The Bell System Technical Journal

Vol. XXI *June, 1942* *No. 1*

The Future of Transoceanic Telephony*

By OLIVER E. BUCKLEY

President, Bell Telephone Laboratories

WHEN Sir William Thomson saw the newly invented telephone of Alexander Graham Bell at the Philadelphia Centennial Exposition in 1876, he stated that "it was the greatest marvel hitherto achieved by the telegraph." Recognizing that the limitations of the first crude instruments would soon be removed, he remarked that "the invention is yet in its infancy and is susceptible of great improvements," and also said "with somewhat more advanced plans and more powerful apparatus, we may confidently expect that Mr. Bell will give us the means of making voice and spoken words audible through the electric wire to an ear hundreds of miles distant." Lord Kelvin lived to see these prophecies rapidly proved true. Had he lived only a few years longer, he would have seen the quality of transmitted speech brought close to perfection, and he would have seen the hundreds of miles extended to thousands.

That Lord Kelvin should have looked upon the telephone as an improvement on the telegraph was natural, for that is the way in which Bell approached it. Bell was experimenting with his harmonic telegraph when he invented the telephone. He was extending the possibilities of the telegraph by making use of a wider band of frequencies than were employed in the systems of Wheatstone and Morse. With its sufficient range of frequencies, Bell's system proved capable of transmitting speech as well as simple signals. Thus, telephony was born from telegraphy by an expansion of the band of frequencies employed in the electrical transmission of intelligence.

More recently, further expansion of the frequency band has been associated both with improvement of quality of transmitted speech and with multiplication of the number of conversations which can be simultaneously transmitted. So far have these developments progressed that today we can transmit speech overland as perfectly as we may desire for any distance we may choose, and we may do so with hundreds of conversations at once over a single coaxial line.

In fact, we have gone further and have so broadened the frequency band

---

* Thirty-third Kelvin Lecture. Read on April 23, 1942 for Dr. Buckley by Vice President Sir A. Stanley Angwin before the Institution of Electrical Engineers.

that vision as well as speech may be carried long distances overland. It does not require a bold stretch of the imagination to predict that some day we shall have multiple-channel transmission of television as today we have multiple-channel transmission of speech. Conductors consisting of hollow pipes offer promise of such application.

Indeed, frequency band width has become almost a commodity to the communication engineer. Telegraphy takes a certain band width as measured in cycles per second for a particular speed of signaling as measured in words per minute. We may assign about 100 cycles to a printing telegraph machine that prints 60 words per minute. To get intelligible telephone communication requires about 1000 cycles, though Bell's original instrument probably used effectively somewhat less than that, which accounts for Kelvin's difficulty in understanding certain words over it. Commercial telephony has benefited from a gradually widened frequency band, until now we look on 3000 cycles as a reasonable engineering standard. This does not provide, however, for perfect transmission of speech. The frequencies in speech commonly cover about 8000 cycles, though frequencies above 4000 contribute little to either intelligibility or quality. Music requires more band width than speech. The range of the normal human ear is about 15,000 cycles, and perfect transmission of music requires that band width. In practice, however, one loses little of esthetic value if the music is limited to 8000 cycles. In fact, most listeners cannot readily distinguish music transmitted over an 8000-cycle band from that over a 15,000-cycle band. Television requires from 20,000 cycles for a recognizable human face, to 20,000,000 or more for vision as well defined as in standard cinema practice, but in present commercial practice the band width is limited to about 3,000,000 cycles.

With a transmission line of given band width, we can allocate its available frequency range to telegraphy, telephony or television as we will. Twenty printing telegraph channels with adequate separation cost in frequencies about as much as one ordinary telephone channel, but for a television channel we must pay the price, in frequencies, of 1000 telephone or 20,000 printing telegraph channels.

It is of the extension of the range of frequencies and their availability for transoceanic communication and particularly transatlantic telephony that I wish to speak. We shall first review what have been the consequences of the extension of frequencies and then explore some of the future possibilities of further expansion of band width in transoceanic communication. Both radio and wire systems must be included to make the picture complete.

The first messages to cross the transatlantic cable were sent at the rate of two words per minute. In frequencies, this means a band width of less than one cycle per second. Lord Kelvin with his mirror galvanometer and later with his siphon recorder and other improvements increased the effective band width and raised the speed to three words per minute. He showed also how to design cables to raise the frequency limits further, with the result that ocean cables were soon made to work at much higher speeds.

Other important advances in terminal apparatus and in methods of operation followed. The application of the duplex principle, permitting simultaneous operation of the cable in the two directions, practically doubled its traffic capacity. Improvements in methods of correcting the signals for the distortion introduced by the cable, and the introduction of mechanical means of sending resulted in a further increase of a third in traffic capacity. Another increase of about the same amount was realized when the cable magnifier was introduced in the early part of the present century.

A major advance ensued with the introduction of the permalloy loaded cable in 1924. The advantages of inductive loading for reducing the attenuation in long circuits had been known for some time and some applications of the Krarup or continuous method of loading had been made to short submarine telephone cables. No practical means, however, of applying this principle to ocean cables was available, since for telegraph frequencies loading with iron wire was not advantageous because of its low permeability. The discovery of permalloy, a material with very high permeability at low flux densities, together with the invention of means for protecting the loading material from the severe stresses that it would otherwise encounter at the ocean bottom, made it possible to build a cable with many times the band width of corresponding non-loaded cables. The increase in traffic over the cable was, however, less than proportional to the increase in frequency range, because duplexing the loaded cable involves a greater sacrifice of one-way speed than is the case for the non-loaded cable. The fastest loaded transatlantic cable has an effective frequency band of over 100 cycles per second and can carry four times as much traffic as a non-loaded cable of the same size and length.

Development of permalloy loading for telegraph cables naturally led to consideration of the possibilities of a loaded telephone cable to span the Atlantic. Whether viewed as an extension of frequency from 100 cycles to the 3000 cycles needed for high grade telephony, or as an extension of distance, the step was a formidable one. The longest deep-sea telephone cable reached only 105 nautical miles from Key West to Havana, where three cables continuously loaded with iron wire were laid in 1921. The transatlantic span called for a minimum of 1350 miles via the Azores, or 1800 miles by the more direct route from Newfoundland to Ireland. It was obvious that such a step could not be accomplished by mere structural changes. New materials were required. For years, a systematic search had been made to improve the properties of electric and dielectric materials for use in submarine cables. By 1928 sufficient progress had been made in the development of materials and in the structure of the cable itself to permit seriously undertaking a transatlantic telephone cable. A decision was than made to embark on a test of a section of such a cable under practical conditions.

In determining the requirements for this cable, it was decided to engineer it for the Newfoundland-Ireland route rather than for the route via the Azores. The longer link made the cable more difficult and its cost per mile higher, but the total cost and considerations of operation and maintenance favored the more direct route.

The structure proposed for the Newfoundland-Ireland telephone cable was of the single-core type with a continuously loaded central conductor and a concentric return conductor similar to that of the Key West-Havana cables, but different in materials and dimensions. Instead of a serving of iron wire or permalloy to provide inductance, there were used four layers of very thin perminvar tape. Perminvar is an alloy which, in the form of loading tape, has a permeability and resistivity suitable for telephone use and at the same time has very low hysteresis, which helps in preventing distortion of speech due to magnetic modulation. The loaded conductor was insulated with paragutta rather than with gutta percha. Paragutta is a mixture of specially purified and deproteinized rubber, deresinated balata or gutta percha, and some wax. It has a dielectric constant 15 per cent lower than the gutta percha in the Key West-Havana cables, and leakance at telephone frequencies about one-fifteenth as great.

The cable was designed on the basis of as high attenuation as would be permitted by considerations of noise at the receiving end and usable power at the sending end. Since the attenuation of such a cable increases rapidly with frequency, only the noise at the high-frequency end of the speech band is significant; here the noise is entirely of thermal origin, for static and other external interference are eliminated by shielding. The sending power is limited by magnetic hysteresis and there is little advantage in applying more than about 50 volts. Most of the power can be concentrated in the high frequencies by placing, at the sending end, part of the network which corrects the distortion of the cable. By these means, it is possible to set a permissible overall attenuation as high as 165 db for a top frequency of 3,000 cycles. This far exceeds attenuation permissible in other wire telephone practice. The cable was designed to give this attenuation with the most economical disposition of materials within practical limits. Its core comprised a loaded central conductor of 800 pounds of copper and 95 pounds of perminvar per nautical mile, 720 pounds of paragutta insulation, and a concentric copper return conductor of 1700 pounds, making the cable much heavier than any that had previously been laid in great depths.

A 20-mile section of this cable was made in 1930 under the supervision of Bell Telephone Laboratories engineers by the Norddeutsche Seekabelwerke in Nordenham, Germany. There it was loaded aboard the cable ship Norderney and taken to a location in the Bay of Biscay where a depth of 2,500 fathoms was conveniently available. This depth was greater than would be encountered on the proposed cable route. The 20-mile section was paid out on the sea floor and its open-end impedance measured over the telephone range of frequencies. From these measurements, changes in its electrical parameters could readily be deduced. The cable was then pulled in and carried to Frenchport Harbor near Belmullet, County Mayo, Ireland, whence it was laid out to sea to permit measurement of terminal noise. Measurements of impedance both from the ship and from the shore showed the cable to be quite unimpaired both at 2½ miles depth and after recovery and relaying in shallow water. Measurements of noise from shore, however, showed that the location was unsuitable for this type of cable because of the rocky bottom. Such a cable is somewhat microphonic on account of the strain sensitivity of perminvar and the terminal sections of the cable must lie quietly on a soft bottom if a low terminal noise level is to be assured.

Simultaneously with the experiments on the cable, experiments were conducted with the terminal apparatus in the laboratory in New York. An artificial line had been built closely simulating the proposed cable in electrical performance, and over this speech was transmitted at the levels intended for the cable. The method of operation was extreme as well as novel. The high attenuation made it impossible to balance the simulated cable for two-way talking, and voice switching had to be used. At both ends the receiving apparatus was normally connected to the line. The speaker's voice currents caused his end of the line to be switched to the transmitting apparatus. Arrangements were devised to avoid loss of speech during the switching interval, and to minimize interference due to the persons at the two ends of the cable speaking almost simultaneously. The time required for speech to travel over such a cable is not negligible. In this case it was about a tenth of a second. This is long enough to be noticed but not serious enough to count as a major disadvantage.

All of the measurements in the laboratory, at sea, and from shore joined in giving assurance of the technical soundness of the proposal to install a cable of this type. Its performance would have been superior to that afforded by radio. The cost, however, would have been much greater than that for a radio circuit. The cable system from Nova Scotia via Newfoundland and Ireland to Great Britain promised to cost about $15,000,000. When the project was first considered, the radio connection had been subject to frequent interruption and the cable was regarded as an economically justified supplement to the radio services as they then were.

Postponed temporarily because of general business depression, the cable project was later postponed indefinitely because, in the face of improvements in transatlantic radio communication, so expensive a cable to carry a single conversation could no longer be justified. Today it seems improbable that such a cable will ever be laid across the Atlantic. Fortunately, other cable possibilities have in the meantime been developed which look more attractive. Before going into these, however, let us review the development of transatlantic radio telephony and estimate some of its future possibilities.

The development of radio communication, even more strikingly than that of wire communication, has been characterized by widening of its spectrum. In fact, starting with Marconi's low-frequency transatlantic experiments of 1901, the spectrum has widened until today it provides some thousands of megacycles. Only a small portion of this range is, however, available for transoceanic communication. There are utilized only two isolated ranges each of which, by comparison with the total radio band, is comparatively narrow. The low-frequency or long-wave range is a band some tens of kilocycles wide with a top of about 100 kilocycles. This low-frequency portion of the spectrum was intensely cultivated during the first two decades of the present century and by the close of that period had become rather densely populated with radio transmissions.

In the 1920's the band of frequencies useful for long distances was widened several hundredfold by the discovery that long-distance transmission could be carried on by short waves, that is, by frequencies in the range 3 to 30 megacycles. This discovery put transoceanic radio communication on its present world-wide basis. Short waves not only contributed greatly to the communication band width but contributed as well as to the demand for service by reducing costs, since the apparatus required for short-wave circuits proved to be less expensive than that for long waves.

The transatlantic telephone like its telegraph predecessor started in the relatively cramped long-wave band and then moved into the freer region of the short-wave range. It was in 1915, 14 years after Marconi had spanned the Atlantic by radio telegraph, that speech was first sent across the oceans from Arlington, Virginia to Paris and to Honolulu. This achievement, somewhat beclouded by the events of the first World War, was the result of a plan to talk across the ocean which was definitely undertaken by Bell System engineers after they had successfully established wire telephone communication across the North American continent. For its accomplishment there were evolved the first high-power vacuum tubes and the first master-oscillator, power-amplifier tube transmitter. This experiment was thus, in a technical sense, the forerunner of modern radio telephony, including broadcasting as well as transoceanic telephony.

It required much study of radio transmission and many further improvements in apparatus technique before speech could be projected across the Atlantic with sufficient clarity and reliability to be truly serviceable. In the long-wave range and for transatlantic distances, radio is seriously limited in two respects: first, the level of noise is high, particularly in summer, due to the frequency of occurrence of thunderstorms in northern latitudes; and second, the received signals become weak during the sunset and sunrise periods. The ionized layers of the upper atmosphere upon which long waves depend for their guidance around the earth are then going through the transition from daylight to nighttime condition. The development of the water-cooled high-power vacuum tube made possible high-power amplifiers to deliver tens of kilowatts needed to lift the signal higher above the level of atmospheric noise. The influence of static was further reduced by the use of directive receiving antennas. Additional improvement was provided by the technique of single-sideband transmission first used on wires. These developments and others assured fairly reliable telephone connections and in 1927 public service was opened jointly by the General Post Office and the American Telephone and Telegraph Company. The carrier frequency was 60 kc corresponding to a wave-length of 5000 meters.

The opening of the first transatlantic short-wave telephone circuit in 1928 followed close on the heels of the long-wave circuit, and was followed in turn by the establishment of additional short-wave circuits in 1929. These were years of increasing disturbance in short-wave transmission and about 1929 preparations were started on both sides of the Atlantic for a second long-wave channel to diversify facilities and thus improve the continuity of the service. This work had not progressed very far, however, before there came in sight opportunities for greatly improving short-wave transmission. Accordingly, the project of the second long-wave circuit was deferred and, upon the actual realization of the short-wave improvements years later, was postponed indefinitely. More recently, experiments have been conducted which have demonstrated the feasibility of transmitting two channels at different frequencies using the same transmitting equipment. This may lead ultimately to more economical provision of an additional long-wave circuit.

Short-wave radio has certain outstanding advantages over long waves for transoceanic service; these are less attenuation, lower noise and a wider frequency band in which to operate. The first two factors together with the ability to obtain readily a high degree of antenna directivity, result in considerable economies. In the case of a radio telephone connection between the United States and England, the cost of short waves under present practice for an approximately comparable quality of service is about one-half that of long waves. The wider band provides an opportunity for the service to expand, for although the short-wave band is by no means unlimited in extent and within the last few years has become increasingly congested, much can be accomplished by careful and coordinated planning for use, and in any case it is hundreds of times wider than the long-wave band. These advantages have been reflected in the rapid growth of short-wave transoceanic telephony, as indicated by the fact that by the beginning of 1939 there were in service throughout the world about 170 important long-distance short-wave telephone circuits, of which five were in regular use between the United States and Europe. There has grown up also a host of short-wave broadcasting channels, the better coordination of which has yet to be worked out.

Certain disadvantages of short-wave transmission must nevertheless be reckoned with. The greatest by far is its susceptibility to complete or partial interruption at certain times, particularly around the maxima of the 11-year sunspot cycle. This weakness of short-wave transmission is fresh in the minds of many of us because we are only now emerging from one of these maxima. Short waves are also affected adversely by various types of signal distortion which arise from the circumstance that the signal picked up at a receiving site is usually made up of several components which have traveled over different paths. Sometimes these paths all lie along the same great circle but involve different numbers of reflections between the earth and the Heaviside layer. Sometimes signal components arrive over other than the great-circle path. Occasionally components travel along the longer of the two great-circle paths between the transmitter and receiver or even clear around the world, producing a distinctive phenomenon known as "round-the-world-echo." Interference between waves arriving over different paths results at times in "general fading" caused by variations in the level of the whole band and at other times in "selective fading" in which portions of the speech band are affected differently.

These phenomena and their causes have been widely and intensively studied ever since the advent of short waves and as a result much progress has been made in improving short-wave telephone transmission. Single-sideband transmission has been helpful in eliminating a particularly disagreeable type of mutilation common to double-sideband transmission wherein the fading out from time to time of the carrier signal gives rise to a harsh, grating character of received speech. The multiple-unit steerable antenna recently developed in Bell Telephone Laboratories, and known familiarly as "musa," has been found useful in reducing speech distortion accompanying wave-interference effects. The musa reduces selective fading by combining signals arriving over different paths or by eliminating all signals except those arriving over one path. By providing additional antenna directivity it makes possible operation in periods of reduced signal strength, though it does not eliminate circuit interruptions at times of very severe disturbance. Single-sideband transmission and the musa are now in regular use on the New York-London telephone circuits.

The combination of modern short and long-wave technique now provides across the Atlantic a telephone service which is acceptable for general commercial use, though admittedly of somewhat variable quality. Complete interruption of short-wave service with inadequate long-wave service to fall back on, remains the most serious limitation. Neither the reliability nor the quality of transatlantic telephony is yet up to the standards of a well-constructed and well-maintained wire line.

As matters stand today the short-wave bands, carefully used, could be made to handle a very substantially increased load. The long-wave band is much more restricted yet it is on the long waves that we now depend when short waves fail, as they do at times. Marked improvement in reliability of present systems, or some altogether new and independent system, is needed to provide an alternative group of circuits to insure availability of service as users come to rely on it in their business and social intercourse. This is one of the important factors that led to the development of the transatlantic telephone cable.

To predict the future development of transoceanic telephony is presumptuous, to say the least. So rapid has been the advance in the art of communication and so revolutionary have been the discoveries in this field that one is quite unwarranted in setting any limits to the progress that may be achieved. However, there are some developments that have progressed far enough in the laboratory to discuss with reference to their early application; also there are pertinent indications as to the future of transoceanic telephony apparent from consideration of developments which have occurred in long-distance overland telephony.

Perhaps the most significant recent development in land-line telephony is that of broad-band transmission over open wires, cables and coaxial conductors. Broad-band transmission means the treanmission by carrier methods of a considerable group of telephone bands on closely spaced channels. Over open-wire lines and over pairs in lead-covered cables, 12 telephone bands spaced at 4000-cycle intervals are commonly transmitted in a group occupying a total band width of 48,000 cycles. With coaxial conductors, the band has been increased to 2,000,000 cycles giving frequency space for some 500 telephone channels and it may be expanded still further when more channels are required.

The application of broad-band methods to transoceanic radio telephony may be anticipated with some confidence. To achieve it requires broad-band amplifying systems capable of delivering high power without distortion. Commercial success has already been achieved with small numbers of channels in the Holland-East Indies and the United States-England single-sideband systems. More recently, by applying the principles of negative feedback, Bell Laboratories engineers have developed a short-wave transmitting amplifier of 200 kw capable of handling 12 or more closely spaced telephone channels.

One might visualize the broad-band transatlantic radio telephone system of the future as being built up of successive groups of these 12-channel blocks. The number of groups that might be used simultaneously is, of course, limited. Over any path where radio transmission depends upon reflections between the ionosphere and the earth, Nature sets a rather definite limit on the range of frequencies that is usable at any given time. In effect, there is provided a transmission path between transmitter and receiver which is capable of passing a broad but nevertheless limited band of frequencies. Frequencies above this band are not consistently returned to earth from the ionized regions. Frequencies below this range are absorbed. The high-frequency end is marked by a sharp cutoff, while there is a more gradual diminution of effectiveness at the low-frequency end. The position in the spectrum of the useful band shifts with time of day, season of the year and phase of the solar cycle. Its width varies, too, being narrow at night and wider during the day.

Thus, for example, when the sun is over the mid-Atlantic in summer there is available a useful band of frequencies about 4 megacycles wide, extending from about 14 to 18 megacycles. It is not sharply defined on its lower side, and its position in the spectrum varies with the season of the year and the sunspot cycle. But we may say roughly that nature provides at any one time, at least during the most useful hours of the day, a band width of the order of 4 megacycles. If this entire range could be utilized for telephony over this particular path, and were subdivided sharply into telephone bands of 4 kilocycles width, there could be realized 1,000 telephone channels. These might be used in any of the several ways, as to two-way transmission and as to the points at which they terminate.

But public service transatlantic telephony is not the only service requiring these important short waves. There are many other uses of them such as radio telegraphy, ship-to-shore telephony, airplane communication and navigation, and overseas broadcasting. There are also other natural barriers than the Atlantic to be bridged in this manner, and these short waves because of their world-wide effect and despite the directivity that can be imparted to them, cannot be counted upon to be duplicated very often for simultaneous use at different locations throughout the world. So we must allow for the available 4 megacycles to be divided to meet a large number of requirements, perhaps none more important than the Atlantic route, yet collectively of great consequence. Let us say that, in view of all the other requirements, public service telephony across the Atlantic deserves something like a tenth of the total facilities in this band. This would mean an allotment of 400 kilocycles or 100 one-way telephone channels, yielding 50 or more two-way circuits realizable under the natural limitations of the medium and the other requirements placed upon it.

Of course, the demand for such a number of transatlantic telephone circuits will depend in large measure upon the economy with which they can be realized, but the estimate serves at least the purpose of pointing out that short waves can provide physical facilities for a volume of telephone communication far beyond that now obtaining. Surely we can anticipate with confidence a great growth of transatlantic telephone traffic, but in proportion as the demands for service grow and we come closer to the realization of the ultimate physical possibilities, the more serious becomes the threat of interruption to this service by magnetic storms.

These conclusions lead us to reconsideration of the transatlantic telephone cable as an auxiliary to short-wave systems. It is readily apparent now, however, that a single-channel cable such as we projected in 1929 would be of little value in supplementing a radio telephone service of so many channels as there may be in the future. To be of any real value in this situation, the cable also must be capable of carrying a considerable group of telephone channels. It was toward such a possibility that we turned when the project of a single-channel cable was suspended. We have made considerable progress in that direction, and I would like to tell you about it, if you will excuse my presenting a proposal which has still many elements of speculation in it.

It was obvious at the start that a multi-channel telephone cable to cross the ocean would have to be provided with intermediate repeaters since even a single-channel cable without repeaters required going to practical extremes in structural design. Consideration of mechanical difficulties ruled out locating the repeaters elsewhere than on the ocean bottom. Problems of laying and lifting made it obvious that the repeater housing should, if possible, be incorporated within the cable structure and treated as a part of the cable rather than as an appendage to it. Hence we were led to develop a small-diameter cylindrical housing to be incorporated as a part of the cable underneath its armor. The whole structure had to be flexible so that it could be bent around a cable drum and passed over the bow or stern sheave of a cable ship.

