The Bell System Technical Journal

Vol. XVII *October, 1938* *No. 4*

Ultra-Short-Wave Transmission and Atmospheric Irregularities

BY C. R. ENGLUND, A. B. CRAWFORD AND W. W. MUMFORD

Results of an ultra-short-wave fading study are here reported. Transmission was carried out in the range of 1.6 to 5.0 meters, over a 70 mile (112.6 kilometer) ocean path, on 106 days during a period of two years. Both horizontal and vertical polarizations were used and during part of the time a 6-megacycle amplitude, 120-cycle, frequency modulated transmission was added, for the cathode-ray tube observation of the frequency characteristics of the radio path. On 45 mornings records were taken, on vertically polarized radiations, during the flight period of the Mitchel Field Weather Bureau plane.

Fading was found present practically all of the time. Amplitude changes up to 40 db and fading rates up to 5 fades per minute were found. Simultaneous transmission of the same wave in two polarizations, and of two waves of different wave-length in the same polarization showed that the horizontally polarized component was practically always, and the shorter wave-length one was usually the worse fader of the pair. The greater part of the time there was no correlation between the fading of these radiation pairs; occasionally, however, and for the slow, smooth amplitude, undulating type of fading, coincidence was observed. The frequency sweep patterns showed multiple signal components to be present, with various degrees of relative phase retardation.

A tentative explanation is proposed for these phenomena. This theory assumes the presence of a refracted-diffracted signal component, transmitted along the earth's surface and calculable in the manner of Wwedensky, Van der Pol and Gray, and one or more signal components reflected from air mass boundaries. The airplane results are shown to be in reasonable agreement with the frequency sweep observations. Boundary heights from 5.5 kilometers down to 1.9 kilometers are measured; below 1.9 kilometers other boundaries are indicated. The receiver band, flat over two megacycles, sets the low height limit of resolution of reflecting boundaries at 1.9 kilometers. Most of the boundaries are at the lower heights.

A discussion is given of some observations of signal fading at various wave-lengths which have been reported by other observers, and which are apparently referable to the same mechanism as is here proposed.

Introduction

IN an earlier paper [1] experimental data were presented which indicated that the transmission of ultra-short-wave signals was dependent upon the state of the atmosphere, in particular upon its water vapor content. The present paper contains the results of a continuation of this work where a two-year survey of ultra-short-wave transmission over a 70-mile (112.6 km.) ocean path was carried out. Transmission was had on 106 days during this period.

In planning this work, preparation was made for seeking a correlation between atmospheric structure and signal intensity; but from the very first transmission fading was found, and this fading was so persistent and intense that the work became essentially a fading study.

In the following paragraphs there are discussed, in the order named, Antennas and Locations; Apparatus and Operation; General Characteristics of Fading, with samples of records taken; Polarization Effect on Fading, with sample records; Wave-length Effect on Fading, also with records; Distance and Antenna Height Effects on Fading; Frequency Sweep Patterns of Fading, with sample records; and the logs taken during the flights of the U. S. Weather Bureau airplane for taking free air data. The presentation of experimental data is then interrupted to present a theory which explains several of the experimental observations. This is followed by further experimental results and checks, and concluding remarks.

Antennas and Locations

Figure 1 shows the layout of the radio circuit. The transmitter was erected at Highlands, New Jersey on the edge of a steep hillside. This edge made an angle of about 45° with the transmitter-receiver direction. Below the edge of the hill lay a strip of land slightly above sea level (seven to eight feet) and beyond was Sandy Hook Bay. The altitude at the antenna foot was 119 feet. The antennas consisted of a vertical rhombic terminated in its surge impedance with carbon lamps, a horizontal rhombic with the same termination, an unterminated inverted "Vee" antenna and a half-wave doublet. This doublet was equipped with a flexible transmission line which permitted it to be raised to the top of the antenna supporting mast. These antennas were supported on a central 60-foot (18.3 meter) lattice mast surrounded by four 30-foot (9.15 meter) poles.

The receiver was located on a plot of land at East Moriches, Long Island, New York. This plot was immediately at the edge of Moriches Bay and was only slightly (approximately four feet) above sea level. The same antenna equipment was supplied here as at the transmitter. Except for the transits across Sandy Hook, Fire Island Beach and Smith Point, the wave path was over sea water. A second receiving site at West Sayville, at the edge of Great South Bay, was briefly occupied, using portable receiving equipment. This site was 52¾ miles (85 km.) from Highlands.

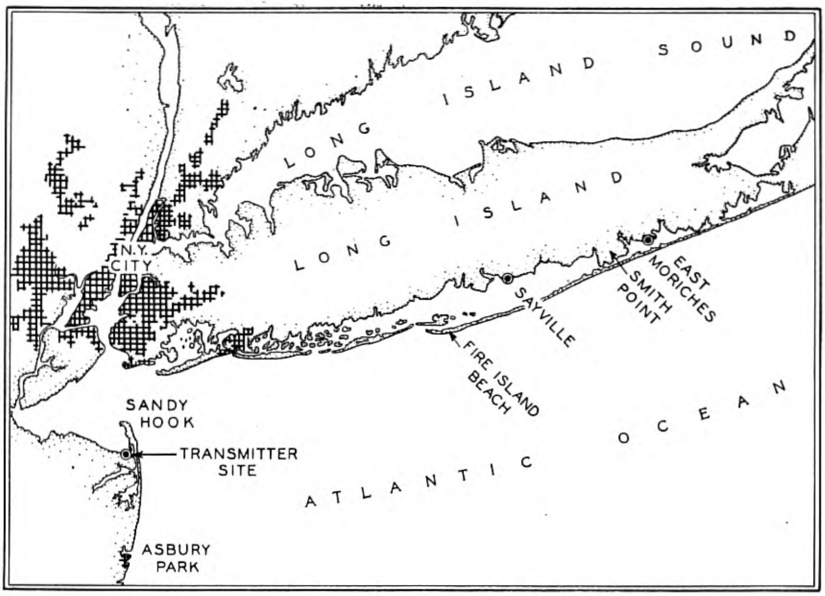


Fig. 1—Map of ultra-short-wave transmission path between Highlands, New Jersey, and East Moriches, Long Island.

Apparatus and Operation

In all, three transmitters were installed at Highlands. The first one, of 100 watts output, covered the wave-length range of 5.0 to 3.5 meters. It was equipped with a motor-driven single-turn short-circuit loop which, coupled with the tank circuit coil, produced a 120-cycle frequency modulation of six megacycles amplitude. For calibration purposes there was added a low-gain double-detection receiver which used an intermediate frequency of one megacycle and was connected so as to pick up an input from the transmitter. The beating oscillator of the receiver was set for the center of the transmitter frequency sweep and the receiver output triggered a gas tube connected to the transmitter tube grids. The transmitter grids thus received a voltage pulse each time that the transmitter frequency passed through one megacycle above or below the beating oscillator frequency. Each transmitter frequency sweep was thus marked with two pulses spaced two megacycles apart.

The second transmitter had Lecher wire tuning elements, covered the wave-length range of 3.5 to 1.2 meters and had a power output of 30 watts at 1.5 meters. It was in operation simultaneously with the first transmitter for six months and then was replaced by transmitter No. 3.

The third transmitter was coil tuned, covered the wave-length range of 4.9 to 2.8 meters and had a power output over this range of 55 watts down to 35 watts. It was operated simultaneously with the first transmitter except for the first six months.

All three transmitters were aıranged for voice modulation through a simple grid input, and the first one was thus used for one-way communication during the entire period of operation.

Normally, unmodulated waves were transmitted and were observed as rectified direct current in the output of the double detection receivers. These receivers had attenuators, variable in steps of 1 db, in the intermediate frequency amplifier circuits and the attenuators were geared to the pens of manual recorders. The operators kept the output current constant by means of the attenuators just mentioned, and there resulted a record of signal amplitude versus time. Some use was made of the Esterline-Angus type of milliampere recorder for automatic recording but no linear scale recorder of this type could handle the amplitude range of the fading encountered.

For the reception of the frequency modulated transmission a tuned radio-frequency receiver, with a three-megacycle band-width centered on 66 megacycles (4.55 meters), was constructed and its rectified output was applied to one pair of plates of a cathode ray oscillograph. A linear sweep voltage, manually synchronized with the transmitter 60-cycle power voltage, was applied to the second pair of plates. The oscillograph pattern thus pictured the frequency-amplitude characteristic of the radio circuit in toto. Over the frequency range where the receiver band was flat (two megacycles) the curve gave the apparent ether characteristic. With a motion picture camera this characteristic was permanently recorded.

Fading Characteristics, General

The fading was always slow compared with that observed on short waves. Except for the rapid fluctuations produced by airplane reflections, a record speed of $\frac{5}{8}$ inch (1.6 cm.) per minute was sufficient. This was our standard speed. Amplitude changes up to 40 db and fading rates up to 5 fades per minute were observed.

It is difficult to describe the fading in any other way than by the records. From a transmission standpoint a curve giving the per cent of time during which the signal is above the abscissa value is useful.

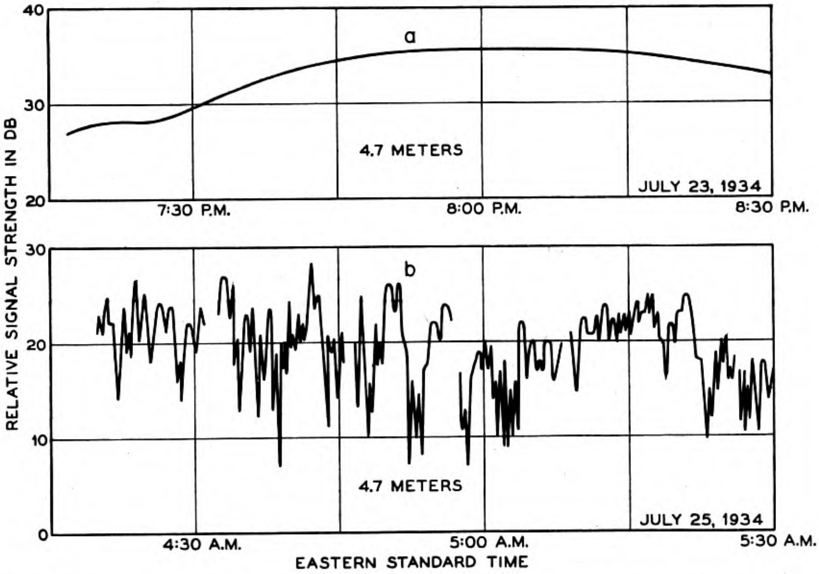


Fig. 2—Fading extremes, vertically polarized transmission; inverted "V" antennas.

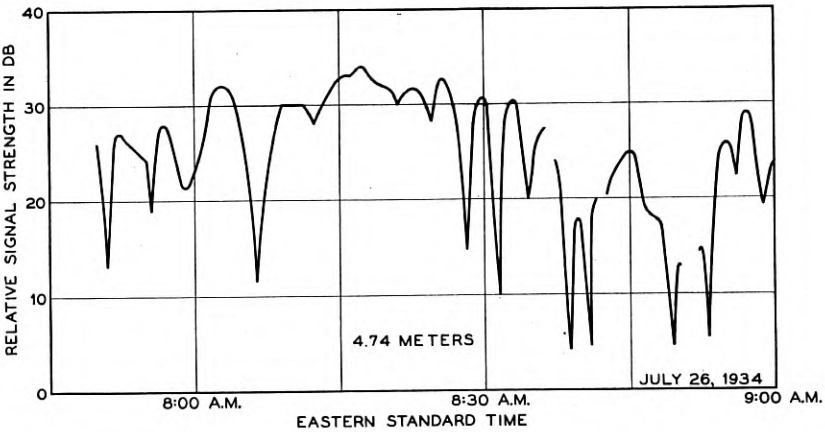


Fig. 3—Extreme amplitude, normal fading rate, vertically polarized transmission; inverted "V" antennas.

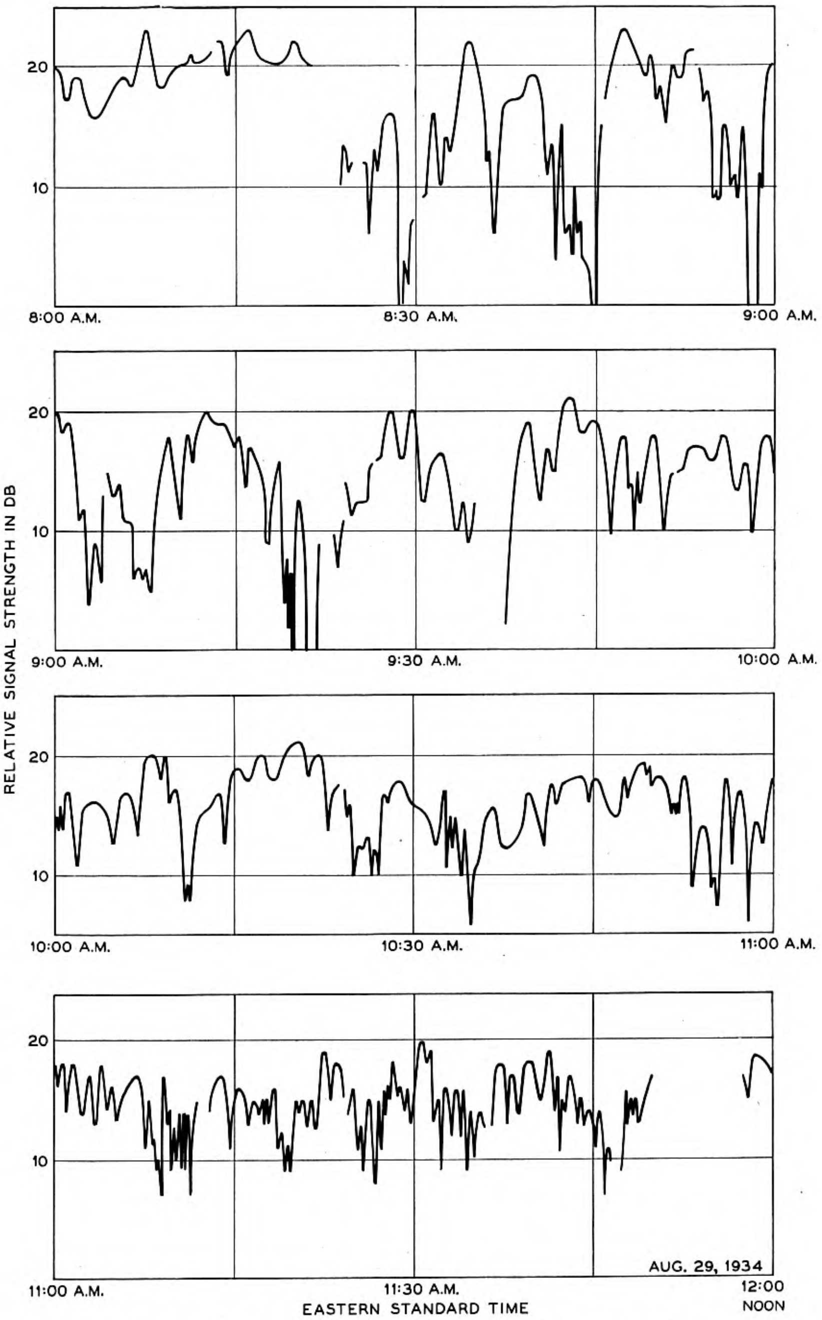
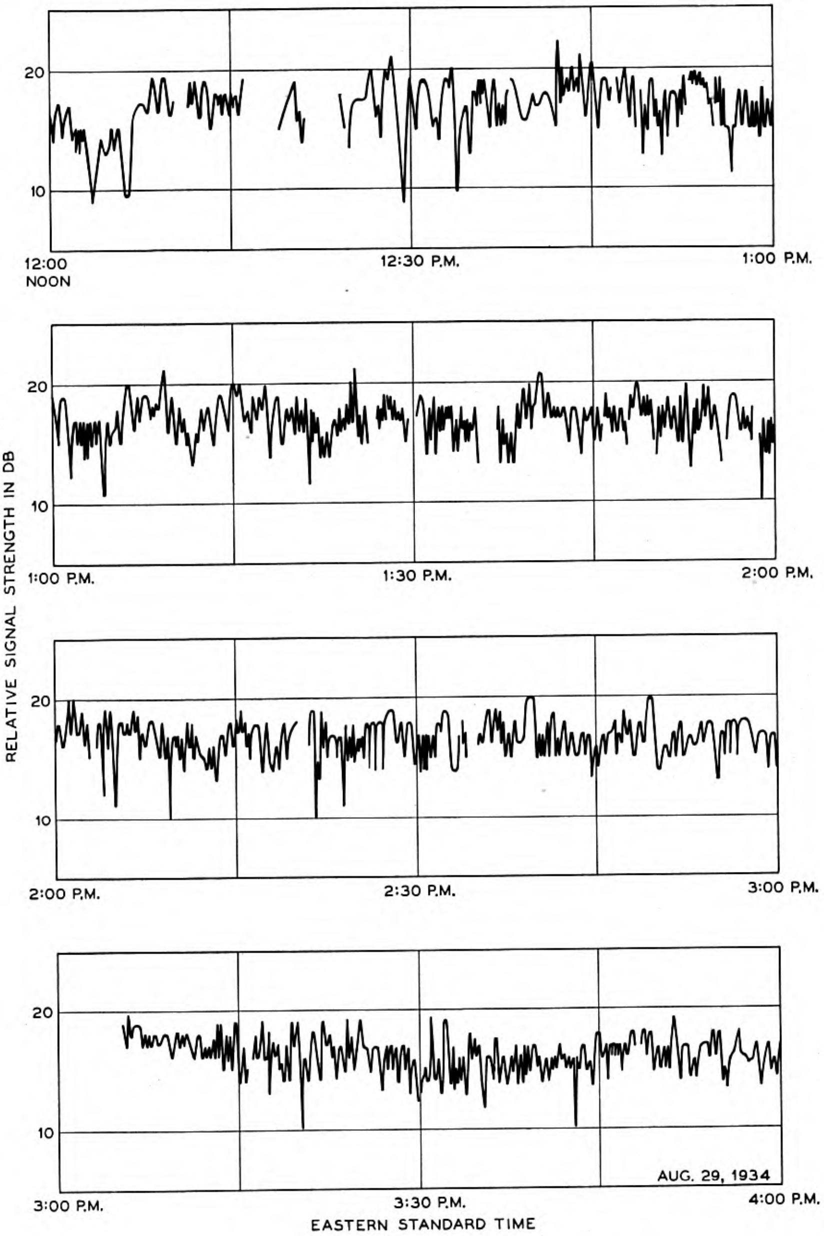


Fig. 4—Development of "scintillation" fading on vertically polarized



transmission, 4.74 meters wave-length; inverted "V" antennas.

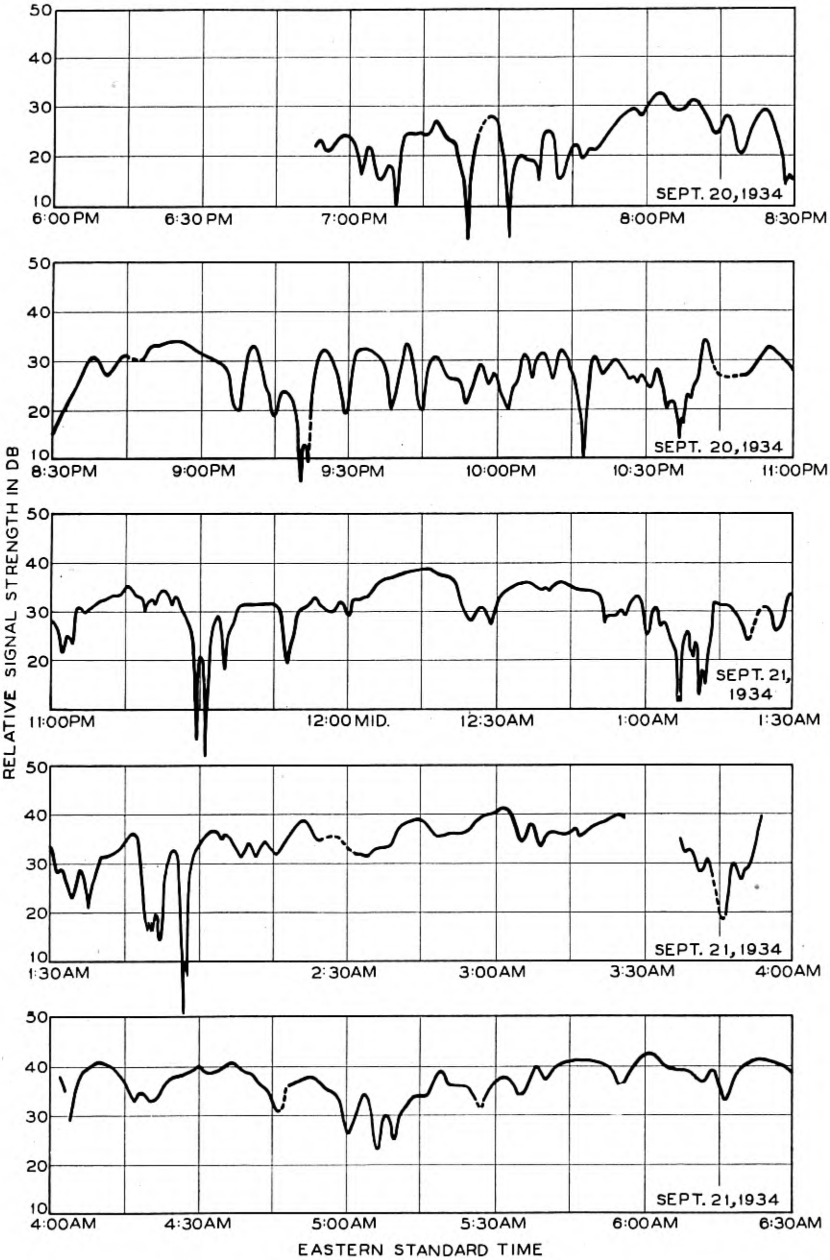
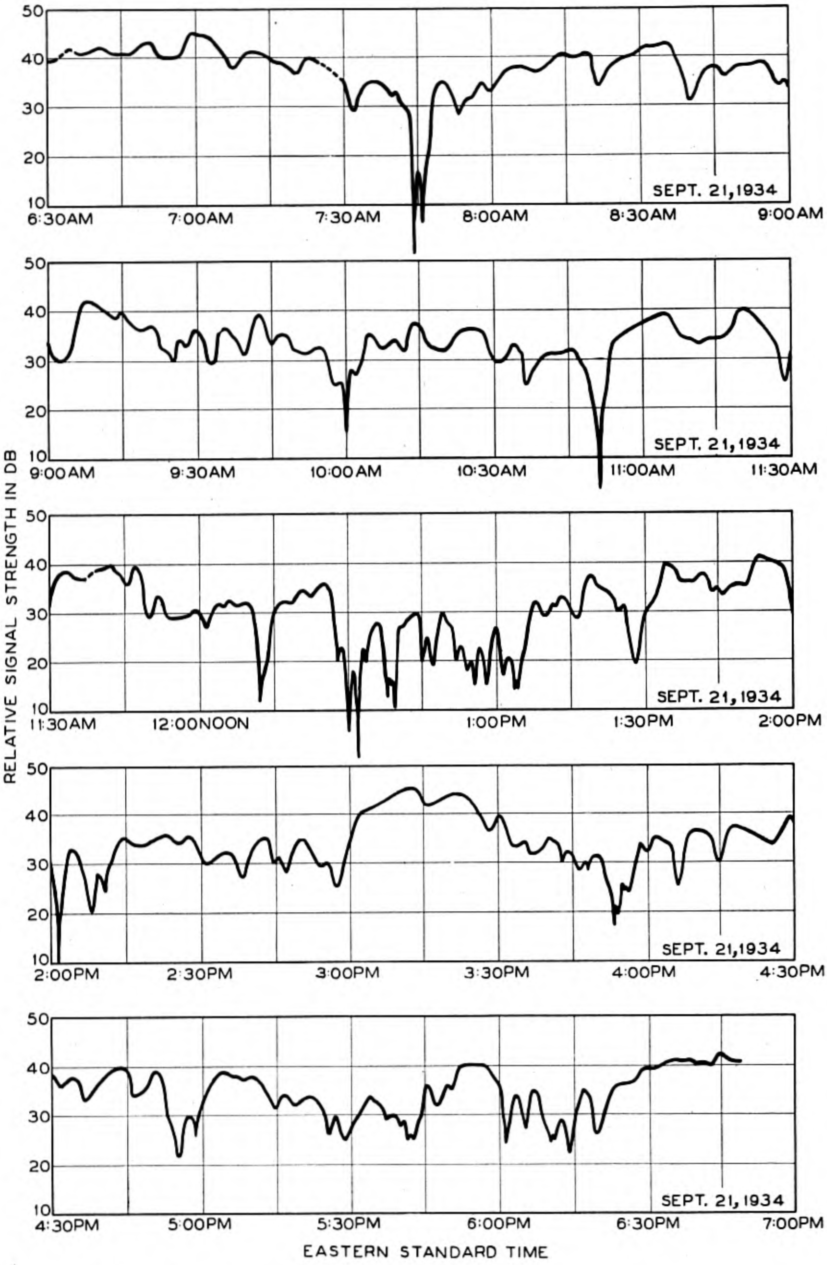


Fig. 5—Twenty-four hour run, vertically polarized



transmission, 4.74 meters wave-length.

Such a curve can also serve to check on the theoretical explanation of the cause of fading in certain cases. Thus if the fading is due to the combination of two radiation components in assigned random amplitude relation and arbitrary or random phase relation, a curve can be calculated from probability considerations and compared with the experimental curve.[2] Such a simple mechanism was inadequate for our fading most of the time. Moreover, the fading changed enormously from day to day. It is hoped that the samples given in the figures will give an adequate idea of this phenomenon.

Only rarely was fading practically absent for periods of an hour or two. Such a period is illustrated in Curve *a*, Fig. 2. Two days later the extreme fading of Curve *b* was recorded. It is significant, as will later appear, that the non-fading situation was the one of higher signal. The amplitude range of curve *b* is nearly normal; the fading rate is much greater than normal, for vertical polarization. In Fig. 3 the fading rate is normal but the amplitude range is excessive. In Fig. 4 a characteristic type of fading, which we have termed "scintillation," is recorded. In this case the fading, initially erratic and of a fairly wide amplitude range, subsides in a characteristic manner to a steady, fast rate oscillation, or scintillation, of moderate amplitude.

In Fig. 5 a 24-hour run is recorded. The rambling erratic character of the fading is well shown here. Characteristic deep short-period minima occur at intervals, occasionally they are twinned, some of them have a fine structure at the bottom. There are several "dropouts" where the signal practically disappeared.

No sunrise-sunset variations in fading were noticed, though looked for. Diurnal variations could not be established since automatic recording was not available. A seasonal falling off in average signal was noticed in the winter; the 1.6 meter wave, because of its normally low level, dropped below the noise level in the winter of '34–'35. No effect of ocean waves, clouds, or other visible weather phenomena could be established. It is true, however, that to be certain of the non-effect of such phenomena as clouds, a cloud observer at the midway point should have been present. In so far as cloud layers make air mass boundaries visible they may well affect the transmission. Cloud bottoms which represent merely the adiabatic dew point level should apparently not cause much signal reflection at these wavelengths.

Effect of Polarization on Fading

After some preliminary experimenting it was found that comparisons of two transmissions were worthless unless made on simultaneous recordings. The recorders were therefore fitted with telechron motors operating on a circuit of the Patchogue division of the Long Island 60-cycle power network. The resulting timing was faultless and by transmitting the same radiation on crossed antennas, and receiving the vertical and horizontal components separately, a comparison was obtained.

In general the horizontal component showed the worse fading, more fades per minute and greater amplitude range. This was always true when the fading on vertical polarization was bad. There was then no noticeable coincidence between the two. When the fading had a smooth long period fade, or "roller," superposed on a short period oscillation, or "fine structure," there was at times coincidence between the roller components. Occasionally, with fine structure absent and

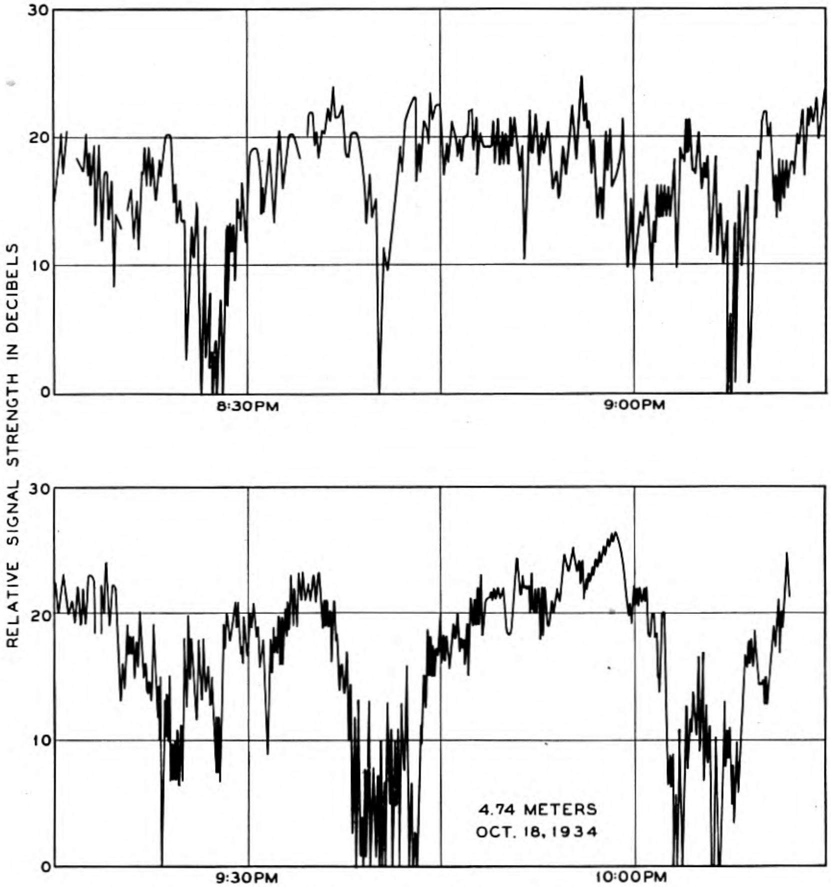


Fig. 6—Composite bad fading, horizontally polarized transmission.

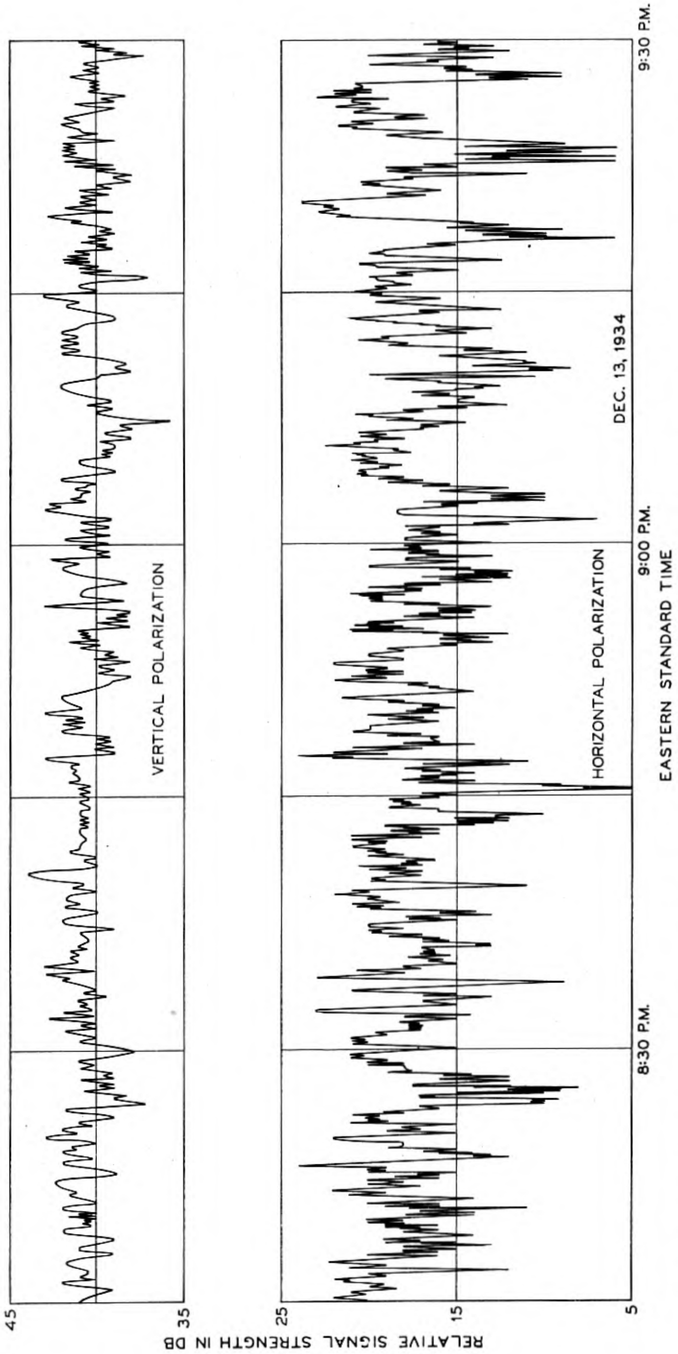


Fig. 7—Comparison of simultaneous bad fading on horizontally and vertically polarized transmissions, 4.74 meters wave-length.

moderate roller fading, a good coincidence between the two records resulted. This is discussed later.

Figure 6 is a sample of fading on horizontal polarization, at its worst. This particular specimen shows the superposition of roller and fine structure fading very well. No vertical-polarization record was taken along with this. Figure 7 shows a typical example of fading simultaneously observed on vertical and horizontal polarization during bad fading conditions. There is no coincidence. Figure 8, on the other hand, records an unusual condition when a mild roller type of fading shows a good coincidence on two polarizations.

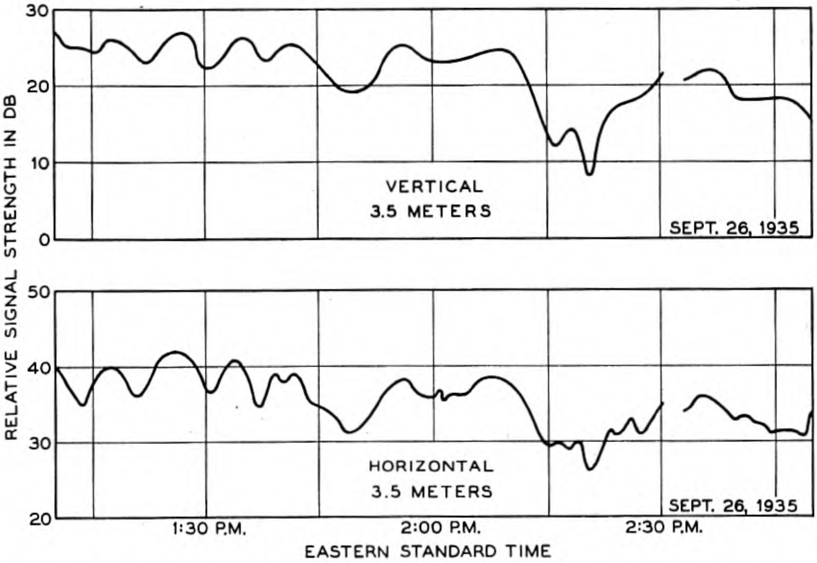


Fig. 8—Comparison of simultaneous mild roller fading on horizontally and vertically polarized transmissions.

Effect of Wave-length on Fading

The double wave-length records are not as contrasty as the double polarization ones. In general the shorter wave has the worse fading, either as higher fading rate, greater amplitude oscillation or both, and the greater the wave spacing the more certain this is to be true. Exceptions have occurred, however, where the fading was much the same, and one record was obtained where the fading rate on 4.7 meters was noticeably greater than on 4.5 meters.

Our first simultaneous records were taken at a wave-length ratio of 3 to 1 (4.7 to 1.58 meters) where the fading on the shorter wave was always worse. The remaining observations were confined to wave-length ratios of 1.5 to 1 and less. A comprehensive set of records was obtained for moderate to small wave-length spacings, down to 1 per cent difference. These records are all for vertical polarization. The few records taken on horizontal polarization happened to be obtained when the horizontal fading was much worse than the vertical fading and the records are too rough for good comparisons.

For these small wave-length-difference records the types of fading are more likely than not to be similar on the two wave-lengths. That is, the fading rate and amplitude excursion will average up much the same. More rarely, there will be a similarity between the two fading tracks which is evident to the eye, sometimes as a "retarded" similarity. Occasionally, and usually on the roller type of fading, there will be a marked coincidence between the two records; this coincidence will be better the milder the fading and the smaller the wave-length spacing. Genuine identity was never recorded on different wave-lengths even down to 1 per cent difference. With scintillation, coincidence was difficult to demonstrate; a similarity on the major swings was all that was shown.

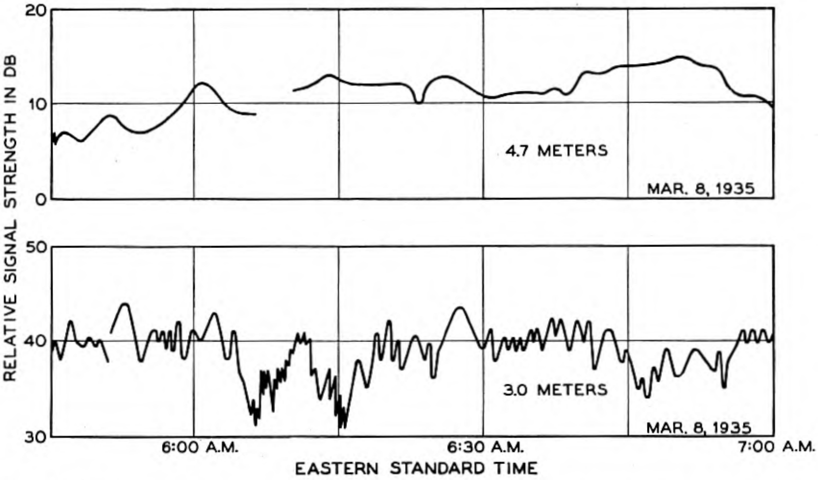


Fig. 9—Comparison of simultaneous fading on two well spaced wave-lengths, vertically polarized transmission.

Figure 9 shows a very marked difference between 4.7 and 3.0 meter fading. This is one of our most contrasty records. Figure 10 shows very slow fading, on two occasions, with wave-length differences of approximately 1 and 4 per cent respectively. There is good coincidence. Figure 11 shows active fading on short rollers for 4.7 and 4.65 meters, a wave-length difference of approximately 1.1 per cent. There is agreement in major features. Figure 12 shows a case of scintillation superposed on mild rollers, again for 4.7 and 4.65 meters. The time scale is here magnified three times. An in and out similarity can be seen, especially for the rollers. In the section on theory these similarities are further discussed.

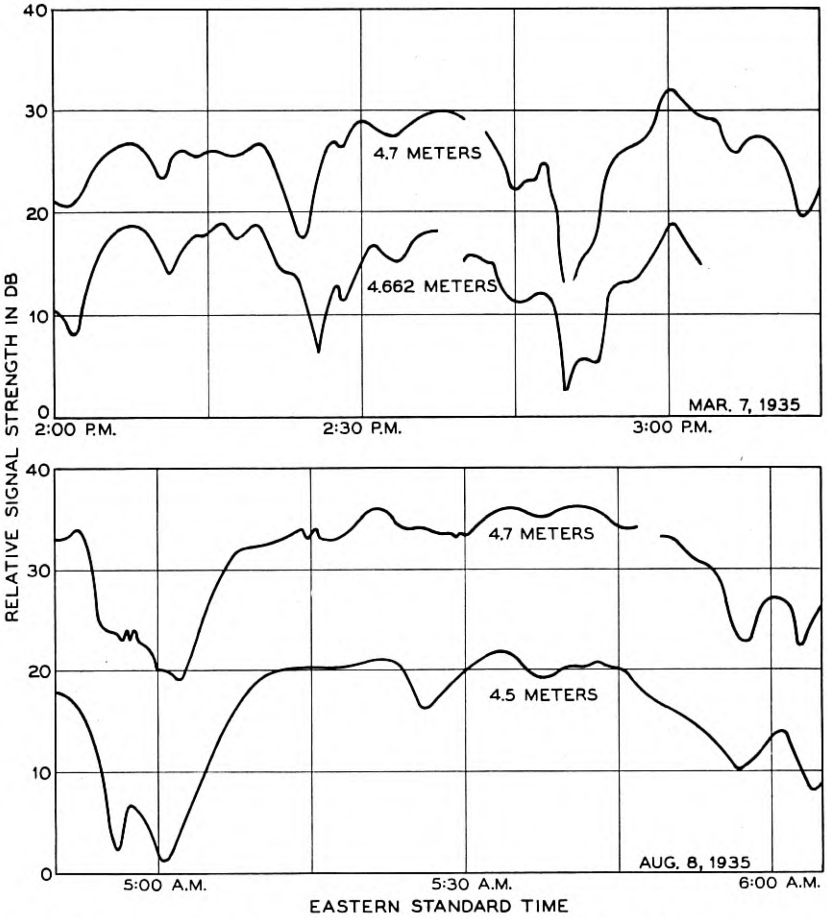


Fig. 10—Comparison of simultaneous slow fading on two slightly different wave-lengths, vertically polarized transmission.

