal equipment, meters, etc., and the technical operator at the New York end.

It is evidently desirable that the transmitting station shall always put out maximum power whether the speaker has a loud or a weak voice or is near or distant from the transmitting station. One of the duties of the technical operator is to bring this about. By proper indicating meters he knows the power level of the speech going to his transmitting station and he keeps this at the point where it will just completely load the transmitting station.

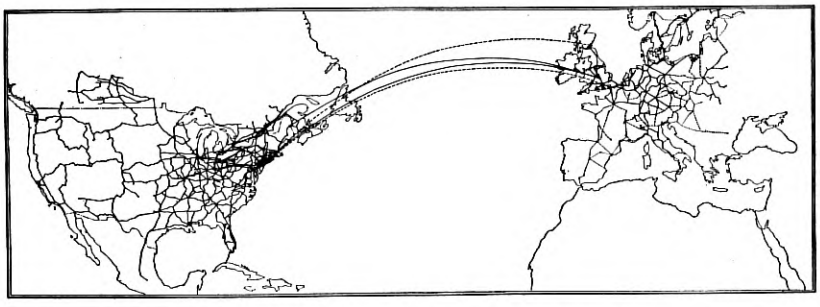


Fig. 16—Fig. 1 with radio connections added

With this brief discussion then let us assume we have progressed from the conditions of our first illustration to the conditions shown here in which the two great telephone systems are joined together by the radio links.

To complete the story, a few more words as to the changes and improvements which development work now under way is expected to give.

As the system stands to-day it does not offer the privacy which ordinary wire connections offer. While ordinary broadcasting sets are not of the proper type to receive messages from this system, it is comparatively easy, with sets designed for the purpose, to pick up one side of the conversation over the system by listening to the transmitting station in the same country as the listener. Because of the voice-actuated devices, the other side of the conversation has to be picked up directly from the distant country, which is a much more difficult thing to do.

To give this system a high degree of privacy is difficult, particularly with the long waves. The frequency range in which the long waves are situated is, as already stated, narrow and well filled so that any proposition that widens the required frequency range, such for example as shifting the carrier frequency rapidly, cannot be employed. Methods have been developed, however, and equipment is far advanced on a

system that is expected to give conversations over this radio channel a sufficient degree of privacy. Certain features of the new privacy method will probably be in experimental use within a few months. It will be at least six months and possibly a year before the complete privacy system is in full operation.

The feature of this whole transatlantic service which worries the engineer most, however, is the matter of reliability. After the engineer has done the best possible in transmitting and receiving stations he is confronted by the fact that transmission through space and noise conditions vary so much that thousands of times as much transmitting power as would be sufficient under good conditions may be inadequate to get through under poor conditions. His only defense is to use a considerable number of wave-lengths which tend not to get into difficulties at the same times or under the same conditions.

Considering first the short waves we find as already noted there is sufficient difference between the transmission characteristics of different parts of the range, between say 10 meters and 30 meters, so that the reliability is considerably improved by designing the stations to use any one of three or more frequencies in this range.

We shall be very happy if by such use of a number of short wave-lengths and by further improvements in technique the reliability of short wave channels can be made such as to some day eliminate altogether the necessity of the long wave channel with its much more extensive plant. There are a number of projects going ahead in the world for the establishment of long distance transoceanic telephone circuits employing short waves alone. With a reasonable further development of the short wave art such service will undoubtedly prove well worth giving.

Telephone service is, however, necessarily an exacting service, particularly since the subscribers participate directly in each connection. Moreover we are dealing here with the joining of North America and Europe which, commercially and otherwise, is of so large importance as to justify perhaps much more exacting technical requirements than any other transoceanic connection.

So far, the data available regarding the short waves do not suggest that they ever will give a reliability of service comparable to that for similar distances over land wire circuits. It is our present expectation, therefore, that the giving of suitable service between America and Europe will require the continuation of the long waves even though such waves demand a much more extensive and complicated plant than do the short waves. In addition to the long waves we shall also want the very best we can get from the short waves. By the combination

of this one long and several short waves we believe that ultimately the service, except for the three summer months of high static, will be but little interfered with by electrical weather, and that even for these months the service will be operative for better than 90 per cent of the time.

You will understand, of course, that the connection to-day from London will be entirely by long waves, the short waves not being ready for commercial operation. The connection from America to England will also in all probability be by long waves, although the short waves are being held in readiness as an emergency routing.

Transatlantic Telephony—Service and Operating Features

By K. W. WATERSON

SYNOPSIS: This paper describes some of the differences in operating practice on the two sides of the Atlantic and plans which were worked out for taking account of them in the handling of commercial transoceanic calls. Difference in the language is also another problem which has required solution. Data are included giving an idea as to the extent to which the transatlantic connection was used during its first year, there having been established during this time a total of something over 2,300 connections.

THE introduction of telephone communication between Great Britain and the United States required the fitting together of the practices of two telephone organizations. The development of usage between subscribers in the two countries involves questions of different telephone habits and experience. It may help to define the problem of setting up a service of this kind if at the start I mention one or two of the more important characteristics of long distance service in the two countries which illustrate outstanding points of difference.

In Great Britain, only number service is available, that is, a service under which the telephone administration undertakes merely to obtain a connection with a specified telephone and on which the message toll charge is assessed in all cases where an answer is obtained from the telephone called whether or not the person desired is there. In the United States, this same number service is available at about the same initial rates as in England. In addition, we have a so-called person-to-person service on which, for a charge approximately 25 per cent above that for number service at the longer hauls, we undertake to obtain connection with a particular person who is specified in the order for service. In case of inability to reach the particular person desired, the full message charge is not assessed, but a so-called report charge is made, which is about 25 per cent of the station charge, tapering off in percentage to a maximum charge of one dollar. This difference in the class of service available in the two countries is a matter of importance because of the fact that our experience here in America shows that on the longer hauls and at the higher rates about 85 per cent of the calls are on a person basis, whereas at short hauls the large majority of calls are for a number only.

In both countries the toll rates provide for an initial talking period of 3 minutes. In Great Britain, additional use of the line is charged

for on the basis of 3-minute units and for each the charge is the same as the initial rate. In the United States, additional use is charged for on a one-minute basis—each minute's charge being about one-third the initial—a finer measure of actual use and one which, particularly at the higher rates, makes long distance telephoning considerably less expensive.

In the United States, the general practice is to allow subscribers to talk as long as they wish except on rare occasions due to emergencies such as storm breaks. In Great Britain, subscribers are notified at the end of 3 minutes and are limited to a maximum use of 6 minutes if other calls are waiting. This difference in the allowable length of long distance telephone conversations has developed out of basic differences in toll service policy as regards the provision of plant and the resultant speed of service. In the United States, we plan to give a very rapid service and we provide toll line plant to meet these needs. This policy seems to best meet the needs and desires of American users and to have been a large factor in the rapid development of our toll business. As a result, practically all toll and long distance calls are placed at the time when the connection is wanted and subscribers are often impatient if their calls are not completed immediately. In Great Britain, the plan has been to maintain as high efficiency as practicable in toll line plant and this naturally results in a somewhat slower long distance service. British subscribers are accustomed to longer delays than ours in obtaining connections. There is considerable advance booking. Under this condition, the limitation of the talking period, which I have already mentioned, is a practicable means of making the service available to as many users as possible and also of avoiding possible cases of unfair use of the lines by certain individuals to the exclusion of others. This difference in practice regarding the allowable length of conversation is a matter of particular moment in connection with a service like the transatlantic service for the reason that our experience has shown a definite tendency for users to talk for longer periods on the long haul, higher rate business. This is probably indicative of the greater importance or different nature of this class of telephone communication. Whatever the reason, calls at 250 miles average under 5 minutes, at 500 miles—$5\frac{1}{2}$ minutes, at 1,000 miles—6 minutes, and transcontinental calls—$6\frac{1}{2}$ minutes.

In Great Britain, distances are relatively short. London-Glasgow, for example, represents one of the important longer haul routes, and the air-line distance is about 350 miles. On international calls, London to Berlin is one of the longer hauls at which service is available

and this is under 600 miles. In Great Britain, there is relatively small development of telephone usage at these distances. In the United States, on the other hand, we have transcontinental service over some 3,000 miles with considerable business at this and other long distances. Our connection with the Cuban Telephone Company has also given us experience with very long haul service. So in the matter of special long distance problems and in the development of long distance telephone usage, the experience has been largely on this side of the water.

The service arrangements for this transatlantic undertaking were made through discussions carried on in London by representatives of our organization and officials of the British Post Office. While there were a good many problems to be worked out, there was the usual result, when both parties desire to cooperate and to discover the best solution, that an agreement was soon reached. It was decided that the service needs of transatlantic telephony would best be met by a single class of service with one rate covering either number or particular person usage. Experience in the Bell System had indicated that on long haul business of this nature, practically all calls would be for a designated person and this has been borne out in the transatlantic usage. The rate between Great Britain and twelve states in the northeastern part of our country was fixed at $75 for 3 minutes, with an additional minute charge of $25. A report charge, of which I have already spoken, was fixed at $10 for use in certain cases where particular persons called cannot be reached.[1] The British Post Office preferred to apply the same rate to England, Wales and Scotland. Because of its wide expanse and expensive land line plant, the United States was divided into five zones for fixing additional land line charges over and above the New York terminal rate. These rate zone lines follow state lines. The zone rates go up in $3 steps as we draw away from New York. Zone rates follow reasonably well the land line charges for service from New York. This zoning plan was adopted to simplify the means of quoting and computing the transatlantic rates abroad as compared with superimposing the more finely measured land line rates on the New York terminal charge.

For communications extending beyond the initial 3-minute period, the plan of charging on a single-minute basis was adopted, as it seemed the most equitable, particularly in view of the distances and charges involved.

[1] "Reduced rates for transatlantic service became effective March 4th superseding those mentioned in this paper. To illustrate the extent of the reductions, a three minute call from New York to London which was initially $75. is now $45."

In setting up service and operating arrangements, it was necessary to give consideration to different conditions which would exist dependent upon the volume of business to be handled over the transatlantic channel. We had to consider operating practices which could be used satisfactorily either under conditions of high load and possibly delayed service or of light load and fast service. As a means of insuring service to as many users as possible in periods of heavy business, agreement was reached that there should be a 12-minute limit on individual usage in case other calls were awaiting assignment to the radio channel. So far, there has been no occasion to enforce this limitation. The limitation of 12 minutes was adopted instead of the usual 6-minute limitation common in British telephone practice for the reason that due to the relatively long talk periods on business of this kind, the 6-minute limitation would have resulted in interfering with too large a proportion of these communications.

One problem of interest involved in the transatlantic service was that of fixing rates which allowed of satisfactory expression either in terms of English pounds and shillings or in American dollars. For this purpose, 4 shillings was considered the equivalent of an American dollar. The rate from London to New York, for example, is £ 15 for 3 minutes and £ 5 for each additional minute. Our zone rate steps of $3 for the initial period and $1 for each additional minute were so set in order to allow of even dollar and even shilling quotations for the zone charges. The rate from Cleveland, for example, which is in our second zone, to London is $78 for 3 minutes and $26 for each additional minute. The same rate quoted from London is £ 15: 12 s. for 3 minutes and £ 5: 4 s. for each additional minute. Rate treatment of this kind was thought desirable, not only to allow of easy expression of rates in either English or American money, but also to avoid odd cents in our service charges.

Another problem had to do with the fixing of the hours of service so that the service would be most valuable and usable with due regard to the five hours difference in time between New York and London. At the time the service was opened, limitations on the use of the Rugby sending station for telephone transmission made it possible to keep the channel open only 4½ hours during the day. The hours from 8:30 A.M. to 1:00 P.M., New York time, which correspond with 1:30 to 6 o'clock, London time, were adopted as allowing the maximum overlapping of the London and New York business day. Later, it became possible to extend the hours of operation so that now the service is available 10½ hours—7:30 A.M. to 6 P.M., New York time, which is 12:30 to 11 P.M., London time. The fact that

both London and New York are on a daylight saving schedule in the summer months has required some shifting of the hours of service as these time rearrangements are effected on the two sides of the water.

The operating arrangements set up for the handling of this business provide traffic control operation at the New York and London long distance offices. These offices have direct access to the radio channel via the radio stations at Rocky Point and Houlton and at Rugby and Cupar where technical operators have the transatlantic channel under constant supervision and control. The New York and London long distance offices have special equipment arrangements necessary for connecting the radio channel and the land lines. On calls terminal at New York or London, the operation is similar to that on other terminal calls. On calls involving points beyond New York and London, the New York and London operators assume control, holding the land lines in readiness for prompt connection to the transatlantic channel, supervising the connection and fixing the amount of chargeable time, special measures being provided to protect the user from overcharges that might result from conversations being longer than otherwise necessary because of static and other atmospheric disturbances. The operating method is set up to require a minimum of time on the transatlantic channel for passing calls back and forth and preparing connections.

The personnel necessary to operate the transatlantic circuit is probably not generally appreciated. While two operators in London and two in New York can readily handle the calls themselves, there are six stations, three in each country, for operating and controlling the radio channel and from 35 to 40 men are needed for this work. This force could, of course, handle much more business than is now offered.

Just as the experience with special long distance operating problems and with the development of long distance usage had been largely on this side of the water, so in the matter of international telephone arrangements the experience had been largely with the British Post Office. We have had connection with Canada for many years, but in none of our interchange of business with Canadian companies have we encountered the problems incident to European international communication. The British Post Office, on the other hand, has communication with many countries on the continent and has played its part in the various European conventions and conferences looking to the betterment of international telephone agreements and communication in Europe. So their experience was particularly helpful in shaping up the contract arrangements. In general, the contract

between the British Post Office and the American Telephone and Telegraph Company covers such matters as responsibilities of the two administrations, classes of service, rates, broader operating provisions and settlement matters.

Turning now to the question of the results which are being obtained in this transatlantic service. During the first year, something over 2,300 connections were established. This is an average of about 7 a day, if we include Saturdays, Sundays and holidays, on which days as a general rule the flow of telephone traffic is relatively low. Usage is not very different east and west, something like 55 per cent of the business having originated on this side. Some business from the other side is, of course, from traveling Americans. After the first two months, January and February, when the business amounted to about 250 messages a month and was affected largely by formal openings and curiosity calls, the traffic fell off, and during the summer it was not more than half as great as it had been in the first two months. This may have been due partly to falling off in business activities and possibly also partly to the fact that more atmospheric difficulties are experienced during the summer and the service is then somewhat less dependable. As a matter of fact, there was less atmospheric difficulty than we had anticipated. Starting with September, the business has shown a steady increase. On Christmas Day there were 44 messages.

About half of the transatlantic calls are between New York and London. Over 70 per cent of them originate or terminate in New York City and the remaining calls involve points scattered over the rest of the country. Considering the type of usage of this transatlantic service, nearly half of the calls appear to be of a social nature. As to calls for business purposes, banks and brokerage concerns account for the greatest use so far.

In general, the quality of speech transmission has been more satisfactory than the preliminary tests indicated it would be possible to maintain throughout the year. The radio link is, of course, under careful observation throughout the service period and is not assigned for commercial use unless it appears that reasonably good communication will be obtained. Except for two summer months when atmospheric conditions made telephone communication impossible on an average of about 2 hours a day, the lost time due to static and other such troubles in the radio channel has been relatively small.

Except for brief periods on individual days, the traffic volume has not been sufficiently high to result in any problem in providing a fairly prompt service. At times, and particularly during the summer

months, individual calls have been delayed due to the fact that at the time they were offered, atmospheric conditions made it impossible to use the transatlantic channel. As the business develops, it will doubtless be necessary to adopt special measures for evening the flow of business throughout the period that the transatlantic channel is open for service. At such time as traffic develops to a point where some artificial leveling of the load is required, we would expect this service to involve advance bookings and longer delays than we are accustomed to here in the United States in our internal services. Pending the availability of other transatlantic channels through the use of short wave lengths or otherwise, I do not believe that this type of service would necessarily be seriously objectionable or deterrent to business development. As a matter of fact, a good many calls are now filed in advance.

Differences in the English language as spoken in London and New York became evident as soon as our New York operators were placed in communication with the operators in London. Each group expressed some concern as to what the other was doing to their language. I believe the London operators were inclined to think the broken English spoken by the telephone operators in Holland was sometimes easier to understand than New York City English.

The self-confidence of Americans evidenced itself in the considerable number of calls filed for the nobility, cabinet ministers and other men in the public eye. The fact that most of these calls were accepted by the persons called, indicated their willingness to play the game. Other evidences of this same confident attitude were the suggestions from individuals that they be given a free call so that we could capitalize on the publicity that they would put into their advertising. Others advised us after using the service that, for a consideration, they would allow their names to be used in our publicity material.

I have spoken of some of the operating problems in setting up the transatlantic service and of our experience in handling this service since its inauguration about a year ago. The operating and service arrangements have worked out satisfactorily. The service as a whole has been considerably better than we had anticipated. The volume of business is small but the business now being handled is in line with our general experience in the development of long distance usage. In a situation of this kind, full consideration should be given to the fact that, generally speaking, potential traffic volumes decrease with distance. Telephone service like the transatlantic is a new means of communication. It will not only take time for potential users to become convinced that satisfactory communication can be carried on

by telephone but time is required before they will break away from dependence on other means of communication such as the cables and mails with which they have had long experience and which may have appeared to meet their needs.

The development of our transcontinental business is of interest as this route may be considered as close a parallel as we can find to the transatlantic situation. For several years after the opening of the transcontinental service, the business was small but it has since greatly increased.

The transatlantic channel is a radio channel and this suggests a possible lack of privacy such as is obtained in ordinary telephone communication. Actually, the chances for conversations being picked up by persons to whom they would be of interest or value are rather remote, but this possibility has doubtless had some deterrent effect on the development of business. It is expected that these deterrent factors will be removed in the near future through the introduction of new equipment arrangements which will assure a high degree of privacy on these overseas conversations.

Possibilities for growth must be present in a communication system at the terminals of which we have New York and London, the largest business centers in the world, both English-speaking. On this side, the service has already been extended beyond the United States to Canada and Cuba, and will be extended to Mexico. On the other side the service has recently been extended beyond England, Wales and Scotland to important cities in Belgium, Holland, Germany and Sweden. Further extensions are under consideration to other important continental cities between which and this country there is undoubtedly potential business. As the service is extended beyond Great Britain, a language problem appears. So far about 5 per cent of the conversations are not in English. For the time being, we are relying on the London operators for smoothing out language difficulties in establishing connections and the problem is, of course, not a new one to them. We are planning, however, to set up an operating force here in New York which can communicate in their own language with users who are not speaking in English.

Not only from a technical viewpoint but in other respects we are gratified at the results of the first year's operation of the transatlantic service and we look forward with confidence that this service will be, not only the quickest, but an essential factor in communication between the old world and the new.

Phase Distortion and Phase Distortion Correction

By SALLIE PERO MEAD

SYNOPSIS: The importance of the rôle played by the steady state phase characteristics of long cable circuits has recently been emphasized in telephone and telegraph transmission. In this paper an analytical exposition of the theory of phase distortion is followed by a consideration of various methods of phase distortion correction with particular reference to terminal phase compensating networks and to the application of the lattice network to the loaded line as a terminal phase corrector.

1. INTRODUCTION

FROM the standpoint of ideal quality, a transmission system must be so designed that the received currents, which represent the transmitted signal, shall be a faithful copy of the corresponding currents which enter the transmission system at the sending end; that is, the transmission system must be distortionless. For relatively short distances the deleterious effects of phase distortion are not appreciable but as the range of transmission is increased a point may be reached where the impairment of quality becomes so serious as to reduce the commercial efficiency of the circuits. This fact was first recognized on long submarine telegraph circuits. With regard to telephony, the importance of the question of the quality of received speech was initially emphasized by the advent of the efficient telephone repeater,[1] making possible the increased length of the modern telephone system. This increased length, involving the necessity for circuits of higher quality, led to the development of improved loading designs.[2]

It has long been recognized as a principle of good telephone or telegraph transmission that the variation of attenuation over the range of speech or signal frequencies should be minimized. With the increased length of circuits, however, this requirement alone was found insufficient to insure good quality and *phase distortion* over the range of essential frequencies was found to play an important rôle. Thus steady state phase characteristics have attained prominence in the engineering of long cable circuits and in the application of this technique to telegraph, telephone and picture transmission.

Distortion in the variation with frequency of the phase difference between the current at the receiving end and that at the sending end of the transmission line, as well as distortion of the amplitude characteristic of the received as compared to the initial current, gives rise

[1] See reference 1.
[2] See reference 2.

to so-called transient effects. That is, the currents, after arriving at the distant end of the communication circuit, require an appreciable time, varying with the impressed frequency, to build up and, under certain conditions, may never build up to anything remotely resembling the transmitted currents in the short interval during which the latter exist. This effect, moreover, may produce serious impairment in the quality of the received speech or signal even when the line is so designed that the steady state attenuation of all currents in the essential range is substantially constant.

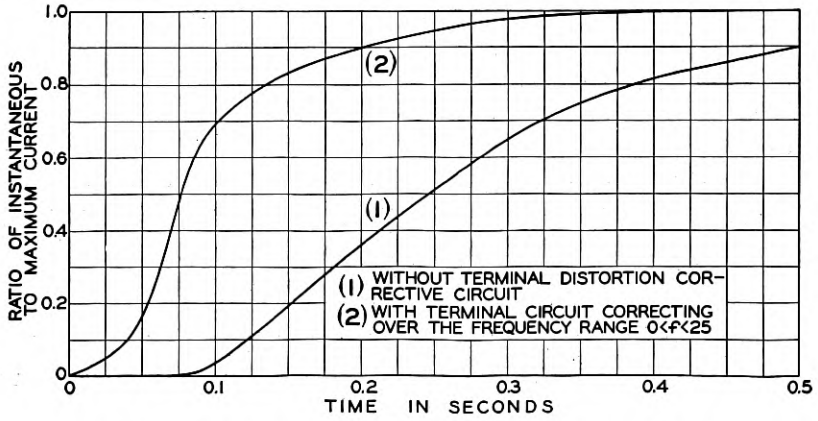


Fig. 1—Building-up of current on 1500 mile telegraph cable

As an illustration of the transient distortion on a transmission line, consider the indicial admittance, $A(t)$, of the cable of length l miles, and of resistance R and capacity C per mile; that is, the received current in response to a unit d.c. voltage applied at the sending end at time $t = 0$. This is given by [3]

$$A(t) = \frac{2}{Rl} \frac{e^{-1/y}}{\sqrt{y}},$$

where

$$y = \frac{4t}{RCl^2}.$$

Curve (1) of Fig. 1 represents relative values of $A(t)$ on a cable 1,500 miles long. (This is the same cable whose phase characteristic is shown in Fig. 5, the inductance being ignorable in determining the indicial admittance.) The departure of the received current from the abrupt wave front of the impressed d.c. voltage is clear.

[3] See reference 3.

The distortion on the loaded cable from the transient point of view is represented in Fig. 2, which shows the envelope of the building-up of current of frequency $\omega/2\pi$ in the Nth section of a periodically loaded line.[4]

The oscillograms of Fig. 3 show the received current in response to sinusoidal waves on a 600-mile medium heavy loaded (H-174) line.[5]

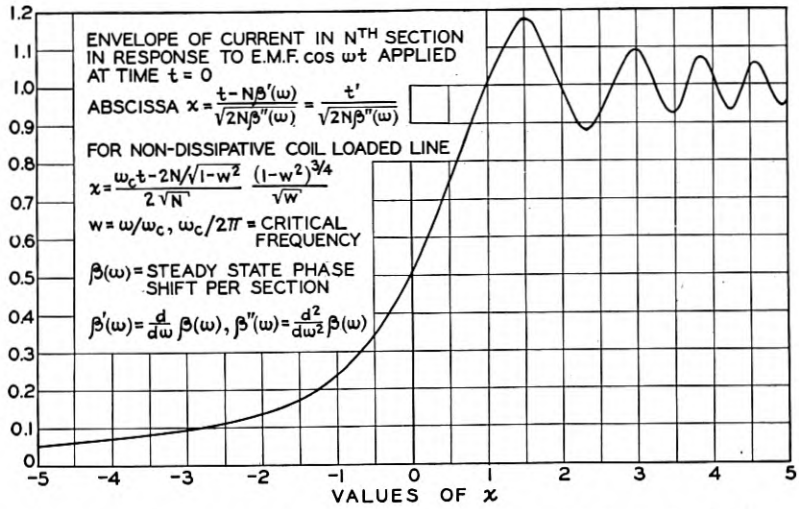


Fig. 2—Building-up of alternating currents in long periodically loaded line

Figs. 3*a* and 3*b* are for frequencies of 1,000 cycles and 1,500 cycles respectively and Fig. 3*c* is for a compound wave made up of 800 and 1,600 cycles. The last oscillogram shows clearly the greater delay of the higher frequency. Due to the relative weakness of this component the building-up transients while quite apparent are not pronounced. It will be observed that an appreciable time has elapsed in each case before the received wave has built up to the amplitude or frequency of the steady state.

The detrimental effects of transient distortion were first studied in the long submarine cable. Early in 1918 a theory of distortion correction was developed by John R. Carson from the transient point of view. The principle was arrived at that the *modified arrival curve* of prescribed form (a square-topped wave in the case of long line telegraphy) may be obtained by combining derivatives with respect to time of the *datum arrival curve* (the current obtained at the end of

[4] See reference 4.

[5] In this symbolism "H" refers to coil spacing of 6,000 ft., and the following number gives the inductance in millihenrys.

the cable itself) in the proper proportions to insure the steepness of building-up of the arrival curve and the maintaining of the tail of the wave. Terminal corrective networks [6] in combination with vacuum tubes were designed to obtain the requisite derivatives. Such a terminal corrective network is shown in Fig. 4, where the

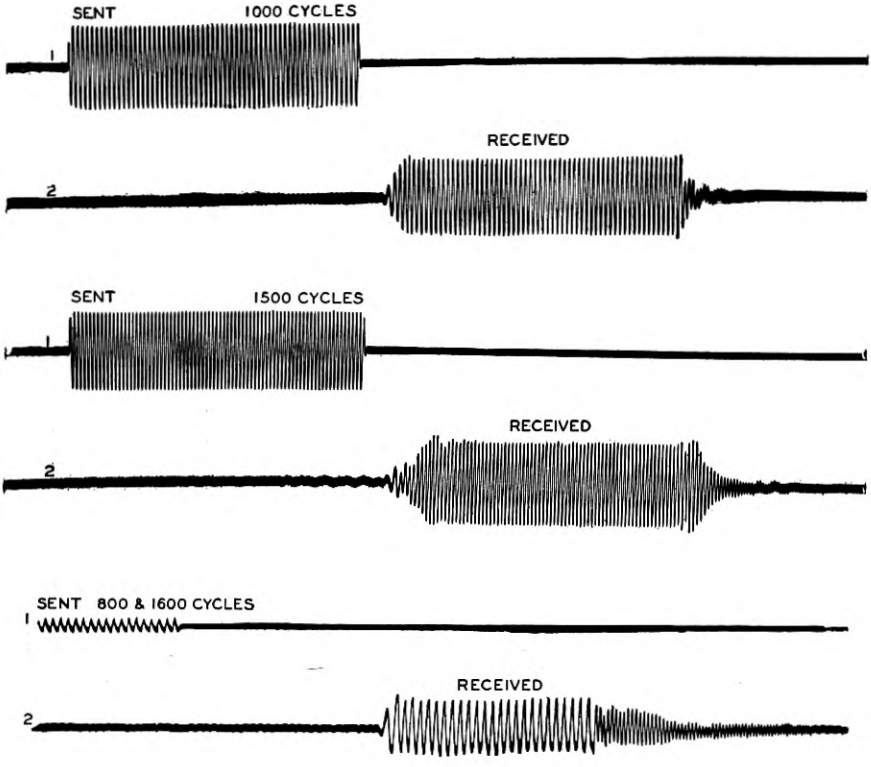


Fig. 3—Transient distortion in a 600 mile length of H-174 loaded cable

resistance R_1 is a distortionless one-way thermionic tube and $V(\omega)$ the output voltage. Examination of this network in the light of the more recent study of phase distortion correction from the steady state point of view has shown that it does also correct the phase distortion of the cable and provide some attenuation equalization as well.

Although the design of corrective networks on the basis of the steady state phase has usually been found more simple than on the transient basis, a knowledge of the arrival current or voltage as an

[6] See references 5 and 6. Also, for the development of distortion corrective circuits with vacuum tube amplifiers, or "signal shaping vacuum tube amplifiers" as they are called, in connection with their application to the new permalloy cables of the North Atlantic, see references 7 and 8.

explicit time function is sometimes essential, in the last analysis, to indicate the degree of correction afforded. This information may be supplied to sufficient precision by the first and succeeding derivatives with respect to frequency of the steady state phase, which, as explained

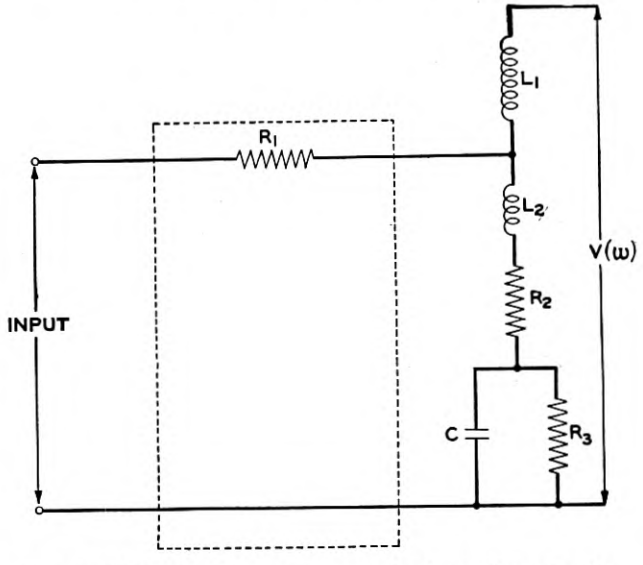


Fig. 4—Distortion corrective circuit for long telegraph cable

below, are extremely useful criteria of the improvement in the time and steepness, respectively, of building-up over a finite range of frequencies. Nevertheless, to visualize the actual effect of the applied phase compensation upon the arrival current or voltage due to an impressed e.m.f. of a specific frequency requires the evaluation of the explicit time function.

It is the object of this paper, following an analytical exposition of the theory of phase distortion, to consider various methods of phase distortion correction with particular reference to terminal phase compensating networks and the application of the lattice network to the loaded line as a terminal phase corrector.

II. Phase Distortion

1. *Steady State Theory of Phase Distortion*

The mathematical theory of phase distortion in signaling systems may be explained briefly from the steady state point of view. We suppose that it is required to transmit a signal $f(t)$, which can be represented by the Fourier integral,

$$f(t) = \frac{1}{\pi}\int_0^\infty F(\omega)\cos[\omega t + \theta(\omega)]d\omega, \tag{1}$$

and that the transfer impedance of the transmission system is given by

$$Z(i\omega) = |Z(i\omega)|e^{iB(\omega)},$$

where $\omega/2\pi$ is the frequency. The received current is, then,

$$I(t) = \frac{1}{\pi}\int_0^\infty \frac{F(\omega)}{|Z(i\omega)|}\cos[\omega t + \theta(\omega) - B(\omega)]d\omega. \tag{2}$$

$F(\omega)$ exists for all values of ω from zero to infinity but, practically, $F(\omega)$ is negligible except over a finite range which is determined by the nature of the signal. For program transmission, for example, the essential frequencies are now considered to lie in a band from about 100 to 5,000 cycles, while for slow speed telegraphy they lie in a band between zero and 10 or 20 cycles per second. If we suppose, then, that the essential frequency band extends from $\omega_1/2\pi$ to $\omega_2/2\pi$, we may replace equations (1) and (2) by

$$f(t) = \frac{1}{\pi}\int_{\omega_1}^{\omega_2} F(\omega)\cos[\omega t + \theta(\omega)]d\omega \tag{3}$$

and

$$I(t) = \frac{1}{\pi}\int_{\omega_1}^{\omega_2} \frac{F(\omega)}{|Z(i\omega)|}\cos[\omega t + \theta(\omega) - B(\omega)]d\omega. \tag{4}$$

Now suppose that within the band of essential frequencies, $\omega_1 < \omega < \omega_2$, we have

$$|Z(i\omega)| = R \tag{5}$$

and

$$B(\omega) = \omega\tau \pm n\pi,$$

where R and τ are constants and $n = 0, 1, 2, \cdots$. Then we may write

$$I(t) = \pm\frac{1}{\pi R}\int_{\omega_1}^{\omega_2} F(\omega)\cos[\omega(t-\tau) + \theta(\omega)]d\omega, \tag{6}$$

$$= \pm\frac{1}{R}f(t-\tau),$$

$I(t)$ being positive or negative according to whether n is even or odd. Whence the received current is proportional in amplitude to the applied signal and merely delayed in time by the 'transmission time' τ. Thus the received current has the same wave form as the applied

signal or the transmission is distortionless. Accordingly, we have the following proposition.[7] *The necessary and sufficient condition for the practically distortionless transmission of signals in communication systems is that, over the essential range of frequencies contained in the transmitted signal, the transfer impedance of the transmission circuit be equalized both as regards amplitude and phase; that is, the amplitude must be constant and the phase angle linear in the frequency, with a value, when the frequency is zero, of* $\pm n\pi$, *where* $n = 0, 1, 2, \cdots$.

For many years the variation of the phase angle with frequency was ignored. Research in distortion correction was directed to devising networks [8] so designed that $|Z(i\omega)|$ would be a constant, R, over the range of essential frequencies. Assuming that this condition is fulfilled by the transducer but that

$$B(\omega) = \omega\tau + \sigma(\omega) \pm n\pi,$$

where $\sigma(\omega)$ is non-linear in the frequency, we may write (4) as

$$I(t) = \pm\frac{1}{\pi R}\int_{\omega_1}^{\omega_2} F(\omega)\cos[\omega(t-\tau) + \theta(\omega) - \sigma(\omega)]d\omega. \qquad (7)$$

In formula (7) the *amplitudes* of the component frequencies of the arrival curve are, within a constant, the same as those in the impressed signal $f(t)$. The *wave form* of the arrival curve, owing to the presence of the phase $\sigma(\omega)$, may, however, be widely different from that of the impressed signal.[9]

2. *Examples of Phase Distortion in Transmission Systems*

Let us consider the frequency-phase angle characteristic of the two important transmission systems, the submarine telegraph cable and the loaded line.

The cable of characteristic impedance $k = \sqrt{(R + i\omega L)/i\omega C}$ and propagation constant $\gamma = \sqrt{(R + i\omega L)i\omega C}$ (with negligible leakage) is assumed terminated in its characteristic impedance at $x = l$ so that reflection is suppressed. The transfer impedance $Z(i\omega)$ is then

$$\begin{aligned} Z(i\omega) &= ke^{\gamma l} \\ &= |Z(i\omega)|e^{iB(\omega)}, \end{aligned} \qquad (8)$$

[7] See reference 9.
[8] See reference 10.
[9] In telephone transmission it is not at all certain that preservation of wave form is essential. It is essential, however, that the components of different frequencies build up at approximately the same time. It is further demonstrated in the section on 'Loading Systems' below that $\sigma(\omega) = 0$ is the necessary and sufficient condition to fulfill the latter requirement.

where $B(\omega) = \omega\tau + \sigma(\omega) \pm n\pi$ and is the phase angle of the transfer impedance, provided, as we shall assume, that k is approximately a constant. Whether k is a constant or not, $B(\omega)$ represents the difference in phase between the currents at the sending and receiving ends. $B(\omega)$ is given by

$$B(\omega) = l\omega\sqrt{LC}\sqrt{\tfrac{1}{2}[1 + \sqrt{1 + (R/\omega L)^2}]}. \tag{9}$$

Fig. 5 shows $B(\omega)$ and $\sigma(\omega)$ in radians for a 500-mile length of cable whose constants are:

resistance $R = 2.74$ ohms per mile,
inductance $L = 0.001$ henry per mile,
capacity $C = 0.296$ microfarad per mile,

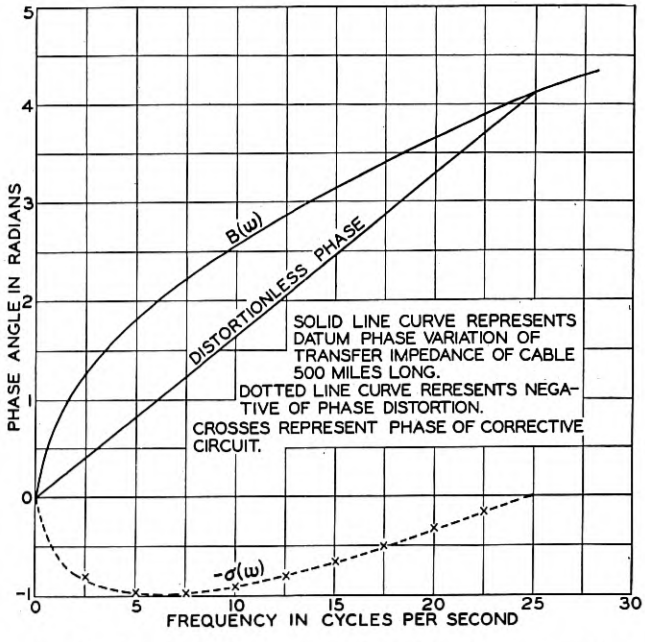


Fig. 5—Phase distortion correction on long telegraph cable

and for the frequency range, 0–25 cycles per second. The straight line, $\omega\tau \pm n\pi$, representing the distortionless phase characteristic is not fixed except that it must pass through the origin or $\pm n\pi$ at zero frequency. It is here chosen to pass through the origin and to have the same value as the cable phase itself at $f = 25$ c.p.s. The dotted curve representing $-\sigma(\omega)$ then shows the departure (with the sign reversed) of the cable phase from the distortionless characteristic.

With regard to the loaded cable, we have similarly the phase angle $B(\omega)$ of the transfer impedance of the cable of length l miles with load impedance Z and smooth line constants R, L, G and C per mile, given rigorously by

$$B(\omega) = \text{imag. comp.} \frac{l}{s} \cosh^{-1} \left[\cosh \gamma s + \frac{Z}{2k} \sinh \gamma s \right], \qquad (10)$$

where

s = spacing of load coils in miles,

$\gamma = \sqrt{(R + i\omega L)(G + i\omega C)}$,

$k = \sqrt{(R + i\omega L)/(G + i\omega C)}$.

It has been found, however, that dissipation and the distributed nature of the line constants have no appreciable effect on the phase

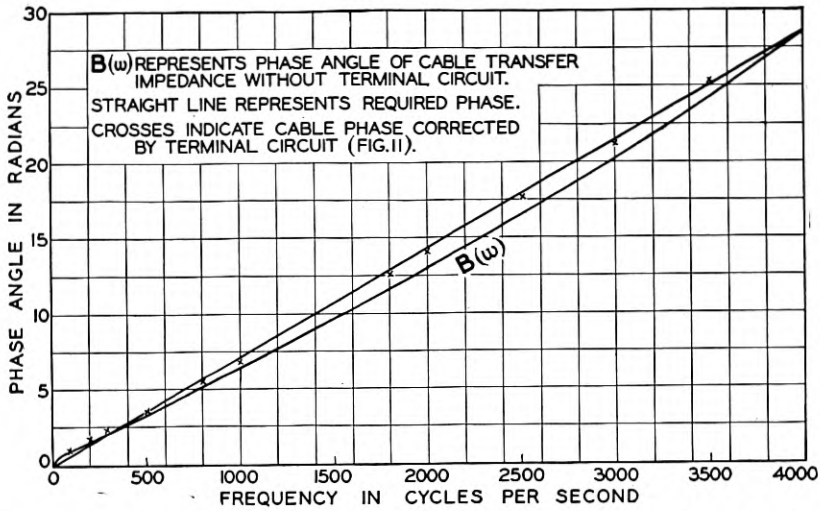


Fig. 6—Phase distortion correction on 20 mile 19 gauge H-44-25 loaded cable, $0 < f < 4000$

variation below 4,000 cycles. Hence, as a close approximation, we may take $B(\omega)$ for the non-dissipative cable without distributed inductance; i.e., simply,

$$B(\omega) = 2N \sin^{-1} f/f_c, \qquad (11)$$

where

$$f_c = \frac{1}{\pi\sqrt{L_0 C_0}},$$

N = number of sections,

L_0 is the coil inductance and C_0 the lumped line capacity per secti on

Even on the light loaded lines designed especially for good quality on long repeatered circuits, the phase distortion is appreciable. The nominal cut-off of these circuits is about 5,600. Fig. 6 shows the phase characteristic of the transfer impedance of a section of side circuit of 19 Gauge H-44 cable only 20 miles long. On Fig. 10 is represented the negative of the phase distortion, $\sigma(\omega)$, obtained by taking $n = 0$ and $\tau = B(\omega_m)/\omega_m$ where $\omega_m/2\pi$ is taken as the highest essential frequency, in this case 4,000 cycles.

In speaking of a pure sinusoidal wave of only one frequency, a phase shift of more than 2π radians or one cycle would be meaningless since every cycle is identical to the preceding and the following cycles. To consider the variation of phase shift over a range of frequencies, however, the total phase shift at any frequency as compared to that at the lowest frequency of the range is required.

III. Phase Distortion Correction

1. *Terminal Networks: Application to the Submarine Cable*

The device of a terminal network having a compensating phase distortion, that is, a network having the phase angle of transfer impedance,

$$\phi(\omega) = [\omega\tau' - \sigma(\omega) \pm n\pi], \tag{12}$$

over the frequency interval $\omega_1 < \omega < \omega_2$ (τ' being a constant), is theoretically the most simple and, in practice, is probably the most flexible and effective method of phase distortion correction. Such a distortion corrective network, in series combination with the transducer in which the attenuation has been equalized, produces an arrival curve

$$I(t) = \frac{1}{\pi R}\int_{\omega_1}^{\omega_2} F(\omega) \cos [\omega(t - \tau - \tau') + \theta(\omega) \pm n\pi]d\omega, \tag{13}$$

which is proportional to

$$f(t - \tau - \tau')$$

provided, of course, that there is no reflection at the transducer terminals. The constant phase angle $\pm n\pi$ does not affect the sinusoidal wave form but merely changes the sign of the wave if n is odd.

The terminal phase corrective network or phase compensator is applicable, at least theoretically, to any type of phase distortion correction and may be supplementary to other forms of correction

such as loading, for instance. Fig. 7 is a schematic diagram of the arrangement of the given transducer of transfer impedance $Z(i\omega)$ with phase angle $B(\omega)$ and the terminal phase distortion corrective network of transfer impedance $N(i\omega)$ with phase angle $\phi(\omega)$. In

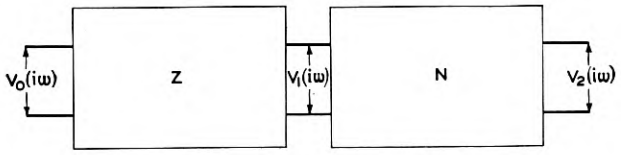


Fig. 7

response to the impressed voltage $V_0(i\omega)$, the voltage $V_1(i\omega)$ at the output terminals of the transducer, which is assumed proportional to the current, is then

$$V_1(i\omega) = \frac{1}{|Z(i\omega)|} e^{-iB(\omega)} V_0(i\omega)$$

and the final voltage $V_2(i\omega)$ is

$$V_2(i\omega) = \frac{1}{|N(i\omega)|} e^{-i\phi(\omega)} V_1(i\omega).$$

Thus

$$\frac{V_2(i\omega)}{V_0(i\omega)} = \frac{1}{|Z(i\omega)|} \frac{1}{|N(i\omega)|} e^{-i[B(\omega)+\phi(\omega)]}. \tag{14}$$

In practical applications, it is usually found advisable to take both τ' and n of equation (12) equal to zero. Then the required phase characteristic, $\phi(\omega)$, of the transfer impedance of the corrective network is

$$\phi(\omega) = -\sigma(\omega).$$

The function $-\sigma(\omega)$ is drawn in Fig. 5 for the submarine cable where $0 < \omega < \omega_m$ and $\omega_m/2\pi = 25$ cycles per second. This may be represented quite closely analytically by the expression

$$\phi(\omega) = \tan^{-1} \frac{ax}{1 + bx^2}, \tag{15}$$

where

$$x = \frac{\omega}{\omega_m} \left(1 - \frac{\omega_m^2}{\omega^2} \right). \tag{16}$$

Thus

when $\omega = 0$, $x = -\infty$ and $\phi = 0$,

when $\omega = \omega_m$, $x = 0$ and $\phi = 0$,

and

when $0 < \omega < \omega_m$, $-\infty < x < 0$ and $\phi < 0$.

Physically, this is realizable in the circuit of Fig. 8 consisting of a resonant element L, C, where $\omega_m = 1/\sqrt{LC}$, in parallel with a resistance R_1. The final voltage is taken across the resistance R_2 and the resistance r represents a vacuum tube. This unilateral element permits of suitable amplification and prevents reflection at the cable terminals so that the network may be designed without regard to any reaction upon the cable.

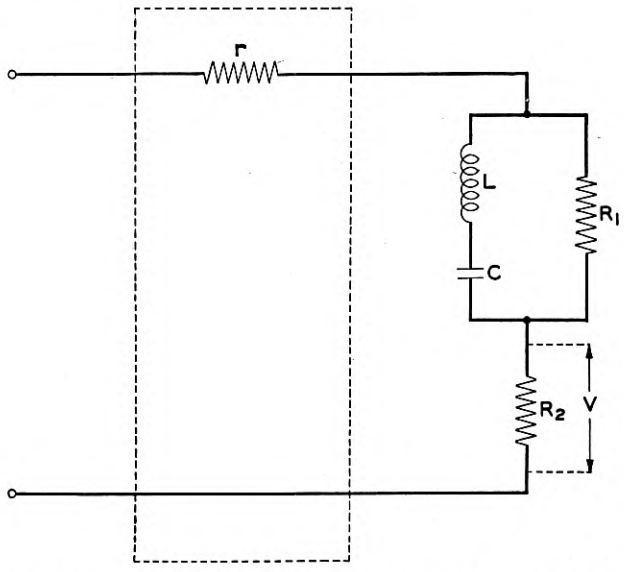


Fig. 8—Distortion corrective circuit for long telegraph cable

One section of the network of Fig. 8 with the values of the constants:

$$r + R_2 = 5{,}000 \text{ ohms}, \qquad L = 26 \text{ henrys}$$
$$R_1 = 51{,}100 \text{ ohms}, \qquad C = 1.56 \text{ microfarads}$$

is used when $l = 500$ miles. The phase of this network is shown in Fig. 5 also. Another equal network section may be added for each additional 500 miles of cable but there is no necessity, of course, for the sections to be equal. If it contains a one-way thermionic tube, each section may be added without affecting what has gone before, and the resultant phase angle will be simply the sum of all of the phase angles of the separate parts.

The improvement in the building-up of the indicial admittance accompanying the use of the phase compensator is evident from

Fig. 1 on comparing curve (2) with curve (1). Curve (2) is computed from the formula [3]

$$A(t) = \frac{2}{\pi}\int_0^\infty \frac{\alpha(\omega)}{\omega} \sin t\omega d\omega, \tag{17}$$

where $\alpha(\omega)$ is the real component of the transfer admittance of cable and network combined (equation (14)).

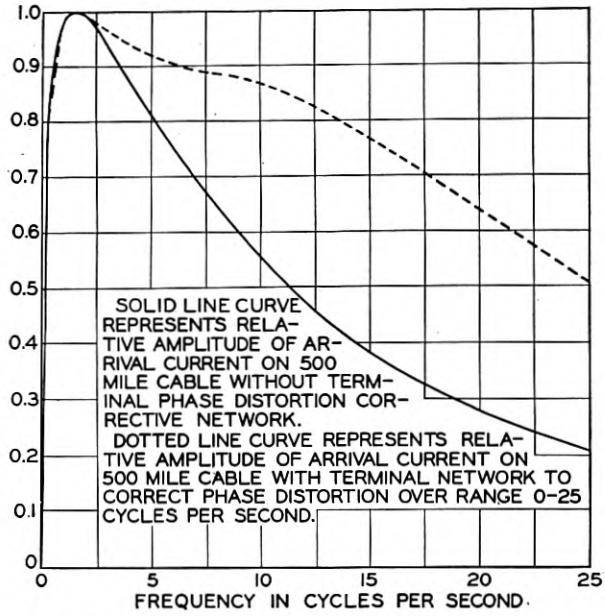


Fig. 9—Amplitude variation on long telegraph cable

The curves of Fig. 9 show that this network affords some attenuation equalization as well as very good phase correction. Amplitude and phase correction, as we have seen, are analytically independent processes. Nevertheless, some arrangements may, theoretically, be designed to correct amplitude and phase simultaneously. A method for so designing a network similar to the one under discussion at present has been developed by O. J. Zobel.[10] In such cases, however, in order to obtain physically desirable values in practical applications, it has usually been found necessary to design the network for one purpose, thereby automatically obtaining some improvement in the other respect, as in the present instance.

The maximum phase displacement obtainable with one section of

[10] This is discussed in a forthcoming paper by O. J. Zobel.

this network is $\pi/2$ radians. Thus, it becomes necessary to use more than one section on a long cable but it is also advantageous from the point of view of flexibility of design. By adopting a standard unit, for a 500-mile section of cable, for instance, the networks may be readily accommodated to different lengths of line.

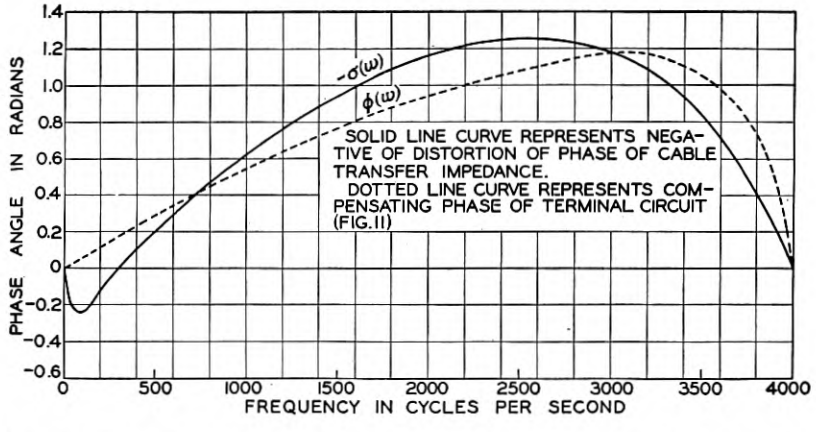


Fig. 10—Phase distortion correction on 20 mile 19 gauge H-44-25 loaded cable, $0 < f < 4000$

It is interesting to observe that the analytical expression

$$\phi(\omega) = \tan^{-1}\frac{ax}{1 + bx^2} \tag{15}$$

is positive when $0 < \omega < \omega_m$, and zero when $\omega = 0$ or ω_m, provided

$$x = \frac{\omega/\omega_m}{1 - \omega^2/\omega_m{}^2}. \tag{18}$$

Hence it will correct the phase distortion on the loaded cable as shown in Fig. 6. The required phase shift is obtainable in the type of network shown in Fig. 11. This network, however, has the disadvantage of tending to increase the attenuation distortion of the loaded cable rather than to equalize it.