The structure of the repeater housing which was devised comprises first a succession of pressure-resisting steel rings each having a diameter of about $1\frac{1}{2}$ inches and a width of $\frac{3}{4}$ inch. Over these is slid a succession of thinner steel rings of the same width but so placed as to overlap the joints of the inner rings. So assembled, the rings form an articulated cylinder about seven feet long. To exclude water, there is placed over this cylinder an annealed copper tube with water-tight seals at its ends. The details of the seal are of the greatest importance. It combines a strictly hermetic seal, in which the conductors are brought out through glass, with a plastic seal through which diffusion of water vapor would be extremely slow, should the glass seal fail. Joined to the copper cylinder, and extending over the cable core for several feet, is a tapered copper sheath which serves to distribute bending strain and protect the conductor joint at the seal. Containers of the type described have been tested at pressures considerably higher than would be encountered in a transatlantic cable. They have also been subjected to repeated bending around a six-foot drum without failure.

Within the repeater housing the elements of the repeater are separately contained in plastic cylinders about six inches long, loosely fitting inside the inner steel rings. Connections between these units are made with flexible conductors.

A repeater must, of course, be supplied with power and, as it is impracticable to provide a primary source of power in such a small housing, power must be fed to the repeater over the cable from a direct-current supply. The supply voltage is one of the limiting considerations in the design of such a cable system. It must not be so high as to endanger the insulation of the cable or repeater elements. An operating potential-to-ground of 2000 volts oppositely poled at the ends of the cable was assumed. Power would be supplied on a constant current basis so that fluctuations of earth-potential would not cause variations of current-supply. The repeater elements were designed to withstand the anticipated voltage-to-earth. Tests of cable-core and joints over a long period of time have shown no observable change under this impressed voltage.

The difficulties of lifting a deep-sea cable for repairs are such as practically to prohibit frequent access to the repeaters for maintenance. Hence, the repeater must be provided with elements which will rarely, if ever, require attention. A period of 20 years without replacement of parts was assumed as a reasonable requirement.

The problem of life and maintenance is principally the problem of a rugged long-lived vacuum tube. Ordinary vacuum tubes have limited service-life on account of evaporation of material from thermionic cathodes. By making the level of transmitted signals relatively low, the space current may be kept very small. By making the cathode surface relatively large, this small current can be obtained at a temperature so low that the cathodes of the tubes may be expected to last for a very long time. This is a different approach to the tube problem than has ordinarily been made. New types of tubes based on these principles were developed and put on life tests more than five years ago. As yet they have shown no evidence of deterioration, and one now may be reasonably sure from their behavior and from physical considerations of a life of at least ten years. There is good reason to think that they should last several times that long, but further observation will be required before a life of as much as 20 years steady operation can be confidently predicted. The tubes must also be more rugged than ordinary vacuum tubes since the cable will be subjected to considerable vibration and perhaps to heavy blows in the course of laying and lifting, though the tubes can be protected to some degree by resilient mountings.

Other elements of the repeater structure such as coils and condensers are also subject to special requirements both electrical and mechanical. These requirements have been met in a preliminary way and the assembled repeater in its housing subjected to mechanical tests in the laboratory.

Although the electrical requirements of such a cable are very severe, there are some respects in which the submarine telephone repeater is simpler than a land-line repeater. The temperature at the bottom of the ocean is nearly constant; consequently, the repeater does not have to be regulated to compensate changes of cable characteristics with temperature. Also, once the cable is laid, it is in a very quiet place, and except in shallow water near shore is not likely to be disturbed. True, the electrical characteristics of the cable may show effects of aging, but over a long period of time changes are not great, and they can be allowed for by providing some margin in the electrical design.

In the circuit of the repeater the heating filaments of the amplifying tubes are placed in series with the central cable conductor. The fall of potential through the heater filaments provides the plate potential for the tubes. Appropriate networks compensate for variation of cable attenuation with frequency. A negative feedback circuit gives a high degree of stability over a wide band of frequencies and minimizes the effect of variations of tube characteristics. It is interesting to note that the amplification provided by a single tube could drop to a tenth its normal value with scarcely appreciable effect on the performance of the repeater.

The number and spacing of repeaters depends of course on the length and design of the cable. For a cable 2000 miles long to connect Newfoundland and Great Britain there was calculated a core comprising 516 pounds of copper per mile insulated with 370 pounds of paragutta, surrounded by a return conductor of 600 pounds. This is like the core of the 1930 Key West-Havana telephone cable but somewhat smaller. Paragutta was assumed as the insulating material because of extensive experience with it. By using for the calculation the characteristics of one of the newer synthetic insulating materials a somewhat more favorable design would have been obtained. On this cable 47 repeaters spaced 42 miles apart would provide for the transmission of a band 48,000 cycles wide.

The repeater is a one-way device and to provide two-way conversations two cables have been assumed, one directed eastward and the other westward. This is the simplest solution of the two-way problem but it is not inconceivable that the problem could be solved with a single cable. Using two cables, each transmitting 48,000 cycles, the number of telephone circuits will depend on the band assigned per channel. If we adhere to the present best land-line practice, and assign 4000 cycles per channel there would be room for 12 telephone circuits. For a small sacrifice of quality the number could be materially increased. Even as many as 24 fairly satisfactory circuits could be provided by assigning only 2000 cycles per channel.

Although in Bell Laboratories we have gone a considerable distance in the design of a broad-band repeatered submarine telephone cable, and have developed many of the essential parts, I would not wish to give the impression that all the problems of such a cable have been solved, or that the time has come to proceed with its construction and installation. Indeed, it is only by building trial sections of such a cable and subjecting them to repeated punishment more severe than a cable is likely to encounter, that the problems can be fully recognized. Extensive electrical tests will also have to be made on a complete assembly of repeaters with artificial lines simulating sections of cable. These steps have yet to be taken.

A submarine cable requires a degree of care and precaution in engineering such as is required in few other situations. It is usually not possible to provide large factors of safety, and yet failure of a single part such as a break in the conductor or a leak in the insulation completely destroys the operation of the whole system. Experiences of over eighty years since the failure of the first attempt at an Atlantic cable have led to the development of practices which give good assurance of the reliability of cables of simple construction, but when a device such as the proposed repeater is made a part of the structure a new set of hazards is introduced. Whether these hazards can be guarded against well enough to justify the risks of such a cable project remains to be seen, but I am optimistic that by a sufficiently thorough job of cable manufacture and a well planned program of trials, the hazards can be reduced to an acceptable degree. It will take some years to reach this point, and at best it must be expected that some degree of hazard will still remain. Submarine cables, like all things that go to sea, can never be completely dissociated from some chance of disaster.

As to the costs of such a cable project for establishment of broad-band wire telephony to England via Newfoundland, only the roughest sort of estimates can be made at this time. However, even applying annual charges somewhat higher than have commonly been used for cables, it appears that the total cost per telephone circuit for the system of two cables with associated equipment will be comparable with that of prospective short-wave radio systems. A considerable increment of cost of cable over that of radio would be justified by the better quality of transmitted speech and the very significant advantage of privacy. Added to this is the value of the cable as a supplement to radio systems to provide against their failure. Indeed it is possible that once the cable were in service radio would be looked on as a supplement to it.

The comparison of cable and radio telephony is not easy to make. It is the composite of cable and radio that assures continuity of service since while radio is sensitive to disturbances accompanying magnetic storms, and cable less so, radio service is not so exposed to the possibility of interruption by mechanical accident or malicious intent. An advantage of radio systems is their flexibility, whereby new routes can be established or old routes abandoned without incurring excessive costs. Further, provision can be made for expansion of radio facilities as required without having to install so large a complement of circuits at the outset. The prospect of the combined radio and cable system is a happy one in that it affords the advantage of both types of facility.

In the foregoing discussion I have treated the transoceanic telephone problem principally as the transatlantic problem and more particularly as the problem of connecting North America and Great Britain. Community of language and many interests lend particular emphasis to that connection, but it is, after all, only one of the many transoceanic links required to build the world-wide telephone network of the future.

When we come to look at other situations, the relative advantages of radio and cable weigh differently. Shortwave radio links have a great advantage in affording direct connection between points on the globe far apart, and the tendency has been to establish short-wave connections directly between large centers rather than through extensive land-line links, particularly where political boundaries have been involved. There has thus grown up an extensive network of single-channel short-wave radio connections operated at low power, giving good service part of the time but not to be depended on all of the time. Most of these connections are over routes which would not support broad-band systems such as I have discussed. The introduction of broad-band methods for transoceanic radio telephony will tend to favor centralizing radio traffic at a smaller number of more important radio terminals, but it is hardly to be expected that all transoceanic radio traffic will thus be concentrated. Even with radio systems dispersed rather than centralized, broad-band cable may still serve as an effective supplement to radio not only between North America and Great Britain but also between North America and all of Europe, with land lines extending the circuits to all important centers of the European continent. It is to be expected, too, that the cable will find important application in other locations than across the north Atlantic. Notably, this type of cable is particularly promising for trans-Mediterranean service. Indeed, the same principles of construction which are proposed for the transatlantic cable may be applied over much shorter distances. With some modification of design, the repeater can be incorporated in lead-covered cables for shallow seas and afford transmission advantages of carrier as well as the economy of broad-band.

If one tries to imagine the world-wide transoceanic network of the future, he may well envisage a net comprising a large number of light linkages plus a small number of heavy linkages over the most important routes. The light linkages will represent direct short-wave single-channel or twin-channel connections using relatively small power. The heavy linkages will comprise highly developed powerful broad-band short-wave radio systems making full use of frequency and directional diversity supplemented by broad-band submarine cables and in a few cases by long-wave radio as well.

From purely physical considerations, it appears feasible to provide all of the facilities for telephone connection between all points on the earth that its inhabitants are likely soon to require. To what extent these facilities will actually be developed will depend on demand and that, to a considerable extent, on cost. It will be interesting to survey briefly this question of prospective demand to see whether after all it promises to be great enough to justify the installation of broad-band cable and radio systems such as are here proposed.

There are so many factors that contribute to telephone demand that it is impossible to make any very reliable estimate. In addition to cost, there are factors of differences in time, in language and in telephone habits and also the factors of community of interest and speed of service. Similar factors affect the demand for telegraph service, but the transatlantic telegraph habit has had more time to mature fully and may reflect more accurately than the telephone the demand that exists for rapid communication between Europe and America.

One possible way to estimate what the future may have in store for transatlantic telephony is to compare the flow of telegraph traffic, say between London and New York, with that between New York and some west-coast American city, and then to examine how intensively telephone service has been developed relative to telegraphy over the two routes. Because of its comparative stability over a period of years, New York-

San Francisco traffic provides an interesting basis of comparison. The distance and difference in standard time between these two cities compare fairly well with those between New York and London. Difference in community of interest is compensated to some degree by the difference in size of London and San Francisco.

This comparison may be made on two bases not very different in character, but leading to widely different results. In the first, let us compare the two routes as regards telegraph traffic, using as our measure the total number of words transmitted in a single years. In the second, let us use as our measure the number of public service telegraph messages, excluding such telegraph business as is comprised under the headings of press service, leased-wire service and code and cipher messages. In each case the estimate is based on terminating messages and excludes traffic routed via the cities named. Data for the year 1937 are available and this particular year has some further advantage in that it represents something between the peak of the 1929 era and the trough of the succeeding depression.

On the first basis of comparison we find that the total number of telegraph words transmitted between New York and San Francisco in 1937 was approximately the same as that between New York and London. On the second basis we find that the number of telegraph messages was about seven times as great between New York and London as between New York and San Francisco. The wide discrepancy between the two comparisons is doubtless accounted for partly by rates and partly by the character of business and social intercourse. Of the two the second, which is based on plain-word public-service messages, would seem to be more significant in relation to potential demand for telephone service. The information transmitted in press and coded telegraph messages and over leased wires is presumably business of record. Public message telegraphy, as a somewhat closer approximation to the informal exchange of ideas by telephone, may be a better index of telephone demand.

On the basis of these figures we may speculate that the potential demand for telephone connection between New York and London is somewhere between one and seven times that between New York and San Francisco. Actually, in the year 1937 the telephone traffic between New York and San Francisco was about three times that between New York and London. Thus it would appear that not more than a third, and possibly not more than a twentieth, of the potential telephone demand has been realized.

If we assume, as seems reasonable, that the same ratio of potential to realized demand exists for all European-North American connections as for the New York-London connection, we may estimate that in place of the five pre-war telephone circuits across the North Atlantic there will be needed from fifteen to one hundred circuits. Which of these figures proves to be the better measure will doubtless depend greatly on costs that can be achieved but I do not think that I can fairly be accused of excessive optimism in predicting a demand for forty or more telephone circuits in the reasonably near future if full advantage is taken of technical possibilities already in view to decrease costs and improve reliability of service.

In estimating the demand for growth it may be a mistake to attach too much importance to cost of service. Speed and reliability are, within limits, just as important. When it becomes possible to pick up a telephone and get a reply within two minutes, which is about the normal time for a long distance connection in the United States, and when the connection provides the clarity and freedom from noise of a local telephone call, then the transoceanic telephone service will, I believe, be used to a degree not even approached at present.

To provide this indicated increase in number of circuits, and to approach land-line standards of reliability and quality of service will demand utilization of all three types of transmission systems: short-wave, long-wave and repeatered cable. Considerations of cost, flexibility and directness of connection suggest that the bulk of the transatlantic business will be handled on the short waves, but any service important enough to justify so large a group of circuits as has been estimated would have to live up to a higher standard of reliability than short-wave circuits alone can provide. A cable between America and Britain would provide this reliability, acting as insurance against serious interruptions of service that would result from a simultaneous failure of all the short-wave facilities during periods of magnetic storm. It would in addition set a high standard of transmission performance in competition with short waves. The cable and short-wave circuit groups plus a few long-wave circuits should provide a high degree of reliability and excellent transmission at a level of cost such as would assure the continued growth of the service.

It may not be necessary to wait until the growth of transatlantic telephone business provides enough traffic to utilize fully a cable of the type described. When once the engineers are ready to give reasonable assurance of the cable, I believe that it will not have to await complete economic justification, because of the tremendous importance which it would have in insuring privacy and continuity of transatlantic telephone service. What the cable really waits on is technical development. To achieve this is fairly straightforward, since there do not appear to be any insurmountable difficulties. There is still much to be done and many difficulties must be overcome before the broad-band repeatered cable can be installed but it does not impress me as a more difficult problem than many that have been solved in the past.

In developing this picture of transoceanic telephony I have endeavored to stay within the realm of engineering fact, and not to count on products yet to be born from the inventor's fancy. Indeed, I may have been overconservative for there are already partly developed inventions which might greatly modify the picture. One such is the vocoder, an instrument which, in a sense, compresses speech into a narrow band. More accurately it dissects speech, transmits it in code and recreates it at the other end of the line. With vocoders a hundred or more simultaneous conversations might be carried by a pair of repeatered cables. While the vocoder would transmit the primary elements of conversation it would not provide all of those qualities of speech which words alone do not convey. The vocoder gains in band width at the cost of naturalness of speech, but even so, it may find important application.

Other inventions may extend the band width available for transoceanic communication far beyond the range here discussed. Projects such as repeatered ultra-short wave radio systems and undersea wave-guides, which today appear fantastic, may some day come within the range of practicability.

The electrical channels over which peoples of one continent hold their more urgent communication with those of another have always been of surpassing technical interest. Ever since the first electrical impulses to carry words across the ocean were traced in the wavering beam of Kelvin's mirror galvanometer, the improvement of these channels has been a fruitful field for scientist and inventor. But these paths for the transmission of intelligence have a wider significance than mere technical achievement. They are strands of an ever-growing bond that unites widely separated continents. The newest of these strands, the overseas telephone, has yet to reach its maturity. Not until conversations can be carried on as easily and reliably between continents as between cities within a continent, can we claim that the art of transoceanic telephony has come of age. When this time arrives, we shall probably realize as we look back that the half-dozen telephone circuits of the 1930's formed indeed a slender thread to bind together in speech the people of North America and those of Europe. Some tens of kilocycles of band width may then appear as inadequate as the slowly dispatched words over the first transatlantic cable appear to us today.

Diamond Dies for the High-Speed Drawing of Copper Wire*

By **H. N. PADOWICZ**

ESSENTIAL to the drawing of copper wire at any speed are the dies to effect the desired reduction steps. It can be readily surmised that this item is one of major importance at drawing speeds of 10,000 and 12,000 feet per minute which are being used in the copper wire drawing plants at the Kearny Works and the Baltimore Works of the Western Electric Company.[1] In these machines diamond dies are used to draft 12 and 14 A.W.G. supply wire to the final sizes of 19, 22, 24 and 26 A.W.G., respectively. As pointed out in the paper[1] describing these machines in the above noted plants, the maximum possible drawing speed is limited by the stresses set up in the take-up reel rims. Drawing dies in themselves should not place any limitations on the wire drawing speeds if the factors of heat generated and the rapid movement of lineal wire surfaces are logically considered and provided for.

In 1924 the manufacture of copper wire at a drawing speed of 2500 feet per minute in a new type of wire drawing machinery, developed and designed by the Western Electric Company, was started in its Chicago plant.[2] At that time copper wire was being generally produced at speeds ranging from 800 to 1200 feet per minute. A study of the manufacture of diamond dies for use in these machines developed that dies suitable for this work required a differently shaped "approach," a better polish and a shorter "land" than those that were available for low-speed operation. These same factors are still the important items which must be considered for today's drawing speed of 10,000 and 12,000 feet per minute.

The technique of making diamond dies for the drawing machines which operate up to about 5,000 feet per minute is now well established. The specifications covering the dies for this purpose are known and available to the trade which is well qualified to produce them. However, the opportunity and necessity for diamond dies to draw copper wire at the noted high speeds did not exist prior to the recent development of these high-speed machines and for this reason the industry was not familiar with the necessary die requirements.

The Kearny Wire Mill was set up to purchase finished mounted dies from outside suppliers and to recut them when oversize. These dies are usually acquired at the smaller final sizes as governed by production schedule requirements and enlarged to the larger sizes for both finisher and line die use.

---

* Published in *Wire and Wire Products*, October, 1941.

A large part of the success attained was due to specifications defining the die requirements and the setting up of adequate inspection means to check the new and recut dies for these requirements. Prior experience of this company and the cooperation of our suppliers greatly aided this policy.

The following details of diamond wire drawing dies have been found to be essential to the desired performance of these tools in our drawing operations:

- Shape and Quality of the Diamond Stone
- Stone Size
- Mounting
- Contour of the Die Hole
- Polish and Finish
- Pounds Pull to Draw the Wire through the Die

SHAPE AND QUALITY OF DIAMOND STONES

Round, clear stones of sound structure are selected, and these should be free from cracks, pits, carbon spots, porosity or any flaws and imperfections which would affect the present or recut drawing surfaces of the die. *Round* is the trade term for stones of polyhedral sides, approximating a spherical form as distinguished from flats and irregular, unsymmetrical shapes.

Usually these stones are octahedral in habit with smooth rounded corners. The coloration and clarity of the mounted stone should permit inspection of all drawing surfaces, relief angle, approach angle and bell. A close-grained stone as determined by the crystal growth lines seems best.

STONE SIZE

The size of stone for specified hole sizes, consistent with the quality of stone previously noted, has been well standardized for the copper wire industry by the American wire die manufacturers. Good die performance has been obtained from these sizes. The average stone size of these manufacturers has been specified for our requirements.

MOUNTING

The mounting encasing the stone is a vital factor in the life of diamond wire drawing dies. Diamonds when used for drawing dies ultimately fail by breakage (i.e. cracks, spalling, chipping out) and not by erosion of the wearing surfaces. Diamonds due to their formation and crystalline structure and to the physical properties resulting therefrom, are particularly prone to breakage when subjected to fatigue, impact and disrupting tensile stresses such as those commonly met in wire drawing operations. To compensate for this weakness it is necessary to effectively encase the stone in a mounting which adequately supports it on all sides. The methods of mountings are many and varied. Mountings which are made by hot pressing, forging or upsetting to obtain definite bonding of the component parts were found to be superior to those made by brazing, puddling, or casting. High-strength metals processed with the proper technique to effect good metal flow and alloying bonding are preferable. The interface of the stone and metal should be free from porosity and crevices. Figure 2 depicts a defective mounting due to poor metal flow which caused the early breakage of the stone.

CONTOUR OF THE DIE HOLE

The contour of the die hole is shown in Fig. 1. It is the familiar radial or parabolic type which is commonly used in the copper wire industry with good results. The straight angles associated with cemented carbide dies do not exist in the present day commercial diamond dies. The noted angles are arcs and the values cited refer to the average chords subtending these arcs. In agreement with the findings of comparatively recent English investigations[3] of the theoretical factors affecting wire die performance, it was found that small changes of angle do not appreciably influence the die pull. Whatever differences do exist are blanketed out by changes in the die frictional forces. The permissible variations in angle are consistent with the average found in today's best commercial dies. In recutting dies ob-

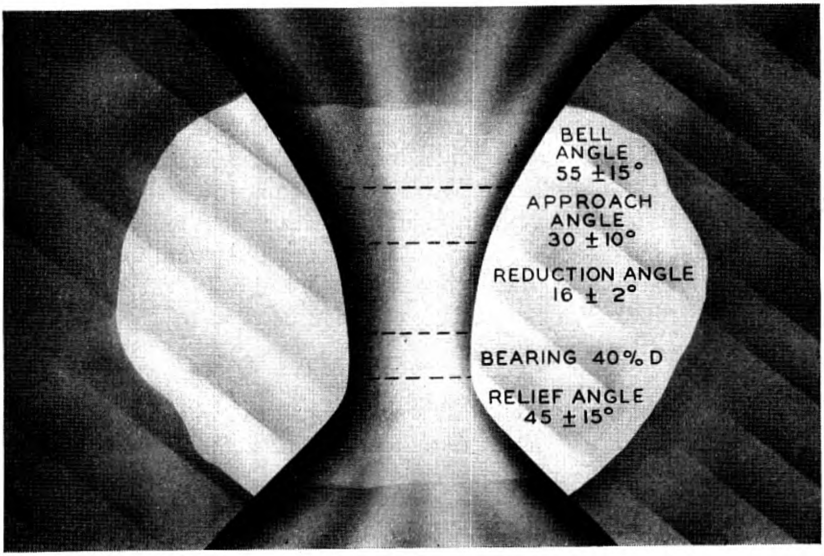


Fig. 1—Contour of the die hole

tained from different sources the variations narrow after one or two recuts. This is to be expected when standardized recutting practises are followed.