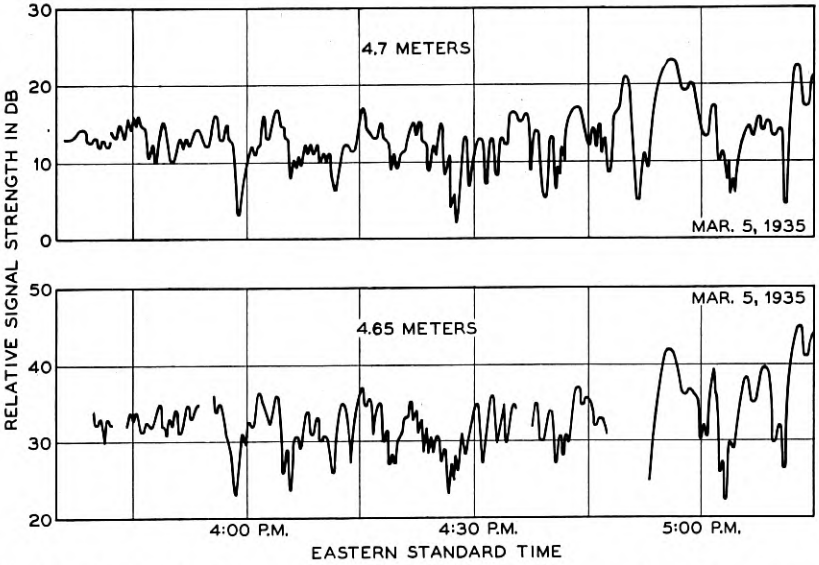


Fig. 11—Comparison of simultaneous active fading on two slightly different wave-lengths, vertically polarized transmission.

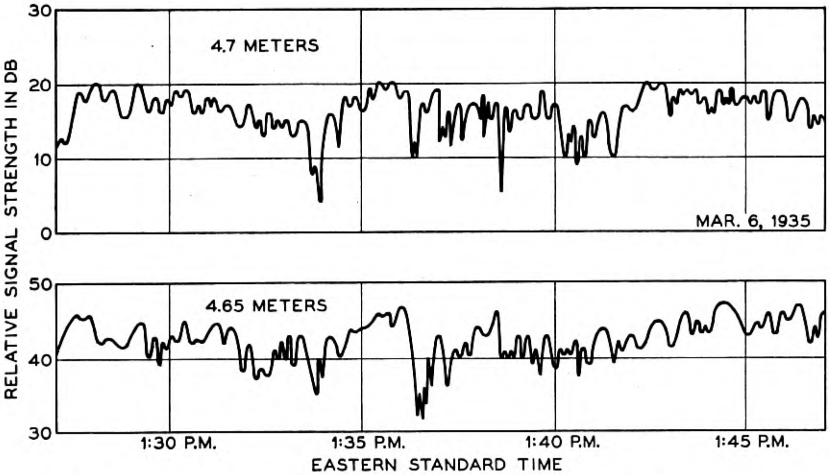


Fig. 12—Comparison of simultaneous "scintillation" fading on two slightly different wave-lengths, vertically polarized transmission. The time scale has been expanded.

Effect of Distance and Antenna Height on Fading

In planning this work a survey for a receiving site was made by means of a portable receiver in a car. Later, simultaneous reception was had at East Moriches and West Sayville, on three days. The survey data were not sufficient to establish any proposition beyond the statement that the signal strength fell rapidly with distance, with the intensity of fading coming up as the signal fell. The simultaneous two-distance recording showed random fading between the two records with less fading amplitude at the shorter distance. The fading rate was about the same. Unfortunately the recording took place under scintillation conditions, thus giving very poor records for comparison purposes.

By mounting two linear doublets on the 60-foot lattice mast simultaneous recording at two heights was carried out. For the two doublets (horizontal, at 14 and 52 feet respectively), a signal level difference of 12 db was observed, in favor of the higher antenna. The fading on the two records was identical. It may be added that, on calibrating the car receiver at East Moriches before moving to West Sayville, identical fading records were obtained with the two antenna systems 150 feet (45.7 meters) apart and substantially broadside to the radiation.

Frequency Sweep Patterns of Fading

The frequency sweep patterns were of many types, from slow to fast fading and from shallow to deep fading. Apparent path differences from 600 meters down to a few meters occurred. The patterns were usually complicated, indicating that more than two components were present. There is no reason to believe, however, that they were not all due to wave interference.[13]

On three days the predominant pattern was simple enough to be referable to two waves with a path difference consistently greater than 75 meters. These will be referred to later. In Figure 13 are given three sample runs illustrating a two-component pattern, a three-component pattern with two of the components forming a small path difference pair, and a multiple component pattern. The receiver characteristic is dotted in on one curve of each set.

Logs During Weather Bureau Airplane Flights

On forty-five mornings recording was carried out during the period of flight of the Mitchel Field Weather Bureau plane. This plane takes off about dawn every morning, when flying is possible, and by means of a meteorograph obtains records of air pressure, temperature and humidity, up to an altitude of about five kilometers. A record of the fading, on 4.7 meters and vertical polarization, was obtained for each of these mornings. In addition, on twenty-six mornings frequency sweep patterns were photographed at or shortly after the time of flight. These sweep patterns were all on horizontal polarization.

From the meteorograph data, kindly furnished us by the United States Weather Bureau, the dielectric constant of the air has been calculated [1] and plotted as a function of the altitude. On twenty-four days there were, above an altitude of 400 meters, changes in the dielectric constant curves equivalent to discontinuities of $\Delta\epsilon \geqq 10^{-5}$. Heights up to 3200 meters were recorded for these. Typical curves

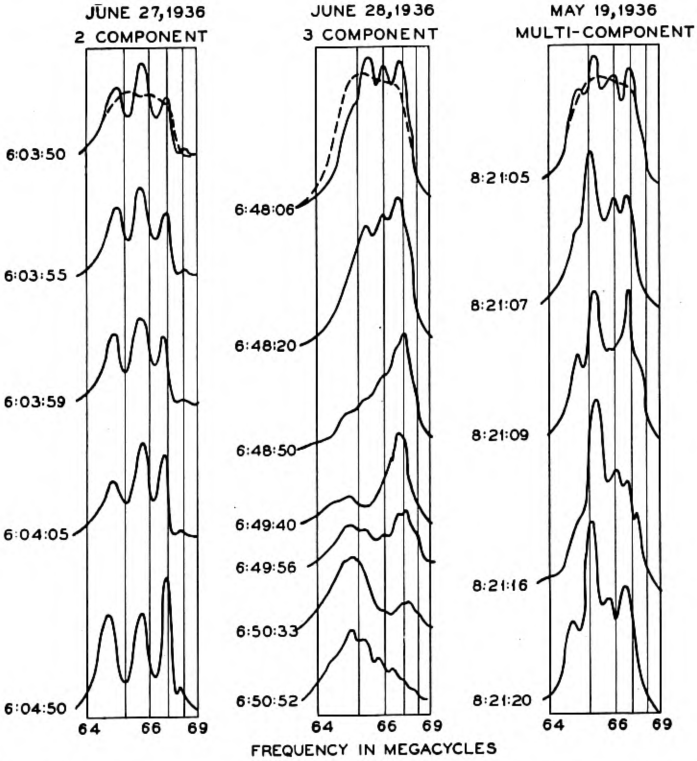


Fig. 13—Three sequences of frequency sweep patterns. Horizontally polarized transmission, 4.55 meters mean wave-length.

are given in Fig. 18 on the left-hand side. On four days there were small boundaries with $\Delta\epsilon < 10^{-5}$; on two days there were possible but not definite boundaries, the experimental points being too widely separated in altitude for precision; on five days there were possible boundaries below 400 meters altitude and on ten days the refractive index-height relation was an approximate exponential one without any evident boundaries. These data will be referred to later.

Theory

The fading phenomenon was explicable in several ways. In our previously cited work [1] we found that variable atmospheric refraction was present, the airplane carried receiver being up where the refracted-diffracted field strength was high and dominant. In general variable refraction would be expected to be a slow phenomenon, operating in hours, or even days, rather than in minutes, and much too slow to explain five-cycle-per-minute oscillations, for example.

Another explanation was air-mass boundary reflection (or refraction),[3] such a boundary readily explaining the rate of signal variation. No Kennelley-Heaviside layer reflection was in question; this had been quickly ruled out by the experimental data. When, therefore, we elected to transmit the frequency modulated signal, already described, and the oscillograph revealed a cyclic maximum-minimum frequency characteristic of the other path itself, it was evident that there was no possibility other than wave interference left—interference presumably between a direct-diffracted and one or more boundary-reflected components.

These boundaries have apparently not been positively identified at longer wave-lengths and for that reason we have tried to get some further experimental contact with them. Attempts, since the closing down of the Atlantic Highlands-East Moriches circuit, to demonstrate an air-mass boundary, any boundary whatever, by high-angle transmission, have failed. No reflected components have appeared. Of course an illy defined, or diffuse, boundary will operate in this manner since only for near grazing incidence can such a boundary give the appearance of a discontinuity for the incident radiation.

If we assume such a boundary a few kilometers up, and assign to it a relatively small discontinuity in index of refraction, compared with that of an earth or sea water boundary, then the four components of Fig. 14 will be the only important boundary reflected ones for a radio circuit such as ours. We now, fortunately, have theoretical formulæ [4, 5] for computing the diffraction of an ultra-short-wave radiation around the earth and the amplitude of the direct-diffracted component can be calculated at once.

That is, it can be calculated at once if the air mass has no refractive bending effect upon the radiation trajectory. Since such a bending effect is certainly present at times, and is equally certainly variable, even if only slowly, it must be taken into account.

If the refractive index of the air varies as a power of the distance to the earth's center, it has been shown [6] that the actual state of affairs can be duplicated by a homogeneous atmosphere over an earth, the radius of curvature of which is greater than that of the actual earth and is calculable from the exponent of the height variation function. With this "effective" earth radius, the formulæ already mentioned become usable. If the air refractive index does not vary as a power of the distance to the center of the earth we must take that exponent which gives the best first order fit over the height covering the refracted wave front, the alternative being a prohibitive complication of the theory.

A plausible physical picture of the fading mechanism can now be set up. If we lump the four boundary reflected components in one, and plot as a function of the distance, we have curves "A" of Fig. 15.

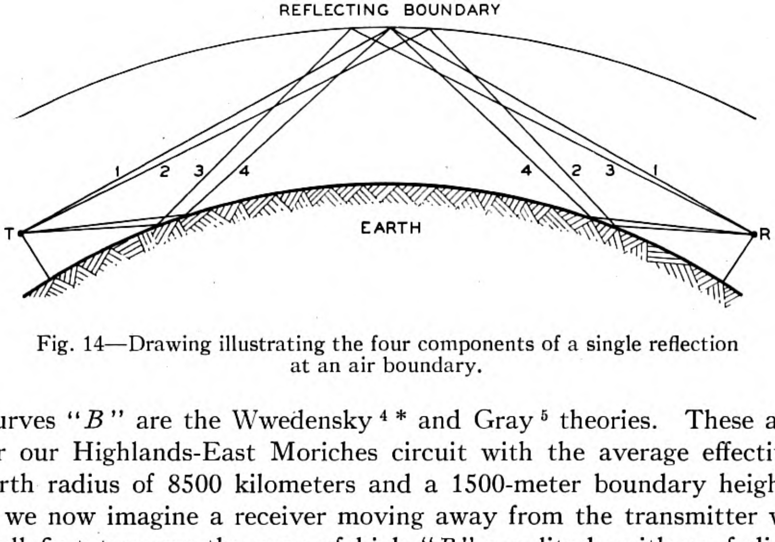


Fig. 14—Drawing illustrating the four components of a single reflection at an air boundary.

Curves "B" are the Wwedensky [4]* and Gray [5] theories. These are for our Highlands-East Moriches circuit with the average effective earth radius of 8500 kilometers and a 1500-meter boundary height. If we now imagine a receiver moving away from the transmitter we shall first traverse the zone of high "B" amplitude with no fading present. The signal amplitude will, for any given near-by point, and for any given antenna ampere-meters, depend on the height of the antenna above the ground and the ground constants. As the distance to the transmitter increases, the falling "B" curve approaches the rising "A" curve in ordinate and we enter a disturbed region where, for any instability of the boundary, more or less complete interference can result and fading will occur. (One such instability occurs when

* There is an error in the formula, as given by Wwedensky. It is corrected here. See appendix II.

a boundary with an irregular surface is carried past the reflection zone by the normal motion of the atmosphere.) A further increase in transmitter distance and the "B" or residual curve drops out of the picture leaving only the "A" curve and, presumably, fairly steady signal amplitude conditions. The location of these zones of undisturbed and disturbed signal will vary from day to day as: (1) the reflection coefficient and height of the layer change, (2) the effective radius of the earth changes. The effect of the height of the layer is shown in Fig. 16.

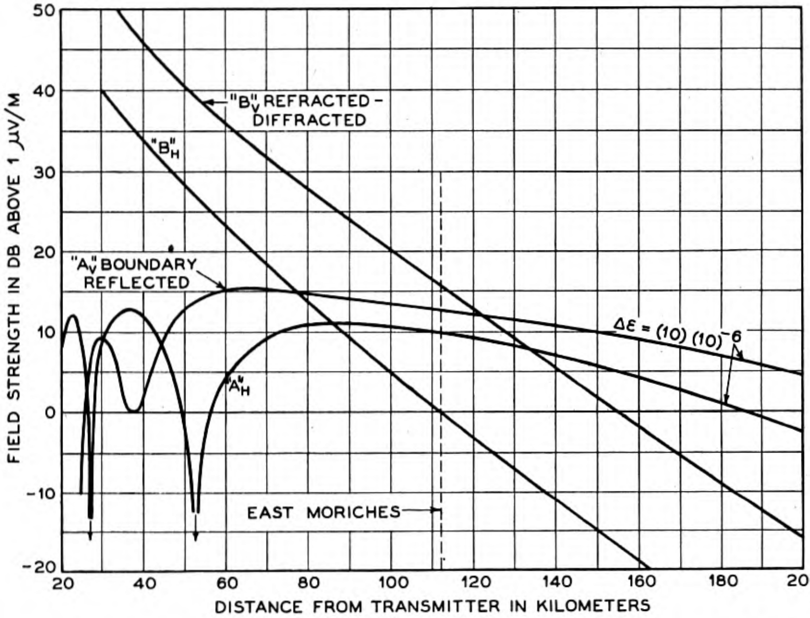


Fig. 15—Calculated curves for air boundary reflected and earth refracted-diffracted radiation components, in both vertical and horizontal polarization. Short doublet antennas, 1 kw. power radiated, wave-length 4.7 meters, $\sigma = 5 \times 10^{-11}$ (E.M.U.) and $\epsilon = 80$ for sea water. Height of transmitter antenna 42 meters, of receiver antenna 5 meters, air boundary height 1500 meters, effective radius of earth 8500 kilometers.

Since the major lobe of the polar characteristic of any simple antenna, such as ours, is directed forward and away from the earth, the signal intensity at the reflecting boundary surface will be comparatively high and will, in some measure, make up for a small reflection coefficient. For longer waves, such as broadcast waves, the high level of the "B" curve will move the disturbed zone so far out that the low residual signal level and the Kennelley-Heaviside layer reflections will conceal or mask the atmospheric boundary reflections. Several observations which can be ascribed to such boundaries have nevertheless been published.[7, 8] Obviously, only boundaries lying considerably higher than those discussed here will give the path differences to produce the same type fading at these longer waves. At the same time the apparent diffuseness of a boundary will fall off with increase in wave-length, thus removing the restriction of reflection to near grazing incidence angles only.

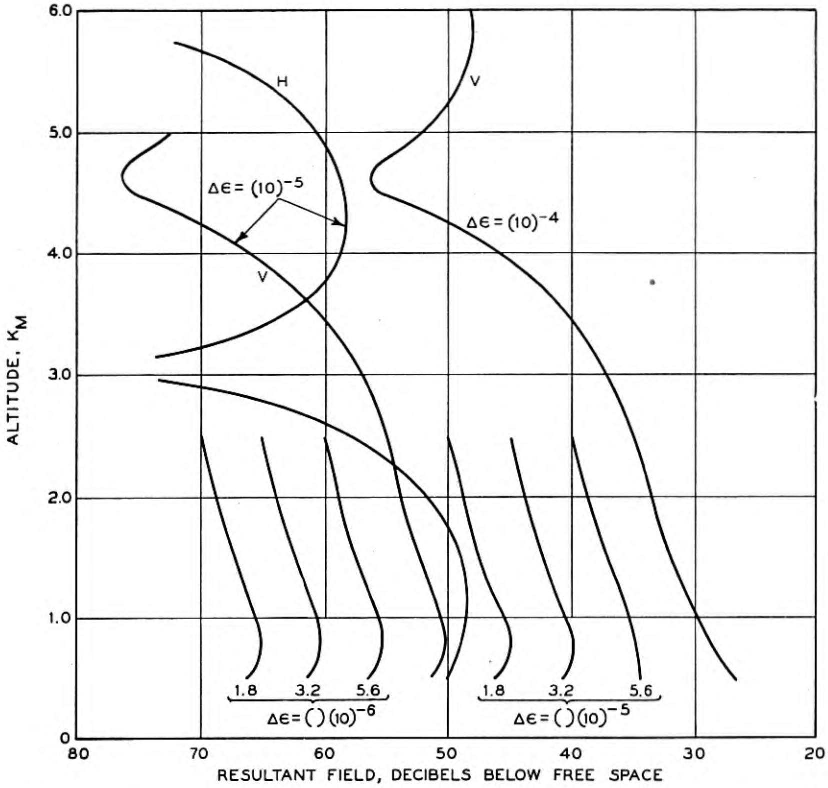


Fig. 16—Calculated field strength curves showing the effect of air boundary height and density on the reflected radiation component, for the Highlands East Moriches circuit. Transmission path 112 kilometers, over sea water, wave-length 4.7 meters, polarization both vertical and horizontal. Vertical antennas 42 and 5 meters high, horizontal antennas 45 and 9.5 meters high, respectively.

This tentative mechanism also explains several other observed features. Thus, for a given type of boundary instability, the fading rate will increase as the wave-length decreases. Furthermore, since the slope of the "*B*" curve increases as the wave-length decreases, the disturbance zone is effectively moved nearer the transmitter and the probability of increase in fading amplitude is enhanced. The usual increase of fading with decrease in wave-length is thus explained. When the wave-length difference is small, on the other hand, the fading type should be much the same on both wave-lengths, as was generally found. The lack of coincidence would arise from the fact that the path difference being a considerable number of wave-lengths, a small wave-length change can introduce a marked randomness in fading without appreciably affecting the type.

As has been mentioned earlier, a multiple of reflecting boundaries is the normal condition, rather than that of a single boundary. This circumstance, without invalidating the explanations already given, makes a further elaboration of the theory possible. The "roller" type or component of fading, in particular, requires explanation. In addition to the smooth signal modulation, from which the name has been derived, this type of fading is characterized by showing more or less frequent deep minima or drop-outs and these are often distinctively twinned. Further, the roller component is that component of fading which shows coincidence, in spite of wave-length or polarization differences. Such coincidence indicates small path difference and this is what we have when a double boundary or stratum exists. Such a stratum would give two "*A*" components and, if of variable thickness, would, as it was carried along by the prevailing air currents, give the steep, deep, minima at phase opposition thickness. Further, if the stratum contour were that of a hump, thick enough to carry the second "*A*" component past phase opposition to the first one, the twinned minima would result as the hump entered and left the reflection zone. Occasionally the two "*A*" components would add properly, with the residual "*B*" component, to give complete extinction, a result less likely from the phase addition of a single "*A*" and the "*B*" component.

This explanation of "roller" fading assumes, tacitly, that the "*B*" component is, at the time, relatively subdued, that is, the disturbance zone has moved inwards due to an increase in the reflection coefficient of the layers or to a decreased "effective" earth radius. The fine structure often appearing at the bottom of a prolonged roller minimum corroborates this, the mutual cancellation of the two "*A*" components having uncovered, so to speak, the weaker "*B*" component with its much shorter traversed path.

With the "roller" condition characteristic of high "*A*" component signal amplitude, the "scintillation" condition would be characteristic of low "*A*" component signal amplitude, the relatively steady "*B*" component having superposed on it a small amplitude, variable phase, "*A*" component. A relatively low mean amplitude value and the coincidence of scintillation conditions with conditions of convective instability of the atmosphere would thus be explained. All the scintillation records came on days of relatively high wind and convective instability. A turbulent atmospheric condition would dissipate or attenuate any boundaries, especially the lower ones. The rapid flutter about the mean amplitude value is the normal expectation from a high, turbulent, low reflection coefficient boundary.

Our two polarization results are qualitatively explicable on the mechanism proposed. As can be seen in Fig. 15, the change from vertical to horizontal polarization results in a relative lowering of the "*B*" curve without much change in the "*A*" curve, which should result in increased fading. For our circuit and a boundary at 1500 meters the relative "*B*" vs. "*A*" drop is 13 db.

As Fig. 16 shows, the variations of the "*A*" components with height are markedly different for the two polarizations. The "A_V" component falls steadily with height up to 4700 meters; the "A_H" component has a deep and sharp minimum at 3000 meters after which it rises again. Since most of our observations concerned boundaries at 2000 meters or less, this high altitude disparity between "A_H" and "A_V" does not affect our explanation. The disparity between vertical and horizontal fading should be much more marked for high boundaries than for low boundaries.

Further Experimental Curves and Checks

The curves given have illustrated the variability in the fading, a variability which no short period of recording can encompass. The tentative explanations proposed have been shown to be in accord with several of the features characteristic of this fading. Certain other experimental results will now be adduced which offer further verification along somewhat different lines.

For the forty-five mornings on which simultaneous recording was carried out during the United States Weather Bureau plane flight, we have calculated, from the airplane data, the values of the "*A*" and "*B*" components. As stated earlier, there were twenty-four days when boundaries above 400 meters altitude, and of sufficient distinctness to be fairly accurately estimated ($\Delta\epsilon \geqq 10^{-5}$) were shown by the meteorograph records. For these the "*A*" components have been computed. By taking the dielectric constant gradient for the first half kilometer, the effective earth's radius was determined and inserted in the Wwedensky formula to give the "*B*" component. These calculated values

("*A*" component as triangles, "*B*" component as circles) are plotted on Fig. 17 together with the maximum and median * observed values. These latter are joined by lines. For the 10 mornings on which no boundaries were evident the calculated "*B*" component appears to be some 8 db higher than the observed values. With this correction the agreement between observed and total calculated fields is fairly good. A partial explanation of this 8 db discrepancy may lie in the fact that the ocean water trajectory assumed in the calculation differs from the actual one by the land terminals and the three tongues of land intervening.

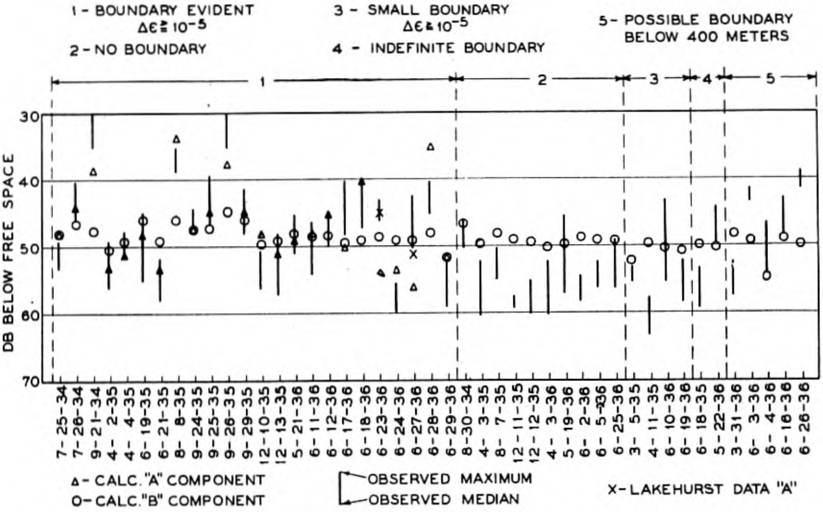


Fig. 17—Comparison of "*A*" and "*B*" radiation components, calculated from the U. S. Weather Bureau free air data, with measured maximum and median signal strengths. Vertically polarized transmission.

On the twenty-six morning frequency sweep runs there were only three on which the predominant sweep pattern was simple enough to be interpreted as due to two components with path difference greater than 75 meters. For those days a series of measurements of the film patterns was made by determining the frequency spacing between a maximum and a minimum and calculating the resulting path difference and boundary height. The dielectric constant-height function was also calculated from the Weather Bureau data. These curves are

* The signal is half of the time greater and half of the time less than its median value. For random phase with "*B*" component equal to "*A*" component the resultant median value signal is $\sqrt{2} \times A$ or 3 db up; it falls from this value to "*A*" as "*B*" decreases to zero.

plotted in Fig. 18 with the calculated boundary heights set down at the right hand, spread out in time of observation. The boundary height coincidence is pretty definitely located in this manner. Many of the more complicated frequency sweep patterns carried a fine struc-

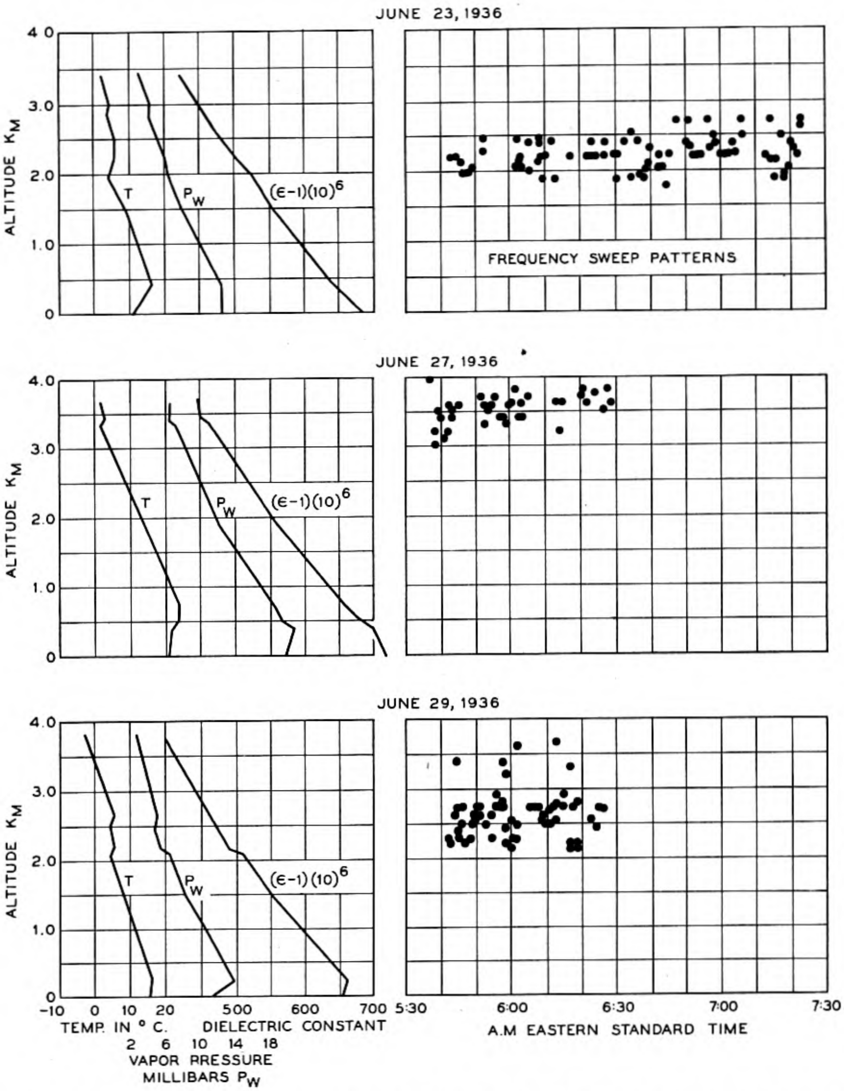


Fig. 18—Comparison of boundary heights shown by the U. S. Weather Bureau free air data, with boundary heights measured from frequency sweep patterns. Horizontally polarized transmission.

ture which indicated weak boundaries at higher altitudes up to, roughly, 5.5 km.; most of the patterns, however, were characteristic of layers below two kilometers. The path difference corresponding to two kilometers is 85 meters. The theoretical limit of resolution of the amplifier band for a maximum to minimum frequency spacing is $\Delta l = 2(C/\Delta f)$ where Δl = path difference, C = velocity of light and Δf = frequency band. For $\Delta f = 2 \times 10^6$ cycles this gives 75 meters, and hence boundary heights at and below 1900 meters are unresolvable by our receiver. It is a remarkable result that the bulk of the disturbing boundaries should lie so low.

It was mentioned earlier that several observations referable to air-mass boundaries have been published. In addition there have been reports, for three consecutive years, of long distance ultra-short-wave reception by American amateurs [9] during the month of May. We have copies of the U. S. Weather Bureau atmosphere cross-sections for several of these days and have been curious enough to examine them. On May 9, 1936, during the long distance amateur reception, there was an extensive boundary at 4 km. between an upper Superior air mass and a lower Tropical Gulf air mass. On May 15, 1937, a similar boundary at 4–5 km. had a Superior air mass above a wedge of Transitional Polar to Tropical Atlantic air. Below this at 3–4 km. lay a Transitional Polar Continental air mass.

On June 11, 1936, when Colwell and Friend [7] report an extra strong 0–2 km. "C" reflection, a subsiding Transitional Polar Pacific air mass lay above a Transitional Polar Continental air mass with the boundary at about 1.5 km. On June 29, 1936, when they reported a very strong 3.5 km. "C" reflection, there existed four wedge-shaped air masses with a Superior air mass over a Transitional Polar air mass at 3–4 kilometers. The wave-lengths used were 186, 125 and 86 meters approximately.

These coincidences may or may not be significant but it is very questionable that any boundaries at such altitudes are due to either electron or gas ion distributions.

The characteristic properties of North American air masses have been published,[11] as average summer and winter values, and show some marked seasonal differences. The greater dielectric constants for summer conditions are due chiefly to greater water content.

For a single air-mass distribution, horizontal stratifications are at a minimum and the radio transmission is via the "B" component. This component can be calculated from the corresponding effective earth radius. The table below gives this radius for three important air mass types.

Air Mass Type	Effective Earth Radius	
	Summer	Winter
Tropical Gulf—T_g	$1.53 \times R$	$1.43 \times R$
Polar Continental—P_c.....	$1.31 \times R$	$1.25 \times R$
Superior—S..............	$1.25 \times R$	$1.25 \times R$

"R" = actual earth radius

The boundaries between different air-mass types furnish discontinuities adequate for radio reflections. The greater the stability of the boundary, the more abrupt it is likely to be. In general, when "S" air overlays either "T_g" or "P_c" air, the resulting boundary is stable. Possible discontinuities, for the three types discussed, may be summarized in the following table. Here the positive sign means that the radiation originates in the more refractive medium. For stability the lower medium is the denser though not necessarily the more refractive.

Altitude	$\Delta\epsilon \times 10^6$					
	Summer			Winter		
	S/T_g	S/P_c	T_g/P_c	S/T_g	S/P_c	T_g/P_c
1.0 Km.............	100	20	-80	55	25	-30
2.0 Km.............	50	10	-40	50	15	-35
3.0 Km.............	30	10	-20	35	10	-25

Concluding Remarks

The characteristics of this seventy-mile circuit indicate that for ultra-short-wave transmission it rates as a long distance one. If we assume that the air refraction is on the average such that the effective earth's radius is 4/3 the actual one, then the receiving station lay 1400 feet below the line of sight from the transmitter. This is equivalent to 0.57° below the horizon. The reception, using high effective-height antennas, was good; there was, however, very little lee-way left, above set noise, for reception with simple doublet antennas. Any longer circuit will require to be terminated on elevations such as to keep the intermediate horizon height down. The fading was too slow to be noticeable on amplitude modulated speech unless a deep minimum or drop-out occurred.

The circuit was probably unusable for television, most of the time. A system adhering to the R.M.A. standard [12] of 441 lines on an interlaced 60-cycle scanning will have a unit time element of 0.17 microseconds. This corresponds to a path difference of 51 meters and only a fraction of this is necessary to produce a ghost. A rough estimate of the boundary height range involved in our fading is one-half to five-and-one-half kilometers. The corresponding path difference range is 8 to 580 meters. As the fading records show, no matter whether the "*A*" or the "*B*" component predominated, the other component was usually present in amplitude only second to the other. It may be pointed out that where a standing wave system exists,[10] reflected components with much larger path differences than those recorded here are almost certain to be found.

Appendix I

In the Wwedensky[4] paper the author applies his theory to one of the experimental curves from a previous paper of ours. He uses the normal earth radius "*R*," however, without any correction for air refraction. If we assume, as a more probable effective earth radius, the value $4/3R$,[6] the agreement with our curve is markedly improved.

Appendix II

In the first Wwedensky paper, *Tech. Phys.* U. S. S. R. Vol. 2, p. 632, 1935 eq. (7, 1) the sign of the term $|\tau_m|^2 \sin 2\theta_m$ should be minus.

Appendix III

The fading produced by moving bodies such as airplanes has been referred to in one of our earlier papers.[10] It happened one day, during the present investigation, that fading of this type appeared when mechanical recorders were being used and, by speeding up the paper, a record in two polarizations was obtained. The airplane itself (or other cause) was not visible. The results are given in Fig. 19. Again the horizontal component was the worse one. At first the two fadings, both fine and coarse components, were in step; later they passed entirely out of step where the fading was so rapid as to smear the paper. These "airplane" fadings were observed, off and on, at other times but were not recorded.

References

1. Englund, Crawford and Mumford, *Bell System Technical Journal*, Vol. 14, p. 369, 1935.
2. Brown and Leitch, *Proc. I. R. E.*, Vol. 25, p. 583, 1937; Norton, *Proc. I. R. E.*, Vol. 26, p. 115, 1938.
3. Ross Hull, *Q.S.T.*, Vol. 21, p. 16, 1937, May.
4. B. Wwedensky, *Tech. Phys.* U. S. S. R., Vol. 2, p. 624, 1935; Vol. 3, p. 915, 1936; Vol. 4, p. 579, 1937.

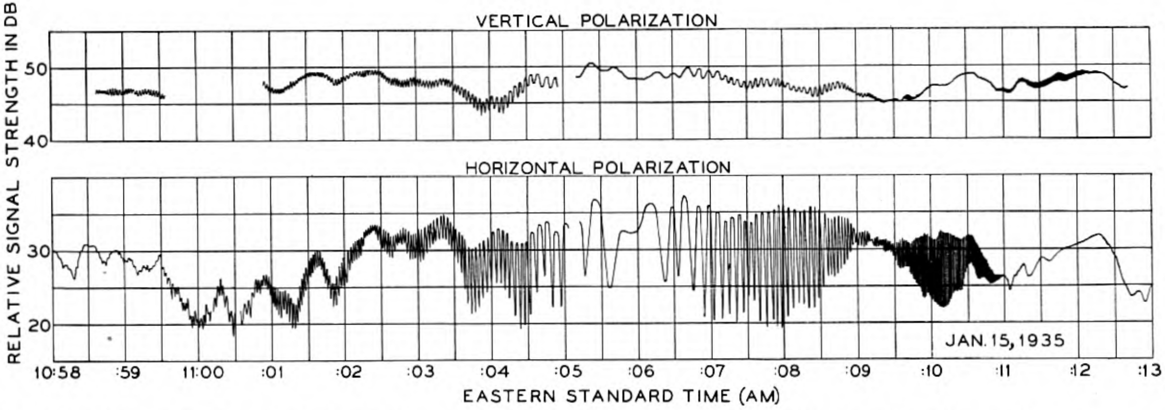


Fig. 19—Records, in two polarizations, of a transient high speed fading attributable to radio reflection from a moving airplane.

B. van der Pol and Bremmer, *Phil. Mag.*, Vol. 24, p. 141, 1937; Vol. 24, p. 825, 1937.
5. Miss M. C. Gray, paper to be published.*
6. Schelleng, Burrows and Ferrell, *Proc. I. R. E.*, Vol. 21, p. 427, 1933.
7. Colwell and Friend, *Nature*, Vol. 137, p. 782, 1936; *Phys. Rev.*, Vol. 50, p. 632, 1936; Colwell, Friend, Hall and Hill, *Nature*, Vol. 138, p. 245, 1936; Friend and Colwell, *Proc. I. R. E.*, Vol. 25, p. 1531, 1937.
8. Watson Watt, Wilkins and Bowen, *Proc. Roy. Soc.*, A, Vol. 161, p. 181, 1937.
9. *Q.S.T.*, Vol. 21, p. 27, 1937, July.
10. Englund, Crawford and Mumford, *Proc. I. R. E.*, Vol. 21, p. 464, 1933.
11. H. C. Willett, *Bull. Amer. Meteor. Soc.*, Vol. 17, p. 213, 1936.
12. Beal, *Television*, Vol. 2, p. 15, 1937, R.C.A. Inst's. Press.
13. Englund, Crawford and Mumford, *Nature*, Vol. 137, p. 743, 1936.

* The case of vertical polarization is treated by references 4, that of horizontal polarization by reference 5.

Amplitude Range Control

By S. B. WRIGHT

The art of controlling the amplitude range of telephone signals involves recognition of certain characteristics in addition to those used to specify the performance of ordinary transducers. Fundamentally, three kinds of characteristics are necessary to distinguish different range control devices. They are (1) the steady-state input-output characteristics, (2) the time actions, and (3) the range over which they function. In some cases, several secondary characteristics may be of interest, but they need not be considered in determining to which class a particular device belongs.

This paper discusses and classifies these characteristics.

Introduction

IN a "non-linear" transducer, the output power is not proportional to the input power. Consequently, the ratio of maximum to minimum power at the output differs from that at the input. But the ratio of maximum to minimum power is an expression of amplitude range. A device designed to alter this ratio may be called a *range controller*.

In telephony the term *range controller* includes many devices [1] having specific names, such as limiters, volume control devices, range reducers, compressors, vogads, expandors, etc. These devices have many properties in common with telephone repeaters, and a repeater may be considered as a special case in which any non-linearity which may exist between the output and input is unintentional.

The purpose of one type of range controller is to reduce the range of significant intensities of signals applied to a telephone circuit so as to ease the requirements of the transmission medium with respect to overloading and noise interference. Such a device is placed at the transmitting end of the circuit. When the range is compressed at the sending end of the circuit it may sometimes be desirable to expand it at the receiving end to the original range. This is done with a device having, in general, the same dynamic characteristic as the compressing device, but a range change which is complementary. The purpose of the expandor is to reduce the noise heard by the listener as well as to compensate for whatever characteristic signal modification occurred in

[1] For numbered references, see end of text.

the process of compressing the original wave. Sometimes an expandor is used at the receiving end to reduce the gain in the intervals between the main signals even when no compressor is employed. This is an example of using a range controller to correct defects in the medium.

As is well known, the performance of a repeater is specified by such characteristics as impedance, amplification, frequency band, noise, and output carrying capacity. The performance of a non-linear device involves some additional characteristics. The primary ones are (1) the slope of the input-output curve, which tells how the range is changed, (2) the dynamic operation, which tells the manner in which the output varies with time following a given change in input, and (3) the range, which tells the region over which the device exercises control.

It may be helpful to imagine a range controller as an amplifier in tandem with an adjustable attenuator, the loss of which may be changed either instantly or slowly to follow in some predetermined fashion changes in the signal. For simultaneous operation, this device could put out a wave which is a simple function of the input, but if the operation were delayed by a definite interval the device would be required to respond in a complex fashion in accordance with a recollection of what had occurred in the signal during the delay period. Such delayed adjustment would be very crude for intervals comparable with the periods of fundamental speech frequencies. To obtain practical regulation of the delayed type it is necessary to increase the delay beyond this range and base the control upon the amplitudes of the syllables. When the delay is increased to a point where it is comparable with the syllabic periods its usefulness is again reduced.

Part 1—Control Ratio

Fundamental Characteristics

Figure 1 shows how waves may be altered by a device having a given output-input characteristic, assuming the operation is instantaneous. As this figure is plotted on a db scale, only the stronger portions of positive values of the wave are shown. A similar diagram could be drawn for negative values. The output-input characteristic, although a straight line in this kind of diagram, would of course be parabola-like if plotted on a current or voltage basis. By selecting points, such as A (or B), on the input wave and determining the relative outputs A'' (or B''), the corresponding resultant wave is obtained. In this case, the resultant has a flatter top than the original sine wave, and this illustrates the capabilities of the device in increasing weak signals with respect to the strong ones and also suggests that distortion may accompany the transformation. Such effects depend upon the slope of the output-input characteristic.

The control ratio of a range controller might be defined as the output range in db divided by the input range in db within the non-linear region of interest. The ratio is obtained in such a way as to eliminate transient effects, i.e., using steady-state sine waves.

Typical Control Ratios

Figure 2 shows some typical output-input characteristics for various transducers having control ratios between zero and infinity. While these typical characteristics are straight lines there is nothing to prevent a range controller having a control ratio which varies with input. However, when complementary action is required at the receiving end it is more readily obtained when the control ratio is constant. Also, some physical elements used in the design of range controllers are most readily adapted to a straight line characteristic.

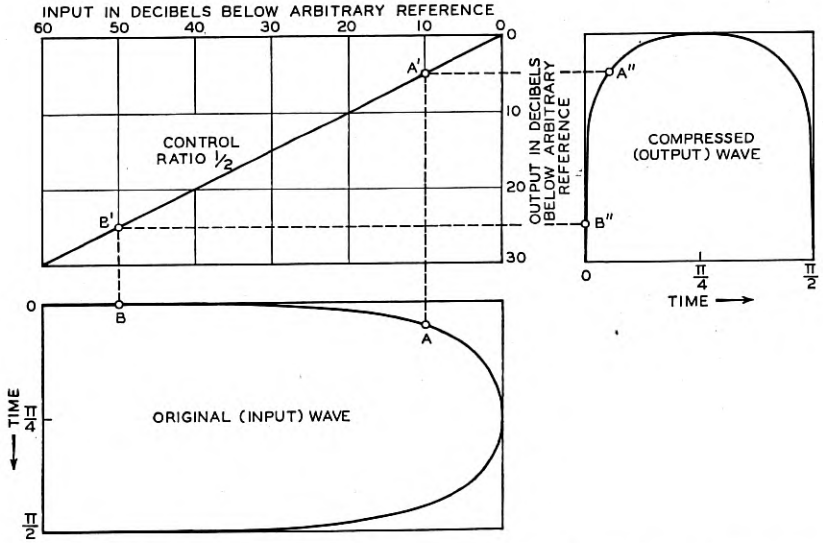


Fig. 1—The signal modification caused by a non-linear transducer depends upon the slope of the output-input characteristic.