When amplification is not required, it is undesirable to use the device of an amplifier in each section of network for the sole purpose of eliminating terminal reflection. As the characteristic impedance of a transmission line may usually be regarded as a constant resistance over the range of essential frequencies, the same purpose may be accomplished by applying networks whose characteristic impedance

is also a constant resistance of the same value as the characteristic impedance of the line. A number of general recurrent networks of the so-called 'constant resistance' type [11] are known and such a

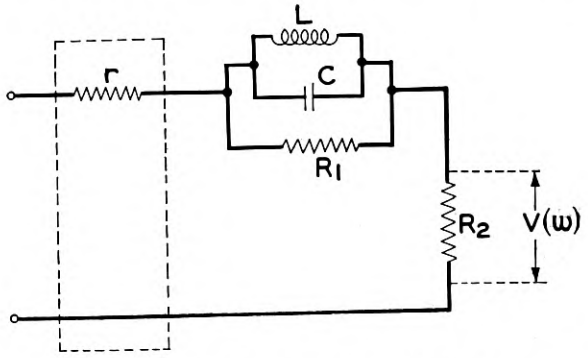


Fig. 11—Phase distortion corrective circuit for loaded cable. For 20 miles of 19 gauge H-44-25 loaded cable, r + R_2 = 5000 ohms, R_1 = 101,000 ohms, L = 0.460 henry, C = 0.0216 microfarad.

network [10] has been applied to correct the distortion on the submarine cable. The arrangement is shown in Fig. 12. It will be observed that z_{11} and z_{21} are inverse networks of impedance product R^2; that

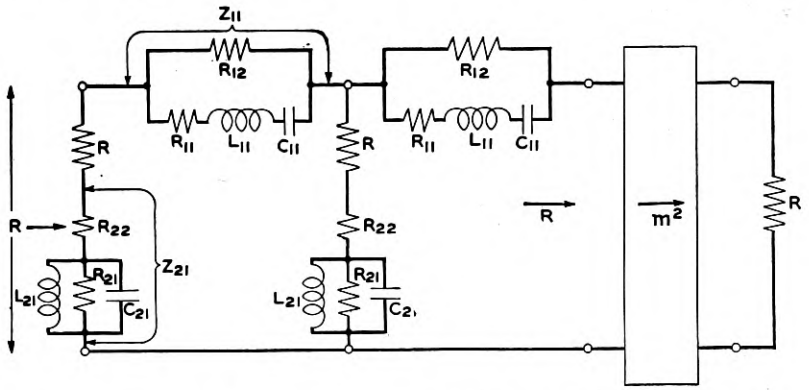


Fig. 12—Constant resistance type corrective circuit for long telegraph cable. m^2 represents distortionless amplifier

is, $z_{11}z_{21} = R^2$, which is, in general, the requisite relation among the network constants to provide a constant resistance characteristic impedance, R. In practical performance the networks of Figs. 8 and 12, each having the basic arrangement z_{11}, are equivalent. Thus the

[11] See reference 11.

function of eliminating terminal reflection, accomplished in the former by the vacuum tube, is accomplished in the latter by the additional impedance elements.

2. *Loading Systems*

The use of loaded cable circuits for the transmission of programs and for the transmission of pictures introduced some interesting problems in connection with phase correction at the higher frequencies. In this connection, a study of the building-up of sinusoidal oscillations in long loaded cables led to the establishment of two propositions which are of great value in comparing communication systems with respect to the quality of transmission.[12] These two propositions relate the variation with frequency of the steady state phase to the duration and nature of the transient distortion. They are applicable to all types of periodic loading, with or without terminal phase compensators under two restrictions; namely, (1) the line must comprise at least 100 loading sections and (2) the transducer as a whole must be approximately equalized as regards absolute steady state values of the received current in the neighborhood of the applied frequency. The successive derivatives of the total phase angle $B(\omega)$ with respect to ω will be denoted by $B'(\omega)$, $B''(\omega)$, $B'''(\omega)$. $B(\omega)$ may be understood to represent the sum of the phase differences due to transmission over the line and a terminal distortion corrective network, as well as the phase difference on the line alone. The propositions follow.

Case I: $B''(\omega) \neq 0$ and $\sqrt{B''(\omega)/2!}$ large compared with $\sqrt[3]{B'''(\omega)/3!}$.

The envelope of the oscillations in response to an e.m.f. $E \cos \omega t$ *applied at time* $t = 0$ *reaches 50 per cent. of its ultimate steady value at time* $t = B'(\omega)$ *and its rate of building-up is inversely proportional to* $\sqrt{B''(\omega)}$.

Case II: $B''(\omega) = 0$, $B'''(\omega) \neq 0$ and $\sqrt[3]{B'''(\omega)/3!}$ large compared with $\sqrt[4]{B^{IV}(\omega)/4!}$.

The envelope of the oscillations in response to an e.m.f. $E \cos \omega t$ *applied at time* $t = 0$ *reaches 1/3 of its ultimate steady value at time* $t = B'(\omega)$ *and its rate of building-up is inversely proportional to* $\sqrt[3]{B'''(\omega)}$.

These propositions furnish supplementary verification of the condition with regard to phase variation already established as a requirement for distortionless transmission; namely, that the phase vary linearly with the frequency or

$$B(\omega) = \omega\tau,$$

[12] See reference 4.

where τ is a constant. It follows immediately that

$$B'(\omega) = \tau.$$

Thus, when the steady state value of phase propagation is in linear relation with the frequency, all signals of any frequency whatever reach their proximate steady state in the same interval of time τ. These propositions make it possible to calculate two important criteria of the transmission properties of the line: (1) the variation with respect to frequency of the time interval τ required for the current to build up to its proximate steady state value and (2) its rate of building-up at time $t = \tau$. The first is very important in telephony but has no significance in telegraphy since only one frequency is transmitted, whereas the second is important in both telephony and telegraphy. While the formulas underlying these propositions are approximate, they are sufficiently accurate for a study of the comparative merit of different types of transmission systems or for the design of loaded lines.

For the purpose of reducing transient distortion on a long loaded cable of N sections, the loading and critical frequency $\omega_c/2\pi$ may be designed on the basis that the two time intervals,

$$t_1 = 2N/\omega_c \tag{19}$$

and

$$t_2 = \frac{2N}{\omega_c}\left(\frac{1}{\sqrt{1 - \omega^2/\omega_c^2}} - 1\right), \tag{20}$$

should be less than given quantities determined by experience. The two expressions t_1 and t_2 represent, respectively, (1) the time of transmission of d.c. current or the nominal time of transmission and (2) the excess of the time of building-up of the frequency $\omega/2\pi$ to approximately one half the steady state value over the nominal time of transmission.[13] The right hand member of formula (20), although initially determined by Carson by a method entirely dissimilar to that in his later investigation, is, it will be noted, given by

$$t_2 = B'(\omega) - B'(0) = T - T_0 \tag{21}$$

and

$$t_1 = B'(0) = T_0. \tag{22}$$

The nominal time of transmission t_1 is significant from the standpoint of echo effects but, with regard to phase distortion correction alone, emphasis is placed upon the quantity $t_2 = T - T_0$ as the more important factor. The enormous improvement in this respect of the

[13] See reference 2.

system of light loading over medium heavy loading is shown in Fig. 19. Below about 3,200 cycles, $T - T_0 < .0005$ second on a 50 mile light loaded line. It is obvious that the higher the frequency the greater the distortion.

Another loading system proposed as a means of providing desirable phase characteristics is the lattice loaded line.[14] This consists of the use of a section of the network of Fig. 14 as the periodic loading unit instead of the ordinary coil unit. Putting $C = rC_0$, where C_0 is the capacity of the line between loading units per section, we have for this system, when non-dissipative,

$$f_c = \frac{1}{\pi\sqrt{LC_0}}, \tag{23}$$

$$B(\omega) = 2N \sin^{-1} \frac{f}{f_c} \sqrt{\frac{1+r}{1+r(f/f_c)^2}}, \tag{24}$$

$$B'(\omega) = \frac{N}{\pi f_c} \sqrt{1+r} \frac{1}{[1+r(f/f_c)^2]\sqrt{1-(f/f_c)^2}} \tag{25}$$

and

$$T_0 = B'(0) = \frac{N}{\pi f_c} \sqrt{1+r}. \tag{26}$$

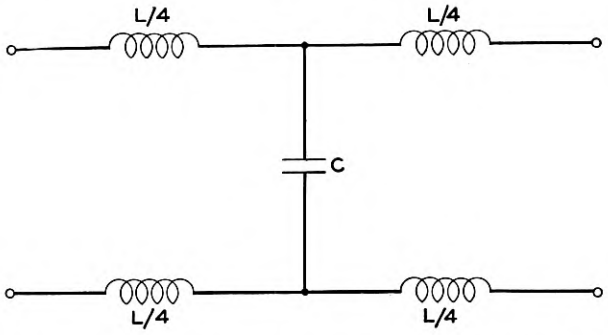


Fig. 13*a*—Section of standard non-dissipative coil loaded line. L = .044 henry, C = .0705 microfarad (for H-44 loading on 19 gauge side circuit.)

With the constants of the lattice loading system chosen to give approximately the same attenuation per mile as the standard loading, as in Figs. 13*a* and 13*b*, the time $t_2 = T - T_0$, it will be seen from Fig. 19, is less for it than that for the standard loading system for frequencies up to about 3,500 cycles but rapidly becomes greater thereafter. A combination of coil and lattice loading obtained by

[14] This was proposed by D. A. Quarles of the Bell Telephone Laboratories in unpublished memoranda.

alternating the coil and lattice loads affords some improvement in this respect. Another possible means of reducing the time interval $T - T_0$ at the higher frequencies is to reduce the spacing of the

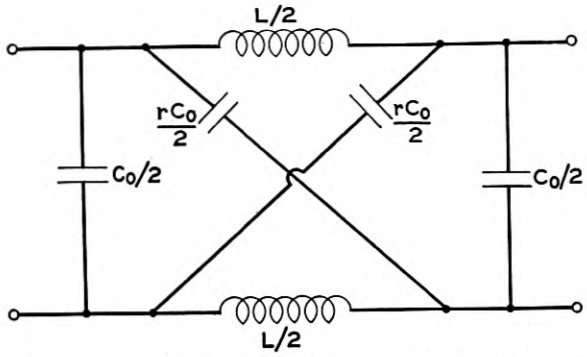


Fig. 13b—Section of non-dissipative lattice loaded line. L = .066 henry, C_0 = .0705 microfarad, r = .50 (19 gauge cable).

loading coils on the ordinary loaded line, keeping the attenuation per mile the same as before. As this change increases N and f_c in the same ratio, the effect is to reduce $T - T_0$.

3. *Lattice Type Terminal Network*

Instead of being used to replace the coil unit in the long periodically loaded line for the purpose of reducing the phase distortion, the lattice network [15] may be applied to the coil loaded line as a terminal phase compensator, a use to which it is peculiarly adapted by virtue of the nature of its characteristics.[16] These have been known for many years. The ideal non-dissipative simple lattice type recurrent network, shown in Fig. 14, with series inductance L and crossed capacity C per section, has a pure resistance characteristic impedance

$$K = \sqrt{L/C}, \tag{27}$$

a phase angle

$$\phi(\omega) = 2N \tan^{-1} (\omega\sqrt{LC}/2), \tag{28}$$

where N is the number of sections, and zero attenuation for all frequencies.[17] These relations readily follow from the general formulas (32)–(35) given below. The phase angle (Fig. 15), therefore, is

[15] The possible usefulness of the lattice network as a phase shifting device was pointed out in a general way by G. A. Campbell. See references 12 and 13.

[16] In this connection see references 12, 13 and 14 to the work of Karl Kupfmuller of the Siemens-Halske Company of Berlin, who independently applied the simple lattice network in the same way.

[17] See reference 13.

positive and varies from zero to π per section. Thus it is of a general form suitable to compensate for the distortion in the cable phase. One or more sections of the simple lattice network when properly

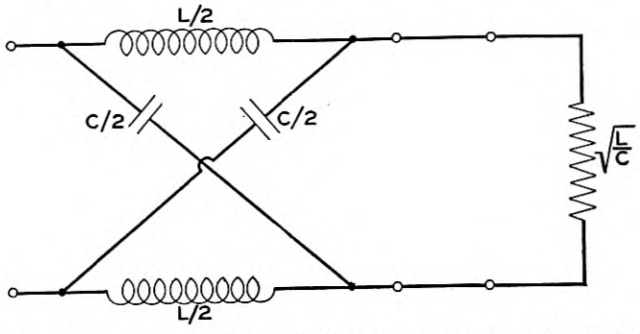


Fig. 14—Phase distortion corrective circuit for loaded cable: simple lattice type

terminated may then be connected directly to the loaded cable without appreciable reflection loss and without affecting the attenuation characteristic of the cable.

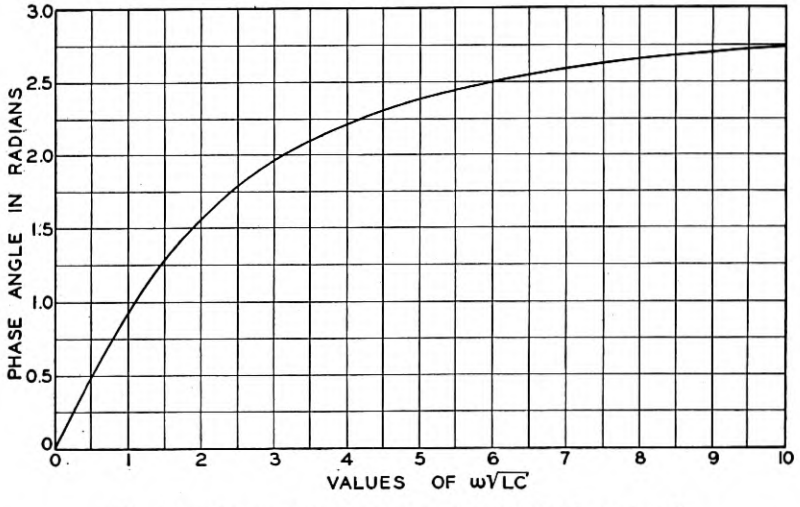


Fig. 15—Phase angle per section of simple lattice network

Proceeding to the transient aspect, the time T required for the current in response to the e.m.f. of frequency $\omega/2\pi$ to build up to 50 per cent. of its steady state value is given by

$$T = \phi'(\omega) = N\sqrt{LC}\,\frac{1}{1 + \left(\frac{\sqrt{LC}}{2}\,\omega\right)^2} \tag{29}$$

and

$$T - T_0 = \phi'(\omega) - \phi'(0)$$
$$= N\sqrt{LC}\left(\frac{1}{1+\left(\frac{\sqrt{LC}}{2}\omega\right)^2} - 1\right). \tag{30}$$

The variation of $T - T_0$ in terms of the general parameter $\omega\sqrt{LC}$ is shown in Fig. 16. As T is a maximum at zero frequency and decreases with increasing frequency, its variation tends to equalize the delay on the cable. It will be demonstrated later that a more complicated type of lattice network is available to supplement the simple lattice type to improve the equalization still further.

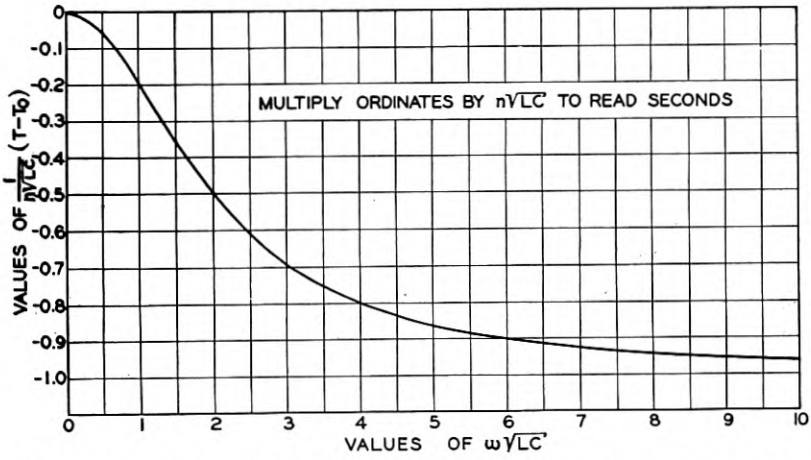


Fig. 16—Time of building-up, $T - T_0$, for simple lattice network of n sections

To provide partial delay equalization by means of the simple lattice network, phase angle variation alone need be considered without taking into account its derivatives. As pointed out above, the distortion $\sigma(\omega)$ of the cable phase $B(\omega)$ is given by

$$\sigma(\omega) = B(\omega) - \omega\tau \pm n\pi,$$

where τ is a constant representing the slope of the required distortionless phase. In the present case n will be zero. The constant τ also represents the time in which the current will build up to proximate steady state on the corrected cable. It is desirable, usually, that this time be as short as possible, and, moreover, the smaller the value of τ, the smaller will be the amount of distortion to be corrected,

remembering, of course, that $\sigma(\omega)$ is to be negative. On the other hand, a consideration of the nature of the phase variation, $\phi(\omega)$, of the lattice network as compared to the distortion of the cable phase

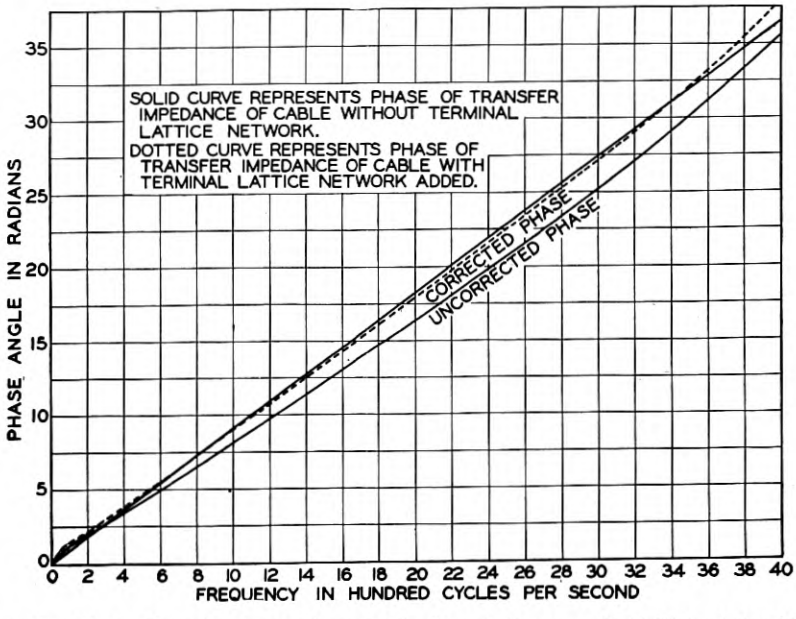


Fig. 17—Phase distortion correction on 25 miles of 19 gauge H-44-25 loaded cable

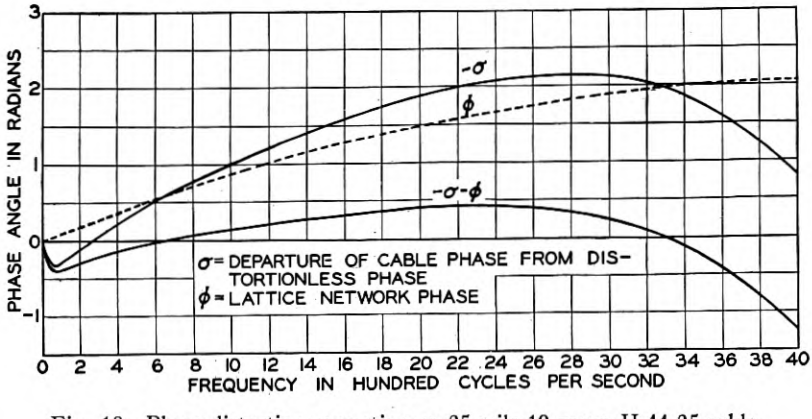


Fig. 18—Phase distortion correction on 25 mile 19-gauge H-44-25 cable

corresponding to various values of τ, leads to a choice of τ such that approximately the maximum cable distortion occurs at a frequency somewhat below the upper frequency of the essential range (Figs. 17 and 18).

While $\phi(\infty) = \pi$, it is apparent from Fig. 15 that ϕ increases very slowly for values of $\phi > 2.5$ and it is found that a distortion angle of not more than about 2.5 radians can well be compensated for with each section of lattice network. $\sigma_m/2.5$ determines roughly the number of sections required where σ_m represents the maximum distortion for the total length of line. It is only essential that the total number of sections be included where correction is desired. The corrective structure may be divided and one or more sections located at convenient points throughout the length of the cable whose phase is to be corrected. In the latter case a desirable arrangement on a repeatered cable is to insert at each repeater point a sufficient number of sections to correct the distortion on the cable length between two successive repeaters.

If the network is connected directly to the line, i.e., without the interposition of a unilateral element such as a vacuum tube, the necessary condition

$$K(\omega) = R,$$

where R is a constant representing approximately the characteristic impedance of the cable, imposes one limitation upon the constants L and C, leaving one other to be fixed by the phase. For this condition, put

$$\phi_a(\omega) = -\sigma_a,$$

where σ_a is the value of $\sigma(\omega)$ at a frequency $f_a = \omega_a/2\pi$ near the upper limiting frequency of the correction range and at which it is desirable to have exactly $\sigma(\omega) = -\phi(\omega)$. Substituting for K and ϕ_a in (27) and (28) and solving, gives

$$\begin{aligned} L &= -\frac{2R}{\omega_a}\tan\frac{1}{2}\sigma_a, \\ C &= -\frac{2}{\omega_a R}\tan\frac{1}{2}\sigma_a. \end{aligned} \tag{31}$$

An application of this method of design to the 19-gauge H-44-25 cable is shown in Figs. 17 and 18. This design requires two sections of lattice network at each repeater point of constants

$$\begin{aligned} L &= .116 \text{ henry}, \\ C &= .181 \text{ microfarad} \end{aligned}$$

per section, the distance between repeaters being 50 miles.

While the design above effects considerable improvement, it is

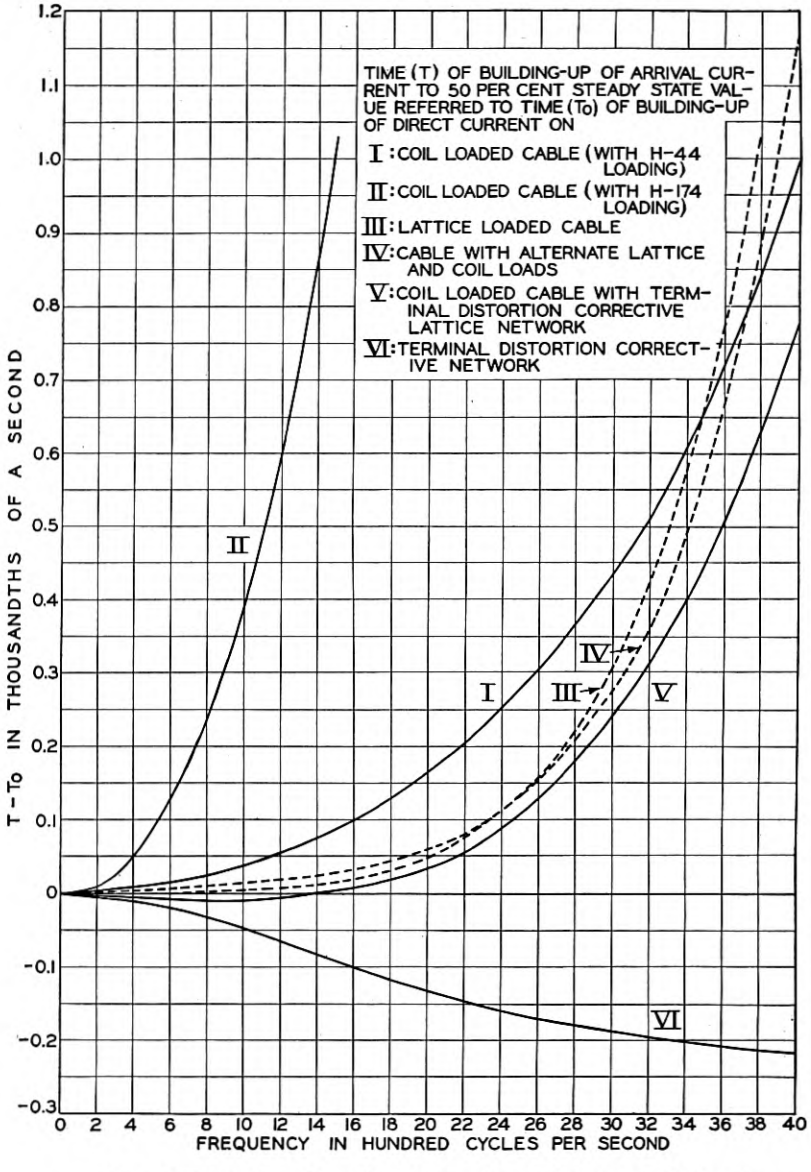


Fig. 19—Transient distortion on 50 miles of 19 gauge loaded cable

evident from a consideration of the duration of the transient distortion that the correction is not perfect, although it is appreciably better than that afforded by either lattice loading or a combination of lattice and coil loading. In Fig. 19 the delay in the time of building-up of any frequency $\omega/2\pi$ over the time of building-up of a direct current is shown for these different systems. Since the physical desideratum is a minimum constant delay for all frequencies, and the delay is quite sensitive to small deviations from the distortionless steady state phase angle, the former is probably a better basis of design and comparison than the latter.

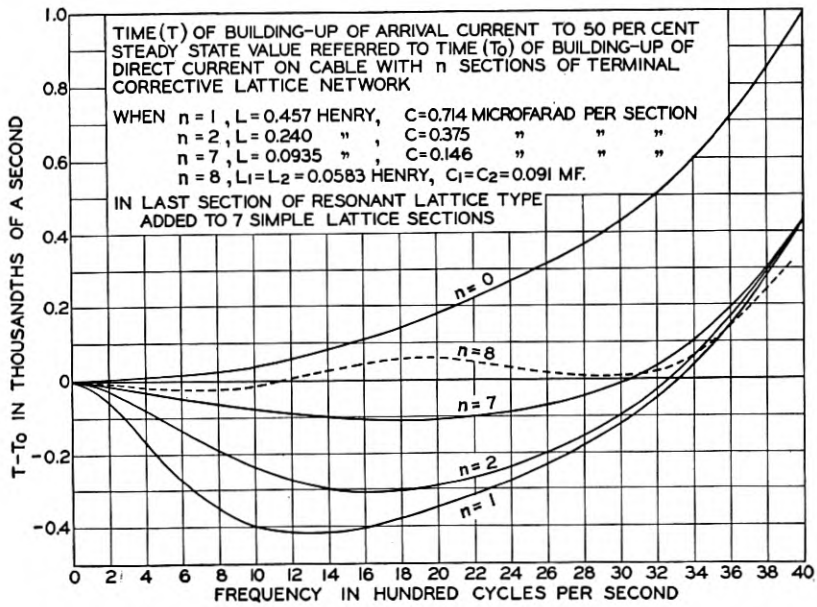


Fig. 20—Transient distortion on 50 miles of 19 gauge light loaded cable

Suppose that it is required to reduce the delay at 4,000 cycles to the value or to less than the value of the delay on the uncorrected cable at 3,000 cycles. The procedure will be to solve equation (30) for $\sqrt{LC}$ in terms of N and the required value of $T - T_0$ at the given frequency $\omega_a/2\pi = 4{,}000$, substitute these values in equation (30) and compute $T - T_0$ over the essential frequency range. It is immediately apparent from Fig. 20 that the improvement at the higher frequencies is at the expense of the intermediate frequencies where the delay is reduced too much. This disadvantage is lessened by increasing the number of sections but this method is uneconomical

and only partially successful because a saturation point is soon reached in the gain obtained with more apparatus.

This difficulty may be overcome by adding networks of the more complicated lattice type shown in Fig. 21.[18] This may appro-

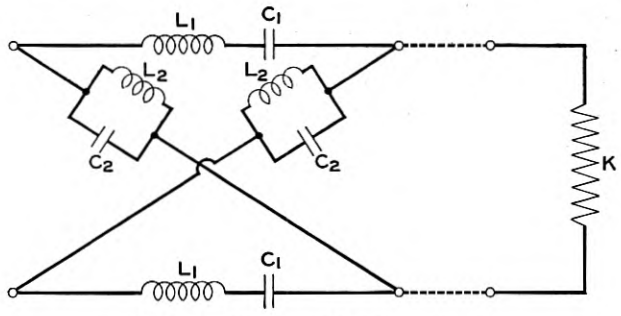


Fig. 21—"Resonant" lattice network

priately be called the 'resonant' lattice type. Its characteristics are most easily derived from the general lattice network having the impedance $z_1/2$ in each series branch and the impedance $2z_2$ in each diagonal shunt branch. Since the characteristic impedance, K, of any section of line is equal to the square root of the product of the open- and closed-circuit impedances and the propagation constant, Γ, per section, is equal to the anti-hyperbolic tangent of the square root of the quotient of the closed-circuit impedance divided by the open-circuit impedance,[19] we have, for the lattice network, in general,

$$\cosh \Gamma = 1 + \frac{2z_1}{4z_2 - z_1} \tag{32}$$

or

$$\tanh \frac{\Gamma}{2} = \frac{1}{2}\sqrt{z_1/z_2},$$

$$K = \sqrt{z_1 z_2}. \tag{33}$$

Thus the requirement that the characteristic impedance be a real constant will, in general, be fulfilled provided

$$z_2 = R^2/z_1, \tag{34}$$

where R is a real constant approximately equal to the characteristic impedance of the cable. This gives

$$\tanh \frac{\Gamma}{2} = \frac{z_1}{2R}. \tag{35}$$

[18] This suggestion was made by H. Nyquist.
[19] See reference 12.

Now if $z_1/2$ is the impedance of a series resonant circuit, we may write

$$\frac{z_1}{2R} = i\omega\frac{b}{2\omega_0} + \frac{b\omega_0}{i2\omega}, \tag{36}$$

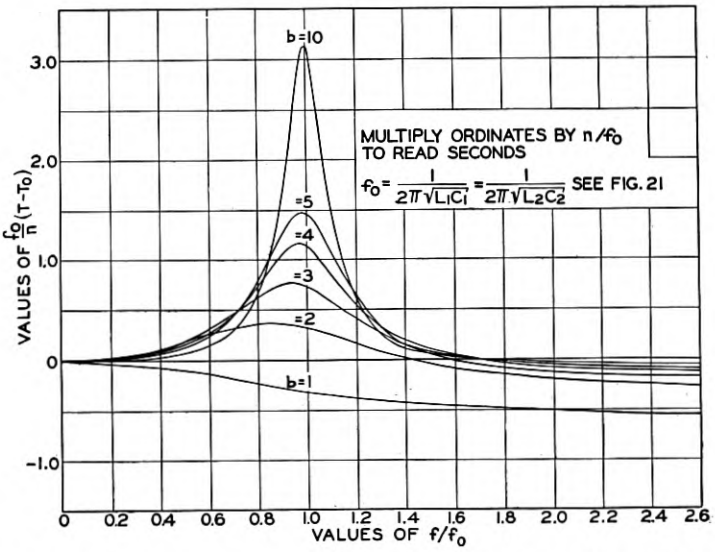


Fig. 22—Time of building-up, T − T_0, for resonant lattice type network of n sections

where b and ω_0 are constants. Then

$$\tanh\frac{\Gamma}{2} = \tanh i\frac{\phi}{2} = i\frac{b}{2}\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right), \tag{37}$$

or

$$\phi = 2\tan^{-1}\frac{b}{2}\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right),$$

$$T = \frac{\frac{b}{\omega_0}\left(1 + \frac{\omega_0^2}{\omega^2}\right)}{1 + \frac{b^2}{4}\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right)^2}, \tag{38}$$

and

$$T_0 = \frac{4}{b\omega_0}. \tag{39}$$

Then, from equations (34) and (36),

$$L_1 = \frac{bR}{2\omega_0}, \qquad C_1 = \frac{2}{b\omega_0 R}, \tag{40}$$

and

$$L_2 = \frac{2R}{b\omega_0}, \qquad C_2 = \frac{b}{2R\omega_0}.$$

The expression for T or $T - T_0$ will have a maximum at $\omega = \omega_0$ if b is large enough. The value of the maximum may be increased by increasing b and its location on the frequency scale may be changed by varying ω_0. A family of curves of $f_0(T - T_0)$ for different values of the parameter b are shown in Fig. 22 with the abscissa f/f_0.

To illustrate the combination of the resonant lattice type with the simple lattice type, add one section of the former with $b = 2$, $f_0 = 2{,}190$, to the seven sections of the latter already applied to the

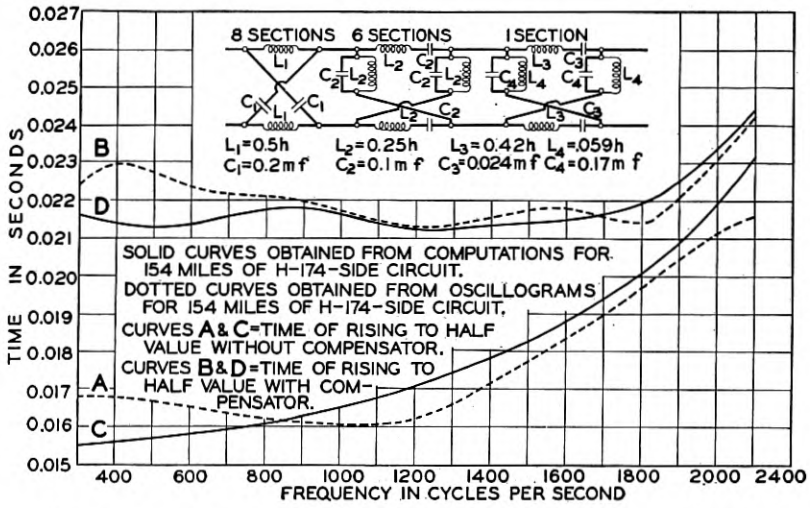


Fig. 23—Transients in loaded lines with and without phase compensator

cable (Fig. 20). The dotted curve for $T - T_0$, which results, is practically zero over most of the frequency range. The value of f_0 was determined so that the delay curve for the resonant lattice type would be complementary to the curve for the remainder of the circuit.

The combination of the two types of lattice network is effective in correcting even the large amount of phase distortion which results in transmission over the medium heavy loaded cable. Fig. 23 (which is reproduced by courtesy of Mr. H. Nyquist) shows a design for use on 154 miles of H-174 side circuit. The experimental results, obtained from oscillograms, are seen to be in close accord with the computed results.

The improvement due to correction of phase distortion by means of the phase compensator circuits employed for picture transmission is illustrated in Figs. 24*a*, 24*b* and 24*c*. Fig. 24*b* is from a negative which was transmitted over a 352-mile H-174 circuit without phase

correctors. Fig. 24*c* is the same print sent over the same circuit after the circuit had been equipped with phase correctors. Fig. 24*a*, which was sent locally, is shown for comparison. Fig. 24*c* is seen to be practically as clear as Fig. 24*a* and both are substantially better than Fig. 24*b*.

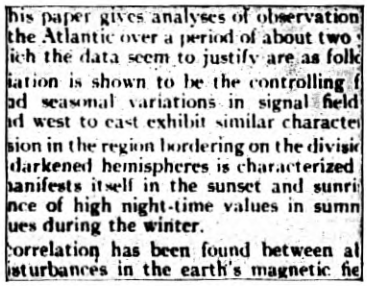

his paper gives analyses of observation
the Atlantic over a period of about two
ich the data seem to justify are as folk
iation is shown to be the controlling f
id seasonal variations in signal field
id west to east exhibit similar characte
sion in the region bordering on the divisi
darkened hemispheres is characterized
anifests itself in the sunset and sunri
nce of high night-time values in summ
ues during the winter.
correlation has been found between al
isturbances in the earth's magnetic fie

Fig. 24*a*—Print transmitted locally

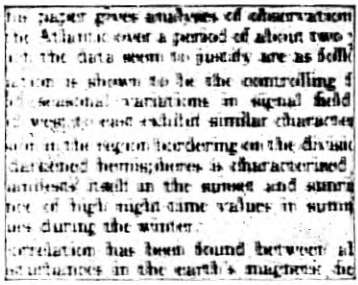


Fig. 24*b*—Print transmitted over 352 mile H-174 loaded cable without corrector

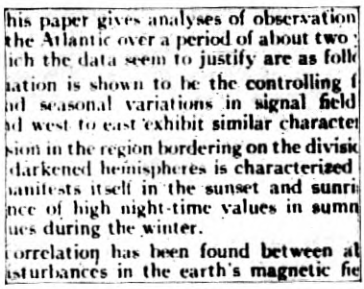

his paper gives analyses of observation
the Atlantic over a period of about two
ich the data seem to justify are as folk
iation is shown to be the controlling f
id seasonal variations in signal field
id west to east exhibit similar characte
sion in the region bordering on the divisi
darkened hemispheres is characterized
anifests itself in the sunset and sunri
nce of high night-time values in summ
ues during the winter.
correlation has been found between al
isturbances in the earth's magnetic fie

Fig. 24*c*—Print transmitted over 352 mile H-174 loaded cable with terminal phase corrector

References

1. "Telephone Transmission over Long Cable Circuits." (A. B. Clark, *Jour. A. I. E. E.*, Vol. 42, No. 1, 1923, and *B. S. T. J.*, Jan., 1923.)
2. "Loading System." (Carson, Clark and Mills, U. S. Patent No. 1,564,201, Dec. 8, 1925.)
3. "Theory of the Transient Oscillations of Electrical Networks and Transmission Systems." (Carson, *Trans. A. I. E. E.*, Feb., 1919.)
4. "Building-up of Sinusoidal Currents in Long Periodically Loaded Lines." (Carson, *B. S. T. J.*, Oct., 1924.)
5. "Distortion-Correcting Circuit." (Carson, U. S. Patent No. 1,315,539, Sept. 9, 1919.)
6. "Receiving System for Telegraphic Signals." (Mathes, U. S. Patent No. 1,586,821, June 1, 1926.)
7. "The Loaded Submarine Telegraph Cable." (Buckley, *B. S. T. J.*, July, 1925.)
8. "The Application of Vacuum Tube Amplifiers to Submarine Telegraph Cables." (Curtis, *B. S. T. J.*, July, 1927.)
9. "Electric Circuit Theory and the Operational Calculus." (Carson, McGraw-Hill Book Co., 1926, 1st ed., p. 185.)
10. "Attenuation Equalizer." (Hoyt, U. S. Patent No. 1,453,980, May 1, 1923.)
11. "Electrical Network and Method of Transmitting Electrical Currents." (Zobel, U. S. Patent No. 1,603,305, Oct. 19, 1926.)
12. "Physical Theory of the Electric Wave-Filter." (Campbell, *B. S. T. J.*, November, 1922.)
13. "Maximum Output Networks for Telephone Substation and Repeater Circuits." (Campbell and Foster, *Trans. A. I. E. E.*, Vol. 39, 1920, p. 258.)
14. "Die Technik der Telegraphie und Telephonie in Weltverkehr." (Lüschen, *E. T. Z.*, July 31, 1924.)
15. "Über Einschwing Vergänge in Pupinleitungen und ihre Verminderung." (Küpfmüller und Mayer, Wissenschaftlich Veröffentlichungen aus dem Siemens-Konzern, Vol. V, Sect. 1, 1926.)
16. "Die Erhöhung der Reichweite von Pupinleitungen durch Echosperrung und Phasenausgleich." (Küpfmüller, *E. N. T.*, Vol. 3, No. 3, 1926.)

High-Speed Ocean Cable Telegraphy

By OLIVER E. BUCKLEY

SYNOPSIS: The invention of permalloy and its application to submarine cables have led to the installation of transoceanic cables of many times the traffic-carrying capacity of the former non-loaded cables. This paper relates briefly the history of the development of permalloy-loaded cables and discusses certain outstanding problems concerned with their design, construction and operation. In a concluding general survey the field of usefulness of loaded submarine telegraph cables is considered.

To a considerable extent the paper is a critical summary of material previously published by members of the staff of the Bell Telephone Laboratories. Its scope is indicated by the sub-titles as follows:

Loaded Cables Now in Service
Historical Remarks
Permalloy and Its Application to Cables
Principles of Design of Loaded Cables
Principles Involved in Operation
Apparatus for Restoration of Signals
Apparatus for Automatic Operation
Electrical Measurements of Loaded Cables
A General Survey

VOLTA devised his famous pile in 1799. Less than 60 years later, in 1858, the first telegraph message was sent over an Atlantic cable. Now nearly 70 years have passed since the remarkable feat of transatlantic telegraphy was first accomplished. Although the art of cable telegraphy may therefore be considered old, it cannot be said ever to have stopped growing. At all periods of its growth it has offered an interesting field for technical endeavor. An added interest was attached to it a little more than 25 years ago when Marconi, by his famous demonstration of transatlantic radio telegraphy, introduced a competitor. With the birth of this new child of science arose the question as to whether the art of telegraphing over cables would not ultimately die. But radio too required time to grow, and it is only within very recent years that there has been occasion for serious concern as to the future of the older art. Now a new advance has been made on the side of the cables and the race for supremacy in transoceanic communication has taken a new turn. The advance to which I refer is the introduction of the high-speed permalloy-loaded cable, and it is with regard to this advance that I wish to speak.

My object is to tell briefly what has been accomplished with cables of the permalloy-loaded type and to describe some of the outstanding features of development which have led to this accomplishment. No attempt will be made in what follows to discuss all phases of cable design and construction, but my remarks will be confined principally to those aspects of cable telegraphy with which the work in the Bell Telephone Laboratories has been concerned.

Loaded Cables Now in Service

There are at present seven high-speed ocean cables of the permalloy-loaded type in operation. Together they have a length of nearly 15,000 miles, which represents about five per cent of the total ocean cable mileage of the world. Their location and lengths are shown on the map in Fig. 1. With the exception of the Cocos Island-Perth (Australia) cable of the Eastern Extension Telegraph Company, these loaded cables are comprised in three transoceanic lines, two crossing the Atlantic and one crossing the Pacific.

The first loaded ocean telegraph cable was the New York-Horta (Azores) cable of the Western Union Telegraph Company which was laid in September 1924. The great success attained with it led to the installation of others, among which was the 1926 Horta-Emden cable of the Deutsch Atlantische Telegraphengesellschaft. The New York-Horta-Emden line thus formed provides not only for carrying a large volume of messages between America and Germany, but also gives a connection with the Italian cable at Horta. This line is now provided with a 5-channel multiplex printing telegraph equipment which is operated at a speed of about 1500 letters per minute and is expected ultimately to be operated at a considerably higher speed. Four of the five channels of this line provide direct communication between New York and Emden, and the fifth serves for messages relayed at Horta.

A second transatlantic line is formed by the New York-Bay Roberts and Bay Roberts-Penzance cables of the Western Union Telegraph Company which were laid in 1926. Each of these cables is capable of carrying more than 2500 letters per minute, but at present this line is being operated at only about one half that speed. The construction of operating equipment to realize the full 2500 letters per minute is now in process. The combined traffic-carrying capacity of these two transatlantic loaded lines is nearly as great as that which was previously provided by the sixteen older non-loaded cables which served to connect North America with Europe prior to 1924.

The Pacific Cable Board, also in 1926, installed loaded cables to parallel its non-loaded cables of 1902, connecting Bamfield, Fanning Island and Suva. A speed of over 1200 letters per minute has been reported for each section of this transpacific line. This speed is nearly four times that which was afforded by the older non-loaded cables over the same route.

Such an extension of facilities for transoceanic communication would be expected to have a pronounced effect on both the cost of

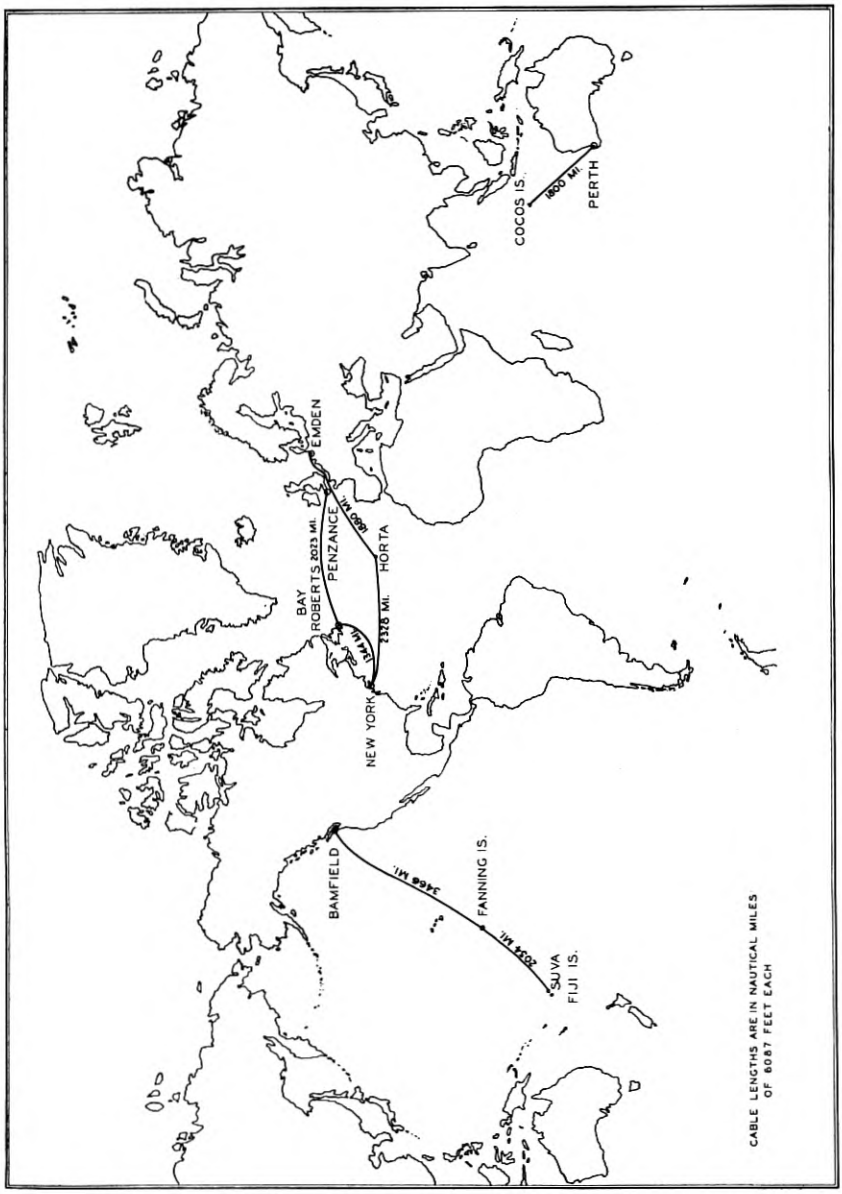


Fig. 1—Map showing location of loaded ocean cables

communication and the amount of communication between continents and, indeed, such an effect has already been experienced, significant reductions in rates having been made within the past year. In view of what has already been accomplished it seems not unlikely that the further introduction of loaded cables will completely revolutionize the whole transoceanic communication situation.

Historical

Like nearly all other important technical advances, that of the loaded telegraph cable is the outgrowth of contributions of many investigators and engineers working in different fields of endeavor. The history of the development of the idea of inductively loading transmission lines, from its original conception to recent times, is well known and need not be gone into here. An excellent theoretical analysis of the problem of transmitting signalling impulses over an ocean cable has been made by Malcolm, who in 1917, in his book on "The Theory of Submarine Telegraph and Telephone Cables," went so far as to predict that heavy continuous loading was the next great advance in the telegraph-cable art to be expected.

The practical accomplishment of the permalloy-loaded cable came about as a result of researches conducted in the laboratories of the American Telephone and Telegraph and Western Electric Companies, now known as the Bell Telephone Laboratories. Our interest in the problems of submarine cables was a natural part of our interest in all phases of electrical communication and was at first concerned principally with the application of vacuum tube amplifiers to ordinary telegraph cables. Considerable progress in the development of amplifiers for this purpose was made in the laboratory as early as 1914. The great demand for cable communication which the World War brought about led to further activity in this direction and to extensive tests of vacuum tube amplifiers on the cables of the Western Union Telegraph Company.

From these tests it was found that although the vacuum tubes would provide any desired amplification of signal strength and although by the combination of vacuum tubes and suitable electrical networks the distortion of the received signals could be corrected to any desired degree, relatively little actual gain in traffic capacity could be obtained by these means, since the real limit to cable speed was not distortion but interference. Although some improvement in cable speed could have been achieved by refinement of means for duplex operation and by improved means to eliminate extraneous interference, it was quite apparent that to obtain any great advance over the existing art would require a modification of the cable itself.

For over fifty years the cable had remained substantially unchanged in character, though great advances had been made in methods of operation. Inductive loading, which was proposed by Heaviside in 1887, was the obvious means for obtaining increased cable speeds, but no one had found a way to realize the advantages of loading as applied to ocean cables. Loading with evenly spaced coils as proposed by Pupin and used on land lines presented difficulties in laying and maintenance which practically prohibited this method. Continuous or Krarup loading by a wrapping of iron wire around the conductor of the cable was mechanically feasible but the amount of inductance which could be obtained in this way was not sufficient to justify its use. In order to make continuous loading advantageous for long ocean cables there was needed a material which could be much more easily magnetized than iron. Fortunately we had at hand as an aid to solving this problem the extraordinary magnetic material, permalloy, an alloy possessing magnetic permeability many times that of iron, even at the low magnetizing forces produced by the feeble currents of a telegraph cable. It is on permalloy that the loaded cable depends primarily for its success.