BELL ANGLE

The bell angle should permit ready ingress of the compound solution to wash out those materials (dust, slivers, sludge, dirt, etc.) which tend to accumulate here and have an abrasive action when pulled through the die by the wire. The bell angle is highly polished to minimize the packing of the noted materials. This polish can be more easily achieved in a smooth tapering angle than in one of a wide flare.

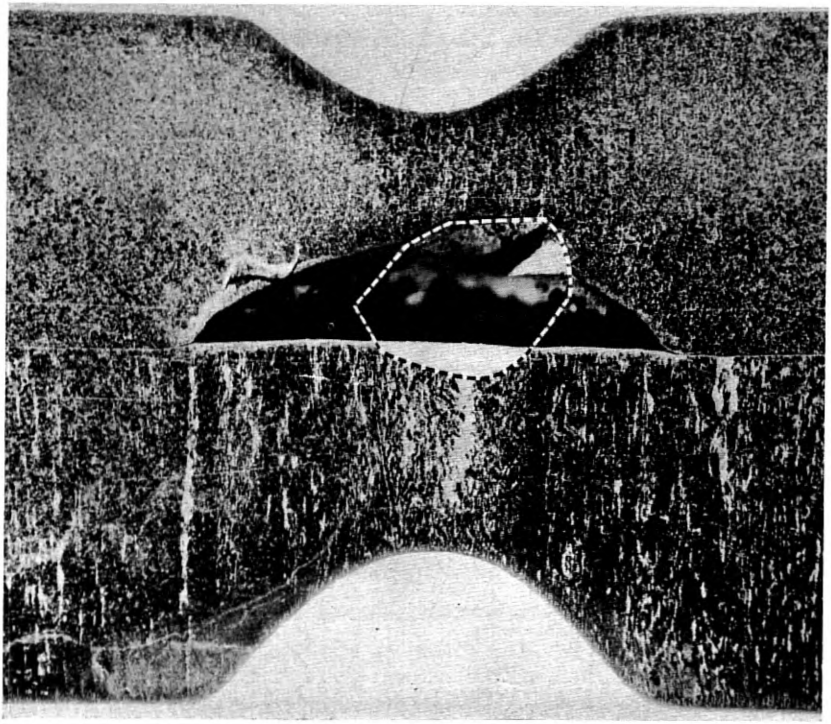


Fig. 2—Defective diamond die mounting due to poor metal flow

APPROACH ANGLE

At high drawing speeds, appreciable wire whip and vibration are present. Also, the higher centrifugal forces tend to "throw out" the wire in relation to the normal axis of the die. An approach angle not too wide in relation to the reduction angle is necessary to guide the wire concentric to the drawing cone.

REDUCTION ANGLE

In this angle, the most important sector of the die, the reduction of diameter, except for the sizing accomplished in the bearing, is carried out. Following the practice of the trade a 16° angle is used. The depth of this area is designed to have the entering wire hit well within it to prevent excessive chipping and undercutting in the pressure ring and to maintain the desired die pull values. The line of contact of the entering wire and the reduction area is located at 65–70% of the vertical height of the reduction angle.

In general, the ratios of the vertical heights of the reduction, approach and bell angles are as follows: 1:0.7:1.5 for dies of 20–26 A.W.G. inclusive and 1:0.5:1 for the larger sizes. The dimensions of the approach and bell angles are of course dependent on the stone size and the number of previous recuts. The heights of the two latter angles may vary appreciably from the above noted values provided that adequate lubrication and "washing out" of the die can take place.

BEARING

A bearing length of about 40% diameter is formed. It is checked by the die pull requirements. The bearing is purposely drilled to obtain a slight front taper of about 2°. The slight taper aids the metal flow and reduces "sucking".

RELIEF ANGLE

Here again, the wire vibration necessitates a smooth angle which will not tear the wire leaving the die. Recutting for good reproduction of contour and die pull also requires a smooth angle. Proper care should be taken to prevent a cup shaped depression which is commonly formed by jamming a diamond chip in the back of the die as a fast intermediate step to final sizing.

POLISH AND FINISH

The degree of polish of the die drawing surfaces is a most important factor in determining satisfactory die performance. In wire drawing there is encountered a vicious repetitive cycle consisting of (1) the wire abrades due to the condition of the wire and die surfaces, (2) the resultant abraded metal particles pack in the die throat and gall the incoming wire, and (3) the heat and the conditions created by this action again adversely affect the wire and the die surfaces to repeat the cycle. There have been cases noticed where the fine copper particles have been compacted by the heat and pressure to form a solid conical mass which has welded to the wire. Deep grooves in the reduction angle and bearing invariably follow back to the pressure ring or irregularities in the die surfaces in which the metal particles, dirt, etc. have lodged.

Characteristic of copper and other non-ferrous metals is its ready flow under pressure such as is present in the die reduction angle. This condition causes die packing, metal galling and the resultant wire scoring to take place at surprisingly slight irregularities in the die drawing surfaces. This is particularly true at high drawing speeds due to the rapid lineal movement of the wire surfaces. Again this emphasizes the necessity for a high polish. The harder ferrous metals tend to bridge across these small irregularities rather than to flow into them.

Highly polished surfaces free from scores, traces of ripping rings and scratches are desired. All the die surfaces with the exception of the relief angle and top flare of the bell should be so finished. The polished surfaces should be a smooth curve as undulating surfaces are undesirable.

TABLE I

Diamond Die Diameter	Max. Pounds Pull	Diamond Die Diameter	Max. Pounds Pull
13 A.W.G	113	20 A.W.G.	25
14 "	92	21 "	21
15 "	75	22 "	17
16 "	60	23 "	13.5
17 "	49	24 "	11
18 "	39.5	25 "	9
19 "	32	26 "	7

Dies after full use to oversize will retain their original high polish if no breaks in the stone, i.e. chips, cracks, etc. have occurred. This polish will also be present in the wear eroded areas.

POUNDS PULL TO DRAW THE WIRE THROUGH THE DIE

Die pull has long been known to the trade as a factor which greatly influences wire drawing operations. However, it has not been customary to recognize its relation to the drawing die itself and it has not been applied as an everyday simple means of checking a die's possible performance. The Western Electric Company has considered for a long time this die characteristic in its wire drawing developments[4] and practical values for use in the present high-speed drawing machines have been established. Listed in Table 1 are the values used for these dies. It should be added that, while die pulls are important, these are not critical and a deviation of +10% can be used but this would result in greater power consumption and more frequent wire breaks. The noted figures are 45–50% of the wire breaking strength based on A.W.G. reduction steps. Die pull is a cumulative check of the following die characteristics: contour as correlated to the reduction angle and bearing length, and the polish of the drawing surfaces. In conjunction with the die diameter, it affords a means of grouping the dies in matched balanced sets and obviates actual try out of these dies in the machines.

INSPECTION METHODS

Suitable inspection methods and equipment to check the die requirements contribute much to good die performance. Microscopes have been used with good results by some manufacturers of fine wires and by several of the more progressive die manufacturers for some time for this purpose. The Western Electric Company has adopted this practice. The binocular wide field microscope has proved to be an indispensable tool for inspection. It is used to the complete exclusion of loupes. The stereoscopic effect, the wide field and the long focal distance features of this instrument make satisfactory examination possible. We have found a magnification of 30× to be most practical for the previously noted die sizes. Good agreement to observations by different persons can be had by the use of this apparatus. This is a rare occurrence in the case of loupes. It is surprising that this inexpensive tool which removes much of the guessing and so-called art in diamond die manufacture has not been universally utilized by wire drawers and die manufacturers.

Frequent examination of finished dies and dies in the various stages of recutting is expediently accomplished by microscopic observation. Routine die contour checks are visually made in this manner. Periodic checks of die angles and die contour are accurately made with a contour projector. Shadowgraphs of 100× magnification are made of die impressions formed of a soft metal.

When polished to the proper degree, it is possible to examine a mounted die under a microscope with a suitable light source and discern the internal flaws in the stone. The examination of stones removed intact from the mounting will also check this point. Mountings of metals having a low reflectivity will require a light source from above to illuminate the stone.

The die pull is checked on a commercial instrument shown in Fig. 3. Auxiliary equipment and gauges have been added to it to make it more versatile for this use. The device consists of a piston which fits snugly into a hydraulic chamber filled with a suitable liquid. The piston is recessed at the outer end to hold the die. The die pull causes the movement of the piston which builds up the registered pressure. The wire is reduced one A.W.G. step and the pressure against the die holder is noted as indicated on

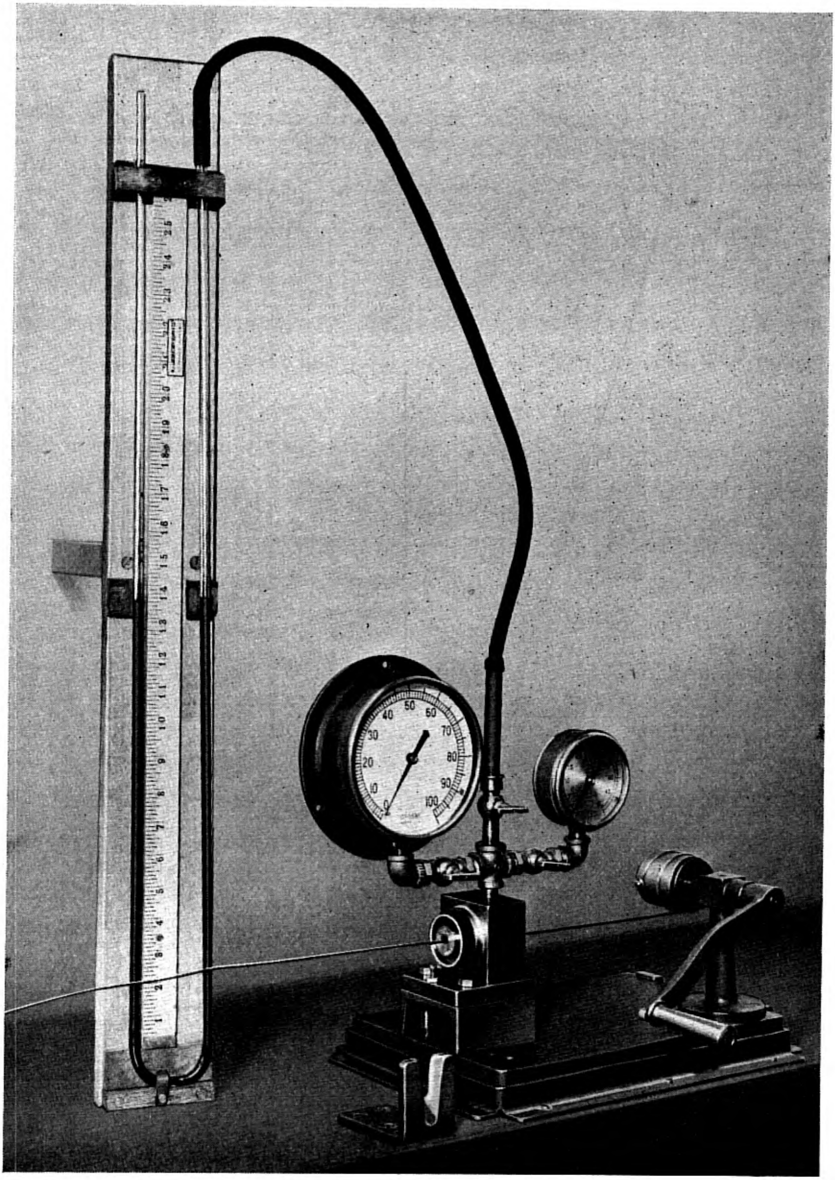
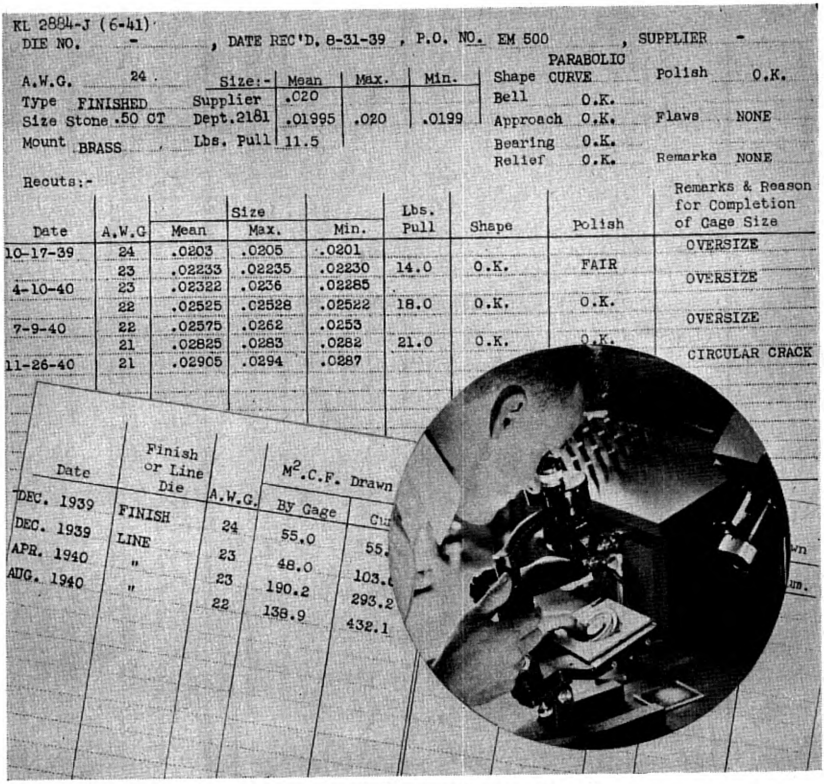


Fig. 3—Die pull indicator

the adjoining gauge. Small die sizes having die pulls below five pounds are checked by means of a manometer filled with an appropriate liquid. When measuring die pulls of low magnitude, the device should be calibrated for the frictional drag of its moving parts.

Die Records

Die control is greatly enhanced by comprehensive records. All dies are logged on individual record cards. A facsimile of each side of this card is shown in Fig. 4. As noted, the individual die characteristics as received, at oversize, and after recutting to final disposition are recorded. These cards are periodically correlated to the daily machine performance sheets by mechanical tabulation to obtain the die life.



KL 2884-J (6-41)

DIE NO. -, DATE REC'D. 8-31-39, P.O. NO. EM 500, SUPPLIER -

A.W.G. 24	Size:-	Mean	Max.	Min.	Shape	PARABOLIC CURVE	Polish	O.K.
Type FINISHED	Supplier	.020			Bell	O.K.		
Size Stone .50 CT	Dept.2181	.01995	.020	.0199	Approach	O.K.	Flaws	NONE
Mount BRASS	Lbs. Pull	11.5			Bearing	O.K.		
					Relief	O.K.	Remarks	NONE

Recuts:-

Date	A.W.G	Size Mean	Size Max.	Size Min.	Lbs. Pull	Shape	Polish	Remarks & Reason for Completion of Cage Size
10-17-39	24	.0203	.0205	.0201				OVERSIZE
	23	.02233	.02235	.02230	14.0	O.K.	FAIR	
4-10-40	23	.02322	.0236	.02285				OVERSIZE
	22	.02525	.02528	.02522	18.0	O.K.	O.K.	
7-9-40	22	.02575	.0262	.0253				OVERSIZE
	21	.02825	.0283	.0282	21.0	O.K.	O.K.	
11-26-40	21	.02905	.0294	.0287				CIRCULAR CRACK

Date	Finish or Line Die	A.W.G.	M^2.C.F. Drawn By Gage	M^2.C.F. Drawn Cu
DEC. 1939	FINISH	24	55.0	55.
DEC. 1939	LINE	23	48.0	103.
APR. 1940	"	23	190.2	293.2
AUG. 1940	"	22	138.9	432.1

Fig. 4—Facsimile of a die record card

Shop Die Control

All drawing dies are grouped in sets for use and matched for balance in relation to their diameters. This is necessary because of the minimum slip characteristics of the drawing machines. It is also a convenient means to reduce machine down time. The sets are used in the drawing machines as units. Replacements and rematchings are made in the Die Shop. Definite oversize diameter limits consistent with best enlargement practice and overall operating efficiency have been set for each finisher and line die size. The die sets are removed from the machine when any one die, with the exception of the finisher, has reached the specified limit. Usually two finishers are used for each set of 26 A.W.G. and one plus for the other final gauge sets before returning these units to the Die Shop. This procedure has greatly enhanced the economics of die costs, operating efficiency and wire quality. It has also made possible the issuance of dies by the Die Shop for shop use without the necessity for preliminary "tryout" in the drawing machines.

RECUTTING PRACTICE

Diamond die recutting is a series of lapping operations in which the abrasive is diamond dust for obvious reasons and the lap or carrier is a suitable pin or wire, usually steel, depending upon the specific operation. The theory and practice of lapping tool gauges to a fine finish is applicable here. The diamond dust imbeds in the lap, is held there temporarily stationary and cuts the die stone when it is forced in contact with it. As its position is fixed relative to the lap, but moving with respect to the stone, it cuts the die surface. However, due to the great hardness of diamond, appreciable wear of the lap takes place and its shape is rapidly changed. In ripping where the shaping of the die contour is done, the pin must be frequently reground. It is necessary to have a hard material for the lap to keep the diamond dust working in contact with the die surface and to retain its shape for an appreciable time. Also, some diamond dust would imbed deeply in a soft material and be removed in the grind.

Successful recutting and reproduction of die characteristics require definite procedures and controls. Specific hole size enlargement, surface requirements, frequency of lapping pin grinds, diamond dust additions, etc. are essential for each operation to achieve the desired results.

Standard commercial machines are used for the various enlarging and refinishing operations. In general, recommended trade practices are followed.

Ripping (the roughing operation which removes the pressure rings, chips, etc. and reshapes the die contour) is most important as it determines the die contour and the subsequent degree of polish. We have found the RPM of the drilling pin spindles to be a determining factor. Spindle RPM's from 4,000 to 12,000 have been tried, and 5,000 RPM has been found to be most satisfactory. Other important elements are the diamond dust size, frequency of pin grinds, speed and kick of the reciprocating vertical motion. A definite sequence of these factors is necessary to achieve the desired results. A gradual breakdown of the diamond dust effecting a cutting and lapping action as controlled by the centrifugal forces dispersing it in the liquid medium takes place. The steel lapping pins are accurately ground to a definite straight included angle and circular arc on a special grinder. The ripped die surface should be a smooth, dull matte surface free of deep rings and scores.

In polishing, the die contour has already been established and it is expedient to have the lap of a material which will rapidly shape itself to the contour to be polished to obtain maximum contact. Here the loading of the lap will hold the diamond dust in contact with the work. A viscous medium for the dust will also help. On the wire polishing machines, the springback of the wire on flexing is important. It was found that a wire resiliency tester will aid to evaluate the desired wire properties. Diamond dust of a uniform and fine grain size is vital to the polishing operation. High RPM in the case of revolving die mounts is beneficial provided it does not "throw out" the diamond dust from the working area.

Sizing of the bearing requires a spindle RPM about 100% higher than that used in ripping. Other factors in this operation are the diamond dust, frequency of its addition and the frequency of the pin grinds.

All die recutting machines should be periodically checked and maintained to minimize vibration.

DIAMOND DUST

Another important factor in the recutting operations is the diamond dust abrasive. This is especially true in the finishing operations of polishing and sizing where no appreciable breakdown of the original dust takes place. Uniformity of particle size range and grain size determine the efficacy of the abrasive. Diamond dust graded by air flotation is now being used with good results. It was found that this material was more satisfactory for our use than that obtained by sedimentation in liquid media. Recent investigations in the separation of microscopic size dry powdered material have been actively carried out in the ceramics and powdered metal industries. The Journal of the American Ceramic Society and the Bureau of Mines publications contain noteworthy papers describing practices which could be applicable to diamond dust grading. Stokes[5] law of fall for microscopic size particles will also hold true for the separation of diamond dust by means of air.

DIE LIFE

In theory and practice, die life resolves itself into a problem of wear. The wear is due to the movement of the wire surface over the drawing die area in contact with the wire. In turn the wear is proportional to the forces acting on these surfaces, their area, the condition of these surfaces and the coefficient of friction. This is dependent on the nature of the wire, the die materials and the lubrication of the contacting surfaces which in turn is affected by their condition. Considered on this general basis, the die life is proportional to the wire lineal footage. A good criterion for determining the overall diamond die performance is the total useful life in million conductor feet. This in turn will be proportional to the number of recuts obtainable before breakage renders the die non-usable.

Although diamond die life is theoretically a function of wear, actually it is limited by the failure of the die stone by breakage and "chipping out." Due to the peculiar physical properties of diamonds some breakage will always occur. This condition is also affected by fatigue caused by vibration, impact, thermal stresses and other factors. An important element influencing die breakage which is often overlooked is the detrimental effect produced by improper maintenance of the drawing machine parts. An appreciable reduction in die breakage was obtained by proper periodic maintenance checks.

Table II

B & S Ga.	Kearny Wire Mill 1/39–8/40					
	Finishers		Line Dies		Average*	
	M Lbs.	M.^{2}C.F.	M Lbs.	M.^{2}C.F.	M Lbs.	M.^{2}C.F.
26	35	45	—	—	35	45
25	—	—	90	93	90	93
24	92	75	105	86	97	79
23	—	—	162	108	162	108
22	136	70	159	82	148	76
21	—	—	226	92	226	92
20	—	—	250	81	250	81
19	250	64	284	73	277	71
18	—	—	368	75	368	75
17	—	—	595	96	595	96
16	—	—	740	95	740	95
15	—	—	1040	106	1040	106
Average M.^{2}C.F. per die gauge step						85

Note: Some of the results listed above appear to be inconsistent with the expected trends. This is attributed to the many variables which exist in normal Wire Mill Operations.

* Based on Kearny's distribution of gauges.

Kearny's Wire Mill diamond die life data are listed in million conductor feet per die per gauge step in Table II. The finisher and line die life is tabulated separately to account for the allowable oversize diameter limits for each type. No data are available at the present time to permit listing of the total useful life in million conductor feet.

STRAINS IN DIAMOND STONES

Rough diamond stones as well as those removed after use in drawing dies have been examined in polarized light to determine the presence of internal strains. In general, the clear flawless stones of alluvial origin were free of strains. Mined stones of South African origin which are commonly found in dies of French manufacture did show in several cases very definite strains. These latter stones were of poor quality. Due to the high index of refraction of the diamond which is much higher than available high refractive liquids, and to local surface strains which distort the interference figures, it was impractical to establish a satisfactory procedure to investigate this subject.

CRYSTALLOGRAPHIC STUDIES

In several cases exceptional die performance was noted, three to four times the average, which could not be accounted for by our regular inspection methods. Also, several dies which chipped initially in use, consistently chipped at subsequent recut sizes after enlargement had completely removed the visible defect. Again and most important is that the majority of our dies are rendered non-usable by the breakage of the die stone, which occurs in a characteristically peculiar manner. These discrepancies are attributed to the crystallographic structure of diamonds. It is well known that the reticular density (atom spacing), hardness and cleavage are greater along certain crystallographic planes.