Compressors (that is, devices having control ratios less than 1) may be divided into two classes: (1) Complete * and (2) Incomplete. In a complete compressor (control ratio = 0) the output is held constant within the range of the device. This control ratio gives a maximum

* This is not usually of practical interest but is useful as an ideal limit of operation.

of possible noise improvement and also a maximum of signal modification. There is, however, no information in the compressed signal which would serve to indicate how much compression occurred. Consequently, if it were desired to restore the original range, it would be necessary to transmit this information in addition to the compressed

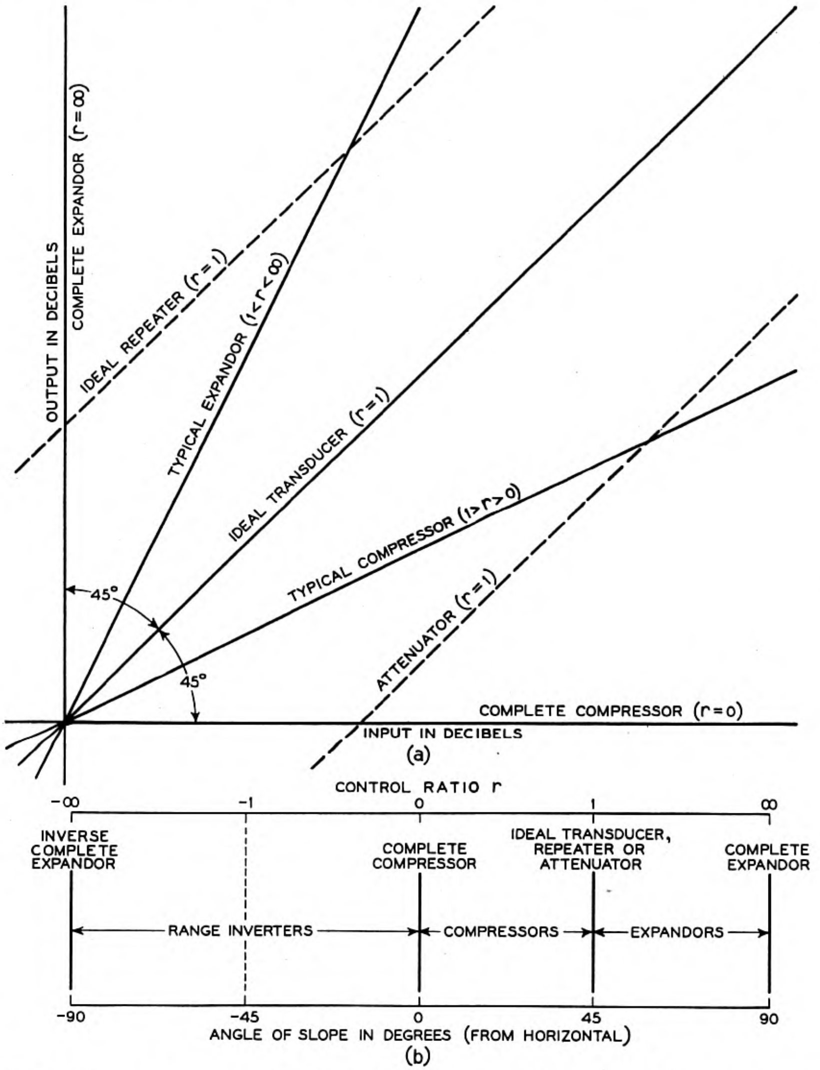


Fig. 2—If transducers are classified with respect to the slope of the input-output characteristics, several fields of action with definite demarcations result.

signal. The gain of the restoring device would be guided by this auxiliary information. Hence, the device used to pass the information along is called a "pilot channel." Various types of pilot channels are listed in Part 4 as secondary characteristics of the control.

When the control ratio is between 0 and 1 the compression is incomplete. A wave compressed in this manner has the property of being able to cause re-expansion at the receiving end since the output amplitude bears a definite relation to the original, assuming constant transmission over the intermediate circuit.

In the field of expandors having a control ratio between one and infinity the signal modification is opposite to that of compressors. Thus a convenient method is available for restoring the original wave shape by using an expandor having a control ratio which is the reciprocal of that of the compressor at the sending end.

Effects of Control Ratio

The control ratio is useful in determining the effectiveness of a device in improving transmission in the presence of noise in the medium. When noise alone is acting on the device, the noise determines the action in a manner similar to speech. When both noise and speech are present, the action is determined by the sum of the two. Thus, room noise applied with the speech will be compressed or expanded exactly as if it were part of the speech. In the case of a compressor used at the sending end of a noisy circuit, an input range of say 60 db might be compressed to 20 db, by using a control ratio of 1/3 over the entire input range. At a point where the strongest signals are unchanged, the weaker signals would then be 40 db stronger than when the compressor was omitted. The improvement of the signal and applied noise with respect to noise in the medium thus depends on the difference in ranges at the input and output which depends on the control ratio.

A large part of the usefulness of an expandor is in changing the apparent ratio of speech to the noise heard in the absence of speech, since the noise is generally weaker than speech and is made even less compared to speech by expansion. This is in spite of the fact that at any instant the signal-to-noise ratio is the same at the output as at the input. When the noise is comparable with the speech in amplitude, or when the noise is so weak as to be negligible without a controller, there can be no improvement in the noise conditions in using these devices. Between these two limits, the noise improvement rises to a maximum value also determined by the control ratio, and the time actions and range to be discussed.

A receiving range controller also changes variations in the transmission medium in proportion to the control ratio.

Part 2—Time Actions

Instantaneous Control

A device having a given control ratio might have its gain changed simultaneously with the applied e.m.f. The signal modification would become greater as the control ratio departed farther from unity and the modified signals would approach rectangular wave shapes at the limiting control ratios. Unless instantaneous compression is limited to a very small part of the signal range, an incomplete instantaneous expandor (inverse rooter) is required at the distant end which does the reverse of what is done at the transmitting end to restore the signal to substantially its original form. Due to the characteristics of the compressed signals, however, a transmission bandwidth without appreciable amplitude or phase distortion of two to three times the normal is necessary for high quality transmission.

Rectified Control

To avoid the necessity of transmitting such a wide band of frequencies, as well as to permit the use of a single device without restoring, in which case the distortion is limited to a value which is permissible from the standpoint of a listener, practical devices do not operate instantaneously. Instead, the gain is controlled by the charge on a condenser, which is controlled by rectified waves. The action of such an arrangement will now be discussed.

Consider a wave formed by subtracting two sine waves equal in amplitude, one having a frequency 10 per cent less than the other.* A portion of such a wave is shown in Fig. 3*a*. This wave is equivalent to a cosine wave of frequency one-half the sum of the two frequencies, as shown by the instantaneous voltages of Fig. 3*a*, multiplied by a secondary wave (envelope) of frequency one-half the difference of the two original frequencies.

The instantaneous voltages of the wave of Fig. 3*a* vary from a positive maximum through zero to a negative maximum. Curve *a* of Fig. 4 is a summation of most of the instantaneous e.m.f.'s of Fig. 3*a* with respect to their occurrence. About 99 per cent of the instantaneous voltages are in the ranges shown, the remainder being in the range between the upper and lower halves of Fig. 4.

* This illustration is not directly comparable with speech, but it contains some of the attributes which are comparable in this analysis, besides being readily reproducible and relatively simple.

Figure *3b* indicates values for the same wave in which the negative ordinates have their signs reversed by means of an ideal full-wave rectifier. The resulting wave contains frequencies which were not present in the original, prominent among them being second and higher harmonics of the original. The range of instantaneous values shown on Curve *b* of Fig. 4 is only half that of the instantaneous voltages. About 99 per cent of the values lie in a 60 db range.

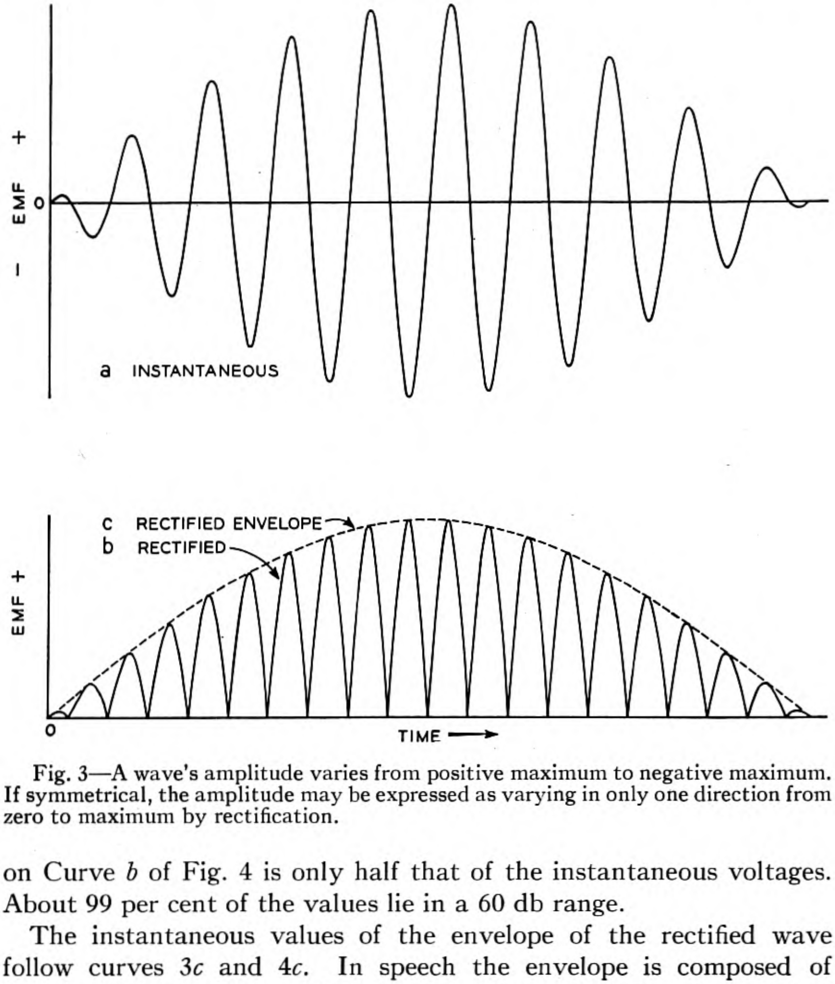


Fig. 3—A wave's amplitude varies from positive maximum to negative maximum. If symmetrical, the amplitude may be expressed as varying in only one direction from zero to maximum by rectification.

The instantaneous values of the envelope of the rectified wave follow curves *3c* and *4c*. In speech the envelope is composed of many rather low frequencies which are determined by the rates of enunciation of syllables. For this reason they are sometimes called the syllabic frequencies. If it were possible to make the control vary as a function of the envelope, the result of using a control ratio of 1/2 on the wave of Fig. *3a* would be as shown in Fig. *5c*. This was obtained by multiplying the original wave by a factor which is inversely proportional to the rectified envelope. For comparison, the original wave is shown in Fig. 5*a*, and the result of instantaneous compression by the same control ratio in Fig. 5*b*. It is assumed that the arbitrary reference voltage which is not changed by compression corresponds to the maximum value of the input wave, although any other value might be used instead. It is evident from Fig. 5 that envelope com-

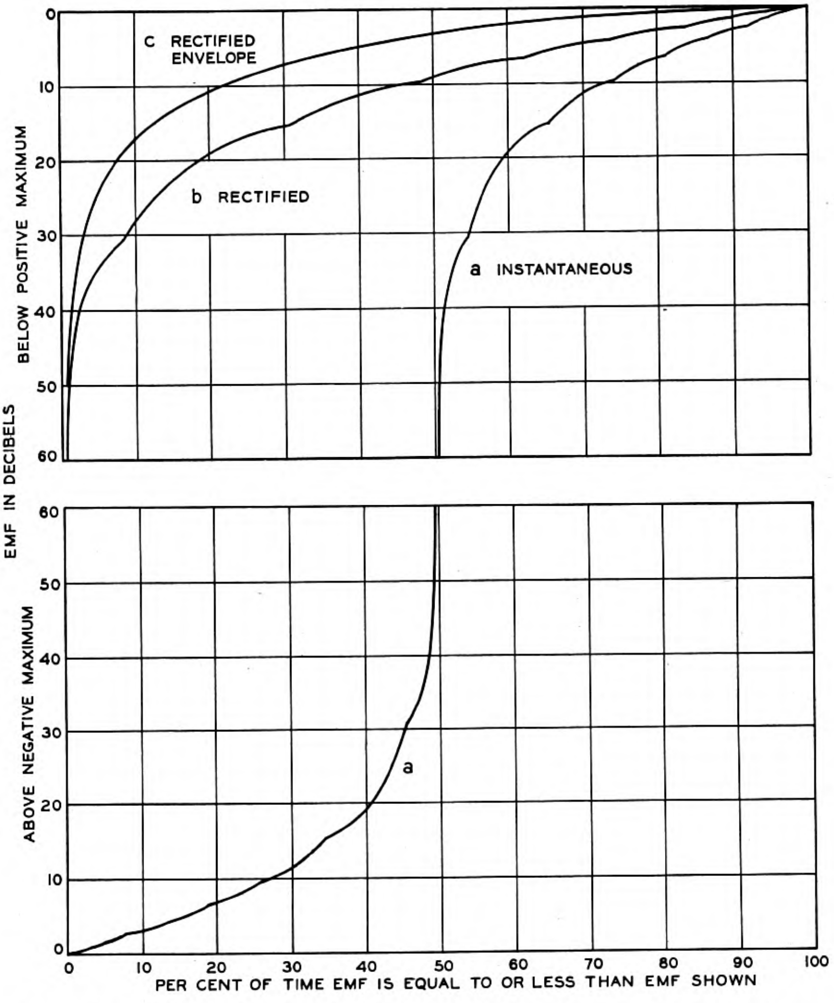


Fig. 4—The amplitude range of the wave of Fig. 3 is infinite on a db scale but most of the values are bunched in a much smaller range.

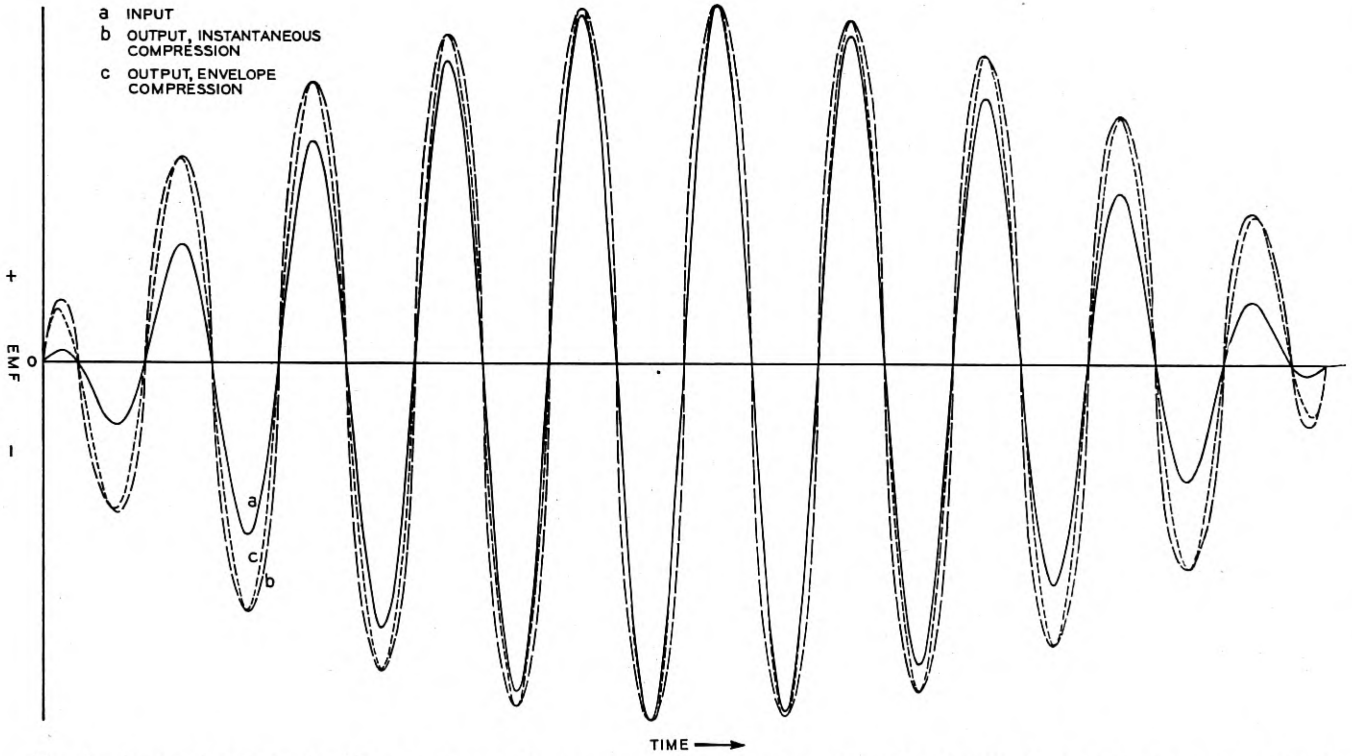


Fig. 5—A wave is compared with two corresponding compressed waves using a control ratio of 1 : 2. For instantaneous action the loops are relatively wider than the original, while in envelope compression the shape is more nearly retained. In both cases the amplitudes of the weak peaks are increased relative to the strong ones.

pression would result in less distortion than instantaneous compression. The extra frequencies formed that were not present in the original wave are the envelope frequencies, so that the additional band required to transmit this wave faithfully is negligible.

Dynamic Operation

The measurements [2] and adjustments of speech amplitudes in common use are made with devices that integrate the effects of the wave over certain time intervals. They do this in a rather complicated manner, however, so that it is difficult to express the resulting quantities in terms that are generally understood.

In the measuring instruments the rectified voltages are impressed on a condenser before being sent through a meter. The readings of the meter are, therefore, proportional to the voltage on the condenser modified by the damping of the meter. The voltage is made up of the sum of the effects of all the instantaneous voltages that have been applied to the condenser from the beginning of time to the instant under consideration. These effects die out so rapidly, however, that the instantaneous voltage on the condenser is practically determined by the voltages received in the immediate past. The condenser may be said to have a memory but a short one. In range control devices, the condenser forms the voltage which determines the amplification of the device.

To distinguish this voltage on the condenser from the applied voltage at any instant, we may call the former an "impression" of the original wave. If the time constant RC is small we get strong impressions similar to the rectified applied wave and its envelope, and if it is large we get weak impressions quite different from the applied wave but something like the rectified envelope.

Figure 6 shows the impressions of the wave of Fig. 3*a*, using four different values of time constant RC as compared to P, the period * of the envelope. Figure 7 shows smoothed summation curves of the impressions of Fig. 6 formed during the time $P/2$. Comparing this with Fig. 4, it is evident that the "bunching" effect for the distribution of impressions is largely between those for the rectified instantaneous and envelope curves. For the longer time constants, i.e., weak impressions, this is not the case for the weaker e.m.f.'s.

* This is twice the duration of Fig. 3, since only half a cycle is illustrated. It is assumed that C is completely discharged at the time this wave is applied. In practice, the rectifier impedance varies with the applied e.m.f. so that the results are not as simple as in this illustration. In general, the time actions are different depending on whether the applied e.m.f. is increasing or decreasing.

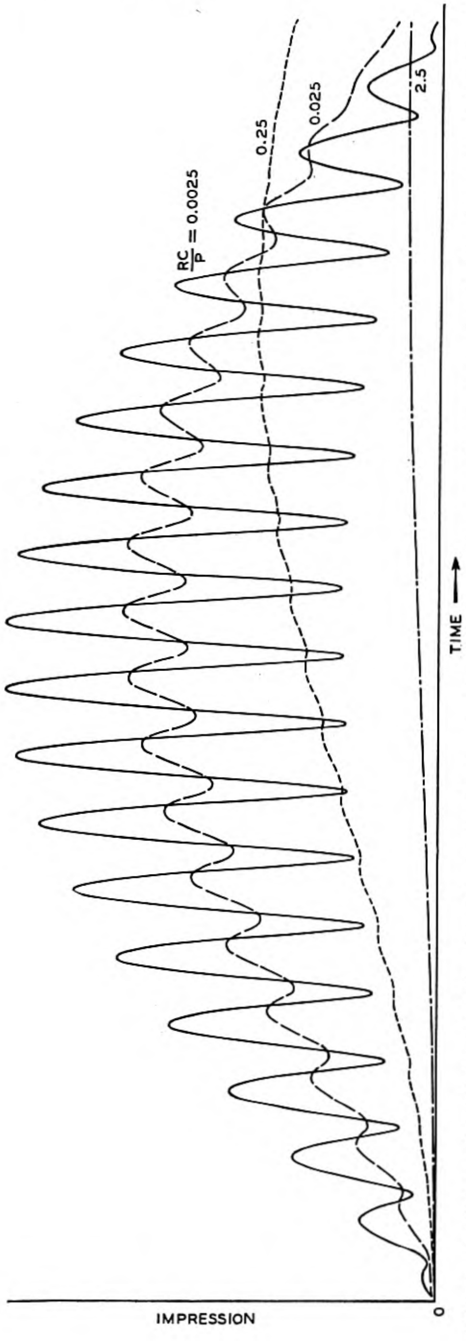


Fig. 6—An integration of Fig. 3*b* differs from the original wave and lags behind it by increasing amounts as the time constant is increased.

Referring again to Fig. 6, it will be seen that for the two smaller values of RC/P the impression curves are composed of (1) the envelope frequency, (2) double the fundamental frequency, and (3) a small delay which can often be neglected. An approximation to envelope compression is therefore possible by choosing RC/P to be in the proper range, i.e., .0025 to .025, and making the output vary as a root or power ot the impressions thus formed.

Figure 8 shows the result of compressing the wave of Fig. *3a* by using the impressions of Fig. 6 to determine the amplification. It was assumed that the amplification varies in inverse proportion to the square root of the impression. The resulting waves for $RC/P = .0025$ and .025 (medium impressions) are recognizable as something like the original wave. However, for the larger values of RC/P (weak impressions), the distortion at the beginning of the wave is quite large. This is because the impressions are formed so slowly that a longer time is required to drive the gain down to the desired value.

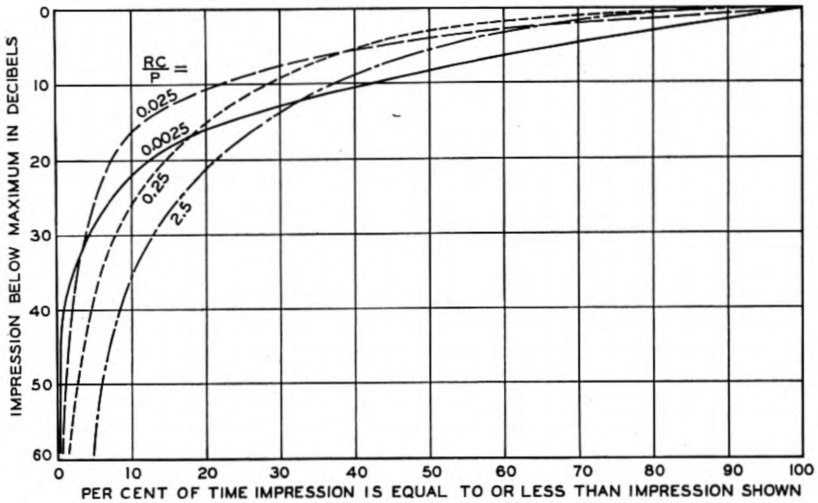


Fig. 7—The amplitude ranges of the impressions shown in Fig. 6 are bunched differently, depending on the time constant. A "volume" measurement means that a given impression is exceeded a small percentage of the time. In speech the peaks are relatively higher than in the wave illustrated.

In order to compare impression compression with instantaneous compression, the ordinates of Fig. 5 and 8 were plotted in Fig. 9. This shows that the greatest possibilities of bunching the waves into a narrow range result from the use of instantaneous compression (*b*), since the ratio between any value and the maximum is modified by the

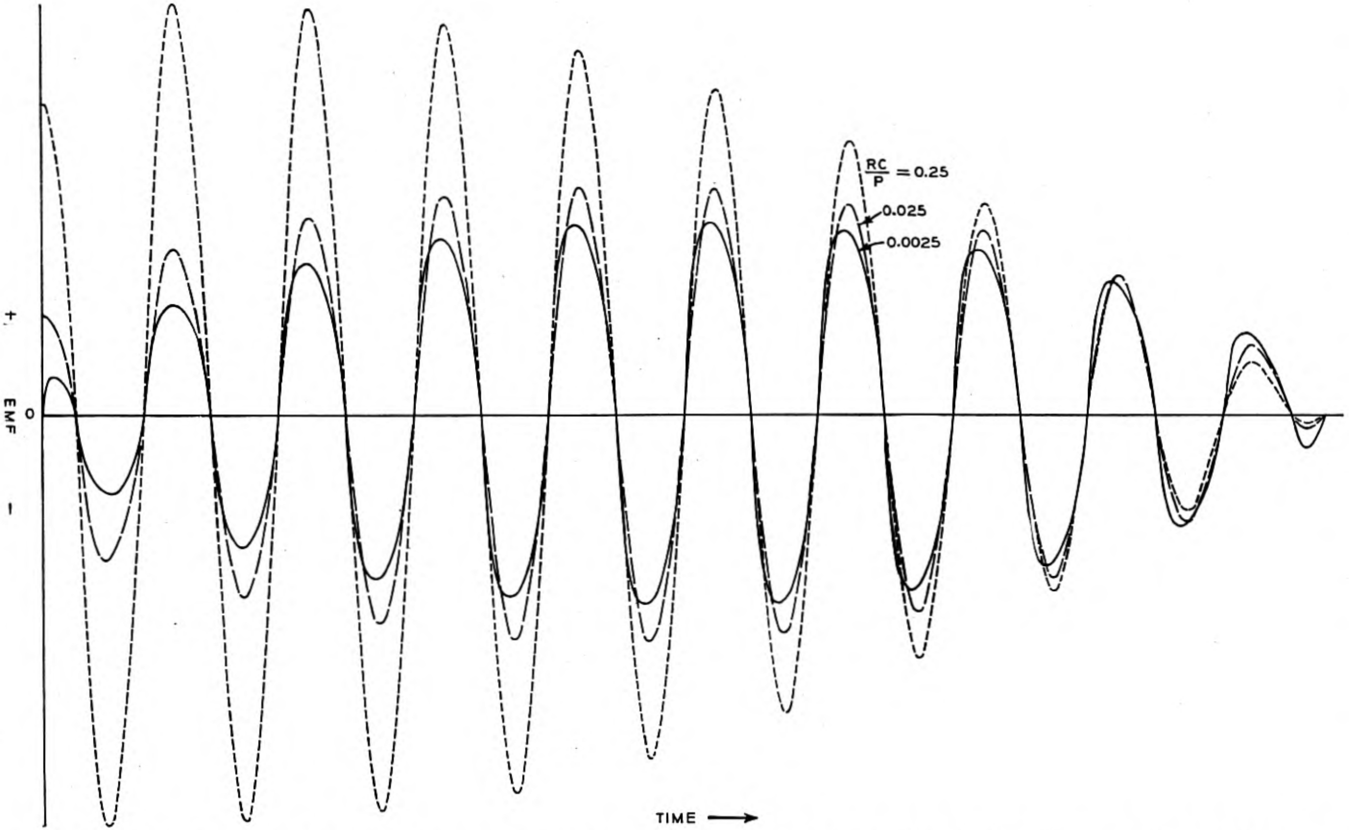


Fig. 8—The wave of Fig. 3*a* is shown here as compressed (1 : 2) by gain changes determined by the impressions of Fig. 6. The smallest time constant gives a result approaching instantaneous compression as in Fig. 5*b*. The next size approaches envelope compression as in Fig. 5*c*. The largest size shown results in considerable distortion, particularly at the beginning. The distortion of the fourth is too large to be shown on the same scale.

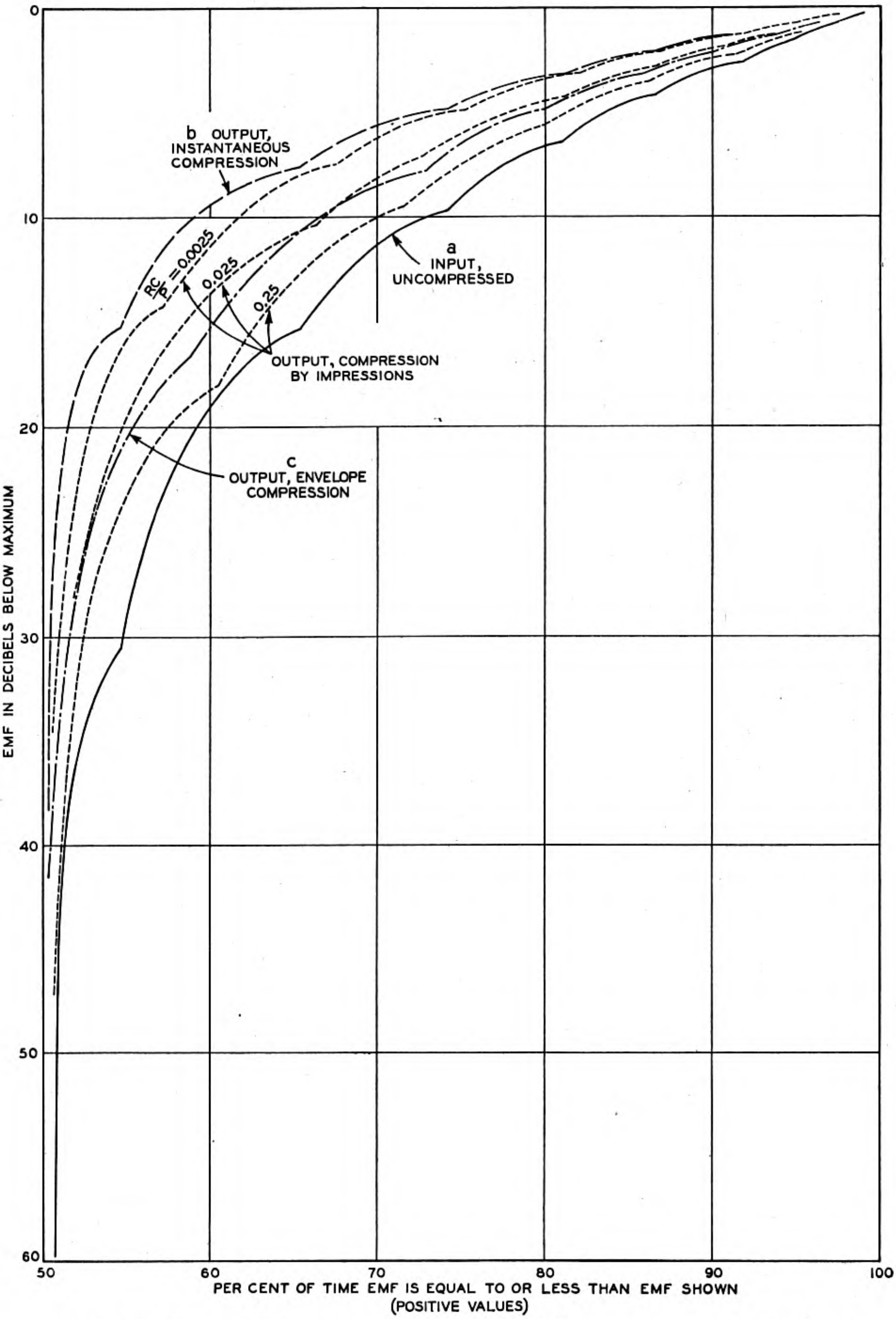


Fig. 9—The amplitude ranges of the compressed waves of Fig. 8 are shown, together with those of Fig. 5. The amount of signal modification (and noise improvement) for any amplitude below maximum may be correlated readily with the time constant.

control ratio. In the case of envelope compression (*c*), the lag causes a reduction in the amount of compression of the instantaneous voltages and the result is seen to be about half way between curves *a* and *b* of Fig. 9. The remaining curves of Fig. 9 are representative of the device [3] used on the long-wave transatlantic radiotelephone circuit.

Volume Control

To avoid both a large range and also the necessity for a continuous record, practical speech amplitude measuring instruments are not directly concerned with either instantaneous, envelope, or impression voltages. Instead a value is determined, corresponding to an impression which is exceeded only a small percentage of the time. This is the principle underlying speech measurements with "volume indicator" type of instruments. In the case of speech, which is much more complex than the simple wave we have discussed, curves like Fig. 9 are steeper, i.e., there are relatively more peaks and a larger range to complicate the problem.

A particular device capable of compressing according to the requirement that the dynamic "volume" range should be reduced, is attained by a combination of several separate range controllers. One is provided to reduce the gain very rapidly when the output volume is too high. A second increases the gain, at a much slower rate, when the impressions formed on the condenser are consistently too low. A third disconnects the condenser from the input when the applied voltage is very small, so that the distortions inherent in change of gain by weak impressions will occur only at times of large and sudden decreases of volume. In the device [4] employed to control volumes applied to a radio transmitter at Norfolk, Virginia, a fourth control provides for rapid partial compression of high peaks, thus improving the modulation. It is unnecessary to re-expand for the purpose of restoring the intelligibility, since the distortion is virtually limited to a change in loudness.

Part 3—Range

Range controllers, like repeaters and attenuators, are limited as to the input range they can accept and the output range they can provide. These limits may be due to thermal noise at the low end and output carrying capacity at the high end. Heretofore, in this paper, the terms "input" and "output" have been purposely left somewhat vague so as to be as general as possible. However, the limits of input and output of a range controller take on particular significance when it is considered that the signal input range may differ both from the input range of the device, and also from the range over which control is exercised.

This control range may be defined as the difference between the maximum and minimum values of an applied wave over which a device is designed to function in a specific non-linear manner. It is usually expressed in db, and may apply to any measure of the applied signal, such as instantaneous voltage, rms steady-state sine waves, or a dynamic measure such as volume. The values dividing the controlled range from the uncontrolled ranges may be referred to as the "control points."

Certain advantages in some cases have been found from restricting the control range. This is accomplished by placing one or both of the control points inside the useful amplitude range. The position of the control point may be moved arbitrarily over a wide range by putting an ordinary repeater (or attenuator) in tandem with the range controller. A given amount of compression at the high amplitude end of the range gives a real signal-to-noise advantage for a much greater proportion of applied e.m.f.'s than the same amount of compression at the low end of the range. In either case the distortion would be less than that of a full range compressor. When expandors with limited range are used, they are subject to the limitation that variations in the medium are increased, but to a lesser extent than full range expandors.

Part 4—Classification of Range Controllers—Secondary Characteristics

Table I, page 536, suggests how the conceptions of control ratio, time actions and range already discussed might be employed to distinguish a variety of devices. In cases where more than one device is covered by a given control ratio and time action the distinction is that the ranges are different. The names of devices used in this table are those which have been used in the past to distinguish the devices one from another.

Nomenclature

Using the above conceptions of the three primary characteristics, it has been found possible to devise a notation to distinguish all the known devices in this field. As an example of how this proposed system of nomenclature would be applied, Table II, page 537, gives three columns. Column 1 sets forth the arbitrary names that have been used in the past to distinguish certain devices which have come into use. Column 2 gives descriptive names which specify the three fundamental characteristics. In column 3 each device is named by three symbols defining the three fundamental characteristics, and a

TABLE I
Classification of Range Controllers

Typical Time Actions	Compressors ($r < 1$)		Expandors ($r > 1$)	
	Full Range	Limited Range	Full Range	Limited Range
Instantaneous	Rooter	Peak Chopper, Voltage Limiter	Inverse Rooter, (Squarer)	Voice Operated Relays, Cross-talk Suppressor
Syllabic	Compressor	Limited Range Compressor, Peak Limiter	Expandor	Noise Reducer
Volume	Vogad, Range Reducer	Volume Limiter, Half Vogad	Range Restorer	

classification which tells what the device is designed to do. In this system the numbers preceding the letters specify the input control range in decibels, and the position of a horizontal bar indicates the position of the main signal range with respect to the control range.

The letters specify the time actions and in the case of vogads, where several time actions may be combined, an arbitrary combination of letters would be used. The final numbers specify the control ratio, and in the case of vogads, where this might be different depending on whether the input was increasing or decreasing, both values are given, the former first.

In this system, definitions of time actions are prerequisite and by way of illustration, the following symbols have been used:

I represents instantaneous, meaning very fast adjustment of device
S represents syllabic, meaning moderate speed adjustment of device
V represents volume, meaning a combination of controllers which produces adjustment of device in response to dynamic speech so that the output volume is approximately determined by the input volume.

Secondary Characteristics

In addition to their three primary characteristics, range controllers may have a number of secondary features which are sometimes important. The outstanding ones are:

1. Bias

A *neutral* range controller is one which holds its setting during the quiet periods between words and sentences and which changes its gain

TABLE II

Comparison of Nomenclature for Range Controllers

Col. (1) Arbitrary [1]	Col. (2) Systematic	Col. (3) Symbolic
1. Vogad	Full Range 45 db Volume Compressor	45 *VSI* 23–18 Compressor
2. Vogad Combined with Syllabic Compressor	Full Range 45 db Volume Compressor	45 *VSSI* 23–18 Compressor
3. Volume Limiter	High Range 15 db 1 : 5 Volume Compressor	$\underline{15}V5$ Compressor
4. Compandor	Full Range 60 db 2 : 1 Syllabic Compandor	$60S2$ Compandor
5. Noise Reducer	Low Range 20 db 2 : 1 Syllabic Expandor	$\overline{20}S2$ Expandor
6. Limited Range Compressor	High Range 10 db 1 : 2 Instantaneous Compressor	$\underline{10}I2$ Compressor
7. Peak Limiter	High Range 12 db 1 : 5 Syllabic Compressor	$\underline{12}S5$ Compressor
8. Peak Chopper	High Range 6 db 1 : 100 Instantaneous Compressor	$\underline{6}I100$ Compressor
9. Crosstalk Suppressor	Low Range 10 db 10 : 1 Instantaneous Expandor	$\overline{10}I10$ Expandor
10. Rooter and Inverse Rooter	Full Range 70 db 2 : 1 Instantaneous Compandor	$70I2$ Compandor
11. Vodas (Singing Suppressor Relay)		$0I\infty$ Expandor
12. Syllabic Vodas		$0S\infty$ Expandor

only when the waves acting on it differ from those just received. This condition sets a new requirement on the range controller which can usually be met by a combination of control circuits.

A *biased* controller is one which returns to a setting corresponding to some fixed or biased intensity when speech is not passing and adjusts itself each time speech begins. A simple compressor is biased since with no input it generally takes a setting of maximum gain so as to be right for the weakest waves that might be applied in its working range or below the working range. It is also possible to bias a range controller so as to have minimum gain, or any other intermediate value when no waves are applied. An important secondary characteristic is the rate at which the device returns to the desired "bias" point.

Any of the devices listed in the tables might be neutral or biased in either direction, thus increasing the number of possible arrangements.

2. *Behavior Outside of Range*

For inputs outside the working range of a range controller it is important to provide that the amplification of these waves does not cause them to be modified so as to be out of proportion to output signals in the main range. In some cases this is met by choosing a device which follows the same law all the way to zero current. In others, the device may act as a linear transducer, i.e., with range factor of one outside the working range. Various other combinations of control ratios can, of course, be employed.

3. Pilot Channel

In all complete compressors some form of pilot channel is necessary to control the re-expansion if this is required. If the gain changes are slow, the pilot channel may include an operator who changes the gain of the receiving device in a manner complementary to that of the sending device based on aural or visual signals. If the gain changes are too rapid for the operator to follow, the receiving gain may be changed automatically.

The pilot channel itself may be a direct or alternating current of variable amplitude or frequency, or in case of carrier or radio, it may be the carrier frequency. In sound reproduction a pilot channel might be a pilot track on the record.

Summary

In an amplitude range control system, the following characteristics must be specified in addition to the usual repeater characteristics, to determine its design and performance:

1. The steady-state control ratio, which determines how much control is obtained and whether restoration can be made automatically or not.
2. The manner in which the output varies with time, following a given change in input.
3. The range over which it is to function.

In specific cases the following should also be considered:

4. The action of the device for inputs outside the working range.
5. If the device is a complete compressor, the type of pilot channel for restoring.
6. The action of the device when signals are removed.

Acknowledgment

The computations used to obtain Figs. 3 to 9, inclusive, were made by Miss Marian Darville.

References

1. "Devices for Controlling Amplitude Characteristics of Telephonic Signals," A. C. Norwine. Presented at A. I. E. E. Pacific Coast Convention, Aug. 9–12, 1938.
2. "Speech Power and Its Measurement," L. J. Sivian, *B. S. T. J.*, Vol. VIII, pp. 646–661.
3. "The Compandor—An Aid Against Static in Radio Telephony," R. C. Mathes and S. B. Wright, *B. S. T. J.*, Vol. XIII, pp. 315–332.
4. "A Vogad for Radio Telephone Circuits," S. B. Wright, S. Doba, A. C. Dickieson. Presented at I. R. E. Convention, June 1938.