Although permalloy provided the means to give the cable the desired high inductance, the mere wrapping of this metal around the copper conductor of the cable was far from providing a practical solution of the problem of high-speed ocean telegraphy. To achieve this solution required the solution of very many subsidiary problems, concerned not only with the making of a cable but also with the transmission of signals over it and with the means for its practical operation. Work on all these phases of the problem of the loaded cable was actively pursued in our laboratories and in the field from the time of the first proposal, made in July 1919, to load a trans-oceanic cable with permalloy to the successful completion and operation of the New York-Horta cable.

During the first two years of this period our investigations were conducted wholly in the laboratory where hundreds of experimental lengths of loaded conductors were made and tested to convince ourselves that a permalloy-loaded cable could be manufactured and laid successfully. During the same period studies of the signal distortion of a loaded artificial cable and means for correcting distortion were carried on. Simultaneously methods of high-speed operation of loaded cables were developed, with the result that we were convinced from our laboratory experiments, not only that a permalloy-loaded cable could be made and laid successfully, but that it could also be operated commercially at the high speeds which we had predicted.

Having gone this far in the laboratory, it was decided to bring the results of our investigations to the attention of one of the cable operating companies with the object of securing a practical trial of a permalloy-loaded ocean cable.

On being shown what could be accomplished with permalloy loading, the Western Union Telegraph Company was quick to take advantage of this means of extending its cable facilities, and shortly thereafter arrangements were made whereby the Telegraph, Construction & Maintenance Company, Ltd., was to manufacture a cable for the Western Union Telegraph Company, using permalloy loading material supplied by the Western Electric Company and applied and treated under the direction of Western Electric engineers.

As a part of this undertaking, it was decided that prior to laying a complete transoceanic length of cable it would be desirable to make, lay and test a shorter length in order to obtain experience in manufacture and a test of its mechanical and electrical properties after it had suffered the extreme treatment to which a deep-sea cable is subject in laying. Accordingly, for such an experiment, 120 miles of cable of the same type which it was proposed to use for a transoceanic length was laid in a loop from the south shore of Bermuda in October 1923. Very thorough tests were made jointly by Western Electric and Western Union engineers to determine whether the electrical characteristics had been affected by laying and what attenuation and distortion were actually produced by such a cable. The results obtained were in excellent agreement with our predictions, and accordingly manufacture of the full 2300 miles required to connect New York and Horta was at once undertaken.

The New York-Horta cable which, like the 120-mile trial cable, was manufactured by the Telegraph, Construction & Maintenance Company with permalloy loading material applied and treated under the technical direction of Western Electric engineers, was laid in September 1924. Within an hour after the cable had been turned over to our engineers for test, a speed of 1500 letters per minute was obtained with the terminal apparatus which had been designed and provided in advance. In this case the messages were received on a high-speed siphon recorder of special design. Shortly thereafter, with the same apparatus, a speed of over 1900 letters per minute was obtained.

The speed of the cable having been demonstrated, commercial operation was quickly established with temporary operating equipment utilizing siphon recorders in conjunction with vacuum tube amplifiers. This type of operation was continued for about two

years during which the engineers of the Western Union Company and the Bell Laboratories worked together on the development of a multi-channel printing telegraph system which would adapt the operating methods previously developed by the laboratories to the needs of the telegraph company. In October 1926 the present five-channel printing telegraph apparatus was put into use.

Demands for other high-speed loaded cables quickly followed the successful demonstration of the New York-Horta cable. The Western Union Company arranged for the manufacture of the cables for the New York-Bay Roberts-Penzance route by the Telegraph, Construction & Maintenance Company, Ltd. On these cables permalloy supplied by the Western Electric Company was again used. The Norddeutsche Seekabelwerke A-G. arranged with the Western Electric Company for the supply of permalloy loading material and for technical assistance to manufacture the Horta-Emden cable for the Deutsch Atlantische Telegraphengesellschaft. The Pacific Cable Board arranged to have the cables for its Bamfield-Fanning Island-Suva line manufactured by two British companies and obtained licence therefor from the Western Electric Company. The shorter section from Fanning Island to Suva was made by Siemens Brothers & Co., Ltd., with permalloy supplied by the Western Electric Company and applied and treated with the technical direction of their engineers. The long section from Bamfield to Fanning Island was made by the Telegraph Construction & Maintenance Company, Ltd. All of these cables were laid in 1926.

Permalloy and Its Application to Cables

Permalloy, which made possible this radical change in the cable art, is the invention of G. W. Elmen of the Bell Telephone Laboratories. Since descriptions and explanations of some of its properties have already been given in papers by various members of the Bell Laboratories' staff,[1] the present discussion will be limited to a brief statement of its outstanding characteristics which are of consequence in connection with its use on cables.

The name "permalloy" has been applied to alloys of iron and nickel of more than about 30 per cent nickel content characterized by extraordinarily high magnetic permeability at very low magnetizing

[1] H. D. Arnold and G. W. Elmen, *Jour. Franklin Inst.*, Vol. 195, pp. 621–632, May 1923; *B. S. T. J.*, Vol. II, No. 3, p. 101.
O. E. Buckley and L. W. McKeehan, *Phys. Rev.*, Vol. 26, pp. 261–273, Aug. 1925.
L. W. McKeehan, *Phys. Rev.*, Vol. 26, pp. 274–279, Aug. 1925.
L. W. McKeehan and P. P. Cioffi, *Phys. Rev.*, Vol. 28, pp. 146–157, July 1926.
L. W. McKeehan, *Phys. Rev.*, Vol. 28, pp. 158–166, July 1926.

forces. The manner in which the initial permeability of these alloys varies with composition, when heat-treated in a particular way, is shown in Fig. 2, which is taken from the Arnold and Elmen paper. The magnetic properties of the alloys of this series depend in an extraordinary degree on their previous mechanical and thermal history. In general, high initial permeability is obtained by rapid cooling after a thorough softening of the metal by heating at a high temperature,

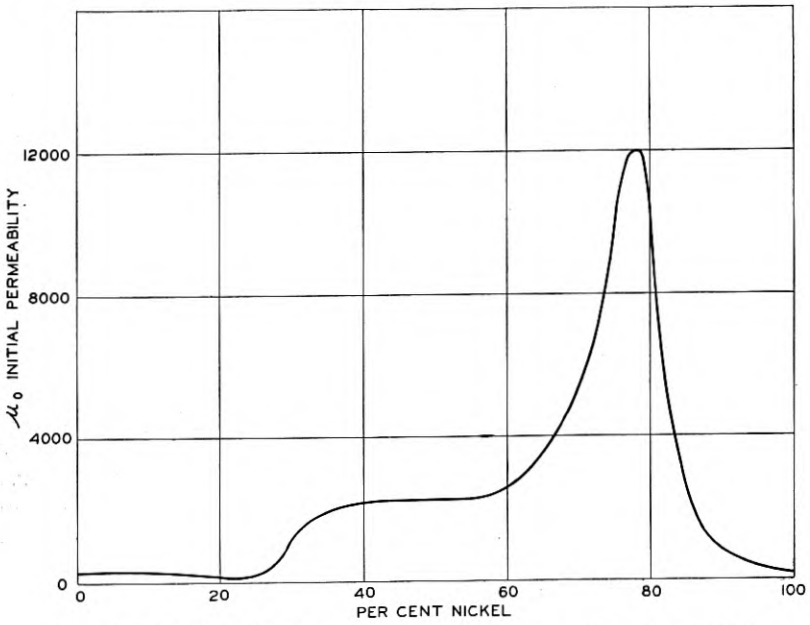


Fig. 2—Variation of initial permeability with composition of permalloy

this effect of rapid cooling being particularly marked on the compositions in the region of 80 per cent nickel. By control of the composition and heat treatment an initial permeability of more than 12,000 has been obtained with an alloy of $78\frac{1}{2}$ per cent nickel and $21\frac{1}{2}$ per cent iron, whereas iron or nickel alone ordinarily have initial permeabilities of only about 200 or 300. It is the high initial permeability of permalloy that is most important in its use on cables, though such an initial permeability as 12,000 would be even higher than is generally desired for a telegraph cable. For use on cable conductors permeabilities of the order of from 2000 to 5000 have been desired and obtained in practice.

Another important property of permalloy with regard to its use on cables is its resistivity, since high resistivity prevents excessive eddy-

current loss. The resistivity of the whole nickel-iron series of alloys is higher than that of either iron or nickel. By adding a third element, for example chromium, to the nickel and iron and keeping the ratio of nickel to iron about 4 : 1, a combination of very high resistivity and very high initial permeability may be obtained in the same alloy.

The permalloy used in the New York-Horta cable contained about 79 per cent nickel and 21 per cent iron with a small amount of manganese to make it more malleable. The permeability of this alloy as used on the New York-Horta cable was about 2300, its resistivity being about 16 microhm-cms. On the Horta-Emden cable, the New York-Bay Roberts-Penzance cables and the Fanning Island-Suva cable permalloy containing about 80 per cent nickel, 17.5 per cent iron, 2 per cent chromium and 0.5 per cent manganese was used. With this alloy an initial permeability of about 3700 was obtained. Its resistivity is about 38 microhm-cms.

The permalloy loading material used on the New York-Horta cable was in the form of a thin tape 0.006 inch (0.015 cm.) thick and 0.125 inch (0.32 cm.) wide applied in a closely wound helix surrounding the conductor. On the Horta-Emden cable tape 0.0059 × 0.098 inch (0.015 × 0.25 cm.) was used. On the New York-Bay Roberts and Bay Roberts-Penzance cables the tape thickness was 0.0055 inch (0.014 cm.) and the widths were 0.079 inch (0.20 cm.) and 0.123 inch (0.31 cm.) respectively. On the Fanning Island-Suva cable the permalloy is in the form of a wire of 0.011 inch (0.028 cm.) diameter applied in a single closely laid helix. The northern section of the Pacific cable from Bamfield to Fanning Island is reported [2] to be loaded similarly with "Mumetal" wire of 0.010 inch diameter, made by the Telegraph, Construction & Maintenance Company.

Very good results have been obtained with both tape and wire loading. The tape has the advantages of costing less to apply and of possessing greater mechanical strength, whereas the wire has the advantages of lower eddy-current loss and of being less affected by the earth's magnetic field. With either wire or tape loading the component of the earth's magnetic field parallel to the cable sets up magnetic induction in the helical loading material and consequently reduces its effective permeability for the small magnetizing forces of the signalling current. This reduction of effective permeability by the earth's magnetic field is greater, the greater the angle of lay with which the loading material is applied. Consequently this effect is generally greater with tape loading than with wire loading. Whether

[2] E. S. Heurtley, *Electrician*, Vol. 98, pp. 348–350, Apr. 1, 1927. See also *P. O. Elec. Engrs. Jl.*, Vol. 20, pp. 36–40, Apr. 1927.

tape or wire should be used is, in the end, an economic problem since any disadvantage of one with regard to the other may be compensated for by increasing the size of the copper conductor.

Permalloy has another property which it is important to consider in connection with its use on cables, namely, its great sensitiveness to mechanical strain. Strain of deformation applied to it will modify its magnetic characteristics, and very great changes in its permeability for small magnetizing forces may be produced by strains well within the mechanical elastic limit. Consequently in making the cable it is necessary to insure that the permalloy shall be as free as possible from strains of deformation. There are two principal ways in which the permalloy used for loading may be subject to such strains. The first comes in the manufacture of the loaded conductor and the second in the laying of the cable.

Since permalloy is so strain-sensitive it must be annealed after it has been applied to the conductor. Accordingly the hard-worked metal is wrapped around the copper conductor and the conductor is thereafter passed continuously through a furnace, maintained at approximately 900° C., and from the furnace into a cooling tube. The lengths of the furnace and cooling tube and the rate of passage of the conductor are so chosen as to insure that the loading material will get the necessary softening in the furnace and will be cooled at the proper rate in the cooling tube. Even though the permalloy is thus annealed on the conductor, it still might well be subject to considerable strain, since the copper, on being heated to such a high temperature, expands more than the permalloy and tends to weld to it and, on contracting, would bend the permalloy tape near the spots where welding occurs. To prevent this action the loading material is applied very loosely and means are taken to prevent adhesion of the permalloy to the copper. In spite of the great sensitiveness of the permalloy to mechanical strain, the loaded conductor after heat treatment stands ordinary handling very well without much loss of permeability. However, if it were insulated by the methods which have been used in the past in making deep-sea cables, it would lose much of its inductance on laying on account of the effect of the great pressures to which a cable is subjected.

To prevent reduction of the permeability and consequent loss of inductance on laying, it is necessary to provide that pressure on the insulating material shall produce only true hydrostatic pressure on the permalloy with no tendency to deform it. This result has been accomplished by vacuum-impregnating the permalloy-loaded conductor with a semi-fluid compound which fills all the interstices of

the conductor and also forms a layer a few thousandths of an inch thick on the outside of the loading material. The gutta-percha insulation may then be extruded over the impregnated conductor with the assurance that the semi-fluid compound will serve to equalize the pressure on the permalloy. Numerous compounds have been proposed and used for this purpose, that on the New York-Horta cable being of an asphaltic type. It is essential, of course, that this compound be sufficiently viscous at temperatures at which the gutta-percha is applied to permit extruding the gutta-percha around it and that it will also be sufficiently fluid at the temperature of the sea bottom, which may be as low as 2° C., to permit readjustment of the pressure on the permalloy. When a loaded conductor insulated in this manner is subjected to high pressures at low temperatures, it

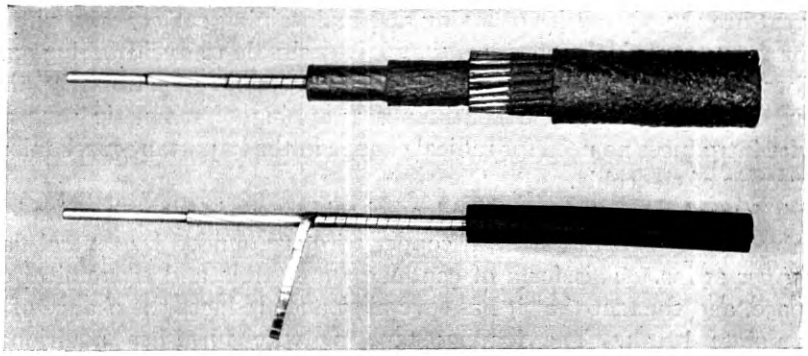


Fig. 3—Permalloy-loaded cable. Above, section of deep-sea type from New York-Horta cable. Below, section of core showing permalloy tape partly unwound

may be found that the inductance drops when the pressure is first applied but in a few minutes the compound flows so as to equalize the pressure on the permalloy and the inductance quickly comes back to its original value.

Outside of the insulated conductor or "core" of the cable the permalloy-loaded cables which have been made are quite like ordinary non-loaded cables, so there is no need to go into the further details of cable construction here. Fig. 3 shows a section of the deep-sea portion of the New York-Horta cable.

Principles of Design of Loaded Cables

There are two principal aspects of the design of a submarine cable—mechanical and electrical. Mechanically, the cable must be so designed as to insure that the conductor shall be continuous, that its

insulation shall be maintained and that the core, which comprises the conductor and insulation, shall not be damaged in laying or in subsequent repairing operations. Electrically, the cable must be so designed that it will serve properly to transmit signals. Thus the electrical design is concerned with the size of the conductor and the amount and characteristics of the loading material and insulation, whereas the mechanical design is concerned with the mechanical characteristics of the conductor and its insulation and with the jute and armor wire which serve to protect the core and give the cable the necessary strength. These two aspects of design cannot, of course, be considered quite independently of each other and both are in the ultimate analysis controlled by economic considerations. It is convenient for our present purposes, however, to consider them separately.

The mechanical design of cables is a well-established art and so great are the difficulties of laying and maintaining cables, even under the most favorable conditions, that it is desirable to avoid taking any liberties with this phase of cable construction. Fortunately the method of loading by a continuous wrapping of magnetic tape or wire introduces no need for radical change in the important mechanical features of the cable.

The copper conductor of the loaded cable is, as in many non-loaded cables, composed of a central copper wire surrounded by several flat copper strips. This form of conductor is flexible and economical of space, and the fact that it has several strands reduces the chance of a complete break. The loading tape or wire furnishes additional protection in this regard.

The thickness of gutta-percha must be sufficient to insure the integrity of the insulation at all points. It is, in fact, this consideration which established the amount of insulation used on the loaded cables which have been laid, since consideration of the theoretical economic optimum thickness of gutta-percha would in each case have demanded less gutta-percha than is considered safe. In this regard the insulating problem of the loaded cable is like that of the non-loaded cable.

The disposition of jute and armor wire around the core is determined wholly by mechanical considerations as in the case of the non-loaded cable for which the practice is fairly well standardized. Unlike the non-loaded cable, however, the loaded cable, in its electrical behavior, is affected somewhat by the presence and character of the armor wire as will be described later.

As is well known, the electrical behavior of *non-loaded* cables is determined almost wholly by their resistance and capacity and consequently the only important features to consider from the electrical

standpoint have been the size of the conductor and the thickness of insulating material. The electrical design of a loaded cable is, however, somewhat more complicated since in addition to copper resistance and electrostatic capacity we have here to be concerned with the inductance added by the loading material and also with added resistance factors which are introduced by its use. The problem of electrical design, therefore, involves determining not only the size of the copper conductor, but also the electrical and magnetic characteristics and the shape and dimensions of the loading material, as well as the electrical characteristics of the insulating material, which will give the highest speed of operation consistent with the mechanical and cost limitations which are imposed.

Since the object of the electrical design is to secure high operating speed, it is essential to consider what are the factors which limit speed and how they are taken into account. This subject has already been treated in some detail in previous papers,[3] and only a general review of the principal factors involved in the electrical design will be undertaken in the present paper.

In the history of cable development prior to the introduction of the permalloy-loaded cable various physical factors at different times limited the speed of operation which could be obtained with a long ocean cable. These were principally distortion of signals, sensitivity of receiving apparatus, limited safe sending voltage, inaccuracy of duplex balance, and extraneous interference from both natural and man-made sources. With the development of cable amplifiers and of improved means of signal shaping, the factors of distortion and limited sensitivity of receiving apparatus were effectually eliminated and at the time when the development of the permalloy-loaded cable was undertaken the speed of long cables was limited in most cases by the accuracy with which artificial lines could be made to balance cables in duplex operation. In some cases where extraneous interference was unusually severe the limit of speed was set by that factor combined with the limit of sending voltage which was usually placed at about 50 volts by extreme concern for the safety of the cable insulation.

It was by no means obvious which of these several factors should be considered in the electrical design of a loaded cable. With the vacuum-tube amplifier available to amplify the weak received signal to the degree necessary to operate recording mechanisms, there was no practical limit to the sensitiveness of receiving apparatus. It was, however, necessary to consider distortion as a possible limit to speed.

[3] O. E. Buckley, *Jour. A. I. E. E.*, Vol. XLIV, pp. 821–829, August 1925, *B. S. T. J.*, Vol. IV, No. 3, pp. 355–374, July 1925; J. J. Gilbert, *B. S. T. J.*, July 1927.

It is interesting to note in this connection that, though, in most previous proposals to load long telegraph cables, loading had been advocated primarily as a means of reducing distortion, practical consideration of the problem uncovered new types of distortion which were absent in the non-loaded cable. The nature of distortion of signals by a non-loaded cable was well understood, the problem having been solved long ago by Lord Kelvin. The distortion of a loaded cable is a much more complex affair since there are involved in it not only the effects of distributed inductance, capacity, resistance and leakance of the ideal cable for which the distortion is readily calculable, but also the factors of change of inductance and resistance with frequency and current, and the effects of magnetic hysteresis which are unavoidable in a practical loaded cable. Though the effect of these factors on distortion could be approximated by theoretical analysis it was considered necessary to have experimental proof that a signal could be restored in shape after passing over a loaded cable and it was primarily on this account that tests were made with an artificial loaded line. These tests showed that even the distortion of a loaded cable could be corrected by using suitable terminal networks in connection with the vacuum tube amplifier.

With the factor of distortion thus eliminated there remained duplex balance, sending voltage and received interference as possible limits to the speed of the loaded cable.

Duplex balance would, of course, set the limit of speed of operation if the cable were to be operated simultaneously in two directions as is commonly done with non-loaded cables, since it would obviously be more difficult to build an artificial line electrically equivalent to a loaded cable with its variable inductance and resistance than one equivalent to a non-loaded cable in which only resistance and capacity have to be considered. Even with non-loaded cables the difficulty of balancing is so great that the double-duplex speed is usually much less than twice the possible simplex speed and with the loaded cable, which is more difficult to balance, the relative gain in traffic capacity to be obtained by duplexing is certain to be less than with non-loaded cables. On the other hand, simplex, or one-way, operation offers very great advantages especially when used in connection with automatic operation, since it disposes of the necessity for an intricate and costly artificial line and permits dividing the full traffic capacity of the cable most efficiently to accommodate the traffic it must carry, which with most transoceanic cables is usually unequal in the two directions. For these reasons it was decided to design the first loaded cable primarily to secure efficient simplex operation. Subsequent

experience has well justified this procedure for the cables which have been made.

The problem of designing a loaded cable was thus reduced to proportioning its component parts so as to secure the desired speed of operation under the conditions imposed by the limitations of sending voltage and received interference. Considerations of safety limit the sending voltage to about 50 volts, and terminal interference as ordinarily experienced requires that the received signal shall have an amplitude of a few millivolts. The risk of increasing the sending voltage to several hundred volts would not necessarily be serious but little advantage could be gained by taking this risk since, with the materials and type of construction used, higher sending voltage would involve increased hysteresis and eddy-current losses and consequently would not result in a proportionately higher received voltage. It is, however, possible to reduce the received interference by proper termination and this is of great importance in cases where the interference is severe.

The nature of cable interference and methods of reducing it have been discussed in a paper by J. J. Gilbert [4] in which is described the method which has been used to decrease the terminal interference on the loaded cables which have been laid. This method consists in using, as the earth connection for the receiving apparatus, a "balanced" sea-earth, terminating in deep water. With ordinary cables the common practice has been to provide as the earth connection a sea-earth core, similar to the main core and sheathed with it, but extending only a few miles from shore to a point where the sea-earth conductor is connected to the sheath of the cable. While this type of earth greatly reduces the interference picked up in and near the cable terminal, it does not completely eliminate it. Almost complete elimination of the effects of disturbances originating between the termination of the sea-earth core and the shore may be obtained by providing a terminal impedance between the sea end of the sea-earth conductor and the sheath of the cable. For a non-loaded cable a combination of condensers and resistances would be required to make up such a terminal impedance, but for the loaded cable a very close approximation is secured by a simple resistance of a few hundred ohms. A few hundred feet of manganin wire, insulated like the rest of the conductor and joined to the end of the sea-earth core, serves this purpose admirably. This type of construction has been used on the New York end of the New York-Horta and on all terminals of the

[4] J. J. Gilbert, *B. S. T. J.*, Vol. V, No. 3, pp. 404–417, July 1926. See also *Electrician*, Vol. 97, p. 152, August 1926.

New York-Bay Roberts-Penzance cables, on both ends of the Horta-Emden cable, and it has also been used in the loaded cables of the Pacific Cable Board.

With the maximum sending voltage determined and with the received voltage necessary to work through interference known, the cable can be designed to give the desired speed of operation. More specifically it is necessary to provide that the attenuation for frequencies essential to the formation of the signal shall be materially less than the attenuation corresponding to the ratio of the sending voltage to the interference at the receiving end. This condition can be met by establishing the attenuation of the cable for one particular frequency related to the speed of signalling. The relation between this frequency and the speed in letters per minute depends of course on the code and method of operation used. In the case of the New York-Horta cable the fundamental frequency of a series of alternate dots and dashes of the cable code, that is, one half the center hole frequency, was used as a basis for design. For this frequency a voltage attenuation of e^{-10}, corresponding to 87 TU, can be safely assumed for recorder operation under conditions of interference such as are encountered on the New York-Horta cable. With the Baudot type of code and using the most improved apparatus, that is including a synchronous vibrating relay, a voltage attenuation of $e^{-9.5}$, corresponding to 82 TU, may be assumed for the frequency resulting from assigning 1.25 cycles to a character of the Baudot code.

The computation of the attenuation of a loaded cable requires, of course, only the substitution in the ordinary telegraph equation of the specific values of inductance, capacity, resistance, leakance and frequency which apply to the particular cable in question. The method of calculation of these electrical quantities has been discussed in previous papers and need not be repeated here.

The design of the cable is thus reduced to proportioning the elements of its construction so as to obtain the most economical cable of a given attenuation at a given frequency. The thickness of insulating material is, as has been noted above, determined practically by mechanical considerations. The electrical characteristics of the insulating material are effectively limited by the quality of gutta-percha, account being taken of its dielectric leakance which is of considerable effect on the behavior of the loaded cable though usually of almost negligible effect on non-loaded cables. With the possibilities of insulating materials thus limited the problem of electrical design reduces practically to determining the size of the conductor and the composition, size and shape of the loading material.

The desirable qualities in the loading material from the electrical point of view are high initial permeability, high resistivity and constancy of permeability in the range of magnetizing forces concerned. The exact composition of permalloy which would give the best combination of these properties would, of course, be different for different cables but for practical reasons it is desirable to choose a composition which approximates the optimum for general use. Having determined on a particular alloy, the optimum size of conductor and thickness of loading material may readily be computed on the basis of its known electrical and magnetic characteristics. With the compositions of permalloy which have been used, the optimum thickness of the layer of permalloy for a long ocean cable generally lies in the range from 0.005 inch to 0.010 inch which is fortunately convenient from the mechanical point of view. If less than the optimum thickness is assumed, the inductance will be too low and the consequent required conductor diameter will be too large. On the other hand, if more than the optimum thickness is assumed, the increase of eddy-current resistance and the effect of dielectric leakance will more than offset the gain due to the increased inductance.

In determining the optimum thickness of the permalloy it is, of course, essential to include all the resistance factors which are of consequence. In addition to eddy-current resistance and the effect of dielectric leakance there are the factors of hysteresis resistance and sea-return resistance which must, in particular, be taken into account.

The effect of hysteresis on attenuation is felt only near the sending end of the cable since over most of the length of the cable the current is so small that the hysteresis is negligible. Its effect near the terminals may be calculated by the method of successive approximations which takes account of the falling off of current and the change of hysteresis resistance with current amplitude. Ordinarily the effect of hysteresis becomes negligible beyond the first one or two hundred miles from the sending terminal. Within that range it may add as much as 10 TU to the total attenuation of the cable for the high-frequency components of the signals.

By sea-return resistance is meant the resistance which is contributed by the sea water and armor wire around the core of the cable. In low-speed non-loaded cables this factor may be safely neglected since the return current of low-frequency signals spreads out through such a great area around the cable that the resistance contributed by the sea water is negligible. With the high-frequency signals of the loaded cable, however, the return current tends to concentrate in the sea water close to the cable and much of it flows in the armor wires.

The result is a loss of energy which introduces resistance in the cable circuit, this resistance being much greater than if the armor wires were absent.[5]

The relations between sent and received voltage for some of the cables which have been laid are shown by the curves in Fig. 4, in which

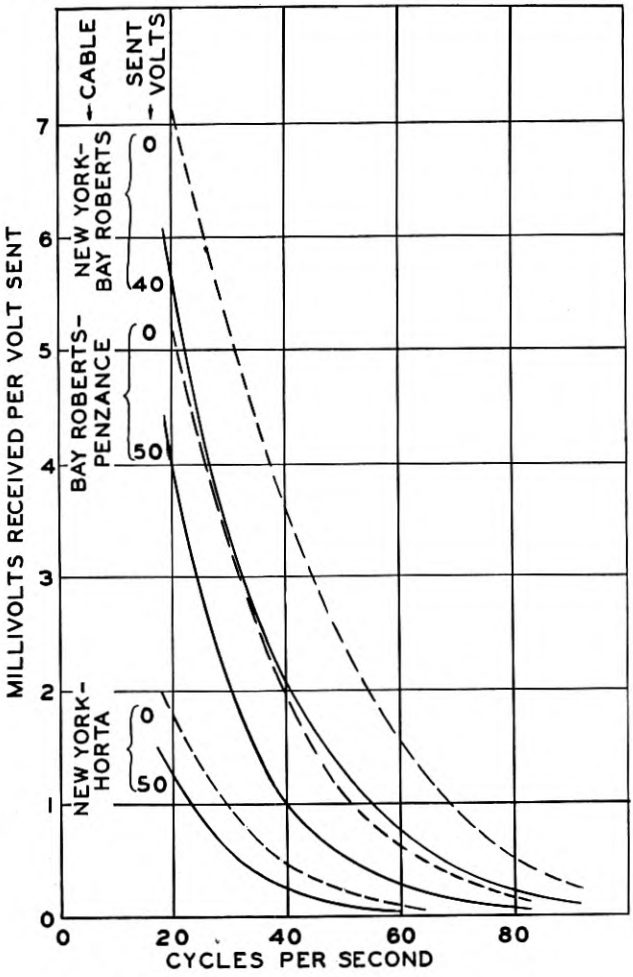


Fig. 4—Received voltage-frequency curves

the dotted curves give these ratios for zero sending voltage, obtained by extrapolation. These curves were obtained experimentally from the laid cables.

[5] See Carson and Gilbert, *Journ. Franklin Institute*, Vol. 192, pp. 705–735, 1921; *Electrician*, Vol. 88, pp. 499–500, 1922; *B. S. T. J.*, Vol. 1, pp. 88–115, July 1922.

Principles Involved in Operation

To realize practically the full benefit of the high speeds of operation of which loaded cables are capable, required the development of new types of terminal apparatus. Although many of the functions performed by the apparatus on loaded cables are similar to those involved in the operation of ordinary cables, many new problems were introduced by the higher speed of the loaded cable and by its peculiar electrical characteristics. Also new means were required to secure efficient two-way working.

With both loaded and non-loaded cables the following steps are involved in operation: translation of messages into signal impulses and the application of these impulses to the cable; correction of distortion or, as it is commonly called, signal shaping; amplification of the feeble received impulses; and reconversion of the restored received impulses into messages. The requirements to be met in accomplishing the first and last of these steps with the loaded cable are different from those in the case of the non-loaded cable, principally on account of the higher speed of the former. The requirements to be met in signal shaping and amplification are different for the loaded cable both because of its peculiar distortion and because of its high speed of operation.

The means commonly employed on non-loaded cables for sending messages involves translation of a message, usually by machine methods, into electrical impulses of the standard cable code in which a dot of the continental Morse code is represented by a positive impulse of definite duration and a dash is represented by a negative impulse of the same duration. A train of impulses of equal length but of varying polarity is thus applied to the cable at the sending end. This train of impulses is distorted and greatly attenuated by the cable but is partially restored in shape and size by terminal apparatus and is finally received on a siphon recorder which makes a record in the form of a wavy line on a paper strip. In the form and spacing of the humps and depressions of this wavy line an expert operator recognizes the positive and negative impulses which were applied at the sending end and which he is able to translate into the original message.

The necessary correction of distortion of signals on ordinary cables is accomplished by simple electrical networks at the terminals. Advantage is also taken of the mechanical characteristics of moving coil instruments. The fundamental principles[6] involved may be

[6] For a more detailed discussion of the principles of correction of distortion as applied to non-loaded cables see J. W. Milnor, *Jour. A. I. E. E.*, Vol. XLI, pp. 118–136, 1922.

roughly summed up as follows. For a line to transmit signals without distortion it would be necessary for all frequency components of the signals to be attenuated to the same degree and also for the delay or time-lag of transmission to be the same for all these frequencies. Legible signals can, however, be received if the attenuation of the combination of cable and apparatus increases with frequency, provided the increase in attenuation up to a certain value of frequency is not too great. Frequencies higher than this value are not required to form the received signal. For example, in the case of cable-code operation, a legible signal will be received if the attenuation at 1.5 times the fundamental or dot frequency is as much as 5 or 10 times the attenuation at the lowest frequencies involved in the signal, and if still higher frequency components are reduced to an inappreciable amplitude as a result of transmission through the system. The dots and dashes of the received signal will in this case be recorded as rounded but readily recognizable humps or depressions in the line traced on the siphon recorder strip.

Now the attenuation of the cable for the various frequency components is not uniform but increases rapidly with frequency. For example, on a particular transatlantic non-loaded cable a frequency of 2 cycles per second is received from the cable with one seventieth the amplitude which it had at the sending end, whereas for 4 cycles per second the received amplitude is one four hundredth of the sent amplitude and for 8 cycles per second it is only one five thousandth. The function of the distortion-correcting networks and apparatus is to attenuate the lower frequencies more than the higher ones so that the combination of cable and terminal apparatus will attenuate all frequencies up to a certain value approximately alike. The process of signal shaping may thus be regarded as one of attenuation equalization for a limited frequency band extending upward from zero. With the networks and apparatus employed on non-loaded cables the same means which serve approximately to equalize attenuation serve also to equalize time-lag. For frequencies higher than those required to form legible signals it is desirable to reduce the received current to as low a value as possible, since at such high frequencies the currents induced by sources of interference are usually stronger than those which belong to the signals. Accordingly the exclusion of these high frequencies makes the received signals more legible in being less affected by external disturbances.

The electrical networks for correcting distortion may be applied at either the sending or receiving end of the cable, or may be divided between the two ends. In ordinary cable practice it is common to

use a condenser in series with the cable at the sending end and to provide further means for signal shaping at the receiving end. The use of partial sending-end shaping has also been found desirable for the loaded cable though a modified circuit arrangement has been found more effective than the simple sending condenser.

Within recent years it has become common practice in the operation of cables to employ means for amplifying the received signals prior to relaying or recording them. This has been necessitated by the limited sensitivity of relays and recording instruments. Most of the amplifiers which have proved successful have been instruments of the moving-coil type in which a slight motion of the coil of a D'Arsonval galvanometer is caused to control a much larger source of power than that which is required to move the coil. Instruments of this type possess an advantage in that their mechanical inertia and stiffness may be used to assist in the processes of signal shaping and interference elimination. On account of mechanical limitations they are not, however, well adapted to operate at the high speed of the loaded cable.

Vacuum-tube amplifiers have been used to a limited extent on non-loaded cables and have many advantages over the moving-coil instruments, notably in their mechanical ruggedness and in the large amount of amplification which can readily be obtained with them. For use on loaded cables they have a further great advantage in that they have no frequency limitations within the range employed on cables and serve as well for high-speed cables as for low. By the use of suitable electrical networks in connection with the vacuum tubes the signals may be restored in shape, and interfering disturbances outside of the signal range of frequencies may be eliminated. A vacuum-tube amplifier which combines means for amplification, correction of distortion, and elimination of interference has been called a "signal-shaping amplifier."

With the combination of sending-end shaping network, loaded cable and signal-shaping amplifier, means are provided for conveying signals in the form of combinations of electrical impulses from one terminal to the other. Any type of telegraphic apparatus for converting messages into signals and reconverting signals into messages may be applied to complete the steps involved in one-way operation. None of the standard types of cable or land-line apparatus, however, are well adapted to meet the needs of commercial operation at the speed of the fastest loaded cables; to gain the full advantage permitted by the cable requires apparatus of special design. Special provision is also required to permit two-way operation.

There are two principal ways in which two-way working may be secured: messages may be sent simultaneously in the two directions or the cable may be used alternately in either direction. The first method is commonly called duplex and the second, simplex. Although, as was pointed out earlier in this discussion, the loaded cables which have been laid were designed primarily for simplex operation, it would be entirely possible to operate them duplex; but to do so would require the employment of an artificial line having nearly the same impedance as the cable over the range of frequencies involved in the signals. The speed of duplex operation would, of course, depend on the accuracy with which the artificial line could be made to balance the cable and this would be largely a matter of cost. Simplex operation, if the reversal of direction is made automatic, has much to recommend it over duplex. It does not require an expensive and complicated artificial line which would need frequent readjustment and it permits using the full speed of the cable to the best advantage to accommodate traffic. Means for reversing the direction may readily be associated with means for automatic printing operation and many of the objections to simplex working which are commonly thought of by the cable engineer do not apply when the reversal is thus made automatic.

Apparatus for the high-speed automatic operation of loaded cables has been described in recent papers by A. M. Curtis and A. A. Clokey. The Curtis paper [7] deals principally with the apparatus for signal shaping and amplification, while the Clokey paper [8] describes the special methods and apparatus for automatic printing telegraph operation. Some of the outstanding features of both classes of apparatus will be discussed in the following sections of the present paper.

Apparatus for Restoration of Signals

A typical circuit diagram of a loaded cable with its terminal networks for signal shaping and amplification is shown in Fig. 5. For the sake of simplicity the circuit details required for two-way operation have been omitted. Such a circuit arrangement applied to a transoceanic permalloy-loaded cable serves to connect a telegraph transmitting instrument with a receiving or recording instrument for one-way operation nearly as effectively as they could be connected by an overland telegraph line.

[7] "The Application of Vacuum Tube Amplifiers to Submarine Telegraph Cables," *B. S. T. J.*, July 1927.

[8] "Automatic Printing Equipment for Long Loaded Submarine Telegraph Cables," *B. S. T. J.*, July 1927.

At the sending end in place of the usual sending condenser there is employed the network N_1 shown in the figure. The condenser C_s may have a capacity of from 30 to 80 microfarads. It is shunted by a resistance R_1 of several thousand ohms. The resistance R_2 connecting the sending end of the cable to earth may be of the order of 100 ohms, and serves approximately to equalize the input impedance of the system over the important range of frequencies. The desirability of the resistance R_2 is peculiar to the loaded cable and is occasioned by the manner in which its characteristic impedance varies with frequency.

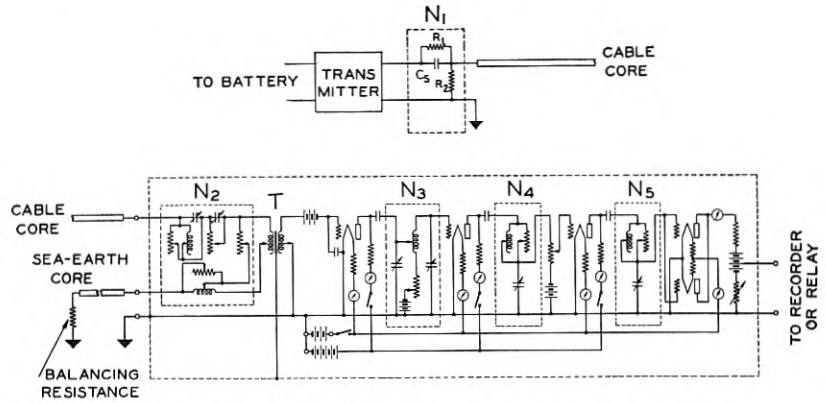


Fig. 5—Terminal networks for signal shaping and amplification

Other sending-end circuit arrangements can, of course, be used and networks combining inductances with capacities and resistances have been effectively employed. The sending-end shaping network may even be dispensed with entirely and all of the shaping done at the receiving end. There are, however, certain conditions under which this leads to the production of distortion due to hysteresis in the magnetic material of the cable and in general it is preferable to reduce the current flowing into the cable by employing sending-end shaping networks which reduce the amplitude of the low-frequency components of the signal.

The circuits employed at the receiving end for completing the process of signal shaping and for amplifying the signals may conveniently be considered in three parts, the receiving shaping network N_2, the shielded transformer T, and the amplifier which includes the interstage shaping networks N_3, N_4 and N_5.

The receiving network N_2 provides means for correction of a considerable part of the distortion introduced by the cable and in so

doing reduces the peak voltage which is applied by the signals to the primary of the transformer T and also the peak voltage which is applied to the grid of the first vacuum tube. By insertion of this shaping network between the cable and the transformer, overloading and consequent distortion are prevented.

The transformer T permits insulating the amplifier and its batteries from the cable and thereby allows the amplifier to be connected directly to earth [9] and to be effectively shielded from local electrical disturbances. Without the transformer or other means to insulate the amplifier from the cable it would be impossible to use an earthed amplifier and at the same time to secure the advantage of the balanced sea-earth in eliminating interference. The requirements for this transformer are very severe since it must be effective for frequencies as low as 0.2 cycle per second and at the same time must be constructed so that it will not pick up the external electrical disturbances generally prevalent in cable stations. The use of a permalloy core and a permalloy shield has made it possible to meet these requirements in an instrument occupying less than one third of a cubic foot.

Connected between the successive stages of the amplifier are the signal-shaping networks N_3, N_4 and N_5. These networks serve both to adjust the shape of the signal and to reduce the effects of interference outside of the signal range. Considerable advantage is gained from the fact that there are in the entire system five signal-shaping networks, each separated from its neighbors by either the cable or the vacuum tubes. This arrangement permits independent adjustment of the separate networks with very little interaction between them and greatly facilitates the systematic correction of signal shape.

The values of the various resistances, inductances and capacities in the networks at the receiving end depend, of course, on the cable as well as on the type of telegraph apparatus employed; for this reason most of the important circuit elements are made adjustable. The adjustments are made by trial, but in spite of the apparent complexity of the networks, which are more elaborate than would be required for any given cable with fixed operating requirements, the adjustments necessary to adapt the apparatus to any particular conditions can be made quite systematically. After the shaping adjustments required for a particular cable have been worked out, which usually takes not more than a few days, the amplifier can be adjusted for any speed in the range of the cable in a few minutes.

[9] The earth connection for the amplifier is preferably made to a short "sea-earth" conductor terminated on the cable sheath at a few miles from shore. The same earth conductor may be used for a transmitting earth.

The output of the amplifier may be applied to a siphon recorder or to relays, as desired, and the amount of amplification may be adjusted over a wide range to meet the requirements of any particular case. In general the power amplification needed for automatic operation of a loaded cable at its maximum speed is of the order of 10,000,000 times, which corresponds to 70 TU.

The external appearance of the signal-shaping amplifier is shown in Fig. 6. All of the receiving circuit elements shown in Fig. 5

Fig. 6—Signal-shaping amplifier

are contained in its shielded case which is made of ample size so that all of the essential apparatus units within it are readily accessible. Great care has been used in the design and construction of the amplifier unit to protect the circuit elements within it from moisture and to prevent leakage or electrostatic coupling. The output terminals of the amplifier may be connected directly to a siphon recorder or to a suitable relay. However, when the amplifier is used for the operation of relays and multiplex printing telegraph apparatus, there is associated with it an additional piece of apparatus called the relay control desk

which is shown in Fig. 7. In this unit is provided means for control and adjustment of the relays and also means to compensate the type of signal distortion commonly described as the "wandering zero" which results from the inability of the system as shown in Fig. 5 to transmit direct current.

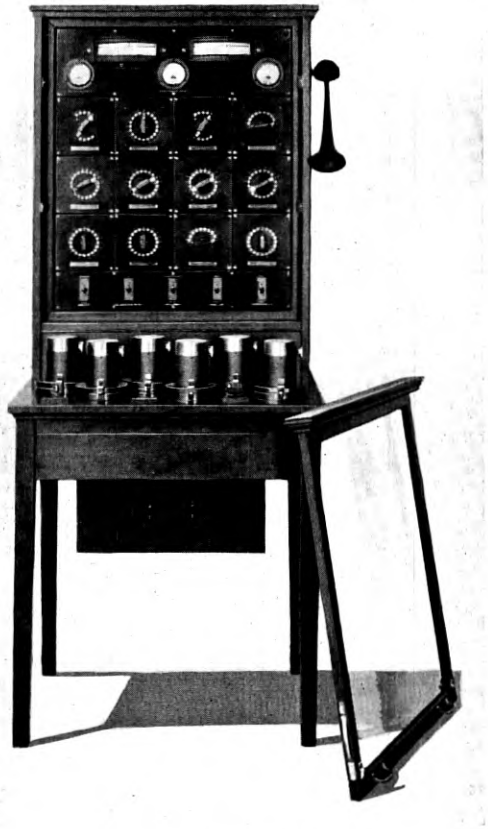


Fig. 7—Relay control desk

Amplifiers of the type described are in commercial operation at all terminals of the Western Union and Deutsch Atlantische loaded cables. In this extensive commercial use they have been shown to require considerably less maintenance than the moving-coil instruments which are commonly used on non-loaded cables, and in fact the loss of time in operation due to troubles in the amplifiers has been almost entirely negligible. It is of interest to note that it has been possible

on numerous occasions to operate cables with these amplifiers during the entire course of severe thunder storms with the loss of only an occasional letter due to lightning discharges.

Apparatus for Automatic Operation

The first operating tests of the New York-Azores cable were made with the signal-shaping amplifier described in the preceding section. For these tests cable-code operation with a siphon recorder was employed, this type of operation being chosen because it would

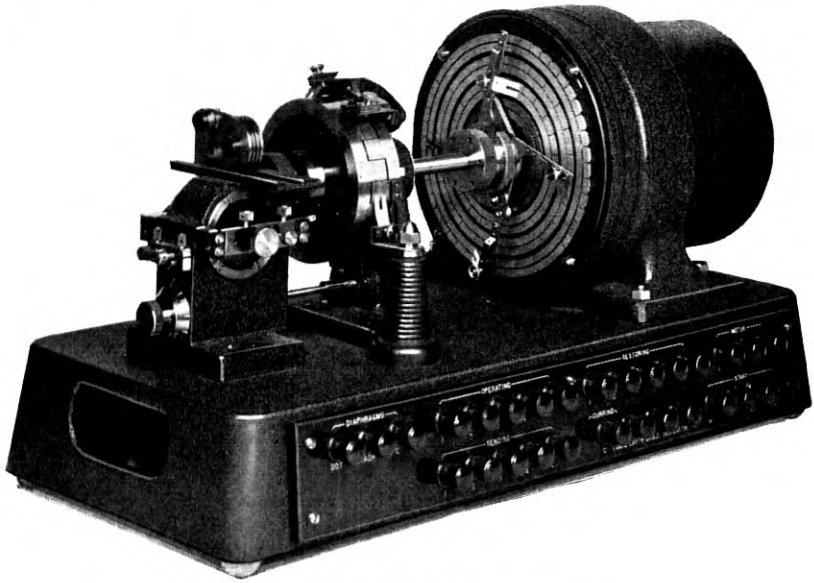


Fig. 8—High-speed cable-code transmitter

permit direct comparison of the behavior of the loaded cable with that of ordinary cables. The ordinary cable-code transmitters and siphon recorders were, however, incapable of operation at the predicted speed of over 1500 letters per minute and a new transmitter and recorder had to be provided for testing and demonstrating the operation of the new cable.

The high-speed transmitter which was developed for these tests is shown in Fig. 8. This transmitter makes use of the ordinary perforated tape used with standard types of cable transmitters but instead of opening and closing contacts by mechanical means it employs pneumatic means for this purpose, the perforated transmitting tape being utilized in the manner of the perforated sheet in a player-piano.

A commutator and relays associated with the pneumatic apparatus serve to equalize the lengths of the transmitted signals and to provide any desired ratio of "marking" to "spacing." This transmitter is capable of operating at speeds up to about 2500 letters per minute.

The high-speed siphon recorder is shown in Fig. 9. It differs from the standard instrument in many respects. A very light moving coil is supported horizontally in the strong field of an electromagnet by a bifilar suspension. A very light rigid arm attached to the coil carries a siphon pen only about 2 cm. long which writes on ordinary

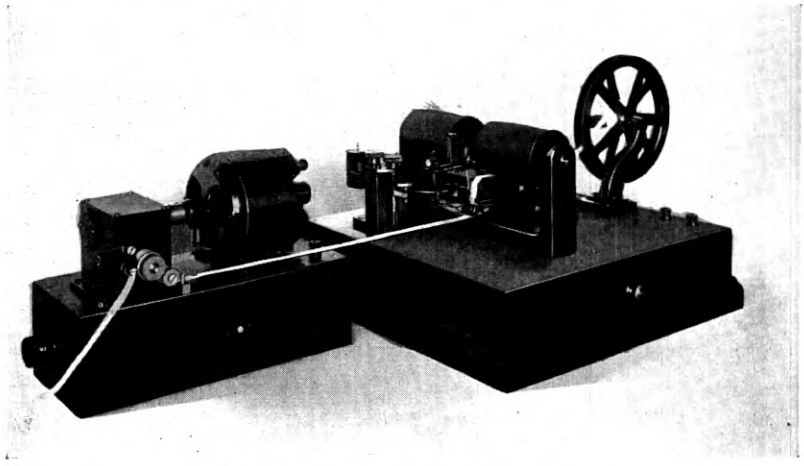


Fig. 9—High-speed siphon recorder

recorder tape drawn rapidly over a vertical table. This instrument may also be operated at 2500 letters per minute with cable-code and makes a record similar to that of the standard siphon recorder.

Both of these instruments and the signal-shaping amplifier were provided in advance of laying the first permalloy-loaded cable and were used on the first tests. A record of an early test message made on the New York-Horta cable at a speed of 1920 letters per minute is shown in Fig. 10.

Since this first cable terminated at the Azores Islands where there was no immediate demand for the full speed of which the cable was capable, the first commercial operation was conducted at a speed of only about 800 letters per minute. This was obtained with a standard cable-code transmitter and a standard type of recorder used with the signal-shaping amplifier. The cable was operated alternately in the two directions as required to accommodate traffic, the reversal of

direction of operation being controlled manually. While this type of operation served well to carry the limited traffic then available, it was not suited for efficient operation of the cable at its maximum speed, both because of the practical difficulty of dividing the rapidly received recorder tape among the three or more operators who would be required to translate it, and because of the delays resulting from manual control of reversal of direction.

To make efficient use of a high-speed telegraph cable requires some means of adapting it to the practical limitations of machines and operators, preferably by the provision of a number of separate channels of operation, each of which may be worked at a speed con-

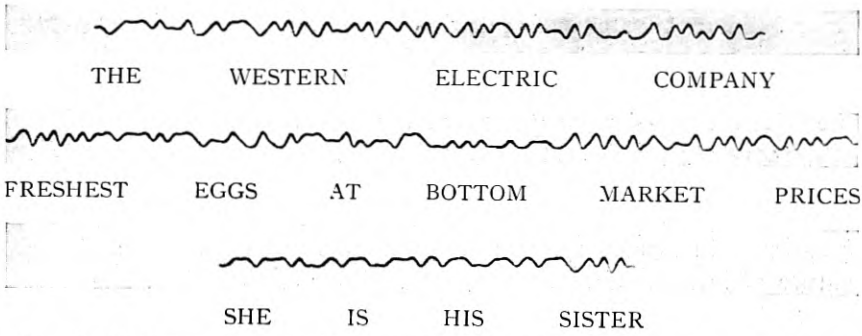


Fig. 10—Test message transmitted over New York-Horta cable at a speed of 1920 letters per minute, Nov. 14, 1924

sistent with the pace of a single operator at each end of the cable. With such multi-channel operation it is obviously necessary to provide means for either simultaneous two-way working or automatic means for direction reversal which shall not interfere with the independent operation of the several channels. Also it is very desirable to provide for automatically printing the received messages.