Investigations as to whether or not the relation of the planes of cleavage to the direction in which the hole is drilled has any practical bearing on die life were made by the Western Electric Company in 1929[4]. The x-ray was used to determine the orientation of the crystal planes. The results of a limited test at that date, in which nineteen dies were drilled at right angles and nine parallel to the plane of cleavage, showed no superiority of these positions. The diamond stones used for dies then were of different origin and quality than those now commonly empoyed for this purpose. Further studies, with particular reference to the relation of crystal structure to die breakage, are now under way for reasons noted previously, and some of our results to date are cited here having a bearing on die performance.

It is well known that hardness in a diamond varies with the crystal planes. However, since the contour of the die hole consists of cylindrical and conical surfaces, it is questionable if this plane property could be effectively utilized in drawing dies.

An x-ray method[6] on a shop production basis is now being used to orient quartz plates preparatory to their cutting along certain crystallographic planes incident to their manufacture into radio and carrier frequency control filters. A similar technique was applied to determine the location of the axis of the die drawing hole with relation to the diamond structure. The orientation of the drilled hole was determined by means of a back reflection Laue photograph.

Diamond stones which meet our drawing die requirements are mostly octahedral in shape, a common crystal form of this material. These are usually alluvial in character and predominantly of Brazilian origin with the exception of a few Southwest African stones.

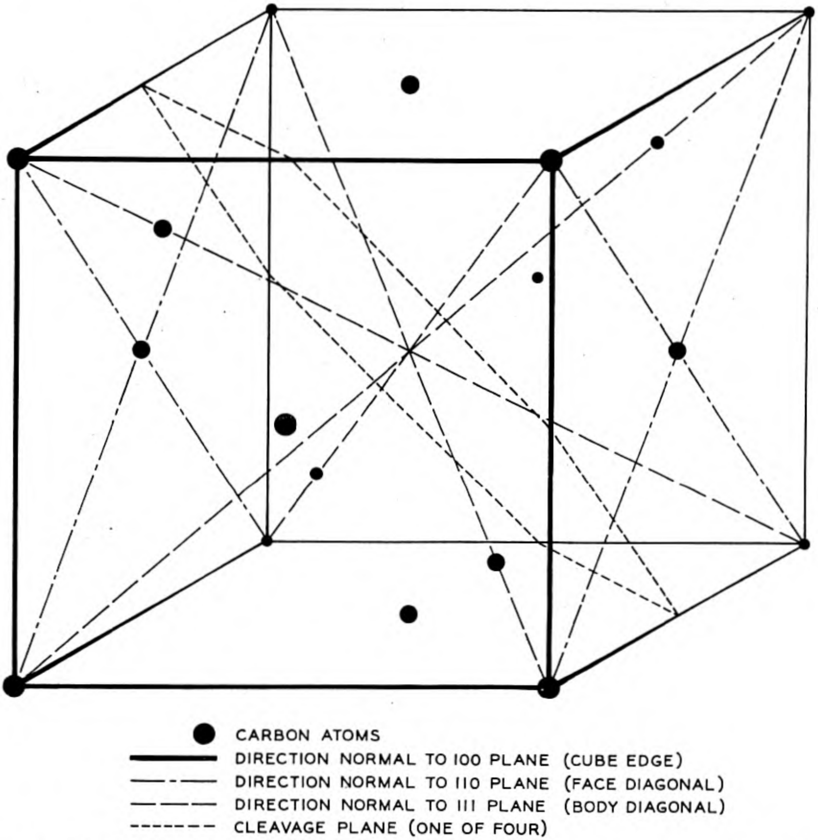


Fig. 5—Diamond structure and the directions of drawing hole axis to the crystallographic planes

Our examinations of numerous die stones have shown that the positioning of these is random with regard to a predetermined drilling axis. About 75% of those studied were drilled normal to the 111 orientation. The axis of the drawing hole in these cases was in the direction of a body diagonal of the cube representing the crystal structure of the diamond. The 111 planes, the cleavage planes for the diamond, are those which appear most frequently as the natural faces of the most common form of the diamond crystal, the octahedron. The disposition of the stone in the above manner would be logically expected to take place in the mounting methods generally used by the American die manufacturers since the stone would probably rest on a flat side, an octahedral face. This placement for very obvious reasons also facilitates cutting true the starting cone for the drilling operations.

A few stones had been drilled normal to the 110 and 100 planes or at an angle to the 111 plane. Holes normal to the 110 plane, which is the face of the rhombic dodecahedron, another diamond crystal form, are along a face diagonal of the cube. Holes normal to the 100 plane, which is the cube face, are along a cube edge. Very few of these were found. This is as expected since natural crystals with faces parallel to the 100 plane are rare. This face is usually produced by sawing or "bruting."

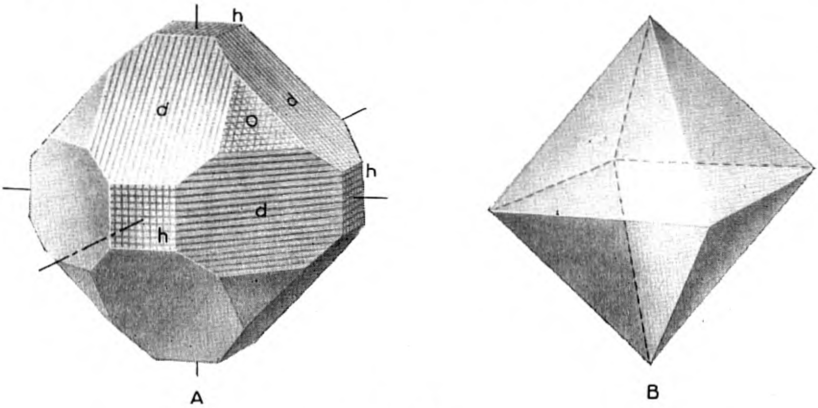


Fig. 6[7]

A—Diamond crystal showing grain and faces (crystallographic planes)

h. Cube face	100 plane
d. Rhombic dodecahedron	110 plane
o. Octahedron	111 plane

B—Diamond octahedron

As previously noted, diamond die life is restricted by the breakage and "chipping out" of the stone. The manufacture of the dies with due consideration of the peculiar crystallographic structure of the diamond to minimize this condition should greatly enhance its performance. Findings to date substantiate this conclusion. Seventy per cent of the dies which were drilled normal or at a small angle to the 111 plane failed due to cracks parallel to this plane and normal to the drawing axis. The crack usually occurred at the pressure ring and its plane coincided with the cleavage plane, the weakest one in the diamond crystal. The stone in many cases on removal from the metal mounting separated into two sections. The majority of the dies failed in this manner.

Examination of worn oversize dies invariably shows that the erosion and the effect of the forces incident to wire drawing are concentrated at the pressure ring where the entering wire contacts the reduction area. Here, the abrasive action of the wire and die contacting surfaces, and the disrupting stresses are most evident. An analysis of the forces present in a wire drawing die indicates that secondary stresses normal to the drawing axis are set up in this sector. In dies drilled normal to the 111 plane these stresses, coupled with the thermal, impact and fatigue forces converging in this area, would tend to shear the diamond stone parallel to this plane. Overdrawing and poor operating conditions would of course accentuate this effect and bring it about prematurely.

The previously mentioned dies drilled normal to 110 and 100 planes did not crack in the manner noted above. These failed due to chipping and spalling. Cracks, when present, were small and inclined to the drawing axis. Dies drilled normal to 110 plane have cleavage planes parallel to the die hole axis. In one case a stone drilled in this manner cracked along this plane. Data to date indicate that better die life was obtained from these dies than from those drilled normal to the 111 plane.

The noted studies indicate that the orientation of the drawing hole is of practical importance with regard to stone breakage, the limiting factor of die life. A better understanding of its relation to die performance should achieve appreciable economies. No conclusions have been made as yet with regard to the exact desired orientation of the drawing axis. In Fig. 5 is depicted a unit cube of a diamond structure. Here is shown the possible directions of drilling and the location of the cleavage plane. Figure 6 shows the diamond crystal faces and planes.

Whether or not the aforementioned die breakage characteristics are local to our drawing operations and die sizes is not known. We invite comments from other wire manufacturers whose equipment and practices vary significantly from those used at the Kearny Wire Mill.

Acknowledgment

The author wishes to acknowledge and thank Mr. F. E. Haworth, of the Bell Laboratories, for his x-ray determinations and Mr. W. W. Broughton, of the Quartz Crystal Development Group of this Company, for his many suggestions with reference to the crystallographic studies.

References

1. "Design and Operation of a New Copper Wire Drawing Plant," H. Blount and J. E. Wiltrakis, *Wire and Wire Products*, October 1940. Bell Telephone System Technical Publications *Monograph* B-1273.

2. "Developments in the Manufacture of Copper Wire," J. R. Shea and S. McMullan, Western Electric Company, Inc. Jour. A. I. E. E., Vol. XLVI, February 1927. *Bell System Technical Journal*, Vol. VI, April 1927.
3. Stresses in Wire Drawing (Digest of research work conducted at Sheffield University, England), A. C. Blackall, *Wire and Wire Products*, August 1933.
   Fundamentals of Wire Drawing, F. C. Thompson, Victoria University of Manchester England, *The Iron Age*, October 13, 1908.
4. "Development of Wire Drawing Methods," J. R. Shea, Western Electric Company, Inc. Part three of papers presented at the Hawthorne Works of the Western Electric Company, Inc., Chicago, Illinois, July 10–12, 1929, to the Summer School for Engineering Teachers sponsored by the Society for the Promotion of Engineering Education.
5. A quantitative physical law governing the fall of small particles in various media due to gravity. *Handbook of Chemistry and Physics*, p. 2224, 24th Edition 1940–41.
6. "Manufacture of Quartz Crystal Filters," G. K. Burns, Western Electric Company, Inc. Bell Telephone System Technical Publications *Monograph* B-1253.
7. "Variation of Hardness in the Diamond," E. H. Kraus and C. B. Slawson, University of Michigan, *The American Minerologist*, November 1939.

Using Double Sampling Inspection in a Manufacturing Plant

By D. B. KEELING and L. E. CISNE

THE necessity for quality control in a manufacturing plant arises from the fact that all units of product cannot be made identical. To limit variations and attain controlled uniformity some sort of inspection must be established. It has been the experience of the Western Electric Company that quality control may be attained most economically by the use of a sampling inspection wherein only a portion of the entire output is examined for desired quality characteristics.

Advantages which have been gained through the use of sampling inspection, and with no adverse effect on previously existing quality levels, are: a reduction in the cost of inspection by economies in inspection time; a reduction in the amount of scrap produced by making available for supervisory action useful records of the results of inspection; and as an end result, the attainment of uniform quality of a satisfactory level.

It is the purpose of this paper to provide a detailed method of procedure that has proved successful in establishing and maintaining one type of sampling—the "Average Outgoing Quality Limit" Double Sampling Plan. Statistically determined tables of lot sizes and corresponding sample sizes which guarantee a certain degree of protection have been used by the Western Electric Company for approximately fifteen years. They were furnished by the Bell Telephone Laboratories and have recently been made generally available in an article published in the January 1941 issue of the Bell System Technical Journal.[1] A typical sampling table, reproduced from the preceding article, is shown in Fig. 1.

Briefly stated, the AOQL Double Sampling Plan involves the examination on a "go—no go" basis of a specified number of articles taken at random from a large group. The acceptance or rejection of this group is usually made on the basis of results obtained from the first sample alone. How-

---

[1] The AOQL Double Sampling Tables specifically referred to in the present article are Tables DA-0.1 to DA-10.0 shown on pages 49-61, inclusive, of a preceding article: H. F. Dodge and H. G. Romig, "Single Sampling and Double Sampling Inspection Tables", *Bell Sys. Tech. Jour.*, pp. 1–61, Jan. 1941. These tables give sample sizes and allowable numbers of defectives for a variety of AOQL values, lot sizes, and process average values. The sampling table reproduced in Fig. 1 is based on an AOQL value of 1.5 per cent defective. In tables prepared for shop use, it has been found preferable to use a notation slightly different from that shown in Fig. 1, specifically to use AN instead of c to represent "allowable number of defectives" and to use SS instead of n to represent "sample size". The shop notation is used in the present article. It should be noted that, in the original article, C was generally referred to as an allowable number of "defects", since primary attention was given to inspection of a single characteristic. See footnote 2 for explanation of the terms "defective" and "defect".

TABLE IV CONT'D: DOUBLE SAMPLING LOT INSPECTION TABLES—BASED ON STATED VALUES OF "AVERAGE OUTGOING QUALITY LIMIT"

TABLE DA-1.5

AVERAGE OUTGOING QUALITY LIMIT = 1.5%

Process Average %	0–.03						.04–.30						.31–.60						.61–.90						.91–1.20						1.21–1.50					
Lot Size	Trial 1		Trial 2			p_t %	Trial 1		Trial 2			p_t %	Trial 1		Trial 2			p_t %	Trial 1		Trial 2			p_t %	Trial 1		Trial 2			p_t %	Trial 1		Trial 2			p_t %
	n_1	c_1	n_2	n_1+n_2	c_2		n_1	c_1	n_2	n_1+n_2	c_2		n_1	c_1	n_2	n_1+n_2	c_2		n_1	c_1	n_2	n_1+n_2	c_2		n_1	c_1	n_2	n_1+n_2	c_2		n_1	c_1	n_2	n_1+n_2	c_2	
1–15	All	0	–	–	–	–	All	0	–	–	–	–	All	0	–	–	–	–	All	0	–	–	–	–	All	0	–	–	–	–	All	0	–	–	–	–
16–50	16	0	–	–	–	11.6	16	0	–	–	–	11.6	16	0	–	–	–	11.6	16	0	–	–	–	11.6	16	0	–	–	–	11.6	16	0	–	–	–	11.6
51–75	23	0	11	34	1	10.5	23	0	11	34	1	10.5	23	0	11	34	1	10.5	23	0	11	34	1	10.5	23	0	11	34	1	10.5	23	0	11	34	1	10.5
76–100	26	0	14	40	1	9.4	26	0	14	40	1	9.4	26	0	14	40	1	9.4	26	0	14	40	1	9.4	26	0	14	40	1	9.4	26	0	14	40	1	9.4
101–200	31	0	18	49	1	8.4	31	0	18	49	1	8.4	31	0	18	49	1	8.4	35	0	35	70	2	7.5	35	0	35	70	2	7.5	35	0	35	70	2	7.5
201–300	33	0	22	55	1	8.0	33	0	22	55	1	8.0	38	0	37	75	2	7.0	38	0	37	75	2	7.0	38	0	37	75	2	7.0	40	0	60	100	3	6.6
301–400	34	0	21	55	1	7.9	34	0	21	55	1	7.9	39	0	41	80	2	6.9	39	0	41	80	2	6.9	42	0	63	105	3	6.3	42	0	63	105	3	6.3
401–500	35	0	20	55	1	7.8	35	0	20	55	1	7.8	39	0	46	85	2	6.9	44	0	71	115	3	6.1	44	0	71	115	3	6.1	46	0	94	140	4	5.8
501–600	35	0	20	55	1	7.8	40	0	45	85	2	6.8	40	0	45	85	2	6.8	44	0	71	115	3	6.1	44	0	71	115	3	6.1	47	0	98	145	4	5.7
601–800	35	0	20	55	1	7.8	41	0	49	90	2	6.7	46	0	74	120	3	6.0	46	0	74	120	3	6.0	49	0	101	150	4	5.6	85	1	125	210	6	5.0
801–1000	36	0	19	55	1	7.8	42	0	48	90	2	6.5	47	0	78	125	3	5.9	47	0	78	125	3	5.9	50	0	105	155	4	5.5	90	1	125	215	6	4.9
1001–2000	44	0	51	95	2	6.3	44	0	51	95	2	6.3	49	0	81	130	3	5.7	55	0	105	160	4	5.4	95	1	175	270	7	4.6	100	1	230	330	9	4.4
2001–3000	45	0	50	95	2	6.2	45	0	50	95	2	6.2	55	0	110	165	4	5.3	100	1	145	245	6	4.6	110	1	255	365	9	4.1	155	2	345	500	13	3.9
3001–4000	45	0	50	95	2	6.2	50	0	85	135	3	5.5	55	0	115	170	4	5.2	105	1	190	295	7	4.3	160	2	300	460	11	3.8	205	3	405	610	15	3.6
4001–5000	45	0	50	95	2	6.2	50	0	85	135	3	5.5	55	0	120	175	4	5.1	105	1	190	295	7	4.3	165	2	340	505	12	3.7	250	4	480	730	18	3.5
5001–7000	46	0	54	100	2	6.1	50	0	90	140	3	5.4	60	0	155	215	5	4.7	110	1	225	335	8	4.2	165	2	375	540	13	3.7	310	5	610	920	22	3.3
7001–10,000	46	0	54	100	2	6.1	50	0	90	140	3	5.4	60	0	160	220	5	4.6	115	1	280	395	9	3.9	170	2	420	590	14	3.6	360	6	660	1020	24	3.2
10,001–20,000	46	0	54	100	2	6.1	50	0	90	140	3	5.4	60	0	165	225	5	4.5	120	1	315	435	10	3.8	210	3	420	630	15	3.6	415	7	835	1250	29	3.1
20,001–50,000	47	0	53	100	2	6.1	55	0	125	180	4	5.0	65	0	195	260	6	4.4	165	2	350	515	12	3.7	225	3	640	865	20	3.3	510	9	1130	1640	38	3.0
50,001–100,000	47	0	53	100	2	6.1	55	0	130	185	4	4.9	115	1	235	350	8	4.0	175	2	440	615	14	3.5	275	4	725	1000	23	3.2	570	10	1400	1970	45	2.9

Fig. 1—Typical sampling table for the AOQL double sampling plan; AOQL = 1.5 per cent defective
(Reproduced from article referred to in footnote on page 37)

ever, if the results from the first sample are not conclusive, an additional sample is examined before disposition of the lot is made.

The particular type of articles to which this plan has been applied are products consisting of individual parts, sub-assemblies, or completed apparatus, which, at the various stages of production where control is necessary, are the result of repetitive operations capable of considerable uniformity. The plan has also been applied to some extent on completed products and purchased materials where there is evidence that the product is of reasonable uniformity even though the quality history is meager or unavailable.

1. STEPS IN SETTING UP A DOUBLE SAMPLING LOT INSPECTION PLAN

1.1 *Analysis of the Production Process*

In order to determine the applicability of the Double Sampling Plan to existing inspection operations it is necessary to examine the manufacturing and inspection processes and all data available regarding past quality performance, such as records of per cent defective, consumers' complaints, etc. The following outline should serve not only as a measuring stick to determine the applicability of Double Sampling but also as an index of the conditions to be met for the successful use of Double Sampling with any inspection operation.

1.11 The *lot*, or group of articles to be examined, should consist of product which is available in its entirety for acceptance or rejection at one time. For sampling purposes, the lot should have characteristics which are the result of a common system of causes. By this it is meant that the lot should, as far as possible, consist of articles made from relatively uniform raw material by operators of equivalent skill and by machines or methods of equivalent precision. If there is evidence of appreciable variation between corresponding machines, operators, or materials, it is desirable to confine a lot to the output of one machine, one operator, or one batch of material, in order to isolate a uniform group of product suitable for sampling inspection.

In brief, Double Sampling may be applied to the output of any repetitive unit operation capable of sufficient uniformity. However, unless immediately essential for economic reasons, it need not be applied at the particular time such an operation is completed, provided succeeding operations do not modify the inspection item under consideration.

1.12 In order to gain the maximum advantage from the use of Double Sampling, it is necessary that lots be as large as the limitation of uniformity will allow so that protection and control may be achieved with a minimum sample size relative to the number of units in the lot. The fact that proportionately smaller sample sizes are used with the larger lot sizes may be seen by reference to the Sampling Table of Fig. 1.

In sampling from larger lots it becomes *increasingly important* in practice to observe certain precautions in order to take care of instances where the lot may not be homogeneous; specifically, each sample should be a group of articles *taken at random from different locations throughout the lot* so that it will represent an impartial cross section of the lot.

1.13 It is necessary for the successful operation of any sampling inspection that at all times there be a known standard of acceptability for the individual article, that is, a definite description of the requirement for each inspection item and reliable measuring equipment against which product may be conclusively checked. Practically, this condition will be realized when the characteristic is defined in such a manner that different observers will obtain consistent results.

1.14 The theoretical background of the plan assumes the repair or elimination of all defectives[2] in samples of accepted lots, as well as all defectives in rejected lots before such lots are passed.

In order to accomplish this, close cooperation between production and inspection personnel is required in assuring that rejected lots are thoroughly cleared of all defectives, since failure to comply with this fundamental part of the procedure destroys the very foundation of the plan.

1.2 *Selection of Proper Double Sampling Table*

Tables are provided for a variety of Average Outgoing Quality Limit (AOQL) values and Process Average classes. The AOQL value is the maximum value of average per cent defective in the product after inspection, which the sampling plan will assure over a long period of time, no matter how defective the product submitted for inspection may be. The process average is the normal per cent defective which is to be expected from the process.

To determine what AOQL valve should be adopted, it is necessary to decide upon a maximum average per cent defective which may be permitted in the product without serious consequences to the user. Product which is of such a nature that defects will be eliminated in subsequent operations may be assigned a rather generous AOQL, and conversely, product which by its relation to the entire assembly may cause considerable inconvenience

---

[2] A *defective* is defined as an individual article that fails to meet the requirements for one or more inspection items. A *defect*, however, is defined as a failure to meet a requirement for a single quality characteristic for which inspection is made. It follows that when several characteristics are inspected, an individual article may have several *defects*, yet be only a single *defective*.

if it fails to meet requirements, usually warrants a strict AOQL. A rather generous AOQL may be assigned to inspection features which are considered relatively unimportant. In other cases the use of sampling may be definitely inadvisable due to the importance of the requirement from a functional standpoint or from the standpoint of the possible effect of a failure upon the safety or health of an individual.

As a guide in the selection of the AOQL, the following table is given, showing values that have been found to be satisfactory for the product listed. These percentage values of AOQL represent per cent of articles defective; if more than one defect is found on an article, the article is counted as one defective. Here, as elsewhere throughout the paper, all figures relate to number of defectives and per cent of articles defective rather than number of defects.