Devices for Controlling Amplitude Characteristics of Telephonic Signals *

By A. C. NORWINE

This paper describes a family of devices which automatically respond to signals and control the circuit amplification in such a way as to improve transmission. Their general characteristics are outlined, their differences explained, and some of their applications are listed.

Introduction

THE transmission of speech energy over electrical circuits is attended by the interesting and sometimes difficult problem of preserving the original signal in spite of limitations in the transmission medium. These limitations include load carrying capacity, interference with other service, noise, change in attenuation with time and many others. Because of special limitations it is sometimes desirable to alter the amplitude characteristics of the speech or other signal energy without, of course, materially lowering its intelligibility. In high quality systems the peak voltage from some speech sounds of a given talker may be over 30 db (some 30 times) higher than from his weakest sounds when there is very little inflection in the speech. Loudness changes for emphasis will increase this range of intensities. Ordinary message systems do not have to contend with quite so wide a range of instantaneous voltages from a single talker, but different talkers under extreme terminal conditions produce about a 45 db range of average voltage, which is additive to that for a single talker. Consequently, a voltage range of about 70 db (over 3000 to 1) must be considered for message circuits.

In order to accommodate such ranges of intensity to certain transmission media such as radio links a new family of automatic devices has been developed. In general all of these contain amplifiers or attenuating networks whose loss or gain is changed according to some function of the applied input and which may have a variety of time sequences in their control circuits. It is hoped that by the classification and description of some of these devices their distinguishing characteristics and fields of usefulness will be made somewhat clearer.

* Presented at the Pacific Coast Convention of A. I. E. E. and I. R. E. in Portland, Oregon, August 9–12, 1938.

We are to be concerned here principally with those elements allied to the telephonic art, although some applications are to be found in other fields. It is not intended to include those voice operated functions which are essentially switching operations although the distinction in some cases becomes exceedingly fine.

Names of volume controlled devices * which have been used in published papers include vogad,†, [2, 3, 4] compandor,‡, [5, 6] and volume limiter.[7] Without direct comparison it may not be obvious how these and similar devices differ. First the apparent similarity of several of these devices will be shown in simple diagrams. Next the more important characteristics of a number of devices will be presented in tabular form, followed by descriptions of the different types. These will then be discussed with particular emphasis on their distinctive qualities, with notes on their variants which have some apparent value.

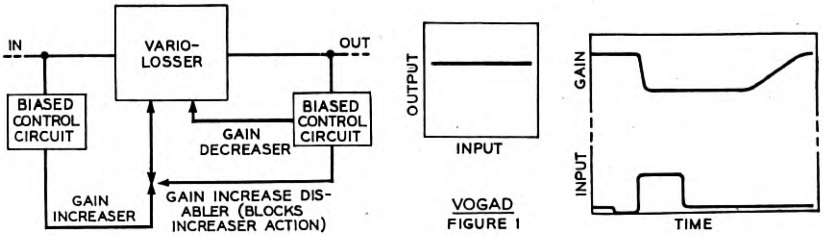
General Characteristics of Volume Controlled Devices

In Figs. 1 to 10 are shown simplified diagrams of some of these devices. While detailed descriptions of them will be deferred till later it may be pointed out that all those shown contain vario-lossers, and all have paths from the main transmission path to control circuits which affect the vario-lossers. A vario-losser usually consists of a balanced pair of vacuum tubes whose gain is changed by varying the grid bias, or of a network of non-linear elements such as copper oxide or silicon carbide whose loss is changed by varying a current through them. In some special cases it may be a mechanically adjusted variable network. The word vario-losser is thus a generic term relating to a circuit whose loss or gain is controllable. A control circuit ordinarily consists of an amplifier and rectifier whose direct current or alternating current output bears a chosen relation to its input. Thus some control circuits are marginal; they produce no control voltage till the input exceeds some critical value, then produce large control voltages for small additional increments of input. These are used, for example, when it is desired to limit the output of a vario-losser to a definite amount. Another type of control circuit produces a current or voltage which is linear with input expressed in decibels. In combination with a vario-losser whose gain is a linear function of control current or voltage one can produce a device whose gain is a linear function in decibels of the input to the control circuit.

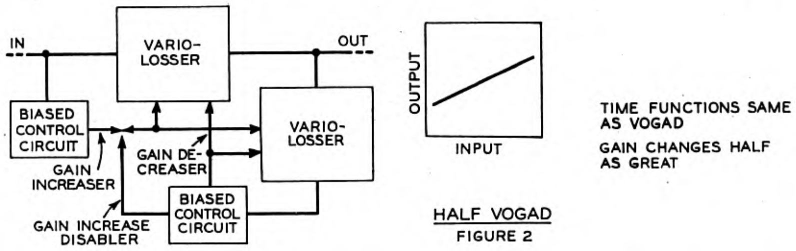
* See the footnote on page 543.

† "*V*olume *O*perated *G*ain *A*djusting *D*evice."

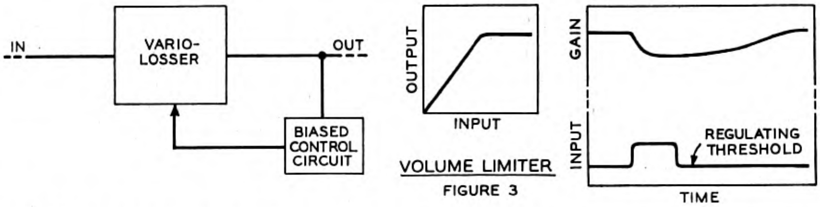
‡ A combination of the names "*Com*pressor" and "Ex*pandor*."



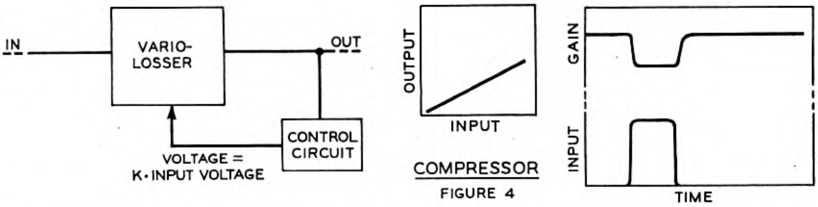
VOGAD
FIGURE 1



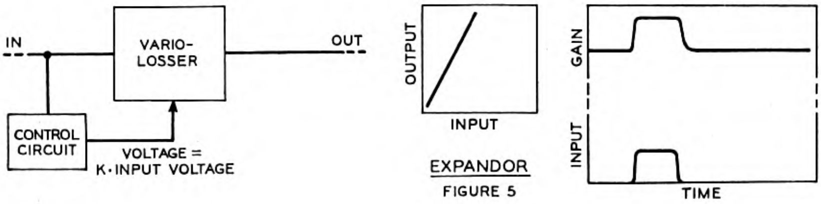
HALF VOGAD
FIGURE 2



VOLUME LIMITER
FIGURE 3



COMPRESSOR
FIGURE 4



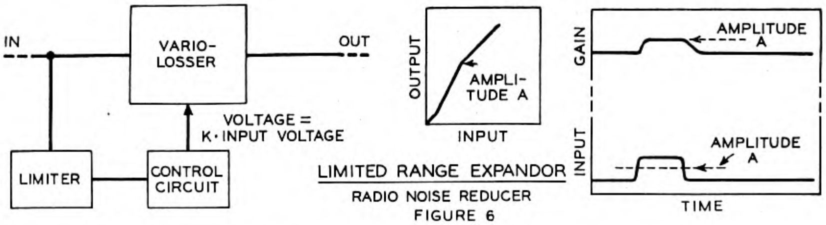
EXPANDOR
FIGURE 5

It will be recognized that if the application or removal of the control energy is retarded, the action of the control circuit may be made quite different on transient inputs than on steady state inputs. It will appear later that this is the important distinction between some of the devices to be discussed and that fundamental differences in their functioning are thus brought about.

Referring to the figures once more it will be noted that some control devices are connected to the transmission path at the input to the vario-losser. These are known as "forward acting" control circuits. Other controls, connected at the vario-losser outputs, are known as "backward acting" control circuits. This is simply convenient terminology to indicate whether the control energy is progressing in the same direction as the main transmission or is progressing in a backward direction after traversing the main path, usually through a vario-losser. Some backward acting controls function to measure the output of the devices containing them and to make whatever adjustments are required. Others are placed in that position to take advantage of the vario-lossers in the transmission paths, i.e., such controls could be replaced by combinations of forward acting controls and extra vario-lossers.

In Table I, nine of the volume controlled devices * which have been developed for various commercial and experimental uses are listed with the functions of voltage, time, and frequency which are employed to obtain their respective performances. There is, of course, some latitude in the choice of these functions for any one device. Pending more complete description of the different types in the following paragraphs this table should be viewed as illustrating the general character of the different circuits and also the range of the variables which already have been employed. For example, it will be seen that instantaneous voltage of the signal wave, its short time average value, peak power, syllabic variations, and long time average power have all been used as criteria of gain settings in different circuits. Some devices change their adjustments only when critical values or ranges are exceeded, while others vary somewhat with every syllable if speech, for example, is being transmitted. Some are linear transducers to all but low or high amplitudes while others reduce or increase the output range from that at the input. It will be seen that proper choices of times for gain increase and gain decrease in combination

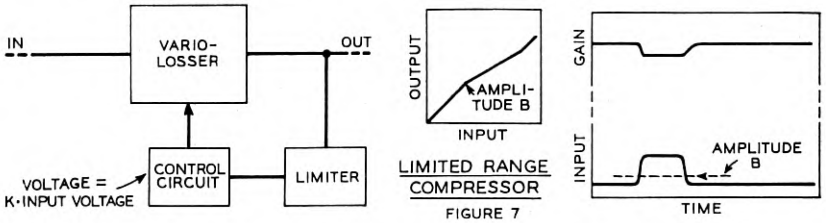
* The names employed do not follow an entirely logical classification, but they are given here because they have had considerable usage. For the same reason the term *volume controlled devices* is used, although to be strictly correct it might better be *sound energy controlled devices*, for example, for not all the devices operate in accordance with *volume* as measured by the well-known class of visual reading meters called *volume indicators*.



LIMITED RANGE EXPANDOR

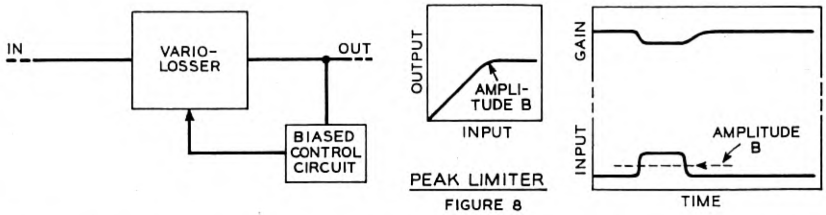
RADIO NOISE REDUCER

FIGURE 6



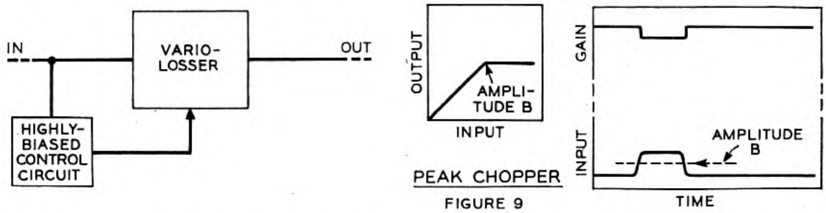
LIMITED RANGE COMPRESSOR

FIGURE 7



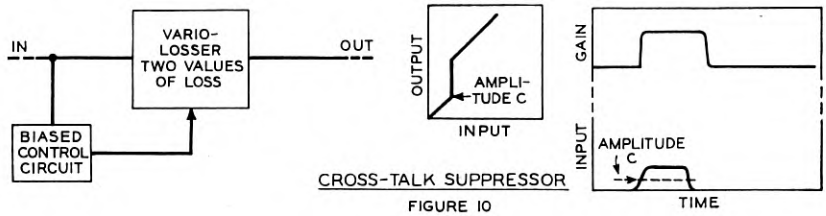
PEAK LIMITER

FIGURE 8



PEAK CHOPPER

FIGURE 9



CROSS-TALK SUPPRESSOR

FIGURE 10

TABLE I

CHARACTERISTICS OF VOLUME CONTROLLED DEVICES

Device	Gain Controlled by	Frequency of Adjustment	Time Required for Gain Changes		Ratio of Output Range to Input Range†	Position of Controls	Volume Range Controlled*	Frequency Range Controlled	Frequency Range Causing Control	Part of Input Range Causing Operation
			Gain Increase	Gain Reduction						
1. Vogad	Average volume	Infrequent. Gain fixed between transmissions	After a few words—sometimes 8 to 10 words	After one or more words	Approx. 0	At input and output	Large	Full band transmitted	Full band	All
2. Volume Limiter	Average volume over part of input range	Relatively infrequent. Approaches max. gain in silent periods	Slow and continuous	After one or more words	1 up to operate point. 0 above	At output	Moderate	Full band	Full band	High amplitude only
3. Compandor *a.* Compressor . . . . *b.* Expandor	Syllabic variations	Continuous at syllabic rate	On each syllable	On each syllable	1/2 to 1/5 5 to 2	Compressor output expandor, input	Large	Full band	Full band	All
c. Special Compandor	do	do	do	do	do	ditto or expandor control over separate channel	do	High frequency only or multiband	High frequency only or multiband	All or high amplitude only
4. Radio Noise Reducer (Limited Range Expandor)	Syllabic variations over part of input range	Each syllable	On each syllable	On each syllable	2 for low amplitudes. 1 for high	At input	Moderate	Full band	Full band	Variable at low amplitudes only. Fixed gain to high amplitudes

TABLE I (*Continued*)

Device	Gain Controlled by	Frequency of Adjustment	Time Required for Gain Changes		Ratio of Output Range to Input Range†	Position of Controls	Volume Range Controlled*	Frequency Range Controlled	Frequency Range Causing Control	Part of Input Range Causing Operation
			Gain Increase	Gain Reduction						
5. Limited Range Compressor	Syllabic variations over part of input range	Each syllable within the operating range	Fast	Fast	1 up to operate point, then 0 to 1/2, then 1	At input or output	Small	Full band	Full band	High or intermediate amplitudes
6. Peak Limiter...........	Syllabic peaks	Infrequent	After about one word	On single syllable	1 up to operate point, then approaches 0	At output	Small	Full band	Full band	High amplitude only
7. Peak Chopper..........	Instantaneous peak voltage	Relatively infrequent	Instantaneous	Insta ntaneous	1 up to operate point, then 0	At input	Very small	Full band	Full band	High amplitude
8. Crosstalk Suppressor.....	Voltage exceeding a specified value low in input range	Relatively frequent	Fast	After one or two words	1 except at point of discontinuity	At input	Very small	Full band	Full band	All above specified low value
9. Rooter and Inverse Rooter	Instantaneous voltage	Continuous	Instantaneous	Instantaneous	Ordinarily 1/2 and 2	Integral with vario-losser	Moderate	Full band	Full band	All

* Outside these ranges most of the devices tend to be linear transducers except for higher volumes applied to "Limiters."

† It is important to note that these ranges are measured in the same units as the respective control circuits measure: viz., "volume," "syllabic power," "voltage," etc.

with certain gain control criteria make possible a wide variety of signal altering means to meet different requirements.

Description of Devices in Table I

With this introduction to the combinations of characteristics which are possible it should be less difficult to distinguish between the specific devices discussed in the following paragraphs, which, in addition to describing the devices, contain some comments which should assist in visualizing their forms and their operation.

1. The *vogad* (Fig. 1) is a device which will maintain at its output speech volume[1] which, over a certain range of input, is relatively independent of the speech volume applied to its input and which, in the ideal case, will not change its gain during periods of no speech input. It makes little or no alteration in the ratios of maximum and minimum instantaneous to average voltages of the speech.

2. The *volume limiter* (Fig. 3) is a device which is a linear transducer for all speech volumes up to a critical value, beyond which all input volumes produce essentially the same output volume. It is essentially different from the vogad in that its gain approaches the maximum value when input is removed.

3. The *compandor* (Figs. 4 and 5) is composed of a *compressor* and an *expandor*. A *compressor* is a device whose input-output characteristic on a decibel scale has a slope less than unity * and whose gain or loss is variable under control of the input energy at a time rate which will permit it to follow the syllabic rate of change of speech energy. Similarly, an *expandor* is a device whose input-output curve has a slope greater than unity and whose gain is variable at a syllabic rate under control of the input energy. Thus very shortly after all input is removed the gain of a compressor is maximum and the loss of an expandor is maximum. The reciprocal of the compressor characteristic slope is spoken of as the compression ratio, and the slope of the expandor characteristic is spoken of as the expansion ratio.

4. The *radio noise reducer*[8, 9] (Fig. 6) combines the functions of an expandor which operates in the range of amplitudes where noise and weaker speech sounds lie and a linear transducer which comes into play for all amplitudes exceeding a critical value, which can be set to best suit the atmospheric noise conditions. In other words, the radio noise reducer is a limited range expandor. Inputs which are below the expandor range are subject to transmission at the minimum gain.

5. The *limited range compressor* (Fig. 7) is a device whose operating

* That is, if the input increases by x db the output increases by less than x db.

range includes a region within which compression at a syllabic rate can take place; at other inputs the device is a linear transducer. Its connecting diagram and time functions are the same as those shown in Fig. 5 except that the control circuit contains a limiting device, so that compression takes place in only a portion of its input range, analogous to the action of the limited range expandor of Fig. 6. As a special case the *limited range compressor* may have no linear range above its compression range, thus becoming one type of peak limiter.

6. The *peak limiter* (Fig. 8) is a device whose gain will be quickly reduced and slowly restored when the instantaneous peak power of the input exceeds a predetermined value. The amount of gain reduction is a function of the peak amplitude, and in practice is usually intended to be small to prevent material reduction of the range of intensity of the signal.

7. The *peak chopper* (Fig. 9) is a device which prevents transmission of peak amplitudes exceeding a critical amount, an essential characteristic being that the loss it inserts is completely determined by the instantaneous voltage of the signal. That is, its operating and releasing times are substantially equal to zero.

8. The *crosstalk suppressor* (Fig. 10) is a device which normally presents a prescribed loss to transmission, which loss is removed rapidly when the input amplitude exceeds a certain threshold and is reinserted at a definite time after the input is removed. It reduces low amplitude unwanted currents such as crosstalk but does not affect amplitudes in the useful signal voltage range. This device differs from the limited range expandor in that the time during which the low loss condition is maintained is considerably greater, so the transition from one gain to the other occurs less frequently.

9. A *rooter* is an instantaneous compressor. Such a circuit can be made to produce an output whose instantaneous voltage is, for example, the square root or some similar function of the instantaneous voltage applied to the input. An *inverse rooter* is an instantaneous expandor whose characteristic is complementary to that of the rooter. A combination of *rooter* and *inverse rooter* will reduce the load requirements on a transmission system between the two units but requires that it transmit a wider band of frequencies than that for the original signal, and that it be essentially free from phase distortion. This does not seem to be an attractive arrangement from a commercial viewpoint and is included here simply as an illustration of one of the possible modifications of signal energy. It is not shown in the group of diagrams.

Variants to the Devices Described

In addition to these there are various devices which are essentially modifications of those described. For example, a half-vogad, Fig. 2, may have the same time functions as a vogad, Fig. 1, but the gain changes in the transmission circuit are half as great for the same range of input volumes. Thus in a vogad the range of gain changes in the transmission circuit is equal to the range of input volumes, so that the output volume is the same for all input volumes. In the case of the half vogad the range of gain changes in the transmission path is one-half the range of input volumes, so the output volume range is one-half that of the input. It is also possible to construct a vogad whose output volume range is any desired fraction of the input range. As another example of modification of the devices described, for special applications it may be desirable to incorporate a certain amount of syllabic compression in a vogad.

Communication circuits which have separate paths for oppositely directed transmission between the two terminals are usually operated at such an overall loss that with ordinary terminations there will be little tendency for circulating currents to build up to a "singing" condition. Sometimes there may not be a great deal of margin, however, so that volume controlled devices added to such circuits must add loss at some point to counterbalance whatever gain is put in at some other point. Thus a vogad inserted at the transmitting side of one terminal of such a circuit to amplify speech energy from weak talkers must be supplemented by a "reverse vogad" in the receiving side of the circuit. The reverse vogad is simply another vario-losser which is operated upon by the vogad control circuit in such a way that it always has a loss numerically equal to the gain of the vogad. Any vogad gain will be compensated by the reverse vogad loss, so no greater tendency to sing will be effected by the addition of the combination to the circuit. In like manner half vogads must be used with compensating reverse half vogads.

Combinations of some of the devices also have interesting characteristics. For example, a combined radio noise reducer and peak limiter at the receiving end of a circuit would suppress noise and would also reduce the amplitude of excessively high amplitude signals. Likewise, a vogad, compressor, and peak chopper in tandem in the order named could be made to reduce the range of input signals by a very large amount for transmission over a medium having only a small range between noise and maximum permissible signal. In this case it would be practically impossible to recover the original signal range at the receiving terminal of the medium, but the intelligibility of speech over such a system has been shown in the laboratory to be good.

Special compandors for high quality service may require compression and expansion which vary with frequency. The exact characteristics will depend upon band width, program material and transmission medium. For transmission media in which the noise reproduced at the receiving end is principally at the higher frequencies an unusual effect is obtained if the usual variety of compandor is used. Low frequencies unaccompanied by high frequencies will cause a gain change in compressor and expandor, thus changing the background of high-frequency noise which is not masked by the low-frequency signal energy. The resulting swishing noise has been given the onomatopœic name of "hush-hush effect." To avoid this, recourse may be had to split band compandors in which the compression and expansion is done only at high frequencies or separately for low and high frequencies. The successful application of the latter method is, however, more difficult than it appears from its simple description.

Distinguishing Characteristics

It is important to distinguish between the half vogad, Fig. 2, and the compressor, Fig. 4. As shown in Table I the latter operates on syllabic variations and the former on the average volume of the input. Thus the half vogad reduces the range of output volumes to one-half that at the input while the compressor reduces the range of syllabic power at its output to one-half that at the input. In other words, the compressor reduces the ratio of peak to average power on constant volume speech, while the half-vogad simply adjusts for that volume and does not alter the peak ratio. There is, of course, the additional important difference that the half-vogad retains its gain setting during silent periods while the compressor, by virtue of having followed the syllabic power, has its maximum gain during silent periods.

Volume limiters, Fig. 3, may be mistaken for vogads, Fig. 1, because during speech input above a certain value the two may produce the same output volume. They both employ something like a measurement of average power over periods longer than a syllable to determine their gain settings. The important difference is that a vogad retains its gain setting when speech currents are not present, while a volume limiter approaches its maximum gain during such periods. In terms of the output resulting from a range of input volumes there is another important difference if the volume limiter operates over only part of the input range: the vogad reduces the width of the distribution curve of volumes to a very small value, while the volume limiter moves all the area under the distribution curve above a certain point to the region near that point, which is its limiting volume. This is illustrated in Fig. 11, in which the calculated modifications of a volume distribution by a vogad and by a volume limiter are shown. In the cases "without volume control" and "with a vogad" the distributions are normal, and the standard deviation, σ, has its usual statistical significance. With a volume limiter, only volumes above the limiting volume are affected, and these higher volumes are redistributed according to a normal law whose standard deviation is 1 decibel, as stated in the figure.

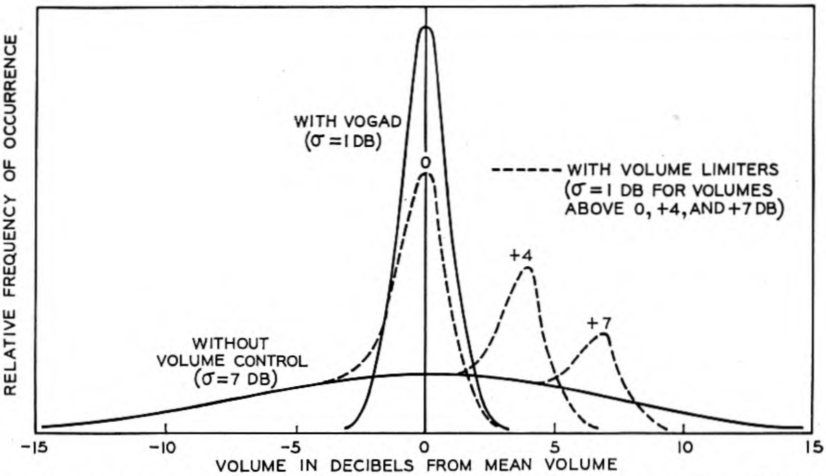


Fig. 11—Modification of volume distribution by use of a vogad or a volume limiter.

It is also important to distinguish between a peak limiter and a peak chopper, Figs. 8 and 9. Naturally they resemble one another since they are intended to permit transmission of signals at higher average amplitudes without excessive loading of transmission circuits. However, they are intended for different classes of service and hence are not interchangeable except in some borderline cases. For the highest grade of transmission harmonic production must be negligible and the reduction in amplitude range of signals small and infrequent. Gain changes must be smooth, though rapid enough to compensate for practically any input wave to be expected. These characteristics are found in the peak limiter now being furnished for use on program networks and radio transmitters.[10, 11] For services in which it is desirable to maintain the signal energy at a high value to over-ride noise and in which harmonic distortion must be kept low a peak limiter with somewhat smaller time constants may be used. A high ratio limited range compressor might be suitable in this instance. This device would lower its gain a little more quickly on excessive inputs, and it would also reinsert its gain much more quickly; it would affect the naturalness of the sound of the signal more than the slower peak limiter but it would also cause the signal to over-ride noise somewhat better. In a third variety of service the harmonic distortion introduced by a limiter is a secondary matter, the prime consideration being that the peak amplitude of the signal shall not exceed a specified value. This may be because higher amplitude signals would produce a tremendous increase in distortion or crosstalk into other channels or would damage expensive equipment farther along in the circuit. For these cases we may use the fastest possible type of limiter, the peak chopper, which simply cuts off any peak exceeding a certain value.

The crosstalk suppressor, Fig. 10, is a splendid example of the fine distinction between volume controlled and voice operated switching devices. This device has been described, but in the present state of the art its time functions have not been definitely fixed. If the characteristic of loss versus input is made steep enough and the speed of operation fast enough it will sound like a switching circuit and may in fact be replaced by a relay-switched attenuating network. If made somewhat slower and given a smaller slope of loss versus input it approaches the limited range expandor or noise reducer.

Applications and Expected Advantages

It may be of interest to give some approximate figures on the magnitudes of the advantages to be obtained by the use of some of these devices. It will be understood that the values to be given are simply illustrative, some having been obtained from field service on particular models and some from tests on laboratory equipment under special conditions.

Vogads appear to be most useful in such circuits as transoceanic radio connections, where it is important to properly operate the terminal switching equipment and to transmit over the radio circuit speech energy from loud and weak talkers equally well. It is essential in such cases that noise should not be increased in amplitude during speech pauses, hence the gain retaining feature of the vogad. On such a circuit a vogad will reduce a 45 db volume range to about 2 to 4 db. This is equivalent to expert manual volume control.

Volume limiters are in use at the present time to prevent peaks of speech energy in carrier circuits from "splashing" into telegraph channels.[7] Some 5 to 10 db limiting is allowed on loudest talkers, which causes little degradation of the speech channels but makes possible the use of telegraph on the same carrier system. There is no wide-spread use of volume limiters in point-to-point radio service so far, but in cases in which there is no disadvantage in raising noise in silent periods in speech, such as in push-to-talk installations, proper transmitter loading can be obtained with volume limiters fairly cheaply.

One commercial model peak limiter, used as part of a program amplifier [10, 11] is capable of introducing a considerable amount of compression without overloading on peaks, but for the preservation of adequate program volume range it is being recommended that only 3 db peak limiting be allowed. This, of course, reduces the range of intensity of the program, but from the standpoint of the listeners it is equivalent to doubling the transmitted power or obtaining the same signal-to-noise ratio with half the transmitted power.

Limited range compressors might be used either on land lines to insure full loading or on radio links whose fading is too severe to permit the use of normal compandors. There is no commercial application of either sort at the present time. Peak choppers are, however, used on some high power radio transmitters which might otherwise be temporarily disabled by high peaks in the signal being transmitted.

The chief usefulness of compandors is on radio links in which the transmission of a compressed signal with subsequent expansion permits operation through higher noise or with lower transmitter power. On a long-wave transatlantic radio telephone circuit a compandor with 40 db range has been shown to allow an increase in noise of some 5 db before reaching the commercial limit.[5] With smaller amounts of noise the noise advantage of the compandor approaches half its range in decibels. This benefit is sometimes applied to a reduction of transmitter power.

Radio noise reducers have been used to advantage in connection with short-wave ship-to-shore and transoceanic radio telephone service. In the former, routine transmission rating is given on a judgment basis using a merit scale from 1 to 5, 5 being practically perfect transmission and 1 so poor that intelligibility is very close to zero. It will then be seen that the observed improvement of ½ to 1 point in transmission rating due to the noise reducer is of considerable importance. Perhaps more graphic figures are those for transoceanic service, where the reduction of noise in the receiving path not only reduces the noise heard by the listener but also improves the voice operated switching with the indirect result that at times receiving volume increases of 5 to 15 db are realized.[9]

As has been noted, the radio noise reducer is a special use of an expandor alone. There are also two interesting applications for a compressor alone. The first, which uses a fairly high ratio of compression, has been mentioned as one type of peak limiting device. The second, using a moderate ratio of compression, is in connection with announcing systems for use in very noisy locations. Its effect is to amplify weak sounds more than strong sounds, which considerably improves the intelligibility through high noise. For quiet locations it is of less value, since the speech sounds lose some of their naturalness in this process.

Conclusion

In the course of developing various types of the volume controlled devices which have been described means have been worked out for providing almost any combination of time constants, range of control, and other characteristics which may be required. Some devices for which there were specific commercial applications or useful functional characteristics for experimental work have been constructed, with resulting advantages which have been briefly mentioned. There remain many possible ways to alter the characteristics of signal energy such as speech to which these methods are applicable and which await the special needs of new transmission problems.

Bibliography

1. C. C. I. F. White Book, 1 bis, pp. 77, 343.
2. C. C. I. F. White Book, 1 bis, pp. 251–3.
3. "A Vogad for Radio Telephone Control Terminals," S. Doba, Jr., *Bell Laboratories Record*, Oct. 1938, Vol. 17, No. 2, pp. 49–52.
4. "A Vogad for Radio Telephone Circuits," S. B. Wright, S. Doba, Jr., and A. C. Dickieson, Presented at *I. R. E.* Convention in New York, June 18, 1938; to be published in *Proc. I. R. E.*
5. "The 'Compandor'—An Aid Against Static in Radio Telephony," R. C. Mathes and S. B. Wright, *Elec. Engg.*, 1934, Vol. 53, No. 6, pp. 860–6; *Bell Sys. Tech. Jour.*, July 1934, Vol. 13, No. 3, pp. 315–32.
6. "The Voice Operated Compandor," N. C. Norman, *Com. and Br. Engg.*, Nov. 1934, Vol. 1, No. 1, pp. 7–9; *Bell Lab. Record*, Dec. 1934, Vol. 13, No. 4, pp. 98–103.
7. "Volume Limiter Circuits," G. W. Cowley, *Bell Lab. Record*, June 1937, Vol. 15, No. 10, pp. 311–15.
8. "A Noise Reducer for Radio Telephone Circuits," N. C. Norman, *Bell Lab. Record*, May 1937, Vol. 15, No. 9, pp. 702–7.
9. "Radio Telephone Noise Reduction by Voice Control at Receiver," C. C. Taylor, *Elec. Engg.*, Aug. 1937, Vol. 56, No. 8, pp. 971–4, 1011; *Bell Sys. Tech. Jour.*, Oct. 1937, Vol. 16, No. 4, pp. 475–86.
10. "Higher Volumes Without Overloading," S. Doba, Jr., *Bell Lab. Record*, Jan. 1938, Vol. 16, No. 5, pp. 174–8.
11. "A Volume Limiting Amplifier," O. M. Hovgaard, *Bell Lab. Record*, Jan. 1938, Vol. 16, No. 5, pp. 179–84.

For the sake of completeness the following references are included, although no allusion has been made to them under the specific device-names used in this paper.

12. "Über automatische Amplitudenbegrenzer," H. F. Mayer, *E. N. T.*, 1928, Vol. 5, No. 11, pp. 468–72.

13. "High Quality Radio Broadcast Transmission and Reception," Stuart Ballantine, *Proc. I. R. E.*, May 1934, Vol. 22, No. 5, pp. 564–629.
14. "Expanding the Music," A. L. M. Sowerby, *Wireless World*, Aug. 24, 1934, Vol. 35, No. 8, pp. 150–2.
15. "Extending Volume Range," *Radio Engg.*, Nov. 1934, Vol. 14, No. 11, pp. 7–9, 13.
16. "Amplitudenabhängige Verstärker," W. Nestel, *E. T. Z.*, 1934, Vol. 55, No. 36, pp. 882–4.
17. "An Automatic Volume Expandor," W. N. Weeden, *Electronics*, June 1935, Vol. 8, No. 6, pp. 184, 5.
18. "Die Arbeitsweise der selbsttätigen Regelapparaturen," H. Bartels and W. G. Ulbricht, *E. N. T.*, 1935, Vol. 12, No. 11, pp. 368–79.
19. "Practical Volume Expansion," C. M. Sinnett, *Electronics*, Nov. 1935, Vol. 8, No. 11, pp. 428–30, 446.
20. "Light-bulb Volume Expandor," *Electronics*, Mar. 1936, Vol. 9, No. 3, p. 9.
21. "Simplified Volume Expansion," W. N. Weeden, *Wireless World*, Apr. 24, 1936, Vol. 38, No. 17, pp. 407–8.
22. "Practical Volume Compression," L. B. Hallman, Jr., *Electronics*, June 1936, Vol. 9, No. 6, pp. 15–17, 42.
23. "Notes on Contrast Expansion," Gerald Sayers, *Wireless World*, Sept. 18, 1936, Vol. 39, No. 12, p. 313.
24. "Contrast Amplification: A New Development," W. N. Weeden, *Wireless World*, Dec. 18, 1936, Vol. 39, No. 25, pp. 636–38.
25. "Overmodulation Control and Volume Compression with Variable-mu Speech Amplifier," W. B. Plummer, *Q. S. T.*, Oct. 1937, Vol. 21, No. 10, pp. 31–33.
26. "Limiting Amplifiers," John P. Taylor, *Communications*, Dec. 1937, Vol. 17, No. 12, pp. 7–10, 39–40.
27. "Low Distortion Volume Expansion Using Negative Feedback," B. J. Stevens, *Wireless Engr.*, Mar. 1938, Vol. 15, No. 174, pp. 143–9.
28. "Distortion Limiter for Radio Receivers," M. L. Levy, *Electronics*, Mar. 1938, Vol. 11, No. 3, p. 26.
29. "Automatic Modulation Control," L. C. Waller, *Radio*, Mar. 1938, No. 227, pp. 21–6, 72, 74.
30. "An AVE Noise Silencer Unit," McMurdo Silver, *Radio News*, May 1938, Vol. 20, No. 11, pp. 46, 55.

The Exponential Transmission Line *

By CHAS. R. BURROWS

The theory of the exponential transmission line is developed. It is found to be a high pass, impedance transforming filter. The cutoff frequency depends upon the rate of taper.

The deviation of the exponential line from an ideal impedance transformer may be decreased by an order of magnitude by shunting the low impedance end with an inductance and inserting a capacitance in series with the high impedance end. The magnitudes of these reactances are equal to the impedance level at their respective ends of the line at the cutoff frequency.

For a two-to-one impedance transformer the line is 0.0551 wavelengths long at the cutoff frequency. For a four-to-one impedance transformer the line is 0.1102 wave-lengths long at the cutoff frequency, etc.

The results have been verified experimentally. Practical lines 50 meters and 15 meters long have been constructed which transform from 600 to 300 ohms over the frequency range from 4 to 30 mc. with deviations from the ideal that are small compared with the deviations from the ideal of commercial transmission lines, either two-wire or concentric.

When an exponential line is used as a dissipative load of known impedance instead of a uniform line it is possible to approach more nearly the ideal of constant heat dissipation per unit length. This makes it possible to use a shorter line.

THE exponential line may be defined as an ordinary transmission line in which the spacing between the conductors (or conductor size) is not constant but varies in such a way that the distributed inductance and capacitance vary exponentially with the distance along the line. That is, the impedance ratio for two points a fixed distance apart ıs independent of the position of these two points along the line. A disturbance is propagated down an exponential transmission line in the same manner as it would be down a uniform line with the additional effect that the voltage is increased by the square root of the change in impedance level and the current is decreased by the reciprocal of this quantity.

The exponential line has the properties of a high pass impedance transforming filter. The cutoff frequency depends upon the rate of

* Presented before joint meeting of U. R. S. I., and I. R. E., Washington, D. C., April 1938. Published in *Communications*, October 1938.

taper. As the frequency is increased the transfer constant * approaches the propagation constant of the equivalent uniform line. At sufficiently low frequencies the only effect of the line is to connect the input to the load.

Above cutoff the magnitudes of the characteristic impedances at any point are approximately equal to the nominal characteristic impedance * at that point but their phase angles (in radians) differ by an amount which at the higher frequencies is equal to the cutoff frequency divided by the frequency in question. The ratio of input impedance to the input impedance level * of an exponential line terminated in a resistance equal to the impedance level at the output always remains within the range from $1 - f_1/f$ to $1/(1 - f_1/f)$ for frequencies, f, greater than the cutoff frequency, f_1. For a 2 : 1 transformation this means that the input impedance remains within ± 6 per cent of the desired value for all frequencies above that for which the line is a wave-length long. For a 4 : 1 transformation under the same conditions the irregularities are twice as great.

A transforming network having deviations from the ideal of the order of $\pm\ (f_1/f)^2$ may be made by connecting an inductance in parallel with the low impedance terminal and a capacitance in series with the high impedance terminal. The magnitudes of these reactances are such that their impedances are equal to the impedance levels of the line at their respective ends at the cutoff frequency. Or expressed in another way the capacitance is equal to $2/(k - 1)$ times the electrostatic capacitance of the line and the inductance is the same factor times the total loop inductance of the line where k is the impedance transformation ratio of the line.

Figure 1 shows the theoretical input impedance-frequency characteristics for 2 to 1 step-up and step-down exponential lines. Curve 1 is for the line with a resistance termination. At low frequencies the input impedance is equal to the load impedance while at high frequencies the line approaches an ideal transformer. Curve 2 is the input impedance of the line terminated with the appropriate resistance-reactance combination. The improvement in the input impedance characteristic for frequencies above the cutoff frequency is evident. At the lower frequencies the input impedance does not approach the terminal reactance but approaches the reactance of the capacitance of the line in parallel with the series terminal capacitance for the step-up line and the reactance of the inductance of the line in series with the shunt terminal inductance for the step-down line. The improvement is not as great as apparent from the figures because the phase angle is

* See appendix for definition of terms.

not improved proportionally. This is easily remedied by completing the impedance transforming network with the appropriate reactance at the input. The resulting input impedance is shown in curve 3. In the "pass" frequency range the maximum reactive component is of the same order of magnitude as the deviation of the impedance from the ideal.

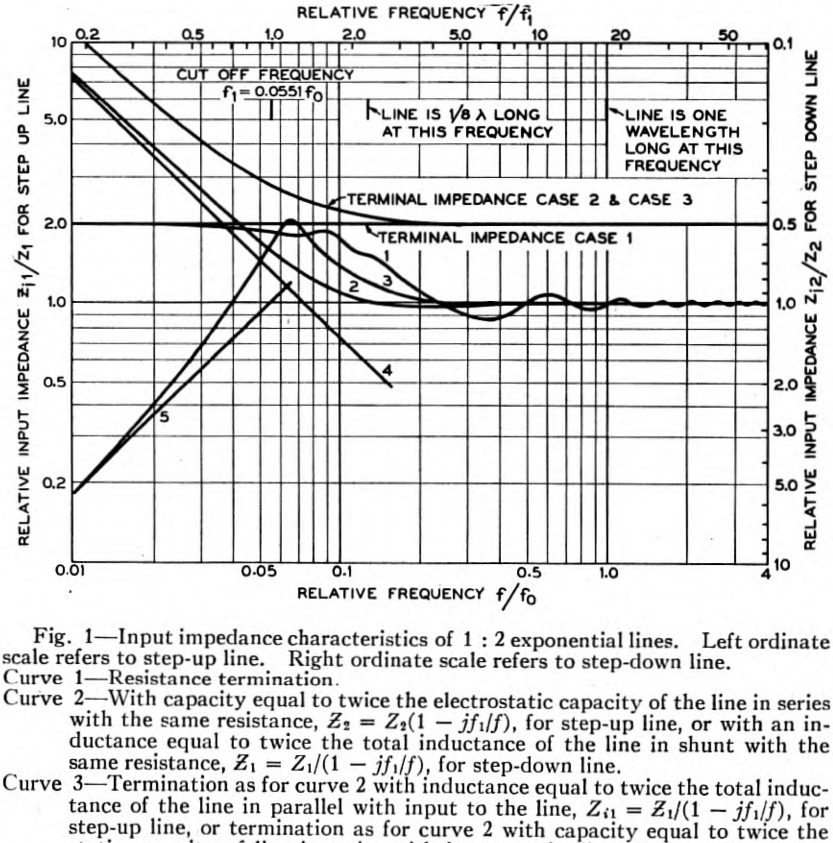


Fig. 1—Input impedance characteristics of 1 : 2 exponential lines. Left ordinate scale refers to step-up line. Right ordinate scale refers to step-down line.
Curve 1—Resistance termination.
Curve 2—With capacity equal to twice the electrostatic capacity of the line in series with the same resistance, $Z_2 = Z_2(1 - jf_1/f)$, for step-up line, or with an inductance equal to twice the total inductance of the line in shunt with the same resistance, $Z_1 = Z_1/(1 - jf_1/f)$, for step-down line.
Curve 3—Termination as for curve 2 with inductance equal to twice the total inductance of the line in parallel with input to the line, $Z_{i1} = Z_1/(1 - jf_1/f)$, for step-up line, or termination as for curve 2 with capacity equal to twice the static capacity of line in series with input to the line, $Z_{i2} = Z_2(1 - jf_1/f)$.
Curve 4—Asymptotic value of impedance of capacity of line in parallel with termination for case 2 for step-up line, or asymptotic value of impedance of inductance in series with termination for case 2 for step-down line.
Curve 5—Impedance of shunt inductance added at input for case 3 for step-up line, or impedance of capacity added in series at input for case 3 for step-down line.