There are two principal methods which have been used to secure multi-channel operation with a single telegraph line—the carrier current method and the multiplex distributor method. By the former the separate channels are obtained by the modulation of separate carrier frequencies in accordance with the telegraphic signals, the line being simultaneously shared by all the channels; by the latter the line is passed in rotation from one channel to the next so that the line time is in effect divided equally among the several channels. Either method or a combination of the two can be applied to a loaded cable. The carrier current method has for several years been used on the loaded cables of the Cuban-American Telephone Co. between

Key West and Havana and has also been used on some non-loaded cables and quite extensively on land lines. The multiplex distributor method is used widely on land lines and has also been used to some extent on non-loaded cables. Of the two the multiplex distributor method makes more effective use of the line when the frequency-range is limited to about 100 cycles per second or less and the carrier current method is more effective when a considerably wider frequency-range is available. Since the frequency-range provided by the New York-Horta cable extended to about 60 cycles per second, the multiplex distributor method was the more effective means for providing multiple channels on this cable and was accordingly adopted.

With the multiplex distributor method of separating channels several different systems of operation employing different signal codes are possible and several different codes have been practically applied. Among these are the cable-code, the three-unit three-element code and the five-unit two-element or Baudot-type code. To determine which of the several possible systems can give the greater speed of operation is an extremely complex problem since it requires consideration not only of the number of characters or letters and their frequency of occurrence in messages but also of the line characteristics and the nature of interference. From the practical point of view, however, the multiplex system, which employes a code of the Baudot type, has the great advantage of availability of perfected transmitting and printing apparatus and, in view of this advantage, there seems little doubt of this being the best system for the immediate practical realization of the possibilities of a loaded transoceanic cable. In this system the line-time is divided into as many parts as there are channels of communication and each of these parts is divided into five units. The line is thus used in effect to transmit five successive signal units of either positive or negative polarity from one transmitter to its corresponding receiver, thereby sending one letter or character over one channel. It is next used to send similarly another letter on another channel and so on until a letter has been sent over each channel, whereupon a second letter is started over the first channel.

Although multiplex distributors for land lines had long been available, the standard apparatus was not suitable for realization of the full advantage of the permalloy-loaded cable. This was appreciated from the first, and long before the manufacture of a loaded cable was started the development of a system for operating it was undertaken. In several important respects the apparatus developed for the cable is different from that used on land lines.

Two-way operation is provided by automatic reversal of the direction

of sending. This is accomplished by driving from the multiplex distributor a reversing mechanism which switches the cable from sending eastward to sending westward or vice-versa at regular intervals without the loss or mutilation of a character on any channel. To adapt the apparatus to the demands for traffic, the intervals of reversal are made capable of variation over a considerable range so that the system can be used, for example, alternately one minute eastward and ten minutes westward or three minutes eastward and three minutes westward, only about five seconds being lost at each reversal.

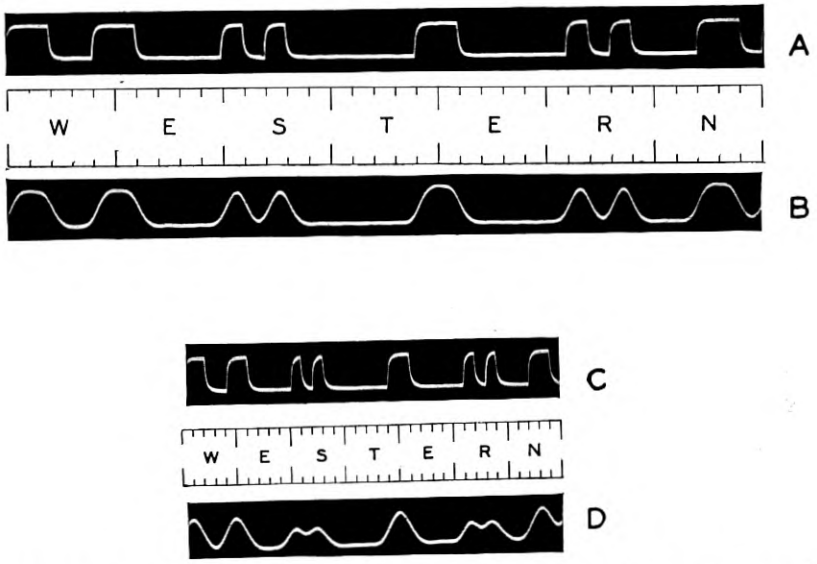


Fig. 11—Sent and received signals. *A*, Signal, sent at speed suitable for plain relay operation; *B*, Received signal shaped for simple relay; *C*, Signal sent at twice speed of *A*; *D*, Received signal shaped for vibrating relay.

(These records were made in the laboratory with an artificial line and accordingly do not show the interference which would be present in the case of a cable operated at maximum speed.)

To secure the maximum speed of operation, use is made of the "synchronous vibrating relay," a method of signal restoration developed in the course of our laboratory studies of apparatus for loaded cables. The synchronous vibrating relay takes advantage of the principle of the Gulstad vibrating relay, which has been extensively used on both land-lines and cables, but possesses a further advantage in that use is made of the synchronous multiplex distributor to secure the most effective application of this principle.

To describe and explain the circuits and apparatus of the syn-

chronous vibrating relay would be beyond the scope of this paper but the way in which it permits an advantage in speed of operation may readily be appreciated from a consideration of the signals shown in Fig. 11.

Consider first the conditions in plain relay operation without the use of the vibrating relay principle. The signal train *A*, Fig. 11, represents the word "western" as translated into the code used in the multiplex printing telegraph system. If this word is transmitted over the combined cable and distortion-correcting networks at a suitable speed for plain relay operation, it will be received in the form, *B*, in which a transmitted impulse of unit length has resulted in a received impulse of about the same amplitude as that of impulses two or more units long. A simple relay operated by the signal train, *B*, will substantially reproduce the original transmitted train, *A*.

Consider now the signal train, *C*, in which the same word "western" is transmitted at twice the speed of *A*. With the same adjustment of cable and terminal networks it will be received in the form, *D*, in which impulses of two units of length are received with the same amplitude as that with which the unit length impulse was received in *B*, whereas the amplitude of a succession of received reversals of unit length in *D* is reduced nearly to zero. Obviously, if *D* were applied to a simple relay, it would not cause the original signal train, *C*, to be reproduced. However, *C* can be reproduced from *D* by means of the synchronous vibrating relay which is arranged to supply impulses of unit length locally, unless prohibited from so doing by currents due to impulses of two or more units of length. One may regard the cable and terminal networks as converting the transmitted two-element (plus and minus) signals into three-element (plus, zero and minus) signals which the vibrating relay reconverts into two-element signals, and in this way permits operation at a speed which is much higher than is possible with a plain relay. With the Gulstad relay or with minor modifications of it, the locally interpolated impulses are supplied from a local vibrating circuit and do not always occur at exactly the right time to be most effective. With the synchronous vibrating relay, these impulses are controlled by the distributor and are therefore introduced at precisely the right time. It is interesting to note that this can be done in a system in which the incoming signals control the rate of the distributor.

Another feature of the apparatus for the loaded cable is its high degree of precision and refinement. The cost of a cable relative to that of even the most refined apparatus is so great that no considerable sacrifice of speed can be justified by ordinary economies in apparatus.

Accordingly, the efficiency to be gained by extreme precision has been sought, and to achieve this desired precision has required radical departure from the design used in land line apparatus. A photograph of one of the distributors used on the New York-Horta-Emden line is shown in Fig. 12. On this line three 5-channel distributors are used, one each at New York, Horta and Emden. Within a few seconds after one of five operators at New York prepares a perforated

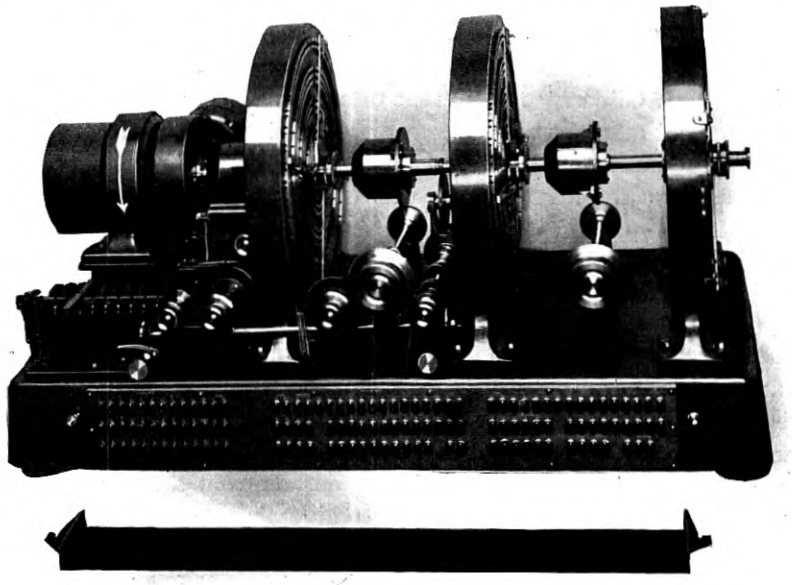


Fig. 12—Multiplex distributor used on New York-Horta-Emden line

strip on a machine resembling a typewriter, the message appears in typewritten form in Emden on a strip ready for delivery or retransmission over a land line.

While the system which I have roughly described is one which was developed principally with regard to use on a particular cable, the principal features embodied in it are applicable to any long loaded cable of the type discussed in this paper. The details of the apparatus which should be used on any cable are of course dependent on the particular requirements to be met and each installation must be engineered for its special needs if the full benefit of loading is to be realized.

Electrical Measurements of Loaded Cables

To check the assumptions made in the design of the first cables, and to obtain the information necessary for the design of the ultimate operating equipment, extensive electrical measurements were made on the three Western Union cables after they had been laid. From an analysis of these measurements the several electrical parameters of the cables were determined. To do this required new apparatus and methods, the development of which was by no means a small part of the total effort involved in the first project. Since a review of some of the methods of measurement has been given in a recent paper by J. J. Gilbert,[10] the present discussion will be limited to the apparatus and methods which seem to be of particular interest.

One of the most important tools in all our investigations concerned with the permalloy-loaded cable was the string oscillograph shown in Fig. 13. From the start it was recognized that an instrument would be needed which would give an accurate record of the manner in which the currents and voltages which were being studied changed with time and, in fact, the first step in the experimental investigation of the cable problem was to search for a suitable oscillograph. Fortunately it was not necessary to look far. A string oscillograph which had been developed for sound-ranging of artillery during the World War was quickly modified and devoted to more peaceful purposes. The present instrument differs in many details from the original but retains the invaluable asset of ability to give almost instantaneously, completely developed and fixed, a distortionless picture of a wave involving any frequencies up to about 300 cycles per second. In a study like that of signal shaping, involving the determination of the effect of numerous slight changes in adjustment of the apparatus, the advantage of such an instrument is obvious. This instrument was used both in the early studies of signal correction with the laboratory artificial cable and in the later measurements on the laid cables, and today is a useful adjunct in cable stations where it serves to show the character of the cable signals at any stage of their conversion into impulses for the recording instruments. A feature of particular value in studying phenomena such as extraneous interference on cables is that the oscillograph permits taking a continuous record over a period of several minutes.

Other special instruments which I shall only mention, but which were developed especially for the cable experiments, were an attenuation meter and a low-frequency vacuum-tube oscillator to give

[10] "Determination of Electrical Characteristics of Loaded Telegraph Cables," *B. S. T. J.*, July 1927.

voltages of constant frequency as low as 2.5 cycles per second and of very pure wave form.

To determine the various electrical parameters of the laid cable wholly from measurements at its terminals was practically impossible

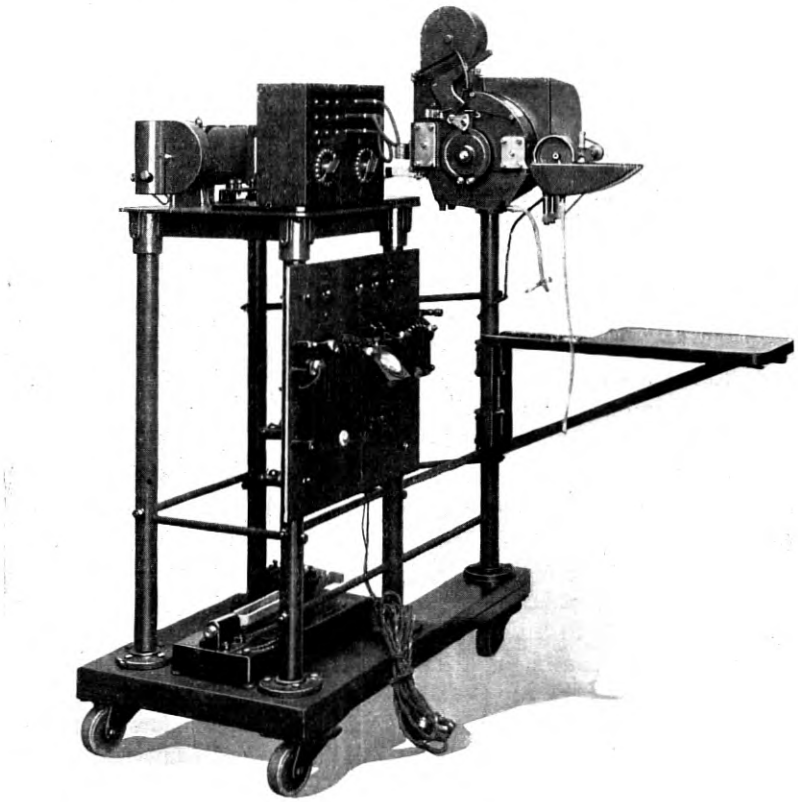


Fig. 13—String oscillograph

on account of the complicated manner in which the resistance and also to some degree the inductance and leakance vary with frequency. However, by combining the results of factory measurements with the results of measurements made at the cable terminals it was possible to determine the fundamental characteristics of the laid cable with a fair degree of accuracy. The method consists essentially in measuring, at a number of frequencies ranging from 5 cycles per

second to the highest frequency possible, the attenuation and delay or time-lag of trains of reversals transmitted over the cable.

Attenuation was measured by transmitting square-topped reversals of constant voltage and frequency and measuring the received current by means of either an amplifier and thermocouple or an amplifier and the string oscillograph, the latter method being preferable at high frequencies where the effect of interference would prohibit accurate measurement by means of the thermocouple.

The delay for steady alternating currents was measured by transmitting short trains of reversals alternately from the two ends of the cable, the transmitted and received trains at each terminal being recorded by means of the string oscillograph on a continuous strip of paper. Thus in a single cycle of operations the record at one terminal would show a transmitted train followed by a received train, while the record at the other terminal would consist of a received train followed by a transmitted train. Since a certain time is required for the establishment of the "steady state" condition at the receiving end, it was found desirable to base the measurement of the time of arrival and departure on a point well along in the train, say at the fifth to tenth cycle. The difference in elapsed time between sending and receiving at the two ends of the cable then gave twice the steady-state delay or "time of propagation" for the particular frequency for which measurements were made.

Both attenuation and delay were measured for several current values and by extrapolation to zero current the values of these quantities corresponding to a very small current amplitude could be determined. Measurements made on cores in the factory under various conditions of temperature and hydrostatic pressure gave a value of capacity for the cable and from this and the measured delay the inductance could be computed. Knowing the inductance and the capacity, the effective resistance of the cable at various frequencies could be computed from the measured values of attenuation, the value of leakance being estimated from factory measurements. This value of effective resistance should agree with the value obtained by adding together the various known components of resistance. Of the components of resistance, the copper resistance and the resistance introduced by the loading material can be computed from measurements made in the factory. The sea-return resistance can be computed from theoretical formulas and the effect of reflections from irregularities along the cable can likewise be estimated.

For the cables on which such measurements have been made, the values of effective resistance obtained from cable measurements agree

to within less than 5 per cent with the values computed as described. Part of this difference is probably due to the fact that the computed values of sea-return resistance are smaller than those actually encountered on the cable. A possible explanation of this effect is that the electrical resistivity of the earth beneath the cable is considerably higher than was assumed in the theoretical development of the formula for sea-return resistance.

A General Survey

The preceding discussion has referred principally to the progress in certain lines of development which were chosen as best suited to accomplish the result of high-speed ocean cable telegraphy. It is of interest now to consider in a general way the field of application of loaded telegraph cables and the nature of modifications which might be made in their construction and operation.

All of the loaded telegraph cables to which I have referred are relatively long. That permalloy loading should have been applied first to long cables is the natural consequence of the facts that the need for increased cable speed was principally between points far apart and that the greatest economic gain from loading could be obtained with long cables. Where high-speed cable operation was desired between points only a few hundred miles apart, it could readily be obtained with a non-loaded cable by merely making the cable large enough to give the required speed. Accordingly for short cables the operating speed was determined either by the demand for communication or by limitations of terminal apparatus. But where the necessary length was of the order of 2000 miles or more, even the heaviest cables which were considered practicable to lay and maintain were limited by the inherent characteristics of non-loaded cables to relatively low speeds, and it was accordingly for such great distances that the manifold speed advantage of the permalloy-loaded cable was of the greatest value.

To give a fair numerical estimate of the advantage of loading long cables is extremely difficult since the result depends so much on the basis of comparison and the limitations of size and operating requirements which are imposed. Probably the most nearly fair basis of comparison would be the relation of cost to speed for the old and for the new cables, but to make such a comparison requires data on cost which is forever changing. An interesting basis of comparison from the technical point of view is the ratio of the traffic capacity of a non-loaded cable operated duplex to that of a loaded cable operated

simplex, both cables being of the same length and size. The latter condition will be met if the diameter of the loaded conductor measured over the permalloy is the same as that of the copper conductor of the non-loaded cable and if the thickness of gutta-percha is the same for both. On this basis one can say that for cables of lengths from about 2000 to 3500 miles the loaded cable has approximately five times the traffic capacity of the corresponding non-loaded cable, this gain being obtained, of course, with a relatively small increase of cost.

A similar comparison might be made for shorter cables but it would have relatively little significance since in any practical case the loaded cable would probably not be made to have the greatest possible speed consistent with practicable size but would be designed with regard to the limitations of terminal apparatus or connecting lines. The problem becomes more complex as the assumed length is reduced since the shorter is the cable the greater are the number of possible ways of obtaining the desired speed and the more is the speed dependent on terminal equipment. It is, however, safe to say that, where the demand for communication is sufficiently great, loading will prove advantageous for cables of all lengths down to perhaps 100 miles or less, but for cables much less than 2000 miles long the electrical design of any particular cable will depend greatly on the use which is to be made of it.

In view of the great gain due to loading long cables it is most probable that all very long cables of the future will be loaded and it is likewise probable that long cables will be used in some cases where previously several short non-loaded sections with repeating apparatus would have been used. Loading will also be used to a considerable extent on shorter cables but it should not be expected that all of the shorter cables will be loaded since there are many cases where the demands for communication which can now be foreseen are so limited that they can be met more economically by non-loaded than by loaded cables.

In Malcolm's prediction of the loaded ocean cable, to which I have previously referred, he went so far as to suggest that even though the first loaded ocean cable would probably be of the continuously loaded type, ultimately coil-loading might be resorted to. Malcolm, of course, was not in a position to take into account the effect of such a radically new material as permalloy, and with the materials which were known to him coil-loading appeared to offer possibilities which continuous loading did not. It is interesting therefore to examine the present apparent merits of coil-loading with regard to its application to transoceanic cables.

An obvious great difficulty with coil-loaded deep-sea cables lies in the mechanical problem of laying a cable to which coils are attached or in which coils are inserted in a way to give a mechanical irregularity. Unless the coils could be made extremely small their presence would certainly interfere with passing the cable smoothly through the paying-out machinery. Cable laying and repairing are sufficiently difficult and hazardous under the most favorable conditions and any alteration in cable structure which would make these tasks more difficult is certainly to be avoided if possible. Permalloy cores for loading coils might, however, to some degree eliminate this objection to coil-loading, since with a permalloy core the loading coils may be made smaller with the result that less difficulty would be caused by the increased size of the cable at the points where the coils were inserted.

The problems of maintaining good insulation and sound joints at the loading coils are probably much more serious. Conductor joints in a cable are frequently subject to considerable stress and even with the relatively simple joints required for ordinary deep-sea cables trouble is occasionally experienced. With loading coils inserted in the cable both the coils and the joints between the coils and the core must be subject to great stress, and since the coils must be many in number to be effective, the probability of faults with even the best imaginable construction would be greatly increased.

From the electrical point of view an apparent advantage of coil-loading is that it might conceivably permit adding the required inductance without introducing so much a.c. resistance and thereby permit more closely approximating the ideal loaded cable. On the other hand, coil-loading has an electrical disadvantage which has not generally been appreciated but which is of serious practical consequence. This disadvantage lies in the distortion of signal-shape arising from the lumped character of the line. With uniform continuous loading the line is electrically smooth; such a line may introduce distortion but this distortion can be compensated for by terminal apparatus. With coil-loading the line is, in effect, a network of as many sections as there are loading coils. Such a line introduces a new type of distortion which arises in the so-called filter oscillations. Although it is theoretically possible to compensate for this effect by terminal networks, the circuits required are extremely complex and practically the limit of speed is set by the frequency of signal impulses at which filter oscillations begin to cause serious distortion. This effect can be practically eliminated by making the distances between coils sufficiently small, but as the distance between coils is diminished the otherwise possible advantages of coil-loading are likewise diminished.

Even if it could be shown that all of the apparent objections to coil-loading could be overcome, I think it is highly improbable that coil-loading would be resorted to for long deep-sea telegraph cables. Continuous loading has been given a practical trial and has proved successful and does not add greatly to the cost of a cable. Coil-loading involves risks which there is now no need to assume and its economic advantage, if any, is certainly small in proportion to the whole cost of a cable installation.

Though continuous loading as applied in several particular instances has been successful, there is no occasion to assume that the development of the art of continuous loading is completed. Modifications in continuous loading can be introduced with relatively little risk and are justified if an economic advantage can be shown. Also cable construction, apart from the loading, may be modified so as to realize more completely the advantages which loading affords. It is therefore of interest to consider some of the ways in which continuously loaded cables of the future might be different from those of the present.

Loading materials can be produced with different magnetic properties and to a limited extent the resistivity of alloys may be altered by control of composition. Higher permeability is not necessarily desirable since with increased permeability goes also increased effective resistance due to energy losses in the permalloy. In the case of any particular cable with practical limitations of dimensions, materials and costs there is an optimum permeability which, in general, is lower the higher is the frequency for which the cable is designed. Short cables designed for very high frequencies will accordingly require lower permeability than has been required for the long cables which have been made. For cables of all lengths and speeds a high degree of constancy of magnetic permeability with regard to magnetizing force is desirable. With the New York-Horta cable the inductance increases about 50 per cent when the current is increased from 0.001 to 0.1 ampere. The relative increase is less for some of the later cables, owing to improvements in loading. A loading material of high electrical resistivity is, of course, always advantageous.

There are other ways of applying continuous loading than that of Krarup. For example, the magnetic material may be electroplated onto the conductor. Such modifications may eventually come into use but the need for them does not appear to be great in the case of long deep-sea cables and there is not much economic incentive for their development. Accordingly I do not believe that changes of this type are likely to alter greatly the possibilities of submarine cables for telegraphy over long distances.

The cost and physical characteristics of insulating materials are, of course, factors of great importance. With few exceptions gutta-percha or compounds consisting mostly of gutta-percha have been used for long submarine cables. The cost of the gutta-percha insulation is a large part of the whole cost of a cable and the fact that its cost is high leads to using the least amount consistent with maintaining safe insulation. If a very much cheaper substitute or one of superior electrical properties were available, the basis of design of loaded deep-sea cables might be somewhat changed.

Even the sheath of armor wires is capable of considerable improvement as regards its effect on the behavior of a loaded cable. As pointed out previously, the sheath introduces electrical resistance due to the fact that it carries some of the return current which at high frequencies tends to concentrate around the cable. Armor wire of higher resistivity would introduce less resistance in this way or an electrical improvement might be obtained by consideration of this effect in the mechanical design of the cable sheath. At very high frequencies where the return current is closely concentrated around the cable the armor wire has an opposite effect and an improvement can be obtained by decreasing its resistance or by the addition of other conductors in parallel with it as was done on the Key West-Havana telephone cables. Some electrical resistance is introduced by the magnetic coupling between the sheath and the conductor due to the helical shape of the loading material and the armor wires. This effect is small in the cables which have been made but might be large in higher speed cables and should be taken into account in the design of such cables. A similar effect results from the use of the ordinary teredo tape on loaded cables.

With improvements in materials and construction which permit higher operating speeds and with the demand for more efficient means of handling a large volume of cable traffic, greater importance will doubtless be attached to duplex or simultaneous two-way operation, and it is of interest to consider some of the ways in which duplex operation might be secured. Of course there is no reason to assume that the loaded cables which have already been laid will not eventually be operated duplex although they were designed primarily with simplex operation in view. From the studies of both types of operation which we have made it appears more economical for the present to operate the existing transatlantic loaded cables one way at a time. Indeed this type of operation with automatic reversing apparatus possesses many advantages over the ordinary duplex methods applied to non-loaded cables. If, however, a loaded cable were designed

originally with regard to duplex operation, the possible advantage of applying duplex apparatus to it would obviously be greater than it is for any of the existing loaded cables.

There are many ways in which the design of a cable might be modified to make duplex operation more advantageous than it is on the present cables, and the problem is much too complex to permit very detailed discussion here. Improvements in constancy of inductance with variation in current and in reduction of alternating-current resistance factors would be of obvious advantage. A tapered cable with high inductance in the middle and low inductance at the ends would also have advantages in this connection and to a limited extent tapering has already been applied to the Pacific Cable Board's loaded cables by arranging the component parts of these cables during manufacture with regard to their inductance.

One of the most attractive methods of duplexing, which would also provide great flexibility of operation, is to use carrier current operation in one direction and ordinary telegraph operation in the opposite direction. Non-loaded cables are ordinarily duplexed by balancing the cable at each end with an artificial line which permits separation of the weak incoming signals from the strong outgoing ones and the limit of speed is usually set by the accuracy with which the cable may be balanced by the artificial line. By using carrier current operation in one direction and ordinary telegraphic operation in the opposite direction the incoming and outgoing signals may be separated by the combined use of artificial lines and frequency filters, in the manner long since employed on the Key West-Havana cables for carrier currents above the voice-frequency range. To design a cable for carrier current operation would, of course, require consideration of its behavior at much higher frequencies than those employed on the existing long loaded cables and would probably call for very high resistivity loading material applied in a very thin layer.

The recent spectacular development of radio both for telephony and telegraphy has raised in the minds of all the question as to whether there is any future left for ocean cable telegraphy. Opinions on this question will doubtless differ. My own opinion is that for short distances across the sea, where the demand for communication is considerable, cables always have offered more economical and satisfactory communication than radio and will probably continue to do so. For long over-sea distances I believe the cables would have faced a serious situation in competition from radio had not permalloy loading been brought forth. Now permalloy loading has so reduced that part of the total cost per word for which the cable itself is responsible

that the financial advantage of radio can never be very great. It has yet to be shown that radio telegraphy can furnish as reliable and satisfactory service as is now provided by the cables. How long the cables will continue in the leading position remains for time to tell, but it is significant that the cable companies have gone courageously ahead with new projects, and it is evident that only a much higher degree of perfection of radio communication than has yet been attained can permit wresting from the cable the advantage which it has so long maintained.

The Present Status of Wire Transmission Theory and Some of its Outstanding Problems

By JOHN R. CARSON

SYNOPSIS: The rapid development in the technique of wire transmission and the increasing complexity of the problems involved calls for a more adequate theoretical guide and a more rigorous transmission theory. This paper gives an account, practically without mathematics, of classical transmission theory and its limitations; of the several ways the problem may be attacked more fundamentally and rigorously, and the lines along which transmission theory must be extended, as the writer has come to view the problem in the light of his own experience.

IN the present paper the term *wire transmission theory* will be understood to mean the mathematical theory of guided wave propagation along a system of parallel conductors; which is supposed to be geometrically and electrically uniform throughout its length. The theory of wave propagation along such a system is of fundamental theoretical and practical importance to the communication engineer and presents some extremely interesting and difficult problems to the mathematician. The development of the elementary or classical theory will first be briefly sketched, after which the rigorous mathematical theory will be discussed together with some of the important unsolved problems.

Historically wire transmission theory goes back to the early work of Kelvin and Heaviside. It is based on the simple idea that a transmission line (say consisting of two similar and equal wires in which equal and opposite currents flow) can be represented as consisting of uniformly distributed series inductance and resistance and shunt capacitance and leakance, these concepts deriving from electrostatics and elementary circuit theory. In accordance with this idea, if X denote the axis of propagation, the current I and voltage V are related by the familiar equations

$$\left(L\frac{d}{dt}+R\right)I = -\frac{\partial}{\partial x}V, \tag{1a}$$

$$\left(C\frac{d}{dt}+G\right)V = -\frac{\partial}{\partial x}I. \tag{1b}$$

Writing these in the usual form

$$\begin{aligned} ZI &= -\frac{\partial}{\partial x}V, \\ YV &= -\frac{\partial}{\partial x}I, \end{aligned} \tag{2}$$

where Z is the uniformly distributed series impedance and Y the shunt admittance per unit length, it is easy to show that I and V satisfy the differential equations

$$\left(\gamma^2 - \frac{\partial^2}{\partial x^2}\right) I = 0, \qquad \left(\gamma^2 - \frac{\partial^2}{\partial x^2}\right) V = 0, \tag{3}$$

where $\gamma^2 = ZY$. The solution of these equations is

$$I = Ae^{-\gamma x} - Be^{\gamma x}, \qquad V = kAe^{-\gamma x} + kBe^{\gamma x}. \tag{4}$$

$\gamma = \sqrt{ZY}$ is called the propagation constant and $k = \sqrt{Z/Y}$ the characteristic impedance of the line. A and B are integration constants which must be so chosen as to satisfy the boundary conditions (continuity of current and potential at the line terminals). The first term represents a direct wave, the second a reflected wave, their relative values depending on terminal reflections and the terminal impressed electromotive forces.

We see therefore that in accordance with elementary or classical transmission theory, the current and potential waves are both expressible as unique simple exponentially propagated direct and reflected waves, the values of which are determined by the continuity of current and potential at the line terminals. The characteristics of the line appear only through two parameters, the propagation constant γ and the characteristic impedance k.

Generalizing the preceding, consider a system of n parallel wires, parallel to the surface of the earth. The differential equations for such a system, in terms of elementary transmission theory, are [1]

$$\sum_{k=1}^{n} Z_{jk} I_k = -\frac{\partial}{\partial x} V_j \quad (j = 1, 2 \cdots n), \qquad \sum_{k=1}^{n} Y_{jk} V_k = -\frac{\partial}{\partial x} I_j \quad (j = 1, 2 \cdots n). \tag{5}$$

Here the physical system is represented by the parameters Z_{jk} and Y_{jk}, the Z parameters being the series impedances (self and mutual) and the Y parameters the shunt admittances. If the differential operator $\partial/\partial x$ is replaced by γ, thus confining attention to *exponentially* propagated waves, and if either the potential V or the current I is

[1] See references 9 and 10.

eliminated from (5), we get a set of n homogeneous equations in I or V, the determinant of which must vanish for a non-trivial solution. This determinant is a function of γ^2 and it has in general n roots in γ^2, indicating n possible modes of propagation, with corresponding characteristic impedances k_{jk}. The general solution is then of the form

$$\begin{aligned} I_j &= \sum A_{jk}e^{-\gamma_k x} - B_{jk}e^{\gamma_k x}, \\ V_j &= \sum k_{jk}A_{jk}e^{-\gamma_k x} + k_{jk}B_{jk}e^{\gamma_k x}. \end{aligned} \tag{6}$$

Here A_{jk}, B_{jk} are integration constants, the number of independent constants being $2n$. These are determined by the $2n$ boundary conditions at the physical terminals of the system; that is, the continuity of the n currents and n potentials. The solution represents n direct and n reflected current waves, which in general are propagated with different attenuations and different phase velocities.

The conclusions derivable from the classical theory sketched above may be summarized as follows: In a system of n parallel conductors there are in general n modes of propagation, that is, n direct and n reflected waves, which may be termed the normal modes of propagation. The distribution of the wave energy among the n modes of propagation or n component waves, is determined by the boundary conditions, which are essentially the continuity of the currents and potentials of the n wires. The system is supposed to be completely specified by the self and mutual series impedances and the self and mutual shunt admittances of the n conductors, while in the solution for the waves the physical and electrical characteristics of the line enter only through the propagation constants and corresponding characteristic impedances.

Before analyzing the theoretical basis of the preceding elementary theory, and showing its limitations, an interesting and practically important extension will be briefly touched on. The equations of the theory given above presuppose that the *impressed* electromotive forces are concentrated at the terminals of the system, and that in the line itself the electric and magnetic fields are due entirely to the currents and charges of the conductors, and consequently that the distribution of current and charge is determined entirely by their own fields. Suppose, however, that the system is in addition exposed throughout its length to an impressed field, from some disturbing source; then the preceding theory must be modified to take into account the effect of this additional field. To take the simplest case, consider a single wire parallel to the surface of the earth (ground return circuit). Let us suppose that this wire is exposed to an arbitrary field specified by an

axial electric force f at the surface of the wire and an impressed potential F (line integral of electric force to ground). The differential equations are then [2]

$$\begin{aligned} ZI &= -\frac{\partial}{\partial x} V + f, \\ YV &= -\frac{\partial}{\partial x} I - gF. \end{aligned} \tag{7}$$

(Here g is a shunt admittance. See reference 10.) Writing

$$\gamma = \sqrt{(Li\omega + R)(Ci\omega + G)},$$
$$k = \sqrt{\frac{Li\omega + R}{Ci\omega + G}},$$

this reduces to the differential equation

$$\left(\gamma^2 - \frac{\partial^2}{\partial x^2}\right) I = \frac{\gamma}{k} f + g \frac{\partial}{\partial x} F. \tag{8}$$

The solution of this equation and its practical significance have been discussed in a recent paper. The resulting analysis is of considerable practical importance in connection with the theory and design of the wave antenna and the problems of 'cross-talk' and induction.

The elementary or classical theory sketched above is essentially based on the simple concepts of electric circuit theory and its beautiful simplicity is a consequence of the fact that it is approximate only. For example the circuit parameters are only approximately calculable from the geometry of the system and its electrical constants and then only when the problem is treated as a two-dimensional one in which the variation of current and charge along the system is ignored as well as the finite velocity of propagation of their fields. Going further, it is by no means evident that even the *form* of the equations is rigorous. (We shall find that the form is rigorously valid only in an ideal case.) In the extension of elementary transmission theory, then, the first problem, as the writer sees it, is to examine the conditions under which the specification of the system by series impedance and shunt admittance parameters is justified; that is, to establish the conditions under which the classical *form* of the differential equations is valid. The second phase of this problem is to formulate a general method for calculating these circuit parameters in terms of the geometry and electrical constants of the system. The investigation of these problems leads to still further problems, arising from the fact

[2] See reference (10).

that the solutions of elementary theory, when valid, are only *particular solutions*, and therefore do not, in general, represent the complete wave.

In taking up this problem it is necessary to discard the simple concepts underlying classical transmission theory and attack the problem, *ab initio*, by aid of Maxwell's equation. Otherwise stated, our problem is to find solutions of the wave equation which satisfy the boundary conditions at the surfaces of the conductors, that is, the continuity of the tangential component of E and H, and therefore represent physically possible waves.

To put the matter otherwise, we shall place ourselves in the position of a mathematician, unacquainted with circuit theory or classical transmission theory, for whom the laws governing propagation of electromagnetic waves are formulated only by Maxwell's equations. His procedure in developing the theory of transmission along wires would be totally different from the way the theory has actually been developed. Starting with Maxwell's equations he would find that the electric and magnetic vectors satisfy a partial differential equation called the *wave equation*. He would then search for particular solutions of the wave equation which satisfy the geometry and electrical constants of the system, and therefore represent physically possible waves. The results of such a mode of approach to the problem are sketched below.

To formulate the problem concretely, consider a system of n parallel conductors, parallel to the (plane) surface of the earth, and extending along the positive X axis. The conductors may have any cross-sectional shape desired, but it is expressly assumed that they do not vary electrically or geometrically along the axis of transmission X (except at points of discontinuity or the terminals); that is to say, the transmission system is *uniform* along the axis of transmission.

Now in any medium of conductivity σ, permeability μ and dielectric constant ϵ, the electric and magnetic vectors satisfy the *wave equation*[3]

$$\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}-\nu^2\right)F=0, \tag{9}$$

where

$$\begin{aligned}\nu^2 &= 4\pi\sigma\mu i\omega-\omega^2/v^2,\\ v &= 1/\sqrt{\epsilon\mu},\\ \omega &= 2\pi \text{ times the frequency},\\ i &= \sqrt{-1},\end{aligned} \tag{10}$$

and F may be any component electric or magnetic vector.

[3] See reference (11).

We now suppose that solutions of the type

$$F = f(y, z)e^{(i\omega t - \gamma x)} \tag{11}$$

exist, where $f(y, z)$ is a two-dimensional wave function satisfying the two-dimensional wave equation

$$\left(\frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) f = (\nu^2 - \gamma^2) f. \tag{12}$$

In other words we search for *exponentially* propagated waves of this type; that is, waves which involve the spatial coordinate x only exponentially. It is well known that solutions of this type exist when the transmission system is uniform along the X axis.

The mathematical analysis of the problem outlined above is dealt with in detail in my paper 'The Rigorous and Approximate Theories of Electrical Transmission along Wires' (ref. 11) and the outstanding conclusions of that analysis are as follows:

The *form* of the differential equations of classical transmission theory is rigorously valid, that is, the system is specified rigorously by its self and mutual series impedances and shunt admittances, only for the ideal case of a system consisting of perfect conductors embedded in a perfect dielectric. In this case $\nu^2 - \gamma^2 = 0$ in the dielectric; $\nu^2 = \infty$ in the conductors, and the propagation constant γ is $i\omega/v$, indicating unattenuated transmission with the velocity of light, $v = 1/\sqrt{\epsilon\mu}$. The wave is a pure plane guided wave, and the electric and magnetic fields are derivable from two wave functions, one a linear function of the conductor charges and the other a linear function of the conductor currents, the determination of which, in terms of the geometry of the system, is reduced to the solution of a well-known potential problem.

Such a system, the ideal for guided wave transmission, is of course unrealizable, since there are always losses in both conductors and dielectric. For efficient transmission, however, the losses must be small and the guided wave must approximate the plane wave of the ideal case. Let us suppose, therefore, that the losses in the system are so small that *in the dielectric*, in the neighborhood of the conductors, we can set $\nu^2 - \gamma^2 = 0$, and that *in the conductors* the conductivity is so high that $\nu^2 - \gamma^2$ may be replaced by ν^2 without appreciable error. Under the circumstances where these approximations are valid it is found that the electric and magnetic fields in the dielectric and the current distribution over the cross-sections of the conductors are likewise derivable from two wave functions which are linear functions

of the conductor charges and currents respectively. The first of these is determined in terms of the geometry of the system by the solution of the same two-dimensional potential problem as in the ideal case, while the second is determined in terms of the geometry and electrical constants of the system, by a generalized two-dimensional potential problem.[4] Otherwise stated, to the approximations explained above the system may be regarded as specified by self and mutual series impedances and self and mutual shunt admittances, and these are calculable by the solution of the two-dimensional potential problems. The solution of the differential equations leads, precisely as in the classical theory, to an nth order equation in γ^2, indicating n modes of propagation. Moreover, the n corresponding waves, which will, for reasons explained below, be termed the *principal waves*, are *quasi-plane*. This means that, in the dielectric the axial electric intensity is in general small compared with the electric intensity in the plane normal to the axis of transmission; or, more broadly stated, the departure of the waves from true planarity is due entirely to dissipation in conductors and dielectric. A plane wave is here understood to mean a wave in which $E_x = H_x = 0$.

Now it is important to observe that in arriving at the foregoing result we have introduced at the outset approximations and assumptions regarding the order of magnitude of the propagation constant γ which depend on the assumption that the transmission losses are small. Fortunately these assumptions are justified, and the resulting approximate solutions are valid to a high degree of accuracy, in those systems which can be employed for the *efficient* guided transmission of electromagnetic energy; otherwise stated, the mathematical restrictions correspond to the actual requirements for efficient transmission. If, however, either the conductors or the dielectric become sufficiently imperfect, the approximations introduced and the resulting wave solutions become increasingly inaccurate and unreliable.

Suppose now that we attack the problem in a still more fundamental way: discard the assumptions regarding the order of magnitude of γ, introduced above, and attempt to deal with the problem and the solution of the wave equation in its general form. The case then is entirely different and vastly more complicated. In general, the solution can not be carried out, but a few simple systems have been studied and the results of this analysis may be generalized as follows:[5] in a system of n parallel conductors there exist, in addition to the n principal modes of propagation, an n-fold infinity of other modes of propagation,

[4] See reference (8).
[5] See reference (5).

which will be termed *complementary* modes of propagation. In general, the corresponding *complementary* waves differ from the *principal* waves in that they are not quasi-plane and are very rapidly attenuated. Consequently it appears that as regards the currents and charges, and the fields *near the conductors*, the effect of the complementary waves is usually appreciable only in the neighborhood of the physical terminals of the system so that at a distance from the terminals, usually small, they are represented with sufficient accuracy by the *principal* waves alone. At a great distance from the conductors, however, it appears that the errors resulting from ignoring the fields of the *complementary* waves may be large; in fact the complementary waves must be expressly included to take into account the phenomena of radiation.

The practical as distinguished from the theoretical importance of the foregoing resides in the fact that the principal waves corresponding to those of elementary theory represent the transmission phenomena accurately only at some distance from the physical terminals of the line and then only in the neighborhood of the wires. This defect may be of small practical consequence when the conductors all consist of wires of small cross section. When, however, conductors of large cross sections, or the ground, form part of the transmission system, the theory may be quite inadequate for some purposes. In particular, in calculating inductive disturbances in neighboring transmission systems at a considerable distance it may lead to large errors.

The discussion given above is based in part on a mathematical analysis of simple representative systems, in part on inferences from physical considerations. Unfortunately a direct frontal attack and rigorous solution of the general problem appears impossible. For example, in addition to finding the infinitely many modes of propagation the corresponding infinitely many complementary waves must be so chosen as to satisfy the boundary conditions at the physical terminals. In the classical theory these boundary conditions are simply the continuity of currents and potentials; in the rigorous formulation of the problem they are the continuity of E_y, E_z, H_y, H_z throughout the entire boundary plane ($x = 0$). Even to formulate these conditions involves specifying the impressed field throughout the plane and this is never given explicitly in technical transmission problems. While, therefore, the theory sketched above leads to inferences and conclusions of importance, the writer is convinced that some more powerful and indirect mode of attack on the problem must be devised; a rather hopeful possibility along this line will be briefly described.

As stated above, it is a reasonable inference from the general theory, that the complementary waves modify the *current* and *charge* waves appreciably only in the immediate neighborhood of the physical terminals, at least in most actual transmission systems. The essence of the method to be described consists of taking advantage of this fact and directly calculating the fields of the principal current and charge waves by means of their retarded potentials,[6] instead of employing for calculations the principal wave fields as given by the solution of the wave equation. This will now be explained in more detail.

In any transmission system energized by impressed forces introduced through terminal networks, the electromagnetic field may be analyzed as follows: (1) the impressed field, (2) the field of the terminal currents and charges, and (3) the field of the line currents and charges proper. The impressed field may be supposed to be concentrated in the terminal network, and the field of the terminal currents and charges may be supposed to be relatively unimportant except in the neighborhood of the terminals; what we are essentially concerned with is the field of the line currents and charges. Now let us suppose that we have calculated the principal wave in the system in the usual manner; corresponding to the resulting current and charge distribution, there will then be an unique corresponding field distribution determined by the solution of the wave equation, and this field is propagated in precisely the same way as the currents. But now suppose that we calculate the field of this current and charge distribution directly by means of their retarded potentials. We will find that the field so calculated is analyzable into two components: (1) a field identical with that given by the solution of the wave equation, and propagated in the same manner as the currents and charges, and (2) an additional field propagated in an entirely different way and for systems of small dissipation much more rapidly attenuated at least in the neighborhood of the conductors. We find further that the field of the principal current and charge wave does not correctly satisfy the boundary conditions at the surfaces of the conductors, which indicates that there must exist a compensating current and charge distribution. However, it appears that this compensating distribution will be relatively small and concentrated in the neighborhood of the terminals, so that we infer that its field, as calculated from its retarded potentials, can be ignored. Under such circumstances the inductive field (and the radiation field) is calculable by means of the retarded potentials in terms of the principal wave of current and charge alone.

[6] See reference (7).

To recapitulate this mode of attack, first determine the distribution of line currents and charges by means of elementary theory; that is, determine the principal wave distribution of currents and charges. Secondly, calculate the field of this current and charge distribution by means of the retarded potentials. This will give in addition to the field calculable from elementary theory an additional field the existence of which is not recognized by elementary theory. In brief, this mode of attack is based on the argument that the actual distribution of current and charge in the system is given with sufficient accuracy by elementary theory, but that in calculating the field at a distance, corrections must be introduced.

As might be expected this mode of attack presents formidable difficulties particularly when the ground plays an important rôle in the transmission phenomena. On the other hand, the analysis of a few of the simplest cases has been quite encouraging and leads one to hope that the method may at least be successfully applied to calculating the orders of magnitude of corrections which must be introduced in such important problems as, for example, inductive disturbances, in neighboring transmission systems.

The foregoing may appear to many as highly academic and theoretical. The writer's actual experience with practical transmission problems has convinced him, however, that the extension of wire transmission theory along the lines indicated above is urgently needed.

References

The papers listed below represent recent work which deals directly or indirectly with the problems discussed in the text. The relatively large number of the writer's own papers which are listed merely reflects the fact that very few specialists are working on the advanced problems of wire transmission theory.

1. "Radiation from Transmission Lines." (Carson, *Jour. A. I. E. E.*, Oct., 1921.)
2. "Radiation from Transmission Lines." (Manneback, *Trans. A. I. E. E.*, 1923.)
3. "A Generalization of the Reciprocal Theorem." (Carson, *B. S. T. J.*, July, 1924.)
4. "Das Reziprotät Theorem der drahtlosen Telegraphie." (Sommerfeld, *Jahrb. d. drahtl. Tel. u. Tel.*, 1925.)
5. "The Guided and Radiated Energy in Wire Transmission." (Carson, *Trans. A. I. E. E.*, 1924.)
6. "Über das Feld einer Unendlich langen Wechselstromdurchflossenen Einfachleitung." (Pollaczek, *E. N. T.*, 3, 1926.)
7. "Electromagnetic Theory and the Foundations of Electric Circuit Theory." (Carson, *B. S. T. J.*, Jan., 1927.)
8. "A Generalized Two-Dimensional Potential Problem." (Carson, *Bull. Am. Math. Soc.*, May–June, 1927.)
9. "Electromagnetic Waves, Guided by Parallel Wires." (Levin, *Trans. A. I. E. E.*, 1927.)
10. "Propagation of Periodic Currents over a System of Parallel Wires." (Carson and Hoyt, *B. S. T. J.*, July, 1927.)

11. "The Rigorous and Approximate Theories of Electrical Transmission along Wires." (Carson, *B. S. T. J.*, Jan., 1928.)
12. As a general reference, the treatise "Electrical and Optical Wave Motion," by Bateman, published by the Cambridge University Press, may be consulted with profit.

APPENDIX

The mode of attack outlined in the latter part of the text will be illustrated by an application to the simplest possible case.

Let the transmission system consist of a wire of radius a whose axis coincides with the X axis, and a coaxial cylinder of internal radius b. Both conductors are supposed to be perfectly conducting, while the dielectric in the space between ($a \leq \rho \leq b$) is supposed to be perfect. For this system we know that the principal wave is transmitted without attenuation with the velocity of light c; that is to say, $\gamma = i\omega/c$, where ω is 2π times the frequency.

We suppose that the system extends for an indefinite distance along the positive X axis so that reflected waves are absent. The principal current and charge waves are then:

$$I = I_0 e^{-i\beta x}, \qquad Q = Q_0 e^{-i\beta x}, \tag{1a}$$

where β denotes ω/c, and $i = \sqrt{-1}$. From the relation

$$\frac{1}{c}\frac{\partial}{\partial t} Q = -\frac{\partial}{\partial x} I$$

it follows that

$$Q = I. \tag{2a}$$

Now by definition the retarded potentials are

$$\Phi = \int \frac{q}{r} e^{-i\beta r} dv \quad \text{(Scalar)},$$

$$A = \int \frac{u}{r} e^{-i\beta r} dv \quad \text{(Vector)},$$

where q and u denote the charge and vector current density respectively, r is the distance between the contributing element dv and the point at which the potential is to be calculated, and the integration is extended over the entire system of currents and charges. In terms of the retarded potentials the magnetic and electric intensities E and H are given by

$$\begin{aligned} H &= \text{curl } A, \\ E &= -\text{ grad } \Phi - i\beta A. \end{aligned} \tag{3a}$$

To formulate the retarded potentials of the system under consideration we have recourse to the Sommerfeld integral

$$\frac{e^{-i\beta r}}{r} = \int_0^\infty J_0(\rho\lambda)e^{-|x-x'|\sqrt{\lambda^2-\beta^2}}\frac{\lambda d\lambda}{\sqrt{\lambda^2-\beta^2}}, \tag{4a}$$

where $\rho = \sqrt{y^2+z^2}$ and J_0 is the Bessel function in the usual notation.