Description	*Requirements*	*AOQL*
Machine Screws	5 Dimensions	2.0%
Hexagon Nuts	Visual Inspection after Zinc Plating	2.0%
Twin Eyelets	6 Dimensions and 4 Visual Requirements	3.0%
Relay Coils*	Inductance and Electrical Breakdown	1.0%
Misc. Completed Electrical Apparatus	Resistance	.5%

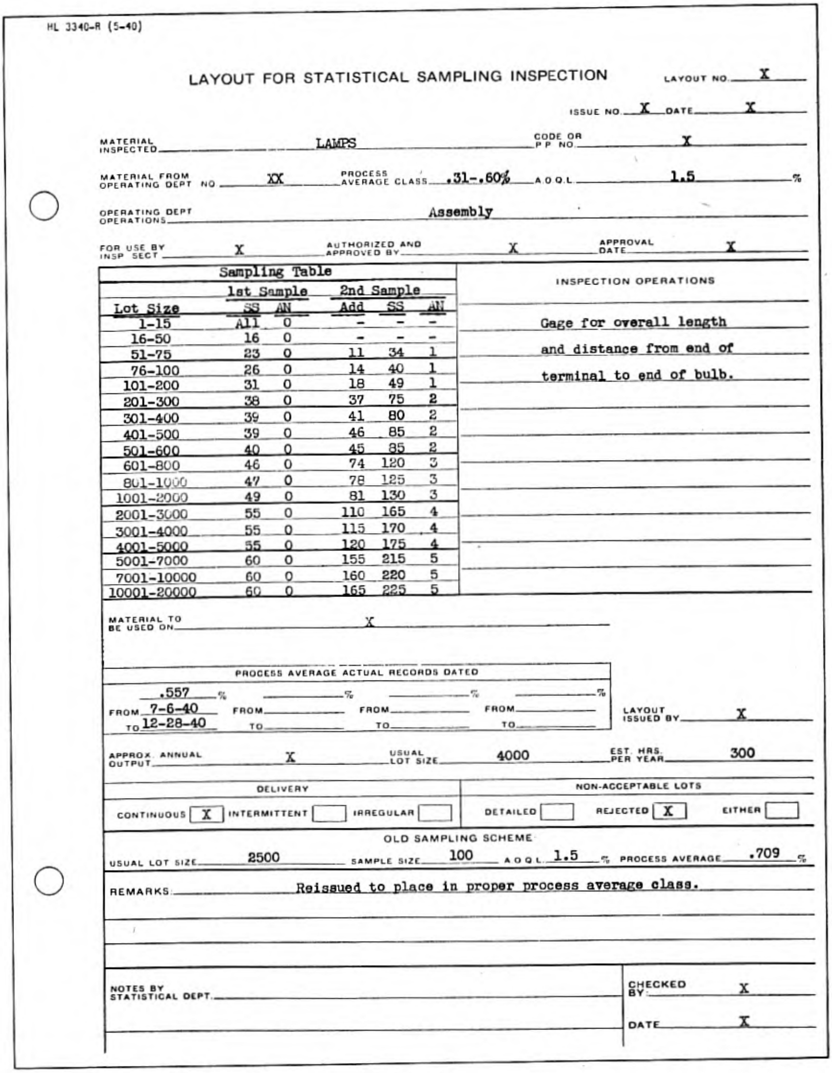
The process average is commonly determined by summarizing the results of the first samples inspected during a representative period and may usually be obtained if there has been a previous inspection with associated records. In case suitable records are not available, an approximation may be made on the basis of an examination of a number of random samples selected from product of current manufacture. This will be a tentative figure and may require revision when data accumulated from the operation of the Double Sampling Plan are available.

1.3 *Issuance of the Inspection Layout—Fig. 2*

After the AOQL value has been established, definite instructions in the form of a Sampling Layout (See Fig. 2) should be provided for the use of the inspector. The layout should contain a list of the inspection items for which inspection is required. It should also contain a copy of the sampling table selected, a description of the material to be inspected, AOQL value, process average and other information which is of importance either in using or in maintaining the Double Sampling Plan.

---

* This is a process check for these requirements which is supplemented by another sampling inspection after assembly.



HL 3340-R (5-40)

LAYOUT FOR STATISTICAL SAMPLING INSPECTION

LAYOUT NO. X

ISSUE NO. X DATE X

MATERIAL INSPECTED LAMPS

CODE OR P P NO. X

MATERIAL FROM OPERATING DEPT NO XX

PROCESS AVERAGE CLASS .31-.60%

A.O.Q.L. 1.5 %

OPERATING DEPT OPERATIONS Assembly

FOR USE BY INSP SECT X

AUTHORIZED AND APPROVED BY X

APPROVAL DATE X

Sampling Table

	1st Sample		2nd Sample		
Lot Size	SS	AN	Add	SS	AN
1-15	All	0	-	-	-
16-50	16	0	-	-	-
51-75	23	0	11	34	1
76-100	26	0	14	40	1
101-200	31	0	18	49	1
201-300	38	0	37	75	2
301-400	39	0	41	80	2
401-500	39	0	46	85	2
501-600	40	0	45	85	2
601-800	46	0	74	120	3
801-1000	47	0	78	125	3
1001-2000	49	0	81	130	3
2001-3000	55	0	110	165	4
3001-4000	55	0	115	170	4
4001-5000	55	0	120	175	4
5001-7000	60	0	155	215	5
7001-10000	60	0	160	220	5
10001-20000	60	0	165	225	5

INSPECTION OPERATIONS

Gage for overall length
and distance from end of
terminal to end of bulb.

MATERIAL TO BE USED ON X

PROCESS AVERAGE ACTUAL RECORDS DATED

.557 %	%	%	%
FROM 7-6-40	FROM	FROM	FROM
TO 12-28-40	TO	TO	TO

LAYOUT ISSUED BY X

APPROX. ANNUAL OUTPUT X

USUAL LOT SIZE 4000

EST. HRS. PER YEAR 300

DELIVERY

CONTINUOUS ☒ INTERMITTENT ☐ IRREGULAR ☐

NON-ACCEPTABLE LOTS

DETAILED ☐ REJECTED ☒ EITHER ☐

OLD SAMPLING SCHEME

USUAL LOT SIZE 2500 SAMPLE SIZE 100 A.O.Q.L. 1.5 % PROCESS AVERAGE .709 %

REMARKS Reissued to place in proper process average class.

NOTES BY STATISTICAL DEPT.

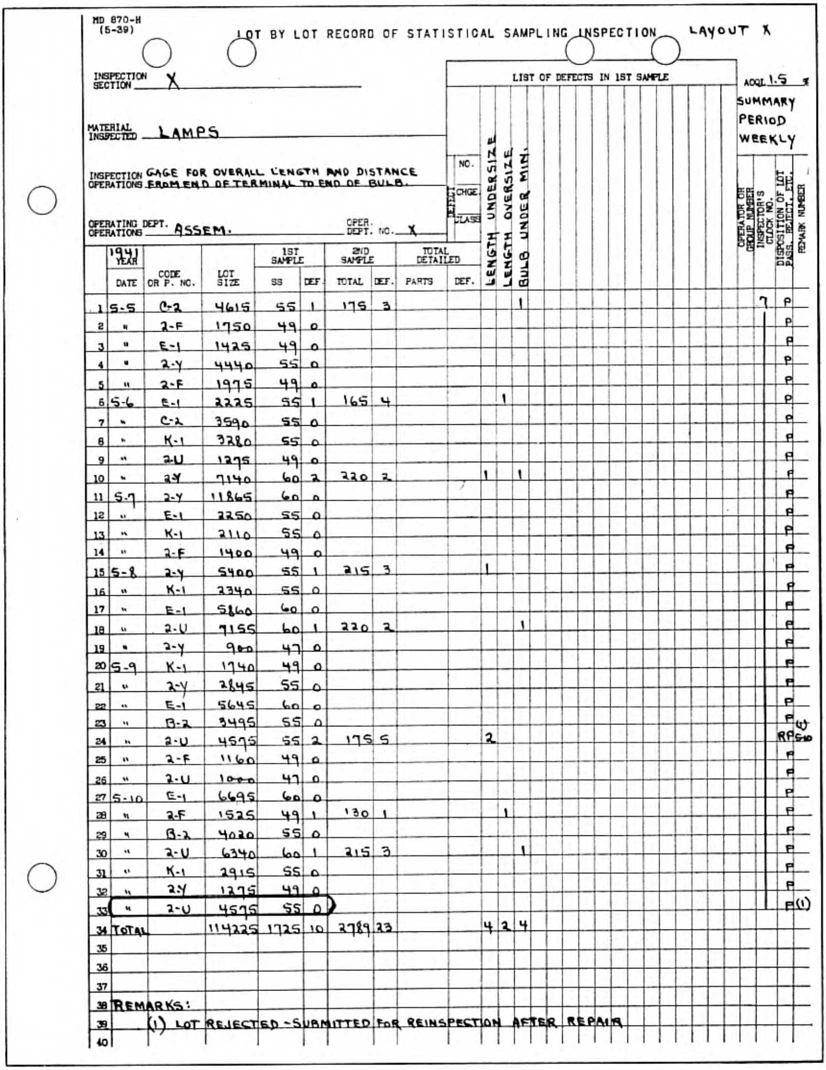
CHECKED BY X

DATE X

Fig. 2

1.4 *Records*

This section contains a description of the running records and clerical operations which are used for supervisory control. The records are "Lot by Lot Record of Statistical Sampling Inspection", and "Summary of Results of Inspection", shown in Figs. 3 and 4, respectively.



MD 870-H
(5-39)

LOT BY LOT RECORD OF STATISTICAL SAMPLING INSPECTION LAYOUT X

INSPECTION SECTION X

MATERIAL INSPECTED LAMPS

INSPECTION OPERATIONS GAGE FOR OVERALL LENGTH AND DISTANCE FROM END OF TERMINAL TO END OF BULB.

OPERATING DEPT. OPERATIONS ASSEM. OPER. DEPT. NO. X

LIST OF DEFECTS IN 1ST SAMPLE

DEFECT NO. CHGE. CLASS

AOQL 1.5 %

SUMMARY PERIOD WEEKLY

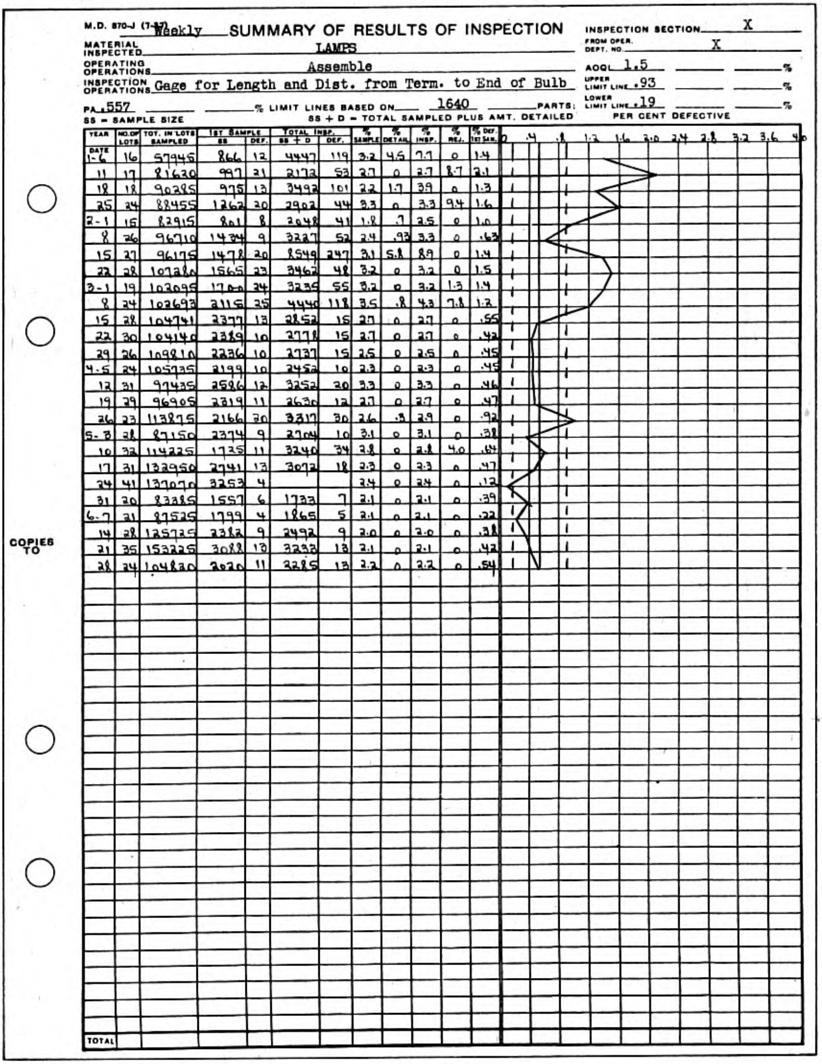
	1941 YEAR			1ST SAMPLE		2ND SAMPLE		TOTAL DETAILED								
	DATE	CODE OR P. NO.	LOT SIZE	SS	DEF.	TOTAL	DEF.	PARTS	DEF.	LENGTH UNDERSIZE	LENGTH OVERSIZE	BULB UNDER MIN.	OPERATOR OR GROUP NUMBER	INSPECTOR'S CLOCK NO.	DISPOSITION OF LOT PASS. REJECT. ETC.	REMARK NUMBER
1	5-5	C-2	4615	55	1	175	3					1		7	P	
2	"	2-F	1750	49	0										P	
3	"	E-1	1425	49	0										P	
4	"	2-Y	4440	55	0										P	
5	"	2-F	1975	49	0										P	
6	5-6	E-1	2225	55	1	165	4				1				P	
7	"	C-2	3590	55	0										P	
8	"	K-1	3280	55	0										P	
9	"	2-U	1275	49	0										P	
10	"	2-Y	7140	60	2	220	2			1		1			P	
11	5-7	2-Y	11865	60	0										P	
12	"	E-1	2250	55	0										P	
13	"	K-1	2110	55	0										P	
14	"	2-F	1400	49	0										P	
15	5-8	2-Y	5400	55	1	215	3			1					P	
16	"	K-1	2340	55	0										P	
17	"	E-1	5860	60	0										P	
18	"	2-U	7155	60	1	220	2					1			P	
19	"	2-Y	900	47	0										P	
20	5-9	K-1	1740	49	0										P	
21	"	2-Y	2845	55	0										P	
22	"	E-1	5645	60	0										P	
23	"	B-2	3495	55	0										P	
24	"	2-U	4575	55	2	175	5			2					RP	(1) 5-10
25	"	2-F	1160	49	0										P	
26	"	2-U	1000	47	0										P	
27	5-10	E-1	6695	60	0										P	
28	"	2-F	1525	49	1	130	1				1				P	
29	"	B-2	4020	55	0										P	
30	"	2-U	6340	60	1	215	3					1			P	
31	"	K-1	2915	55	0										P	
32	"	2-Y	1275	49	0										P	
33	"	2-U	4575	55	0										P	(1)
34	TOTAL		114225	1725	10	2789	23			4	2	4				
35																
36																
37																

38 REMARKS:
39 (1) LOT REJECTED - SUBMITTED FOR REINSPECTION AFTER REPAIR
40

Fig. 3

The Lot by Lot Record of Statistical Sampling Inspection—Fig. 3

On this form the inspector records the results of his inspection at the time of his observation. An examination of Fig. 3 will indicate how the form is filled out.



M.D. 870-J (7-37) Weekly SUMMARY OF RESULTS OF INSPECTION

INSPECTION SECTION X

MATERIAL INSPECTED LAMPS

FROM OPER. DEPT. NO. X

OPERATING OPERATIONS Assemble

AOQL 1.5 %

INSPECTION OPERATIONS Gage for Length and Dist. from Term. to End of Bulb

UPPER LIMIT LINE .93 %

PA .557 % LIMIT LINES BASED ON 1640 PARTS;

LOWER LIMIT LINE .19 %

SS = SAMPLE SIZE

SS + D = TOTAL SAMPLED PLUS AMT. DETAILED

PER CENT DEFECTIVE

YEAR	NO. OF LOTS	TOT. IN LOTS SAMPLED	1ST SAMPLE		TOTAL INSP.		% SAMPLE	% DETAIL	% INSP.	% REJ.	% DEF. 1ST SAM.
			SS	DEF.	SS + D	DEF.					
DATE 1-6	16	57945	866	12	4447	119	3.2	4.5	7.7	0	1.4
11	17	81630	997	21	2172	53	2.7	0	2.7	8.7	2.1
18	18	90285	975	13	3492	101	2.2	1.7	3.9	0	1.3
25	24	88455	1262	20	2902	44	3.3	0	3.3	9.4	1.6
2-1	15	82915	801	8	2048	41	1.8	.7	2.5	0	1.0
8	26	96710	1434	9	3227	52	2.4	.93	3.3	0	.63
15	27	96175	1478	20	8549	247	3.1	5.8	8.9	0	1.4
22	28	107280	1565	23	3462	48	3.2	0	3.2	0	1.5
3-1	19	102095	1700	24	3235	55	3.2	0	3.2	1.3	1.4
8	24	102693	2115	25	4440	118	3.5	.8	4.3	7.8	1.2
15	28	104741	2377	13	2852	15	2.7	0	2.7	0	.55
22	30	104140	2389	10	2778	15	2.7	0	2.7	0	.42
29	26	109810	2236	10	2737	15	2.5	0	2.5	0	.45
4-5	24	105735	2199	10	2452	10	2.3	0	2.3	0	.45
12	31	97435	2586	12	3252	20	3.3	0	3.3	0	.46
19	29	96905	2319	11	2630	12	2.7	0	2.7	0	.47
26	23	113875	2166	20	3317	30	2.6	.3	2.9	0	.92
5-3	28	87150	2374	9	2704	10	3.1	0	3.1	0	.38
10	32	114225	1725	11	3240	34	2.8	0	2.8	4.0	.64
17	31	132950	2741	13	3072	18	2.3	0	2.3	0	.47
24	41	137070	3253	4			2.4	0	2.4	0	.12
31	20	83385	1557	6	1733	7	2.1	0	2.1	0	.39
6-7	21	87525	1799	4	1865	5	2.1	0	2.1	0	.22
14	28	125725	2382	9	2492	9	2.0	0	2.0	0	.38
21	35	153225	3088	13	3233	13	2.1	0	2.1	0	.42
28	24	104820	2020	11	3285	13	2.2	0	2.2	0	.54
TOTAL											

COPIES TO

Fig. 4

The Summary of Results of Inspection—Fig. 4

This form is kept with the inspection layout and the Lot by Lot Records. Entries are made as indicated on Fig. 4. Clerical operations involved in making the necessary computations are explained in the following paragraphs.

1.41 *PA* (*Process Average*)

The Process Average should be computed at least once every six months, and more frequently when conditions warrant. Only data accumulated since the last computation should be used. When it is known that the quality of the product has changed significantly during the period, use only the data collected since the change.

Record the results of all process average computations and the periods covered by them in the space provided on the sampling layout.

1.42 *Periodic Totaling of Lot by Lot Record and Posting on Summary of Results Form, Fig. 4*

The following columns on the Lot by Lot Record should be totaled at intervals, and may, for contrast, be marked in red:

No. of Lots	*Lot Size*	*1st Sample*		*2nd Sample Total*	*List of Defects in 1st Sample Total by Columns*
		SS	*Def.*		
X	X	X	X	X	XXXX etc.

Totals after approximately 20 entries are generally considered satisfactory. The required totals are then posted to the Summary Form, Fig. 4.

1.43 *Per Cent Defective*

The purpose of this figure is to show the average quality of the product as received by the inspector during the period covered and is obtained by dividing the total number of defectives found in first samples by the total number of articles inspected in first samples, and multiplying by 100. Only the results of the first samples are used, in order to accord equitable treatment to all lots.

1.44 *Control Chart*

The per cent defective is plotted on the graph at the right of the summary form (Fig. 4). Control Limit Lines are drawn around the process average to indicate the variation that may be expected due to random sampling. These control limits are determined by the following formulae[3]:

$$\left.\begin{array}{l}\text{Upper Control Limit}\\ \text{for fraction defective}\end{array}\right\} = \bar{\text{p}} + 2\sqrt{\frac{\bar{\text{p}}\,(1-\bar{\text{p}})}{\text{n}}}$$

$$\left.\begin{array}{l}\text{Lower Control Limit}^{4}\\ \text{for fraction defective}\end{array}\right\} = \bar{\text{p}} - 2\sqrt{\frac{\bar{\text{p}}\,(1-\bar{\text{p}})}{\text{n}}}$$

---

[3] The considerations involved in the establishment of control chart limits are discussed in ASA Standards Z1.1–1941 and Z1.2–1941, "Guide for Quality Control and Control Chart Method of Analyzing Data". In this case, 2-sigma limits have been chosen. For this particular application within the manufacturing plant over a period of years, this choice has appeared to strike an economic balance with respect to the net consequences of two kinds of "errors" that may occur in practice; namely, looking for trouble that does not exist and not looking for trouble that does exist.

[4] If this result is negative, the lower control limit is to be taken as zero.

Where $\bar{p}$ = Process average fraction defective

n = Average number of articles in first samples inspected during the summary period.

For example, the records for a typical period of 10 weeks might show the following:

Total number of articles inspected in first samples = 16,400
Process average fraction defective = $\bar{p}$ = .00557 or .56%

$$\text{Thus, } n = \frac{16{,}400}{10} = 1640 \text{ articles}$$

Hence:

$$\text{Upper Control Limit} = .00557 + 2\sqrt{\frac{.00557\,(1 - .00557)}{1640}}$$

" " " $= .00557 + .00368$

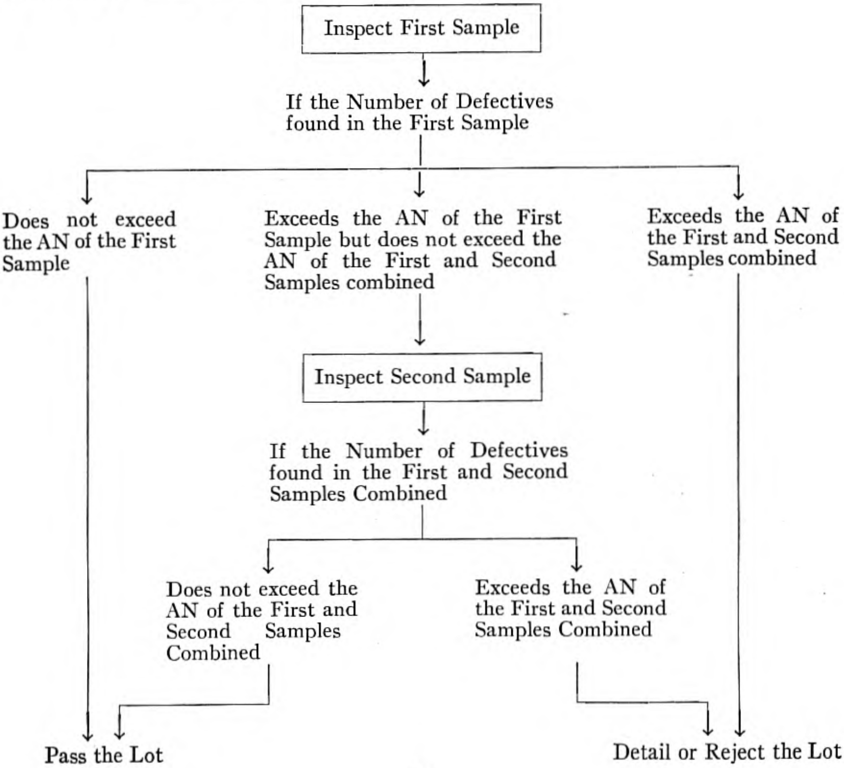
" " " $= .00925 = .93\%$

Lower Control Limit $= .00557 - .00368$

" " " $= .00189 = .19\%$

2. APPLICATION OF THE DOUBLE SAMPLING PLAN BY THE INSPECTOR

The following flow chart illustrates in sequence the basic steps involved in the inspection of a lot.