Besides its application as an impedance transforming network, the exponential line may be used as a "resistance" load of constant known impedance that has a high capability for dissipating power. As such it is capable of dissipating more power in the same length of line than the uniform line. If x is the maximum attenuation in nepers that can be obtained with a uniform line without overheating, the same length of exponential line will have an attenuation of $(e^{2x} - 1)/2$ nepers.

Exponential lines of the proper length have properties similar to half-wave and quarter-wave uniform lines. The input impedance of an exponential line an even number of quarter wave-lengths long is equal to the load impedance times the impedance transformation ratio of the line. When the length of the line differs from an odd multiple of a quarter wave-length by an amount that depends upon the frequency and load impedance, the input impedance is equal to the product of the terminal impedance levels divided by the load impedance.

Mathematical Formulation

The telegraph equations for the exponential line may be solved by the methods employed in the problem of a uniform line. The resulting equations for the voltage and current at any point along the line are

$$v_x = Ae^{-\left(\Gamma - \frac{\delta}{2}\right)x} + Be^{+\left(\Gamma + \frac{\delta}{2}\right)x} = Ae^{-\left(\Gamma - \frac{\delta}{2}\right)x}\left[1 + \frac{B}{A}e^{2\Gamma x}\right] \qquad (1)$$

and

$$i_x = \frac{A}{Z_0}\frac{\Gamma - \frac{\delta}{2}}{\gamma}e^{-\left(\Gamma + \frac{\delta}{2}\right)x} - \frac{B}{Z_0}\frac{\Gamma + \frac{\delta}{2}}{\gamma}e^{+\left(\Gamma - \frac{\delta}{2}\right)x}$$

$$= \frac{A}{Z_0}\frac{\Gamma - \frac{\delta}{2}}{\gamma}e^{-\left(\Gamma + \frac{\delta}{2}\right)x}\left[1 - \frac{B}{A}\frac{\Gamma + \frac{\delta}{2}}{\Gamma - \frac{\delta}{2}}e^{2\Gamma x}\right], \qquad (2)$$

where

$\delta = \dfrac{\log_e z/z_0}{x} = \dfrac{\log_e y_0/y}{x} = \dfrac{\log_e Z/Z_0}{x}$ is the rate of taper,

$Z_x = \sqrt{z/y} = Z_0 e^{\delta x}$ is the *surge* or *nominal characteristic impedance* of the exponential line at the point x which is equal to the characteristic impedance of the uniform line that has the same distributed constants as this line has at the point x,

$\gamma = \sqrt{zy} = \sqrt{z_0 y_0}$ is the *propagation constant* of any uniform line that has the same distributed constants as this line at any point. It is independent of the point along the line to which it is referred, and

$\Gamma = \sqrt{\gamma^2 + \delta^2/4} = \alpha + j\beta$ is the *transfer constant* of the exponential line.

$+\gamma$ and $+\Gamma$ refer to the values of the indicated roots that are in the first quadrant.

If these equations are compared with those for a uniform transmission line it is found that the *propagation constant* is $\Gamma - \delta/2$ for voltage waves traveling in the positive x direction and $\Gamma + \delta/2$ for voltage waves traveling in the negative x direction. For current waves the corresponding *propagation constants* are $\Gamma + \delta/2$ and $\Gamma - \delta/2$. In the terminology of wave filters, Γ is the *transfer constant* and δ is the impedance *transformation constant.* $\delta/2$ is the *voltage transformation constant* and $-\delta/2$ is the *current transformation constant.* The real and imaginary parts of Γ, α and β are the *attenuation* and *phase constants* respectively.

An important parameter is

$$\nu = j\frac{\delta}{2\gamma},$$

which for a non-dissipative line is the ratio of the cutoff frequency to the frequency, as can be seen if we write the transfer constant as

$$\Gamma = \gamma\sqrt{1 - \nu^2},$$

where the indicated root is in the fourth quadrant. For a non-dissipative line ν is real and the transfer constant is real or imaginary depending on whether ν^2 is greater than or less than unity. Hence the exponential line is a high pass filter whose cutoff frequency, f_1, is that frequency for which $\nu = \pm 1$. The transfer constant is then less than that for a uniform line by the factor $\sqrt{1 - \nu^2}$ so that both phase velocity and wave-length are larger for the exponential line than for the uniform line by the reciprocal of this factor.

If we terminate this line at $x = l$ with an impedance $\mathcal{Z}_l = v_l/i_l$, the ratio of the reflected to direct voltage wave is found to be

$$\frac{B}{A} = -\frac{1 - (\mathcal{Z}_l/Z_l)(\sqrt{1 - \nu^2} + j\nu)}{1 + (\mathcal{Z}_l/Z_l)(\sqrt{1 - \nu^2} - j\nu)}\, e^{-2\Gamma l}, \tag{3}$$

where the coefficient of the exponential is the *voltage reflection coefficient.*

There will be no reflection if

$$\mathcal{Z}_l = Z_l/(\sqrt{1 - \nu^2} + j\nu) = Z_l^{+}, \tag{4}$$

which becomes $Z_l e^{-j\sin^{-1}\nu}$ above the cutoff frequency for non-dissipative lines. This is the magnitude of the forward-looking *characteristic impedance* at $x = l$ as can be seen by dividing the first term of (1) by the first term of (2). Curve 1 of Fig. 2 gives the characteristic impedance of a non-dissipative exponential line looking toward the high impedance end as a function of frequency. At infinite frequency the characteristic impedance is a resistance equal to the nominal characteristic impedance but as the frequency is decreased the phase angle of the characteristic impedance changes so that its locus

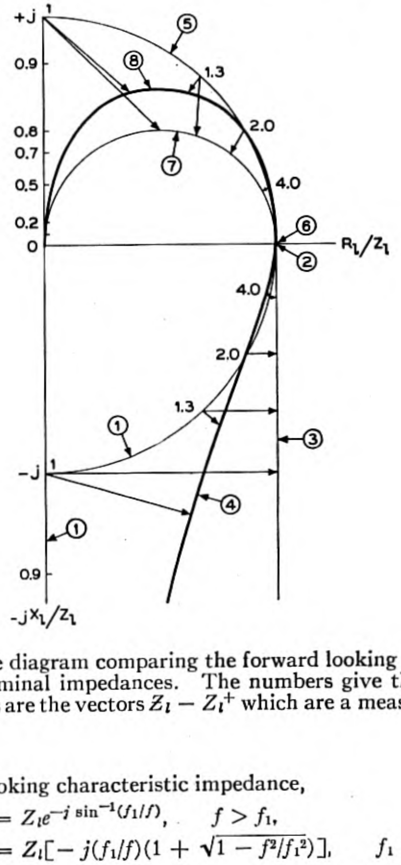


Fig. 2—Impedance diagram comparing the forward looking characteristic impedance with various terminal impedances. The numbers give the frequency relative to cutoff. The arrows are the vectors $\mathcal{Z}_l - Z_l^+$ which are a measure of the magnitude of the reflection.

A. Step-up line.

Curve 1—Forward looking characteristic impedance,

$$Z_l^+ = Z_l e^{-j \sin^{-1}(f_1/f)}, \qquad f > f_1,$$
$$Z_l^+ = Z_l[-j(f_1/f)(1 + \sqrt{1 - f^2/f_1^2})], \qquad f_1 > f;$$

Curve 2—Resistance termination, $\mathcal{Z}_l = Z_l$;
Curve 3—Capacity resistance termination, $\mathcal{Z}_l = Z_l(1 - jf_1/f)$;
Curve 4—Capacity, resistance and inductance termination adjusted for no reflection at twice the cutoff frequency and at infinite frequency;

B. Step-down line.

Curve 5—Forward-looking characteristic impedance,

$$Z_l^+ = Z_l e^{+j \sin^{-1}(f_1/f)}, \qquad f > f_1,$$
$$Z_l^+ = Z_l[+j(f_1/f)(1 - \sqrt{1 - f^2/f_1^2})], \qquad f_1 > f;$$

Curve 6—Resistance termination $\mathcal{Z}_l = Z_l$;
Curve 7—Inductance resistance termination $\mathcal{Z}_l = Z_l/(1 - jf_1/f)$;
Curve 8—Inductance, resistance and capacity termination adjusted for no reflection at twice the cutoff frequency and at infinite frequency.

is the circular arc. At and below cutoff it is a pure reactance. If the load is a resistance equal to the nominal characteristic impedance at the terminal as indicated at 2 of Fig. 2, there will be no reflection at infinite frequency, but as the frequency is lowered there will be an increasing impedance mismatch with its accompanying reflected wave.

This reflection may be materially reduced by inserting a condenser in series with the resistance load as shown by curve 3. Further improvement results from more complicated networks. Curve 4 shows the effect of adding an inductance in shunt with the resistance load of the resistance-capacitance combination. The arrows indicate the resulting impedance mismatch which is a measure of the reflected wave.

The characteristic impedance looking toward the low impedance end is the inverse of that looking in the other direction as shown by curve 5. Shunting the resistance load with an inductance gives the impedance curve 7. Adding a capacitance element gives curve 8.

Division of (1) by (2) and substitution of the result of (3) gives the following ratio for the impedance looking into the line at the point x to the impedance level at that point,

$$\frac{Z_x}{Z_x} = \frac{K(\sqrt{1-\nu^2} - j\nu) + 1 + [K(\sqrt{1-\nu^2} + j\nu) - 1]e^{-2\Gamma(l-x)}}{K + j\nu + \sqrt{1-\nu^2} - [K - \sqrt{1-\nu^2} + j\nu]e^{-2\Gamma(l-x)}}, \tag{5}$$

where $K = Z_l/Z_l$ is the ratio of the load impedance to the impedance level at the terminal. Here as before the indicated root is in the fourth quadrant.

Network Characteristics

Three parameters are required to specify the characteristics of an exponential line of negligible loss: (1) the cutoff frequency, f_1, (2) the length of the line which is perhaps best specified as the frequency, f_0 = velocity of light/length of line, for which the line is one wavelength long, and (3) the impedance level at some point along the line. We will designate the impedance levels at the low and high impedance ends of the line by Z_1 and Z_2 respectively, and their ratio Z_2/Z_1 by k.

When the line is terminated in a resistance equal to the impedance level at the output (5) reduces to

$$\frac{Z_1}{Z_1} = k^{\cos 2\zeta} e^{2j\zeta} \frac{1 + j \tan \zeta \, k^{-\cos 2\zeta}}{1 - j \tan \zeta \, k^{\cos 2\zeta}}, \qquad \nu > 1, \tag{6}$$

$$\frac{Z_1}{Z_1} = e^{-2j\xi} \frac{1 + \tan \xi \, e^{j\left(2\xi + \frac{\pi}{2} - \eta\right)}}{1 + \tan \xi \, e^{j\left(-2\xi - \frac{\pi}{2} - \eta\right)}}, \qquad \nu < 1. \tag{7}$$

for frequencies below and above cutoff respectively. Here $\eta = - j2\Gamma l$ is twice the electrical length of the line in radians, $\sin 2\zeta = 1/\nu$, $\sin 2\xi = \nu$ and $\cos 2\xi$ is ratio of the electrical length of the line to that of a uniform line of the same physical length. For the step-down line the corresponding ratios are the reciprocal of the above expressions. These ratios are plotted in Fig. 1.

When $f \rightarrow 0$, $\mathcal{Z}_1 = kZ_1 = Z_2 = \mathcal{Z}_2$ and the only effect of the line is to connect the load to the input. Above cutoff the magnitude of the input impedance oscillates about the nominal characteristic impedance and the phase angle oscillates about the value $- 2\xi (\approx - f_1/f$ for $f \gg f_1)$ which goes from $- \pi/2$ to 0 as the frequency increases indefinitely from cutoff. The variation of the input impedance with frequency is shown for two lines of different length but the same rate of taper in Fig. 3. The magnitude of the oscillations depends only on the rate of taper and decreases with increase in frequency. The impedance varies between $(1 + f_1/f)$ and $1/(1 + f_1/f)$. The positions of the maxima and minima, however, are determined by the length of the line. They occur respectively at those frequencies for which the line is approximately 1/8 of a wave-length more than an even or an odd number of quarter wave-lengths long. The phase angle is usually negative but has a small positive value when the line is approximately a half wave-length long.

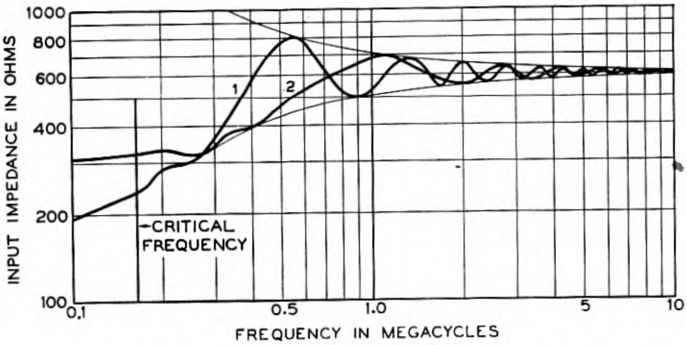


Fig. 3—Input impedance characteristics.
Curve 1—150 : 600 ohm line, 100 meters long.
Curve 2—300 : 600 ohm line, 200 meters long.
Both lines have the same rate of taper.

The locations of these maxima and minima are the same as would result from terminating a uniform line in an impedance whose magnitude is the same as the characteristic impedance but has a small reactive component. This suggests adding a compensating reactance to the resistance load. From (3) the best single reactive element is found to be a condenser whose impedance is equal to the impedance level at the cutoff frequency. This gives a value of $K = 1 - j\nu$ which when substituted in (5) shows that the input impedance is to a first approximation a constant times the terminal impedance. To correct for the reactive component of the input impedance an inductance having an impedance jZ_1/ν which is equal to the input impedance level at cutoff is shunted across the input. The resulting impedance transforming network consists of an exponential line with a series capacitance at the high impedance end and a shunt inductance at the low impedance end. When terminated in a resistance load at either end equal to the impedance level at that end the input impedance, to a first approximation, is a resistance equal to the impedance level at the input end. In fact the deviations of the input impedance from the ideal for transmission in one direction are just the reciprocal of those for transmission in the other direction.

The magnitudes of the series capacitance and shunt inductance that give the improved network may be expressed in terms of the electrostatic capacitance and loop inductance of the line. Simple calculation shows that the required series capacitance is equal to $2/(k - 1)$ times the electrostatic capacitance of the line and the required shunt inductance is equal to the same factor times the total inductance of the line.

There is an interesting relationship between these terminations and a simple high-pass filter. The LC product of the shunt and series arms of the filter resonates at f_1. If an ideal transformer with transformation ratio k is inserted between the shunt inductance and the series capacitance, the capacitance becomes C/k and the new LC resonates at $f_1\sqrt{k}$. This is the same frequency at which the series capacitance and shunt inductance that are added to the terminations of the exponential line resonate. Furthermore the reactance of the shunt inductance is equal to the impedance level at the cutoff frequency and the reactance of the series capacitance is equal to the impedance level at the cutoff frequency exactly as in the case of the high-pass filter.

By using the exponential line it is possible to construct a network with properties that no network with lumped circuit elements possesses, namely, a high-pass impedance transforming filter.

Critical Lengths

Besides the characteristics of the exponential line that are substantially independent of the length of the line, it has properties that depend on the length of the line that are analogous to those of a uniform line a half wave-length or quarter wave-length long. For non-dissipative lines above the cutoff frequency (5) becomes

$$\mathcal{Z}_1 = \frac{K \cos\left(\frac{\eta}{2} - 2\xi\right) + j \sin \frac{\eta}{2}}{\cos\left(\frac{\eta}{2} + 2\xi\right) + jK \sin \frac{\eta}{2}} \mathcal{Z}_1. \tag{8}$$

When the line is an integral number of half wave-lengths long ($2\eta = \pi$) this reduces to

$$\mathcal{Z}_1 = KZ_1 = k\mathcal{Z}_2, \tag{9}$$

which says that the input impedance is equal to the impedance transformation ratio times the load impedance. The length of exponential line that corresponds to a quarter-wave uniform line differs from an odd multiple of a quarter wave-length by an amount such that

$$\tan\left(\frac{\eta - (2n + 1)\pi}{2}\right) = \frac{K^2 - 1}{K^2 + 1} \tan 2\xi, \tag{10}$$

for which (8) becomes

$$\mathcal{Z}_1 = \frac{Z_1 Z_2}{\mathcal{Z}_2}. \tag{11}$$

Similar expressions exist for the step-down line, but $1/K$ must be substituted for K in (10) for the length corresponding to the quarter-wave uniform line.

With Dissipation

An exponential line is an improvement over the uniform "iron wire" line as a resistance load that will dissipate a large amount of power.

Provided the attenuation is not too large the current and voltage distribution will be the same as for a non-dissipative line except for the additional power loss so that we may use the equations for an exponential line even though the distributed series resistance and shunt leakage do not vary exponentially with distance.

Suppose that the conductor size and resistance that will just dissipate the desired input power result in an attenuation constant α_0 for a uniform transmission line. To a first approximation the conductors can carry the same current irrespective of the impedance level. The current wave will be given by the first term of equation (2) which becomes

$$i = i_0 e^{-(\delta/2)x - \alpha_0 x},$$

except for a phase factor. In order that the current will not increase, $\delta = -2\alpha$. The actual attenuation "constant,"

$$\alpha_x \sim \left(\frac{R}{2Z_x} + \frac{GZ_x}{2}\right)\left(1 + \frac{1}{2}\frac{f_1^2}{f^2}\right) \sim \alpha_0 e^{2\alpha_0 x}\left(1 + \frac{1}{2}\frac{f_1^2}{f^2}\right), \qquad (13)$$

will increase with distance down the line so that the current will decrease but not as rapidly as with a uniform line. The total attenuation in nepers is approximately

$$\left(1 + \frac{1}{2}\frac{f^2}{f_1^2}\right)\int_{x=0}^{l} \alpha_0 e^{2\alpha_0 x} dx = \left(1 + \frac{1}{2}\frac{f^2}{f_1^2}\right)\left(\frac{e^{2\alpha_0 l} - 1}{2}\right). \qquad (14)$$

At the point where the attenuation of the uniform line is 6 db the tapered line has an additional attenuation of 7 db above the uniform line or a total attenuation of more than twice. The current has been reduced to less than half. Here an improvement may be made by increasing the dissipation by either changing the wire size or resistivity of the conductor. A greater improvement would result from changing the resistivity because then the capacity for heat dissipation would be the same. Suppose, however, that one conductor material is to be used throughout and the dissipation capacity is proportional to the wire surface; then at this point the wire size could be reduced to 1/2, doubling the attenuation factor. It is already 4 times that for the uniform line, so this increases it to 8 times. The resulting total attenuation is 30 db in a length that would have less than 7 db if the line were uniform. If this attenuation were required the length of line could be reduced by a factor of about 4.4. Of course the spacing is very close at the end of this line, but the line could be shorted at the end. This would approximately double the current at the end, but here again the current carrying capacity of the line is more than double the current traveling down the line. With the line shorted the reflected current would be 60 db down, which would not affect the input impedance appreciably. For the first 13 db of attenuation the impedance of the line would be relatively free from changes due to changes in spacing resulting from wind, etc. When the spacing is small enough to be affected by wind, vibration, etc., the attenuation will be great enough to suppress these small irregularities.

Experiment

In order to verify the foregoing theoretical development, measurements have been made on several experimental lines. Figure 4 shows the results of measurements on two such lines. These lines were constructed of No. 12 tinned copper. At the low impedance end the strain was taken by a victron insulator which also served as a line spreader and terminal mounting. At the high impedance end the strain was taken by 1/4″ manila rope without other insulation. The line spacing was adjusted by "lock stitch" tension insulators spaced 1 meter and 1/2 meter apart on the low and high impedance end respectively of the 9-meter line. The 3-meter line was supported at the 1/4, 1/2, 2/3, 3/4 and 7/8th points.

The impedance was measured by the substitution method. To facilitate the substitution of the reactive component of the line it was bridged by an antiresonant circuit. Pencil leads calibrated on direct current were used as the resistance standards. Type BW IRC 1/2 watt resistances were used for terminations. The solid circles of Fig. 4 represent measurements on the 9-meter line. The agreement with theory is as good as is usually found for actual "uniform lines." In order to check the theory further toward the lower frequency end—beyond the range of the measuring equipment—measurements were made on a 3-meter line. These measurements are shown by the open circles. The agreement with theory is not so good, but here the lengths of the connecting leads are an appreciable fraction of the length of the line.

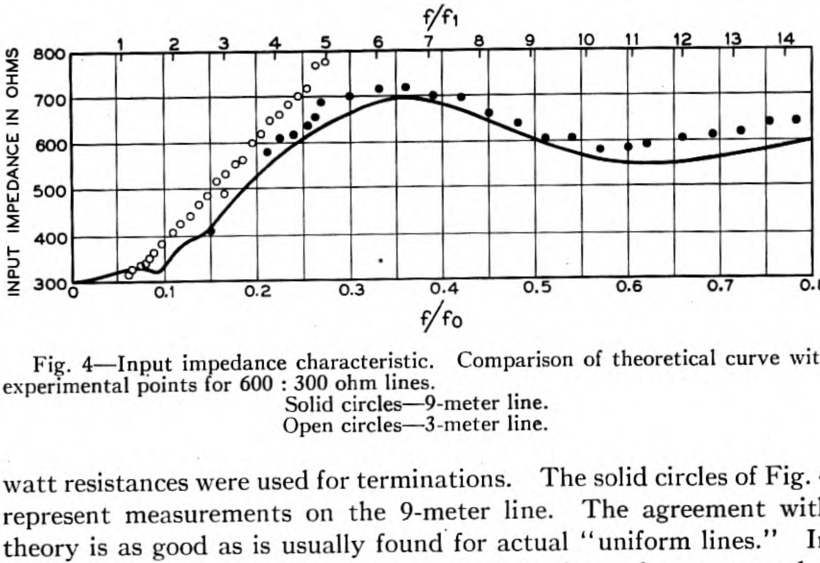


Fig. 4—Input impedance characteristic. Comparison of theoretical curve with experimental points for 600 : 300 ohm lines.
Solid circles—9-meter line.
Open circles—3-meter line.

Preliminary tests on a full size model of exponential line impedance transformer showed deviations from the theoretical that might be attributed to improper termination, irregularities along the line, irregularities introduced at the change in conductor size or capacitance of the spacing insulators. Since it was impossible to determine which of these was the predominant cause of the deviations from the ideal, it was decided to introduce each of these factors one at a time. This test was made on a 600 : 300 ohm line constructed of No. 6 copper wire with lockstitch insulators except at the terminals. The correct termination was obtained by tests on a uniform 300 ohm line with the same physical structure at the termination. Of necessity the tying of the wire to the strain insulators at the end introduced a shunt capacitance which augmented the inherent additional capacitance due the "end effect." This additional capacitance is equal to that of a short length of line.

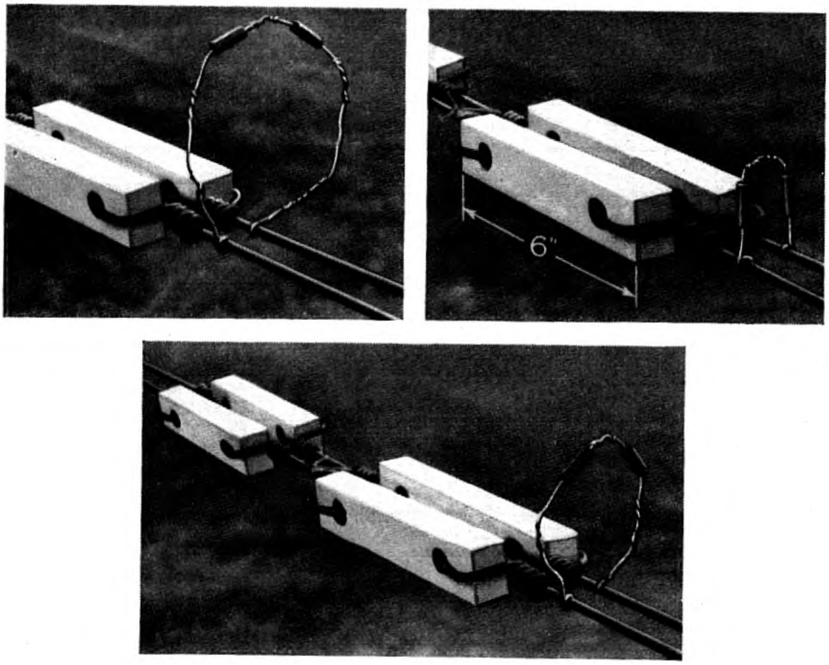


Fig. 5—Photographs of terminations of 300 ohm line.
upper right for curve 1
lower for curve 2 } of Fig. 6.
upper left for curve 3

If the correct amount of inductance is inserted in series with the resistance load the combined effect of the additional capacitance and inductance becomes the same as the addition of a small length of line for all frequencies up to those for which this length of line is an appreciable fraction of a wave-length. Accordingly a small amount of inductance was inserted in series with the resistance as shown in the right picture of Fig. 5. The input impedance of the uniform line with this termination is given by "Experimental Curve 1" at the bottom of Fig. 6.

A three-inch length of No. 18 wire was inserted as shown in the lower picture of Fig. 5 and "Experimental Curve 2" resulted. This reduced the irregularities in the input impedance to about half, so another three inches were inserted, resulting in "Experimental Curve 3." Here the maxima and minima are displaced, indicating that the effect of the stray capacitance has been reduced to the same order of magnitude as that due to the deviation of the resistance from the desired value. This

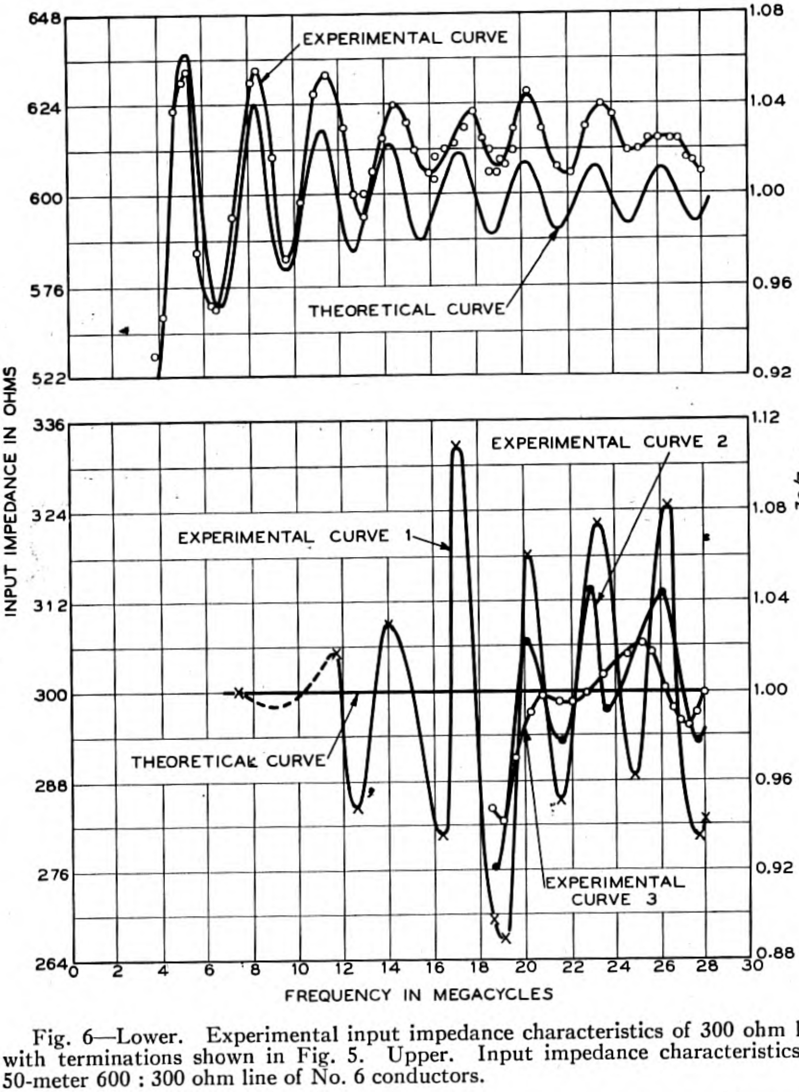


Fig. 6—Lower. Experimental input impedance characteristics of 300 ohm line with terminations shown in Fig. 5. Upper. Input impedance characteristics of 50-meter 600 : 300 ohm line of No. 6 conductors.

termination was accordingly removed to the exponential line, resulting in the "Experimental Curve" at the top of Fig. 6. It agrees within experimental error with the "Theoretical Curve." The slight vertical displacement of the experimental curve at the higher frequencies is attributed to deviations in the impedance of the pencil lead, which was used as a resistance standard, from a pure resistance equal to its direct current value.

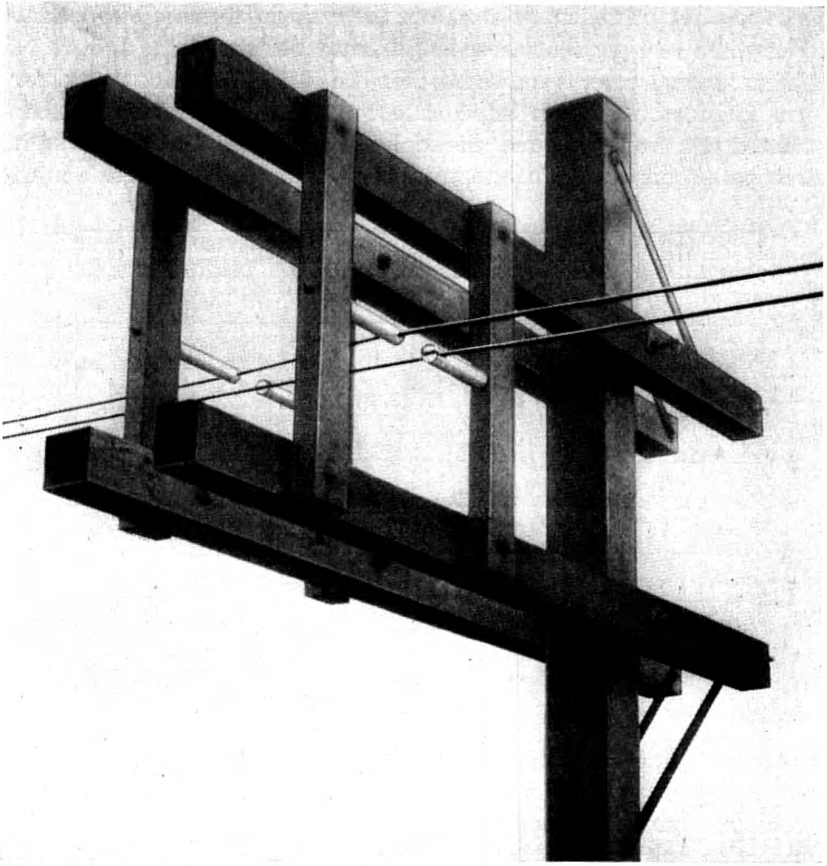


Fig. 7—Photograph of one of the changes in conductor size.

To increase the power carrying capacity of the exponential line, one was built with larger wire size at the lower impedance end. This increased the breakdown voltage by increasing the spacing and conductor diameter and at the same time increased the current carrying capacity by decreasing the resistance and increasing the heat dissipating capacity of the conductors. This was a 600 : 300 ohm line constructed of 20 meters No. 6 wire, 10 meters 1/4″ tubing and 20 meters 3/8″ tubing. Here again the correct termination was determined by measurements on a 300 ohm uniform line of 3/8″ tubing. The total length of terminating loop that gave the best termination was 6½″ in this case compared with 10½″ for the 300 ohm line of No. 6 wire. Since no attempt was made to reduce the variations in input impedance to less than ± 1 per cent these lengths may be as much as an inch off.

These measurements indicated that the exponential line would perform satisfactorily as an impedance transformer if it could be constructed to have the desired mechanical features without impairing its electrical properties. The greatest difficulty appeared to reside in the insulators. Special isolantite insulators were designed that would be satisfactory commercially and still keep the additional capacity to a reasonable value. Figure 7 shows the construction of the line at the supporting poles where the conductor size changes.

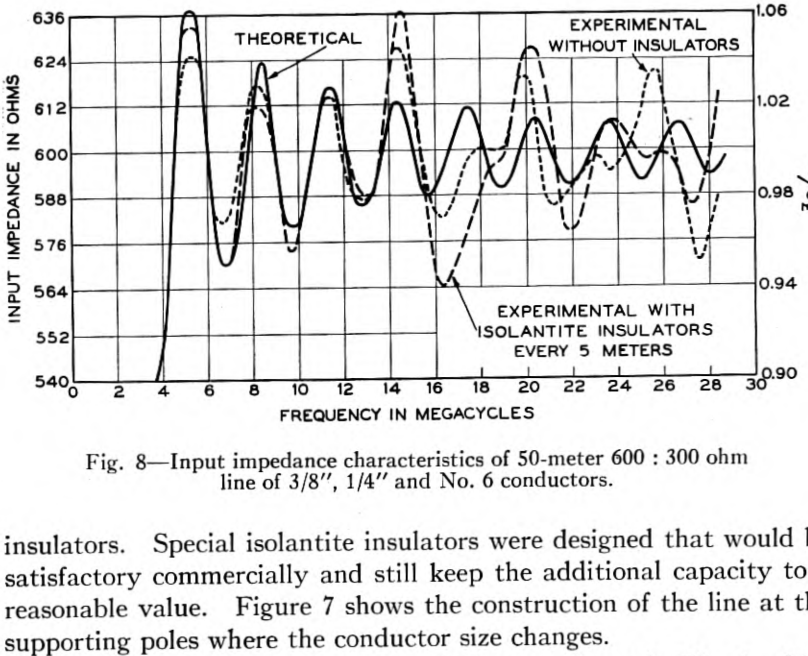


Fig. 8—Input impedance characteristics of 50-meter 600 : 300 ohm line of 3/8″, 1/4″ and No. 6 conductors.

The results of measurements on this line are shown in Fig. 8. The solid curve was calculated from the equations developed earlier. The two broken curves are the results of measurements on the line, one without insulators and one with insulators. While the insulators affect the line somewhat they do not increase the deviation from the ideal appreciably. [The improvement in the agreement between experiment and theory in this set of curves over that in Fig. 4 is presumably due to the fact that the comparison resistance for Fig. 8 consisted of 3-IRC resistances instead of the pencil lead. With the fixed IRC resistance it was, of course, impossible to adjust the standard to exactly the same value as the unknown. In this case the small difference was determined by using the slope of the rectifier voltmeter calibration.] This line has a maximum deviation from the desired input impedance of ± 6 percent for all frequencies above 4.2 mc. (Measurements were made up to 28 mc.) The phase angle of the input impedance was found to be zero within the accuracy of measurement. From theory the phase angle would be expected to vary between − 0° and + 3°.

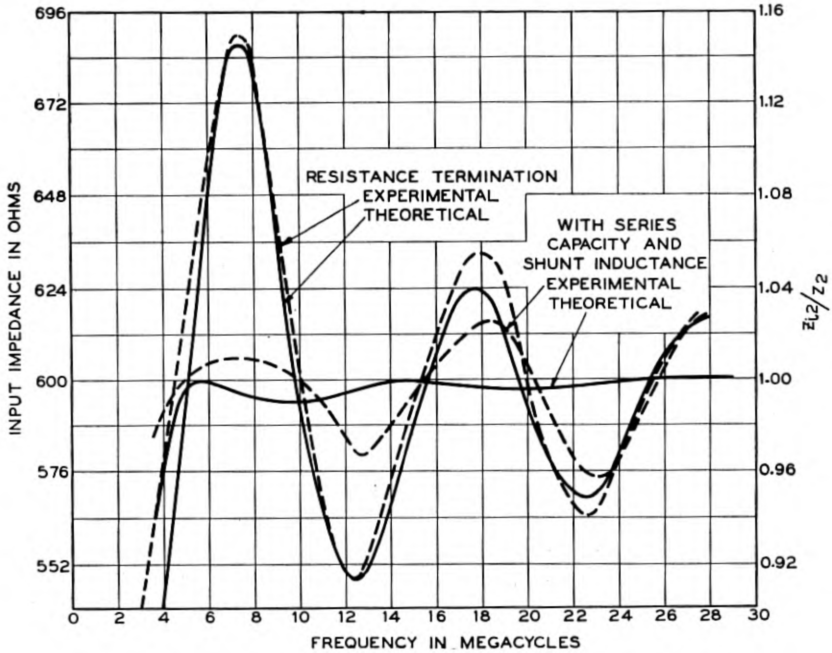


Fig. 9—Input impedance characteristics of 15-meter 600 : 300 ohm line of No. 6 conductors

The curves of Fig. 9 refer to a 600 : 300 ohm line of No. 6 wire 15 meters long. With resistance termination this line has rather large variations in the input impedance but with the addition of the proper reactances the input impedance is flatter than the longer line with resistance termination. At the lower frequencies where the variations in the input impedance were large without the reactive networks, their addition gives approximately the expected improvement. At the higher frequencies the inductance was approximately anti-resonated by its distributed capacity and the input impedance approaches that for the resistance termination.

Conclusion

Theory indicates that the exponential line may be used as an impedance transformer over a wide frequency range. The results of experiment show that the desired characteristic can be realized in practice. Among the applications of the exponential line may be mentioned its use in transforming the impedance level back to its original value after the paralleling of two transmission lines feeding two antennas. It could be used to transform the input impedance of a rhombic antenna down to the usual 600-ohm level of open wire transmission lines. If twin coaxial lines are used inside the transmitter building to eliminate undesired feedback, coupling, etc., the exponential line could be used to transform from the highest practical impedance level of such lines to a practical level of the more economical open wire lines for use outside the building.

Appendix

The exponential line is a non-uniform line so that the terms "characteristic impedance" and "surge impedance" of an exponential line are not synonymous. The terms "surge impedance"[1] and "nominal characteristic impedance"[2] may be used synonymously for the characteristic impedance of the uniform line that has the same distributed constants as the non-uniform line at the point in question. Expressed as functions of the distributed "constants" of the line they are the square root of the ratio of the distributed series impedance to the distributed shunt admittance at the point along the line in question. It will be expedient to refer to the nominal characteristic impedance as the impedance level at the point in question. Schelkunoff[3] has defined the characteristic impedances as the ratio of voltage to current at the point in question for each of the two traveling waves of which

[1] The term "surge impedance" is defined by A. E. Kennelly on page 73 of "The Applications of Hyperbolic Functions to Electrical Engineering Problems" (McGraw-Hill 1916) as follows: "The surge impedance of the line is not only the natural impedance which it offers everywhere to surges of the frequency considered, but it is also the initial impedance of the line at the sending end." Hence the "surge impedance" should be independent of the configuration of the line except at the point in question and in particular it should be equal to that for a uniform line constructed so as to have the same dimensions everywhere as the non-uniform line has at the point in question.

[2] The word nominal as used here has the same meaning as in "nominal iterative impedance" as used by K. S. Johnson in "transmission circuits for telephone communication" (Van Nostrand 1925).

[3] S. A. Schelkunoff, "The Impedance Concept and its Application to Problems of Reflection, Refraction, Shielding and Power Absorption," *Bell System Technical Journal, 17*, 17–48, January, 1938.

the steady state condition is composed. At each point an exponential line has two characteristic impedances which are different and depend upon the frequency as well as the position along the line.

Because of the change of impedance level, the propagation constants for the voltage and current differ, so that it is convenient to consider the transfer constant[4] which may be defined as half the sum of the voltage and current propagation constants.

[4] Compare with the definition of "image transfer constant" as given by K. S. Johnson in "Transmission Circuits for Telephone Communication."

The Bridge Stabilized Oscillator *

By L. A. MEACHAM

A new type of constant frequency oscillator of very high stability is presented. The frequency controlling resonant element is used as one arm of a Wheatstone resistance bridge. Kept in balance automatically by a thermally controlled arm, this bridge provides constancy of output amplitude, purity of wave form, and stabilization against fluctuations in power supply or changes in circuit elements. A simple one-tube circuit has operated consistently with no short-time frequency variations greater than ± 2 parts in 10^8. Convenient means are provided for making precision adjustments over a narrow range of frequencies to compensate for long-time aging effects.

Description of the circuit is followed by a brief linear analysis and an account of experimental results. Operating records are given for a 100 kc. oscillator.

Introduction

THE problem of improving the stability of constant frequency oscillators may be divided conveniently into two parts, one relating to the frequency controlling resonant element or circuit, and the other to the means for supplying energy to sustain oscillations. The ideal control element would be a high-Q electrical resonant circuit, or a mechanical resonator such as a tuning fork or crystal, whose properties were exactly constant, unaffected by atmospheric conditions, jar, amplitude of oscillation, age, or any other possible parameter. The ideal driving circuit would take full advantage of the resonator's constancy by causing it to oscillate at a stable amplitude and at a frequency determined completely by the resonator itself, regardless of power supply variations, aging of vacuum tubes or other circuit elements, or the changing of any other operating condition.