Applying this integral to the system of currents and charges under consideration, and remembering that they are surface currents and charges at $\rho = a$ and $\rho = b$ respectively, we get without difficulty, for $x \geq 0$,

$$\begin{aligned}\Phi = Q_0\int_0^\infty J_0(\rho\lambda)[J_0(a\lambda) - J_0(b\lambda)]e^{-x\sqrt{\lambda^2-\beta^2}}\frac{\lambda d\lambda}{\sqrt{\lambda^2-\beta^2}} \\ \times\int_0^x e^{x'[\sqrt{\lambda^2-\beta^2}-i\beta]}dx' \\ + Q_0\int_0^\infty J_0(\rho\lambda)[J_0(a\lambda) - J_0(b\lambda)]e^{x\sqrt{\lambda^2-\beta^2}}\frac{\lambda d\lambda}{\sqrt{\lambda^2-\beta^2}} \\ \times\int_x^\infty e^{-x'[\sqrt{\lambda^2-\beta^2}+i\beta]}dx',\end{aligned} \tag{5a}$$

which reduces to

$$\begin{aligned}\Phi = 2Q_0e^{-i\beta x}\int_0^\infty J_0(\rho\lambda)[J_0(a\lambda) - J_0(b\lambda)]\frac{d\lambda}{\lambda} \\ - Q_0\int_0^\infty J_0(\rho\lambda)[J_0(a\lambda) - J_0(b\lambda)]\frac{e^{-x\sqrt{\lambda^2-\beta^2}}}{\sqrt{\lambda^2-\beta^2}-i\beta}\frac{\lambda d\lambda}{\sqrt{\lambda^2-\beta^2}}.\end{aligned} \tag{6a}$$

Since the currents are entirely axial, we have also $A_y = A_z = 0$, and from (2a)

$$A_x = \Phi. \tag{7a}$$

The first integral in Φ represents a potential wave propagated along the X axis in precisely the same way as the current and charge; it will therefore be termed the *homogeneous* potential wave. We find further that the field derivable from the *homogeneous* potentials is precisely the *principal wave* field, as given by the particular solution of Maxwell's equation, corresponding to $\gamma = i\beta$.

The second integral in Φ represents a potential wave propagated in an entirely different manner, and dying away for sufficiently large values of x. The corresponding field may be called, for want of a better term, the *heterogeneous* field, since its mode of propagation is quite different from that of the current and charge. It is this field

which represents the correction which must be added to the field of elementary theory.

It is beyond the scope of this brief appendix to discuss this solution in detail. It may be said, however, that, while the integrals representing the heterogeneous field can not be solved in finite terms, their properties can be approximately and qualitatively deduced without much difficulty.

One point of interest may be noted; the homogeneous wave is plane, that is, the axial electric intensity is everywhere zero. If we apply to the preceding formulas the relation

$$\begin{aligned} E_x &= -\frac{\partial}{\partial x}\Phi - i\beta A_x \\ &= -\left(\frac{\partial}{\partial x} + i\beta\right)\Phi, \end{aligned}$$

we get, for the heterogeneous field,

$$E_x = -Q_0 \int_0^\infty J_0(\lambda\rho)[J_0(\lambda a) - J_0(\lambda b)]e^{-x\sqrt{\lambda^2-\beta^2}} \frac{\lambda d\lambda}{\sqrt{\lambda^2 - \beta^2}}.$$

The integral term in this expression is simply the retarded potential of a ring of point sources located on the circle $x = 0$, $\rho = a$ minus the retarded potential of a corresponding ring of point charges located on the circle $x = 0$, $\rho = b$. Since this field does not vanish at the conductor surfaces $\rho = a$ and $\rho = b$, it is clear that a compensating charge and current distribution must exist.

Contemporary Advances in Physics—XV

The Classical Theory of Light, First Part

By KARL K. DARROW

FOR twenty years and more we have been hearing continually about the conflict between the corpuscular and the undulatory theories of light, and it is possible that for years to come we may be hearing about a similar contest between the wave-theory and the particle-theory of matter. Furthermore, there are intimations that if an adequate theory either of light or of matter ever is attained, it will involve conceptions of waves which in certain limiting cases approach to conceptions of particles. Already it is established that the appropriate way to attack the typical problems of the atom consists in setting up a wave-equation, and dealing with it in the same manner as one adopts to solve the typical problem of acoustics: how to determine the resonance-frequencies of a piece of elastic matter, such as a taut wire or a drumhead or a column of air in a tube. Therefore it seems opportune to restudy, with care and in detail, the great classical example of a wave-theory highly developed and widely successful—the great theory of light dimly foreshadowed by Huygens, endowed with its essential attributes by Young and Fresnel and Kirchhoff and a host of their coevals, utilized in the design of a multitude of ingenious instruments, perfected by Maxwell and connected with the theory of electricity and magnetism, and serving to this day as the basis for the theory of quanta. So doing, we shall be reminded of many triumphs of the past century of physical research, discoveries which in their time were as exciting as new quantum phenomena in ours; we shall notice certain achievements themselves as recent as those of quanta, and perhaps not less impressive; we shall retrace the reasonings which led to certain conclusions which the quantum-theories, unable to do without them and yet incompetent to derive them, have taken bodily over from their forerunner; we shall reconsider the evidence which in the litigation of a century ago caused the verdict to be rendered in favour of the wave-theory over the particle-theory; and perhaps incidentally we shall be drilling ourselves to test the evidence lately submitted and still to be submitted in the appeal of that case, and in the hearing of that other which impends.

For purposes of drill it might seem better to study the example of water-waves, which are visible; or sound-waves, of which no one denies

the existence, and no one wishes to supplant them with quanta. Ripples on the surface of a pond do furnish a precious example of wave-motion, and I presume that the notion of an undulatory theory was suggested originally by these; but it is precisely because they are visible that they fail to pose some of the questions which in dealing with light and matter are the most perplexing. Watching the leaves and the straws which float upon the surface of the water, one sees that they do not advance with the ripples; they are heaved up and down as the crests and the troughs of the wavelets pass them by. It is evident that the waves are not to be identified with the water; rather they are a form, a profile, a molding of the surface, which moves rapidly along while the substance of the liquid oscillates only a little. Now in this instance of the ripples on the pond, it is the relatively-immobile water which seems substantial and real, while that which is propagated as a wave appears to be merely a shape or a configuration, nothing more than a geometrical abstraction. It would seem strange and whimsical to assert that the liquid is a mere abstraction, but the waves are real. Yet we have to embrace this apparent absurdity in dealing with waves of light.

The example of sound-waves shows forth the paradox quite clearly. One can feel a tuning-fork; when it begins to act as a source of sound, one can see that it is quivering, and with a stroboscope one can even follow the actual course of its motion; it is even possible to see condensations and rarefactions travelling through the air, and there are numberless indirect ways of showing that sounding bodies and sound-transmitting media are matter in vibration. To the eye and to the hand, the body which vibrates is material and substantial, but not so its "vibration"—this word is only a way of saying that the shape, or the position, or the density of the body is undergoing a continuous and cyclic change.

It happens, however, that we also possess a sense for which the vibration is real but the vibrating substance is not. The ear takes no cognizance of the steel of which the tuning-fork is made, nor of the air which carries the undulations; but the ear perceives a tone. One must fully realize that the sense of hearing does not disclose that sound is vibratory. The ear does not report a sensation which goes through a cyclic variation two hundred and fifty-six times (or whatever the frequency of the fork may be) in every second. If it did, it would be perceiving the vibrating medium, not the vibration. The ear reports a sensation which is uniform, unvarying, constant; in fact, it translates a steady vibration into a constant sensation. This we have learned with ease, because of the collaboration of the other senses

which observe the bodies which oscillate. But if no one had ever felt or seen the quiverings of the humming fork, the ringing bell, or the resounding drumhead, we should be handicapped severely for discovering the true nature of sound.

Precisely so handicapped are we for discovering the nature of light. It may be that light is the vibration of a substance; but if so, the eye does not perceive that substance nor anything which fluctuates; it translates the vibration into a constant sensation. Moreover, we have no other sense which perceives that substance. When the filament of a lamp is incandescent, nothing is observed to pulsate on its surface; nothing is observed to go up and down or back and forth in the surrounding vacuum. Our instruments also fail to detect anything of which the vibrations are light. One may measure light with a photographic film or a bolometer; but the undulations—if such there be—are translated, in the one case into a steady rate of chemical change, in the other into a steady flow of electric current. In short, the eye and all our instruments register light as the ear registers tone, and not at all as the eye or the hand may register the quiverings of a sounding body; and therefore they do not report that light is vibratory. And if it be true that tangible matter is itself of the nature of a wave-motion, then the sense of touch must respond to these waves as the eye to light or the ear to sound, not reporting anything vibratory and not perceiving any medium which vibrates, but translating the vibrations into a constant sensation.

Therefore, to test whether light or matter or electricity is a wave-motion, one must make such experiments as could be made to test whether sound is a wave-motion, if there were no instrument able to perceive the vibrations of matter except the ear. Let us then suppose ourselves required to prove, to someone unable or unwilling to use any instrument except the ear, that sound is of the nature of waves; and consider how we should go about it.

The ear, we are told, is able to make distinctions of "loudness," "timbre," and "pitch." The two latter, interesting as they are, are of no immediate concern in this enterprise. It is sufficient to know that the ear makes distinctions in loudness, which according to the wave-theory correspond to distinctions in amplitude of vibration. Again, we have no concern with the exact relation between amplitude and loudness. What matters is that the former controls the latter, and therefore the latter reveals the former. Though the ear cannot detect the cyclic variations of the density and pressure of the air which make the sound, it can detect fluctuations in the amplitude of these cyclic variations. To put this statement into briefer language of

which, much later, there will be a reminiscence: the ear can detect the amplitude, but not the phase, of the vibrations. It follows, then, that we must devise tests of the wave-theory in which the amplitude of the waves shall vary from time to time, or from place to place.

Such tests are easily arranged. Let a pair of tuning-forks be set up not too close together and not too far apart. If sound consists of waves, the spherical undulations broadening outward from each of these separately must fall off steadily in amplitude as they recede, and the sound grow steadily fainter as the listener moves away. So it does; but the fact is equivocal, and cannot be taken as evidence for the waves; if the fork emitted corpuscles of sound, they would scatter apart as they flew away, and fewer would enter the ear the farther off it was placed. If, however, both of the forks are giving voice at once, and trains of spherical waves expanding outward from both, then the amplitude in the air must vary in a curious and striking manner from place to place, alternating between maxima and minima. This is just the sort of test which the ear is excellently fitted to make; being moved (or the mouth of its listening-tube being moved) from place to place in the field where the streams of sound overlap, it reports the fluctuations of loudness which are predicted from the wave-theory. By properly choosing the conditions one or more of the minima may be reduced to zero; loudness added to loudness makes silence. By properly choosing the conditions, maxima and minima may be caused to move in succession across a fixed point, listening whereat the observer hears "beats." All of these are phenomena of *interference*, and many like them are realized with light.

But it is not necessary to produce two overlapping streams of sound, in order to find evidence favouring the wave-theory. One suffices, provided that we try to separate from it a narrow jet or ray. Near one of the forks let a wall be placed, and perforated with a little hole. This seems to be an artifice for producing a constricted beam of sound proceeding like a searchlight straight outward along the line passing from the source through the hole; but it does not work that way. Instead, the tone of the fork is heard everywhere beyond the wall; sound is radiated from the hole in all directions. The aperture becomes itself a sort of secondary source, from which sound emanates sidewise as well as forward.

Precisely similar is the visible behaviour of water-waves (and incidentally of the violent compressional pulses produced in air by explosions, which have been photographed; but we are assuming that our imaginary pupil knows nothing of sound but what he hears!). Circular ripples expand over the surface of still water until one of them

meets a wall with a very narrow opening. Does a narrow segment of the ripple go clean through the opening and continue onward as a sharply-ended crescent? Not so; a new circular or semicircular ripple spreads out from the aperture as a new centre.

This is called, in the science of light, a phenomenon of *diffraction.* Actually, it is a phenomenon which reveals the law of wave-propagation—a law, which in the deceptively simple cases of spherical, circular, or infinite plane waves is artfully concealed. When one sees a circular ripple broadening over the surface of a lagoon, it seems as if each arc of the circular crest were advancing independently and of its own momentum; as if each segment of the circle at a given moment were due entirely to the corresponding segment of the smaller circle which existed a fraction of a second earlier. Nothing could be further from the truth. At a given moment, a given segment of the circle is due to the collaboration of all the segments of the earlier smaller circle; and it will collaborate with all the segments of its own circle to build the future yet larger one; and if isolated from the rest of the circumference, it would build a new family of circular ripples all by itself. Somewhat as the primitive animals which can regenerate their amputated parts, a wave-system seems to possess in each of its elements something of the power to build itself anew.

Such is the nature of ripples on water and sound-waves in air. As for a general definition of wave-motion, perhaps there is no better way of making one than to accept this manner of propagation as the distinctive mark. It may seem strange, however, that there should be any question about definition. Does not everyone know what a wave is? and is not the difference between a wave-theory and a corpuscle-theory made instantly clear by their names?

Well! it would not be hard to compile a series of paradoxical statements, by which to show that our immediate off-hand notion of a "wave" is not by any means sufficiently precise to serve as basis for an elaborate physical theory. Even in the ancient and familiar instance of circular ripples on water, even for students acquainted with the concepts of wave-length and wave-speed, there are possibilities of confusion. It is not expedient to define the wave-length as the distance from one crest to the next, for this is inconstant. It is not expedient to define the wave-speed as the speed with which a crest advances, for this may depend upon the form of the wave. It is injudicious to think exclusively about the profile of the water-surface as a sequence of visible elevations and depressions gliding steadily onward without change of shape; for any part of the profile may alter itself incessantly as it advances, departing more and more

from its original contour till it becomes unrecognizable. Definite as the wave-length and the wave-speed and the waves themselves may seem at times to be, at other times they seem indefinite and indefinable.

Now in the general theory these difficulties are removed, for the attention is focussed first of all upon an abstract entity with a neutral name—the "phase." There is a differential equation, a "wave-equation," governing the phase; and this entity is propagated in a certain very definite way conforming to the vague description which I gave above, and it has a wave-length and a wave-speed. As for the elevations and depressions of the water-surface, they copy the variations of the phase more or less faithfully, and may be computed from these; but except in particular cases the copy is not exact. To the theorist, the ripples upon the water appear as the secondary and imperfect manifestations of an abstract wave-motion, discernible only to the eye of the mind.

That the undulatory theory should introduce an abstraction even into the example whence it sprang, requiring us to imagine waves of phase underlying the tangible waves of the sea, is not at all remarkable. Often in physics a theory evolves in this way. It begins when some one notes a resemblance between two or more phenomena; it continues by the invention of a neutral and colorless mathematical expression for describing the common aspect of all these phenomena; and then the theorists take over the mathematical expression and transform and generalize and extend it, until the theory, which at first was a casual statement that two different things are in some ways much alike, eventually is defined as the entire system of solutions of some differential equation. At present there seems to be no adequate way of defining the term "wave-theory," except to say that wherever a certain differential equation is introduced and solved, there a wave-theory is adopted. Moreover, it is open to anyone to adduce new differential equations more or less like the first one, and define as a wave-theory any which involves the solutions of these. Then a wave-motion is any motion which conforms to one of the solutions; and when it comes to defining a wave, what can anyone say except that under certain restricted conditions a wave-motion may resemble a procession of ripples on water?

Such a consummation may be devoutly wished by the mathematician; to the physicist and to the expositor it is not always so welcome. As a theory increases in scope through increase of abstraction, it loses the picturesqueness which for many minds is its reason for being. One climbs and climbs, and the view indeed grows wider, but the

fascinating details of the landscape are distorted when seen from above, and finally they are lost in the haze of distance. It grows more difficult to lead others to the heights, and sometimes even the explorer cannot retrace his path and return to the firm ground of experience whence he departed. Yet repeatedly in the evolution of physics it happens that a theory, already grown so abstract that it seems almost completely severed from reality, suddenly makes new contact with the world of phenomena by a prediction so novel and daring that except for the far preliminary excursion it would probably never have been conceived; as for instance the existence of quanta, the "Einstein shift" of the lines in the spectrum of the sun, the diffraction of electrons by crystals. Remembrance of such episodes as these is an encouragement, when the path seems devious and steep.

Propagation of Waves

The laws of the propagation of waves—the so-called "laws of diffraction"—are the most important topic with which we have to deal; for they involve the very nature and definition of wave-motion, and in the end the distinction between a corpuscular and an undulatory theory of matter may rest upon these. Let us attend first of all to the making of this distinction.

Imagine, then, a multitude of particles—bullets, or atoms, or sand-grains—all rushing along through space in the same direction with the same speed, say northward with the speed c. Suppose that the location of each is stated for a certain moment of time, say t_0. The question to be asked is a very simple one, of the yes-or-no variety. At an arbitrarily-chosen point P, at an arbitrarily-chosen moment t, will there be a grain of sand or will there not?

It is easy to see what determines the answer. If at the point P at the moment t there is a grain of sand, it must have spent the time-interval extending from t_0 to t in travelling northward along the straight south-to-north path which ends at P, and therefore commences at a point P_0 due south of P and distant from it by $c(t - t_0)$ units of length. If therefore at the moment t_0 there is a grain of sand at P_0, the answer to the question is *yes*. Otherwise it is *no*. No other knowledge is required, or even relevant. It is not necessary to know the location of any of the particles which are not upon the north-south line traversing P. It is not even necessary to know the location of any particle which is upon that line, provided that at the instant t_0 it is surely somewhere else than at P_0. The state of affairs in P at t is controlled by the state of affairs in P_0 at t_0, and by nothing else whatever.

Let the same question be put in another way. Be it supposed that we are required to predict whether or not there will be a particle in the place P at the moment t; and that we are offered our choice of data concerning the places of the particles at any prior moment. Let us choose at random some point P' due south of P, distant from it by r' units of length. Then there is only one piece of information for which we have to ask: is there a sandgrain passing through P' at the moment $(t - r'/c)$, earlier than t by r'/c units of time? Any other information would be not only superfluous, but useless. Had we chosen some point lying south of P at a distance r'', the condition prevailing there at the moment $(t - r'/c)$ would have had no bearing upon the problem; but the condition there at the moment $(t - r''/c)$ would have been all-powerful. Had we chosen some point not lying south of P, nothing happening there at any time would have had any bearing upon the question to be answered.

In fine: at any moment t' there is a corresponding point P' lying south of P, which holds the destiny of the point P at the moment t. That which is predestined to befall in P at t is at every prior instant concentrated, so to speak, at a particular point of space. As time draws on towards t, this point moves on toward P, travelling always along the north-south line—travelling always, let us say, along a certain *ray* which at the proper moment carries it right into P.

All these remarks may seem too evident and trivial to be worth the making; yet they deserve attention, for it is here that the contrast lies between motion of particles and motion of waves, between undulatory theories and corpuscular. If the region around the point P is traversed not by corpuscles but by waves, it is not correct to say that the condition at the point P at the moment t is determined by the condition in some other *point* at some prior moment. Even if the waves appear to be travelling northward with the constant speed c, it is not right to say that the state of affairs prevailing in P at t is controlled entirely by the state of affairs prevailing at the moment $(t - r/c)$ at the point r units southward from P. The destiny of P at t is not travelling towards it concentrated into a point moving along a ray. Under some circumstances it appears to be right to say so; but this is only a semblance, as experiments in other conditions will clearly prove.

Suppose for instance that one is confronted with the task of sheltering the point P, first against corpuscles and then against waves, which are advancing from the south. It seems natural to put some obstacle athwart that particular north-south line which traverses P; for example, to place a solid disc so that its axis lies upon that line. If the disc can arrest all the particles which fly towards it, and cannot deflect

those which do not, then no matter how small it is it shields the point P completely against corpuscles. Not, however, against waves. Suppose that the point P is in water, where the actual waves may be seen; suppose that before the obstacle is dipped in, each of the wave-crests extends straight east and west, and they move straight northward. When the obstacle is inserted southward from P, the water at P does not become perfectly quiet. Apparently the waves curl around the edge of the obstacle, invading the zone behind it which it could have protected perfectly against corpuscles. One cannot stop a wave-motion from reaching a point merely by interrupting with some small obstruction the line along which the waves seem to be approaching it.

Now what this means is simply that, whether the obstacle be present or absent, and even though the undisturbed wave-crests move steadily due northward, the motion at P is not controlled exclusively by the motions at earlier moments at the points due south of P. To put it a little more loosely: the wave-motion at a point arrives not solely from the direction from which the wave-fronts appear to be coming, but from all directions. To put it much more strictly: imagine a sphere drawn, with any radius r, around P as centre. When we were dealing with corpuscles, we found that the state of affairs at the centre of this sphere at the moment t was entirely controlled by the state of affairs at the moment $(t - r/c)$ at one single point on the sphere (the point due south from P). Now that we are dealing with waves, we shall find that the state of affairs at the centre of the sphere at t depends upon the state of affairs all over the sphere at $(t - r/c)$. Every point upon the sphere influences the centre. Every point in the medium which the waves traverse sends forth an influence to every other point; the influence is not instantaneous, but travels from one point to another with the wave-speed c. This "influence" is often called the *wavelet*.

Too much emphasis has been laid, in the foregoing passage, upon the spheres which are centred at P; and this must now be rectified. Any closed surface whatever may be drawn around P, and the state of affairs in P at t will be determined by the state of affairs prevailing all over this surface, S, at certain prior moments t'; only, since the area-elements of S are not in general equidistant from P, the corresponding values of t' are not in general the same for all of them. The distance r, measured from P along any direction to the surface S, is in general a function of direction; consequently the time-interval, r/c, required for a wavelet to arrive at P from S along any direction is itself a function of direction; and so also is t', which is $(t - r/c)$. To every point P' on S corresponds its own value of t'; and if we know the wave-motion in every P' at its proper t', we can determine the wave-motion

in P at t. Therefore, if there is any closed surface in space, everywhere over which the wave-motion is known for all times, it is possible to compute the wave-motion at any point in the volume which that surface encloses.*

This is a feature common to all the familiar examples of wave-motion, and it is suitable for a tentative basis for a general definition of waves.

To formulate it strictly, let s be used as the symbol for any quantity which is propagated in waves. Examples of such a quantity are: the twist of a taut and twisted wire—the lateral displacement of a taut wire or a tense membrane—the excess of the pressure in the air over its average value—a component of the electric field-strength or the magnetic field-strength in a vacuum—the entirely imperceptible and hypothetical entity denoted by Ψ in wave-mechanics.

We write s as a function of x, y, z, and t:

$$s = s(x, y, z, t). \tag{1}$$

Fewer than three dimensions of space will suffice in some cases (e.g., those of the wire and the membrane); in certain problems of wave-mechanics, more than three may be required; but in dealing with sound in air and light in vacuo, three are usually necessary and sufficient. For the time being I will suppose that the speed of the waves is everywhere the same. Interesting things will happen when this assumption is discarded.

What I have loosely called "the state of affairs" in a point $P(x, y, z)$ at a moment t will involve the value of s at x, y, z, and t. Also it may involve the first and higher derivatives of s with respect to space and time, evaluated at x, y, z, t. Which of these derivatives we are required to know is something which might vary from case to case. For the present, we may consider ourselves required to know s and its first derivatives ds/dx, ds/dy, ds/dz, ds/dt.

We are to evaluate s and its derivatives at a point P at a moment t, in terms of the values which s and its derivatives possessed at certain earlier moments over a surface S enveloping P.

Let P be made the origin of our coordinate-system; let x, y, z denote the coordinates of the points on the surface S; let r denote the distance from the origin to any of these points, so that:

$$r^2 = x^2 + y^2 + z^2. \tag{2}$$

Introduce as an auxiliary the function U, defined thus:

* Naturally the surface must not be so drawn that it includes sources emitting waves during the time-interval $(t' - t)$.

$$U(x, y, z, t) = s(x, y, z, t - r/c). \tag{3}$$

The value of U in any point of the surface S at the moment t is the value of s which prevailed in that point at the moment when the "wavelet" started forth which was destined to reach the origin at t. It might be said that an observer, stationed in the origin at the moment t and inspecting the surface by means of the wavelets, observes the values of U instead of the contemporary values of s. Thus a star-gazer viewing the sky perceives, not the stars as they now are or as at some one past moment they all were, but each star separately as it was at some past epoch peculiar to itself; and the apparent arrangement of the heavenly bodies is one which in fact has never existed.

We shall be concerned not only with the value of s, but with the values of the space-derivatives ds/dx, ds/dy, ds/dz, which prevail at each point of the surface at the moment when the wavelet starts forth; for all of these will influence the value of s at the origin when the wavelet arrives there. These may be written as derivatives of U; but one must be careful here, for U is a function of x, y, z not only explicitly, but also implicitly through r; and there is a distinction to be made between total and partial derivatives, a distinction having physical importance.

To grasp this, denote by (x, y, z) the coordinates of some particular point on S, and by $(x + dx, y, z)$ those of a nearby point, and by r and $r + dr$ their respective distances from the origin, and by U and $U + dU$ the values of U in these points at the instant t. Now, U and $U + dU$ are values of s which existed *at different instants of time*, as may be seen by writing down the expressions:

$$U = s(x, y, z, t - r/c); \; U + dU = s(x + dx, y, z, t - \overline{r + dr}/c). \tag{4}$$

Therefore, if I form the total derivative dU/dx in the classical way, I am *not* obtaining the value of ds/dx which prevailed in (x, y, z) at $(t - r/c)$. To obtain this value, I must begin by subtracting the value of s prevailing in (x, y, z) at $(t - r/c)$ from the value of s prevailing in $(x + dx, y, z)$ at the same moment; that is, I must form the difference between $(U + \partial U)$ and U, meaning by the former symbol:

$$U + \partial U = s(x + dx, y, z, t - r/c). \tag{5}$$

I must then divide this difference by dx, and pass to the limit. But this is the classical way of forming the partial derivative of U with respect to x. Therefore the values of the derivatives ds/dx, ds/dy, ds/dz prevailing at the moment of departure of the wavelet which is destined to reach the origin at t, are the partial derivatives $\partial U/\partial x$,

$\partial U/\partial y$, $\partial U/\partial z$. However, the value of the derivative ds/dt prevailing at the moment when the wavelet starts is simply the derivative dU/dt, which we may as well write $\partial U/\partial t$—it makes no difference.

Our definition of wave-motion may now be stated more rigorously. A quantity s is said to be propagated by waves, if its value at the origin at the moment t is determined by the values of U, $\partial U/\partial x$, $\partial U/\partial y$, and $\partial U/\partial z$ over any surface enveloping the origin.

We now turn to another and more familiar definition of wave-motion, which shall presently be shown to fall as a special case under this one.

The Wave-Equation

There is a very celebrated differential equation of mathematical physics, known as "the wave-equation" *par excellence*. Any theory which culminates in this equation is designated as a wave-theory. The foundation of the theory of sound is the proof that the excess of pressure in the air over its average value is subject to this equation. The elastic-solid model of the luminiferous æther was partially suited to explain the phenomena of light, because the compressions and the distortions of an elastic solid conform to the wave-equation. The electromagnetic theory of light was born when Maxwell discovered interrelations between electric and magnetic fields, out of which by transformation a wave-equation could be formed. Undulatory mechanics is based upon an equation of this type which emerges during the process of setting and solving the classical equations of motion.

This wave-equation is:

$$c^2\left(\frac{d^2s}{dx^2}+\frac{d^2s}{dy^2}+\frac{d^2s}{dz^2}\right)=\frac{d^2s}{dt^2}\cdot \tag{6}$$

To demonstrate why it is called a wave-equation and what is the physical meaning of the constant c, it is customary to make a drastic simplification by assuming that the function s depends only on one co-ordinate. Such is the case, for instance, when s stands for the transverse displacement of an endlessly long taut string initially parallel to the axis of x; likewise, when it stands for the excess of the pressure of the air over its average value, and this excess is constant over every plate normal to the x-direction—a condition known as that of "plane waves." Then the wave-equation assumes the form:

$$c^2\frac{d^2s}{dx^2}=\frac{d^2s}{dt^2}\cdot \tag{7}$$

There are infinitely many solutions of this equation, and among them

are all [1] the functions of the pair of variables x and t, in which these variables appear coupled together into the linear combination $(x - ct.)$ Using f as the general symbol for a function, we may write

$$s = f(x - ct). \tag{8}$$

When such a relation prevails, any value of s which occurs at a given place x at a given moment t recurs at any other moment t' at another place x', distant from x by the length $(t' - t)/c$. *All* of the values of s existing at t are found again in the same order at t', but they have all glided along the x-direction through the same distance $(t' - t)/c$. The form, the profile, the configuration of the string are moving along with the speed c, although the substance of the string is oscillating only a little, and not even parallel to the x-direction. Now this is the property which to a certain degree of approximation ripples on water display; this in fact supplies the elementary and restricted definition of wave-motion, out of which by generalization and extension the wave-theory has grown.

Thus we see that there is reason for calling (7) a wave-equation, and identifying the constant c with the speed of the waves. Yet there are also solutions of (7) which are not of the form (8), and these do not correspond to an unchanging profile of the string travelling along with a constant speed, though by mathematical artifice they may be expressed as a summation of such; and nothing is easier than to find solutions in two or three dimensions of the general equation (6) which do not bear the least resemblance to a regular procession of converging, flat, or diverging waves. The question then arises: is there a feature common to all solutions of the "wave-equation" fitted to serve for a general definition of wave-motion?

I will now show—in the manner of Kirchhoff and Voigt—that there is such a feature, and it is precisely the one already proposed as a definition for wave-motion. If s be a function conforming to (6), and U a function related to s according to (3), then the value of s at any point at any moment is determined by the values of U and its partial derivatives $\partial U/\partial x$, $\partial U/\partial y$, $\partial U/\partial z$, over any surface surrounding that point.* The proof is long and intricate; bur for anyone who desires appreciate the nature of wave-motion, it is not superfluous.

To prove the theorem we have to manipulate the vector (call it W) of which the components are:

[1] Exceptions being made for functions which do not have derivatives, and other curiosities of the mathematicians' museum.

* The necessary requirements for continuity in s exclude sources of light from the region of integration.

$$W_x = \frac{1}{r}\frac{\partial U}{\partial x}, \qquad W_y = \frac{1}{r}\frac{\partial U}{\partial y}, \qquad W_z = \frac{1}{r}\frac{\partial U}{\partial z}. \tag{9}$$

Like U, it is a function of x, y, z not only explicitly, but also implicitly through r; we must therefore discriminate with care between total and partial derivatives. For reference, here are the formulæ[2] connecting derivatives of the one type with those of the other:

$$\frac{d}{dx} = \frac{\partial}{\partial x} + \frac{\partial r}{\partial x}\frac{\partial}{\partial r} = \frac{\partial}{\partial x} + \frac{x}{r}\frac{\partial}{\partial r} = \frac{\partial}{\partial x} + \cos(x, r)\frac{\partial}{\partial r}, \tag{10 a}$$

$$\frac{d}{dy} = \frac{\partial}{\partial y} + \frac{\partial r}{\partial y}\frac{\partial}{\partial r} = \frac{\partial}{\partial y} + \frac{y}{r}\frac{\partial}{\partial r} = \frac{\partial}{\partial y} + \cos(y, r)\frac{\partial}{\partial r}, \tag{10 b}$$

$$\frac{d}{dz} = \frac{\partial}{\partial z} + \frac{\partial r}{\partial z}\frac{\partial}{\partial r} = \frac{\partial}{\partial z} + \frac{z}{r}\frac{\partial}{\partial r} = \frac{\partial}{\partial z} + \cos(z, r)\frac{\partial}{\partial r}, \tag{10 c}$$

$$\frac{d}{dr} = \frac{\partial}{\partial r} + \frac{\partial x}{\partial r}\frac{\partial}{\partial x} + \frac{\partial y}{\partial r}\frac{\partial}{\partial y} + \frac{\partial z}{\partial r}\frac{\partial}{\partial z}$$

$$= \frac{\partial}{\partial r} + \cos(r, x)\frac{\partial}{\partial x} + \cos(r, y)\frac{\partial}{\partial y} + \cos(r, z)\frac{\partial}{\partial z}. \tag{10 d}$$

The procedure consists in forming the expression for the true divergence of W, to wit:

$$\operatorname{div} W = \frac{dW_x}{dx} + \frac{dW_y}{dy} + \frac{dW_z}{dz}, \tag{11}$$

and integrating it over the volume comprised between two surfaces: outwardly, the surface S over which the values of U are preassigned, and which envelops the origin at which the value of s is to be computed; and inwardly, an infinitesimal sphere centred at the origin.

It will turn out that the volume-integrals of the various terms either vanish, or else may be converted into area-integrals over the two surfaces. Now the area-integral of any function f over the surface of a sphere of radius R may be written as

$$A = 4\pi R^2 \bar{f}, \tag{12}$$

in which $\bar{f}$ stands for the mean value of f over that surface—a statement

[2] In deriving the first three of these formulæ, use the relation $r^2 = x^2 + y^2 + z^2$ in evaluating $\partial r/\partial x$, $\partial r/\partial y$, $\partial r/\partial z$. In deriving the last, remember that in forming a derivative with respect to r at a point P, the increment dr is always measured along the line extending from the origin through P, for which line $x/r = \cos(x, r) = \text{const.}$; $y/r = \cos(y, r) = \text{const.}$; $z/r = \cos(z, r) = \text{const.}$; hence $\partial x/\partial r = \cos(x, r)$, etc. Or one may arrive by geometrical intuition at the formula,

$$d/dr = \cos(r, x)\,d/dx + \cos(r, y)\,d/dy + \cos(r, z)\,d/dz,$$

from which (10 d) may be obtained by means of (10 a, b, c).

which, of course, is merely the definition of f. The essential thing is that if the sphere is infinitesimal, then in the limit $\bar{f}$ becomes the value f_0 which the function f possesses at the centre of the sphere. If f_0 is finite at the origin, A vanishes in the limit; but if f varies inversely as the square of the distance from the origin, A approaches in the limit a finite value differing from zero. Upon this property our demonstration will depend.

Developing by means of (10 *a*, *b*, *c*) the expression given in (11) for div W, we find:

$$\begin{aligned}\operatorname{div} W = \frac{1}{r}\left(\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2}+\frac{\partial^2 U}{\partial z^2}\right)\\ +\frac{1}{r^2}\left(x\frac{\partial^2 U}{\partial x\partial r}+y\frac{\partial^2 U}{\partial y\partial r}+z\frac{\partial^2 U}{\partial z\partial r}\right)\\ -\frac{1}{r^3}\left(x\frac{\partial U}{\partial x}+y\frac{\partial U}{\partial y}+z\frac{\partial U}{\partial z}\right).\end{aligned} \tag{13}$$

The second and third terms on the right may next be beneficially transformed by means of (10 *d*), using first U and then $\partial U/\partial r$ as the argument of the derivatives in that equation:

$$\frac{x}{r}\frac{\partial U}{\partial x}+\frac{y}{r}\frac{\partial U}{\partial y}+\frac{z}{r}\frac{\partial U}{\partial z}=\frac{dU}{dr}-\frac{\partial U}{\partial r}, \tag{14 a}$$

$$\frac{x}{r}\frac{\partial^2 U}{\partial x\partial r}+\frac{y}{r}\frac{\partial^2 U}{\partial y\partial r}+\frac{z}{r}\frac{\partial^2 U}{\partial z\partial r}=\frac{d}{dr}\frac{\partial U}{\partial r}-\frac{\partial^2 U}{\partial r^2}, \tag{14 b}$$

and so finally we arrive at

$$\operatorname{div} W=\frac{1}{r}\left(\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2}+\frac{\partial^2 U}{\partial z^2}-\frac{\partial^2 U}{\partial r^2}\right)+\frac{1}{r^2}\frac{d}{dr}\left(r\frac{\partial U}{\partial r}-U\right) \tag{15}$$

as the expression to be integrated over the volume between S and the infinitesimal sphere.

Now *owing to the nature of the function U, the first term of the expression vanishes.* This is responsible for our theorem; for the volume-integrals of the remaining terms can easily be translated into surface-integrals over S and the infinitesimal sphere, from which it will follow that the value of s at the origin is determined by the values of U and its derivatives over S; but if this first term should remain, its volume-integral could not be thus transformed, and we should find that the value of s at the origin was influenced by the values of U all through the space which S encloses.

That the term in question does actually vanish is easily proved. For on the one hand it follows, from the coupling of t and r into the linear combination $(t - r/c)$ in the argument of U, that

$$\frac{\partial^2 U}{\partial t^2} = c^2 \frac{\partial^2 U}{\partial r^2}, \tag{16}$$

and on the other hand it follows, from the facts that the partial derivatives of U at any point and moment have the same values as the corresponding derivatives of s at the same point at some other moment, while the derivatives of s at *every* point and moment conform to (6)—from these it follows that

$$\frac{\partial^2 U}{\partial t^2} = c^2 \left(\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} + \frac{\partial^2 U}{\partial z^2} \right). \tag{17}$$

Therefore the first term of the right-hand member of (15) is zero everywhere, and we have to perform the volume-integration only over the second:

$$\int \operatorname{div} W dV = \int \frac{1}{r^2} \frac{d}{dr} \left(r \frac{\partial U}{\partial r} - U \right). \tag{18}$$

Employ spherical coordinates for the integration; then the element of volume is $dV = r^2 \sin \theta d\theta d\varphi dr$, and we have:

$$\begin{aligned} \int \operatorname{div} W dV &= \int d\theta \sin \theta \int d\varphi \int dr \frac{d}{dr} \left(r \frac{\partial U}{\partial r} - U \right) \\ &= \int d\theta \sin \theta \int d\varphi \left[\left(r \frac{\partial U}{\partial r} - U \right)_S - \left(r \frac{\partial U}{\partial r} - U \right)_R \right]. \end{aligned} \tag{19}$$

This signifies that the long narrow volume-element comprised within any elementary solid angle $dw = \sin \theta d\theta d\varphi$, and limited at its two ends by the surface of the sphere and the surface S, contributes to the volume-integral the difference between the values of $(r\, \partial U/\partial r - U)$ at its two extremities. Completing the integration by considering all of these volume-elements together, we see that the volume-integral therefore becomes a pair of angle-integrals, those of the function $(r\partial U/\partial r - U)$ over the surface S and over the sphere. We may transform the first of these into an area-integral by reflecting that the elementary solid angle dw intercepts upon the surface S the area-element dS, given by the equation:

$$-dS \cos (n, r) = r^2 dw. \tag{20}$$

in which (n, r) stands for the angle between the normal to dS and the radius r drawn to dS from the origin. We must choose positive

directions for these lines. Let the radius be taken as pointing outward, and the normal as pointing inward towards the volume over which we have integrated. Then the angle is greater than 90° and not greater than 180°, its cosine is negative, and the negative sign must be prefixed to the left-hand member of (20) that dS may be positive. We make this transformation in (19), and the first of the angle-integrals becomes:

$$\int d\theta \sin\theta \int d\varphi \left(r\frac{\partial U}{\partial r} - U \right)_S = \int_S dS \cos(n, r) \left(\frac{1}{r}\frac{\partial U}{\partial r} - \frac{U}{r^2} \right). \quad (21)$$

The second of the angle-integrals in (19) relates to the infinitesimal sphere. We transform it as we did the first. Now, however, the process is more simple, for the radius r is constant and equal to R, and the angle (n, r) is zero; hence

$$dS = R^2 \sin\theta d\theta d\varphi \quad (22)$$

and the second angle-integral becomes:

$$\int d\theta \sin\theta \int d\varphi \left(r\frac{\partial U}{\partial r} - U \right)_R = \int dS \left(\frac{1}{R}\frac{\partial U}{\partial r} - \frac{U}{R^2} \right). \quad (23)$$

Here we meet the situation for which equation (12) was introduced—the integration of a function over an infinitesimal sphere. Denote by f the integrand in (23), viz.,

$$f = \frac{1}{R}\frac{\partial U}{\partial R} - \frac{U}{R^2}, \quad (24)$$

by $\bar{f}$ the mean value of f over the sphere of radius R, by U_0 the value of U at the centre of the sphere, which is the origin. As R approaches zero, the surface-integral in (23) approaches a limit A_0 which coincides with the limit approached by $4\pi R^2 f$:

$$A_0 = \lim_{R=0} 4\pi R(\overline{\partial U/\partial R}) - 4\pi U_0. \quad (25)$$

Unless the mean value of $\partial U/\partial R$ should vary as the first or a higher power of $(1/R)$—a possibility which must be guarded against—the first term on the right of (25) will vanish. Under this restriction, then, A_0 is equal to $-4\pi U_0$. Now at the origin U is identical with s, by definition (equation 3). Consequently U_0 is identical with the value of s at the origin at the moment t—the very thing which we set out to calculate. For this—let it be called s_0—we have attained the following equation:

$$4\pi s_0 = \int \operatorname{div} W\, dV + \int dS \cos(n, r)(\partial U/\partial r). \quad (26)$$

We still have a volume-integral in the formula; but there is a very noted theorem whereby it may be transformed with sign reversed into a surface-integral over the two surfaces, S and the sphere, which bound the region of integration. According to Gauss' Theorem, any vector function satisfying certain simple conditions of continuity throughout a region enclosed by a surface enjoys this property: its volume-integral through the region is equal to the area-integral, over the enclosing surface, of the projection of the vector upon the direction of the *outward*-pointing normal. This latter is the same in magnitude and opposite in sign to the projection upon the direction of the *inward*-pointing normal, which it is traditional to prefer. The theorem is not valid, if the vector should exhibit certain singularities within the volume; one of the reasons for introducing the infinitesimal sphere is that the vector W has a singularity at the origin, which point must therefore be excluded from the volume of integration.

Remembering the definition of W (equation 9), we see that its projection upon the direction of the *inward*-pointing normal at any place upon either surface is

$$\begin{aligned} W_n &= W\cos(n, r) = W_x\cos(n, x) + W_y\cos(n, y) + W_z\cos(n, z) \\ &= \frac{1}{r}\left[(\partial U/\partial x)\cos(n, x) + (\partial U/\partial y)\cos(n, y) + (\partial U/\partial z)\cos(n, z)\right] \\ &= \frac{1}{r}(\partial U/\partial n), \end{aligned}$$

in which $(\partial U/\partial n)$ stands for the rate at which the function U, owing to its *explicit* dependence upon x, y, and z, varies as one moves *inward* along the normal to the surface. The distinction drawn in the foregoing pages between partial and total derivatives must be remembered. The partial derivative $\partial U/\partial n$ existing at any point P and moment t is equal to the value which the corresponding derivative ds/dn of the function s possessed at that same point at the earlier moment $(t - r/c)$.

The quantity W_n is to be integrated over the surface of the sphere and over the surface S. However, the integral over the sphere vanishes as the radius of this latter approaches zero, for the same reason—and under the same restriction—as caused the integral of the first term in (25) to vanish. This leaves us with nothing but the integral of W_n over the surface S, so that eventually:

$$\int \operatorname{div} W\, dV = -\int_S \frac{1}{r}(\partial U/\partial n)\, dS. \tag{28}$$

$$4\pi s_0 = \int_S dS\left[\cos(n, r)\frac{\partial}{\partial r}\left(\frac{U}{r}\right) - \frac{1}{r}\frac{\partial U}{\partial n}\right]. \tag{29}$$

We have spoken of the value of s at the origin of coordinates, for mathematical convenience; but in reality the "origin" is any point P, and S is any surface enclosing that point, and s_0 is the value of s in the point P at any moment t, and U is the value of s in any point distant by r from P, evaluated at the moment $(t - r/c)$. Hence (29) may be written thus:

$$4\pi s = \int_S dS \left[\cos(n, r) \frac{\partial}{\partial r} \left(\frac{s(t - r/c)}{r} \right) - \frac{1}{r} \frac{\partial s(t - r/c)}{\partial n} \right]. \quad (30)$$

The task is achieved. It has been proved that when a function conforms to "the wave-equation" it conforms also to the first-suggested definition of a wave-motion, in that its value at any time and place is determined by its anterior values and those of its derivatives over a surface completely enclosing the place. Moreover the actual formula has been derived whereby the value at any point and moment can be computed when the values all over any surrounding surface are known at the appropriate prior moments.

Introduction of the Ideas of Frequency and Wave-length

Hitherto I have spoken chiefly of an extremely abstract "something," denoted by a symbol s, and possessed of the property that its value at any point and moment is built up out of contributions despatched at earlier moments from all of the area-elements of any continuous surface which encloses the point; these contributions being borne as it were by messengers, who travel to the point at a finite speed from the various area-elements whence they depart. Only one physical constant has been introduced, and this is the speed of these messengers. This is the constant which appears in the wave-equation (6), being there denoted by c. It is commonly called the speed of the waves; but, for various reasons which will eventually appear, it had better be called the *phase-speed*. Now there are two other constants familiar in our experience with water-waves and sound; they are *frequency* and *wave-length*. Let us try to import them into the general theory.

At any point of a water-surface over which uniform ripples are passing, the elevation is a periodic function of time; so also are the pressure and the density at any point of a gas through which uniform sound is flowing, or the displacement of either prong of a steadily-humming tuning-fork. Any periodic function of time is either a sine-function, or a composite of sine-functions. It is suitable therefore to begin by analyzing the case in which the function is a sine. Using f to denote any of the quantities above mentioned or anything behaving like them—say displacement, for example—let us write:

$$f = F \sin(nt - \delta) = F \sin \varphi. \quad (41)$$

In this very familiar form, F stands for *amplitude* and n for 2π times the *frequency*, and δ for something which is commonly called the phase; but it will be better to reserve this name for the entire argument of the sine-function:

$$\text{Phase} = \varphi = nt - \delta = \arcsin (f/F), \tag{42}$$

and I shall use it henceforth in this sense.

Now, in general, both the amplitude F and the phase φ vary from point to point—the latter because δ is often a function of position. Consequently f is a function of x, y, z and t; and it is immediately important to find out whether f satisfies the wave-equation. One who is familiar chiefly with the standard one-dimensional case of waves on a string is likely to think that this is true as a matter of course. However, on differentiating f twice with respect to x or y or z and taking due account of the dependence of both F and δ upon these variables, one sees directly that in general it is not true—not unless the functions F and δ conform to definite and sharply restrictive conditions.[3]

This appears a rather disconcerting result. However, if instead of f we envisage f/F—the value of the displacement at each point referred to its amplitude there as a unit, or the sine of the phase—it turns out that the condition under which f/F obeys the wave-equation is far less drastic. Forming the derivatives, we obtain:

$$\frac{\partial^2}{\partial t^2}\frac{f}{F} = -n^2 \sin \varphi, \tag{43}$$

$$\nabla^2 \frac{f}{F} = (\nabla^2\varphi) \cos \varphi - \left[\left(\frac{\partial \varphi}{\partial x}\right)^2 + \left(\frac{\partial \varphi}{\partial y}\right)^2 + \left(\frac{\partial \varphi}{\partial z}\right)^2 \right] \sin \varphi. \tag{44}$$

Evidently, in order that the function f/F shall conform to the wave-equation, it suffices that

$$\nabla^2\varphi = 0, \tag{45}$$

This condition is fulfilled by a variety of functions, including all which are linear in x, y, and z—the case of "plane waves," which as we shall see is one of those permissible when the wave-speed is everywhere the same, as I have been assuming.

Next we will evaluate the speed of the phase-waves, and incidentally we shall be led to a new aspect of wave-motion. When the phase

[3] The general expression for $\nabla^2 f$ is $(\nabla^2 F - F|\nabla\phi|^2) \sin \phi + (2\nabla F \cdot \nabla\phi + F\nabla^2\phi) \cos \phi$. The coefficient of $\cos \phi$ must vanish, if f is to satisfy the wave-equation with real phase-speed. By introducing the notion of "imaginary phase-speed" one may continue to regard the function f as conforming to the wave-equation, even though the coefficient of the cos-term does not vanish.

conforms to (45), it follows from (43) and (44) that

$$\nabla^2 \frac{f}{F} = \frac{1}{n^2}\left[\left(\frac{\partial\varphi}{\partial x}\right)^2 + \left(\frac{\partial\varphi}{\partial y}\right)^2 + \left(\frac{\partial\varphi}{\partial z}\right)^2\right]\frac{\partial^2}{\partial t^2}\frac{f}{F}. \tag{46}$$

The quantity in brackets, being the square of the magnitude of the gradient of φ, shall be denoted by the usual symbol $|\nabla\varphi|^2$. Then for the square of the phase-speed we obtain $n^2/|\nabla\varphi|^2$, and for the phase-speed itself:

$$c = = + n/|\nabla\varphi|, \tag{47}$$

The quantity n is 2π times the frequency. Also, it is the time-derivative of φ, as follows directly from the definition in (41); consequently:

$$c = = (\partial\varphi/\partial t)/|\nabla\varphi|. \tag{48}$$

This is an equation with a very important meaning, which will now be displayed.

Select any point in the medium and any moment of time t_0, and denote by φ_0 the value of φ prevailing then and there. Singular cases excepted, there is an entire surface containing the point in question everywhere over which φ has the same value φ_0. This surface is by definition a *wave-front*. Call it S_0. At a slightly later moment $t_0 + dt$, there will also be a surface everywhere over which the value of φ is φ_0. It will however not be the same surface S_0, but another—a wave-front S_1 so placed that from any point P_0 on S_0 the nearest point on S_1 is reached by measuring the length cdt along the line normal to S_0, in the sense in which φ is decreasing (the sense opposed to the gradient of φ).

To see this, imagine a particle which at the instant t_0 is travelling through P_0 along the direction normal to S_0 in the sense just stated, with a speed to be designated by u. At the instant $t_0 + dt$ it occupies a point where the value of φ then prevailing is given by the formula:

$$\begin{aligned}\varphi_0 + d\varphi &= \varphi_0 - |\nabla\varphi|ds + \left(\frac{\partial\varphi}{\partial t}\right)dt \\ &= \varphi_0 - u|\nabla\varphi|dt + \left(\frac{\partial\varphi}{\partial t}\right)dt,\end{aligned} \tag{49}$$

for in the time-interval dt it travels over a distance $ds = = udt$ along the normal to the surface S_0, and along this normal the slope of the function φ is equal to $|\nabla\varphi|$, and meanwhile at each point of space φ is varying directly with time at the rate $(\partial\varphi/\partial t)$. Now if the imagi-

nary particle happens to be moving with just the speed defined by the equation

$$u = (\partial\varphi/\partial t)/\nabla\varphi = c, \tag{50}$$

the coefficient of dt in equation (49) vanishes; that is, the particle as it moves along keeps up with the preassigned value of φ; but this is the same thing as saying that c is the speed of the wave-front.

The phase-function φ therefore possesses a quality which in itself suggests one aspect of a wave-motion. It is not periodic, neither does it conform to the wave-equation; but each of the surfaces over which φ has any constant value is perpetually travelling. Each of them may be changing continually in size, it may even be changing in shape; but each retains its identity, and if it is completely known for any given instant, its past and future history are determined completely; for each of the area-elements of such a surface is moving at the speed c and in the direction normal to itself, and from the position of each area-element at the moment t we can determine the position of the area-element into which it evolves at the moment $t + dt$, and repeat this process of prediction or retrospect *ad infinitum*. I will speak of this state of affairs as *propagation by wave-fronts*.