An example of the operation of this chart is shown below:

Example:
AOQL = 1.5%
Process Average = .56%
Lot Size = 4615

On consulting the sampling table on the layout, which is information obtained from the table for AOQL = 1.5%, Process Average Column .31–.60% (see Fig. 1), it will be found that for a lot of 4001–5000 parts the following sample sizes (SS) and allowable numbers of defectives (AN) are shown:

Lot Size	*First Sample* *SS*	*AN*	*Add.*	*Total Sample* *Total*	*AN*
4001–5000	55	0	120	175	4

This means that for the first sample 55 parts should be selected and completely inspected for all items covered by the Sampling Layout. In order that the per cent defective may be accurately determined for control purposes, a *complete inspection* of the first sample must be made regardless of how defective it may be. If five or more defectives are found in the sample of 55 parts, the lot should be rejected or detail inspected, or otherwise disposed of as shown on the layout.

If no defectives are found in the 55 parts, the lot should be passed.

If one, two, three, or four defectives are found in the 55 parts, the second sample of 120 additional parts should be selected.

In the combined sample of 175 parts, a total of four defectives is allowed. if a total of five defectives is found before all of the 175 parts are inspected, sampling should be discontinued and the lot disposed of as indicated on the layout.

If less than five defectives are found in the combined sample of 175 parts, the lot should be passed.

2.1 *Counting the Lot Size*

The determination of the lot size may be made by weighing methods or careful estimates. It has been found in practice that an estimate which is within 20% of the true value of lot size is satisfactory for sampling purposes.

2.2 *Counting the Sample Size*

Since the relation between corresponding Lot Sizes and Sample Sizes for a particular AOQL is not linear the same order of accuracy of count does not prevail. If, in using a given sampling table, a sample smaller than that prescribed is taken, the result is to increase the AOQL, and if a sample larger than that prescribed is taken, the result is to decrease the AOQL and increase the cost of inspection.

Counting becomes a simple matter when articles are handled in compartment boxes or when the gage or testing device provides for automatic count Regardless of what method of counting is used there should be agreement between the size of sample selected and sample size indicated in the sampling table.

2.3 *Reinspection of Rejected Lots*

In order to guarantee the protection promised by a particular Double Sampling Plan it is required that a rejected lot be completely cleared of all defects. The rejected material should be repaired and returned for reinspection as one lot. It should then be reinspected for all inspection items using the sample size that would normally apply to a new lot of the same size. If a defect is found during reinspection it is evident that proper repair or sorting has not been accomplished. Such a lot should, of course. not be considered acceptable until all defects are removed.

In the Lot-by-Lot Record, the results of resampling of rejected lots are recorded on a separate line and the entry circled or otherwise identified so that it will not be included with the results of first samples on the Summary Form.

3. SUPERVISION OF THE DOUBLE SAMPLING PLAN

In order that maximum advantage may be gained from the use of Double Sampling Lot Inspection it is necessary that attention be given to the topics listed below. Upon proper attention to these factors depends the effectiveness of the Plan.

3.1 *Changes in AOQL*

There is no assurance that the value originally selected for an AOQL will continue to be the most satisfactory in view of changing factors relating to the product, such as:

3.11 Changes in design of the product or changes in the requirements for inspection items may increase or decrease the trouble caused by the acceptance of defective parts and therefore will occasion a review of the AOQL value.

3.12 New methods of manufacture that change the difficulty and therefore the cost of making a product will necessitate a reconsideration of the AOQL value.

3.13 An excessive number of complaints or other reports of difficulty from succeeding stages of manufacture or from customers may indicate too large an AOQL.

3.14 If no significant changes such as those mentioned above exist and there are repeated lot rejections, it would appear that quality is unsatisfactory. However, it should be determined whether or not succeeding operations can possibly tolerate more defectives than they are actually receiving and, if this is the case, the AOQL may profitably be increased.

3.2 *Changes in Process Average*

The economy of the sampling plan tends to decrease when the level of quality of the product shifts outside the range of process average on which the layout was based. The sampling table on the layout should ordinarily be changed whenever the process average shifts from one range to another. However, before reissuing the layout, the reason for the shift should be determined. If the process average has been reduced, an attempt should be made to make the change permanent; or if it has been increased, the cause should be eliminated.

3.3 *Interpretation of the Control Chart on "Summary of Results of Inspection" Form*

It is very important to review the Control Chart of per cent defective (Fig. 4) frequently as it is an index of the success of the sampling plan. Not all fluctuations of the per cent defective are significant. Even though the quality of the product is controlled, the results of sampling inspection may produce fluctuations in the indicated per cent defective. These variations are measured by a simple control chart on the "Summary" form which employs control limits for values of per cent defective. As long as the plotted points of the per cent defective remain within control limits the fluctuations are no greater than may reasonably be expected from a uniform manufacturing process. However, if a point goes above the upper control limit, the cause may be defective workmanship, defective raw materials, or even a change in the severity of inspection. If a point goes below the lower control limit, the cause may be an improvement in quality or a change in the definition of a defect through misinterpretation or changes in inspection equipment or method of check. If the curve of the plotted points hugs either limit line or shows a definite trend toward one side or the other, a significant change in the quality of the product is indicated.

In order to achieve control of the quality of a manufactured product, direct and immediate action must be taken in order to stem unfavorable trends. The presence or absence of a satisfactory quality level may be detected by means of inspection, but such a level can only be originated and maintained by adequate manufacturing methods and equipment in the hands of a quality-minded producing personnel.

3.4 *Changes in the Definition of a Defect*

Either laxity or severity of inspection may cause the "reported" per cent defective to show significant changes even though the actual per cent defective is unchanged. This may result from a change in the definition of a defect; that is, the same condition may at times be considered defective and at other times acceptable. This happens most frequently in border-line cases. To avoid such variations it is necessary that the condition that constitutes a defect be clearly defined and strictly followed in all inspections.

3.5 *The Abnormal Existence of One Kind of Defect*

When the sampling scheme includes the inspection for several different requirements, the Allowable Number may at times be exceeded because of one kind of defect only. In other words, the lot would be satisfactory if this one kind of defect did not exist. If the same defect persists for several lots, it should receive definite supervisory attention. If substantial improvement is not feasible, it may be convenient to remove the inspection item to a separate sampling layout.

3.6 *Abnormal Distribution of Defectives*

Occasionally there may be reason to believe that a group of parts submitted for inspection is not uniform in quality throughout, that is, not a true lot as defined in the early part of the paper. Such a group should be divided into homogeneous sections and each section sampled separately. However, when this happens, subsequent lots for sampling purposes should be similarly subdivided, that is, they should be confined to the output of one source at a time; for example, one machine, one operator, or one batch of material, etc., based on one system of causes, so that control of quality at each source may be applied, and the consumer protected from receiving spotty product.

In addition to the specific steps followed in establishing and operating a sampling plan it must always be remembered that, since relatively important decisions concerning the acceptance of product hinge upon the results of an examination of a small group of parts, inspection must be conscientious and accurate. In order to guarantee the order of protection promised by the sampling plan, the results of inspection and the prescribed procedure for disposing of individual lots must be regarded with thorough respect.

Entropy

By KARL K. DARROW

MANY of the important ideas of physics are of such recent birth that if they still seem hard to grasp, it may be contended that the world has not yet had the time to assimilate them. Of entropy this cannot be said; with its centenary almost upon us, entropy is ancient compared with most of the concepts which baffle the student today. Yet an aura of mystery seems to envelop it still, and two other things inseparably joined to it: the scale of temperature called absolute, and the Second Law of Thermodynamics. One is driven to wonder why the three of them resist the understanding so tenaciously, and certain reasons are not hard to find. Thus, to speak of the three as "joined together" is too weak, for it implies that they can at least be taken in a certain order for didactic purposes, the student ascending from the one to the next and finally to the last. Actually they are much too tightly interlocked for this, a sort of trinity one and indivisible, which must be apprehended as a whole if ever to be properly grasped at all. Again, the Second Law has been expressed in many different ways, and it is one of the oddest things in science to see how various authorities can claim that the law is obeyed absolutely without exception, while they themselves cannot agree how to state it. A further cause of trouble lies in the unlucky boundary between chemistry and physics, which nowhere harms these sciences more than in the study of entropy. Like the worst of the old-fashioned boundaries of Europe, it wanders capriciously across the natural lines of intercourse and trade, cutting off the traditional chemist from the origin and development of some of his most valuable ideas, cutting off the traditional physicist from some of the finest verifications of the thought of his fore-runners. It will be the principal object of this paper to dwell on these verifications, abolishing the barrier so far as may be feasible.

First to illustrate how the three ideas are tangled up together, I give one of the most useful of the definitions of entropy: *when a system passes in a reversible way from one to another state, at a constant absolute temperature T, its change of entropy ΔS is equal to the heat Q which it absorbs, divided by T.* That both the entropy and the absolute temperature have slipped into this definition is all too obvious; as for the Second Law, it lurks underneath the phrase "in a reversible way." Quite evidently we are invited to master all the three in a single mental operation, but quite as evidently this is impossible. A breach must somehow be made in this hard and unified surface, and the best of scaling-ladders appears to be the ideal gas, as everybody has used it for the purpose.

We therefore receive the definition of the "ideal gas": a substance whereof (*a*) the temperature remains unchanged when the gas is given and seizes the chance of expanding without any hindrance into a vacuous space; and (*b*) the pressure varies inversely as the volume, so long as the temperature remains unchanged. Now rather than being an aid to enter the citadel, this definition appears to presume that we are in the citadel already, since the word of "temperature" is all too prominent in it. But the word comes up only in the phrase ". . . temperature remains unchanged . . ."; so nothing more is implied, than that the onlooker knows how to recognize whether temperature is changing or staying the same. This ability can be his, whether or not he knows about the scale called absolute; and so the definition implies nothing about the scale.

The definition would be an idle collection of words, were there not actual gases conforming to it so nearly, that at least for a time they may safely be taken to conform to it exactly. All gases in fact approach conformity, as the density lessens; and with helium and hydrogen especially, the approach is already close while the density is still so high that there is no trouble at all in using them as thermometers. Now, to use an ideal gas as a thermometer means simply this: P being the pressure of the gas and V its volume and n the quantity of the gas measured in moles, there is the equation,

$$PV = nRT \tag{1}$$

which is a mixture of a new definition and a theorem. The part which is definition is, that the temperature called absolute is taken to be proportional to the product PV. The part which is theorem is, that if at the same temperature we compare the product PV for equal numbers of moles of various gases, its value is the same. What is denoted by R in equation (1) is a universal constant. Its value is 1.985 in the customary units, which are calories per degree per mole. Since now the degree of temperature has appeared in this discourse, I recall that its value is fixed by the convention that there are one hundred degrees from the freezing to the boiling point of water; but on this I need not dwell.

It is tacitly assumed that nothing in this definition will ever clash with the threefold simultaneous definition of temperature and entropy and Second Law, should we ever arrive at it. Taking this for granted, and taking advantage of the breach which has thus been made in the formerly impenetrable surface of our definition of entropy, we now attack the concepts of "state" and "reversible way."

The usual way of dealing with "state" is to say that it is determined completely by any two of the three variables P, V, and T, which are interconnected among themselves by what is called an "equation of state." This takes for granted that one knows what kind of substance one is speaking of, how many moles there are of it, and which phase it is in; for the statement is valid not only for gases but also for liquids, and even (under certain restrictions) for solids. If two or three phases are coexisting, precautions must be taken. The statement is not valid for mixtures of substances which differ; but there is doubt as to just how much it takes to constitute a difference, and isotopes of a single element are usually considered to be the same.

Now let us envisage two states which differ in temperature but are the same in pressure. P is to stand for the common value of pressure, T_0 for the lower and T_1 for the higher of the temperatures. For the utmost in simplicity at the beginning, the substance shall be an ideal gas, and moreover a gas of which the specific heats do not depend upon the temperature; any noble gas under ordinary conditions approaches very closely to these stipulations.

One may operate the transition from (P,T_0) to (P,T_1) in the manner which follows: Let the gas be heated from T_0 to T_1 while the volume of its container remains the same. Its pressure will have gone up meanwhile; let this situation be remedied by piercing a hole through the container wall, into an evacuated chamber prepared in advance of just the right dimensions, so that after the "free expansion" the pressure of the gas will be down again to P. The temperature will still be T_1, for that is one of the qualities of the ideal gas.

Or one may operate the transition in the manner which follows: Let the container be made in advance with a movable wall, say a piston-head against which an outer pressure P—it might be that of the atmosphere—is steadily pressing. Let the gas be gradually heated: as its temperature rises, the piston-head glides gradually outward, increasing the volume of the container at such a rate that the pressure of the gas always remains the same and equal to P. The process is to be stopped when the temperature reaches T_1.

These, then, are two very different ways of carrying through the transition; and the reason for bringing them in is to elucidate the words "in a reversible way" which occur in the phrase defining entropy. The former way includes a process which is spontaneous, rapid, turbulent, explosive; nothing can inhibit it when the piercing of the hole creates the opportunity. The latter way is gradual, languid, crawling, always on the verge of stopping, would instantly stop if the inflow of heat should cease, would reverse its trend if the inflow should be turned into outflow. The former way includes a process during which the pressure of the gas is ill-defined or not defined at all, for the gas is spurting through the hole and reverberating in the chamber formerly empty. During the entire latter process, if it be gradual enough, the pressure of the gas is definite, and related to V and T by the equation (1). The former is an irreversible way of effecting the transition. The latter is *the reversible way* of effecting the transition.

As the heat flows into the gas, its entropy mounts up. Since the temperature is meanwhile changing, we must mentally subdivide the incoming heat into driblets dQ, during the inflow of each of which T remains about the same. The corresponding driblets of energy are given by dQ/T, the total change in entropy due to inflow of heat is given by $\int dQ/T$ integrated from beginning to end of the inflow.[1] By the former way the integral is $\int (C_v/T)dT$ or $C_v \ln(T_1/T_0)$, by the latter way the integral is $\int (C_p/T)dT$ or $C_p \ln(T_1/T_0)$. The two are not the same. Now we see why the phrase "in a reversible way" was necessary in the definition of change-of-entropy. But for that phrase or something similar, we should now have no definition. But having accepted the phrase, we are invited and required to write,

$$\Delta S = \int (C_p/T)\, dT \tag{2}$$

for all transitions of a single substance within a single phase; ΔS signifying the change of entropy, and the use of C_p implying that the conditions are those under which specific heat at constant pressure is properly measured: the pressure of the substance being definite, nothing turbulent or gusty or explosive happening within the substance, and an equal pressure bearing down upon it from the outer world.

We have now the necessary and complete statement for the variation of entropy with temperature, pressure remaining the same; but it has to be supplemented with a statement for the dependence of entropy on pressure, T remaining the same. For this and other purposes, let us return to the irreversible passage between (P,T_0) and (P,T_1)—the passage, rather, of which one stage is reversible but the other not. During the reversible stage, the gain of entropy is $\int (C_v/T)dT$. This falls short of the gain of entropy incurred along the other route as given by (2). However the other route leads reversibly all the way to the goal, while the reversible part of this one brings us only a part of the way, leaving us with the irreversible expansion still before us. By assuming that the remainder of the total gain of entropy demanded by (2) is made up during the irreversible expansion, we rescue the concept of entropy. Now to save and establish the concept

---

[1] It is often said or implied that this formula should not be used unless the process is fully reversible, in the sense that the inflow of heat occurs from a reservoir of temperature identical with that of the gas. As no such precaution is taken when specific heats are measured, and as measurements of specific heat are commonly used for establishing values of entropy, I assume that the limitation is needless.

of entropy, to make it a definite property of the state of a substance, is one of the offices of the Second Law.

It is therefore a part and a consequence of the Second Law to affirm that when an ideal gas undergoes a free expansion, it experiences a gain of entropy despite the fact that it receives no heat from the world without! Nor is the affirmation confined to ideal gases; it would be true of any substance, though in general a free expansion would be attended with a change in temperature. Nothing, therefore, could be more wrong than to repeat our first definition of change-of-entropy with the words "in a reversible way" left out.

One now begins to see why the concept of entropy is so much harder to receive than that of energy. Every scientist is accustomed by now to the "conservation" of energy: whereby it is meant that if the energy of a system rises or falls by any amount, it is because there has been an inflow or an efflux of just that amount from or into the outer world. Nothing of the sort can be said of entropy, of which we have just seen that it may vary even within a system which is having no transactions at all with the outer world. One may not speak of the conservation of entropy excepting under the sharp and severe restriction that all of the processes in the system *and* in the outside world are reversible: and "reversible" must be used in the full sense adumbrated in a previous footnote, whereby no transfer of heat is reversible unless the body whence it comes and the body to which it goes are of identical temperature. Yet nothing of all this contradicts the assertion that entropy and energy are, both of them, uniquely determined functions of the state of a system—functions, therefore, of any two of the three variables P and V and T, for substances of a single kind in a single phase.

Since the equating of entropy-change to inflow of heat divided by temperature is something that positively must not be done for an irreversible process, we must seek other ways to assess it.

One such way has already been stated. If it is possible to start from a particular state, and thence to arrive by reversible ways at both the beginning and the end of the irreversible process in question, all the necessary knowledge is at hand; for now by integrating dQ/T along the two ways we find two quantities, of which the difference is the entropy-change desired.

Applied in the special case which we have been considering, this method has given

$$\Delta S = (C_p - C_v)\, ln\, (T_1/T_0) \tag{3}$$

not, however, a useful expression as yet, since it contains a quantity (T_0) which does not figure at all in the irreversible process in question—a process which, I recall, is the free expansion of one mole of ideal gas, at a constant temperature T_1, from a higher pressure and lower volume which we will denote by P' and V', to a lower pressure and a higher volume which we will denote by P and V. Remembering (1) the equation of state of the ideal gas, and applying it to the reversible heating at constant volume which preceded this expansion, we find

$$T_1/T_0 = P'/P \tag{4}$$

and so,

$$\Delta S = (C_p - C_v)\, ln\, (P'/P) \tag{5}$$

for the entropy-change incurred when one mole of ideal gas expands at uniform temperature so that its pressure falls from P' to P. I do not have to state the temperature, since it has vanished from the equation. I do not have to say that the expansion is irreversible, for if there be such a thing as entropy at all, its alteration depends only upon its values in the initial and in the final state, and not on the manner in which the system has made the passage from initial to final. I ought, however, to recall that we are still supposing a gas of which the specific heats do not depend upon temperature, nor in fact upon any variable whatsoever.

Now we have an expression—equation (5)—for the way in which entropy changes with pressure at constant temperature, and another—equation (2)—for the way in which entropy changes with temperature at constant pressure. We may combine them to get the change of entropy occurring when the gas proceeds, by whatever route, from an initial state (P_1, T_1) to a final state (P_2, T_2). This is,

$$S_2 - S_1 = C_p\, ln(T_2/T_1) - (C_p - C_v)\, ln(P_2/P_1) \tag{6}$$

which, by stripping off the subscripts 2, and gathering into one term I all of the terms containing the subscripts 1, may be written,

$$S = C_p\, lnT - (C_p - C_v)\, lnP + I \tag{7}$$

the expression, for a gas of the special type stated, of entropy as a function of absolute temperature and of pressure. The quantity I is an "additive constant" which will turn out to be one of the major topics of this paper.

All this way we have come without invoking the First Law of thermodynamics, and we might even go further without its aid! But there is no point in deferring it longer, and I wish to be able to convert (7) into a more familiar form by replacing $(C_p - C_v)$ with the constant R of the equation-of-state of the ideal gas. To do this I return to the irreversible process which has so long engaged our attention; the free expansion of the gas from the higher pressure P' to the lower pressure P, its temperature remaining the same. I seek a reversible way of conducting the expansion from the same beginning to the same ending under the same condition of steady temperature. A reversible way of going between the extremes we have indeed already found, but it involves a retracing of our steps: the cooling of the gas to a lower temperature at a constant volume, followed by the heating of the gas to the original temperature at the constant pressure. We wish a forward-going reversible way, and such a one can be found. It is necessary to have the container built with a frictionless piston-head for one wall, and to bear down upon this from without with a pressure always nicely calculated to be equal to the internal pressure of the gas. If under this condition the piston-head is gliding slowly outward it will continue so to glide, and the gas will expand in the gradual, languid, crawling manner, with its internal pressure always definite—the manner which we call "reversible." All that is now required is to know the amount of heat which enters the gas during this process, so that we may divide it by T and so assess the entropy-change. It is from the First Law that we get this information.

The First Law is to be spoken in the form

energy-gain = *"heat in"* less *"work out"*

and written in the form

$$\Delta U = \Delta Q - \Delta W \tag{8}$$

the symbols fitting the words in the way which is obvious.

For an ideal gas, the energy U is independent of pressure or volume, depending on the temperature alone. The reader may or may not take this as a matter of course, but it will be proved later on. We are considering an isothermal expansion, and therefore ΔU is zero, and the problem of evaluating ΔQ is that of evaluating ΔW. Now the "work out"—the work done by the gas upon the outer world—is equal to the pressure bearing down upon the piston-head from the outer world, multiplied by the area of the piston-head and the distance through which it advances. The last two factors multiplied together give the gain in volume of the gas, ΔV; and without a moment's hesitation one usually puts for the first factor the symbol P signifying the pressure of the gas. However, it is wise to hesitate just long enough to realize that by so doing one assumes the *reversible* expansion with all the attributes set forth above. For an irreversible expansion ΔW would not be equal to $P\Delta V$, but less. But assuming the reversible expansion, and remembering that P is not independent of volume as in these last few lines I have tacitly assumed, we find

$$\begin{aligned} \Delta W &= \int P\,dV = RT \int (1/V)\,dV \\ &= RT\,ln\,(V/V') = RT\,ln\,(P'/P) \end{aligned} \tag{9}$$

the unprimed symbols referring to the final state of the expanding gas and the primed symbols to the initial state. Now identifying ΔW with ΔQ, and dividing it by T to get the change-of-entropy, we have for the very quantity ΔS which figured in equation (5) a second expression, *viz.*

$$\Delta S = R\, ln(P'/P) \tag{10}$$

where now R stands in place of $(C_p - C_v)$. We make the substitution into (7), and arrive at our next-to-final form for the entropy of the ideal gas of constant specific heats:

$$S = C_p ln T - R\, ln P + I \text{ (one mole)} \tag{11}$$

I pause to mention that since for any gas PV/T is measurable and so is also $(C_p - C_v)$, the rightness of our assumptions may be tested by ascertaining whether for gases nearly ideal, the one—which for an ideal gas is R—and the other are nearly equal. This is so; and if we wish to pick out a special one among the assumptions for which this shall constitute a special test, then we may set down as having been particularly proved the assertion, that the energy of an ideal gas depends upon its temperature alone and not upon its pressure nor its volume.