This paper, concerning itself principally with the second part of the problem, describes an oscillator circuit which attains a very close approximation to the latter objective. The "Bridge Stabilized Oscillator" provides both frequency and amplitude stabilization, and as it operates with no tube overloading, it has the added merit of delivering a very pure sinusoidal output.

Oscillator Circuit

The bridge stabilized oscillator circuit, shown schematically in Fig. 1, consists of an amplifier and a Wheatstone bridge. The amplifier out-

* Presented at Thirteenth Annual Convention of Institute of Radio Engineers, New York City, June 16, 1938. Published in *Proc. I. R. E.*, October 1938.

put is impressed across one of the diagonals of the bridge, and the unbalance potential, appearing across the conjugate diagonal, is applied to the amplifier input terminals. One of the four bridge arms, R_1, is a thermally controlled resistance; two others, R_2 and R_3, are fixed resistances, and the fourth, $Z_4 = R_4 + jX_4$, is the frequency-controlling resonant element.

In this discussion Z_4 is assumed to represent a crystal suitable for operation at its low-impedance or series resonance. A coil and condenser in series could be substituted, and even a parallel-resonant control element might be used by exchanging its position in the bridge with R_2 or R_3. Operating a crystal at series resonance has the advantage of minimizing effects of stray capacitance.

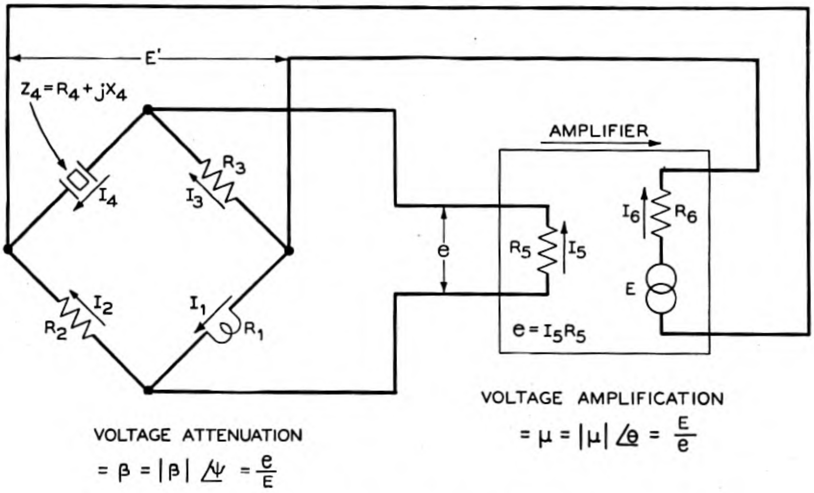


Fig. 1—Schematic circuit diagram of bridge stabilized oscillator.

The bridge is kept as nearly in exact balance as possible. Assuming that R_1, R_2 and R_3 are pure resistances, we may write for exact reactive balance,

$$X_4 = 0,$$

and for exact resistive balance,

$$\frac{R_1}{R_2} = \frac{R_3}{R_4}.$$

In order that the circuit may oscillate, a slight unbalance is required. Accordingly R_1 must be given a value slightly smaller than $(R_2R_3)/R_4$, so that the attenuation through the bridge is just equal to the gain of the amplifier.

It is evident that if all the bridge arms had fixed values of resistance, the attenuation of the bridge would be very critical with slight changes in any arm. This would obviously be undesirable, for the circuit would either fail to oscillate, or else build up in amplitude until tube overloading occurred. The thermally controlled resistance R_1 eliminates this difficulty. This arm has a large positive temperature coefficient of resistance, and is so designed that the portion of the amplifier output which reaches it in the bridge circuit is great enough to raise its temperature and increase its resistance materially. A small tungsten-filament lamp of low wattage rating has been found suitable. It functions as follows:

When battery is first applied to the amplifier, the lamp R_1 is cold and its resistance is considerably smaller than the balance value. Thus the attenuation of the bridge is relatively small, and oscillation builds up rapidly. As the lamp filament warms, its resistance approaches the value for which the loss through the bridge equals the gain of the amplifier. If for some reason R_1 acquires too large a resistance, the unbalance potential e becomes too small or possibly even inverted in phase, so that the amplitude decreases until the proper equilibrium is reached.

This automatic adjustment stabilizes the amplitude, for the amount of power needed to give R_1 a value closely approaching $(R_2R_3)/R_4$ is always very nearly the same. A change in the amplifier gain would cause a readjustment of the bridge balance, but the resulting variation in R_1 or in the amplifier output would be extremely small. The operating temperature of the lamp filament is made high enough so that variations in the ambient temperature do not affect the adjustment appreciably.

No overloading occurs in the amplifier, which operates on a strictly Class A basis, nor is any non-linearity necessary in the system other than the thermal non-linearity of R_1. As the lamp resistance does not vary appreciably during a high-frequency cycle, it is not a source of harmonics (or of their intermodulation, which Llewellyn [1] has shown to be one of the factors contributing to frequency instability).

In contrast to the lamp, an ordinary non-linear resistance, of copper oxide for example, would not be suitable for R_1. A resistance of the thermally-controlled type having a negative temperature coefficient

[1] "Constant Frequency Oscillators," F. B. Llewellyn, *Proc. I. R. E.*, December 1931.

could be used by merely exchanging its position in the bridge with R_2 or R_3.

The frequency control exerted by the crystal depends upon the fact that the phase shift of the amplifier must be equal and opposite to that through the bridge. In the notation of Black,[2] applied to the circuit of Fig. 1,

$$\mu = \frac{E}{e} = |\mu| \underline{|\theta},$$

and

$$\beta = \frac{e}{E} = |\beta| \underline{|\psi}.$$

The condition for oscillation is

$$\mu\beta = 1 \underline{|0},$$

which implies that $|\mu\beta| = 1$ and $\theta = -\psi$.

The vector diagrams of Fig. 2 illustrate the frequency-stabilizing action of the bridge by showing the voltage relations therein for two values of amplifier phase shift, θ. When θ is zero, as in diagram A, the unbalance vector e is in phase with the generated voltage E applied to the bridge input, and thus all the vectors shown are parallel. They are displaced vertically from each other merely to clarify the drawing. The crystal is here constrained to operate at exact resonance.

In diagram B, the amplifier is assumed to have changed its phase for some reason by an amount far in excess of what would be anticipated in practice, θ here having a value of $+45$ degrees. The important point to be observed is that the ratio of θ to the resulting change in the phase angle ϕ of the crystal impedance Z_4 is very large. That is, the crystal is still operating close to resonance in spite of the exaggerated change in the driving circuit. If the gain of the amplifier were greater, the action of the thermally controlled resistance would keep the amplifier output vector E practically the same in length, making the unbalance vector e correspondingly shorter. The angle ϕ would therefore have to be more acute for the same value of θ, and it follows that with increased gain the crystal is held closer to true resonance and the stability is improved.

When θ equals zero, changes in $|\mu|$ do not affect the crystal operating phase, but for any other small value of θ, gain variations cause slight readjustments of the angles between vectors. The amplifier should accordingly be designed for zero phase shift, and also, of course, should have as much phase stability as possible.

[2] "Stabilized Feedback Amplifiers," H. S. Black, *Bell System Technical Journal*, January 1934.

In this discussion the input and output impedances of the amplifier, R_5 and R_6, are assumed to be constant pure resistances. Actually, changes in the tube parameters or in certain circuit elements are likely to affect both the magnitude and the phase of these impedances. It may be shown, however, that such changes have the same effect upon the bridge and upon the frequency as do changes of about the same percentage in $|\mu|$ or θ; therefore all variations in the driving circuit external to the bridge may be assumed for convenience to be represented by variations in its gain and phase.

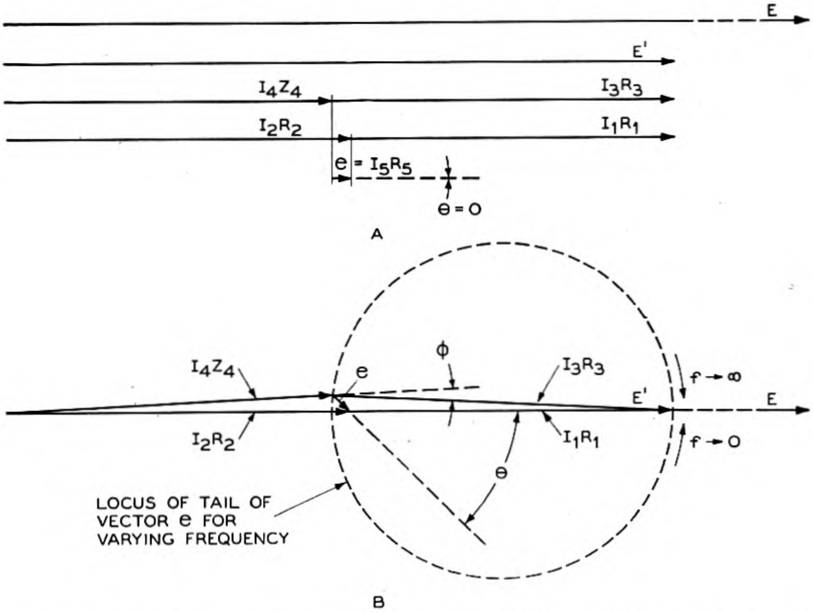


Fig. 2—Vector diagrams illustrating operation of bridge oscillator, with simplifying assumptions that R_5 is large and that E and E' are strictly in phase.

A—At resonance
$Z_4 = R_4 + j0$
$\theta = 0$
$R_1 < R_2 = R_3 = R_4$
B—Above resonance
$Z_4 = R_4 + jX_4$
X_4 Inductive
$\theta = +45°$
$R_1 < R_2 = R_3 = R_4 \ll R_6$

This leniency with regard to R_5 and R_6 does not apply to the other bridge resistances, however. R_1, R_2 and R_3 are directly responsible for the crystal's operating phase and amplitude; they should be made as stable and as free from stray reactance as possible.

The effect of the bridge upon harmonics of the oscillation frequency is of interest. Harmonics, being far from the resonant frequency of the crystal, are passed through the bridge with little attenuation but with a phase reversal approximating 180 degrees, as illustrated by the dotted locus in Fig. 2. Thus if the amplifier were designed to cover a band broad enough to include one or more harmonics and if care were taken to avoid singing at some unwanted frequency, a considerable amount of negative feedback could be applied to the suppression of the harmonics in question.

Circuit Analysis

In the following section, expressions are derived for the frequency of oscillation in terms of the gain and phase shift of the amplifier, the Q of the crystal, and values of the bridge resistances. It is assumed that the latter are constant and non-reactive, and therefore, as explained previously, that all sources of frequency fluctuations apart from changes in the crystal itself appear as variations in $|\mu|$ or θ. Because the bridge oscillator does not rely upon non-linearity in the ordinary sense to limit its amplitude, the analysis can be based reasonably on simple linear theory.

In the near vicinity of series resonance the crystal may be represented accurately by a resistance R_4, inductance L and capacitance C, connected in series. The reactive component of the crystal's impedance is accordingly

$$X_4 = \omega L - \frac{1}{\omega C} = \frac{\omega^2 LC - 1}{\omega C}. \tag{1}$$

Solving for the frequency,

$$\begin{aligned}\omega &= \frac{X_4}{2L} + \sqrt{\left(\frac{X_4}{2L}\right)^2 + \frac{1}{LC}} \\ &= \frac{1}{\sqrt{LC}}\left[\frac{X_4}{2}\sqrt{\frac{C}{L}} + \sqrt{1 + \left(\frac{X_4}{2}\sqrt{\frac{C}{L}}\right)^2}\right] \\ &= \frac{1}{\sqrt{LC}}\left[1 + \frac{X_4}{2}\sqrt{\frac{C}{L}} + \frac{1}{2}\left(\frac{X_4}{2}\sqrt{\frac{C}{L}}\right)^2\right. \\ &\qquad\qquad \left. - \frac{1}{2}\cdot\frac{1}{4}\left(\frac{X_4}{2}\sqrt{\frac{C}{L}}\right)^4 + \cdots\right]. \end{aligned} \tag{2}$$

Near series resonance, $(X_4/2)\sqrt{(C/L)} << 1$. We therefore disregard powers higher than the first in the series expansion above and obtain the close approximation,

$$\omega \doteq \frac{1}{\sqrt{LC}}\left[1 + \frac{X_4}{2}\sqrt{\frac{C}{L}}\right]. \tag{3}$$

The frequency deviation from resonance, expressed as a fraction of the resonant frequency f_0, is thus

$$\frac{f - f_0}{f_0} = \frac{\omega - \omega_0}{\omega_0} \doteq \frac{X_4}{2}\sqrt{\frac{C}{L}}, \tag{4}$$

and in the region of interest, where ωL and $1/\omega C$ are approximately equal,

$$\frac{f - f_0}{f_0} \doteq \frac{X_4}{2\omega L} = \frac{X_4}{2QR_4}. \tag{5}$$

Considering now the bridge circuit, and applying well-known equations,[3] we obtain

$$\beta = \frac{I_5R_5}{E} = \frac{AR_4 - jBX_4}{MR_4 + jNX_4}, \tag{6}$$

in which

$$\left.\begin{aligned} A &= R_5(R_2R_3 - R_1R_4),\\ B &= R_1R_4R_5,\\ M &= (R_1 + R_2)(R_3R_4 + R_5R_6) + (R_3 + R_4)(R_1R_2 + R_5R_6)\\ &\quad + (R_5 + R_6)(R_1R_4 + R_2R_3) + R_5(R_1R_3 + R_2R_4)\\ &\quad + R_6(R_1R_2 + R_3R_4),\\ &\text{and}\\ N &= R_4(R_1 + R_3 + R_5)(R_2 + R_6) + R_1R_4(R_3 + R_5). \end{aligned}\right\} \tag{7}$$

The condition for oscillation, as mentioned previously, is $\mu\beta = 1\underline{|0}$. Putting $\mu = \mu_1 + j\mu_2$, we may write

$$(\mu_1 + j\mu_2)\left(\frac{AR_4 - jBX_4}{MR_4 + jNX_4}\right) = 1, \tag{8}$$

which gives the pair of equations

$$\mu_1AR_4 + \mu_2BX_4 - MR_4 = 0 \tag{9}$$

and

$$\mu_2AR_4 - (\mu_1B + N)X_4 = 0. \tag{10}$$

For the special case in which the amplifier phase shift is zero ($\mu_2 = 0$), these become

$$\mu_1 = \frac{M}{A} = |\mu| \tag{11}$$

and

$$X_4 = 0. \tag{12}$$

[3] "Transmission Circuits for Telephonic Communication," K. S. Johnson, pp. 284–5. D. Van Nostrand Company.

The latter equation indicates that the frequency is then independent of changes in any of the circuit parameters except the crystal, which must operate exactly at resonance.

If the phase of μ differs only slightly from zero, so that μ_2 is very small, then it may be inferred from continuity considerations that the frequency is still very nearly independent of all circuit parameters, except of course variations in θ, the phase of μ. When θ is limited to values for which $\mu_2 BX_4 << \mu_1 AR_4$, (11) still applies closely. Substitution into (10) gives

$$X_4 \doteq \frac{MR_4}{B\mu_1 + N} \cdot \frac{\mu_2}{\mu_1} \doteq \frac{MR_4\theta}{B|\mu| + N}, \tag{13}$$

and finally from (5) and (13) we obtain the frequency deviation in the form

$$\frac{f - f_0}{f_0} \doteq \frac{M\theta}{2Q(B|\mu| + N)}. \tag{14}$$

As noted above, this expression applies accurately only when θ is small, as it should be in a well designed bridge oscillator.

The effect of variations in the amplifier may be examined by differentiating (14). For changes in θ only,

$$\left.\frac{df}{f_0}\right]_\theta \doteq \frac{M}{2Q(B|\mu| + N)}\, d\theta, \tag{15}$$

and for those of $|\mu|$,

$$\left.\frac{df}{f_0}\right]_{|\mu|} \doteq - \frac{BM\theta}{2Q(B|\mu| + N)^2}\, d|\mu|. \tag{16}$$

Equations (15) and (16) have been found to be closely in accord with experiment, although the differentiation is not rigorously allowable (B, M and N being only approximately constant).

In the special case where all the fixed bridge resistances (R_2 to R_6 inclusive) are equal, and $|\mu|$ is large enough so that R_1 has nearly the same value, (14), (15) and (16) reduce to the following:

$$\frac{f - f_0}{f_0} \doteq \frac{8\theta}{Q(|\mu| + 8)}, \tag{17}$$

$$\left.\frac{df}{f_0}\right]_\theta \doteq \frac{8}{Q(|\mu| + 8)}\, d\theta, \tag{18}$$

$$\left.\frac{df}{f_0}\right]_{|\mu|} \doteq - \frac{8\theta}{Q(|\mu| + 8)^2}\, d|\mu|. \tag{19}$$

These expressions show, as did the vector diagrams, that for optimum stability the amplifier phase shift should be made approximately zero, the crystal should have as large a value of Q (as low a decrement) as possible, and the amplifier should have high gain. Linearity in the amplifier is also desirable, to minimize the modulation effects described by Llewellyn.[1] When present, these effects appear as variations in $|\mu|$ and θ.

One of the significant differences between the bridge oscillator and other oscillator circuits is the fact that its frequency stability is roughly proportional to $|\mu|$. This relationship holds at least for amounts of gain that can be dealt with conveniently. Although increased gain is generally accompanied by larger variations in phase, the two are not necessarily proportional. For example, if greater stability were required for some precision application than could be achieved with a single-tube bridge oscillator, and if the constancy of the crystal itself warranted further circuit stabilization, it could be obtained by adding another stage. The phase fluctuations in the amplifier might possibly be doubled, but the value of $|\mu|$ would be multiplied by the amplification of the added tube, giving an overall improvement.

To illustrate the high order of stability provided by a bridge oscillator, let us consider a model composed of a single-tube amplifier and a bridge in which all the fixed resistances are made equal to that of the crystal. We will assume the crystal to have a reasonably high[4] Q of 100,000. The amplifier phase, let us say, is normally zero, but may possibly vary ± 0.1 radian (± 6 degrees), and the value of $|\mu|$, nominally 400, may change ± 10 per cent. From (18) and (19) we find

$$\left.\frac{\Delta f}{f_0}\right]_\theta = \pm \frac{(8)(0.1)}{(100{,}000)(360 + 8)} = \pm\, 2.17 \times 10^{-8},$$

and (when θ has its maximum value of 0.1 radian)

$$\left.\frac{\Delta f}{f_0}\right]_{|\mu|} = \pm \frac{(8)(0.1)(40)}{(100{,}000)(360 + 8)^2} = \pm\, 2.36 \times 10^{-9}.$$

This example represents the degree of stabilization against circuit fluctuations that can be obtained with a simple form of the oscillator operating under poorly controlled conditions. By stabilizing the power supply and other factors affecting $|\mu|$ and θ, and by increasing the gain, the frequency variations arising in the driving circuit may be made negligible compared to the variations found at present in the properties even of the best mounted crystals.

[4] For crystals in vacuum, values of Q as great as 300,000 have been obtained.

Experiment

The circuit diagram of an experimental bridge stabilized oscillator is shown in Fig. 3, and its photograph in Fig. 4. The amplifier unit consists of a single high-mu tube V_1 with tuned input and output transformers T_1 and T_2 and the usual power supply and biasing arrangements. The crystal, mounted in the cylindrical container at the left end of the panel, is one having a very low temperature coefficient of frequency at ordinary ambient temperatures. In Fig. 4 it is shown without provisions for temperature control. A high Q is obtained by clamping the crystal firmly at the center of its aluminum-coated major faces between small metal electrodes ground to fit, and by evacuating the container.

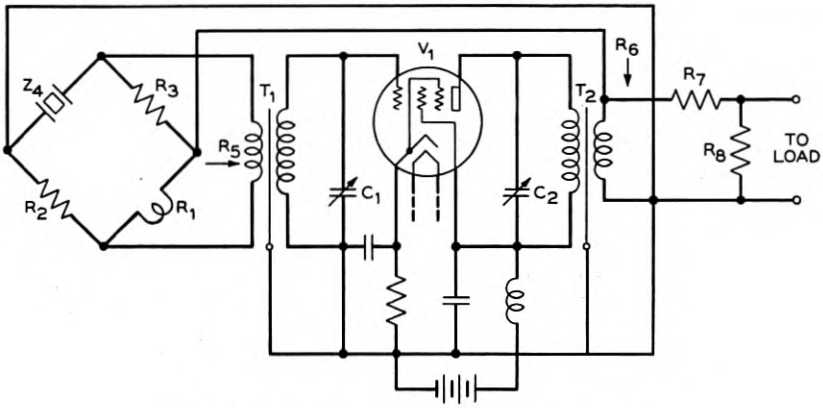


Fig. 3—Circuit of experimental bridge oscillator.

Some of the circuit parameters are listed below:

R_1 = Tungsten-filament lamp,
R_2 = 100 ohms,
R_3 = 150 ohms,
Z_4 = 100 kc. crystal.
Characteristics at resonance:
R_4 = 114 ohms,
$X_L = X_C$ = 11,900,000 ohms,
Q = 104,000,
$R_5 = R_6$ = 150 ohms (approx.),
R_7 = 500 ohms,
R_8 = 200 ohms,
$|\mu|$ = 422 (52.5 *db* voltage gain from e to E).

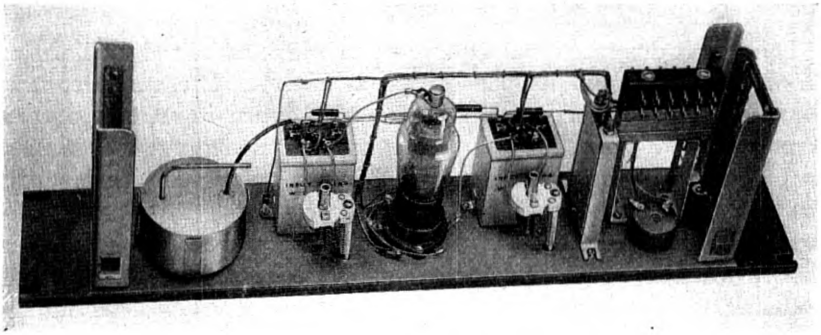


Fig. 4—Experimental bridge stabilized oscillator without provision for temperature control.

Figure 5 shows the resistance of the lamp R_1 plotted against the power dissipated in its filament. The large rise in resistance for small amounts of power is due to the effective thermal insulation provided by the vacuum surrounding the filament and to low heat loss by radiation. The lamp operates at temperatures below its glow point, assuring an extremely long life for the filament.

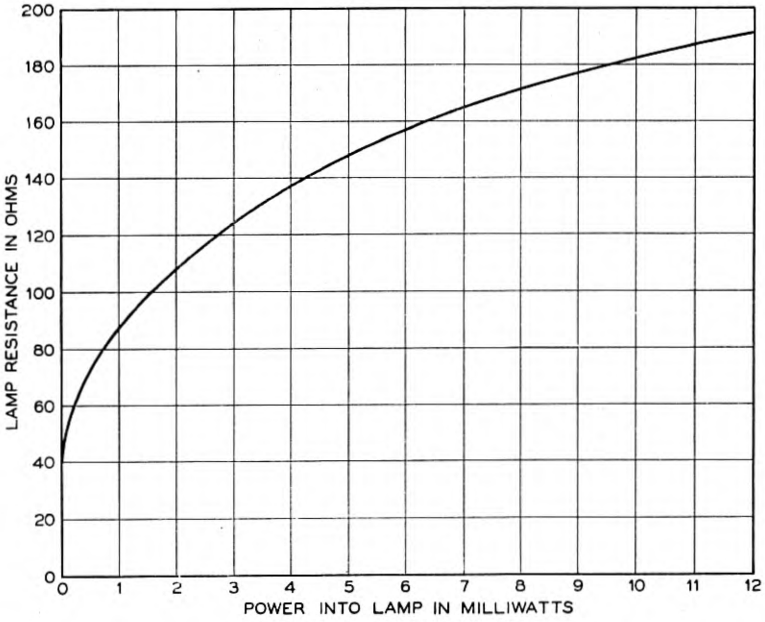


Fig. 5—Characteristic of lamp used for R_1.

The particular value assumed by R_1 in the circuit of Fig. 3 is approximately $(R_2R_3)/R_4 = [(100)(150)]/114 = 131.6$ ohms, and hence from Fig. 5 it follows that the power supplied to the lamp is about 3.7 milliwatts. The r.m.s. voltage across the lamp is computed to be 0.70 volt, and across the entire bridge, 1.23 volt. The power supplied to a load of 150 ohms through the pad composed of R_7 and R_8 is accordingly 0.22 milliwatt, or 6.6 *db* below 1 milliwatt, which is in agreement with measurements shown in Figs. 8 and 9, described below. These quantities are given to illustrate the fact that currents and voltages in this type of oscillator may be calculated readily from the values of the circuit elements, and without reference to the power supply voltages or the tube characteristics except to assume that they give the amplifier sufficient gain to operate the bridge near balance, and that tube overloading does not occur at the operating level.

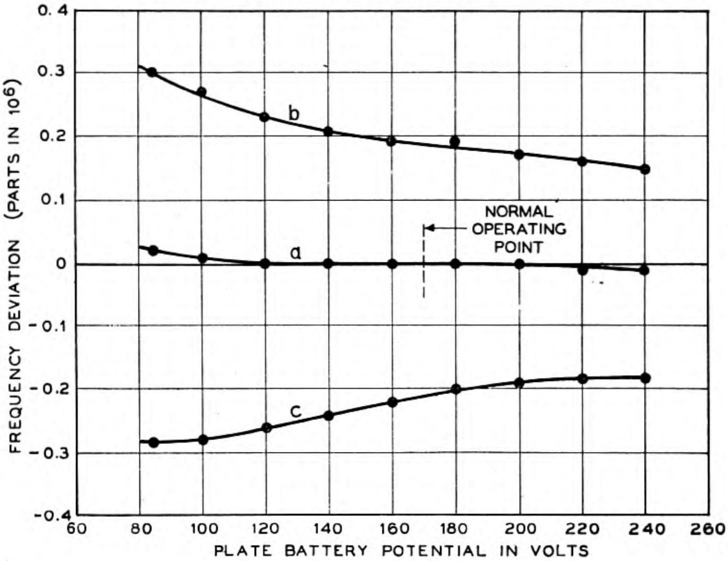


Fig. 6—Oscillator frequency vs. plate battery potential.

a—C_1 and C_2 tuned for maximum amplifier gain.
b—C_1 and C_2 decreased 5%.
c—C_1 and C_2 increased 5%.

Experimental performance curves for the circuit of Fig. 3 are presented in Figs. 6 to 11 inclusive. Figure 6 shows frequency deviation plotted against plate battery voltage for several settings of the grid and plate tuning condensers. For curve *a* the amplifier was adjusted at maximum gain, corresponding approximately to zero phase shift as well. Here the frequency varied not more than one part in one hundred million for a voltage range from 120 to 240 volts. Curve *b* was taken with the two tuning capacitances C_1 and C_2 decreased 5 per cent from their optimum settings, and curve *c* with both capacitances increased 5 per cent. These detunings introduced phase shifts of about ± 40° (± 0.70 radian), decreased $|\mu|$ by 0.8 *db* and changed the frequency, as shown in Fig. 6, approximately ± 2 parts in ten million. Although the analysis should not be expected to apply accurately for such large phase shifts, calculation of the frequency deviations by means of (18) gives ± 1.4 parts in ten million—in fair agreement with the experimental results. As might be expected, curves *b* and *c* show somewhat less stability with battery voltage changes than does curve *a*.

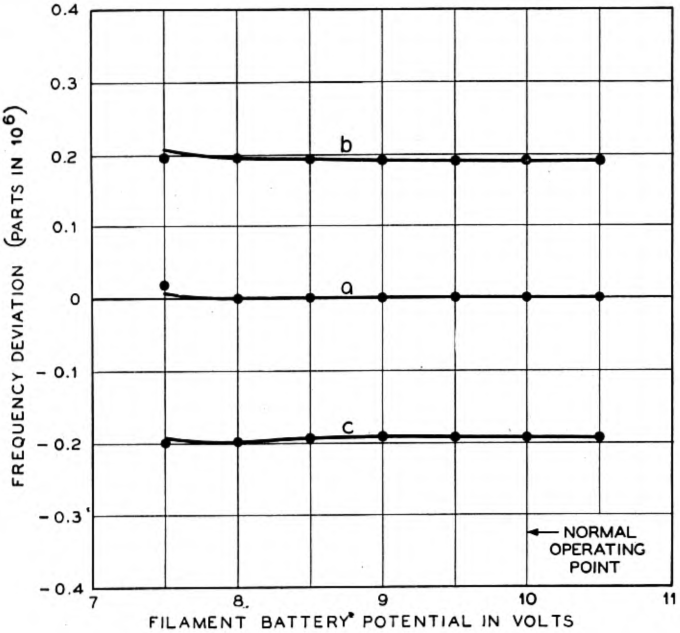


Fig. 7—Oscillator frequency vs. filament battery potential.
a—C_1 and C_2 tuned for maximum amplifier gain.
b—C_1 and C_2 decreased 5%.
c—C_1 and C_2 increased 5%.

Figure 7 presents a similar set of curves for variations of filament voltage. Here, for the "maximum-gain" tuning adjustment, a drop from 10 volts, the normal value, to 8 volts caused less than one part in one hundred million change of frequency, as shown in curve *a*.

In Fig. 8, the gain of the amplifier and the output level of the oscillator are plotted against plate battery voltage, while in Fig. 9 the same quantities are related to filament potential. These curves show that although power supply variations change the amplifier gain, they have but slight effect upon the amplitude of oscillation. This stabilization is produced, as explained heretofore, by the action of the lamp.

The oscillator was designed to work into a load of 150 ohms, its output impedance approximately matching this value. It might be expected that variations in the magnitude or phase angle of the load would affect the frequency materially even though a certain amount of isolation is provided by R_7 and R_8. However, measurements made with (1) a series of load impedances having a constant absolute magnitude of 150 ohms but with phase angles varying between $-90°$ and $+90°$ and (2) a series of resistive loads varying between 30 ohms and open circuit, showed less than one part in a hundred million frequency variation. Graphs of these results have not been included, since they practically coincide with the axis of zero frequency deviation.

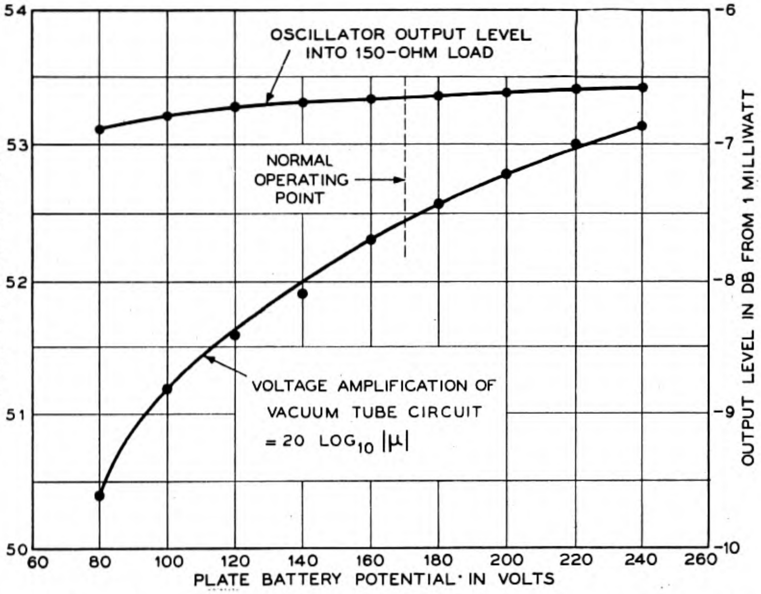


Fig. 8—Amplifier gain and oscillator output level vs. plate battery potential.

The tuned transformers T_1 and T_2 in this experimental model precluded the suppression of harmonics by negative feedback, $|\mu|$ being small at the harmonic frequencies. The tuning itself provided suppression, however, so that the measured levels of the second and third harmonics in the output current were respectively 67 db and 80 db below that of the fundamental. This purity of wave form is of course largely dependent upon the absence of overloading.

To correct any small initial frequency error of the crystal and to allow for subsequent aging, a small reactance connected directly in series with the crystal provides a convenient means of adjusting the frequency as precisely as it is known. This added reactance may be considered as modifying either of the reactances in the equivalent series resonant circuit of the crystal. Figure 10 shows that for a small range of frequencies the change introduced in this manner is accurately proportional to the added reactance. Series inductance, of course, lowers the frequency, while series capacitance raises it. The stability requirement imposed on the adjusting reactance is only moderate, for its total effect upon the frequency should not be more than a few parts in a million.

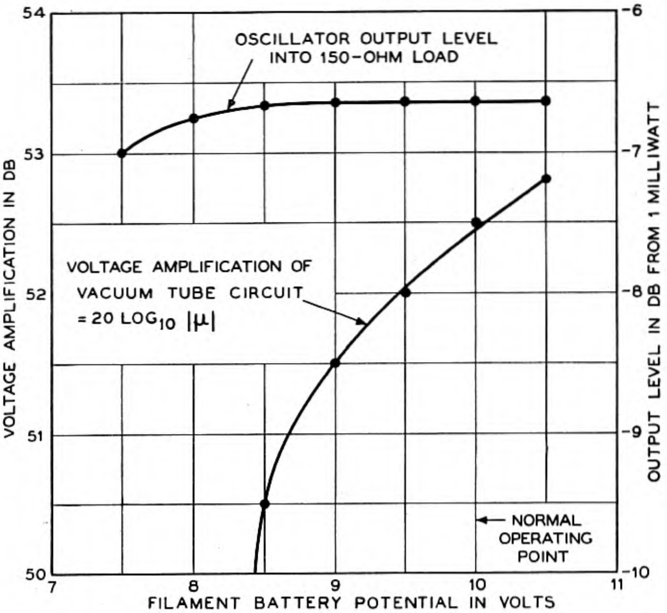


Fig. 9—Amplifier gain and oscillator output level vs. filament battery potential.

The frequency measurements here presented were obtained using apparatus similar in principle to the frequency comparison equipment of the National Bureau of Standards.[5] Frequency differences

[5] "Harmonic Method of Intercomparing the Oscillators of the National Standard of Radio Frequency," E. G. Lapham, *Journal of Research of the National Bureau of Standards*, October 1936, p. 491.

between the oscillator under test and a reference bridge oscillator were read upon a linear scale calibrated directly in terms of frequency deviation. Full scale could be made one part in 10^4, 10^5, 10^6 or 10^7 by means of a simple switching operation. For most of the measurements in this paper the full-scale reading was one part in a million, and the resolution about ± 0.005 part in a million.

By using a recording meter with this measuring set, continuous frequency comparisons between two independent bridge oscillators were obtained over a period of several months. Figure 11 is a photograph of a section of this record. It shows the short-time variations of both oscillators plus a small amount of scattering caused by the measuring equipment itself. The crystals were temperature-controlled in separate ovens, and the power was supplied from separate sets of laboratory batteries controlled to about ± 2 per cent in voltage. Shielding was ample to avoid any tendency to lock in step.

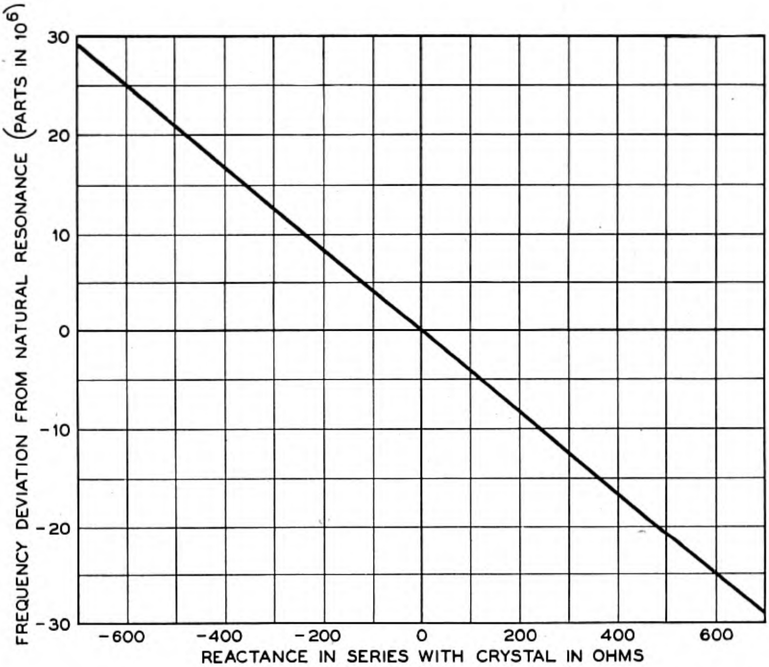


Fig. 10—Frequency of oscillator vs. adjusting reactance.

In addition to these small short-time variations, the oscillators exhibited a very slow upward drift in frequency, attributed to aging of the mounted crystals. This aging decreased in a regular manner with time, the mean drift of one of the crystals being less than one part in ten million per month after three months of continuous operation,

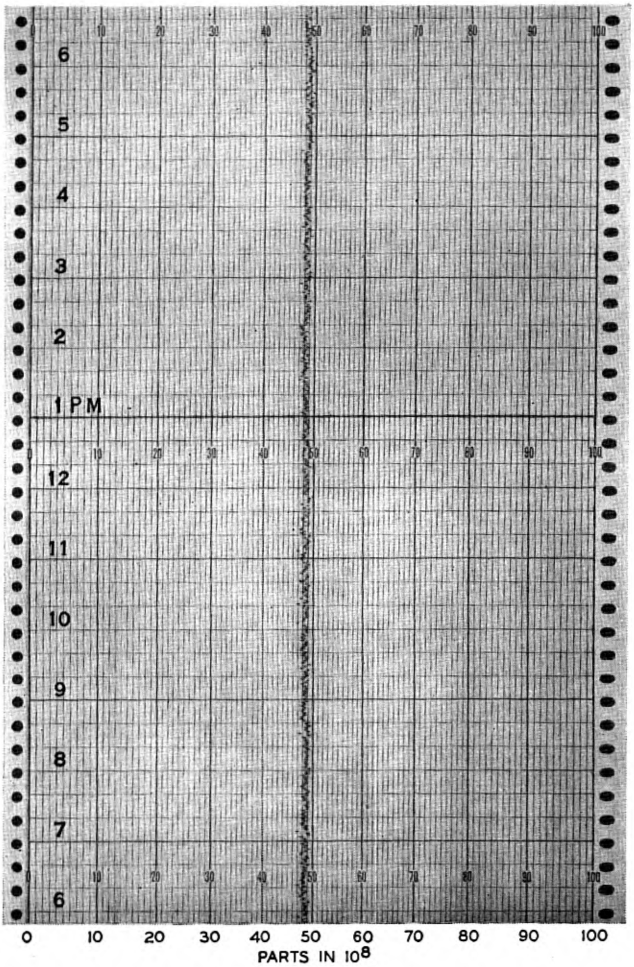


Fig. 11—Record of frequency comparison between two bridge stabilized oscillators. Full scale one part in a million. Variations less than ± 2 parts in one hundred million.

and about a third of this amount after seven months. In most applications, gradual frequency drift is not objectionable even though the required accuracy is very high, for readjustment is merely a matter of setting a calibrated dial.

Application

The bridge stabilized oscillator promises to become a useful tool in many commercial fields as well as in certain purely scientific problems such as time determination and physical and astronomical measurement. It may be used either to increase the frequency precision in applications where operating conditions are accurately controlled, or else to make such control unnecessary, affording high stability in spite of unfavorable conditions.

An interesting application in the field of geophysics has already been made in the form of a "Crystal Chronometer." This chronometer consists of a single-tube bridge oscillator, a frequency dividing circuit, and a synchronous timing motor. It was recently loaned by Bell Telephone Laboratories to the American Geophysical Union and was used with the Meinesz gravity-measuring equipment on a submarine gravity-survey expedition in the West Indies. Although operating under somewhat adverse conditions of power supply, temperature, and vibration, it was reported [6] to be more stable than any timing device previously available, errors in the gravity measurements introduced by the chronometer being negligibly small.

[6] "Gravity Measurement on the U. S. S. Barracuda," M. Ewing, and "Crystal Chronometer Time in Gravity Surveys," A. J. Hoskinson; pp. 66 and 77 resp., *Transactions of the American Geophysical Union*, Part I, 1937.

Effect of Space Charge and Transit Time on the Shot Noise in Diodes

By A. J. RACK

The theoretical analysis of the effect of space charge upon the "shot noise" in a planar diode shows that for practically all operating conditions, the tube noise is equivalent to the thermal resistance noise of the plate resistance at 0.644 times the cathode temperature. Noise in diodes of other than planar shapes is discussed and it is concluded that the same relation holds. It is shown that transit time produces the same high frequency modification for both the thermal and shot tube noise, and that the tube noise is decreased by transit time.

IN the study of noise in vacuum tubes, the effect of the space charge upon the shot noise has been a subject of considerable interest and practical importance. Several papers have been written in which it is shown that the shot noise is decreased by the space charge, and that the tube noise in a diode with space charge is equivalent to the thermal resistance noise of the plate impedance at a temperature slightly greater than half of that of the cathode.[1, 2, 3, 4] The most comprehensive analysis was made by Schottky and Spenke. These authors, employing a different method from the one here presented, have obtained the same general conclusions given in this paper, although they prefer to express the result in the form of a modified shot-noise equation, whereas for reasons developed below, the writer prefers the thermal form. The theoretical analysis and discussion presented here was undertaken to show in more detail the extent of the range of the operating condition for which the thermal resistance equivalent of tube noise is valid and to study the effect of transit time upon both the shot and thermal tube noise.