Having determined by (48) the relation between frequency and phase-speed, we now can give both the definition and the formula for the wave-length. The wave-length λ is by definition the quotient of phase-speed by frequency:

$$(n/2\pi)\lambda = c, \tag{51}$$

and in this special case, the formula for it is:

$$\lambda = 2\pi/|\nabla\varphi|. \tag{52}$$

The reason for giving this quantity a name, and such a name as "wave-length," arises from the best-known and too-exclusively-known special case, that of "plane waves" commonly so called—meaning not only that the wave-fronts are plane, but also that the amplitude is constant over each. Such waves travelling along any direction, say that of x, are described by the expression:

$$f = F \sin(nt - mx), \qquad F = \text{constant}. \tag{53}$$

The wave-fronts—that is to say, the surfaces over which $\varphi = (nt - mx)$ is constant at any moment—are planes normal to the axis of x. These planes are likewise the surfaces over which the displacement f is constant at any moment, and it is tempting to define the wave-fronts as the loci of constant displacement; but this is a coincidence which should be regarded as an accident. At a given moment, any value of sin φ

which is found anywhere repeats itself at intervals $2\pi/m$ all along the x-direction; exactly as, at a given point, any value of sin φ which is found at any moment repeats itself at intervals $2\pi/n$ all through time. Owing to the coincidence aforesaid, any value of f which is found anywhere also repeats itself at spacings $2\pi/m$ along the direction of x. This constant spacing serves as the elementary definition of wave-length; and in this special case the elementary agrees with the general definition, for

$$m = |\partial\varphi/\partial x| = |\nabla\varphi| = 2\pi/\lambda. \tag{54}$$

But there is an almost equally simple case in which the spacing between wave-fronts and the spacing between surfaces of constant f are not the same. I refer to the case of spherical waves of sound or the circular ripples on a water-surface, in which we have:

$$f = F \sin (nt - mr), \qquad F = \text{constant}/r. \tag{55}$$

In this case f/F does not conform to the wave-equation, but the function f does. Nevertheless it is the phase, and not the displacement, which advances steadily outward (or inward) in a sequence of steadily diverging (or converging) spherical wave-fronts which expand or contract with the constant phase-speed c. For any two of these spherical wave-fronts differing in radius by $2\pi/m$, the values of ϕ are the same. The surfaces of constant f are also spheres, but they expand or contract with variable speed, and for any two which differ in radius by $2\pi/m$ the values of f are different. This shows that one must not be misled by experience of plane waves into defining "wave-length" as "distance between points where at the same moment the displacement is the same" but must hold fast to the phase as the central fact of any wave-motion.

If the phase-function ϕ does not vary in space, we have the case of *stationary waves*. The coefficient of cos ϕ in the general expression for $\nabla^2 f$ (footnote on p. 339) now vanishes automatically; the coefficient of sin ϕ reduces to the term $\nabla^2 F$, and this must be equated to $-n^2F/c^2$, which if n and c are preassigned leads to an alternative form of the wave-equation

$$\nabla^2 F + \frac{n^2}{c^2} F = 0 \tag{56}$$

very common in acoustics and in wave-mechanics.

In summary:

We have considered two definitions of wave-motion: *first*, that the state of affairs at any point and moment in the medium is controlled by the state of affairs at earlier moments all over any continuous sur-

face drawn in the medium completely around the point; *second*, that the function which is propagated in waves conforms to the so-called "wave-equation."

We have found that these definitions are compatible with one another, the latter being included under the former.

We have applied them to the case of a function which at any particular point of the medium varies as a sine-function of time, thus:

$$f = F(x, y, z) \cdot \sin \varphi; \qquad \varphi = nt - \delta(x, y, z),$$

and have found:

(*a*) that provided the functions F and δ conform to certain stipulations, the function f will satisfy the wave-equation;

(*b*) that φ itself is propagated by wave-fronts; although there is nothing periodic or vibratory about φ, each surface over which φ possesses any constant value wanders onward through space, changing, it may be, in shape as well as position;

(*c*) that the speed with which the wave-fronts of the phase-function φ travel is the speed at which the contributions, out of which the value of f/F at any point and moment is built up, travel to that point from any environing surface.

A Test of Kirchhoff's Theorem

Having formed the conceptions of plane waves and sine-vibrations and frequency and wave-length, we now can practice on Kirchhoff's theorem by applying it to a problem of which the answer is predetermined, so preparing ourselves for other problems of which the answers can be discovered only by means of the theorem.

Imagine plane monochromatic waves, of frequency $\nu = 2\pi n$, wave-length $\lambda = 2\pi/m$, and phase-velocity $c = \nu\lambda = n/m$, travelling in the positive x-direction in an endless procession through an infinite medium. They are described by the function:

$$s = \cos (nt - mx), \tag{61}$$

which is a solution of the wave-equation (6).

The value s_0 of the function s at the origin at the moment t is by hypothesis

$$s_0 = \cos nt \tag{62}$$

and according to Kirchoff's theorem it is given by the following equation:

$$4\pi s_0 = \int dS \left[\cos (n, r) \frac{\partial}{\partial r} \frac{U}{r} - \frac{1}{r} \frac{\partial U}{\partial n} \right], \tag{63}$$

in which the integrand on the right is integrated over any surface completely enclosing the origin. For any such surface, then, the integral on the right of (63) must be equal to $4\pi \cos nt$.

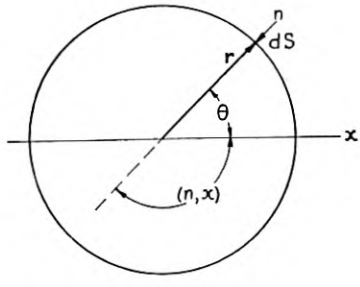


FIG. 1

This we proceed to test for the simplest of such surfaces, a sphere centred at the origin.

Denote by R the radius of the sphere; by θ, the angle between the positive direction of the x-axis and the line drawn from the origin outward through any point P of the sphere. The inward-pointing normal at P is directed oppositely to this line, and hence $\cos(n, r) = -1$ and $\cos(n, x) = -\cos\theta$. The function U and its derivatives are these:

$$\begin{aligned} U &= \cos\left[n\left(t - \frac{r}{c}\right) - mx\right] = \cos(nt - mr - mx), \\ \partial U/\partial r &= m \sin(nt - mr - mx), \\ \partial U/\partial n &= (\partial U/\partial x)\cos(n, x) = (\partial U/\partial x)\cos(\pi - \theta) \\ &= -m\cos\theta \sin(nt - mr - mx). \end{aligned} \tag{64}$$

In each of these we are to set $r = R$ and $x = R\cos\theta$ in preparation for integrating over the sphere. The element of area is

$$dS = 2\pi R^2 \sin\theta d\theta \tag{65}$$

and the limits of integration are 0 and π. Therefore the integral is this:

$$I = 2\pi \int d\theta \sin\theta[\cos\varphi - mR(1 - \cos\theta)\sin\varphi]; \qquad \varphi = nt - mR(1 + \cos\theta), \tag{66}$$

which is easy to evaluate, since on putting z for $(1 + \cos\theta)$ and $-dz$ for $\sin\theta d\theta$ it becomes

$$I = -2\pi \int_2^0 dz[\cos(nt - mRz) - mR(2 - z)\sin(nt - mRz)], \tag{67}$$

which can be integrated almost by inspection (the last term through integration-by-parts) and yields the desired value:

$$I = 4\pi \cos nt, \tag{68}$$

and Kirchhoff's theorem comes triumphantly through the test.

It happens however that these conditions, to which we have applied the theorem with so easy a success, are such that the result is of no value whatever in testing the undulatory theory of light.* As I have said already, the eye and the light-recording instruments register amplitude only, not phase. In the train of plane parallel waves described by (61), the amplitude is everywhere the same, and the instruments must report an impression uniformly intense. Nothing could be learned from them about the frequency or the wave-length of the light, and indeed they could not even show that there is anything periodic in the beam. The situation is no better in such a train of spherical diverging waves as (55) describes. Here the amplitude varies inversely with the distance from the centre of divergence, and the eye must report a gradual smooth decline of intensity as it moves away from that centre. In neither observation is there anything to reveal a periodicity or a wave-length, nor anything to forbid the supposition that a beam of light is a stream of straight-flying particles. What we require is a situation in which the use of Kirchhoff's theorem leads to a peculiar and striking variation of amplitude from place to place in the radiation-field—an amplitude-pattern or vibration-pattern, so to speak, depending in detail upon the wave-length of the waves, and so distinctive that if such a pattern of intensity were actually to be found in a field of light one could not but regard it as forceful evidence for the undulatory theory.

Situations which answer this requirement occur when a broad beam of waves is intercepted by a screen pierced with small apertures. In the space beyond the apertures there is a wave-motion of which the amplitude varies from place to place in a remarkable way, depending in detail upon the wave-length. When light falls onto a screen pierced with small holes, the intensity beyond the holes varies remarkably in space. The variations which are observed agree with those which are predicted from the wave-theory, when the proper value of wave-length is chosen; and this is the method of measuring wave-length. Also it is the method of measuring frequency; for the frequency of light cannot be measured directly; it must be computed by dividing the wave-

* In the actual theory of light there are several distinct quantities s—components of electric and magnetic field strengths—each of which separately conforms to equation (6), and which are interconnected in ways which need not yet concern us.

length into the speed of light. Now all the contemporary theories of the atom and of light are based upon values given for frequencies of radiation; and therefore all of them are founded on the wave-theory of light.

We will consequently next consider the propagation of waves beyond apertures, or what Rayleigh called the "effects dependent upon the limitation of a beam of light."

Propagation of a Wave-Motion beyond Apertures

Suppose that the entire plane $x = 0$ is occupied by a wall acting as a total stop to all the waves which come up against it, except where it is pierced by one or more openings; and for simplicity suppose that the oncoming waves compose a plane parallel monochromatic train,

$$s = \cos(nt - mx), \tag{69}$$

advancing from the side of negative x. The primary question is: what goes on in the region to the positive side of the screen?

The question shall be answered by approximations.

The *first approximation* consists in assuming that the wave-fronts come unaltered up to the screen, and each segment which coincides with an aperture goes straight through it and indefinitely onward, travelling unchanged in a straight line even though the surrounding portion of the wave-front has been blocked. If this were perfectly valid, there would be rectilinear propagation of light; the laws of geometrical optics would always be exact; and there would be no need for any but a corpuscular theory of radiation. Because this approximation is deficient, the wave-theory is required. Yet it is close enough to the truth to seem exact to all but the most careful observation, if the apertures are as wide as windows or even as keyholes.

The *second approximation* consists in assuming that the wave-fronts come unaltered up to the screen, and each segment which collides with a portion of the wall is swallowed up and blotted out of existence, but the wave-motion within each aperture is precisely the same as it would be if the entire wave-front passed intact across the plane $x = 0$. That is to say: the displacement on the front face of the screen is supposed to be given thus:

$$\begin{aligned} &s = \cos nt,\ \partial s/\partial x = m \sin nt \quad \text{wherever there is an aperture,} \\ &s = \partial s/\partial x \qquad\quad = 0 \qquad\quad \text{wherever there is obstruction.} \end{aligned} \tag{70}$$

This is the assumption on which are founded the conventional theories of the passage of light through a hole, or a slit, or a pair of slits, or a

diffraction-grating, or an echelon, or past a straight-edge or a solid disc or any small obstacle. Having made it, one proceeds to determine the value of s at any point beyond the screen by integrating Kirchoff's integrand all over the apertures—all over the vacant places of the screen, as if in those places only the wave-front were intact, and elsewhere it were abolished; as if the wave-front were cut into the pattern of the apertures as by a template, and each of the segments thenceforth propagated according to the law of wave-motion.

This method is fairly easy to apply, at least when the contours of the openings are simple geometrical figures, circles or rectangles for instance. In practice it is used almost always; and its results are ratified by experience. Yet it is not quite accurate.[4] I am not referring here to the ever-present possibility that boundary-conditions chosen for their mathematical simplicity may not properly describe the actual conditions in the physical world. I am referring to a mathematical, that is to say, a logical, difficulty which is inevitably fatal. A function which is equal to $\cos(nt - mx)$ over arbitrary patches of the plane $x = 0$, but is always equal to zero over the remainder of the plane—such a function does not conform to the wave-equation (6). If in an actual case of light passing through a hole in a screen the phenomena conform everywhere to the wave-equation, then the boundary-condition (70) cannot be valid; if on the other hand the state of affairs in the apertures is rightly described by (70), then the wave-equation cannot be valid and Kirchhoff's theorem cannot be applied.

There is no way out of this dilemma. One must either accept the foregoing assumption frankly as an approximation, or else undertake the vastly more difficult problem of solving the wave-equation itself with the boundary-condition $s = 0$ (or some other which is deemed appropriate) for all the opaque area of the screen, and with other boundary-conditions at infinity to settle the direction from which the light is supposed to come. By this method one makes no assumption about the values of s in the apertures; they are part of the solution. But the general problem is formidable, and no less eminent a man than Sommerfeld was required for the solving of even the simplest conceivable case. Later I will mention his solution of the case in which the barrier covers all that part of the plane $x = 0$ which lies to one side of a straight line, the y-axis for instance, and the remainder of the plane is void. Meanwhile, since on the whole the second approximation is a close one, I will adopt it to explore these effects of *diffraction* which first invited the wave-theory of light, as they now are inviting that of matter; which serve to determine wave-lengths of light, and of matter;

[4] In some of the texts on optics this is not made sufficiently clear.

which set the limits for the powers of telescope and microscope, and perhaps for perception altogether; and which are responsible for the haloes and the parhelia of the sky.

Imagine then that the waves which come up to the screen from behind are plane-parallel and monochromatic, and travel in the positive sense of the x-direction, the screen itself occupying the entire plane $x = 0$. In making the "second approximation" aforesaid, we are to regard this plane as a surface where s and its gradient are zero everywhere save over certain patches—to wit, the apertures—and over these are given by the expressions:

$$\begin{aligned} s &= \cos(nt - mx)_{x=0} &&= \cos nt, \\ \partial s/\partial x &= m \sin(nt - mx)_{x=0} &&= m \sin nt. \end{aligned} \tag{71}$$

We are then to determine the value s_0 of s at any field-point P anywhere before the screen—anywhere in the region $x > 0$—by forming Kirchhoff's integral over these apertures:

$$4\pi s_0 = \int dS \left[\cos(n, r) \frac{\partial}{\partial r} \frac{U}{r} - \frac{1}{r} \frac{\partial U}{\partial n} \right]. \tag{72}$$

Over the rest of the plane $x = 0$ the integrand vanishes. Since, however, Kirchhoff's theorem involves an integration over an entire closed surface surrounding P, we ought in strictness to extend the integral over some far-flung surface completing the enclosure; as for instance a hemisphere seated upon the plane $x = 0$, sufficiently great in radius to contain P and all the apertures. This is always neglected, possibly because in practice the wave-motion over such a surface would as a rule be too chaotic to produce any regular effect at P.[5]

In the integrand of (72), r stands for the distance from P to any area-element dS of an aperture; the positive r-direction is measured from P through dS in the direction from front to back; the positive n-direction is the forward-pointing normal to dS, and therefore is identical with the positive x-direction. Remembering the definitions of U and its derivatives, one easily sees that:

$$\begin{aligned} &U = \cos(nt - mr); \\ &\partial U/\partial r = \partial U/\partial x = \partial U/\partial n = m \sin(nt - mr). \end{aligned} \tag{73}$$

It will be convenient to give the symbol θ to the angle between the posi-

[5] Certainly it cannot be argued that the effect from a distant surface is necessarily too small to be noticed at P; we have just seen that in a field of plane-parallel waves it is the same for any spherical surface, no matter how great the radius.

tive x-direction and the line from dS to P, so that $\cos(n, r) = -\cos\theta$. Consequently:

$$4\pi s_0 = \int dS\left[-\frac{1}{r^2}\cos(nt - mr) - \frac{m}{r}(1 + \cos\theta)\sin(nt - mr)\right]. \quad (74)$$

One is tempted to say that the quantity under the integral sign is the contribution made by the element-of-wave-front dS to the value of s at P. This notion facilitates both thought and description, and I will adopt it, but with a warning. The danger is that one may come to think of an element-of-wave-front as an independent entity, capable of existing by itself in the medium regardless of what other elements-of-wave-front adjoin it or stand elsewhere. This is unpermissible, for the same reason which makes the method that I am now expounding an approximate and not a rigorous one. Were one of these elements of wave-front alone in the medium, the function s would not conform to the wave-equation. Therefore if we call the expression

$$ds_0 = \frac{1}{4\pi}dS\left[-\frac{1}{r^2}\cos(nt - mr) - \frac{m}{r}(1 + \cos\theta)\sin(nt - mr)\right] \quad (75)$$

the *contribution* of the element-of-wave-front dS, we must always remember that it cannot be isolated, but—like the donations to certain endowments—is given only under the condition that other elements also contribute.[6]

The contribution of dS, then, is made up of two terms, one varying inversely as r^2 and the other inversely as r/m. At great distances the latter must increasingly outweigh the former; and "great distances" in this context signify those which are much greater than $1/m$—that is to say, very many times as great as the wave-length of the light. Now as the wave-lengths of most kinds of light are less than .001 mm., a field-point where observations can actually be made must necessarily be distant by many wave-lengths from the screen. Hence it is customary to ignore the first term in the expression (75) and in its integral, and write for the contribution of dS:

$$ds_0 = -\frac{1}{4\pi}dS\frac{m}{r}(1 + \cos\theta)\sin(nt - mr). \quad (76)$$

This expression is the approximate description of what in an earlier

[6] The reader may notice that whereas in dealing with a closed surface surrounding the field-point the n-direction was defined as that of the "inward-pointing normal," there is no way of discriminating between the two senses of the normal to an isolated area-element. This causes an ambiguity in the sign of the contribution; for reversing the sense of the normal reverses the signs both of $\partial U/\partial n$ and of $\cos(n, r)$. The ambiguity is always, I think, physically trivial.

passage I called the "influence" or the "wavelet" which spreads out from the element-of-wave-front in all directions.

Examining it factor by factor, one sees:

(*a*) that the amplitude of the wavelet varies inversely as the distance r from the starting-point, which seems natural;

(*b*) that the wavelet is not isotropic, its amplitude diminishing according to the law $(1 + \cos \theta)$ from a maximum value in the forward to zero in the rearward direction. This is commonly stated as the reason why waves can be propagated in one direction only, not necessarily both forward and backward at the same time;

(*c*) that for waves of the same amplitude and different wave-lengths the amplitudes of the wavelets stand in the inverse ratio of the wave-lengths—the shorter the waves, the more powerfully they are diffracted;

(*d*) that the wavelet from any point is constantly one quarter of a cycle in advance of the primary wave, varying as $-\sin nt$ whereas the wave varies as $\cos nt$.

The advance-in-phase and the factor m in the amplitude enter, it is clear, because the "wavelet" represents the second term in (75)—the term which involves the slope $\partial s/\partial x$ of the wave-function, not the wave-function itself. One might say that the cyclic variation of $\partial s/\partial x$ stirs up a relatively far-reaching commotion in the medium, while the disturbance which the cyclic variation of s excites is rapidly attenuated and mostly negligible. Formerly the factor m and the advance-in-phase seemed unnatural and very strange; for they antedated the theorem of Kirchhoff by sixty years, having been forced upon Fresnel before 1820—and this invites an allusion to history.

Though it is in connection with Huyghens' principle that one commonly hears of wavelets, that principle itself amounts to a denial of nearly every quality which we associate with the ideas of wavelet or wave. Not only are the "wavelets" of Huyghens' construction quite devoid of anything undulatory or periodic; the construction itself is based on the assumption that there is only one point on each where the amplitude is appreciable—the point on the prolongation of the normal from the primary wave-front (corresponding in my notation to $\theta = 0$). But to say that a disturbance is transmitted by wavelets such as these is to say that it is transmitted in concentrated form along lines or rays—which is the same thing as saying that it travels like corpuscles. Huyghens' principle in fact leads straight to the doctrine of the rectilinear propagation of light, and fails either to predict or to explain the phenomena which require a wave-theory.* The accredited

* I am not prepared to say that this is true of the applications to crystal optics.

founder of the wave-theory of light invented in reality a novel language for expressing the corpuscular theory!

Fresnel however invested these wavelets of Huyghens with some of the properties which entitle them to the name. He supposed that the amplitude was distributed widely over each, not confined to the point $\theta = 0$, though greatest at that point; he thought that it diminished slowly with increase of θ, though he did not suggest the precise factor $(1 + \cos \theta)$ nor any other; and he thought that it varied inversely as distance. Further, he endowed it with a periodicity. Thus far, he was right. But naturally he supposed that the cause of the wavelet was the cyclic variation of the wave-function s, and therefore he presumed that it started out in consonance of phase with the primary wave; and he did not insert the factor m. However when he came to test his ideas in somewhat the same way as Kirchhoff's theorem has been tested in these pages—by applying them to a case where the required result was known *a priori*—he was unable to derive the proper answer, except by introducing the factor m and the advance-in-phase; and thenceforth they have figured in the theory of diffraction, indispensable and until the day of Kirchhoff inexplicable.

To return to the problem of determining the wave-motion beyond the apertures: under the approximations stated, it is mathematically quite definite. The solution is the value of the integral:

$$s_0 = -\frac{1}{4\pi}\int dS\left[\frac{m}{r}(1 + \cos\theta)\sin(nt - mr)\right], \qquad (77)$$

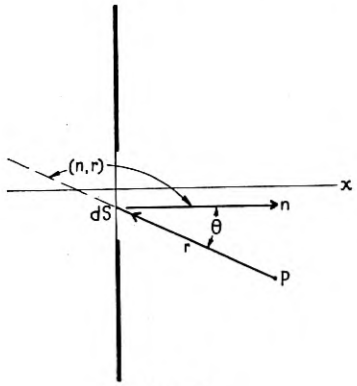


FIG. 2

extended over the apertures; r standing for the length of the line joining the field-point P with the element-of-wave-front dS, and θ for the angle between this line and the perpendicular dropped from P to the plane of the screen.

A simple, instructive, and historically famous example is that of the circular hole.

Diffraction from a Circular Aperture

If the propagation of waves were rectilinear, their amplitude would be constant along every line passing normally across an aperture. Such however is as far as possible from being the truth, as we can easily learn by evaluating the wave-motion along the "axis" of a circular hole—that is, the line passing through the centre of the hole perpendicular to the plane of the screen. Locate the centre of the circle at the origin, so that its axis is the axis of x. Denote by R the radius of the circle, by x_0 the coordinate of the field-point P located anywhere upon the axis. All points on any circle centred at the origin being equidistant from P, we may divide the area of the hole by concentric circles into annular elements-of-area. Denote the radius of such a one by p, its breadth by dp; then for it:

$$dS = 2\pi p dp; \qquad r^2 = x_0^2 + p^2; \qquad \cos\theta = x_0/r, \tag{78}$$

and the limits of integration are $p = 0$ and $p = R$.

The problem is now stated in full; but it is very much simplified if the distance x_0 from screen to field-point is very many times as great as the width R of the aperture; and this in practice is commonly the case. Then to first approximation,

$$r = x_0, \qquad \cos\theta = 1, \tag{79}$$

and these values are close enough to the correct ones to suffice for the multipliers of the sine-function in (77); but in the argument of the sine, r is multiplied by m, and a variation of only half a wave-length in r entails a complete reversal of the function; hence in the argument we must proceed to second approximation, and write

$$mr = mx_0 + mp^2/2x_0. \tag{80}$$

Making these substitutions in (77), we have finally:

$$s_0 = -(m/x_0)\int_0^R p \sin(nt - mx - mp^2/2x_0)dp$$

$$= \cos(nt - mx_0) - \cos(nt - mx_0 - mR^2/2x_0) \tag{81 a}$$

$$= \sqrt{2(1 + \cos mR^2/2x_0)}\cos(nt - mx_0 - a). \tag{81 b}$$

Interpreted, these equations tell the startling fact that along the axis of the hole the amplitude, far from being constant, varies in a

gradual and cyclic way between zero as one extreme and double the amplitude of the unintercepted wave-train as the other. As the field-point is displaced along the axis towards or away from the aperture, as the aperture itself is expanded or contracted, doubled agitation succeeds upon quiescence and quiescence upon agitation; and the opening, far from serving as a window to let a segment of the oncoming wave-train pass unaltered by, acts as an agency for producing a curious pattern of varying amplitudes over the region before it.

Now these are precisely the conditions under which, as I remarked before, one can arrange a test of the wave-theory of sound or light; for here we have the amplitude varying from point to point, in a pattern depending in detail upon the wave-length. Experience of light reveals just such a pattern; when parallel light is shed normally upon a screen pierced with a small and accurately rounded hole, the illumination in the axis of the hole passes alternately through maxima and minima as the observer recedes along it. Fresnel was led in a curious way to discover the minima. The French Academy having offered a competitive prize for a study of diffraction—an action instigated, it appears, by adherents of the corpuscular theory of light, who expected that a thorough knowledge of the phenomena of diffraction would demolish the support which they were vaguely supposed to provide for the wave-theory—Fresnel conducted a research and submitted a memoir which ranks among the classics of physical science. It went for judgment to an illustrious committee of five,[7] one of whom, the very eminent mathematician and physicist Poisson—who had been an upholder of the corpuscular theory—promptly deduced the law of the maxima and minima along the axis from Fresnel's conception of the wavelets. He imparted this prediction to the author of the memoir; and in a note appended to the published version, Fresnel has left it on record that he looked for a minimum and found it "like an inkspot" in the centre of the field before the hole.

Equation (81) shows further that the amplitude at any point upon the axis must vary to and fro between the same two extremes—zero, and double the amplitude of the unhindered waves—as the hole expands or shrinks. Wood has described how this may be observed with an iris diaphragm. For an observer stationed at a fixed point upon the axis at a distance x_0 from the hole, the amplitude falls to zero whenever the radius of the circle has one of the values determined by the condition

$$mR^2/2x_0 = \text{even integer multiple of } \pi, \tag{82}$$

[7] Arago, Biot, Gay-Lussac, Laplace and Poisson. It would be hard to assemble a more distinguished group at any time or place.

that is to say, whenever

$$x_0 = mR^2/2k\pi, \qquad k = 0, 2, 4, 6, \ldots, \tag{83}$$

and attains its maximum value, double the amplitude of the uninterrupted waves, whenever

$$x_0 = mR^2/2k\pi, \qquad k = 1, 3, 5, 7, \ldots. \tag{84}$$

Imagine circles drawn upon the plane of the screen, with their common centre at the origin and their radii R_1, R_2, R_3, . . . prescribed by the equations,

$$mR_k^2/2\pi x_0 = k, \qquad k = 0, 1, 2, 3, 4, \ldots. \tag{85}$$

They divide up the plane of the screen into a tiny central circular area and a series of surrounding rings. These are the "Fresnel zones" relative to the point x_0 where the observer is placed. If the circular hole comprises an odd number of the zones, the wave-motion at x_0 attains its maximum; if an even number, the wave-motion vanishes—there is silence or darkness. It seems as if the first, third, fifth and other odd-numbered zones brought light, and the second, fourth and other even-numbered zones destroyed it.

It is equally easy to find the wave-motion along the axis of an annular opening—that is to say, a circular hole partly filled by a concentric circular stop. Denote by R_0 the radius of the stop and by R the radius of the hole; then the limits of integration in (81) are superseded by $p = R_0$ and $p = R$, and the amplitude along the axis varies thus:

$$A = \sqrt{2[1 - \cos m(R^2 - R_0^2)2x_0]}. \tag{86}$$

This contains the surprising conclusion that the maxima of amplitude along the axis are as great as they would be if the stop were removed, though they may be differently placed. An observer properly stationed should see the light brighten when the obstacle is inserted; it may even be brighter than when the obstacle within the hole and the wall surrounding it are totally removed, leaving no hindrance to the onward march of the waves.

The conclusion still holds good when the boundaries of the circular hole retire to infinity, leaving nothing but an opaque disc in an otherwise uninterrupted stream of plane parallel waves; although the approximations made in the foregoing pages are then no longer valid, and equation (86) is not to be employed. Experience however shows that when a small and accurately rounded circular disc is immersed in a beam of parallel light there is a bright spot—more precisely speaking,

a bright core—along the axis of the geometrical shadow. Poisson forecast this also when Fresnel's memoir came before him, and seems to have thought that it would make an *experimentum crucis*, for another member of the committee—Arago—has recorded that he tested the prediction when Poisson made it. He found the bright spot in the centre of the shadow of a circular disc. It is said that Delisle had found and recorded it already, but the record had slipped into oblivion.[8]

We take up now the problem of determining the wave-motion away from the axis—otherwise expressed, that of determining the distribution-of-amplitude over any plane parallel to the plane by the screen.

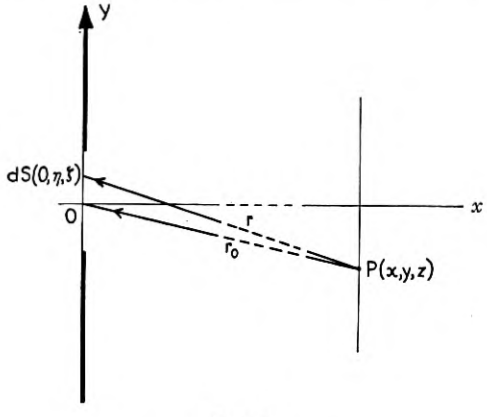


FIG. 3

Denote by (x, y, z) the coordinates of any field-point and by $(0, \eta, \zeta)$ those of any area-element dS of the aperture; by r, as heretofore, the distance from P to dS, and by r_0 the distance from P to the origin. Then

$$r^2 = x^2 + (y - \eta)^2 + (z - \zeta)^2 = r_0^2 - 2y\eta - 2z\zeta + \eta^2 + \xi^2. \quad (87)$$

As heretofore r and r_0 shall be supposed to be very many times as great as the dimensions of the apertures, and therefore as the greatest values attained by η and ζ; therefore, to first approximation,

$$r = r_0, \qquad \cos\theta = x/r_0, \quad (88)$$

[8] It is interesting to notice why an accurately circular disc is required to show the bright spot in its best development. Take the case of the aperture, since we already have its suitable equation (81). A nearly but not quite circular hole may be regarded as made up of sectors, each with a different radius. For each of these the upper limit of the integral in (81) would be different, and therefore the condition (84) for doubled amplitude could not be realized for all at once. There would be a wave-motion along the axis, but not the regular alternation of maxima and minima nor the sharply outstanding brightness at the maxima.

and to second approximation,

$$r = r_0 - (y\eta + z\zeta)/r_0. \tag{89}$$

Into equation (77) we insert the first-approximation values of r and $\cos\theta$ in the multipliers of the sine-function, but the second-approximation value of r into the argument of the sine. Therefore we have, for the value of s at the very distant point (x, y, z), the expression:

$$s = -\frac{1}{4\pi}\frac{m}{r_0}\left(1 + \frac{x}{r_0}\right)\int\int d\eta d\zeta \sin (nt - mr_0 - m\overline{y\eta + z\zeta}/r_0). \tag{91}$$

It is expedient to introduce the three direction cosines of the line extending from the origin to the field-point, the cosines of the angles between it and the coordinate axes:

$$\alpha = \cos (x, r_0) = = x/r_0; \qquad \beta = y/r_0; \qquad \gamma = z/r_0. \tag{92}$$

Then, with a slight additional transformation, we convert equation (89) into:

$$\begin{aligned} s = \text{const.}\ (1 + \alpha)\Big[&\sin (nt - mr_0)\int\int d\eta d\zeta \cos m(\beta\eta + \gamma\zeta) \\ &- \cos (nt - mr_0)\int\int d\eta d\zeta \sin m(\beta\eta + \gamma\zeta)\Big] \\ = \text{const.}\ (1 + \alpha)&[C \sin (nt - mr_0) - S \cos (nt - mr_0)], \end{aligned} \tag{93}$$

the symbols C and S being traditional for these integrals.

The coordinates of the field-point have disappeared, leaving only the cosines which define its direction as seen from the origin. This means that we have here the formula for the wave-motion over any plane parallel to the screen and infinitely far away, in terms of the directions in which its various points are seen. The words "infinitely far away" sound formidable; but it is not necessary to depart for infinity, in order to find a plane where (93) describes the state of affairs. There is an artifice for bringing the infinitely distant plane up to a convenient nearness; an artifice known as a *lens*. When a converging lens is set up before the apertures, the wave-motion predicted by the formula (93) for all points infinitely far away upon the line with direction cosines (α, β, γ)—this wave-motion occurs at the point where the line intersects the focal plane of the lens. Therefore we may regard equation (93) as the description, according to the wave-theory of light, of the distribution-of-amplitude in the focal plane of the lens. (To convert the cosines into coordinates in that plane, it is sufficient to multiply each by the focal length of the lens.)

Returning now to (93), it is evident that the problem is solved when the integrals are evaluated; in particular the amplitude is given by the formula,

$$A = \text{const.}\ (1 + \alpha)\ \sqrt{C^2 + S^2}. \tag{94}$$

Whatever the shape of the aperture or apertures, the values of the integrals can be determined as closely as may be desired; and in two instances which happily are the most frequent and useful—those of the circular and the rectangular openings—the integrations lead directly to familiar functions.

Diffraction Patterns in the Focal Plane of a Lens

If the origin is located at the centre of the circle, the integral S vanishes—for the value of the sine-function contributed by each area-element is annulled by the value contributed by the element symmetrically placed to the other side of the centre—and the integral C for the same reason becomes this:

$$C = \int\int \cos(m\beta\eta) \cos(m\gamma\zeta) d\eta d\zeta. \tag{95}$$

By putting $\gamma = 0$ and then integrating, we shall obtain the distribution of amplitude along the line passing through the centre of the diffraction-pattern and parallel to the axis of y; but this, by reason of the circular symmetry of the entire system, is the same as the distribution of amplitude along any radius passing through the centre of the diffraction-pattern, and therefore is all we need. In the expression so obtained, replace the Cartesian coordinates heretofore used in the plane of the screen by polar coordinates ρ and φ; then we have

$$C = \int\int \rho \cos(m\beta\rho \cos\varphi) d\rho d\varphi, \tag{96}$$

the limits of integration being 0 and R (the radius of the aperture) for ρ, and 0 and 2π for φ.

The integral C is proportional to the Bessel function of order unity of the variable $mR\beta$:

$$C = 2(\pi R/m\beta) J_1(mR\beta). \tag{97}$$

This is a function which like the sine vanishes at intervals, though not at equal intervals. The centre of the diffraction-pattern is therefore encircled by concentric rings over each of which the wave-motion vanishes; between each pair of these there is a zone where the amplitude differs from zero and varies, attaining a maximum somewhere near the middle of the zone. In the focal plane of the lens there are ring-shaped zones of light, surrounded and divided by dark circles; these are the "fringes."

This system of annular fringes is the *image* produced by a lens on which plane-parallel light falls normally through a circular aperture; also when there is no screen before the lens, for being circular it serves as its own aperture. Now plane-parallel light is such as originates in an infinitely distant luminous point, or—what comes to the same thing —any luminous object so distant that neither the curvatures of the wave-fronts proceeding from its various parts nor the angles between the directions in which these lie are appreciably large; a star, for instance. The image of a star in the focal plane of a telescope objective is therefore not a point, however far away the star may be; it is a system of rings. So it is in the eye, the pupil serving as the aperture; but the inner rings are in both cases so narrow and the outer rings so faint that they appear condensed into a point. Magnification of the fringes in the telescope by the eyepiece brings them into view, and so they set a limit to the value of magnification; for it is of no avail to be able to examine an image more minutely if all that can be examined is the consequence of the disturbance produced in the incoming waves by the finiteness of the lens.

The limitation which the law of propagation of light thus sets upon the formation of images is very important. The simplest possible illustration is furnished by a double star. Let the telescope be directed upon such a pair of stars so that the light from one component falls normally upon the lens, the light from the other component at any angle of which I denote the complement by φ; thus φ stands for the angular distance between the two stars in the sky. Now I have not hitherto treated the case of light falling otherwise than normally upon the screen containing the apertures, which in this case is nothing but the plane of the objective. The extension however is immediate. Orienting the y-axis in the plane of the screen so that the direction of propagation of the waves coming from the stars lies in the xy-plane, we have for the wave-function in the region extending up to the aperture from behind:

$$s = \cos(nt - mx\cos\varphi - my\sin\varphi), \tag{98}$$

and therefore in the plane of the aperture ($x = 0$) we have, instead of the values given in (71), these:

$$\begin{aligned} s &= \cos(nt - my\sin\varphi), \\ \partial s/\partial x &= m\cos\varphi\sin(nt - my\sin\varphi), \end{aligned} \tag{99}$$

and for the value of s at any point in front of the aperture we have, instead of (77), this value:

$$s_0 = -\frac{1}{4\pi}\int dS\left[\frac{m}{r}(\cos\theta + \cos\varphi)\sin(nt - my\sin\varphi - mr)\right],$$

and there are corresponding changes in the values of the integrals C and S which determine the amplitude. To first approximation—that is to say, when φ is not too great—the result is, that the diffraction pattern of one star is like that of the other, but shifted sidewise. The angular displacement between the centres of the two fringe-systems is the same as the angular displacement between the two stars. The question now arises: how far apart must the two fringe-centres be, that the two families of rays may be securely told apart?

Such a question of course cannot be definitely answered; the answer would depend upon the acumen and the experience of the observer. The conventional response is, that the two systems of rings are surely distinguishable if the centre of one lies upon the first dark circle of the other. Now the angular radius of the first dark ring, i.e., the value of β for which the Bessel function of (97) first vanishes, is 1.22 times the ratio of the wave-length of the light to the diameter of the aperture; for green light in the largest available refracting telescope this amounts to about an eighth of a second of arc. This then is nearly the least angular separation between two stars which are distinguishable; a pair or a group much closer together would appear as one, not through any avoidable defect of the telescope nor through any insufficiency of the eyepiece but through the laws of propagation of light themselves, working to prevent the formation of an image indefinitely sharp.

For a rectangular aperture the integrals C and S are extremely easy to evaluate. The diffraction-pattern is a criss-cross of dark lines, intersecting at right angles and bounding rectangular areas of light, similar in shape to the aperture but oriented at right angles to it. If the rectangle is prolonged indefinitely and so becomes an infinitely long slit, the diffraction-pattern becomes a sequence of parallel bands separated by dark lines normal to the length of the slit. If then a multitude of identical slits are cut into the screen at equal intervals side by side, a new periodicity is superposed upon the periodicity of the waves, and out of the interaction of these two there come diffraction-patterns much more sharp and striking than any which a single aperture, however shaped, is able to produce. These will be considered in the following chapter.

Recent Developments in the Process of Manufacturing Lead-Covered Telephone Cable [1]

By C. D. HART

THE manufacture of telephone cable consists essentially of insulating copper wire with paper, twisting two insulated wires together to form a pair, again twisting to form a quad if quadded cable is to be made, stranding these pairs or quads into a compact core, removing moisture, covering the core with a continuous sheath of lead or lead alloy, testing the completed cable and packing it for shipment.

In order to bring out clearly some of the recent developments in manufacturing processes it is necessary to review the beginning of the art.

The idea of using cables for telephonic communication goes back to about 1878. In a talk given in London by Dr. Alexander Graham Bell he stated "It is conceivable that cables of telephone wires could be laid underground, or suspended overhead, communicating by branch wires with private dwellings, country houses, shops, manufactories, etc., uniting them all through the main cable with a central office where the wires could be connected as desired, establishing direct communication between any two places in the city."

About two years later, or in 1880, the idea became a fact and wires enclosed in sheath were used across the Brooklyn Bridge.

The insulation used on these early cables was gutta-percha or rubber but these materials were not very satisfactory for land telephone cables. A little later sisal and cotton were used and the cable core was impregnated to prevent the entrance of moisture and then drawn into successive lengths of lead pipe previously extruded and laid out in straight pieces, the different lengths being then joined together by means of plumber's joints. Impregnation was resorted to because it was difficult to obtain a lead sheath which was entirely free from defects.

By about 1890 paper ribbon had been introduced as a substitute for cotton and similar insulations, effecting, of course, a great saving in space and therefore in sheathing material and cost.

Fig. 1 shows a group of insulating machines used about 1892. With these machines paper ribbon was wound from a spool mounted eccentrically with the wire and the insulating speed was necessarily very slow.

[1] Presented at the Regional Meeting of District No. 5 of the A. I. E. E., Chicago, Ill., Nov. 28–30, 1927.

Fig. 2 shows the twisting machines used at that time for twisting pairs. These machines were crude and operated at a low speed.

Fig. 1—Old insulating department about 1892

Fig. 2—Old twisting department about 1890

Fig. 3 is of an old stranding machine consisting of one drum only as the cable cores were built up one layer at a time and the core was run

through the stranding machine as many times as there were layers in the finished core.

The old process of pulling cable core into lead pipe is illustrated in Fig. 4. This picture was posed a few years ago, and the man standing in the foreground was one of a gang who formerly did this work.

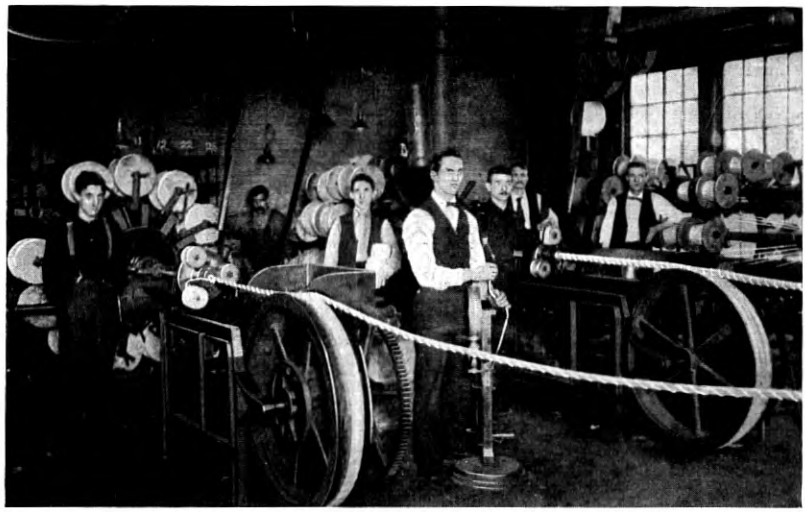


Fig. 3—Old stranding department about 1890

A forward step in design of insulating equipment was made with the use of pads concentric with the wire which permitted very much higher insulating speeds and very much reduced paper breakage. The twisting machines were also modified to reduce uneven twisting and permit greater speed.

Another step was the development of multiple drum stranders permitting a number of layers or complete small cables to be made in one operation. Also, extrusion presses were improved so that a continuous sheath of lead alloy could be extruded directly on to the cable core, eliminating the pulling-in operation.

During the period from about 1900 to 1920 many changes were made to increase output and improve the quality. Improvements in machines made possible the use of thinner and narrower insulating papers so that a greater number of pairs of wires could be placed within the same cross-sectional area, tending greatly to decrease the cost per circuit. Cables made about 1888 contained 50 pairs of 18-gauge conductor. By about 1902 improvements had been made which permitted 606 pairs of 22-gauge wire to be put into a sheath of $2\frac{3}{8}$ in.

inside diameter which is the maximum size of sheath which has been found generally economical in telephone plants in this country. By 1912 further improvements in manufacturing equipment made it possible to use insulating paper of even smaller dimensions and to get 909 pairs of wire into the same diameter of sheath.

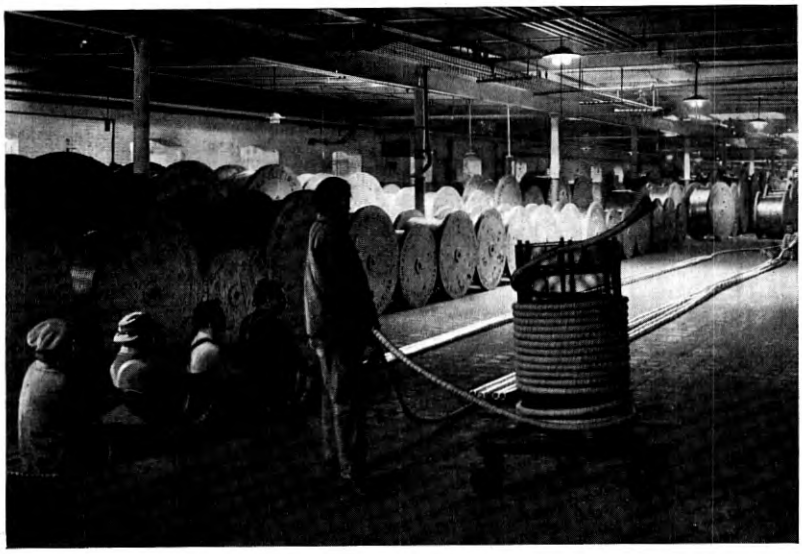


Fig. 4—Pulling core into lead pipe, method used prior to 1894

On account of increased congestion in the densely populated sections of the larger cities, there was continued demand for more pairs of wire per cable, and in 1914 the first 1,212-pair 24 A.W. gauge cables were produced. This 24-gauge wire was insulated with paper $^{5}/_{16}$ in. wide and $2\frac{1}{2}$ mils thick. The mutual capacitance between the two wires of a pair in this cable averages about .079 microfarad per mile, which allows a normal margin below the guaranteed value shown in Table 1.

TABLE 1

A.W.G.	Standard Sizes—Pairs	Average A. C. Capacitance Guarantee m.f. per Mile	Principal Uses
13	11 to 76	.071	Toll entrance and long trunks.
16	11 " 152	.071	
19	6 " 455	.090	Trunk and long subscriber lines.
22	11 " 909	.089	Subscriber lines.
24	11 " 1212	.085	Short subscriber lines.

The insulation withstands a potential test of 500 volts (maximum instantaneous value). The increasing number of pairs per cable and

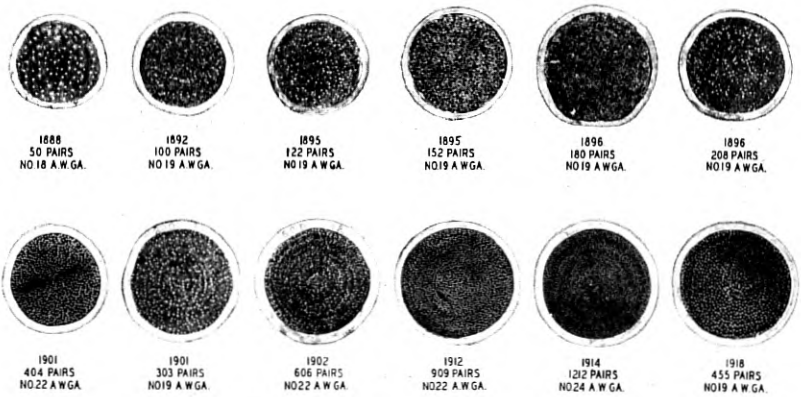


Fig. 5—Principal stages in the development of paper-insulated cable

the corresponding decreasing cost per mile of circuit resulting from the changes described is shown in Figs. 5 and 6.

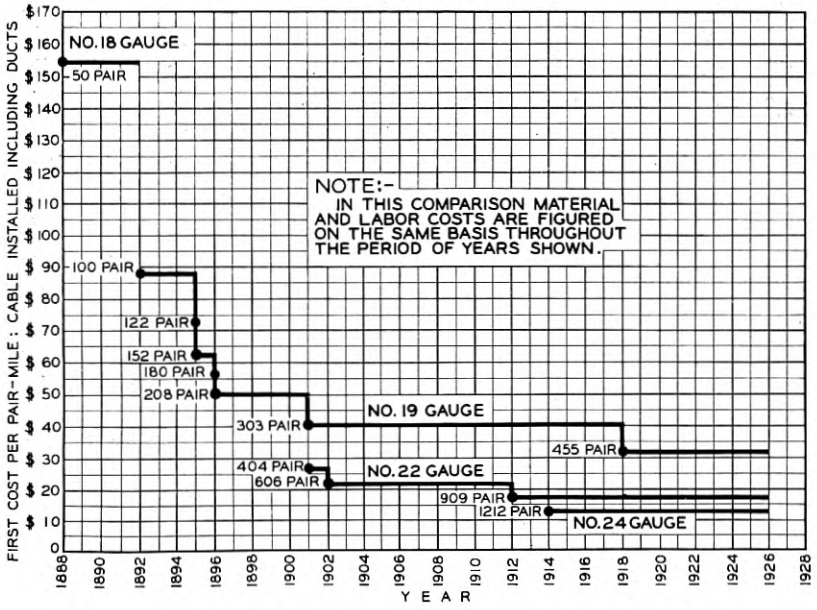


Fig. 6—Cost of a mile of circuit in full-size cable

With the growth of large office buildings and further increases in the demand for telephones in the great cities, even 1,212 pairs of wire per

cable were in some cases found to be inadequate, and in answer to the demand a cable has been developed containing 1,818 pairs of 26 A.W.G. wires within a sheath having an inside diameter of $2\frac{3}{8}$ in.

These wires are insulated with paper $\frac{7}{32}$ in. wide and $1\frac{3}{4}$ mils thick by the use of specially designed insulating heads and, instead of being stranded in reverse layers as is the case with older types of cables, they are first stranded in groups of 101 pairs, 18 of these groups being then cabled together to form a compact core.

This method of cabling, called the "unit" type to distinguish it from the layer type, has several advantages, particularly in splicing in the field. Development work on this 1,818-pair cable is not yet complete but there is no reason to doubt that, if there is a demand for a 2,400-pair cable, the demand will be met.

For convenient reference Table 1 has been shown giving the specified limiting characteristics of some of the standard types of non-quadded cables. From the table it will be seen that the larger gauge cables are used mostly for trunk work and the smaller gauges for connections to subscribers. While the electrical characteristics of these non-quadded cables are of prime importance, they do not demand quite the extreme refinement in manufacturing processes required for quadded cables.

The discussion so far has been confined mainly to cable intended for local service, that is, cable providing conductors to connect subscribers directly with the central office and different offices with one another. Gradually, the network of long lines connecting different exchange areas or cities grew and while the early lines were mostly open-wire, it was necessary to provide cable in and near the larger cities to bring these lines into the central offices. Most of the long lines were operated on the phantom principle where four wires are combined to provide two ordinary pair circuits and a third or phantom circuit which uses the four wires simultaneously. It was, therefore, necessary to provide cable for these toll entrances which could also be operated on the same phantom principle. More recently many long toll lines have been placed for their entire length in cables of this type.