For quite some time I have been referring to a single mole of the gas in question; but for an odd and probably an unexpected reason, it is going to be desirable to make an explicit broadening of these equations to the general case of any number n of moles. The broadening required for (11) is so simple and trite as to seem not worth the doing: we have simply to multiply every term of (11) by n, and so obtain

$$S = nC_p ln T - nR\, ln P + nI \text{ } (n \text{ moles}) \tag{12}$$

But now let us translate this into an expression for S as a function of volume and temperature, by use of the fact that P is nRT/V, and the further fact that R is $C_p - C_v$. We come upon the astonishing equation,

$$S = nC_v ln T + nR\, ln V - nR\, ln(nR) + nI \tag{13}$$

astonishing because the terms to the right of $nR\, ln V$ are not reducible to n multiplied into a constant, but involve a more intricate function of n. Perhaps, after all, we ought not to have taken the simple recourse of broadening (12) by multiplying n into every term of (11) including the last one? Actually it is quite all right; the additive constant in (12) is truly proportionate to the number n of moles; that in (13) is the more intricate function of n which we have just derived.

In Fig. 1 there is shown a "phase-diagram" appropriate to a substance of a single kind, capable of existing as a gas and as a liquid and as a solid. Pressure and temperature are the coordinates along the vertical and horizontal axes, but no coordinate-scales are shown, and no attempt has been made to shape the curves in exactly the manner correct for any particular substance, since all that matters here is the general idea. Envisage any point (P, T) in the area "gas," lying just off the curve which separates that area and the other one marked "solid." This is the point—I will call it the "point of interest"—for which we are to obtain two expressions for entropy S arising from different sources, and find an important result by comparing the two so obtained.

One of the two is of course the right-hand member of (12). It may create surprise that one should be treating the gas as ideal, under conditions where the slightest fall in temperature or rise in pressure would condense it. The approximation, however, may still be a good one, and if it is not close, the equation of state of the actual gas may be used in place of (1).

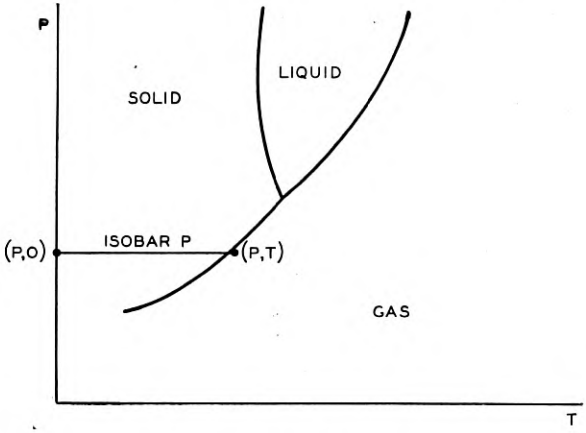


Fig. 1

To form the other expression, we commence at the point $(P, 0)$ where the isobar P which traverses the point of interest reaches the vertical axis, and call the entropy there by the symbol $S(P, 0)$. We proceed along the isobar toward the point of interest, building up the integral $\int (C_p/T)dT$; since we remain in the area called "solid," it is C_p of the solid which concerns us, and we may mark it so. Just before the isobar passes over the curve dividing solid from gas, the entropy arrives at the value

$$S(P, 0) + \int_o^T (C_p^{sol}/T)\, dT$$

but this is not yet the value at the point of interest, for the "divide" is yet to be crossed. At the crossing of the divide a certain amount of heat is absorbed: denote it by L for one mole. If the crossing of the divide is a reversible process, the entropy of the substance goes up by L/T as it turns from solid to gas.

Is the crossing of the divide a reversible process? It is indeed, and that is why phase-diagrams are constructed. Think of the substance enclosed in a container with frictionless piston-head, equal pressures P bearing upon the piston from the world without and from the substance within, and the whole affair at temperature T. If the point defined by P and T lies exactly on one of the curves which partition the diagram off, then the substance within the container may be a mixture of the two adjoining phases in any ratio imaginable. If the piston-head is gliding slowly inward it continues so to glide, P and T remaining steadily the same while all that there is of the phase of greater specific volume[2] is converted gradually into the other. If the piston-head is gliding slowly outward it continues so to glide, P and T remaining constant while the latter phase is converted into the former. All of the attributes of the reversible process are here: gradualness, lack of turbulence and of explosiveness, willingness to go in either sense, willingness to stand indefinitely still at any partway stage. For ease of formulation it is the prince of reversible processes, since P and T are both unchanging while it is going on.[3]

Accepting then the crossing of the divide as a reversible process, we have our second expression for the entropy at the point of interest:

$$S(P, T) = S(P, 0) + \int_0^T (C_p^{sol}/T)\, dT + L/T \tag{14}$$

I pause to quiet the fear that the expression here written down is necessarily infinite, because of T standing in the denominator of the integrand on the right! It is a fact of experience that as T approaches absolute zero, C_p approaches zero for all solids, and with such rapidity that (if extrapolation is safe) the integrand tends to zero and not to infinity. Now comes the climax, which consists in equating the right-hand members of (12) and (14). Referring to a single mole of the substance, and rearranging the terms, we find:

---

[2] Volume per unit mass.

[3] Irreversible transitions from "supercooled" liquid to solid may occur. In such a case the area called "liquid" may be regarded as spreading somewhat over the divide and overlapping a portion of the area called "solid." The values of P and T for any point in this region of overlapping may characterize either a solid or a liquid, though what is called a "liquid" for this purpose may be so stiff and tough as to deserve and bear the name of "glass." If the solid and the liquid are brought into contact the latter may pass into the former (but never the former into the latter) in a manner so nearly explosive as to be clearly not reversible. I should be less than frank were I to imply that this is always so. With the transition from "supersaturated gas" to liquid, the irreversibility is manifest.

$$I - S(P, 0) = R \ln P - C_p^{gas} \ln T + \int (C_p^{sol}/T)\, dT + L/T \qquad (15)$$

All of the terms on the right are measurable quantities, excepting that C_p^{sol} cannot be followed clear down to the absolute zero, so that the piece of the integral extending from zero up to a few degrees absolute must be guessed. Now at the bottom of the accessible temperature-range C_p is already very small, and in most cases the curve of C_p versus T seems to be heading very smoothly toward zero, so that the uncertainty is probably slight. Surprises in the unreachable range are, however, not inconceivable.

Everyone familiar with entropy will have known in advance, and on reading equations (12) and (14) will have remembered, that the additive constants I and $S(P, 0)$ are beyond the reach of all experiment, be it physical or be it chemical. No way can be devised of measuring them, for in chemistry or in physics it is never the entropy of a system in any one state which is measured, but only the difference of the entropies in two different states, and the additive constant is cancelled in the subtraction. So far as (12) by itself or (14) by itself is concerned, each constant is but a vain appendage, and to develop a theoretical value of either would be reasoning in a void. In spite of all this, the difference between I and $S(P, 0)$ is within the reach of experiment. This permits of one, or two, or even of all three of the following situations:

(*a*) If there is a plausible theory of entropy which leads to a value for I, experiment will fix a value for $S(P, 0)$ corresponding to that theory.

(*b*) If there is a plausible theory of entropy which leads to a value for $S(P, 0)$, experiment will fix a value for I corresponding to that theory.

(*c*) If there is a theory which leads to a value for $S(P, 0)$, and there is another and independent theory which leads to a value for I, then experiment can tell whether the two are compatible.

The actual situation is most nearly like the last of these three; and from this viewpoint I will describe it.

Before going on to the theories of I and of $S(P, 0)$ I point out that (15) *is the equation of the vapor-pressure curve*; for P and T occur in it as variables, and it refers explicitly to such paired values of P and T as correspond to points lying on the divide between solid and gas, and P for any such point is called the vapor-pressure of the substance for the temperature corresponding. Measurements of vapor-pressure are therefore the ones which are called to decide on these questions—measurements of vapor-pressure and of specific heat, and of the heat of vaporization. The last-named, the quantity L, need be measured at a single temperature only, for there is a formula which gives its value at any temperature in terms of its value at any other temperature and the specific heats over the range between.

This is often incorporated into (15), and alters its aspect; but until page 71 I will pass over this detail. If the point (P, T) lies beside the liquid-gas divide, the isobar to which (14) refers will cut both this and the solid-liquid divide, and traverse the liquid region. Extra terms will then appear in (14) and in (15), but I leave it to the student to divine them.

THE THEORY OF THE CONSTANT S_0

The theory of the constant S_0 is easy to state, provided that no objection is raised to having it stated in a manner rather too drastic at first, and waiting for the necessary reservations to be added later.

The constant S_0—the entropy at the absolute zero—is taken to be zero for every substance of a single kind.

This is a way of putting, and the strongest possible way of putting, what is known as "Nernst's Law" or "Nernst's Heat Theorem" or even "the Third Law of thermodynamics." Originally expressed in a much milder form nearly forty years ago, it rapidly progressed to the stringent form embodied in these words. As I have suggested already, it is a form too stringent; but the truth lies nearer to it than to the milder phrasings earlier used, and therefore it is justifiable as a commencement.

Notice to begin with that, in the statement as just given, there is no allusion to the pressure or the volume. It is therefore asserted that, at the absolute zero, the entropy of a substance (of a single kind) does not depend on either. This I implied already in the caption to this section, by discarding the symbol $S(P, 0)$ which had previously served for the additive constant in (14), and replacing it by S_0. From the general thermodynamic equations based on First and Second Laws, it can be shown that if this is true for any substance in particular, then certain measurable features of that substance—notably the coefficient of thermal expansion—must be zero at the absolute zero. Now it does appear to be a general rule that this coefficient, and the other features in question, are trending rapidly to zero at the lower end of the accessible range of temperatures; so this, the mildest form of the "Third Law," is well attested.

Notice then that in the statement as given there is no allusion to phase. Thus, if any substance can exist in both the solid and the liquid phase at the absolute zero, its entropy must be the same in both (if the theorem is true). If we insist on fluidity as a quality of a liquid, there is evidently just one such substance—helium, of course. It appears that for this case the theorem is true.[4]

---

[4] At the lowest accessible temperatures, the divide between the "liquid" area and the "solid" area of the phase-diagram is running nearly parallel to the temperature axis, and heading for the ordinate 25 (in atmospheres) on the pressure-axis. Strict parallelism of the divide to the temperature-axis, which is probably realized just before the absolute zero, would imply equality of entropy in the two phases (by one of the thermodynamic equations hinted at above). Another item of evidence is cited on p. 72.

There are substances able to exist in two or more crystalline phases: tin and sulphur are probably the best-known examples. For some of these it is possible to start with one phase at a temperature extremely low: warm the substance up to a temperature of "transition," at which it changes reversibly into the other phase; and cool the new phase down to the temperature at which the experimenter started. Let me denote by S_1 the entropy at the commencement of this process; by S_2 the entropy at its finish; by C_{p1} and C_{p2} the specific heats of the two phases; by T_t the temperature of transition, by L_t the heat absorbed during the transition. We have:

$$S_2 - S_1 = \int_o^{T_t} (C_{p1}/T)\, dT + L_t/T_t - \int_o^{T_t} (C_{p2}/T)\, dT \qquad (16)$$

According to Nernst's Theorem, S_2 and S_1 and consequently their difference should vanish if the extremely low temperature at start and finish were the absolute zero. We should therefore expect the right-hand member of this equation to be at any rate extremely small, if the temperature in question is at the bottom of the accessible range. Such is indeed the case.

It is very evident that the argument just given proves at the very most that the entropies of the two phases are equal at the absolute zero—not that either of them has the particular value *zero*. The like is true for the other arguments thus far cited; and indeed in the earliest phrasings of the "Third Law," no value was assigned to the entropy at the absolute zero—neither the value *zero*, nor any other. Why then are we to adopt, and presently seek to justify, the particular value *zero*? One part of the answer will be the climax of this paper. The other derives from the speculations as to the nature of entropy, which for half a century have been among the most deeply perpended, the most difficult and the most fruitful of the divisions of theoretical physics.

There are two words which dominate these speculations: "probability" on the one hand, "disorder" on the other. Both of these are very familiar words with very familiar meanings, and some tinge of the familiar meaning is in each case carried over into the technical meaning. The technical meanings, are, however, abstruse; and cynical though it may sound, there is no exaggeration in saying that a large part of the speculation consists in trying to find meanings for the one and the other, which can be fruitfully used in the study of entropy. "Disorder" is the word which we shall examine first.

The familiar meaning of the word "disorder" leads straight to one useful consequence. Of all the possible or conceivable states of matter, the one which anybody would choose as the least disorderly is the crystalline state. But moreover, most people would deem the thermal agitation of the atoms in a crystal as a departure from order; therefore the colder the crystal, the closer its approach to perfect order. Perfect order would consist in every atom located at rest in exactly its proper place in the crystalline lattice. If such is the state of affairs at the absolute zero, then a crystalline phase at the absolute zero exhibits perfect order. Now as I have implied without precisely saying, entropy is taken to be a measure of disorder. If this is proper, then it is sensible and correct to say that entropy vanishes for the perfect crystal at the absolute zero. This is a reason, it is in fact *the* reason, for assigning the value *zero* to the additive constant S_0.

If I were to stop my exposition at this point, it would certainly be an impressive conclusion, but hardly a just one. We cannot solve the mysteries of entropy in any so simple a way; in fact they have never been completely solved. Many additions should be made to the foregoing paragraph: some would help out its evident meaning and some would not. Thus, as it stands, it certainly implies that a liquid at the absolute zero—helium offering such a case—must exhibit disorder and therefore have a greater value of entropy than does the crystalline solid. However, it turns out that liquid helium near the absolute zero has the same entropy as solid helium does, and therefore if one value is zero so also must the other be. If instead of liquid helium I had taken the supercooled and toughened liquid which is called a "glass"—glycerol or alcohol in the glassy state, for instance—the result would have been more agreeable. Several of these glasses have been studied with great care, and the right-hand member of equation (16) evaluated for them; it turns out that the entropy near the absolute zero is markedly and indubitably greater for the glass than for the crystal. "Disorder" must therefore prevail in the glass, but the result with liquid helium impedes the physicist from proclaiming that it is simply the obvious disorder of the irregular placing of the atoms in the glass. A mixture of two substances of a single kind, even when this mixture forms a beautiful crystal, may nevertheless present disorder; for the two kinds of atoms may be sprinkled in a thoroughly chaotic manner over the available points of the crystalline lattice. Such a mixture, it has been found, does have an entropy near the absolute zero which is definitely greater than the sum of the entropies of the two substances when unmixed. But it is also possible for the mixture of two substances to be orderly, in the usual sense of the word! Thus in a mixture—"alloy" is the customary word—of equal numbers of atoms of copper and zinc, the copper atoms may all be found upon one lattice and the zinc atoms upon another, the two lattices interpenetrating and interlocking with each other so as to form the lattice of the crystal of the alloy. This is one of the extreme possibilities: the other is, the entirely chaotic besprinkling of the points of the two lattices with atoms of the two kinds. Both can be realized, the transition between the two being known as "the order-disorder transition." The entropy goes down as the transition is made from disorder into order.

The conclusion then is, that we may accept entropy as a measure of disorder and disorder as a way of visualizing entropy, provided that we are prepared to define "disorder" in ways which at least in certain striking cases do not depart impossibly far from its traditional meaning. A consequence of this attitude is, that it is plausible and sensible to attach the value *zero* to the constant here called S_0, the entropy of any substance of a single kind in a crystalline phase at the absolute zero. The words "in a crystalline phase" are a reservation to the original statement. If further reservations become necessary, they will of course have to be made.

THE THEORY OF THE CONSTANT I.

To give even an inkling of the theory of the constant I, it is desirable to take "probability" as the word for which meanings must be found, not too distant from the popular meaning and yet fruitful for the study of entropy. Those who began this process were Gibbs and Boltzmann, working in the closing years of the nineteenth century. Their ideas have since undergone many a transformation, usually in the direction of greater adequacy but also (alas!) in that of greater difficulty. I will follow a route beginning as Boltzmann's did, but carried onward in a manner which became possible about thirty-five years ago, at the time when Nernst's Heat Theorem was being established. It does not lead us quite the whole way to the accepted value of I, so that at the end I shall have to make an extra step without doing more than to indicate whence its justification comes.

We begin by considering a gas in a container of volume V, in equilibrium with itself and with the outer world at a temperature T. "In equilibrium with itself" implies first of all that it is evenly spread throughout the volume of the container—surely one of the earliest of all inbred ideas concerning the behavior of gases. To give a quantitative meaning to this notion of the gas being evenly spread throughout the container, we imagine the volume divided into little compartments or cells of equal volume V_0. The statement then is that the number of atoms in every cell is the same. Putting N for the total number of atoms and Nf_i for the number in the ith compartment,

$$f_i = \text{constant} = V_0/V \tag{17}$$

The quantity f_i is called the "probability" that an atom chosen at random shall be in the ith cell—the first occurrence of the word "probability" with a definite meaning in this discourse.

The next step is to define the entropy in the manner which follows:

$$S = -kN\Sigma f_i ln f_i \tag{18}$$

a definition which in the case of the even distribution of the gas throughout the container reduces to this,

$$S = kN\, lnV - kN\, lnV_0 \tag{19}$$

This is of course a *tentative* definition, to be eschewed if ever it should lead us into contradiction with the known properties of entropy. As yet it leads us into none, for the term in $ln\ V$ corresponds to a similar term in the description of entropy which equation (13) has already supplied. To make the two agree exactly, we have simply to assign a special value to the factor k; and as will be seen at once, this value is

$$k = R/N_0 \tag{20}$$

N_0 standing for the number of atoms per mole, the Avogadro number. Though k is known as Boltzmann's constant, this evaluation was beyond Boltzmann's powers, for in his time the value of N_0 was not known.

The expression (19) contains no allusion to temperature. It is in fact not the entropy in full which has so far been defined, but only what I may call the "contribution of volume to entropy." We have now to account for the contribution of the kinetic energy of the molecules to the entropy of the gas. Thus far I have been able to come by adducing the deeply-inbred conviction that a gas in equilibrium in a container is evenly spread throughout the container. There is no such widely-held conviction about the distribution-in-energy of the molecules of the gas; but to everyone who has studied physics for more than a year or two there will be nothing surprising in the formula which follows. It must be introduced by asking the reader to imagine a three-dimensional space, in which the variables along the three axes are identified not as coordinates in ordinary space, but as components of momentum p_x, p_y, p_z. The momentum which is meant is the momentum of the individual atom, and the axes x, y, z along which its components are taken are axes of a coordinate frame in ordinary space—they might be along three edges of the container, for instance. A point in the "momentum-space" represents an individual atom in respect of its momentum and therefore in respect of its energy, but not in respect of its position.

The momentum-space is now to be divided into compartments of equal volume H_0; but we are not to besprinkle its compartments uniformly with the dots representing the atoms! Instead, when comparing any two of the cells, say the ith and the jth, we are to write

$$f_i/f_j = exp\,[-(E_i - E_j)/kT] \tag{21}$$

which reduces to the statement,

$$f_i = Z^{-1} exp(-E_i/kT) \tag{22}$$

Z being a symbol for the sum,

$$Z = \Sigma_i \, exp(-E_i/kT) \tag{23}$$

which, incidentally, is known as the "partition-function" or "state-sum" of the system—here, of the individual atom of the gas.

In these latest equations, E_i stands for the energy-value corresponding to some point in the ith cell. It is kinetic energy of translatory motion with which we are concerned; therefore E_i is given by the equation

$$E_i = (1/2m)[(p_x^2)_i + (p_y^2)_i + (p_z^2)_i] \tag{24}$$

where now the components of momentum are to be evaluated at some particular point in the ith cell. But at which particular point? And to go further back, just how are the cells of volume H_0 to be designed in the momentum-space? As cubical blocks with their edges parallel to the coordinate-axes, or how? There are problems in which definite answers must be given to these questions, but we shall be able to avoid them. It will be adequate here to conceive of the cells as cubical blocks and the value of E_i as the *average* value of the right-hand member of (24) in the ith cell.

Now we require from (24) the value of Z as defined in (23), to establish the values of f_i as given in (22), to yield finally the value of entropy as given in (18).

Let us form the integral

$$\iiint exp\,(-E/kT)\, dp_x\, dp_y\, dp_z\,, \qquad E = (1/2m)(p_x^2 + p_y^2 + p_z^2) \tag{25}$$

the range of integration extending over the whole of momentum-space. This integral may be described as follows. Let the momentum-space be divided into *cells of unit volume*. Each of these cells of unit volume makes a contribution

$$exp(-\bar{E}/kT)$$

to the integral, $\bar{E}$ standing now for the *average value of E in the cell in question*. The integral is the sum of all of these contributions. Now let us inquire how much of a contribution is made by this same cell of unit volume to the partition-function. This second contribution is made up of $1/H_0$ terms, one for each of the cells of volume H_0 which occupy the cell of unit volume. The values E_i corresponding to these cells will not be exactly equal to the value $\bar{E}$ corresponding to the entire cell of unit volume;

but to the degree of approximation which is now being used, the difference may be neglected. The partition-function is then equal to $1/H_0$ times the integral (25). Now the value of the integral (25) is given in all tables of definite integrals, and in terms of our symbols it amounts to

$$(2\pi mkT)^{3/2}$$

We are now to divide this by H_0, and proceed along the path which has been indicated.

The procedure is simple and straightforward. As a byproduct one finds the result that the energy of the gas—which I have earlier symbolized by U—is equal to $(3/2)kNT$. It follows that the specific heat at constant volume is equal to $(3/2)kN_0$ for one mole, to n times this value for n moles of gas. Utilizing this result, and putting nR for Nk wherever the latter occurs, one duly arrives at (18) all filled out in the proper way. This represents the contribution of the temperature to the entropy; now adding the contribution of the volume from (19), one arrives at *the entropy of n moles of the gas as function of temperature and of volume, as derived by the statistical method*:

$$S = nR\, ln\, V + nC_v\, ln\, T + nR\, ln\, \frac{(2\pi m)^{3/2} k^{3/2}}{V_o H_o} + (3/2)nR \qquad (26)$$

This is now to be compared with the equation (13) for entropy as function of volume and of temperature, embodying the definition of entropy wherefrom we started.