For convenience, the paper is divided into three parts. In the first section is given an exact mathematical treatment of the tube noise at low frequencies in a parallel plane diode for any degree of space charge. A discussion of the final tube noise equation obtained through this analysis, and the extension of these results for the planar diode to any other shape diode is given in Part II, where the presentation is such that the section may be read independently of the theoretical analysis in Part I. Through several approximations, Part III treats the effect of transit time upon tube noise in the planar diode.

Part I—General Low Frequency Analysis

In the development of the general equations for the direct current in vacuum tubes with space charge, account has been taken of the fact that the electrons are emitted from the cathode with Maxwellian velocity distribution. This fact has been verified experimentally by Germer,[5] and the resulting equations for the relation between current and voltage have been derived and investigated by Fry,[6] Langmuir,[7] and others. In the extension of this analysis to tube noise, it is only necessary to assume that the number of electrons emitted with any velocity does not remain constant, but fluctuates with time according to the well-known laws of probability. In the analysis on this basis, the frequencies involved will be considered to be sufficiently low so that any transit time effect is negligible.

Below is given a list of the definitions of various symbols to be used in the tube noise study of a parallel plane diode. The practical system of units is employed throughout.

$n(u_c)du_c$ = instantaneous rate of emission per unit area of the cathode of electrons with initial velocities between u_c and $u_c + du_c$ in the x-direction, regardless of the velocity components in the other directions,

= $n_0(u_c)du_c + \delta(u_c)du_c$,

$n_0(u_c)du_c$ = average rate of emission of electrons with x-directed velocities between u_c and $u_c + du_c$,

$\delta(u_c)du_c$ = instantaneous deviation from average rate of emission,

I = instantaneous anode current per unit area,

V = instantaneous potential with respect to cathode of a plane at a distance x from the cathode,

V' = instantaneous potential with respect to cathode of the potential minimum,

u = instantaneous velocity at x-plane of electrons which had an initial x-directed velocity of u_c at the cathode,

x' = instantaneous position of potential minimum,

e = charge on electron = -1.59×10^{-19} coulombs,

m = mass of electron = 9.01×10^{-28} grams,

h = ratio of dyne cms. to joules = 10^{-7},

ϵ = permittivity of a vacuum in practical units = 8.85×10^{-14} farads/cm.,

k = Boltzmann's gas constant = 1.372×10^{-23} watts/degree Kelvin,

N = average total number of electrons emitted per second per unit area from the cathode,

T = absolute temperature of the cathode.

In the following analysis, it is assumed that the electrodes of the planar diode are infinite in extent, and that the electron emission is random, so that the equipotential surfaces are parallel planes perpendicular to the x-axis.

The potential distribution in such a planar diode operating with space charge is shown in Fig. 1. The origin of coordinates is taken at the cathode, and the potential minimum formed by space charge occurs at a distance $x = x'$ from the cathode. The subscript α will be used to denote the space between cathode and potential minimum while β applies similarly to the space between minimum and anode. Of all the electrons emitted from the cathode only those whose x-velocity exceeds the value u_c' corresponding to the potential minimum can penetrate the barrier and proceed to the anode. Electrons with lesser values of initial velocity will come to rest at a point in the α-region where the potential corresponds to their initial velocity and will then return to the cathode. The anode current density is thus given by

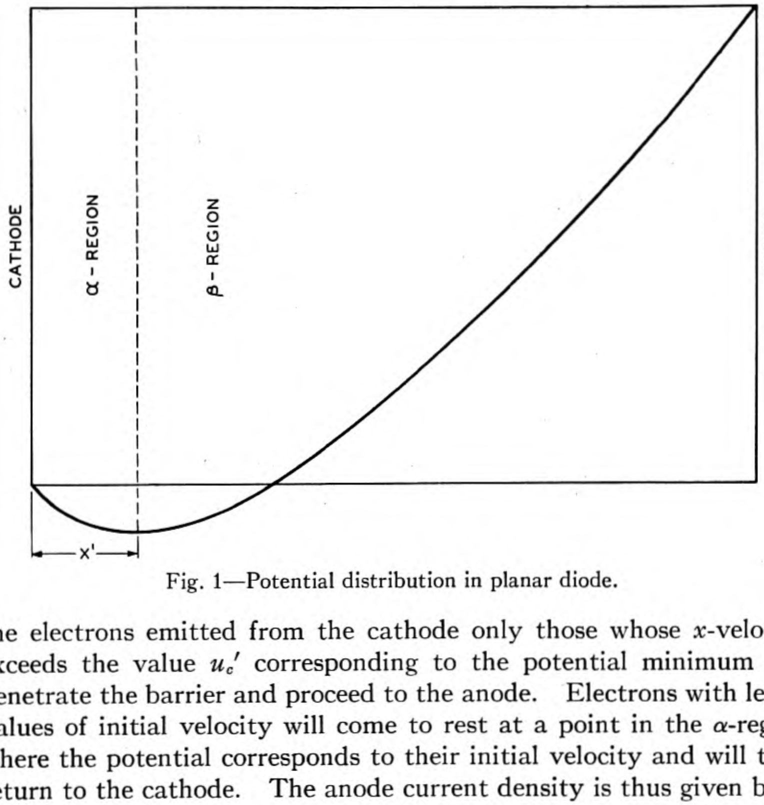


Fig. 1—Potential distribution in planar diode.

$$I = e\int_{u_c'}^{\infty} n(u_c)du_c, \tag{1}$$

while the relation between velocity u and potential V at a given value of x is

$$u^2 = u_c^2 - \frac{2e}{hm}V. \tag{2}$$

A third fundamental relation is Poisson's equation which becomes in the parallel plane case under consideration

$$\epsilon \frac{d^2 V}{dx^2} = - \rho. \tag{3}$$

In the α-region the total charge density is made up of three classes of electrons, namely

1. Those destined to pass the potential minimum and arrive at the anode.
2. Those moving away from the cathode but which will not travel as far as the minimum point.
3. Those returning to the cathode.

Corresponding to each class of electrons, there is an associated current, ρu, so that each of the three densities ρ_1, ρ_2 or ρ_3, may be expressed by a relation of the form,

$$\rho_n = \frac{I_n}{u}. \tag{4}$$

When it is remembered that the potential and velocity at a given value of x are uniquely related through (2), then it is easy to see that the total density for a given plane in the α-region is given by

$$\rho_\alpha = e \int_{u_c'}^{\infty} \frac{n(u_c)}{u} du_c + 2e \int_{v}^{u_c'} \frac{n(u_c)}{u} du_c, \tag{5}$$

where the first term represents the contribution of electrons in class 1 above, while the second term represents the contribution of electrons in classes 2 and 3. The contribution of class 3 is equal to that of class 2. The lower integration limit v of the second term of (5) represents the initial velocity of an electron which would just arrive at the value of x under consideration before coming to rest and starting back toward the cathode and the limit u_c' in both terms represents the initial velocity of an electron which comes to rest just at the potential minimum. Thus, from (2)

$$v = \sqrt{\frac{2e}{hm} V} \quad \text{and} \quad u_c' = \sqrt{\frac{2e}{hm} V'}. \tag{6}$$

In the β-region there is only one class of electrons, so that the density is more simply expressed. Thus,

$$\rho_\beta = e \int_{u_c'}^{\infty} \frac{n(u_c)}{u} du_c. \tag{7}$$

The value of ρ in (5) and (7) may each be expressed in terms of d^2V/dx^2 by the use of (3), and the integration of these two Poisson's relations for the common boundary condition that the electric force is zero at the potential minimum has the following result:

$$\frac{(dV_\alpha)^2}{(dx)} = \frac{2hm}{\epsilon}\int_{u_c'}^{\infty}(u - u')n(u_c)du_c + \frac{4hm}{\epsilon}\int_{0}^{u_c'} un(u_c)du_c, \tag{8}$$

$$\frac{(dV_\beta)^2}{(dx)} = \frac{2hm}{\epsilon}\int_{u_c'}^{\infty}(u - u')n(u_c)du_c, \tag{9}$$

where u' is the electronic velocity at the potential minimum, i.e., $(u')^2 = u_c^2 - (2e/hm)V'$.

At this point the analysis departs for the first time from the classic analyses of Fry[6] and Langmuir,[7] through the introduction of the concept that the instantaneous rate of emission may be expressed as the sum of an average rate of emission plus an instantaneous deviation. That is,

$$n(u_c) = n_0(u_c) + \delta(u_c), \tag{10}$$

transforms (8) and (9) into the following equations:

$$\frac{(dV_\alpha)^2}{(dx)} = \frac{2hm}{\epsilon}\int_{u_c'}^{\infty}(u - u')n_0(u_c)du_c + \frac{4hm}{\epsilon}\int_{0}^{u_c'} un_0(u_c)du_c + \alpha(\delta), \tag{11}$$

where

$$\alpha(\delta) = \frac{2hm}{\epsilon}\int_{u_c'}^{\infty}(u - u')\delta(u_c)du_c + \frac{4hm}{\epsilon}\int_{0}^{u_c'} u\delta(u_c)du_c$$

and

$$\frac{(dV_\beta)^2}{(dx)} = \frac{2hm}{\epsilon}\int_{u_c'}^{\infty}(u - u')n_0(u_c)du_c + \beta(\delta), \tag{12}$$

where

$$\beta(\delta) = \frac{2hm}{\epsilon}\int_{u_c'}^{\infty}(u - u')\delta(u_c)du_c.$$

Since the average rate of emission may be expressed by the Maxwellian relation,

$$n_0(u_c) = 2\alpha N u_c \epsilon^{-\alpha u_c^2},$$

where

$$\alpha = \frac{hm}{2kT},$$

the indicated integrations in (11) and (12) have as a result,

$$\frac{(kT)^2}{(e)}\frac{(d\eta_\alpha)^2}{(dx)} = \frac{Nhm}{\epsilon}\sqrt{\frac{\pi}{\alpha}}\,\epsilon^{-\eta'} \times \left[\epsilon^{\eta} - 1 + \epsilon^{\eta}P(\sqrt{\eta}) - 2\sqrt{\frac{\eta}{\pi}}\right] + \alpha(\delta) \tag{13}$$

and

$$\frac{(kT)^2}{(e)}\frac{(d\eta_\beta)^2}{(dx)} = \frac{Nhm}{\epsilon}\sqrt{\frac{\pi}{\alpha}}\,\epsilon^{-\eta'} \times \left[\epsilon^{\eta} - 1 - \epsilon^{\eta}P(\sqrt{\eta}) + 2\sqrt{\frac{\eta}{\pi}}\right] + \beta(\delta), \qquad (14)$$

where

$$\eta = \frac{e}{kT}(V' - V), \qquad \eta' = \frac{eV'}{kT},$$

$$P(x) = \frac{2}{\sqrt{\pi}}\int_0^x \epsilon^{-x^2}dx.$$

The fact that both $\alpha(\delta)$ and $\beta(\delta)$ are very small greatly simplifies the solution for the distance coordinate x in (13) and (14). The process is to invert the two equations, respectively, extract the square root, and then expand the right-hand side in powers of $\alpha(\delta)$ and $\beta(\delta)$, respectively. This results in expressions for $dx/d\eta$ which can be integrated term by term. However, the small values of $\alpha(\delta)$ and $\beta(\delta)$ allow powers higher than the first to be disregarded, and hence,

$$\frac{e}{kT}x' = \frac{F(\eta')}{\left[\frac{Nhm}{\epsilon}\sqrt{\frac{\pi}{\alpha}}\,\epsilon^{-\eta'}\right]^{1/2}} - \frac{1}{2\left[\frac{Nhm}{\epsilon}\sqrt{\frac{\pi}{\alpha}}\,\epsilon^{-\eta'}\right]^{3/2}}\int_0^{\eta'}\frac{\alpha(\delta)d\eta}{\Phi(\eta)} \qquad (15)$$

and

$$\frac{e}{kT}(x - x') = \frac{f(\eta)}{\left[\frac{Nhm}{\epsilon}\sqrt{\frac{\pi}{\alpha}}\,\epsilon^{-\eta'}\right]^{1/2}} - \frac{1}{2\left[\frac{Nhm}{\epsilon}\sqrt{\frac{\pi}{\alpha}}\,\epsilon^{-\eta'}\right]^{3/2}}\int_0^{\eta}\frac{\beta(\delta)d\eta}{\varphi(\eta)}, \qquad (16)$$

where for convenience

$$\left.\begin{aligned}
F(\eta') &= \int_0^{\eta'}\frac{d\eta}{\left[\epsilon^{\eta} - 1 + \epsilon^{\eta}P(\sqrt{\eta}) - 2\sqrt{\frac{\eta}{\pi}}\right]^{1/2}}\\
f(\eta) &= \int_0^{\eta}\frac{d\eta}{\left[\epsilon^{\eta} - 1 - \epsilon^{\eta}P(\sqrt{\eta}) + 2\sqrt{\frac{\eta}{\pi}}\right]^{1/2}}\\
\Phi(\eta) &= \left[\epsilon^{\eta} - 1 + \epsilon^{\eta}P(\sqrt{\eta}) - 2\sqrt{\frac{\eta}{\pi}}\right]^{3/2}\\
\varphi(\eta) &= \left[\epsilon^{\eta} - 1 - \epsilon^{\eta}P(\sqrt{\eta}) + 2\sqrt{\frac{\eta}{\pi}}\right]^{3/2}
\end{aligned}\right\}. \qquad (17)$$

Up to this point, the present analysis is very similar to that given by Spenke.[3] For the reasons stated above, the two methods digress hereafter.

Since the instantaneous position of the potential minimum depends upon the operating conditions as indicated above, the elimination of this variable by the addition of the two equations (15) and (16) results in a simpler expression for the cathode-anode spacing, namely:

$$\begin{aligned}\frac{e}{kT}x\left[\frac{Nhm}{\epsilon}\sqrt{\frac{\pi}{\alpha}}\,\epsilon^{-\eta'}\right]^{1/2} &= F(\eta') + f(\eta)\\ &\quad - \frac{1}{2\left[\frac{Nhm}{\epsilon}\sqrt{\frac{\pi}{\alpha}}\,\epsilon^{-\eta'}\right]}\left\{\int_0^{\eta'}\frac{\alpha(\delta)}{\Phi(\eta)}\,d\eta + \int_0^{\eta}\frac{\beta(\delta)}{\varphi(\eta)}\,d\eta\right\}\cdot \qquad (18)\end{aligned}$$

To separate the noise or fluctuation component of the potentials from their average values, it is necessary to assume at any plane in the diode that it is possible to express both the instantaneous voltage and velocity as a steady state or average value plus a very small superimposed fluctuation component. Since this assumption does not result in any discontinuities, it is possible to express the instantaneous values of the dimensionless variable as

$$\left.\begin{aligned}\eta &= \eta_0 + \eta_1\\ \text{and} \qquad \eta' &= \eta_0' + \eta_1',\end{aligned}\right\} \qquad (19)$$

where both η_1 and η_1' are very small. From this assumption, it may readily be shown that the d-c. solution and the first approximations to the fluctuation component of the solution for (18) are respectively

$$\frac{ex}{kT}\left[\frac{Nhm}{\epsilon}\sqrt{\frac{\pi}{\alpha}}\,\epsilon^{-\eta_0'}\right]^{1/2} = F(\eta_0') + f(\eta_0) \qquad (20)$$

and

$$\begin{aligned}-\frac{ex}{kT}\left[\frac{Nhm}{\epsilon}\sqrt{\frac{\pi}{\alpha}}\,\epsilon^{-\eta_0'}\right]^{1/2}\frac{\eta_1'}{2} &= \eta_1'\frac{dF(\eta_0')}{d\eta_0'} + \eta_1\frac{df(\eta_0)}{d\eta_0}\\ &\quad - \frac{1}{2\left[\frac{Nhm}{\epsilon}\sqrt{\frac{\pi}{\alpha}}\,\epsilon^{-\eta_0'}\right]}\left[\int_0^{\eta_0'}\frac{\alpha(\delta)}{\Phi(\eta)}\,d\eta + \int_0^{\eta_0}\frac{\beta(\delta)}{\varphi(\eta)}\,d\eta\right]\cdot \qquad (21)\end{aligned}$$

In the last equation, the average or d-c. values of all quantities except η_1, η_1' and $\delta(u_c)$ are to be used.

To avoid the necessity of using long, awkward equations, it will be of great convenience to define several new variables.

Let

$$\left.\begin{aligned} y &= \sqrt{\alpha}\, u' \\ B &= \frac{F(\eta_0') + f(\eta_0)}{2} + \frac{dF(\eta_0')}{d\eta_0'} + \frac{df(\eta_0)}{d\eta_0} \\ C &= \frac{1}{\sqrt{\pi}} \int_0^{\eta_0'} \frac{[\sqrt{y^2 + \eta} - y]}{\Phi(\eta)}\, d\eta \\ D &= \frac{1}{\sqrt{\pi}} \int_0^{\eta_0} \frac{[\sqrt{y^2 + \eta} - y]}{\varphi(\eta)}\, d\eta \end{aligned}\right\}. \tag{22}$$

From the definition for η, given in connection with (13) and (14), it may be shown that

$$\eta_1 = \eta_1' - \frac{e}{kT} V_1, \tag{23}$$

where V_1 is the a-c. anode potential. With the definition given for $\alpha(\delta)$ and $\beta(\delta)$ in (11) and (12), and from the above relations, (21) may be rewritten as

$$\eta_1'B - \frac{eV_1}{kT}\frac{df(\eta_0)}{d\eta_0} = \frac{1}{N\epsilon^{-\eta_0'}}\left\{ \int_{\bar{u}_c}^{\infty} (C + D)\delta(u_c)du_c \right. \\ \left. + \frac{2}{\sqrt{\pi}} \int_0^{\eta_0'} \int_{\bar{v}}^{\bar{u}_c} \frac{\sqrt{y^2 + \eta}\; \delta(u_c)du_c d\eta}{\Phi(\eta)} \right\}, \tag{24}$$

where

$$\bar{u}_c = \sqrt{\frac{2e}{hm} V_0'}, \qquad \bar{v} = \sqrt{\frac{2e}{hm} V_0}.$$

Before (24) can be used, the relation between η_1' and the a-c. anode current, and also the expression for the a-c. plate impedance must be obtained. From the general relation given in (1), the instantaneous current per unit area is

$$I = e\int_{u_c'}^{\infty} n(u_c)du_c = e\int_{u_c'}^{\infty} n_0(u_c)du_c + e\int_{u_c'}^{\infty} \delta(u_c)du_c \\ = Ne\epsilon^{-\eta'} + e\int_{u_c'}^{\infty} \delta(u_c)du_c.$$

Since the instantaneous voltage of the potential minimum may be expressed as a sum of a d-c. component plus a small a-c. fluctuation, the d-c. and first order fluctuation plate current are respectively

$$I_0 = Ne\epsilon^{-\eta_0'}, \tag{25}$$

$$I_1 = -\,\eta_1'Ne\epsilon^{-\eta_0'} + e\int_{\bar{u}}^{\infty} \delta(u_c)du_c. \tag{26}$$

Whence,

$$\eta_1' = \frac{e}{I_0}\int_{\bar{u}_c}^{\infty} \delta(u_c)du_c - \frac{I_1}{I_0}. \qquad (27)$$

This is the desired relation between η_1' and the a-c. anode current.

To find the other desired relation, it is first observed that by (20) and (25), the d-c. voltage-current relation is

$$\frac{ex}{kT}\left[\frac{hm}{\epsilon e}\sqrt{\frac{\pi}{\alpha}}\right]^{1/2} I_0^{1/2} = F(\eta_0') + f(\eta_0). \qquad (28)$$

This is identical with that obtained by Fry and Langmuir. The values of $F(\eta_0')$ and $f(\eta_0)$ have been tabulated by Langmuir.[7]

From this d-c. relation, it may readily be shown that the a-c. plate impedance of the planar diode is

$$r_p = \frac{\dfrac{F(\eta_0') + f(\eta_0)}{2} + \dfrac{dF(\eta_0')}{d\eta_0'} + \dfrac{df(\eta_0)}{d\eta_0}}{\dfrac{eI_0}{kT}\dfrac{df(\eta_0)}{d\eta_0}} = \frac{BkT}{eI_0}\frac{1}{\dfrac{df(\eta_0)}{d\eta_0}}. \qquad (29)$$

Since the noise generator voltage,

$$E = I_1 r_p + V_1,$$

the substitution of (27) and (29) reduces (24) to the following relations:

$$\begin{aligned}\frac{e}{kT}(I_1 r_p + V_1)\frac{df(\eta_0)}{d\eta_0} &= \frac{eE}{kT}\frac{df(\eta_0)}{d\eta_0} = \frac{e}{I_0}\left\{\int_{\bar{u}_c}^{\infty}(B - C - D)\delta(u_c)du_c\right. \\ &\left. - \frac{2}{\sqrt{\pi}}\int_0^{\eta_0'}\int_{-}^{\bar{u}_c}\frac{\sqrt{y^2+\eta}\,\delta(u_c)du_c d\eta}{\Phi(\eta)}\right\}. \qquad (30)\end{aligned}$$

By the additional symbols,

$$\left.\begin{aligned}H(u_c, \eta_0, \eta_0') &\equiv \frac{kT}{I_0\dfrac{df(\eta_0)}{d\eta_0}}(B - C - D) \\ G(u_c, \eta, \eta_0) &\equiv -\frac{2}{\sqrt{\pi}}\frac{kT}{I_0\dfrac{df(\eta_0)}{d\eta_0}}\frac{\sqrt{y^2+\eta}}{\Phi(\eta)}\end{aligned}\right\}, \qquad (31)$$

the noise generator voltage may be expressed in the condensed form,

$$E = \int_{\bar{u}_c}^{\infty} H\delta(u_c)du_c + \int_0^{\eta_0'}\int_{\bar{v}}^{\bar{u}_c} G\delta(u_c)du_c d\eta. \qquad (32)$$

Unfortunately (32) cannot be integrated because the specific value of the instantaneous deviation in the rate of electron emission is unknown. Moreover, as shown by Fry [10] there exists no frequency spectrum for this deviation. The reason is because there is no way of foretelling at a particular instant just when the next electron is going to be emitted from a thermionic cathode. It does not follow, however, that Fourier methods are powerless, as the following argument will show. Imagine that the emission in a thermionic system has been going on for a long time, and place a recorder in the system which makes an oscillograph record of the voltage produced across the tube by the fluctuating current. Let the record be made over a long period of time. Then, it is a perfectly possible thing to analyze the record so obtained into a Fourier spectrum. The result will give no information that the Fourier spectrum which would be obtained on the oscillograph during an ensuing time period of equal length would be the same as the one which has been secured during the past. However, when the mean square value of the Fourier spectrum for the recorded interval is computed, it is found that the mean square value of any two records so obtained is the same when they are both produced by random emission. Moreover, the mean square value within a specified frequency interval is also the same in the two records. These facts result from the random character of the events producing the records, as may be seen even more clearly by examination of the mathematical steps in the following equations, particularly as given in the progression from (37) to (38). It follows, then, that one is justified in concluding that the mean square response of an electrical system to a random excitation can be calculated by the Fourier series method, and that the result so obtained applies equally well either to systems which have been measured in the past, or to those which will be measured in the future, provided only that they both are similar in their configuration and external operating conditions.

To obtain the Fourier expression for the emission deviation, it will be assumed that this function repeats itself after a very long period of time. To find this Fourier Series for the instantaneous deviation from the average rate of emission of electrons with x-directed velocities between u_c and $u_c + du_c$, the very long period of time, L, is divided into P equal intervals of length, τ, where τ is assumed to be mathematically small. The exact number of electrons emitted during any of these intervals of time with velocities between u_c and $u_c + du_c$ per unit area of the cathode is denoted by $n_m(u_c)\tau$. The average rate of emission of electrons in this velocity class may then be defined as follows:

$$n_0(u_c) \equiv \frac{1}{P}\sum_{m=1}^{P} n_m(u_c). \tag{33}$$

During each interval, the deviation from the average rate of emission is

$$\delta_m(u_c) = n_m(u_c) - n_0(u_c). \tag{34}$$

The mean square of the instantaneous deviation for all P intervals is then

$$\overline{\delta^2(u_c)} = \frac{1}{P}\sum_{m=1}^{P} [n_m(u_c) - n_0(u_c)]^2. \tag{35}$$

The value of the mean square emission deviation may be found from the following considerations. In previous studies of pure shot noise, it was assumed that the electrons are emitted from the cathode independently of one another, that is, the probability of an electron being emitted during a very small interval of time depends only upon the average rate of emission and the length of the time interval. The classical theoretical equation developed for the shot noise from this assumption was found experimentally to be correct to well within experimental errors. Hence, the same assumption may be made in this analysis of tube noise as the presence of the space charge near the cathode can have but a negligible effect upon the rate of emission of electrons from the cathode. Thus, since the electrons in any velocity class are also emitted independently of one another, the application of probability theory shows that the mean square deviation in the total number of electrons emitted with velocities between u_c and $u_c + du_c$ in a time, τ, is equal to the average number of electrons of this velocity class emitted in the same time, τ.

That is,

$$\frac{1}{P}\sum_{m=1}^{P} \tau^2\delta_m{}^2(u_c) = \tau n_0(u_c),$$

or

$$\sum_{m=1}^{P} \delta_m{}^2(u_c) = \frac{L}{\tau^2} n_0(u_c). \tag{36}$$

Since the period of time, τ, is mathematically small, it can be assumed that the instantaneous deviation from the average rate of emission is constant in each of the P intervals and equal to $\delta_m(u_c)$. If it is assumed that the function representing the instantaneous deviation repeats itself after a very long period of time, L, the Fourier series for the P square-top pulses comprising the instantaneous deviation function may be shown to be:

$$\delta(u_c) = \sum_{l=-\infty}^{\infty} \sum_{m=1}^{P} \frac{\tau\delta_m(u_c)}{L} \left[\frac{e^{-il\omega\tau} - 1}{-il\omega\tau} \right] e^{il\omega(t-t_m)}, \tag{37}$$

where $\omega = 2\pi/L$, and $t_m = m\tau$.

The mean square deviation may be found by squaring the above expression and averaging the result over the long period of time L. The result is

$$\overline{\delta^2(u_c)} = 2\sum_{l=1}^{\infty} \sum_{k=1}^{P} \sum_{m=1}^{P} \frac{\tau^2}{L^2} \delta_m(u_c)\delta_k(u_c) \left[\frac{e^{-il\omega\tau} - 1}{-il\omega\tau} \right] \times \left[\frac{e^{+il\omega\tau} - 1}{+il\omega\tau} \right] e^{-il\omega(t_k-t_m)}.$$

Since the time-average of the instantaneous emission deviation must be zero over the period, L, the contribution to the mean square from the double summation with respect to k and m is zero, unless $m = k$.

Thus

$$\overline{\delta^2(u_c)} = 4\sum_{l=1}^{\infty} \frac{\tau^2}{L^2} \left[\frac{1 - \cos l\omega\tau}{l^2\omega^2\tau^2} \right] \sum_{m=1}^{P} \delta_m{}^2(u_c). \tag{38}$$

From (36), this equation reduces to

$$\overline{\delta^2(u_c)} = \frac{4}{L}\sum_{l=1}^{\infty} \left[\frac{1 - \cos l\omega\tau}{l^2\omega^2\tau^2} \right] n_0(u_c). \tag{39}$$

The contribution to the mean square of the instantaneous deviation in the electron emission, from the frequencies between f and $f + df$ is given by

$$\overline{\delta_f{}^2(u_c)} = \frac{4}{L} \sum_{l=2\pi f/\omega}^{l=2\pi(f+df)/\omega} \left[\frac{1 - \cos l\omega\tau}{l^2\omega^2\tau^2} \right] n_0(u_c). \tag{40}$$

The limit of this expression as the length of the periodicity, L, is made infinite, may readily be shown to be

$$\overline{\delta_f{}^2(u_c)} = 2n_0(u_c)df. \tag{41}$$

It is now possible to proceed to find the mean square value of the noise generator voltage given in (32) with the aid of (37) and (41). From (37), the noise generator voltage may be expressed as

$$E = \sum_{l=-\infty}^{\infty} \sum_{m=1}^{P} \frac{\tau}{L} \left[\frac{e^{-il\omega\tau} - 1}{-il\omega\tau} \right] e^{il\omega(t-t_m)} \times \left\{ \int_{\bar{u}}^{\infty} H(u_c)\delta_m(u_c)du_c + \int_0^{\eta_0'} \int_{\sqrt{\alpha(\eta-\eta_0')}}^{\bar{u}_c} G\delta_m(u_c)du_c d\eta \right\}. \tag{42}$$

The mean square of this equation for the noise generator voltage may be obtained by finding the average of the square of the expression over a very long period of time; that is

$$
\begin{aligned}
\overline{E^2} = 4\sum_{l=1}^{\infty}\sum_{k=1}^{P}\sum_{m=1}^{P}\frac{\tau^2}{L^2}\left[\frac{1-\cos l\omega\tau}{l^2\omega^2\tau^2}\right]e^{il\omega(t_m-t_k)} \\
\times\left\{\int_{x=\bar{u}_c}^{\infty}\int_{y=\bar{u}_c}^{\infty}H(x)H(y)\,\delta_m(x)\,\delta_k(y)dxdy\right. \\
+\,2\int_{\eta=0}^{\eta_0'}\int_{y=\bar{u}_c}^{\infty}\int_{x=\sqrt{\alpha(\eta-\eta_0')}}^{\bar{u}_c}G(\eta,\,x)H(y)\,\delta_m(x)\,\delta_k(y)dxdyd\eta \\
+\int_{\eta=0}^{\eta_0'}\int_{z=0}^{\eta_0'}\int_{y=\sqrt{\alpha(z-\eta_0')}}^{\bar{u}_c}\int_{x=\sqrt{\alpha(\eta-\eta_0')}}^{\bar{u}_c}G(\eta,\,x)G(z,\,y) \\
\left.\times\,\delta_m(x)\,\delta_k(y)dxdydzd\eta\right\}. \qquad (43)
\end{aligned}
$$

In the above equation, the contribution to the mean square noise generator voltage from the summation with respect to m, for a fixed value of k, is zero unless $m = k$, since the long time average of the emission deviation must be zero. Furthermore, since the electrons are emitted independently of one another

$$
\sum_{m=1}^{P}\delta_m(x)\,\delta_m(y) = 0, \quad \text{unless} \quad x = y.
$$

From these considerations, the contribution to the mean square generator voltage from the second integral in (43) is zero since x and y have no common value. The contribution from the last integral is a bit more difficult to obtain. However, from (41), the contribution to the mean square noise generator voltage from the frequencies between f and $f + df$ can be shown to be

$$
\begin{aligned}
\overline{E_f{}^2} = 2df\left\{\int_{\bar{u}_c}^{\infty}H^2(u_c)n_0(u_c)du_c\right. \\
+\int_{\eta=0}^{\eta_0'}\int_{z=0}^{\eta}\int_{u_c=\sqrt{\alpha(\eta-\eta_0')}}^{\bar{u}_c}G(\eta,\,u_c)G(z,\,u_c)n_0(u_c)du_cdzd\eta \\
\left.+\int_{\eta=0}^{\eta_0'}\int_{z=\eta}^{\eta_0'}\int_{u_c=\sqrt{\alpha(z-\eta_0')}}^{\bar{u}_c}G(\eta,\,u_c)G(z,\,u_c)n_0(u_c)du_cdzd\eta\right\}. \qquad (44)
\end{aligned}
$$

In terms of the variable $y = \sqrt{\alpha}u'$, the average rate of emission may readily be shown to be

$$
n_0(u_c)du_c = \frac{2I_0}{e}y\epsilon^{-y^2}dy.
$$

From this value of the average rate of emission, and from the definition of H and G in (31), and since the plate impedance of the planar diode is given by (28), the final expression for the mean square noise generator voltage may be expressed as follows:

$$\overline{E_f^2} = 4kr_p(\lambda T)df, \tag{45}$$

where

$$\begin{aligned}\lambda = \frac{1}{B\dfrac{df(\eta_0)}{d\eta_0}}\Bigg\{ & \int_0^{\infty} y(B - C - D)^2 e^{-y^2} dy \\ & + \frac{4}{\pi}\int_{z=0}^{\eta_0'}\int_{x=0}^{z}\int_{y=\sqrt{-x}}^{0} \frac{y\sqrt{y^2+x}\,\sqrt{y^2+z}\;e^{-y^2}}{\Phi(x)\Phi(z)}\,dxdydz \\ & + \frac{4}{\pi}\int_{z=0}^{\eta_0'}\int_{x=z}^{\eta_0'}\int_{y=\sqrt{-z}}^{0} \frac{y\sqrt{y^2+x}\,\sqrt{y^2+z}\;\epsilon^{-y^2}}{\Phi(x)\Phi(z)}\,dxdydz\Bigg\},\end{aligned} \tag{46}$$

$$\left.\begin{aligned}
\eta_0 &= \frac{e}{kT}(V' - V_p), \qquad \eta_0' = \frac{e}{kT}V' \\
B &= \frac{F(\eta_0') + f(\eta_0)}{2} + \frac{dF(\eta_0')}{d\eta_0'} + \frac{df(\eta_0)}{d\eta_0}, \qquad r_p = \frac{B}{\dfrac{eI_0}{kT}\dfrac{df(\eta_0)}{d\eta_0}} \\
F(\eta_0') &= \int_0^{\eta_0'} \frac{dx}{\left[\epsilon^x - 1 + \epsilon^x P(\sqrt{x}) - 2\sqrt{\dfrac{x}{\pi}}\right]^{1/2}}, \\
f(\eta_0) &= \int_0^{\eta_0} \frac{dx}{\left[\epsilon^x - 1 - \epsilon^x P(\sqrt{x}) + 2\sqrt{\dfrac{x}{\pi}}\right]^{1/2}} \\
C &= \frac{1}{\sqrt{\pi}}\int_0^{\eta_0'} \frac{[\sqrt{y^2+x} - y]}{\Phi(x)}dx, \\
D &= \frac{1}{\sqrt{\pi}}\int_0^{\eta_0} \frac{[\sqrt{y^2+x} - y]dx}{\left[\epsilon^{-x} - 1 - \epsilon^x P(\sqrt{x}) + 2\sqrt{\dfrac{x}{\pi}}\right]^{3/2}} \\
\Phi(x) &= \left[\epsilon^x - 1 + \epsilon^x P(\sqrt{x}) - 2\sqrt{\dfrac{x}{\pi}}\right]^{3/2}, \\
P(x) &= \frac{2}{\sqrt{\pi}}\int_0^x \epsilon^{-x^2}dx.
\end{aligned}\right\}. \tag{47}$$

Part II—General Discussion

The analysis in Part I shows that as soon as a potential minimum exists, the tube noise in a planar diode is equivalent to the thermal noise of the plate resistance at an effective temperature which is a function of that of the cathode.

In general, the effective value of the diode plate resistance temperature for any operating condition is very difficult to obtain because of the complexity of the final noise equations (45) and (46). However, the limiting value of the ratio of the effective plate resistance temperature to that of the cathode, denoted by "λ" in (45), may be evaluated very readily for certain limiting conditions.

One encounters the first of these conditions when the plate potential and cathode emission are such that the potential minimum has moved just up to the cathode, and is in fact on the point of disappearing. This condition is secured by decreasing the space charge to values less than are required for the formation of a potential minimum away from the cathode. In the equations, it is represented by letting the quantity η_0' approach zero, where η_0' is the natural logarithm of the ratio of the saturation current to the anode current. For this set of operating conditions, all the electrons emitted from the cathode will go to the anode, and hence the condition is appropriate to the study of pure shot noise.

A second condition is obtained when the plate potential is equal in value to the potential of the minimum. Physically, this condition means that the minimum has moved just up to the anode, and requires a negative value for the plate potential. Mathematically, it is represented by a zero value for the quantity η_0, where η_0 is equal to the difference between η_0' and $(e/kT)V_p$. For negative plate voltages greater in magnitude than that of the potential minimum, all electrons having an initial kinetic energy greater than eV_p will reach the anode regardless of the presence of the space charge existing between the two electrodes. For these conditions, the diode becomes a temperature limited current device.

A third limiting condition occurs when the plate potential is large in magnitude compared with that of the potential minimum referred to the cathode. In this condition a potential minimum still exists. It is represented in the mathematics by letting the quantity η_0 become large. This condition represents the normal operating condition for the diode.

As the space charge is decreased, making η_0' very small, from (47), the diode plate impedance becomes very large through the action of $dF(\eta_0')/d\eta_0'$ which becomes infinite as η_0' approaches zero. As all other quantities involved remain finite, the mean square noise current for a very small space charge is

$$\overline{I_1^2} = \frac{\bar{E}_f^2}{(R_p + Z)^2} = \frac{4kTdf}{B\,\dfrac{kT}{eI_0}} \, \frac{\dfrac{df(\eta_0)}{d\eta_0}}{B\,\dfrac{df(\eta_0)}{d\eta_0}} \, B^2 \int_0^\infty y\epsilon^{-y^2}dy \rightarrow 2eI_0df. \tag{48}$$

Thus, as the potential minimum voltage is reduced to zero, the tube noise as given by (45) reduces to the well known shot effect equation.

For some space charge at the cathode, the value of λ in (45) has definite limiting values for both very low and for very large plate voltages. For a very small value of η_0, that is for negative plate voltage, the value of B defined in (47) is very large because $df(\eta_0)/d\eta_0$ becomes infinite as η_0 is decreased to zero. Thus as $\eta_0 \rightarrow 0$

$$\lambda \rightarrow \frac{1}{B^2} B^2 \int_0^\infty y\epsilon^{-y^2}dy = \frac{1}{2}\cdot \tag{49}$$

Hence, for any value of space charge, the effective plate resistance temperature for negative plate voltages is one-half of the cathode temperature, under the restriction that no potential minimum exists between the cathode and anode.

Since the diode is usually operated with a positive plate voltage, the value of the effective plate resistance temperature for a large value of η_0 is of more interest. For η_0' not equal to zero, and a large value of plate voltage, it can be readily shown that the values of $f(\eta_0)$ and of D are much larger than any other quantities involved in the equation for λ. After a bit of mathematical operation, it may be shown that the limiting values for $f(\eta_0)$ and D are

$$D = \frac{\pi^{1/4}}{2^{3/2}}\left[\frac{4}{3}\eta_0^{3/4} + 3\sqrt{\pi}\eta_0^{1/4} + \cdots - (4y\eta_0^{1/4} + \cdots)\right]$$

$$f(\eta_0) = \frac{\pi^{1/4}}{\sqrt{2}}\left[\frac{4}{3}\eta_0^{3/4} + \sqrt{\pi}\eta_0^{1/4} + \cdots\right].$$

From these relations, the limiting value of λ for a large plate voltage is given by

$$\lambda = 3\int_0^\infty y\left[\sqrt{2}y - \sqrt{\frac{\pi}{4}}\right]^2 \epsilon^{-y^2}dy = 3\left(1 - \frac{\pi}{4}\right) = 0.644. \tag{50}$$

Thus, for any value of space charge, as long as a potential minimum exists, a sufficiently large value of plate voltage may always be found for which the effective plate resistance temperature is 0.644 times the cathode temperature.

It is possible to obtain a good approximation for the effective diode temperature for any operating condition by the following method. The values of "C" and "D" in (47) may be found without too much difficulty by graphical integration for several different values of y, η_0 and η_0'. From the tabulated values for $F(\eta_0')$ and $f(\eta_0)$ given by Langmuir, and from the values found for C and D, the integral,

$$S = \int_0^\infty y[B - C - D]^2 \epsilon^{-y^2} dy, \tag{51}$$

may be evaluated by mechanical means for several values of η_0 and η_0'. This gives the first integral in (46).

It was found practically impossible to calculate directly the contribution to λ from the last two integrals in (46). However, a rough approximation to them may be found indirectly by the following method: If the sum of the two integrals is denoted by Q, then (46) may be written

$$\lambda = \frac{S + Q}{B \dfrac{df(\eta_0)}{d\eta_0}},$$

or $Q = B\lambda(df(\eta_0)/d\eta_0) - S =$ function of η_0' only.

For a fixed value of η_0', the solution of the above equation for several values of η_0 should give a constant value for Q. Unfortunately, only the limiting values of λ are known. However, if the limiting value of 0.644 is substituted for λ in this equation, the calculated value of Q, for a fixed value of η_0', should approach a constant value as η_0 is increased since λ does assume the 0.644 value for η_0 sufficiently large. The limiting value of Q calculated in this manner is the desired contribution to λ from the last two integrals in (46). This method of evaluating Q cannot be very accurate since it involves the difference of two quantities of the same magnitude. However, since Q is small compared to the contribution from the first integral in (46), a large error in Q will introduce a much smaller error in the value of λ.

The values of the effective diode plate resistance temperature calculated in this manner for several different operating conditions are shown in Fig. 2. These curves indicate that the effective diode temperature is 0.644 times the cathode temperature for all practical operating conditions. The values of η_0' and η_0 may be determined from the following relations:

$$\eta_0' = \log_e \frac{I_s}{I_p}, \tag{52}$$

where I_s is the saturation current and I_p is the anode current, and

$$\eta_0 = \eta_0' - \frac{e}{kT} V_p = \eta_0' + \frac{1.16}{T} \times 10^4 V_p, \tag{53}$$

where V_p is the anode potential, and T is the absolute cathode temperature.