One of the greatest difficulties in providing this type of cable was that of building it with sufficiently good electrical balance to avoid serious interference or "crosstalk" between the various circuits in the same four-wire group or "quad," such crosstalk being especially liable to occur because practically all of these lines are loaded. For a given degree of imperfection in capacitance balance, crosstalk is much more serious if the line is loaded than otherwise. A very considerable amount of work was necessary to determine the principles of design and manufacture which have the most influence in bringing about the best balance reasonably attainable.

The specified limiting degree of unbalance of the capacitance in quadded cable is indicated in Table 2, and Fig. 7 is a diagram showing the capacitances involved and a brief explanation of them.

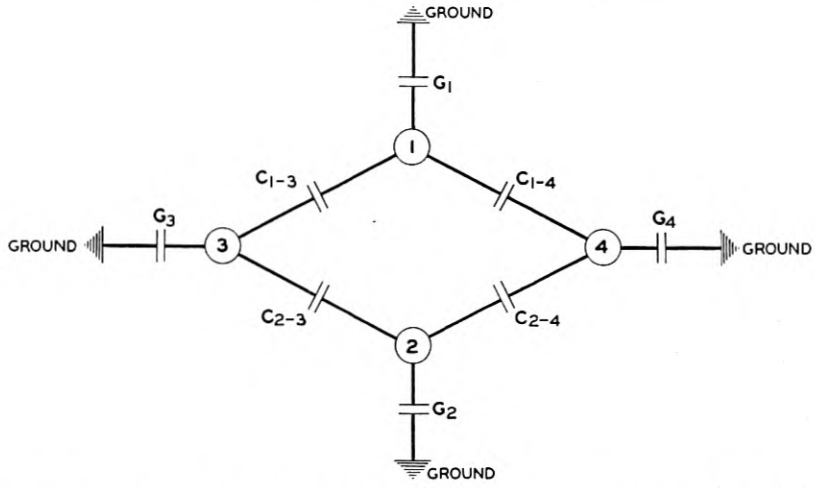


Fig. 7—Diagram showing the capacities involved in capacity unbalances between circuits

TABLE 2

Capacitance in m.f. per mile		Capacitance Unbalance in m.m.f. per 500 ft. length					
Pair	Quad	Side to Side		Phantom to Side		Phantom to Phantom	
Av.	Av.	Av.	Max.	Av.	Max.	Av.	Max.
.068	.112	30	100	120	200	60	600

[1] CLASS I UNBALANCES—PHANTOM TO SIDE

1, 2, 3 and 4 represent the four wires of a quad, of which 1 and 2 form one pair and 3 and 4 form the other pair.

Unbalance between Phantom and Side $1\text{–}2 = 2[C_{1-3}+C_{1-4}-(C_{2-3}+C_{2-4})]+G_1-G_2$
Unbalance between Phantom and Side $3\text{–}4 = 2[C_{1-3}+C_{2-3}-(C_{1-4}+C_{2-4})]+G_3-G_4$

CLASS II UNBALANCES—SIDE TO SIDE

1, 2, 3 and 4 represent the same as in Class I Unbalances.

Unbalance between Side 1–2 and Side $3\text{–}4 = C_{1-4}+C_{2-3}-(C_{1-3}+C_{2-4})$

[1] Capacitance Unbalances involve differences of Direct Capacitances. See G. A. Campbell, *Bell System Technical Journal*, July 1922.

CLASS III UNBALANCES—BETWEEN CIRCUITS IN DIFFERENT QUADS

Unbalances between two phantoms, or between pairs not in same quad, or between a phantom and a pair not in same quad, in each case $= C_{1-4} + C_{2-3} - (C_{1-3} + C_{2-4})$ in which, for

(*a*) *Phantom to Phantom*, 1 represents the two wires connected in parallel to form one pair of a quad, 2 represents the two wires of the quad, and 3 and 4 represent similarly the pairs of another quad.
(*b*) *Pair to Pair*, 1 and 2 represent the two wires of a pair and 3 and 4 the two wires of another pair not in the same quad.
(*c*) *Phantom to Pair*, 1 and 2 represent a phantom as in (*a*) and 3 and 4 a pair as in (*b*).

The type of quad now most commonly used in toll cables in this country is known as the multiple twin type and consists when completed of two twisted pairs which are again twisted around each other. Differently colored wrappings of cotton around the several pairs hold the two wires of the pair together and afford means of identifying various types of quad and pair as used, for example, in the segregation of the circuits operating in different directions in the so-called four-wire circuits.

A type of quad construction different from that described above and commonly known as the "spiral four" type of quad has been used more extensively abroad than here. In this construction four wires are twisted together in such a way that at every position each wire occupies approximately a corner of a square and the two diagonally opposite conductors are used to form a pair.

This construction has the merit of very low mutual capacitance of the pairs, but the disadvantage of very high mutual capacitance of the phantom. It has also been found more difficult with this construction to obtain sufficiently good balance to give satisfactory loaded phantom circuits. This type of quad has, therefore, in some cases been used without utilizing the phantom circuits. The loss of these phantom circuits is less than it might seem at first sight because, on account of the inherently lower pair capacitance for a given space per pair, more wires can be placed in the same space for a given capacitance than with other types of construction.

Another characteristic which under certain conditions is important is the alternating current conductance or leakance. The leakance which is measured in micromhos is that property which determines, under given conditions of potential and frequency, the losses in the insulation. These losses become of greater importance when the cable is loaded than when non-loaded and also of relatively greater importance when the conductors are large because then the dielectric losses become relatively greater in comparison with the lower losses in the decreased resistance of the conductor. For this reason many of the

large gauge loaded toll cables are treated with a special drying process to diminish the leakance.

Under-water Cables

Either quadded or non-quadded cable may be used on occasion for crossing rivers, bays, etc., and in these cases the lead-covered cable is protected by being first served with two or three layers of jute roving impregnated with tar, then wound with galvanized steel armor wire, and again served with jute yarn, impregnated with an asphalt compound, although in many cases at present this outer serving of yarn is omitted. In case of injury causing an opening in the sheath of such a cable, water may enter the interior and interrupt the service. It is also liable to penetrate for a considerable distance and thus ruin a substantial length of cable which it then becomes necessary to replace. To diminish the amount of cable damaged in this way, this type of cable is sometimes made with a very large amount of paper insulation crowded into a small space to make the cable within the lead pipe very dense. The swelling of this paper as it becomes wet tends to retard the penetration of water and to diminish the amount of cable damaged.

This dense core construction has, however, the objection that it tends to produce circuits of lower transmission efficiency on account of the higher capacitance and leakance obtained. For this reason cables for this purpose in many cases are made with less dense core construction similar to that used in land cables but with the core treated so as to provide water barriers at frequent intervals to prevent or greatly diminish the passage of water through the barrier, commonly known as a "plug," so that the damage resulting from an injury to the sheath is substantially confined to the portion between two consecutive plugs.

The Cable Sheath

One of the outstanding developments in cable manufacture which occurred about 1911 was the substitution of 1 per cent antimony in lead cable sheath for 3 per cent tin. The use of tin alloyed with lead for cable sheath had been instituted many years before, as it had been found that such sheath was more durable than sheath composed of lead alone and had better mechanical characteristics.

Exhaustive tests showed that lead-antimony alloy sheath is equal in quality to lead-tin alloy and, although its use required the development of improved methods of mixing and extrusion, it has resulted in large cost savings.

Another decided improvement introduced later was the substitution of vacuum drying ovens for the old gas or steam-heated air ovens.

It was found that the drying time using vacuum ovens was reduced to about one third as compared with hot air ovens, and improved quality and large cost savings resulted.

Before the war the average demand for telephone cable in this country amounted to about two hundred million conductor feet per week. During and after the war this demand steadily increased until now it amounts to about six hundred million conductor feet per week or about thirty billion feet per year, requiring annually forty thousand tons of copper wire, seventy-five thousand tons of lead, and six thousand tons of insulating paper.

Cable-making Machinery

In planning for the manufacture of this quantity of cable, the design of all machinery was reviewed and changes made wherever possible to improve quality or increase output.

A great deal of work was done in improvement of insulating machines, and a ten-head vertical type insulator was developed to replace the older five-head horizontal type for non-quadded light gauge wire. In designing the new machine many improvements were incorporated. The old machines had been built to handle relatively strong paper and heavy wires, and studies indicated that to insulate finer wires successfully with lighter paper, also to run at high speeds without stretching the wire, and to apply a uniform wrapping without backlapping or folding over of the paper and with low breakage per pad the insulators must be rigid, the tension on the wires should be uniform and both supply and take-up mechanisms should operate smoothly.

The relative floor space per head for the ten-head machine including operator's space is about 60 per cent of that taken by the five-head machine but based on production the relative space per unit of production is about 50 per cent. The new machine runs at a head speed of about 3,000 R.P.M., carries a 12-in. pad of paper, and in general is a very substantial machine.

The insulating head, the vital part of the insulating machine, has undergone many changes to accommodate the thinner, narrower insulating papers. One of the most important of these has been to improve the tension mechanism which now consists of a very small multiple disc clutch actuated by a system of levers so that a very light but very uniform tension is applied at all times. This is not only making possible the use of smaller paper ribbon but may permit of changes in the composition of the paper with resultant cost savings. This head is shown in Fig. 8.

Another desirable feature in a paper insulating machine is a bare

wire detector as the insulation sometimes parts after passing through the sizing die or polisher as it is called, separates for a few inches and then picks up and goes on. Many electrical devices have been tried and practically all have the objection of high maintenance cost. A

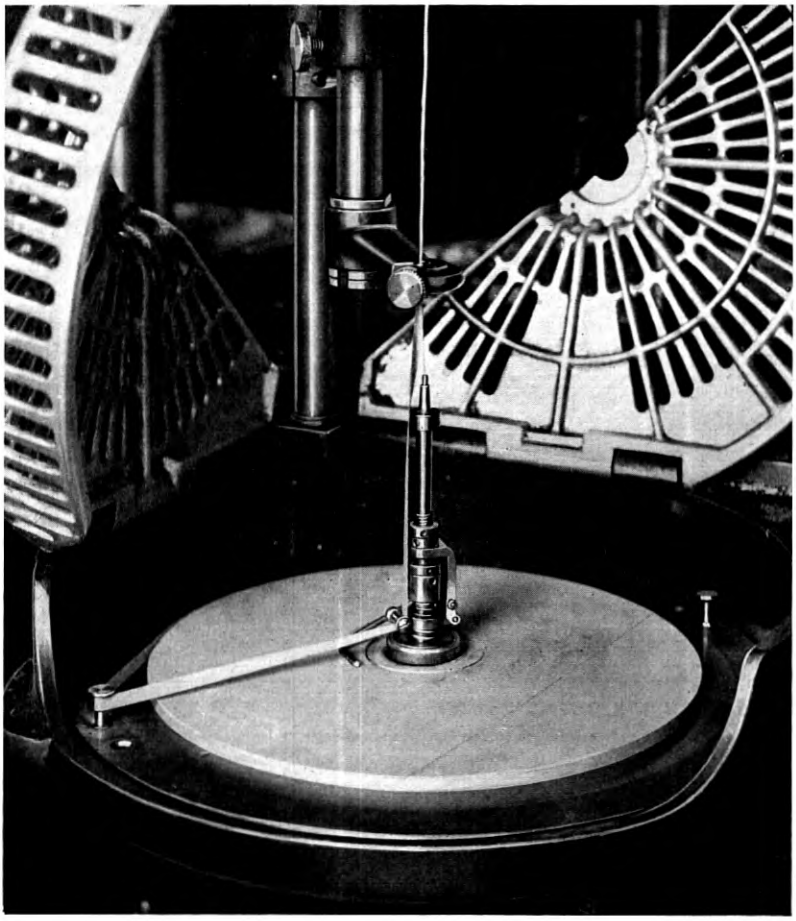


Fig. 8—Paper insulating head

very simple and effective remedy was the installation of a second polisher placed between the capstan and take-up spool which catches broken paper and pushes it back until the operator sees and repairs it.

The insulating machine used for heavy gauge wire is an eight-head machine built along the same general lines as the ten-head machine. This is illustrated in Fig. 9.

The method of splicing the copper wire is by means of a transformer, the low voltage side of which is equipped with clamps for holding the

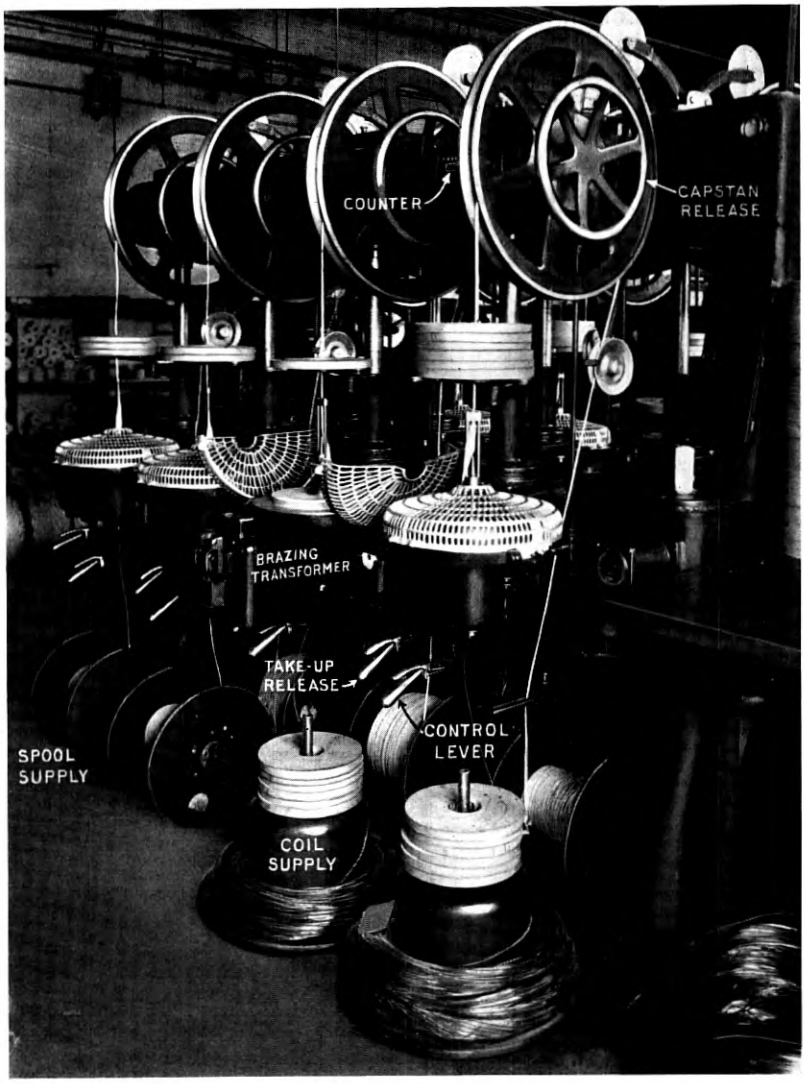


Fig. 9—Heavy wire insulator

two ends of wire which are butted together, heated by electric current and brazed by the application of borax flux and silver alloy solder. The transformer windings are so designed with low internal resistance

that, although different sizes of wire may be handled, the resistance of the wire between the clamps is so large in proportion to the total resistance that it automatically controls the current and prevents overheating of the wire.

Splices in the insulating paper are made by the application of a thin strip of gummed paper.

New twisting machines for non-quadded light gauge wire have been developed and these machines have some unique features which are worth a word of explanation. Fig. 10 shows schematically the old

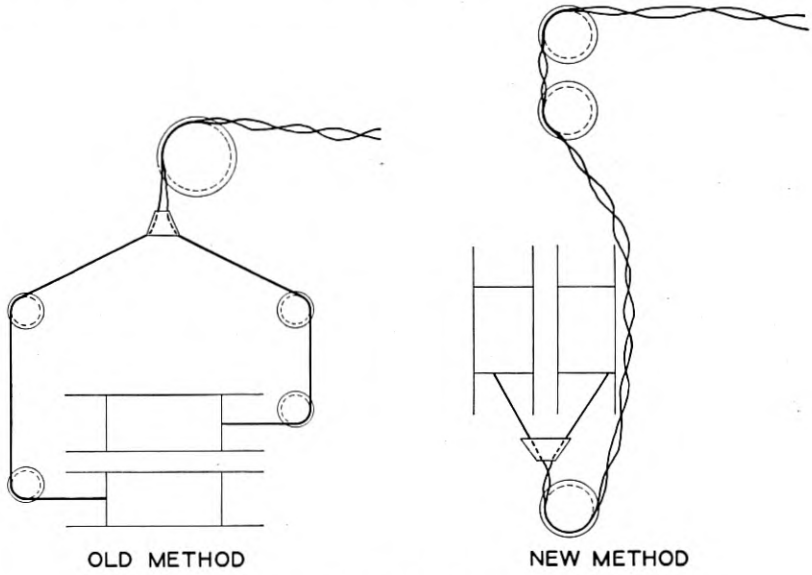


Fig. 10—Schematics of old and new type twisters

type of twister used ten years ago in which the two spools were placed with axes vertical inside of a flier which carried guide bushings through which the wire from the two spools was brought up to the center of the yoke and to the capstan. These machines operated at 500 R.P.M. and produced one twist per revolution. Assuming a 3-in. twist, the output would be about 125 feet per minute. In the new machines the spools are mounted side by side in a flier, the spools not revolving around each other, with axes horizontal, and the wire from each is taken off in a downward direction around a guide pulley and then up through the flier, around another guide pulley and to the capstan. With this arrangement two twists per revolution of the flier are produced and, as the machine is built to operate at 1,000 R.P.M., the out-

put for double the speed of the old machine is four times as great or about 500 feet per minute for a 3-in. twist. Additional features are

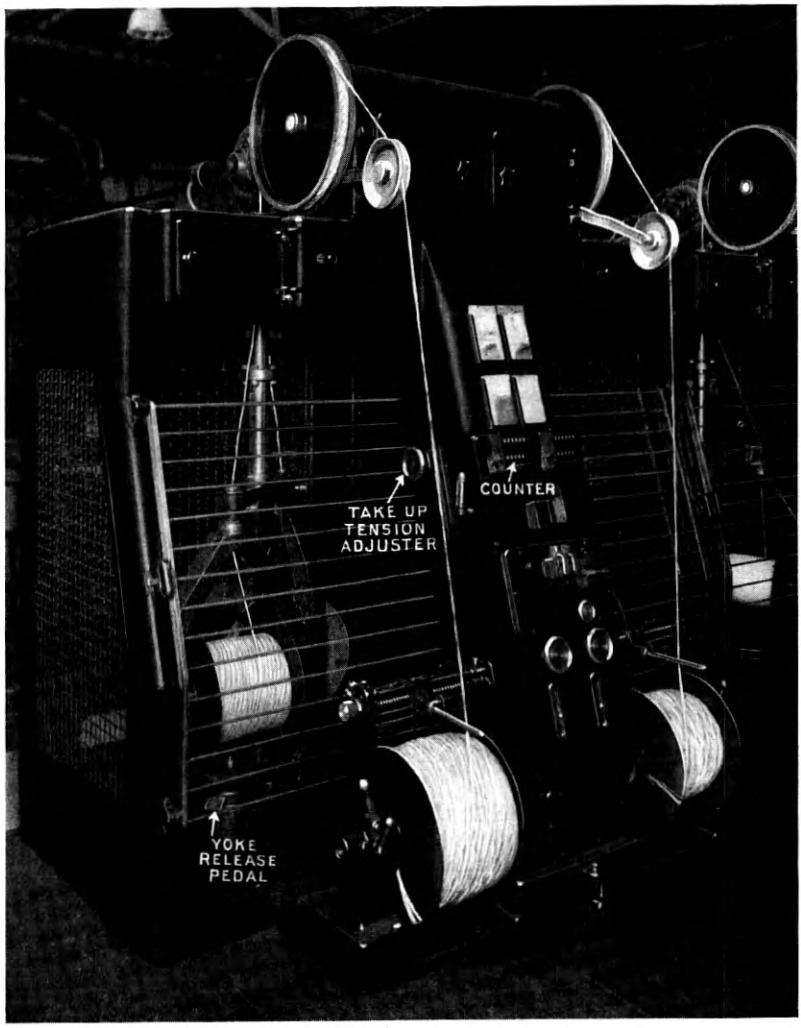


Fig. 11—Combined twisting and quadding machine

special tension devices to insure uniform tension on the wire and supports to assist in loading spools of wire into the yoke.

The twister for pairing and quadding heavy gauge wire in one operation is shown in Fig. 11.

Each spool, containing two conductors, is mounted in a yoke which revolves on its own axis to give the pair twist and the two yokes are revolved around each other to give the quad twist. This is accomplished by an arrangement of change gears from which can be obtained practically any length or direction of twist desired.

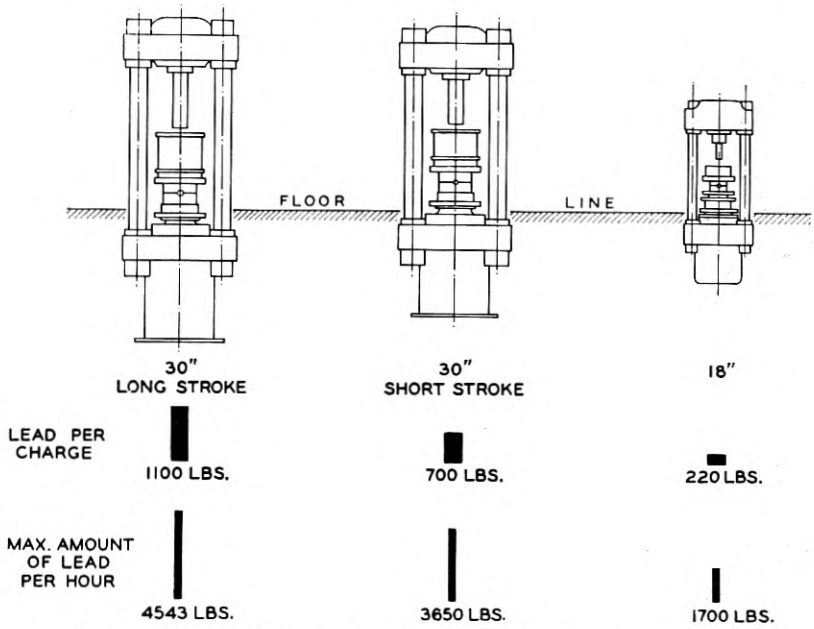


Fig. 12—Schematic showing relative increase in size of lead presses

Modern stranders follow the same general line as the older stranders but the whole design has been reviewed in detail with the view of strengthening and perfecting, and improved tension devices have been developed consisting of a tension arm actuated by the pair which in turn applies a brake to or removes it from the reel head. These are adjusted to give a tension of about three pounds per pair which causes no stretch and prevents over-running. With these, it is possible to run very fine wires at a minimum tension with a maximum smoothness of operation. The drums are gear driven and are capable of running up to 100 R.P.M.

After stranding, the cores are dried under vacuum to remove the moisture from the paper and then are covered with a lead alloy sheath.

It is necessary after the cable is removed from the vacuum drier to keep it in an atmosphere of a low moisture content until the lead sheath is applied. This was formerly accomplished by placing it in an oven

at a temperature of about 160 to 180° F. with a resultant relative humidity of not over 10 per cent. Cables maintained at this humidity would pick up very little moisture but in transit from the vacuum drier to the storage oven some moisture might be absorbed; also working in and out of these hot ovens was not particularly pleasant. Therefore, a method was developed for installing the vacuum driers in such a way that one end opens into an enclosed storage area in which the air is maintained at a temperature of about 100° F. and a relative humidity of less than 10 per cent until the cables are covered with lead. This temperature and humidity are obtained by cooling the incoming air to a dew point corresponding to the temperature and relative humidity desired and then passing it into the oven. A considerable engineering problem was involved in determining the heat given off by the vacuum driers and the hot cables and the additional moisture introduced by infiltration through walls, doors, etc.; also the relation between relative humidity, moisture content of paper and electrical characteristics presented a most interesting field for study.

The method outlined above has proved very satisfactory as the cables do not absorb enough moisture to affect their electrical properties and the conditions in the storage area are not unpleasant; in fact, during the summer time they are somewhat more agreeable than the outside air during periods of high humidity.

The process of applying lead sheath to cable is one which has not undergone any change in principle since sheath was first applied directly to the cable instead of cable being pulled into it. There have been, however, a number of developments tending to improve the quality or increase the output.

In covering a large cable something more than half of the total time of one cycle of operation is taken up by filling the cylinder with lead and cooling under pressure to the point where it can be extruded. The tendency, therefore, has been to build presses with larger lead containers in order to increase the time of extrusion relative to the total cycle.

The diagram (Fig. 12) shows schematically an early type of press, one which was considered standard a few years ago, and one of the presses designed and built recently. Underneath each press is a figure showing the lead content per charge and the relative amount of lead extruded per hour by each of the three presses.

As will have been noted from the diagram, the stroke of the newest type of presses is about one foot longer than that of the former presses although the diameter of the lead container and the diameter of the water ram are the same.

The pressure for operating these presses is furnished by a hydraulic pump, pumping water at six thousand pounds pressure per square

Fig. 13—Electric tractor and trailer for handling cable reels.

inch. Presses were formerly connected to four plunger vertical type pumps, but it was found that more water could be used with the large

Fig. 14—Insulating machines

sizes of cable and, therefore, new pumps were built with six plungers, giving a proportionally greater output. The diameter of the lead ram

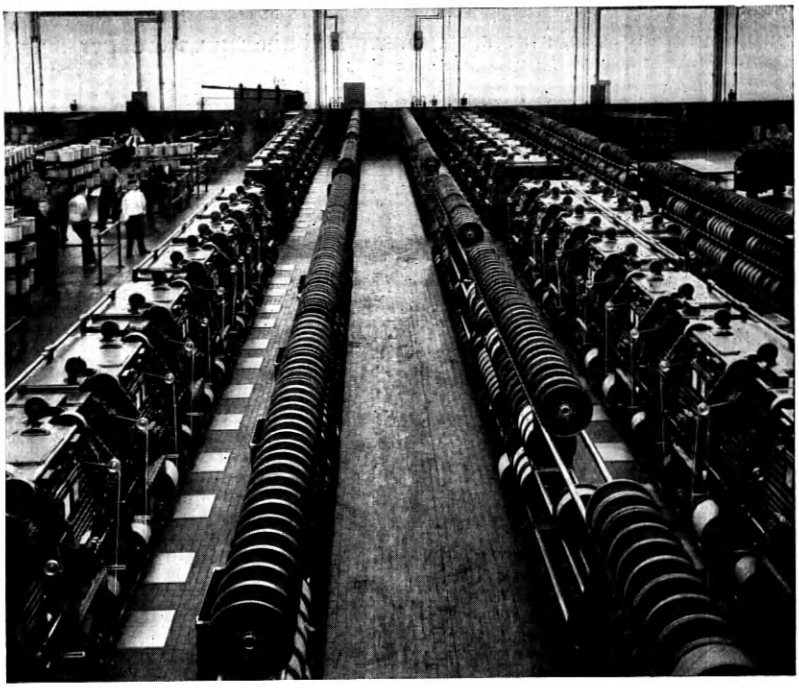


Fig. 15—Twisting machines

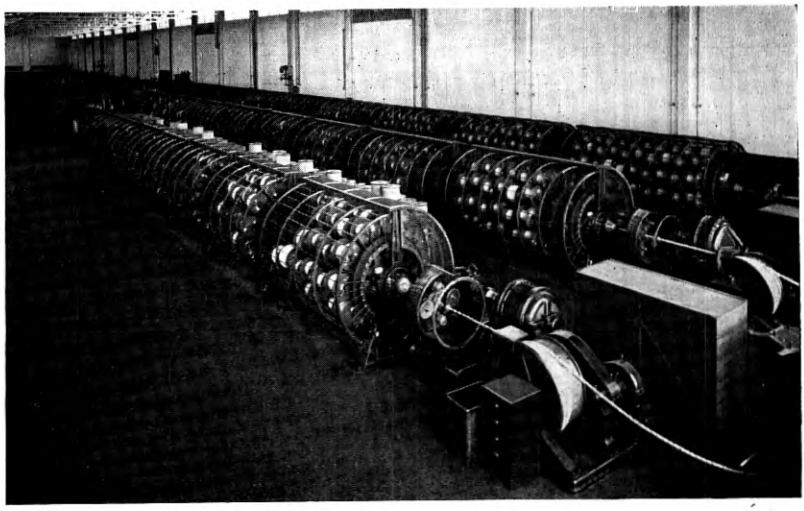


Fig. 16—Stranding machines

Fig. 17—Lead press equipment

Fig. 18—General view of reel yards

is one third that of the water ram, so that the pressure on the lead during extrusion is about 54,000 pounds per square inch.

Aside from increasing output many studies have been made to determine the exact mechanism of lead extrusion, the relative flow of lead in different parts of the extrusion block, the effect of application of heat at different points, etc.

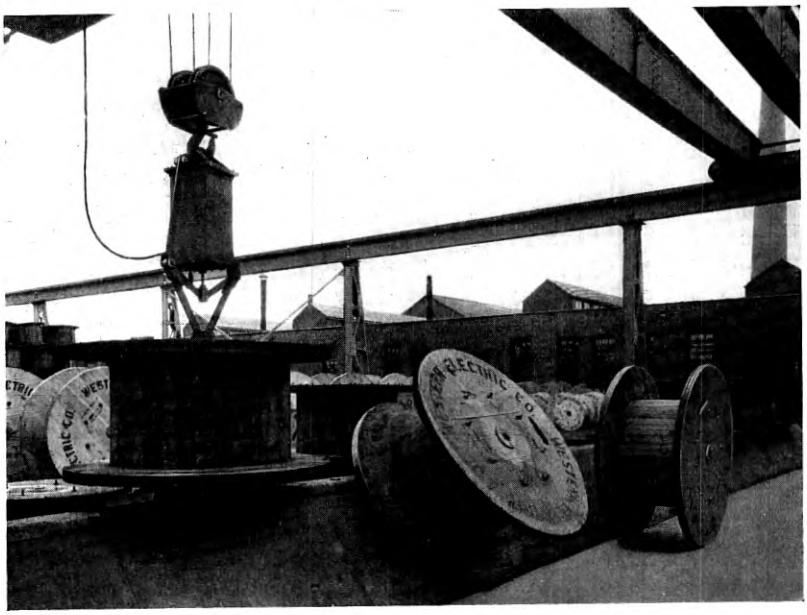


Fig. 19—Delivery of empty reels from yard

An interesting experiment consisted in filling an extrusion block with layers of different colored waxes and noting their flow under pressure. This gave valuable data as to the proper contour of the extrusion chamber.

The concentricity of sheath is affected not only by the contour of the extrusion chamber but also by the manner in which heat is applied; and thickness is affected by temperature and speed of extrusion so that the human element is an important factor, and it is necessary to have thoroughly trained and reliable operators on this kind of work. Temperature indicators are used to show die block temperatures and the temperature of the molten lead is automatically controlled and recorded.

Testing, Storage and Shipment

Handling of lead-covered cable on reels, the total weight of which runs from one to five tons, is a very distinct problem. This handling from press to test is done by a crane which picks up the reels and carries them to the place where they are to be tested for insulation resistance, capacitance, dielectric strength, etc.

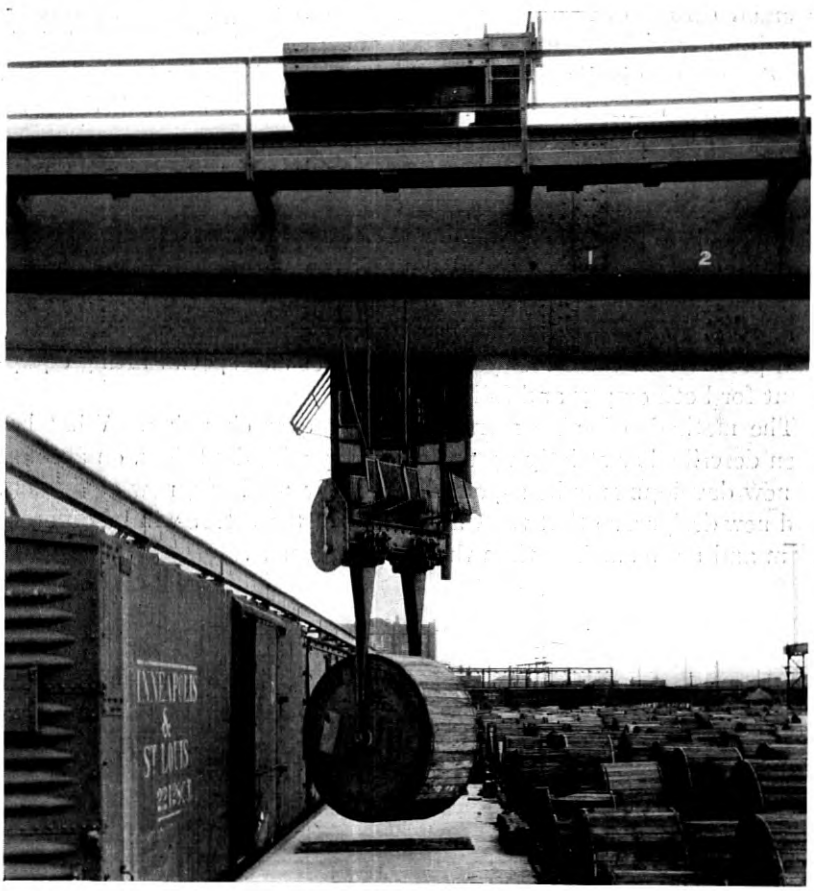


Fig. 20—Crane placing reels on loading platform

After the cables are tested, the ends are sealed and wooden lags are fastened around the periphery of the reels after which the cables are taken to a storage yard until the customer's order is completed, at which time they are shipped. A special tractor and trailer, Fig. 13, has been developed and substituted for manual handling.

Handling cables from the reel yard to the loading platform was a very serious problem, particularly in the winter during snow storms. This was taken care of by the installation of overhead cranes for picking up reels and placing them on the platform.

The lifting mechanism for empty reels consists of a solenoid-operated plunger controlled by the crane operator. The reels are turned on the side, the plunger inserted in the bushing and the operation of the solenoid throws out two lugs which prevent the plunger from being withdrawn and lift the reel. When the reel is to be released, it is put down on an inclined surface which turns it back on to its flanges. This method of lifting empty reels permits them to be stacked one on top of the other and saves storage space.

The lifting mechanism for full reels consists of two side arms with lugs moved horizontally by means of a double-threaded screw and a motor controlled by the crane operator. With this device the crane operator can pick up and put down any reel without the assistance of a ground man.

Figs. 14 to 20 show insulating, twisting and stranding machinery, lead presses and cable reel yard with cranes and special lifting equipment for both empty and full reels.

The methods of cable manufacture are ever changing. What has been described as strictly up to date today will, doubtless, on account of new developments be superseded by new methods, new equipment and new designs, so that the Cable Plant of the future will be different from and more efficient than that of the present.

ERRATA: *Bell System Technical Journal*, April, 1928

Page 327, Table 2—Interchange the number "200" of column 6 and number "600" in column 8.

Page 328, beginning line 4, should read—(*a*) Phantom to phantom; 1 represents the two wires, connected in parallel, of one pair of a quad. 2 represents the two wires in parallel of the other pair of the quad, and 3 and 4 represent similarly the pairs of another quad.

Page 347—Figure 3 should be inverted.

Bridge for Measuring Small Time Intervals

By J. HERMAN

SYNOPSIS: A bridge circuit for measuring time intervals from about one ten-thousandths of a second up to several seconds is described and its operation explained. The device is fairly accurate and easy to operate and gives the results of measurements in fractions of a second directly. Its calibration can readily be determined mathematically since it is dependent only upon the values of certain capacities and resistances used in the measuring circuit.

TO the large family of measuring devices making use of certain principles of electrical balance there has recently been added a new member. This measures the elapsed time between the opening or closing of one set of contacts, and the subsequent opening or closing of another set of contacts, the agency employed for operating the contacts being immaterial. The particular form of the device described below, was designed primarily for use in adjusting the operating and releasing times of the voice-operated switching relays at the terminals of the transatlantic radio telephone circuit. In this form or with minor changes, it is applicable to the measurement of intervals of time in the operation of a large variety of other types of apparatus.

The new time measuring device is simple and easy to operate, its operation consisting merely of opening and closing a key repeatedly and securing a balance by observing a meter. The balance is secured by turning one or more dials which are calibrated in fractions of a second.

A range of measurements extending from about one ten-thousandths of a second up to several seconds is readily obtainable and an accuracy of measurement to within $\pm$ 1 per cent can probably be realized over the greater part of this range if sufficient care is taken in the design of the circuit and the selection of the apparatus. In a fairly rugged type of bridge, now in commercial service (See Figs. 3 and 4), which covers the range from one ten-thousandth of a second to one second, and in which little attempt was made to secure a high degree of sensitivity, the results of measurements are accurate to within $\pm$ 5 per cent for time intervals down to about five thousandths of a second. For time intervals below this value the accuracy decreases rapidly, due partly to the fact that the smallest step provided for on the dials is one ten-thousandth of a second and partly due to the effect of variations in the operating time of the relays used in the bridge.

Because of its simplicity and accuracy, the bridge is especially valuable for making a series of time measurements to determine the

best adjustment and the most desirable circuit condition for the operation of a relay or similar device. With the bridge, this requires very little time, especially since the results of the individual measurements are immediately available to guide the work. With the oscillograph, several hours may be required and the results obtained are available only after developing and analyzing the oscillograms.

The calibration of the bridge is determined by the values of certain capacities and resistances in the measuring circuit. These may usually be selected with sufficient accuracy during manufacture so that the bridge requires no further calibration after it has been constructed. In fact, it has been found practicable to design the bridge so that the steps on a standard decade resistance box correspond to decimal fractions of a second.

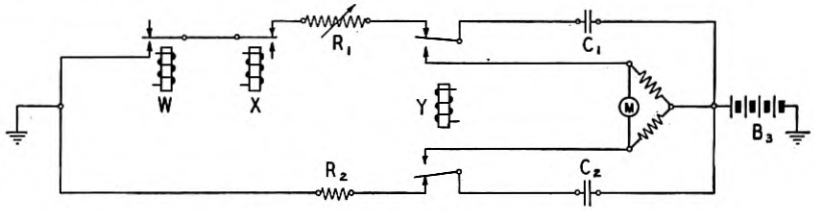


FIG. 1

The principles underlying the measurement of an interval of time will be explained in connection with Fig. 1. Two condensers C_1 and C_2 of unequal capacity are charged from a common battery B_3. The condenser C_1, which has a larger capacity than C_2, is charged through an adjustable high resistance R_1 during the time elapsing between the operation of relay W and the subsequent operation of relay X. This elapsed time is the interval of time to be measured and the charge accumulated on the condenser is an accurate means for doing so. The second condenser C_2 is used merely for comparison purposes. It is charged through a fairly low resistance R_2 and acquires its full charge in a relatively small interval of time.

After the completion of the charging interval, relay Y is operated and the two condensers are discharged simultaneously through a differential meter circuit. If the two charges are equal, the meter will show no deflection, but if they are unequal, it will show a momentary deflection, the direction of which will indicate whether the charge on C_1 is too high or too low. By repeating the charging and discharging process a few times and adjusting the value of resistance R_1 in series with the first condenser, the charges on the two condensers can be made equal. When this condition is obtained, the interval of time during

which the charging took place may be determined from the value of the high resistance.

The relationship between the interval of time of charging and the value of resistance required to make the charges on the two condensers equal is a direct proportion. This will be obvious from an inspection of the general equation for the charge at any instant on a condenser which is being charged through a high resistance. The equation is

$$q = Q(1 - e^{-(t/CR)}), \tag{1}$$

where

q = charge at time, t,

t = elapsed time in seconds since the charging began,

Q = final or maximum charge on the condenser,

C = capacity of condenser in farads,

R = resistance in ohms in series with the condenser,

e = base of Naperian logarithms.

Since q is always made equal to the charge on the comparison condenser and since the charging battery is common to the two condensers, therefore, using the symbols shown in Fig. 1, C_2E may be substituted for q and C_1E for Q. The equation then becomes

$$C_2 = C_1(1 - e^{-(t/C_1R_1)}). \tag{2}$$

As mentioned above, the two condenser capacities are kept constant. Therefore, any change in t requires a proportional change in R_1 in order to satisfy the equation.

Fig. 2 is a schematic circuit diagram of the complete time measuring bridge showing the manner in which it may be connected to a representative type of circuit to be tested (shown by dotted lines). The symbols used to designate the various circuit elements in this figure are the same as those used in Fig. 1 and since the principles of operation have already been explained in connection with the latter figure, a comparison of the two figures will aid materially in understanding the detailed circuit arrangements of the bridge.

As shown in Fig. 2 the bridge is arranged to measure the operating time of a voice operated switching device consisting of a detector and a relay Z in the output circuit of the detector. The input circuit of the detector is connected to the output circuit of an oscillator and may be opened or closed by contacts on one of the bridge relays (relay W). This relay is under the control of key K which, when closed, causes relay W to operate and complete the oscillator connections to the

detector thereby initiating the operation of relay Z. At the same time, another set of contacts of relay W close the charging circuit of condenser C_1 thereby permitting the charging of this condenser through the high resistance R_1. These conditions remain unchanged until the armature of relay Z has reached its m contact and caused the operation

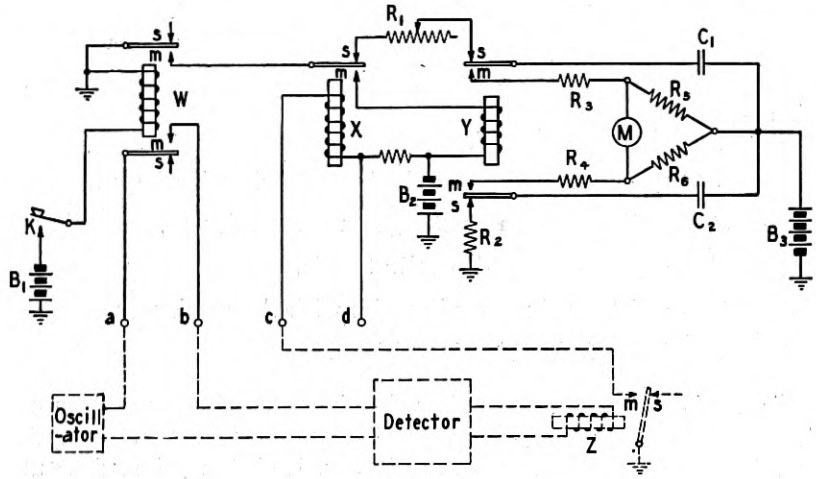


FIG. 2

of relay X. The latter relay first opens the charging circuit of condenser C_1 and then causes the operation of relay Y. The operation of relay Y discharges the two condensers C_1 and C_2 through the differential meter circuit composed of the meter M and the two equal resistances R_5 and R_6. Additional resistances R_3 and R_4 are connected into the discharge circuit to limit the discharge current and prevent sparking at the relay contacts.

If the meter shows no deflection at the instant of discharge, it indicates that the two charges are equal and that the operating time of relay Z is as indicated by the value of resistance R_1. If the meter shows a deflection, the key K should be opened and closed repeatedly and the resistance R_1 adjusted until the meter shows no deflection.

In order to prevent a quick double deflection of the meter when a balance has been reached, it is necessary that the time constant of the two branches of the discharge circuit be approximately alike. For this reason, the ratio $R_4 + R_6$ to $R_3 + R_5$ of the discharge circuit resistances should be approximately the same as the ratio C_1 to C_2 of the condenser capacities.

The releasing time of relay Z; that is, the time required for the armature to leave its m contact after the oscillator is removed from the detector, may be measured by making a few simple changes in the

connections. The oscillator is connected directly to the input of the detector and the terminals a and b of the bridge are connected across the oscillator output. Contact m of relay Z is connected to terminal d, and terminal c is grounded. The closing of key K will then cause a short circuit to be placed across the output of the oscillator, thereby initiating the releasing of relay Z. As soon as the latter occurs, a short circuit is removed from the winding of relay X, allowing this relay to

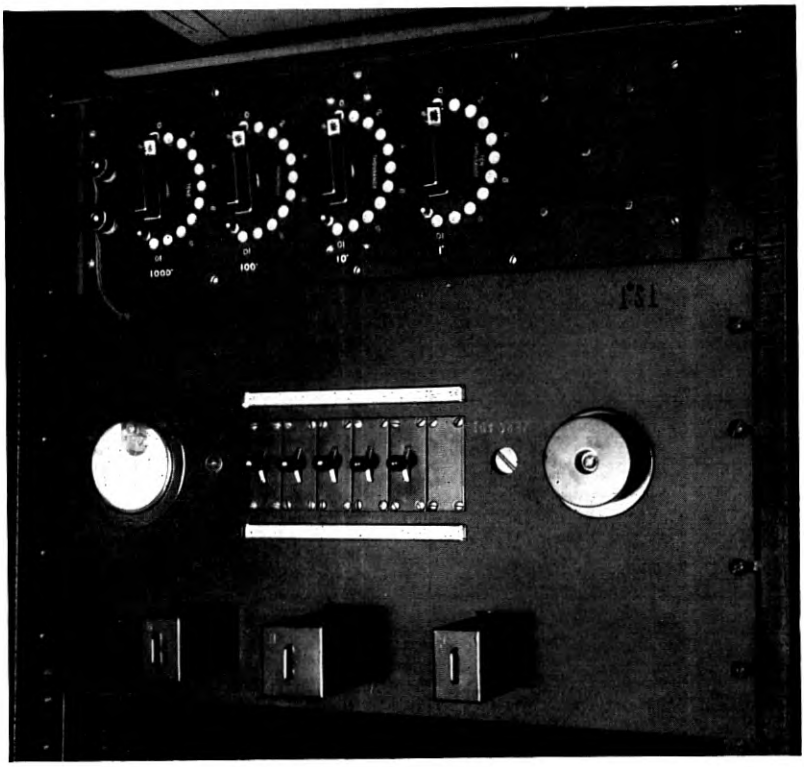


FIG. 3

operate and open the charging circuit of the condenser C_1. Except for the differences noted, the operation of the bridge is the same as described above.

Other conditions of measurement, for example, a measurement of the time required for the armature of relay Z to reach its s contact after a short circuit is applied to the output of the oscillator will be obvious from the two examples given. It should also be obvious that a battery and suitable resistances may be substituted for the oscillator and detector and measurements made in this direct-current circuit in the same manner as above.

Assuming that the bridge has been properly calibrated from theoretical considerations of the constants of the measuring circuit, there are two possible sources of error in the results obtained. These errors are due to the time of operation of the two switching relays W and X. In the case of relay W an error will be caused if its two sets of contacts do not close simultaneously. In the case of relay X the source of

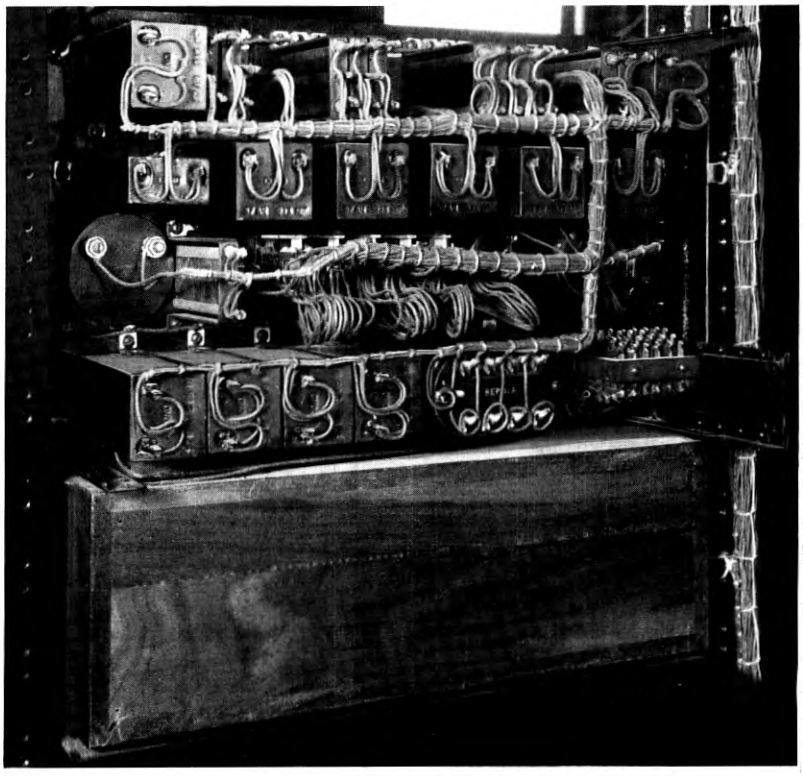


FIG. 4

error is due to the time required for the relay to open the charging circuit at its s contact at the end of the operating or releasing time which is being measured. The combined error due to the operation of the two relays may actually be determined by means of the bridge itself and substracted from the results of the measurements. To determine the error, terminal a is connected to ground and terminal b to terminal c. The key K is then operated in the normal manner and a balance secured on the meter. The readings of the dials will then indicate the error of the bridge.

In order to avoid the necessity of correcting the measurements for the error in the bridge, a value of resistance corresponding to the error may be connected permanently in series with R_1. This should be made variable if a high degree of accuracy is desired, because the operating time of the two relays in the bridge may change slightly from day to day and small readjustments of the resistance will, therefore, be necessary.

The sensitivity of the bridge, that is, the ease with which small intervals of time can be distinguished by the deflection of the meter, is dependent upon the sensitivity of the meter, the voltage of the common charging battery, the value of capacity, and the ratio of the capacity of the large condenser to that of the small condenser. With a particular type of meter, the larger the voltage, the condenser capacities, and ratio of condenser capacities, the more sensitive will be the bridge.

Although the chief use for the bridge up to the present time has been in connection with voice operated relay devices, its future use will undoubtedly be extended to other fields. Some indication of the uses to which it might be put may be obtained from the following suggested applications.

1. For measuring the functioning times of electromagnetic switching arrangements in various kinds of communication and signaling systems.
2. For measuring propagation time at different frequencies over telephone circuits and "lag" in telegraph instruments and circuits.
3. For studying the operation of a machine with a view to improving it by measuring the speed of operation of various parts and the relative time of operation of certain parts with respect to other parts. Electrical contacts would have to be provided temporarily at suitable places on the machine.
4. For maintaining the proper adjustment of time-limit over-load relays, circuit breakers, etc., on power circuits.
5. To determine the rate of acceleration of motors or other machinery by employing a suitable centrifugal contact arrangement.
6. In psychological tests for determining whether the time of response of a person to a particular signal is above or below a required value.