So far as the dependence on T and on V is concerned, all is well! And there seems even to be a prospect of finding a formula for the additive constant. The prospect, though, is still befogged in two ways: by my lack of precision till now as to the magnitudes of V_0 and H_0, and by the absence from (26) of any term convertible into the term $nR\, ln(nR)$ which stands out so prominently in (13).

As to V_0 and H_0: no assumption shall be made about either by itself, but it will be assumed that *their product is equal to Planck's constant h raised to the third power* (third power, because of the three dimensions of space):

$$V_0 H_0 = h^3 \qquad (27)$$

This I will attempt to justify from a fact not even divined when the formula was made.

To divide the momentum-space into cells of definite size, and to allot to the partition-function just one term from each cell—this comes to the same thing as allowing certain discrete momentum-values to the atoms in question, and denying them all values intermediate to these "permitted" ones. By using the words "permitted values" I am recalling the quantum-theory, and it is in fact a part of the quantum-theory which we are now employing, as betokened already by the entry of the symbol h. It is one of the oldest parts of the quantum-theory; but the new fact—comparatively new—is this. Atoms, like all other particles, are attended and governed by waves. These waves, when with their atoms they exist in a limited space such as that which the container offers to the gas, are constrained to what in acoustics is known as "resonance." Not every frequency of vibration is allowed to the air within an organ-pipe, to the wire of a piano or to the membrane of a drum, but only such as have wave-lengths fitting neatly into the compass of the cavity, the wire or the drumhead. The dimensions of these acoustical resonators control the permitted wave-lengths, and these in their turn determine the frequencies. In the case which we are now considering of a container filled with a gas, the dimensions of the container control the wave-lengths associated with the atoms, which are the wave-lengths of resonance. These in turn control the momenta of the atoms, because of the relation between the momentum of a particle and the wave-length of its associated waves—the "Rule of Correlation":

$$p = h/\lambda \tag{28}$$

To say that the momenta of the atoms are those and only those corresponding to the resonant wave-lengths, and to say that V_0H_0 in (26) is equal to h^3—these are equivalent statements. When the former is accepted, so perforce is the latter, and the additive constant in (26) is fully determined. But still it lacks the term $-nR\, ln\, nR$ or $-Nk\, ln\, Nk$ which figures in (13)!

To introduce this term into the theory in a way both logical and simple is not an easy task. The formula at which we are about to arrive is frequently known as the "Sackur-Tetrode formula" after the two physicists of whom (before the first world war) one was the first to approach and the other the first to reach it. Sackur assumed outright that V_0H_0 is inversely proportional to N, while Tetrode subtracted from (26) a term $k\, ln\, N!$—the exclamation-point here not having its rhetorical meaning, but signifying that $N!$ is "N factorial," the product $1 \cdot 2 \cdot 3 \cdots N$. By Stirling's celebrated formula,

$$ln\, N! = N\, ln(N/e) \tag{29}$$

an approximation amply valid for such enormous values of N as are normally here considered. Be it noted that e here stands for the exponential base and not for the electron-charge (in the latter sense it is never used in this article).

To make clear the basis for this subtraction of $k\, ln\, N!$ I should have to go far into the roots of the conception of entropy as probability. The best

I can do for the present is to follow Tetrode by saying that he deemed (26) as it stands to be an overstatement of the entropy, arising because in advancing from the underlying theory to equation (18) one assumes the atoms of the gas to be distinguishable, whereas actually for a substance of a single kind they are indistinguishable. In somewhat the same way one might overstate the entropy of a crystal by supposing it to be composed of atoms no two of which were alike, while actually it was a substance of a single kind. The subtraction of $k \ln N!$ was Tetrode's manner of correcting the overstatement. He did better than perhaps he knew, for while his reasons never commanded universal assent, his choice of a term to be subtracted was ratified first by experiment and then by the "new statistics" which made their appearance in physics some fifteen years ago.

Returning to (26): writing the last term of the right-hand member as $nR \ln e^{3/2}$, and consolidating it with the third term; introducing Tetrode's subtractive term; augmenting this last by a term $-nR \ln k$, and compensating by adding $+nR \ln k$ to the third term—doing all this, one finds,

$$S = nR \ln V + nC_v \ln T - nR \ln (nR) + nR \ln \frac{(2\pi m)^{3/2} k^{5/2} e^{5/2}}{h^3} \qquad (30)$$

Now the additive constant is filled out completely, and ready for whatever test experiment may impose.

To prepare it for the test, we turn back first to equations (12) and (13), and note that the constant there denoted by nI is none other than the fourth term in the right-hand member of (30):

$$I = R \ln \frac{(2\pi m)^{3/2} k^{5/2} e^{5/2}}{h^3} \qquad (31)$$

Continuing onward to (15) we are reminded that no theoretical statement about I is worth anything by itself, since all that data can supply is the value of the combination $(I - S_0)$. A hopeless situation, in appearance! But now it is high time to hearken to what the data say. The data say, to begin with:

For many monatomic gases, the right-hand member of equation (15) is equal to the "statistical" value of I.

This may be taken as meaning two things at once: that (*a*) the statistical theory of the entropy of a gas is right, and (*b*) the entropy of a solid (crystalline and of a single kind, for such are these solidified gases) is zero at the absolute zero. It *is* taken as meaning these things. It might of course also be taken as meaning that both statements are wrong by about the same amount, the errors compensating one another. But so unlikely does it seem that two such different theories should both be wrong and yet by precisely the same amount, that I venture to say that few will be found who are critical enough to insist on the doubt.

This concordance might be shown by giving actual values of the right-hand member of (15), for actual gases at one or more points adjoining the curve which divides solid from gas on the phase-diagram. The usage is, however, different, and it is of some value to abide by the usage, even though I am now required to make one more transformation of (15) which I have hitherto avoided. This involves the auxiliary equation,

$$dL/dT = C_p^{gas} - C_p^{sol} \tag{32}$$

which makes it possible to calculate L for any temperature, provided it be measured at any other temperature and the specific heats be known at all the temperatures in between. We may thus start from any measured value of L, calculate a value L_0 appropriate to the absolute zero, and thereafter we may write the integrals, as always heretofore, extending from zero to T.

$$L = L_o + \int C_p^{gas}\, dT - \int C_p^{sol}\, dT \tag{33}$$

The first of these two integrals is equal for a monatomic gas to $(5/2)RT$ per mole. Since it is L/T which appears in (15), this entails a term $(5/2)R$ on the right of that equation. This term neatly blots out the term $R\, ln\, e^{5/2}$ which is a portion of our statistical value of I as given in (31), so that after all the equation is not much worsened in appearance. It has in fact assumed the form

$$R\, ln\, P = -L_o/T + (5/2)R\, ln\, T - \int (C_p^{sol}/T)\, dT + (1/T) \int C_p^{sol}\, dT + J \tag{34}$$

and here J is yet another additive constant—the additive constant of the vapor-pressure equation in its commonest form—of which this may be said,

$$J = R\, ln\, \frac{(2\pi m)^{3/2}\, k^{5/2}}{h^3} \tag{35}$$

if the statistical theory is sound and if in addition the entropy of a crystalline solid of a single kind at the absolute zero is zero.

It is evident that if the two presumptions are true, then the remainder left behind when $(3/2)\, ln\, m$ is subtracted from the experimental value of J will have a definite value, the same for all gases. This indicates how the test is made, or rather how it is set forth in the literature. The remainder, expressed in units which need not concern us, should be -1.589. Now, data are available[5] for all of the noble gases from helium through xenon; and I quote the values which they furnish for this "remainder":

He 1.59 ± 0.01; Ne (two experiments) 1.56 ± 0.04, 1.59 ± 0.01
Ar (two experiments) 1.61 and 1.61 ± 0.04;
Kr 1.59 ± 0.02; Xe 1.588

The agreements are remarkably good, and when one remembers the contrast and the subtlety of the two theories which they ratify conjointly, one deems them among the most impressive in the whole of physics or of chemistry either. The result for He implies that $S_0 = 0$ for *liquid* helium.

There are data available also for the vapors of several metals, though for these the experimentalist's task is much harder. It will be seen that in most of the cases we may again rejoice in agreements, though within a broader range of experimental unsureness; in a few there is disagreement, and one can only wonder whether the measurer of vapor-pressure underestimated his possible error.

Na (two experiments) 1.57 ± 0.1 and 1.41 ± 0.03;
K 1.47 ± 0.04; Mg. 1.61 ± 0.2; Zn 1.51 ± 0.15;
Cd (two experiments) 1.63 ± 0.1, 1.51 ± 0.1; Hg 1.62 ± 0.03;
Tl 1.40 ± 0.3; Pb (two experiments) 1.7 ± 0.2, 1.21 ± 0.26.

So the situation with respect to monatomic gases is satisfactory on the whole and almost unblemished.[6] But monatomic gases are but a small proportion of all vapors: has not the theory something to say as regards those which are diatomic—such as oxygen, hydrogen, nitrogen, carbon monoxide—and even those which are polyatomic?

The theory does indeed cover these; and the needed enlargement is simple to write down, if not always easy to compute. Thus far the partition-function has contained only the terms supplied by the translatory motions of the molecules in the container—terms which depend on no other quality of the molecules than their masses, and for a given mass are the same whether the molecule is composed of a single atom or of two or of many. Now to the partition-function so composed are to be added the terms deriving from the rotation and from the vibrations of the molecules. Each new term is of the familiar form $exp(-E_i/kT)$, where now E_i is to stand for some permitted energy-value of rotation or of vibration, and the summation is

---

[5] Data from the collation of Ditchburn and Gilmour (*Rev. Mod. Phys.* **13**, 310, 1941), except the value for xenon which is from Clusius.

[6] Here I must say, even if only in a footnote, that for atoms which have angular momentum there is still one more term in the additive constant occurring in the right-hand member of (30). This term may be loosely described as referring to the "disorder" arising from the fact that there are two or more different orientations of the angular momentum, between which the atoms may choose. Its exact value (zero for the inert gases, $nRln2$ for the alkali metals) is derived from the spectrum of the atom. Allowance is made for this term in the computation leading to the value of the "remainder", and the agreements here tabulated for Na and K sustain the allowance. A bad disagreement for tungsten emerges from an early (1924) research by Lange.

to be carried over all values of E_i. There is less of a mental obstacle now than there was in the "translatory" terms, for the doctrine of discrete permitted energy-values in rotation is among the oldest, while that of discrete energy-states in vibration is quite the oldest, of the whole of quantum-theory. To get the energy-values E_i, the spectra are consulted. Spectrum analysis is the most highly-developed of the branches of physics, and the knowledge of the values of E_i which it affords for the better-known of the diatomic molecules—and even for some few of the triatomic—may be described as complete, daring as it may sound.

Values of entropy figured by the statistical way which I have been describing, with additions duly made to the partition-function to allow for rotation and vibration, are called "spectroscopic" entropy-values by Giauque and his colleagues of the Berkeley school. As a rule they are computed for a standard pressure and temperature, often for one atmosphere and 25°C.; let us call these P_0 and T_0. Now as a rule the point (P_0, T_0) does not lie just off the divide which on the phase-diagram separates solid (or liquid) from gas. One may, however, use (11) for getting the entropy-difference between (P_0, T_0) and any such point (P, T) lying along the divide on the side corresponding to gas. At the point (P, T) the entropy may be evaluated by the right-hand member of (14), and to it may then be added the entropy-difference just mentioned. The sum, *computed with the assumption that S_0 is zero*, is what Giauque and his school name a "calorimetric" entropy-value. If the two agree, there is a verification of the conjoined assumptions that the statistical theory is sound, and that the crystalline solid at zero absolute has none of that disorder which is the source of entropy.

I take the following data from Eastman[7] of the Berkeley school. Of his two tables, one comprising gases for which the data are definitely better known than they are for the rest, I quote this one only. The accuracy of the calorimetric value "is in all cases high, errors being of the order of 0.1 throughout." The supposed accuracy of the spectroscopic values is indicated by the number of significant figures to which they are carried.

	Calor.	*Spectrosc.*
N_2	45.9	45.788
O_2	49.1	49.03
HCl	44.5	44.64
HBr	47.6	47.48
HI	49.5	49.4
CO	46.2	47.316
NO	43.0	43.75
H_2O	44.23	45.17
N_2O	51.44	52.581
H_2	29.7	31.23

---

[7] *Chem. Rev.* 18, 257 (1936).

The first five are the stellar examples of concordance, therefore support for the two conjoined assumptions. In the last five the discrepancy is deemed to exceed the possibility of doubt, and some explanation ought to be found.

How is one to explain a "calorimetric" value less than the "spectroscopic" —as occurs in all five of these cases? The comparison, I recall, is essentially between $(I-S_0)$ and I. In all of these five cases, $I-S_0$ turns out to be less than I; therefore we are tempted to assume a "disorder" in the solid, subsisting even unto the absolute zero. The case of hydrogen is the clearest, the disorder being traceable in part to the intermingling of two types of hydrogen molecule known as the "ortho" and the "para" type. In the other cases, the explanations have a vagueness which suggests that they are speculative. This is not a very satisfactory statement to end with. Yet even these discrepancies are so small, that even if there were no closer agreements one would still feel assured that the truth is to be found in some minor alteration of the theory rather in rejecting the theory altogether; and the five excellent agreements, combined with the excellent agreements for so many monatomic gases, convert this feeling of assurance into certitude.

Abstracts of Technical Articles by Bell System Authors

Recent Developments in Burying Telephone Cables.[1] DONALD FISHER and TEMPLE C. SMITH. The term "buried cables" has come to mean those underground cables which have no conduit protection. Due to the accelerated demand for such construction in recent years, much effort has been expended in devising methods and developing machinery for burying cables. One of the earlier methods used in this and some foreign countries was to dig a trench by hand alongside the road; unreel the cable from a moving truck, thus laying it beside the trench; work the cable over into the trench by having 30 to 50 men handle it in relays; splice the cable in the trench, and finally backfill the spoil and tamp it by hand. Later variations of this method introduced one or more of the following units of machine equipment: Power trenching machines: Caterpillar tractors with trailers to straddle the trench, laying the cable directly from the reel into the trench; Drag-line or other types of power backfillers; Power tampers or rollers.

In order to further reduce the number of operations involved, speed up the installation, and reduce the cost, large plow trains have recently been developed which, except for splicing, in ordinary soil complete the job of burying a cable in one pass over the route. The idea of plowing cable into the ground is not new. In fact the great grandfather of all the cable plows was designed by Ezra Cornell long before he established the university. His "ponderous machine" drawn by a "long line of horses" was designed for laying telegraph cable in the early 1840's, but the development was dropped when the simple expedient of carrying wires on poles and insulators was conceived.

The large plow trains recently developed for installing telephone cable are capable of burying either a single cable or a pair of cables together with as many as three properly spaced lightning-protection wires, and of cutting a slot for them as much as 50 inches deep where such a depth is required. To provide the complete plow train has required the design of many pieces of equipment which the word "plow" does not suggest to one's mind. The plows and some of this equipment are discussed in this paper.

A Frequency-Modulated Control-Track for Movietone Prints.[2] J. G. FRAYNE and F. P. HERRNFELD. A 5-mil frequency-modulated track located between sound and picture areas is proposed to control reproduction

---

[1] *Elec. Engg., Transactions Section*, April 1942.
[2] *Jour. S.M.P.E.*, February 1942.

in the theater from one or more sound-tracks. A variation of approximately one octave in the control frequency provides a 30-db change in volume range which may be used in part for volume expansion of loud sounds or as noise reduction for weak sounds. The control-track frequency is varied manually and recorded simultaneously with the sound-track in the dubbing operation, the gain of the monitoring channel being varied in accordance with the control frequency to produce automatically the enhanced volume range desired from the release print. The track is recorded in line with the standard soundtrack and does not require separate printing or reproducing apertures. It is scanned by a separate photosensitive surface, the output being converted from frequency to voltage variations by a frequency-discriminating network identical to that used in the monitoring channel. The output from the network, applied to the grid of a variable-gain-amplifier in the sound channel, controls automatically the volume of the reproduced sound in accordance with that observed in the dubbing operation.

Performance of Ground-Relayed Distribution Circuits during Faults to Ground.[3] C. L. Gilkeson, P. A. Jeanne and J. C. Davenport, Jr. An extensive oscillographic study has been made on power distribution feeders primarily to obtain data useful in the consideration of joint use of poles by power and telephone facilities. Some of the results, chiefly those obtained from three-phase, four-wire, multigrounded neutral feeders equipped with instantaneous ground relays and for immediate breaker reclosure, are believed to be of general interest and are presented herewith. Included are data on the performance of the protective devices utilized for clearing ground faults on the feeders included in the study, the effectiveness and certain limitations of these devices, and characteristics of the faults experienced.

A More Symmetrical Fourier Analysis Applied to Transmission Problems.[4] R. V. L. Hartley. The Fourier identity is here expressed in a more symmetrical form which leads to certain analogies between the function of the original variable and its transform. Also it permits a function of time, for example, to be analyzed into two independent sets of sinusoidal components, one of which is represented in terms of positive frequencies, and the other of negative. The steady-state treatment of transmission problems in terms of this analysis is similar to the familiar ones and may be carried out either in terms of real quantities or of complex exponentials. In the transient treatment, use is made of the analogies referred to above, and

---

[3] *Elec. Engg., Transactions Section,* January 1942.

[4] *Proc. I.R.E.,* March 1942.

their relation to the method of "paired echoes" is discussed. A restatement is made of the condition which is known to be necessary in order that a given steady-state characteristic may represent a passive or stable active system (actual or ideal). A particular necessary condition is deduced from this as an illustration.

Paper Dielectrics Containing Chlorinated Impregnants.[5] D. A. McLEAN, L. EGERTON, G. T. KOHMAN and M. BROTHERTON. Chlorinated aromatic hydrocarbons are used extensively as impregnants for paper in electrical apparatus such as capacitors and transformers. They possess high stability toward heat and oxygen, nonflammability, and desirable electrical properties. However, the present study shows that under special conditions a type of decomposition occurs which liberates products so active chemically that, especially at elevated temperatures, under high potential gradients, and in contact with metals, autocatalytic decomposition reactions are initiated.

When aluminum electrodes are used and d-c. fields are applied, the deterioration manifests itself in a leakage current increasing with time and in the formation of partially carbonized areas in the paper. Evidence is presented that the decomposition starts by the splitting off of hydrogen chloride from the chlorinated compounds. This attacks the electrodes to form aluminum chloride, which decomposes the organic constituents of the insulation.

The deterioration is considered in relation to certain variations in composition; all components of the structure—the paper, the impregnant, and the electrode materials—play important roles in the reactions.

Methods for Measuring the Performance of Hearing Aids.[6] FRANK F. ROMANOW. A hearing aid can be considered as a sound transmission system which is interposed in the path between the source of sound and the listener's ear. As such its performance can be judged by comparing the sound that reaches the ear first through the air path and then through the hearing aid. The experimental procedure to carry out this concept is, however, time consuming. To obtain a simple laboratory technique for comparative purposes, it is proposed, therefore, that the amplification of a hearing aid be measured by placing its microphone in a known sound field and observing the output of the receiver when terminated in an appropriate impedance. For an air conduction receiver this impedance takes the form of a closed cavity. For a bone conduction receiver the termination is an artificial mastoid. Since in a portable hearing aid compactness is desired,

[5] *Indus. & Engg. Chem.*, January 1942.
[6] *Jour. Acous. Soc. Amer.*, January 1942.

the indistorted pressure that can be developed by the set is limited in magnitude. Also the amplification is a function of the variation of the power supply voltages. Hence it is necessary that these characteristics be measured. As an illustration of the different factors involved, characteristic curves for a recently developed vacuum tube aid are shown.

Some Mechanical Aspects of Telephone Apparatus. Part 1—Development of Crossbar Switch as a Typical Unit of Automatic Dial Telephone System.[7] J. D. TEBO and H. G. MEHLHOUSE. One half the telephones in the world are in the United States. A total of 100,000,000 calls a day are made on this equipment, on matters large and small—routine or emergency. Especially in these days of national emergency, communication facilities have become indispensable to our defense organizations, vitally assisting in speeding up the program of national safety.

Americans have grown to accept their telephones as a necessity, seldom realizing the vastness of the system and the spirit of service in the large organization constantly striving to maintain and improve a means of communication already unexcelled. Seldom is it realized that the equipment in the home or office is only a very small part of the facilities required in this business of telephoning. A network of wire literally covering the nation from coast to coast, and intricate switching devices for interconnecting the telephones throughout the nation and most of the world all combine to transmit the spoken word quickly and faithfully.

The vastness and complexity of such a service must of necessity require a wide range of equipment which, in the Bell System, totals 44,000 kinds of apparatus involving 170,000 different parts. Some of the mechanical aspects of such equipment are presented in this paper, particularly those having to do with design and manufacture, jointly handled for the Bell System by the Bell Telephone Laboratories and Western Electric Company. Obviously, the entire range of telephone equipment cannot be included. This paper is therefore intended to deal with a specific telephone switching mechanism, known as the crossbar switch, which is used in the latest dial telephone system.

---

[7] *Mech. Engg.*, May 1942.

Contributors to this Issue

OLIVER E. BUCKLEY, B.Sc., Grinnell College, 1909; Ph.D., Cornell, 1914; D.Sc., Grinnell, 1936. Engineering Department, Western Electric Company, 1914–1917 and 1919–1925; Major, Signal Corps, U. S. A. in charge of research laboratory of Signal Corps in A.E.F. 1917–19; Bell Telephone Laboratories, 1925–. Assistant Director of Research, 1927; Director of Research, 1933; Executive Vice-President, 1936; President, 1940–. Dr. Buckley's personal contributions to the art of communication have been principally in the field of submarine telegraphy and telephony. He is a member of the National Academy of Sciences and of the American Philosophical Society; a fellow of several scientific and engineering societies, and chairman of the Engineering Foundation Board.

L. E. CISNE, B.S., in Physics, University of Illinois, 1941. Western Electric Company, Hawthorne Works, Chicago, Illinois, 1936–. Mr. Cisne has been engaged in inspection engineering.

KARL K. DARROW, B.S., University of Chicago, 1911; University of Paris, 1911–12; University of Berlin, 1912; Ph.D., University of Chicago, 1917. Western Electric Company, 1917–25; Bell Telephone Laboratories, 1925–. As Research Physicist, Dr. Darrow has been engaged largely in writing on various fields of physics and the allied sciences.

DOLORES B. KEELING, Ph.B., University of Chicago, 1921. Western Electric Company, Hawthorne Works, Chicago, Illinois, 1922–. Miss Keeling has spent eight years in industrial relations and twelve in inspection development and statistics.

H. N. PADOWICZ, Ch.E., Rensselaer Polytechnic Institute, 1928. Metallurgical Engineer, Powdered Metals and Wire Products, Westinghouse Lamp Company, 1929–1935; Engineer of Research, Sirian Wire and Contact Company, 1935–1937; Western Electric Company, 1937–. Mr. Padowicz has been engaged in metallurgical work, wire drawing and cable development since joining the Western Electric Company.