Numerous other applications to laboratory and field work might be suggested where such a time measuring device could be employed to considerable advantage. The cases mentioned above are not necessarily the most practical but they serve to illustrate the capabilities of the device as described or when provided with simple modifications.

ERRATA: *Bell System Technical Journal*, April, 1928

Page 327, Table 2—Interchange the number "200" of column 6 and number "600" in column 8.

Page 328, beginning line 4, should read—(*a*) Phantom to phantom; 1 represents the two wires, connected in parallel, of one pair of a quad. 2 represents the two wires in parallel of the other pair of the quad, and 3 and 4 represent similarly the pairs of another quad.

Page 347—Figure 3 should be inverted.

A Method of Rating Manufactured Product

By H. F. DODGE

SYNOPSIS: This paper outlines a method of rating manufactured product. In the particular form here described, the rate has been found very useful for measuring the quality of communication equipment and materials entering the plant of the Bell System. While the primary object is control of quality of finished product, it is proving useful for measuring the workmanship of individual operators and groups of operators engaged in similar production work. Particular attention is directed to the statistical aspects of the rate to show how it can assist in controlling quality.

GIVEN a product whose quality is dependent on a number of diverse characteristics, the following questions and others similar to them frequently require answers. Has quality been satisfactorily controlled? Is there any general trend in quality either upward or downward? How does current quality compare with that of a year ago?

These are questions of importance to the manufacturer. Qualitative answers can often be given on the basis of general knowledge by those familiar with the details of manufacturing performance but such answers tend to be inaccurate or biased. What is often wanted is some statistical index based on quantitative data, a figure which balances the favorable features against the unfavorable to give an overall picture of quality *on the average.*

To get such a picture does not in general require special data. The detailed data obtained in the course of routine inspection, while often used only for the immediate purpose of determining the satisfactoriness of individual lots of product, are just what is needed for the present purposes. These inspections are critical examinations of the features that are essential to proper operation of the product in service. Hence the results are a measure of quality. There are of course many possible ways of classifying and combining this quantitative information, some of which are more efficient than others. The problem is to set up a method of handling the data in a way which will paint as clear a picture as possible of the overall quality.

The rate here described has been found very useful for measuring the quality of communication equipment and materials entering the plant of the Bell System.[1] It recognizes and takes account of the relative seriousness of different types of defects found in the course of

[1] This method of rating is being used extensively by the Manufacturing Department of the Western Electric Company where some of the features outlined in this paper originated.

inspection. For convenience, the rate is made a relative figure which incorporates the features of index numbers used by the economist. Just as the index numbers of cost of living, wages, corn production, etc., indicate current conditions relative to some reference condition as

$$\text{Index Number} = 100\,\frac{\text{Current Cost of Living}}{\text{1914 Cost of Living}},$$

so does the rate reflect current quality relative to that of a selected standard of reference. One of the features of the rate is its assistance in controlling quality, its provision of means for discriminating between chance and non-chance variations from the quality level which should currently be expected.

Character of Inspection

As in other fields much of the inspection work on telephone products consists of critical examinations of essential features to determine whether or not the units of product conform with specification requirements. This is done:

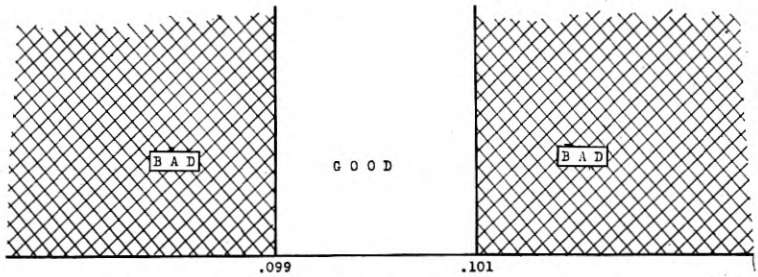
(1) By visual examinations in which obvious defects of material or workmanship are discovered by eye.
(2) By using "Go" and "No Go" gauges or their equivalent, which determine whether a unit does or does not conform with a requirement, or
(3) By using measuring instruments which reveal the numerical magnitude of the characteristic for each unit tested.

To illustrate the last two kinds of inspection, the specification requirement for the capacity of a type of condenser is "not less than .099 microfarad and not more than .101 microfarad." Inspection may be done by the "Method of Attributes," using a test set which shows merely whether the capacity of a condenser is inside or outside of the limits, or by the "Method of Variables," using an indicating or recording meter to show the numerical value of capacity for each test. In these two cases the data, if tabulated, would appear as in Fig. 1. Inspection data used for rating come in both varieties.

Items which Enter the Rate

Commercial measurement of quality by inspection usually consists in a comparison with stated requirements. Starting with the design and a knowledge of what can be accomplished in the shop, allowances for variations in materials, dimensions and salient properties are established in specifications. The aggregate of specification requirements constitutes a standard of quality which the manufacturer holds

before him as an upper limit of attainment. To him, perfect performance is 100 per cent conformance with requirements and the resulting product he regards as of "perfect quality." The rate encompasses this narrow viewpoint of quality and measures the success to the manufacturer in living up to this adopted standard.



CAPACITY (in Microfarads)

INSPECTION BY METHOD OF ATTRIBUTES		INSPECTION BY METHOD OF VARIABLES	
Condenser No.	Observation	Condenser No.	Observation
1	Good	1	.0991
2	Good	2	.1006
3	Bad	3	.0985
4	Good	4	.0995
—		—	
—		—	
—		—	
121	Good	121	.0999
122	Bad	122	.1013
123	Good	123	.0994

Fig. 1—Two methods of measuring the quality of condensers in respect to capacity

The only items which enter the rate are the "defects," i.e. failures to meet requirements, found in the course of inspection. Experience has shown that percentage non-defective, the ratio of perfect parts to the total parts, while useful for certain classes of investigation, is not a very satisfactory yardstick for measuring quality of complex products. This factor fails to take into account two important things:

(1) Defects of different kinds are not equally serious.
(2) Defects of the same kind vary in seriousness according to the degree of departure from specified limits.

Thus a failure to meet a major requirement should have greater weight than a failure to meet a minor one and in like manner the degree of imperfection of a given kind should be taken into consideration.

The rating method recognizes such gradations in seriousness by making use of a system of weighting defects.

Method of Weighting Defects

The seriousness of a defect is judged from the standpoint of the consumer. A defect, if allowed to get into service, means trouble in one form or another, and trouble costs money. Seriousness depends fundamentally upon the evaluation of the loss or expense that would be incurred by using the defective unit. The determination of exact costs of trouble is generally not possible but these costs or, better, the relative costs can be estimated. Such estimates may be based on past experience, judgment, engineering knowledge of service requirements, complaints received from consumers and available information on costs associated with past troubles in service.

A standard set of classes is adopted for defects associated with a given kind of product, the classes being ordered in seriousness and each sufficiently well defined to make the business of classification a fairly simple and uniform process. The following four-fold classification has been found satisfactory for many kinds of telephone products.

Class "A" Defects—Very serious.

Will render unit totally unfit for service.

Will surely cause operating failure of the unit in service which cannot be readily corrected on the job, e.g. open induction coil, transmitter without carbon, etc.

Liable to cause personal injury or property damage.

Class "B" Defects—Serious.

Will probably, but not surely, cause Class "A" operating failure of the unit in service.

Will surely cause trouble of a nature less serious than Class "A" operating failure, e.g. adjustment failure, operation below standard, etc.

Will surely cause increased maintenance or decreased life.

Class "C" Defects—Moderately serious.

Will possibly cause operating failure of the unit in service.

Likely to cause trouble of a nature less serious than operating failure.

Likely to cause increased maintenance or decreased life.

Major defects of appearance, finish or workmanship.

Class "D" Defects—Not serious.

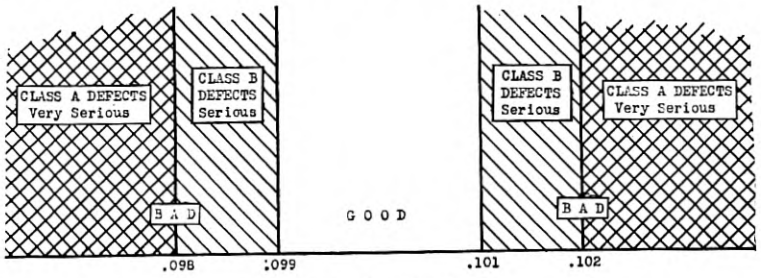
Will not cause operating failure of the unit in service.

Minor defects of appearance, finish or workmanship.

It should be pointed out that the number of classes to be used is arbitrary. Two classes, major and minor, may be sufficient for some

relatively simple products. The number of classes that can logically be used in any case depends upon the accuracy which can be attained in making estimates of relative seriousness.

Before proceeding further it may be well to indicate how the defects for features which are inspected as "variables" are weighted. Take the illustration accompanying Fig. 1. Any failure to meet the commercial limits of .099 and .101 microfarad will result in irregularities in transmission such as the distortion of the words spoken over a telephone line. The greater the departure from these limits the greater is the seriousness from a service standpoint. Strictly the weight for a defect should depend upon the degree of its departure from a limit but the desired result can be approximated to a satisfactory degree of accuracy by classifying the defects into two or more classes. To illustrate, assume two classes as indicated in Fig. 2. Defects falling within the



CAPACITY (in microfarads)

Fig. 2—Classification of defects for variable characteristics

ranges .098 to .099 and .101 to .102 are serious and can be considered as Class "B" defects in a four-fold classification while defects outside of the two outer limits .098 and .102 are Class "A" defects and can be weighted as such.

COMPUTATION OF THE RATE

A defect is weighted by assigning to it a number of "demerits." For a given kind of product each class of defects has a specified weight. Since the *relative* weights are alone of importance, the scale of demerits may be chosen arbitrarily.

The unit of measurement in the rating plan is "demerits per unit." [2] This factor is the simple sum of the demerits per unit contributed by the different types of defects found in inspection.

[2] The "unit" is commonly a physical unit of product such as a piece part, a partial assembly or a finished unit of apparatus or equipment. Exceptions to this rule have been found desirable for certain complicated types of product, such as switchboard sections or installed central office equipment, in which cases the unit may be a natural element of a physical unit such as a soldered connection, a circuit, etc.

$$\text{Demerits per unit} = \frac{w_1 d_1}{n_1} + \frac{w_2 d_2}{n_2} + \cdots \qquad (1)$$

for all types of defects, where

w_1, w_2, etc. = weight (demerits per defect) for defects of type 1, 2, etc.

d_1, d_2, etc. = number of type 1, 2, etc., defects, and

n_1, n_2, etc. = number of units inspected for type 1, 2, etc., defects.

Instead of using equation (1) directly for indicating quality, it has seemed desirable to establish a rate which by its numerical magnitude gives an immediate indication of whether the quality is better or worse

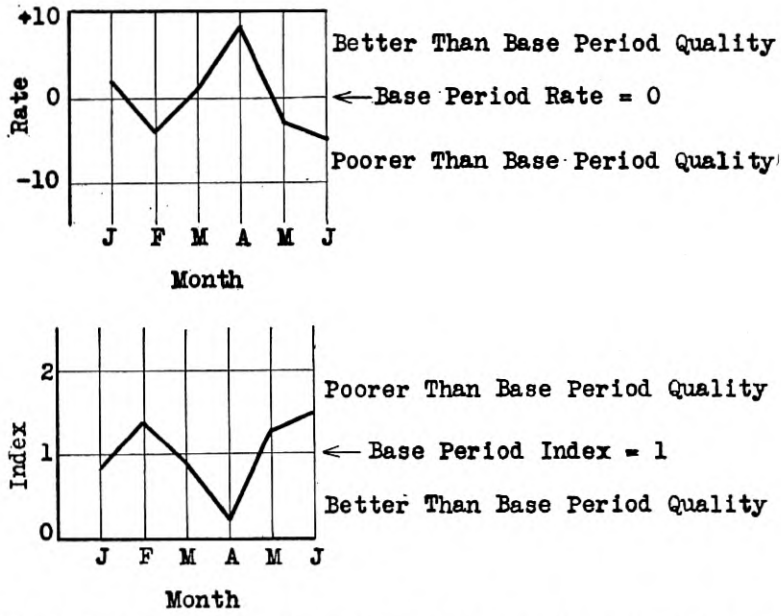


Fig. 3—Relation between index and rate for a given product and period of time

than that of some easily recognized reference condition. Resorting to methods commonly used in constructing index numbers, we therefore select a base period during which the manufacturing conditions and inspection methods were known to be essentially the same as at the present time, determine the demerits per unit for the representative data of that period, and set up the following index:

$$\text{Index} = \frac{\text{Current Demerits per Unit}}{\text{Base Period Demerits per Unit}}. \qquad (2)$$

Since the demerit is an element of badness, this index increases in magnitude as the quality grows worse as shown in the upper chart on Fig. 3. It is preferable to have the rate high when the quality is good and low when the quality is bad. This has been taken care of by using the factor (1 − Index) in the rate equation,

$$\text{Rate} = 10\,(1 - \text{Index}), \tag{3}$$

where the factor 10 is introduced merely to make a convenient scale. This gives a rate of +10 for a product of perfect quality (i.e. no defects found in the material inspected), a rate of zero when current quality is the same as the average for the base period, and a negative rate when current quality is poorer than that of the base period. This equation, as portrayed by the lower chart of Fig. 3, is merely a numerical way of saying "better than" or "poorer than" base period quality and it also tells how much better or poorer.

The choice of base period rests on judgment and knowledge of conditions and must be made with the eyes open. To take care of evolutionary changes in manufacturing conditions for telephone products it has been found desirable to use a *moving* base period [3] of not longer than five years. The use of a somewhat extended period where possible has the advantage of stability in that it tends to smooth out the high and low spots resulting from temporarily abnormal conditions of production such as are liable to recur in the future. The magnitude of the *base period demerits per unit* thus establishes a reference level for quality under *average* conditions.[4]

Quality-Control Feature of the Rate

Rates obtained from week to week or from month to month are used to indicate whether quality has been controlled. If manufacturing conditions are steady and everything is running smoothly, some definite value of rate can be *expected*. But even with a perfectly controlled process, there will be fluctuations above and below the expected rate value, fluctuations resulting from the effects of a large number of causes over which the manufacturer has no control.

[3] By a moving base period of 3 years is meant the three years just preceding the current year. With a moving base period the standard of reference (the denominator of the index) will change slightly at the beginning of each year as one year is dropped and a new one, the preceding, is added to the base.

[4] The average of past experience is sometimes a suitable estimate of *expected* quality but its indiscriminate use for this purpose is to be avoided. For products which are reasonably well controlled this estimate will often serve satisfactorily. Primarily the denominator of the index is a magnitude chosen to represent some standard of reference. The numerical rate obtained at any time reflects quality relative to the standard. It is not essential to the rate that the expectancy feature be stressed in this connection. Expectancy is, however, of importance to the control limits discussed in the subsequent paragraphs.

How low does a rate have to fall before lack of control is indicated? Does a rate of -10 signify that something abnormal has happened? The following discussion gives a method which can be used to detect lack of control.

First of all we must determine the value of rate to be expected, i.e. establish a norm for expected quality. Past experience can usually be used as a guide for this purpose. If the average quality during the base period is considered satisfactory as an estimate of expected quality under current conditions, then the expected rate is 0. If only a portion of past data is judged suitable for this purpose, then the rate figure corresponding to the selected data is the expected value.

The method of establishing limits of expected variation for the rate makes use of statistical methods which have been described elsewhere,[5] but will be briefly reviewed. If the current rate deviates from the expected rate by an amount which is greater than can be attributed to chance, this will be taken as an indication of lack of control.

Just how chance enters the discussion will perhaps be better understood from the following. Each unit of product is the physical result of fashioning and combining various materials by a large number of manual and mechanical operations and processes. Every element in the production process which contributes to the final detailed character of a unit can be considered as a cause. Now the ideal state of affairs, purely conceptual to be sure but nevertheless one which is the goal in all attempts to secure greater uniformity of quality, is one in which each of the elemental causes or groups of causes (affecting a particular trait of the product) functions continuously in the same manner to produce a given elemental effect in the direction of defective quality. Considering overall quality, one group of manufacturing causes is responsible for one type of defect, another group for a second type, etc. The aggregate of these many causes which cooperate to mould the product may be considered as a system of causes. When the concept of constancy-with-time is associated with all of the causes, the system is spoken of as a "constant system of causes," [6] i.e. one whose tendency toward defective quality does not change with time. Product turned out by such a system will be referred to as "uniform product."

For product which is uniform in this sense the rates obtained week

[5] "Quality Control Charts," by W. A. Shewhart, *Bell Sys. Tech. Jour.*, Vol. V, pp. 593–603, October, 1926.

[6] "Application of Statistics as an Aid in Maintaining Quality of a Manufactured Product," by W. A. Shewhart, *Jour. Am. Stat. Ass'n*, Vol. XX, pp. 546–548, December, 1925. It should be noted that the system of causes associated with the *data* used in rating is all-inclusive, encompassing the causes which are responsible for inaccuracies of measurement introduced by inspection as well as the manufacturing causes which affect actual quality.

by week or month by month will fluctuate around some average value according to the laws of chance. For example, in the manufacture of selectors assume conditions are such as to give uniform quality with an expected rate of 0. The rates for batches of selectors turned out weekly will fluctuate about Rate $=0$, the range of variation depending on the number produced each week. A week's output can be regarded merely as a sample of the product which this system of causes would turn out if it were allowed to function in the same manner for an

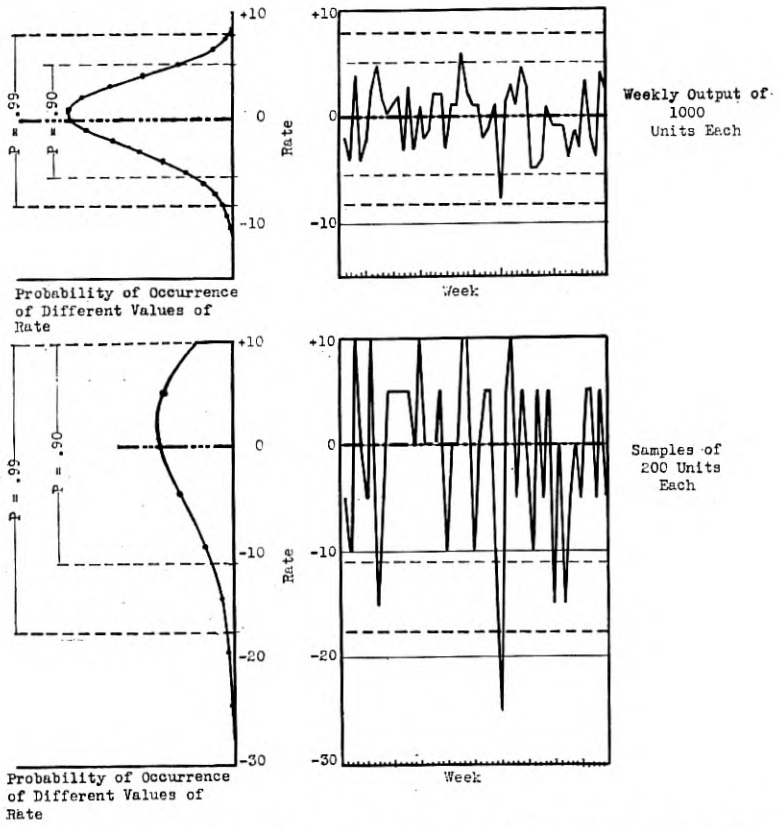


Fig. 4—Typical fluctuations of rates for uniform product whose expected rate = 0

indefinite length of time. The distribution of weekly rates is the same as would be obtained in an ordinary sampling experiment by drawing samples from an infinite warehouse of thoroughly mixed selectors having an average quality represented by Rate = 0. If inspection consists in examining only a percentage of the selectors manufactured, this will be merely equivalent to taking smaller quantities of selectors

from the warehouse and the resulting rates will be spread out more widely around 0 than when the entire product is inspected. These results are exemplified by the two diagrams in Fig. 4.

The probability curves of Fig. 4 represent the basis for setting control limits. The area under the curve between any two limits divided by the total area represents the probability that a single rate will fall between these limits. For a probability of .99 we can say that if the product is controlled at a level corresponding to the expected rate (zero in the illustration) then the chances are 99 in 100 that the current rate will fall within the limits thus established and only 1 in 100 that it will fall outside the limits.

For any product the spread between the limits is governed by two factors, the number of pieces inspected and the value of the above probability. It is necessary therefore to make an arbitrary choice of probability, a choice which will depend on the use to be made of the rate.

The control lines are used primarily to distinguish between those variations which may be attributed to chance causes and those which are more probably the result of some significant change in manufacturing conditions, either production or inspection, and therefore worthy of investigation. The criterion of the suitability of the limits chosen is the percentage of cases falling outside of the limits which on investigation are found to have resulted from some significant departures from current standards of performance.

In setting limits for rates the manufacturer has one point of view and the purchaser another. The manufacturer wishes to detect lack of control as early as possible and is willing to follow up false scents occasionally in his endeavor to prevent the persistence of costly irregularities. The purchaser is more interested in major swings or trends in quality, is not so much concerned with the use of limits for actual control and hence does not desire to instigate fruitless investigations frequently. For many telephone products, experience has indicated that a probability value between .90 and .95 is economical for shop control work while higher values such as .99 or above are better suited for quality reports issued for purposes of general information.

Inasmuch as the rate measures overall quality as determined by a number of different characteristics, its control feature relates particularly to final or partial assemblies of product. This control work should, of course, be preceded by control activities based on the same principles applied to the process inspection data for each of the essential characteristics of the parts which make up the whole.

PARTIAL SUMMARY OF INDIVIDUAL DEFECTS

Type of Defect	Demerit Weight	Base Period 1922	1923	1924	1925	1926 Jan.	1926 Feb.	1926 Mar.	1926 Dec.	1926 Total	Current Year 1927 Jan.	1927 Feb.	1927 Mar.	1927 Aug.
Electrical														
Type 1	100	0	0	0	1	0	0	0	16	24	18	5	14	0
Type 2	100	30	147	168	135	3	20	12	9	171	0	19	0	1
Mechanical														
Type 3	100	1	0	0	0	0	0	0	0	0	0	0	0	5
Type 4	100	0	0	9	9	0	3	1	0	10	1	1	0	0
Mounting and Assembly														
Type 5	35	5	0	1	1	2	11	0	1	44	0	0	0	0
Type 6	35	0	0	0	1	0	1	0	0	7	0	0	0	0
Type 7	20	6	11	8	8	0	0	0	1	5	0	0	1	1
Type 8	20	2	0	21	5	0	0	0	1	13	0	0	0	0
Wiring and Soldering														
Type 9	50	0	0	0	6	0	0	0	0	0	0	0	0	0
Type 10	35	0	0	6	1	0	0	0	0	0	0	0	0	0
Marking														
Type 11	50	3	4	8	8	0	0	0	1	5	0	0	0	2
Packing														
Type 12	50	13	0	0	0	0	0	0	0	0	0	0	0	0
Summary of Total Defects by Classes														
Class A	100	44	162	193	176	3	24	16	27	255	19	28	33	27
Class B	60	21	29	42	21	0	0	5	0	57	1	0	0	9
Class C	25	12	25	55	31	7	25	6	1	131	2	0	2	12
Class D	5	14	36	59	16	1	7	1	6	48	0	0	2	1
No. Inspected*		7,478	24,871	24,524	23,093	1,230	2,184	3,201	3,160	31,385	2,579	2,657	3,424	2,475
Rate		+1.05	+1.64	−0.63	+0.52	+4.9	−5.5	+2.9	+0.3	−1.29	+1.4	−1.7	−0.9	−5.9

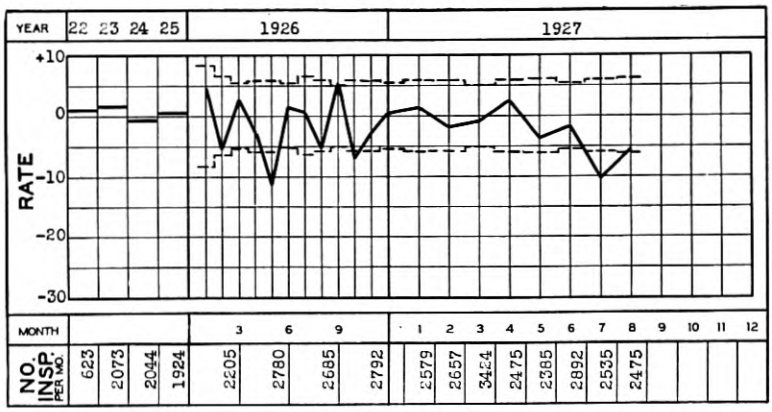
* Same for all types of defects.

Fig. 5—A typical classified summary of the defects found in inspection

Illustrative Examples

Example I—Monthly Rate for General Information Purposes Showing Quality of Finished Product

This example presents a monthly quality report for a kind of telephone apparatus which is manufactured in large quantities for use in



Defect	Wt. w	Base Period			March 1927		
		No. of Defects d	No. Insp. n	Demerits per Unit $\left(\frac{D}{n}\right)=\frac{wd}{n}$	Expected Dem. per Unit $\left(\frac{D}{n}\right)_e$	No. Insp. n	$\frac{w}{n}\left(\frac{D}{n}\right)_e$
Class A.....	100	830	111,351	.7454	.7454	3,424	.02177
Class B.....	60	170	111,351	.0916	.0916	3.424	.00160
Class C.....	25	254	111,351	.0570	.0570	3,424	.00042
Class D.....	5	173	111,351	.0078	.0078	3,424	.00001
Total				Σ.9018			Σ.02380

Computation of Control Limits.—

$$k = \frac{1}{\text{Base Period Demerits per Unit}} = 1.109, \qquad R_e = 0, \qquad K = 3,$$

$$\sigma_{R_e} = 10k\sqrt{\Sigma\left[\frac{w}{n}\left(\frac{D}{n}\right)_e\right]} = 10(1.109)\sqrt{.02380} = 1.711.$$

Limits $R_L = R_e \pm K\sigma_{R_e}$.
$= 0 \pm 3(1.711)$
$= +5.133$ and -5.133.

Fig. 6—A typical rating chart with variable control limits

central office exchanges. The data given in Fig. 5 represent a portion of the summarized results of inspection. The detailed information

given in tabulations such as this is the basic material used in investigation work relating to control of quality. The composite summary at the bottom of the figure is used directly for computing rates.

The quality report for this product is based on the following:

(1) The data are obtained by the check inspection of representative samples of the total product.
(2) The average quality for the base period is taken as the "standard expected" quality for current product. Control lines are thus placed above and below the base period rate of 0.
(3) The limits are computed each month on the basis of the number inspected during that month, using a probability value of .997.

The computations shown below the rating chart of Fig. 6 indicate the work necessary to the determination of the control limits for a given month. The basis of these computations is given in the Appendix. The limit lines on the chart are broken lines merely because the number inspected varies from month to month.

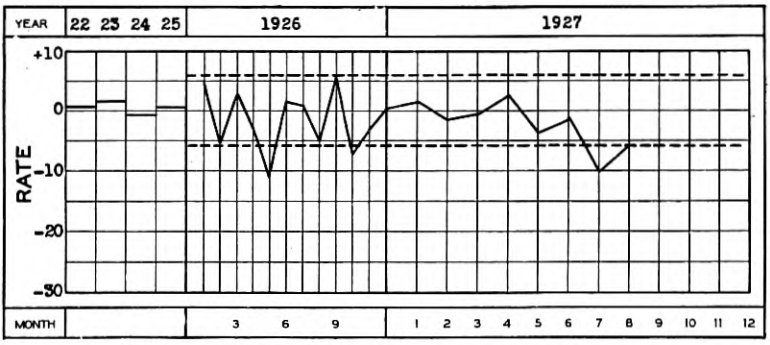


Fig. 7—Rating chart with constant control limits

When production and inspection schedules are fairly uniform so that the number inspected per month remains substantially constant, the limits may be computed once a year using some estimated size of monthly sample and drawn as parallel lines across the chart as in Fig. 7. This approximation is justified when the loss of accuracy thereby introduced is outweighed by the extra charting costs associated with monthly computations of limits.

Example II—Monthly Rates Showing Quality of a Product Before and After a Screening Inspection

Assume the following procedure to be in force.

(1) The shop product is inspected 100 per cent by an inspection group which serves as a screening medium for eliminating defective units.

(2) The product which passes this inspection is subsequently examined on a sampling basis by a check inspector who looks for the same defects.

The data obtained by the screening inspection provide a measure of the quality submitted by the Operating Department. The check inspection data give a picture of the quality of the finished product placed in stock. The rating chart is shown in Fig. 8. In a case of this sort where identical product is handled by two successive inspection groups, it is advantageous to show both rates on the same scale. In this particular instance a rate of 0 corresponds to the base period

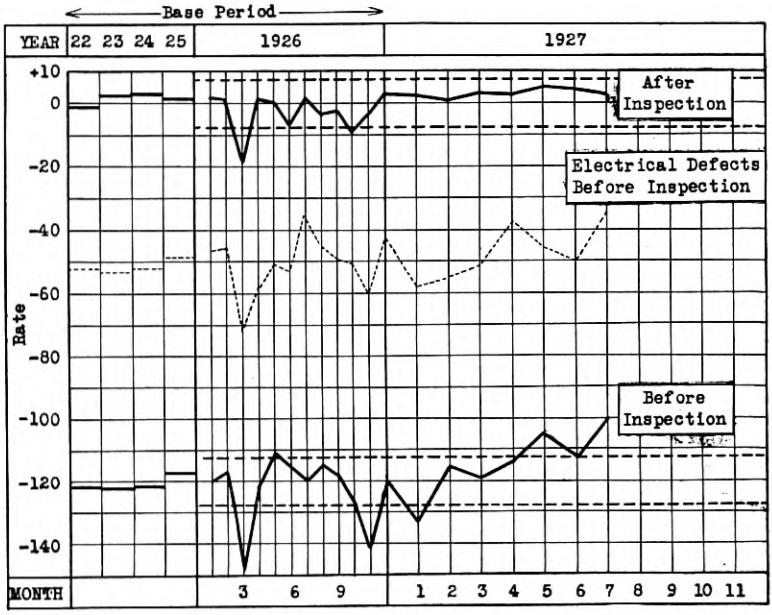


Fig. 8—Rate showing quality of a product before and after a screening inspection

quality of the product submitted to the check inspector. The control limits for the lower rate are drawn above and below the expected level of quality for product submitted to the first group of inspectors and both sets of limits are based on a probability of .95.

The results of the screening inspection can be used directly for controlling the work of the Operating Department. For this purpose it has been found valuable to prepare weekly rates with control limits based on a slightly lower probability value than that used for monthly rates. When the defects can be readily classified into two or more major groups, such as defects for electrical requirements, defects for

mechanical requirements, etc., it has often been found useful to compute sub-rates with respect to such classifications of trouble. A typical sub-rate is indicated by the fine dotted line of Fig. 9. Ex-

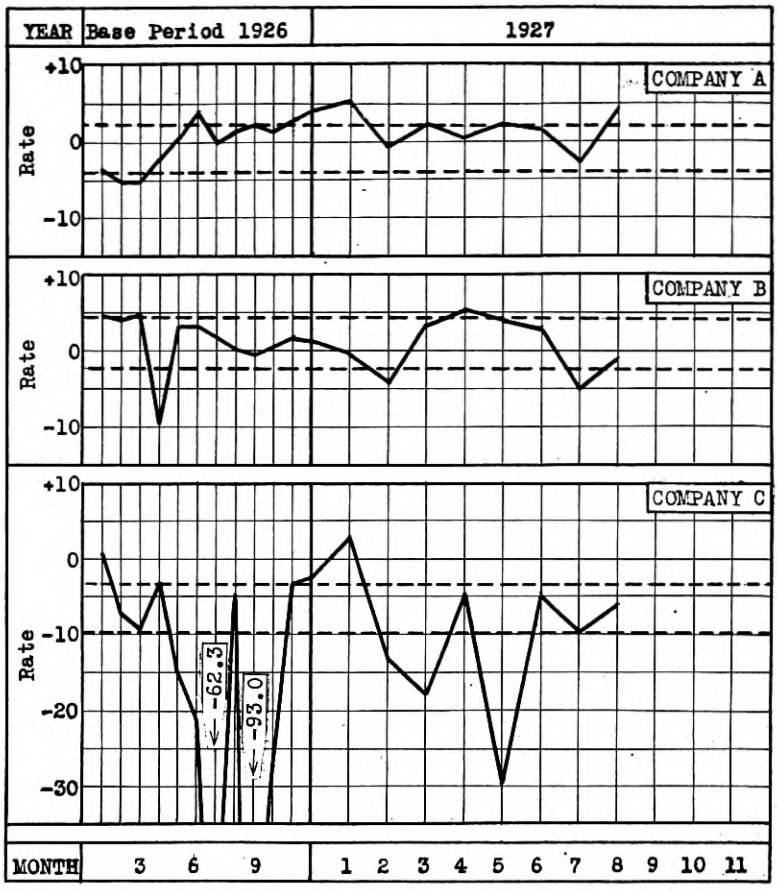


Fig. 9—Rates showing the quality of similar product supplied by three independent companies

perience has also shown the practicability of rating the quality of output of the individual operators and gangs in some classes of work for the purpose of comparing workmanship.

The principal advantage of these latter steps is to provide a ready means for tracing causes of trouble.

Example III—Rates for Comparing the Quality of Similar Product Supplied by Different Organizations

The rate can be used in like manner for comparing the quality of similar product produced by different factories of one company or by different companies when the several manufacturing units are governed by the same set of specifications.

As an example take the case of similar material supplied by three independent companies. There are several factors to be considered in constructing the rates. First of all, some standard of reference must be chosen to represent zero on the rate scale. This might correspond to the average performance of all companies, of the best company, the average of the two best, etc. Secondly, to show the degree of control for each company, the limits should be constructed independently for each company above and below the individual expectancy levels of quality.

The rating chart of Fig. 9 gives a quality picture for a particular kind of material supplied by three independent concerns, two of which are doing consistently better work than the third. Zero rate has been chosen arbitrarily to represent the average base period performance of the two best companies. The rate obtained monthly for any one company thus reflects, by its numerical magnitude, the relation between this company's current quality and that chosen as the standard of reference.

Graphical quality reports of this sort are of value to purchasing organizations in their relations with competing suppliers of similar materials.

Appendix

Computation of Control Limits

Assume

(1) $R = 0$ corresponds to base period demerits per unit.

(2) Control limits are to be set above and below some rate figure, R_e, corresponding to expected demerits per unit. (If expected quality is the same as base period quality, then $R_e = 0$.)

(3) Control limits are to be computed for monthly rates. (The procedure is similar for any other period of time.)

The following symbols will be used:

w = weight (demerits per defect) for a given type of defect.

d = expected number of defects per month, for a given type.

$D = wd$ = demerits for d defects.

n = number inspected for a given type of defect during the month.

$\left(\frac{D}{n}\right)_e$ = expected demerits per unit.

I_e = index for Expected Quality = $\frac{\text{Expected Demerits per Unit}}{\text{Base Period Demerits per Unit}}$.

R_e = rate for Expected Quality.

R_L = control limit value of rate.

$k = \frac{1}{\text{Base Period Demerits per Unit}}$.

σ = standard (root mean square) deviation.

The rate for Expected Quality is

$$R_e = 10(1 - I_e). \tag{1}$$

To find the control limits for the rate, first determine its standard deviation, σ_{R_e}.

$$\sigma_{I_e} = 10\sigma_{I_e}. \tag{2}$$

The index for Expected Quality is given by

$$I_e = k\left(\frac{w_1 d_1}{n_1} + \frac{w_2 d_2}{n_2} + \text{etc. for all types of defects}\right), \tag{3}$$

where d_1, d_2, etc. = expected number of defects per month for defects of type 1, 2, etc., and the subscripts 1, 2, etc., refer generally to the several types of defects.[7]

To find σ_{I_e}, the standard deviation of the index, the d's are considered as independent variables subject to sampling variations. I_e is then a linear function of d_1, d_2, etc., with the constant coefficients $\frac{kw_1}{n_1}$, $\frac{kw_2}{n_2}$, etc. Hence

$$\sigma_{I_e} = \sqrt{\left(\frac{kw_1}{n_1}\right)^2 \sigma_{d_1}^2 + \left(\frac{kw_2}{n_2}\right)^2 \sigma_{d_2}^2 + \cdots}. \tag{4}$$

The values of σ_{d_1}, σ_{d_2}, etc., are evaluated by the following consideration.

Assume that a sample of size N is drawn from a source for which the probability of occurrence of a defect is p. The expected number of defects in the sample is pN, and the standard deviation of the expected number is $\sqrt{p(1-p)N}$. If p is small, the factor $(1 - p)$

[7] In carrying out the computations of rates and control limits, it is convenient to group together all defects having the same "weight" (w) and the same "number inspected" (n), and to let the subscripts 1, 2, etc., of the equations refer to these groups.

can be neglected,[8] which gives $\sqrt{pN}$. Considering the practical case, if the ratio of the number of defects is small compared with the possible number of defects, then the standard deviation of the expected number, d, is equal to $\sqrt{d}$. For many telephone products certain types of defects may occur several times on a single unit of product; for example, when inspection is made for the tension requirement of springs on a relay or for character of soldered connections on a switchboard, then N refers to the number of springs or the number of soldered connections, respectively, inspected during the month. Likewise p refers to the probability of occurrence of a defective spring or of a defective soldered connection. Fortunately, in the determination of the standard deviation it is not necessary to know exactly what p is nor what N is, so long as it is known that p is small (less than .10 for ordinary engineering purposes). This condition is usually satisfied in practice, hence the above result can be used.

Equation (4) then becomes

$$\sigma_{I_e} = \sqrt{\frac{k^2 w_1^2 d_1}{n_1^2} + \frac{k^2 w_2^2 d_2}{n_2^2} + \cdots}. \tag{5}$$

To simplify routine computations, this can be changed in form by removing the factor $\frac{wd}{n}$ from each term (this is merely the Expected Demerits per Unit $\left(\frac{D}{n}\right)_e$ for a given type of defect) which gives

$$\sigma_{I_e} = k\sqrt{\frac{w_1}{n_1}\left(\frac{D}{n}\right)_{e_1} + \frac{w_2}{n_2}\left(\frac{D}{n}\right)_{e_2} + \cdots}. \tag{6}$$

and the $\left(\frac{D}{n}\right)_e$ factors may be computed directly from the totality of data available for establishing the Expected Demerits per unit.

In shorter notation, equation (6) can be expressed as

$$\sigma_{I_e} = k\sqrt{\Sigma\left[\frac{w}{n}\left(\frac{D}{n}\right)_e\right]}. \tag{7}$$

If the number of units inspected is the same for all types of defects, i.e. $n_1 = n_2 = \text{etc.} = n$, equation (7) becomes

$$\sigma_{I_e} = k\sqrt{\frac{1}{n}}\sqrt{\Sigma\left[w\left(\frac{D}{n}\right)_e\right]}. \tag{8}$$

[8] This follows directly from the Law of Small Numbers. Theoretically this result is obtained if p is small, N infinite and pN finite. See any standard text on the subject. Practically this law can be used as an approximation if p is less than .10 and N is greater than 16.

The control limits for the rate are obtained from the equation

$$R_L = R_e \pm K\sigma_{R_e}.$$

(assuming the Normal Law to be a satisfactory approximation for determining probabilities), where σ_{R_e} is given by equations (2) and (6), and where K is a constant whose value depends on the choice of probability.

The following table gives values of K for different values of probability.

Probability	K
.997	3.00
.990	2.33
.955	2.00
.900	1.65
.800	1.28
.683	1.00
.500	.675

Abstracts of Bell System Technical Papers Not Appearing in this Journal

A Note on the Thermionic Work Function of Tungsten.[1] C. DAVISSON and L. H. GERMER. It has been pointed out that the authors in a previous paper [2] had neglected to apply a correction for the "Schottky Effect" to their results. The present note applies this correction which is found to be comparatively small and does not materially affect the results given before. A fuller discussion of the interpretation of the work function measurements previously reported is also included in this note.

The Action of Fluxes in Soft Soldering and a New Class of Fluxes for Soft Soldering.[3] R. S. DEAN and R. V. WILSON. This is a report of basic studies of soldering fluxes undertaken by the authors. It was found that the fluxing action depends on the evolution at soldering temperatures of HCl gas or other halogen acid gases which have been found to be effective soldering fluxes. Based on this discovery soldering fluxes have been found among the organic compounds and the way opened for the development of a truly non-corrosive flux.

Certain Topics in Telegraph Transmission Theory.[4] H. NYQUIST. The author gives results of theoretical studies of telegraph systems which have been made from time to time. Among other things he points out that although the usual method of determining the distortion of telegraph signals is to calculate the transients of the system an alternative method is based on the steady state characteristics. For the first method the telegraph wave is taken as a function of t and for the second as a function of ω. A discussion of the minimum frequency range required for transmission at a given signaling speed is also included in the paper.

Experiments and Observations Concerning the Ionized Regions of the Atmosphere.[5] R. A. HEISING. Experiments are described in which a virtual height of the reflecting ionized region was measured using time lag between radio signals arriving over a direct and the reflected path. Heights were found of 150 to 400 miles with vertical move-

[1] *Phys. Rev.*, Vol. 30, pp. 634–638, Nov. 1927.
[2] Davisson and Germer, *Phys. Rev.*, Vol. 20, 300 (1922).
[3] *Indus. and Eng. Chemistry*, Vol. 19, No. 12, pp. 1312–1314.
[4] Presented, A. I. E. E., February 13–17, 1928. *A. I. E. E. Jour.*, Vol. XLVII, No. 3, pp. 214–216, Mar. 1928.
[5] *Proc. I. R. E.*, Vol. 16, pp. 75–99, Jan. 1928.

ments at rates as high as 20 miles per minute. Other experiments and curves are mentioned which show absorption to be one of the important factors causing poor daylight transmission for wave-lengths around 214 meters. A discussion is given to show that both electromagnetic waves from the sun and β particles must be assumed to produce ionization to explain radio transmission phenomena observed.

The ionization is pictured as beginning at an altitude of about 16 miles and extending upward, and as experiencing diurnal and seasonal variations. The electromagnetic or day ionization occupies a wide region, and is fairly steady except for the diurnal variation. The β particle ionization which is the principal ionization at night occurs continuously. It is, however, less dense than the other ionization and is very variable.

Diffraction of Electrons by a Crystal of Nickel.[6] C. DAVISSON and L. H. GERMER. The scattering of electrons by nickel has been reported on recurrently since 1921. This paper gives the latest results which indicate that a wave-length is in some way connected with the electron's behavior. A rather complete summary of the experiments and conclusions appeared in the *Bell System Technical Journal* for January, 1928, which makes unnecessary a fuller account here.

A General Operational Analysis.[7] W. O. PENNELL. The paper outlines the elements of a general operational analysis. Using p as the symbol for an operator, the author, starting with a general typical defining equation such as $p(ax^b) = \psi(a, x, b)$, proceeds by definitions and theorems to demonstrate the use of operational methods in a large variety of common mathematical operations.

Precision Determination of Frequency.[8] J. W. HORTON and W. A. MARRISON. The relations between frequency and time are such that it is desirable to refer them to a common standard. Reference standards, both of time and of frequency, are characterized by the requirement that their rates shall be so constant that the total number of variations executed in a time of known duration may be taken as a measure of the rate over shorter intervals of time. Frequency standards have the further requirement that the form of their variations and the order of magnitude of their rates shall be suitable for comparison with the waves used in electrical communication.

Two different types of standard which meet these requirements are

[6] *Phys. Rev.*, Vol. 30, No. 6, pp. 705–740, Dec. 1927.
[7] *Jl. of Math. and Phys. of M. I. T.*, Vol. 7, No. 1, pp. 24–38, Nov. 1927.
[8] *Proceedings of the I. R. E.*, Vol. 16, No. 2, Feb. 1928.

described. One consists of a regenerative vacuum-tube circuit, the frequency of which is determined by the mechanical properties of a tuning fork. The other is a regenerative circuit controlled by a piezo-active crystal. Means are provided, in the case of each standard, whereby the recurrent cycles may be counted by a mechanism having the form of a clock, the rate of which is a measure of the frequency of the reference standard.

Data taken over a period of several years with a fork-controlled circuit show that, under normal conditions, its rate may be relied upon to two parts in one million. Data taken over a much shorter time with crystal controlled oscillators indicate that they are about ten times as stable.

Plane Waves of Light. II. Reflection and Refraction.[9] THORNTON C. FRY. This paper extends the study of plane light waves, which was begun in the *Journal of the Optical Society* for September, 1927, to the phenomena of reflection and refraction. It develops the general formulæ for reflection from a plane boundary between two media and for reflection from thin films, paying especial attention to the situations under which "hybrid" polarization occurs. The differences between the reflected and refracted components in the case of dielectrics and metals are illustrated by a number of diagrams.

The paper closes with a discussion of the determination of the optical constants of metals, both from the state of polarization of the reflected light and from the direction of emergence of the ray transmitted through a prism.

Propagation Characteristics of Sound Tubes and Acoustic Filters.[10] W. P. MASON. This paper describes a method for making acoustic propagation measurements, and presents the results of attenuation and velocity measurements of straight tubes and acoustic filters. The ordinary electrical transmission measuring circuit is employed in conjunction with loud speakers and acoustic resistance terminations. The process of measurement consists in obtaining the transmission characteristics of the systems with the device to be measured in the acoustic circuit, then obtaining the characteristics with the device to be measured taken out of the acoustic circuit. The difference between these two measurements gives the transmission characteristics of the device to be measured between the two acoustic resistance terminations. Results obtained from these measurements for straight pipes agree well with the Helmholtz-Kirchoff Law.

[9] *Journal of the Optical Society of America*, Vol. 16, pp. 1–25, January, 1928.
[10] *Physical Review*, pp. 283–295, Vol. 31, No. 2, February, 1928.

Brittleness Tests for Rubber and Gutta-Percha Compounds.[11] G. T. KOHMAN and R. L. PEEK, JR. An insulating material compounded of rubber, gutta-percha, or of similar substances becomes brittle at a temperature, characteristic of the material, below which it may not be used if liable to mechanical stress. This paper describes an apparatus designed to determine this temperature by giving the sample a sharp bend through a fixed angle. The highest temperature at which fracture occurs in this test (the brittle temperature) is found to be nearly independent of the bending angle and the sample's dimensions provided the rate of bending is maintained at a nearly constant (high) rate. A modified form of the apparatus is also described with which the brittle temperature may be determined when the material is under high hydrostatic pressure. The constancy of the brittle temperature when determined under different conditions suggests that it marks a change in the structure of the material.

[11] *Ind. and Eng. Chem.*, Vol. 20, pp. 81–83, January, 1928.

Contributors to this Issue

O. B. BLACKWELL, B.S. in electrical engineering, Massachusetts Institute of Technology. After graduation, he entered the Engineering Department of the American Telephone and Telegraph Company as engineer and in 1919 was made Transmission Development Engineer. Mr. Blackwell has general supervision of transmission developments and by virtue of his position has been prominently associated with progress in long distance wire and radio telephony.

K. W. WATERSON, S.B. in E.E., Massachusetts Institute of Technology, 1898; Mechanical Department, American Bell Telephone Company, 1898; in charge of equipment engineering, 1901; in charge of traffic engineering, 1905; in charge of traffic and equipment engineering, 1906; Assistant Chief Engineer, 1907; Engineer of Traffic, 1909; Executive Officer, Department of Development and Research, 1919; Assistant Chief Engineer, Department of Operation and Engineering, 1920; Assistant Vice President, 1927, in charge of traffic, plant operation and general results divisions, Department of Operation and Engineering.

SALLIE PERO MEAD, A.B., Barnard College, 1913; M.A., Columbia University, 1914; American Telephone and Telegraph Company, Engineering Department, 1915–19; Department of Development and Research, 1919–. Mrs. Mead's work has been of a mathematical character relating to telephone transmission.

OLIVER E. BUCKLEY, B.Sc., Grinnell College, 1909; Ph.D., Cornell University, 1914; Engineering Department, Western Electric Company, 1914–17; U. S. Army Signal Corps, 1917–18; Engineering Department, Western Electric Company (Bell Telephone Laboratories), 1918–. During the war Major Buckley had charge of the research section of the Division of Research and Inspection of the Signal Corps, A. E. F. His early work in the Laboratories was concerned principally with the production and measurement of high vacua and with the development of vacuum tubes. More recently he has directed investigations of magnetic materials and the development of the permalloy-loaded submarine cable.

JOHN R. CARSON, B.S., Princeton, 1907; E.E., 1909; M.S., 1912; Research Department, Westinghouse Electric and Manufacturing Company, 1910–12; instructor of physics and electrical engineering, Princeton, 1912–14; American Telephone and Telegraph Company, Engineering Department, 1914–15; Patent Department, 1916–17; Engineering Department, 1918; Department of Development and Research, 1919–. Mr. Carson's work has been along theoretical lines and he has published several papers on theory of electric circuits and electric wave propagation.

KARL K. DARROW, S.B., University of Chicago, 1911, University of Paris, 1911–12, University of Berlin, 1912; Ph.D. in physics and mathematics, University of Chicago, 1917; Engineering Department, Western Electric Company, 1917–24; Bell Telephone Laboratories, Inc., 1925–. Mr. Darrow has been engaged largely in preparing studies and analyses of published research in various fields of physics.

C. D. HART, M.E., Cornell University, 1906; entered Western Electric Company in Student Course at New York in 1906; transferred to Hawthorne in 1911, development work on the manufacture of lead-covered cable; transferred to Tokyo, Japan, in 1913 to inaugurate the manufacture of lead-covered telephone cable at the Nippon Electric Company; returned to Hawthorne, December, 1915; 1916–20, general foreman of Cable Shops, Metal Finishing Department and Rubber Shops; 1920–23, manufacturing development work; 1923–, Assistant Superintendent of Manufacturing Development.

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 and has been associated with the development of voice-operated switching devices.

H. F. DODGE, S.B., Mass. Inst. Tech., 1916; instructor in electrical engineering, 1916–17; A.M., Columbia University, 1922; Engineering Department of the Western Electric Company and Bell Telephone Laboratories, 1917–. Mr. Dodge was earlier associated with the development of telephone instruments and allied devices, and is now engaged in development work relating to the application of statistical methods to inspection engineering.