For $T = 900°$ K,

$$\eta_0 = \eta_0' + 12.9 V_p. \tag{54}$$

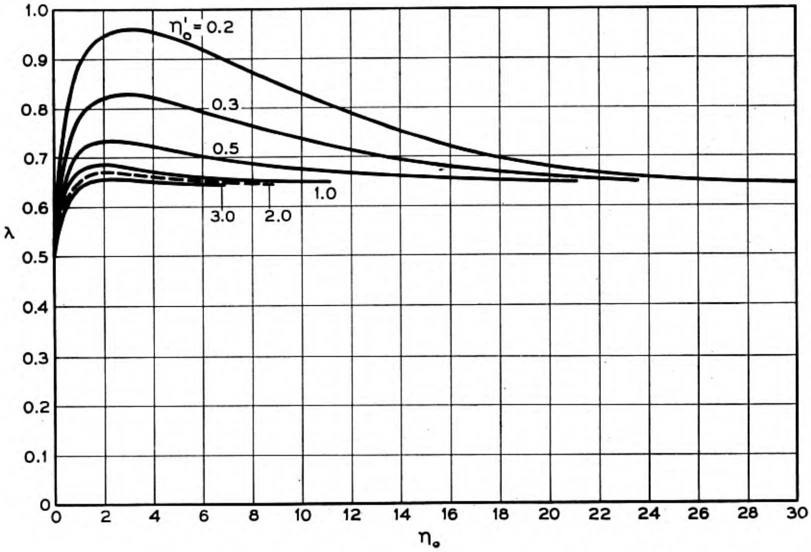


Fig. 2—Effective noise generator voltage of planar diode.

$$\overline{E_f^2} = 4kr_p(\lambda T)df, \qquad \eta_0' = \log_e \frac{I_s}{I_p}, \qquad \eta_0 = \eta_0' - \frac{e}{kT} V_p.$$

For $T = 900°$ K., $\eta_0 = \eta_0' + 12.9 V_p$.

Even for the small space charge condition for which the plate current is eight-tenths of the saturation current ($\eta_0' = 0.2$), the value of η_0 need be greater than about 25 only before λ assumes its limiting value. For a temperature of 900° K., as for oxide coated cathodes, this would require a plate voltage of only two volts. If the plate current were less than eight-tenths of the saturation current for very high plate voltages, then as the plate voltage is reduced, η_0' would increase. For this operating condition λ maintains its limiting value of 0.644 for all except negative values of plate voltages.

The transition between the various effective planar diode plate resistance temperatures is more clearly shown in Fig. 3. In this figure, the natural logarithm of the ratio of saturation current to the plate current is plotted as a function of the plate voltage for several constant values of the coefficient λ. These curves show for a fixed positive value of plate voltage that as the space charge is decreased toward zero, by a reduction in the ratio of the saturation current to the plate current, the value of λ for moderately large values of space charge increases but little from 0.644, and then for a very low space charge, increases very rapidly to its limiting value given by the shot noise

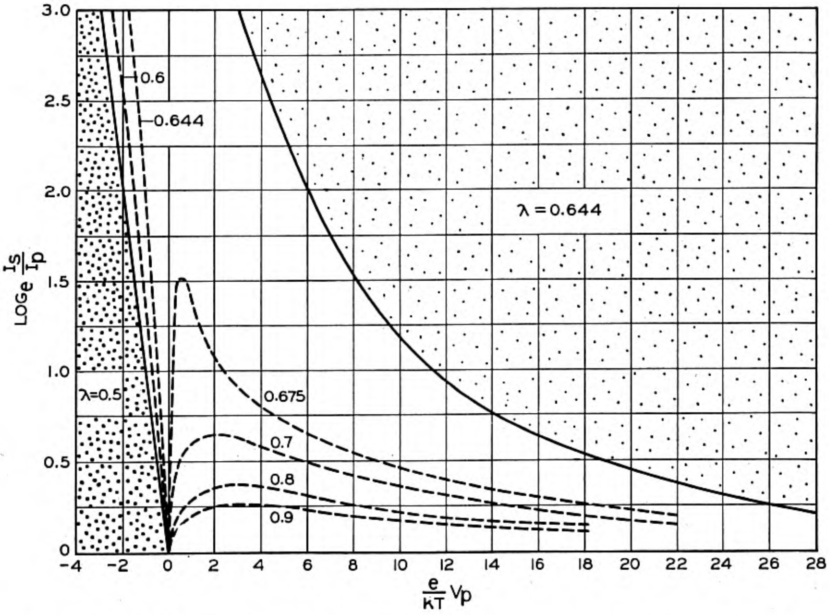


Fig. 3—Modification of effective plate resistance temperature produced by space charge.

$$\overline{E_f^2} = 4kr_p(\lambda T)df.$$

which is represented by the axis of abscissæ. Thus, the value of λ digresses markedly from 0.644 only for the narrow region of operating conditions for which the saturation current is less than 1.25 times the plate current and the plate voltage is less than $28e/kT$ volts. For an oxide coated cathode for which T is 900° K., the effective plate resistance temperature is $0.644T$ for any operating condition for which plate current is less than eight-tenths of the saturation current and the plate voltage is greater than two volts.

For a cylindrical diode, the general method of analysis used in the parallel plane case results in equations which are practically impossible to solve. The difficulty in these equations arises from the fact that tangential as well as radial initial velocities must be considered in obtaining the total anode current. Since it was shown for the planar diode that the effective temperature of the plate resistance is 0.644 times the cathode temperature for practically all operating conditions, all that is really desired in the cylindrical diode solution is the limiting value of the effective tube temperature. This may be found rather easily from a comparison of the cylindrical diode with the planar tube in the following manner.

For a very large space charge, and a high plate potential the radius of an equipotential surface near the potential minimum will be very nearly equal to that of the cathode. Hence, for these operating conditions, the planar diode equations may be applied to this region of the cylindrical diode. In the planar tube, it was shown that for $\eta_0' > 3$, η_0 had to be of the order of unity to obtain the limiting value of 0.644 for λ. If the space charge and plate potential are sufficiently large in the cylindrical diode, the radius of the equipotential surface for which η_0 is greater than unity will practically be equal to that of the cathode. The cylindrical diode may then be divided into two parts, a planar diode between the cathode and the equipotential surface for which $\eta_0 > 1$, and a cylindrical diode formed from the remainder of the tube. In any diode, the only source of noise energy is the cathode from which the noise power is transferred to the anode and external circuit through the mechanism of the initial electronic velocities. Furthermore, the same total noise power must be transferred across any equipotential surface between the cathode and anode. In the planar portion of the cylindrical diode as described above, the total noise power crossing any equipotential surface was shown to be $2.576kTdf$. This same noise power must be transferred across any other equipotential surface in the cylindrical diode. Hence, the effective plate resistance temperature for the cylindrical electrode tube must also be 0.644 times its cathode temperature. From this line of reasoning, it may be shown that the limiting value of the effective temperature for any shape diode is the same as that for the planar tube with the same cathode temperature.

From the experimental data given in his paper, Pearson definitely recognized that the limiting value of the diode plate resistance temperature should be between 0.59 and 0.65 of that of the cathode.[2] The writer understands that North and Thompson of the R.C.A. in an unpublished paper have obtained the same general result for the effect of space charge upon shot noise in diodes.

In a diode, the tube noise may be expressed equally well and with equal correctness either as a modified shot noise or as a thermal resistance noise. In this paper, the thermal resistance viewpoint was taken for two reasons. First, the coefficient "λ," used in the thermal resistance noise equation

$$\overline{E_f^2} = 4kr_p(\lambda T)df,$$

is practically always a constant equal to 0.644, whereas, the factor, "F," used by Schottky and Spenke in their modified shot noise equation

$$\overline{I_1^2} = 2eF^2I_0df$$

is always a function of the operating condition. That is, for the operating conditions for which λ is a constant, F has the following value:

$$F = \frac{1.39}{\left[\log\frac{I_s}{I_p} - \frac{e}{kT}V_p\right]^{1/2}}.$$

The second reason for the selection of the thermal resistance noise relation is that power from the motion of the atoms in the cathode is actually transferred to the plate electrode and external circuit through the mechanism of the initial electron velocities. Hence, the tube noise in a diode with space charge is very similar to a thermal resistance noise.

Part III—Effect of Transit Time

The analysis, in Part I, while giving the correct results for all operating conditions in the ordinary frequency range, is extremely long and cumbersome. It shows, however, that only the limiting values of the effective temperature of the plate resistance are required for most practical cases, and therefore it points the way to make simplifying assumptions which result in a much shorter analysis, and moreover, which allow the analysis to be extended to frequencies so high that electron transit time phenomena become of importance.

Thus the final noise equation in Part I shows that for moderately high anode potentials and for the usual excess of cathode emission, a very good approximation may be had by a consideration of the current-voltage relations existing in the β-region between potential minimum and anode without the necessity of encumbering the analysis by including the α-region between potential minimum and cathode. Moreover, for a large anode potential, the terminal velocities of the electrons at the plate are very large in comparison with their initial velocities for practically all of the electrons. This means that the transit time for the various electrons is practically the same for all of them which leave the cathode within a particular very short time interval, even though the initial velocities of the various electrons are statistically distributed among them. It results that the various individual velocities of the electrons in the β-region may be replaced by an average value, which at the potential minimum may be defined as follows:

$$\bar{u} = \frac{\int_{u_c'}^{\infty} u' n(u_c) du_c}{\int_{u_c'}^{\infty} n(u_c) du_c} \cdot \tag{55}$$

Physically, the meaning of this expression is the average velocity of these electrons which cross a plane in the β-region close to the potential minimum in a unit of time. Inasmuch as the unit of time may be taken to be very small, it follows that (55) expresses the effective instantaneous value of the initial velocity which may, and does, fluctuate as time goes on.

On the basis of an equation of the form

$$I = \rho u - \epsilon \frac{\partial^2 V}{\partial t \, \partial x} \tag{56}$$

the planar diode has been extensively investigated by a number of workers and it has been shown [8] that the relation between current and voltage is completely specified as soon as two boundary conditions are given. These may be the initial velocity and acceleration, or they may equally well be the initial velocity and conduction current ρu. However, the analysis based on (56) applies strictly to the case where all of the charge moves with the same velocity and hence contains a certain approximation when electrons are considered whose velocities have a certain dispersion around some mean value. The error will be small until frequencies are considered which are so high that a large proportion of the electrons which left the cathode in a time interval which is very short compared with the period of the high frequency arrive at the anode in a time interval which is not small compared with the high frequency period. Normally this means that the error is small even for frequencies so high that the majority of the electrons require several cycles to make their transit from potential minimum to anode.

It is convenient to write the resulting equations in terms of d-c. and first order a-c. values where the initial values of d-c. velocity and acceleration are given, but initial values of a-c. velocity and conduction current are employed. The first order a-c. relation derived by Llewellyn may be written in the form

$$-\frac{e}{hm}V_1 = \frac{e}{hm\epsilon}I_1A + \frac{e}{hm\epsilon}q_aB + \mu_aC, \tag{57}$$

where q_a and μ_a are the initial values of fluctuation conduction current and velocity, respectively, while A, B and C are defined by:

$$\left.\begin{aligned} A &= \frac{1}{\omega^4}\left[-i\omega^3x + \frac{eI_0}{hm\epsilon}(2 - 2e^{-i\theta} - i\theta - i\theta e^{-i\theta})\right] \\ B &= -\frac{1}{i\omega^3}[a_a(i\theta e^{-1\theta} + e^{-i\theta} - 1) + u_ai\omega(e^{-i\theta} - 1)] \\ C &= \frac{eI_0}{hm\epsilon\omega^2}[i\theta e^{-i\theta} + e^{-i\theta} - 1] \end{aligned}\right\}, \tag{58}$$

in which $\overset{\theta}{\eta}$ is the transit angle, $\omega\tau$, the transit time being τ, and I_0 is the d-c. current.

In the application of these relations to noise analysis, the initial values of velocity, acceleration, and conduction current must be taken at a point in the β-region beyond the potential minimum, but just as close to it as possible without encountering conditions where electrons may be moving toward the cathode, for the equations apply only to cases where the electrons are moving in one direction only. The initial point is, however, located so near to the potential minimum that the d-c. acceleration in (58) may be taken as zero. When this is the case, it may be shown that the initial conduction current is equal to the total current. In other words, the initial value of displacement current is zero. Under such conditions (57) and (58) reduce to the following expression for the a-c. anode potential in terms of the a-c. component of current and initial velocity:

$$\begin{aligned} V = \frac{I_1}{\omega^4\epsilon}\bigg[\frac{eI_0}{hm\epsilon}\left(\frac{i\theta}{6} + i\theta + 2e^{-i\theta} + i\theta e^{-i\theta} - 2\right) \\ + \omega^2u_a(i\theta + e^{-i\theta} - 1)\bigg] + \frac{\mu_aI_0}{\omega^2\epsilon}[i\theta e^{-i\theta} + e^{-i\theta} - 1]. \end{aligned} \tag{59}$$

The term multiplying the a-c. current I_1 in the above equation is the internal high-frequency impedance z of the planar diode. The last term may therefore be identified with an internal emf. When the initial velocity μ, is expressed in terms of the fluctuation of electron velocity, the term gives the equivalent noise generator, E. Thus

$$E = \frac{\mu_a I_0}{\omega^2 \epsilon} (i\theta e^{-i\theta} + e^{-i\theta} - 1) \tag{60}$$

and the mean-square value of the noise emf. (at a frequency ω) is given by:

$$\overline{E^2} = \frac{\overline{\mu_a{}^2} I_0{}^2}{\omega^4 \epsilon^2} \,|\, i\theta^{-i\theta} + e^{-i\theta} - 1 \,|^2. \tag{61}$$

The problem is now reduced to finding the mean square value of initial velocity fluctuation, $\overline{\mu_a{}^2}$, which corresponds to electrons crossing the potential minimum. This may be done by going to (55) which gives the effective value of the instantaneous initial velocity and separating all quantities, including the lower integration limits into d-c. and a-c. components. Thus

$$\left.\begin{aligned} n(u_c) &= \underline{n_0(u_c)} + \delta(u_c) \\ u_c{}' &= \overline{u_c{}'} + \delta u_c{}' \\ u' &= \overline{u'} + \delta u' \\ \bar{u} &= u_a + \mu_a \end{aligned}\right\}. \tag{62}$$

The result may be expanded in series form and products of the δ's may be disregarded inasmuch as the a-c. components are small in comparison with the d-c. The indicated operations have as a result

$$u_a = \sqrt{\frac{\pi k T}{2hm}} \tag{63}$$

and

$$\mu_a = \frac{e}{I_0} \int_{\bar{u}_c}^{\infty} (u' - u_a) \delta(u_c) du_c. \tag{64}$$

The Fourier analysis may be applied to this in the way outlined in connection with (37) and (41) in Part I and gives the mean-square value of velocity fluctuation corresponding to a frequency interval df as follows:

$$\overline{\mu_f{}^2} = \frac{2e^2}{I_0{}^2} df \int_{\bar{u}_c}^{\infty} (u' - u_a)^2 u_c(u_c) du_c = \frac{4ekT}{I_0 hm} df \left(1 - \frac{\pi}{4} \right) \cdot \tag{65}$$

This may be substituted in (62) giving for the effective noise emf. in the frequency range df

$$\overline{E_f{}^2} = 4kTdf \left[\frac{eI_0 \tau^4}{hm\epsilon^2} \right] \left[1 - \frac{\pi}{4} \right] \left[\frac{1}{\theta^4} \right] \times [\theta^2 + 2 - 2(\cos\theta + \theta \sin\theta)]. \tag{66}$$

The initial average velocity is small so that the low-frequency plate impedance may be written

$$r_p = \frac{eI_0\tau^4}{12hm\epsilon^2}. \tag{67}$$

Thus for any transit angle, the mean square noise generator voltage is given by

$$\left.\begin{aligned} \overline{E_f^2} &= 12\left(1-\frac{\pi}{4}\right)kr_pTdf\left\{\frac{4}{\theta^4}[2+\theta^2-2(\cos\theta+\theta\sin\theta)]\right\}\\ &= 4Sk(0.644T)r_pdf\\ \text{where}&\\ S &= \frac{4}{\theta^4}\left[\,2+\theta^2-2(\cos\theta+\theta\sin\theta)\,\right] \end{aligned}\right\}. \tag{68}$$

For low transit angles, this expression reduces to

$$\overline{E_f^2} = 12\left(1-\frac{\pi}{4}\right)kr_pTdf, \tag{69}$$

which is precisely the limiting value obtained by the much longer, but more rigorous analysis.

It must be understood that (68) is an approximation since the transit time effect in the region between cathode and potential minimum was entirely neglected, and because the validity of the average velocity concept does fail at the very high frequencies.

Some knowledge of the extent of the operating conditions for which the above equations are good approximations may be obtained from the d-c. current-voltage relation. For the boundary conditions assumed, the low frequency current equation derived from the general solution given by Llewellyn reduces to

$$I = \frac{-\,2.33(V-V')^{3/2}}{10^6(x-x')^2}\left[1+2.66\sqrt{\frac{kT}{e(V'-V)}}\right]. \tag{70}$$

This equation was shown by Langmuir to be a very good approximation for the plate current for most operating conditions and fails only for very low values of plate voltages. Thus, it may be concluded that (68) is a good approximation for all operating conditions except for very low plate voltages and a small space charge.

The plot of (68) given in Fig. 4 shows that the magnitude of the mean square noise generator voltage decreases by five per cent only for transit angles as large as one radian.

The effect of transit time on the pure shot noise for a low space charge density and a high plate voltage may be obtained quite readily from (57) and (58). Since for a very small space charge, I_0 and u_a are small, and a_a large, the equations then reduce to the following expression:

$$-V_1 = \frac{I_1}{i\omega\epsilon}\frac{a_a\tau^2}{2} + \frac{q_a}{i\omega\epsilon}a_a\tau^2\left[\frac{1}{-\theta^2}(i\theta e^{-i\theta} + e^{-i\theta} - 1)\right]\cdot \qquad (71)$$

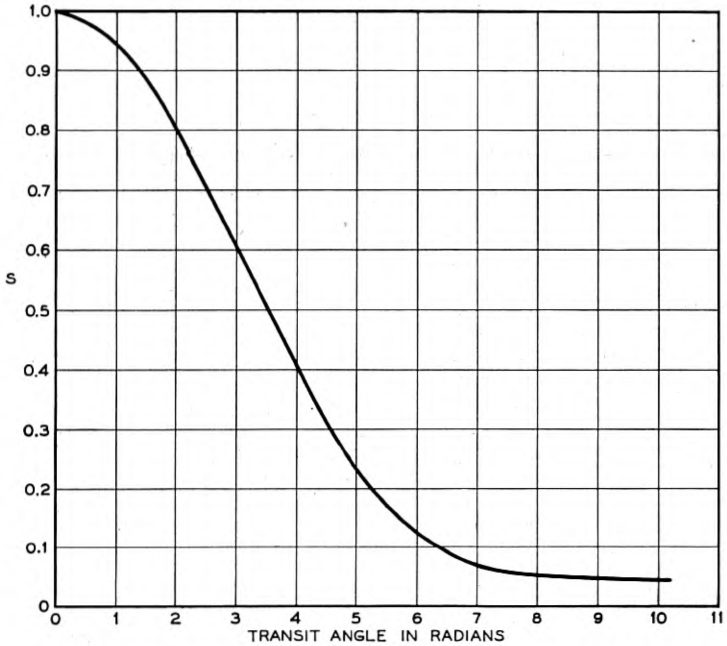


Fig. 4—Effect of transit time on both thermal and shot tube noise.

$$\overline{E_f^2} = 4Sk(0.644T)r_p df,$$
$$\overline{I_f^2} = 2eSI_0 df.$$

For these operating conditions, the transit time in terms of the d-c. acceleration and the electrode spacing is given by

$$x = \frac{a_a\tau^2}{2}\cdot$$

In terms of the external circuit impedance Z_f

$$V_1 = I_1 Z_f, \qquad (72)$$

so that (71) and (72) combine to give

$$I_1 = -\frac{q_a}{1 + \dfrac{Z_f i\omega\epsilon}{x}}\left[\frac{2}{i\theta^2}\left(i\theta e^{-i\theta} + e^{-i\theta} - 1\right)\right]\cdot \tag{73}$$

The mean square shot noise current is thus given by

$$\overline{I_1^2} = \frac{\overline{q_a^2}}{\left|1 + \dfrac{Z_f i\omega\epsilon}{x}\right|^2}\left|\frac{2}{\theta^2}\left(i\theta e^{-i\theta} + e^{-i\theta} - 1\right)\right|^2\cdot \tag{74}$$

The value of the mean square a-c. conduction current at the cathode to be substituted in the above equation may be derived as follows:

The total current emitted from the filament was defined as

$$I = e\int_0^{\infty} n(u_c)du_c = e\int_0^{\infty} n_0(u_c)du_c + e\int_0^{\infty} \delta(u_c)du_c. \tag{75}$$

Hence

$$q_a = e\int_0^{\infty} \delta(u_c)du_c. \tag{76}$$

From (37) and (41), the contribution to the mean square of this conduction current from the frequencies between f and $f + df$ is

$$\overline{q_a^2} = 2e^2 df\int_0^{\infty} n_0(u_c)du_c = 2eI_0 df. \tag{77}$$

With this result, the effect of transit time on the shot noise current is given by

$$\overline{I_f^2} = \frac{2eI_0 df}{\left|1 + \dfrac{Z_f i\omega\epsilon}{x}\right|^2}\left\{\frac{4}{\theta^2}\left[2 + \theta^2 - 2(\cos\theta + \theta\sin\theta)\right]\right\}, \tag{78}$$

where Z_f is the impedance of the external circuit at the frequency f and $x/i\omega\epsilon$ is the capacitive reactance of the diode at the same frequency. Thus the shot noise current is modified by transit time in precisely the same manner as the noise generator voltage for the thermal tube noise.

The effect of transit time upon the shot noise, as indicated in (78), is identical with that obtained by Spenke for the same operating condition of low space charge and high anode potential.[4] Spenke derives this result through a clever application of a Fourier Series in which account was taken of the effect of transit time upon the wave shape of the current induced in the anode by the electron moving from cathode to the plate. The advantage of the method of average velocities used in this paper is that the effect of transit time in both the thermal tube noise and the shot noise may be found.

It is noteworthy that Ballantine in 1928 derived an expression for the effect of transit time upon the pure shot noise which is identical to that obtained in this paper.[9]

In conclusion, the writer wishes to express his appreciation to F. B. Llewellyn whose supervision and numerous suggestions made possible this paper.

References

1. F. B. Llewellyn, "A Study of Noise in Vacuum Tubes and Attached Circuits," *Proc. I.R.E.*, vol. 18, pp. 243–265 (1930).
2. G. L. Pearson, "Shot Effect and Thermal Agitation in a Space Charge Limited Current," *Physics*, vol. 6, pp. 6–9 (1935).
3. W. Schottky and E. Spenke, "Die Raumladungsschwächung des Schroteffektes," *Wissenschaftliche Vëroff entlichungen aus den Siemens-Werken*, vol. 16, pp. 1–41, Aug. 6, 1937.
4. E. Spenke, "Die Frequenzabhangigkiet der Schroteffektes," *Wissenschaftliche Vëroffentlichungen aus den Siemens-Werken*, vol. 16, pp. 127–136, Oct. 8, 1937.
5. L. H. Germer, "Distribution of Initial Velocities Among Thermonic Electrons," *Phys. Rev.*, vol. 25, 795 (1925).
6. T. C. Fry, "Thermonic Current Between Parallel Plane Electrodes," *Phys. Rev.*, vol. 17, 441 (1921).
7. I. Langmuir, "Effect of Space Charge and Initial Velocities," *Phys. Rev.*, vol. 21, 419 (1923).
8. F. B. Llewellyn, "Operation of Ultra High Frequency Vacuum Tubes," *B.S.T.J.*, Oct. 1935.
9. S. Ballantine, "Schrot-Effect in High Frequency Circuits," *J.F.I.*, vol. 206, 159 (Aug. 1928).
10. T. C. Fry, "The Theory of the Schroteffekt," *J.F.I.*, vol. 199, No. 2 (Feb. 1925).

Fundamentals of Teletypewriters Used in the Bell System

By E. F. WATSON

During the past few years the use of teletypewriters has become quite extensive in the Bell System. Simpler and cheaper machines have recently been made available for meeting the simpler service requirements and attachments have been designed to provide additional features for meeting more complex service requirements. This article discusses the fundamental principles and various features of the teletypewriter machines now in common use and explains the more important factors which have been controlling in their development.

WITH the growth of Teletypewriter Exchange Service and the general increase in the use of teletypewriters in private line services of various types, questions frequently asked are: How do teletypewriters operate? What is the "start-stop" system? Why is it used? What is a regenerative repeater?

This article will attempt to answer some of these questions and explain also the fundamental principles and features of teletypewriters and their auxiliary arrangements as now employed in the Bell System. These have been developed to meet the needs of customers for a typed or similar record form of communication and at the same time be suitable for operation in connection with the Bell System plant.

Code

For economical transmission over long distances it is fundamental that only a single wire or transmission channel be required to carry the signals. Furthermore, long experience with manual telegraphy on land lines has proved that reliable and efficient operation is secured by using not more than two conditions on the line, such as current and no current or positive impulses and negative impulses, as contrasted with the use of three or more conditions, or current values. The entire telegraph plant of the Bell System as well as practically all other land line telegraph systems have been built on this two-condition basis.

The familiar Morse code uses sequences of dots and dashes to represent the different characters of the alphabet and meet the above conditions. This code is not well adapted for teletypewriter control, however, since the signals for different characters vary widely in the time they require, from a single dot for the letter E to a combination of several dots and dashes for some of the less frequently used letters or numerals.

For machine operation it has thus far appeared desirable in order to obtain simplest mechanisms and to obtain maximum operating speeds with low line signaling frequencies, to have the signals for the different characters of uniform length, that is, each contain the same number of time units. This condition is met by the five-unit code where each character is identified by the impulses in five units of time, and this is the code normally employed in Bell System teletypewriters. Each of the five units of this code may be either positive or negative, current or no current, or either of two values of current, and the permutations provided are 2^5 or 32. These are sufficient for the 26 letters of the alphabet, a space, carriage return and paper feed signals as well as case shifting signals to bring another set of characters into action so as to include numerals and punctuation marks. A chart of this code as used in Teletypewriter Exchange Service (TWX), is shown below.

FIGURES	—	5/8	1/8	$	3	1/4	&	STOP	8	,	1/2	3/4	.	7/8	9	0	1	4	BELL	5	7	3/8	2	/	6	"	LINE FEED	SPACE	CAR. RET.	LETTERS	FIGURES	BLANK
LETTERS	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z						
PULSE 1	●	●		●	●	●				●	●						●		●		●		●	●	●	●				●	●	
PULSE 2	●		●				●		●	●	●	●				●	●	●			●	●	●				●			●	●	
PULSE 3			●			●		●	●		●		●	●		●	●		●		●	●		●	●			●		●		
PULSE 4		●	●	●		●	●			●	●		●	●	●			●				●		●					●	●	●	
PULSE 5		●					●	●				●	●		●	●	●			●		●	●	●	●	●				●	●	

Fig. 1—Chart of five-unit TWX code.

The keyboard used for sending this code is shown below.

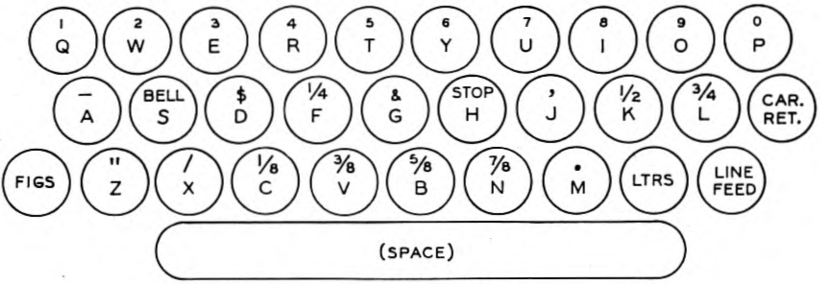


Fig. 2—Chart of TWX keyboard.

It will be noted that this keyboard is similar to the ordinary typewriter keyboard except that there are only three rows of keys instead of four as in the typewriter. In the typewriter keyboard, the lower three rows of keys are used ordinarily for small letters but when a shift key is also operated they type the corresponding capital letters. The fourth or top row of keys carries the numerals and certain punctuation marks. The teletypewriter types capital letters but not small letters so that by using a shift or *Figures* key the upper case position of the letter keys is available for the usual punctuation marks and numerals. Thus only three rows of keys are required on the teletypewriter keyboard. The operation of the *Figures* key sends a signal causing the receiving machine to shift to upper case so that numerals and punctuation marks will be printed until a *Letters* or *Space* signal is sent which restores the machine to lower case.

Start-Stop System

For transmitting the signals of the five-unit code over a telegraph line, it is necessary to have some system of timing so that each of the five impulses may be properly received, identified and interpreted at each receiving station. The start-stop system is used for this purpose. One arrangement of this system using segmented distributors with revolving brushes, is illustrated in Fig. 3.

In this system both sending and receiving brush arms are normally at rest but are maintained under constant torque, tending to rotate them in the direction of the arrows, by constantly running motors driving them through friction clutches. Normally the line circuit is closed and carries current. When a key of the keyboard is operated to send a signal, the start magnet of the sending distributor is energised releasing the sending brush arm and allowing it to rotate. As this brush passes from the stop to the start segment, the line circuit is opened and this open signal transmitted to the receiving station where it causes energization of a start magnet which releases the receiving brush and allows it to rotate.

Both sending and receiving brush arms rotate at approximately the same speeds since they are driven from motors running at approximately the same speeds. These motors are either small synchronous motors driven from constant frequency commercial 60-cycle 110-volt power supply or by commutator type motors, equipped with centrifugal governors to hold them at approximately a constant speed, for use on other commercial a-c. or d-c. supplies.

Now as the sending brush arm sweeps over the sending face, the impulses of the five-unit code, as set up by the particular key depressed, will be transmitted over the line as shown in Fig. 4 for the letter A, and through the action of the rotating receiving brush, Nos. 1 and 2 current impulses will cause the energization of Nos. 1 and 2 selecting

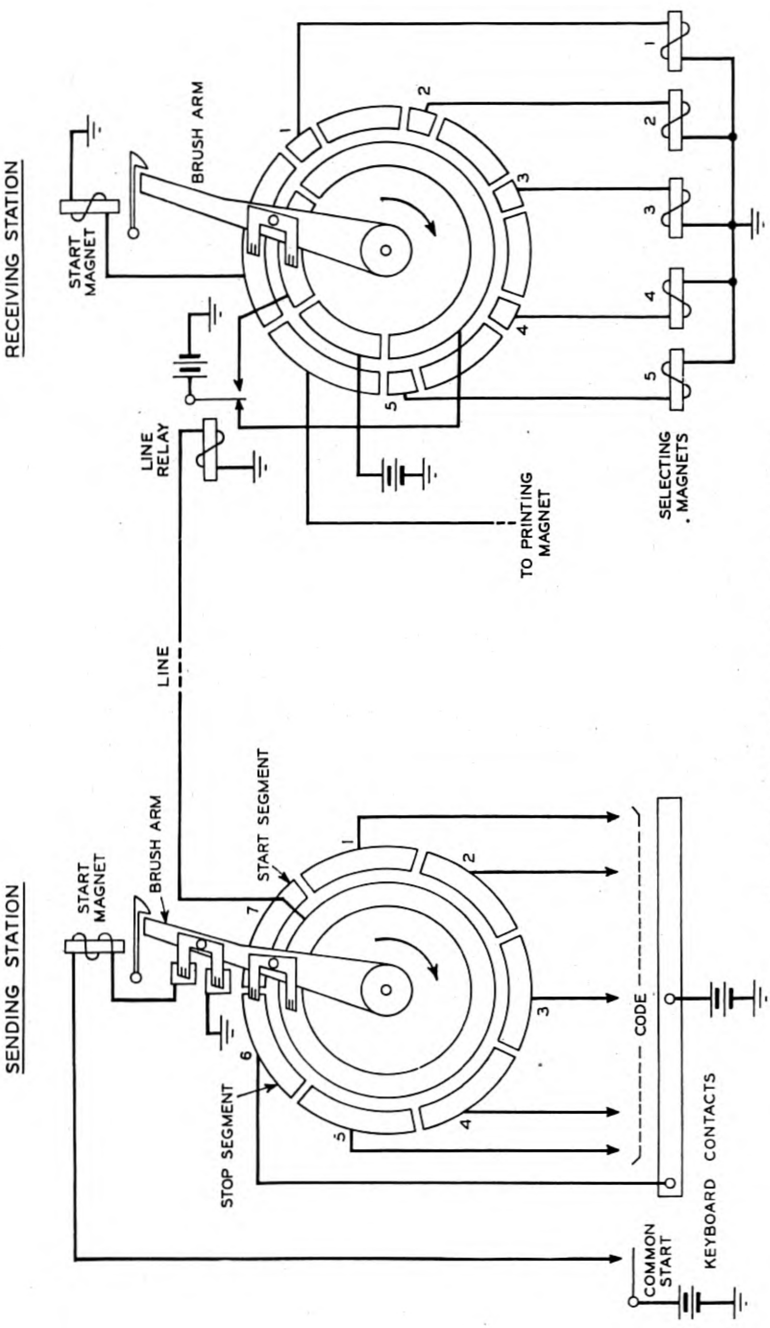


Fig. 3—Simplified diagram of start-stop system.

magnets while Nos. 3, 4 and 5 no-current impulses will not energize selecting magnets 3, 4 and 5.

Similarly other signals for other characters may be sent and received as permutations on the selecting magnets whose operation may in turn control the selection of an individual type bar to be operated to type the desired character.

While brush and segment distributors are shown above for purposes of illustration, in modern machines these are replaced by simple mechanical devices, sending contacts and a single receiving magnet, which function in the same manner but are cheaper, more reliable and easier to maintain.

Advantages of the start-stop system over other systems of timing include its simplicity, the fact that highly accurate speed regulation of the motors is not required, that to start a station it is only necessary to turn on the power, and that lag in the line signals due to time for propagation is automatically compensated for. The lag compensation is automatic because the receiving distributor does not start until the start signal has been propagated and received and then the other signals always follow in a fixed definite time relation. This makes it possible to connect any number of stations to a single circuit and have them intercommunicating; that is, signals sent from any station will be received and printed at all other stations without requiring readjustments, regardless of the distances or circuit complications due to repeaters and the like intervening.

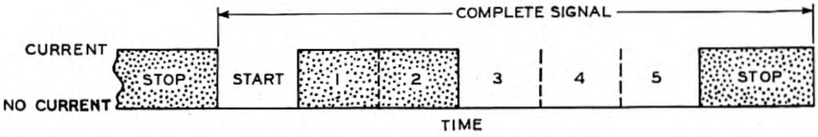


Fig. 4—Chart of line signal for letter A.

To illustrate one of these points further, it may be mentioned that with the start-stop system, if other distortions were not present, it would be theoretically possible to secure perfect operation without errors between two ideal distributors one of which was running either faster or slower than the other by as much as 7 or 8 per cent. This is because a correction is automatically made after each character is transmitted. Since some distortion of the signals is usually experienced in transmission, in order to maintain a large tolerance for such distortions, it is the usual practice to attempt to maintain the speed of each distributor within ± .75 per cent although a much larger variation may be tolerated without causing errors except when it occurs simultaneously with an abnormally large line distortion.

In contrast with this, in a synchronous system such as employed in the usual "Multiplex Printer System," it is necessary that the speeds of distributors be very accurately maintained or errors in transmission will result. This is accomplished by controlling the driving motors from very accurate timing sources such as tuning forks, then testing the speed of the receiving distributor two or more times every revolution from the signals transmitted over the line from the sending distributor, and automatically correcting this speed as required. If it were not for this last mentioned correction in order to receive signals correctly on a two-channel multiplex system for a period of 15 minutes at a speed of 60 words per minute per channel, the receiving distributor speed would have to be held accurate within ± .002 per cent. In other words, in 15 minutes the receiving distributor would make about 5400 revolutions (60 words per minute × 6 revolutions per word × 15 minutes). At the end of the 15 minutes, if it were one-tenth of a revolution ahead or behind its correct position, a No. 3 pulse, for example, would be received on a No. 4 or No. 2 segment resulting in an error. Thus the limit would be that the receiving distributor could not run fast or slow by as much as one-tenth revolution in 5400 revolutions or roughly 2 parts in 100,000.

Tolerance for Distorted Line Signals

During transmission over long lines, telegraph signals may become badly distorted; that is, the dots and dashes may be considerably shortened or lengthened from their correct values. It is essential that the receiving distributors be capable of receiving and interpreting these signals without error. To accomplish this the receiving distributors are arranged so that they are sensitive for the reception of the selecting impulses only for a very short time at the middle of each impulse. The exact location of this sensitive period with relation to the incoming signals is adjustable in each receiving distributor so that it may have maximum tolerance for receiving distorted signals. This situation is illustrated in the diagram, Fig. 5, which shows a receiving distributor with the received signals under various conditions developed, and the sensitive points for the reception of the impulses indicated by small arrows.

It will be noted from this chart that the signals may be very badly distorted (theoretically up to nearly 50 per cent of a pulse length at either the front or rear end of a current pulse), without causing imperfect reception.

In the figure it will be noted that trace (a) shows a perfect signal with the arrows denoting the sensitive points for receiving each pulse located directly at the center of each of the pulses 1 to 5. This is a normal adjustment for a receiving distributor which, without change of this adjustment, must be able to tolerate the distortions of various kinds experienced in service without failure to receive and properly identify each pulse as a No. 1, 2, 3, 4 or 5 pulse. The other traces illustrate certain types of distortions which may be experienced and the conditions existing for their proper reception with this adjustment. These traces will now be explained on a purely theoretical basis assuming an ideal receiving machine without mechanical or other imperfections.

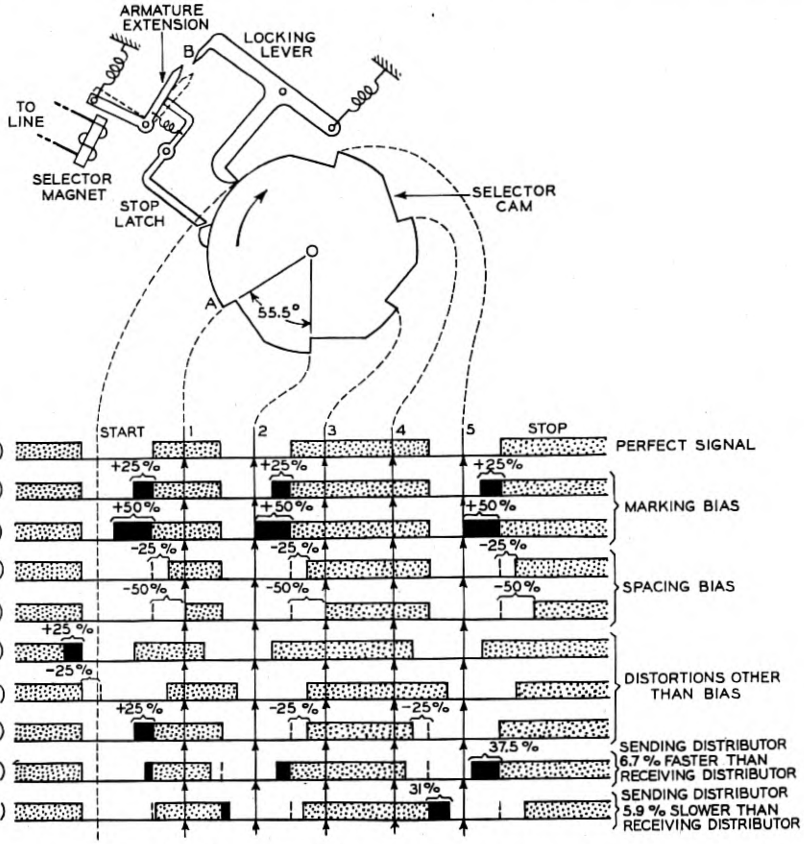


Fig. 5—Effect of distorted signals on reception.

Trace (b) shows the conditions in case of 25 per cent "marking bias" in the received signals; that is, each marking pulse has been lengthened by 25 per cent of one pulse length. It will be noted that under these conditions each pulse will still be properly received and identified.

Trace (c) illustrates 50 per cent marking bias in the received signal, and at this point it will be noted that the No. 3 and stop pulses have been so elongated that they are just on the verge of being erroneously received and identified also as No. 2 and No. 5 pulses. This then is a theoretical limit for proper operation with marking bias without a readjustment of the receiving distributor.

Similarly traces (d) and (e) illustrate respectively the conditions when the received signals have 25 per cent and 50 per cent "spacing bias," that is each marking impulse has been shortened by this percentage of one pulse length. It will be noted that 25 per cent spacing bias can be easily tolerated but that with 50 per cent spacing bias the No. 1 and No. 3 pulses are on the verge of failure to be recorded. This then is a theoretical limit for spacing bias in the signals under this adjustment.

Traces (f) and (g) show the effect of 25 per cent marking and spacing distortions respectively on the rear end of the stop or front end of the start impulse, all other pulses remaining undistorted.

Trace (h) shows the effect of distortions on the selecting impulses alone. By combining this trace with traces (f) and (g) it will be seen that with 25 per cent distortion of the start pulse, 25 per cent distortion of the same sign is the limit for distortion on the front end of marking impulses, or 25 per cent distortion of opposite sign on the rear end of marking impulses. Thus for distortions other than bias, which are apt to affect both start and selecting impulses in the same signal, ± 25 per cent is the theoretical limit of allowable distortion.

Traces (i) and (j) show the effects of speed inaccuracies. From these it will be seen that theoretically the sending distributor could be about 8.9 per cent faster or 9.5 per cent slower than the receiving distributor before errors would be experienced.

In practical machines there are, of course, inaccuracies due to tolerances of manufacture, and other departures from the ideal so that the above mentioned theoretical limits are not reached. However, all machines used in the Bell System are required to tolerate a lengthening or shortening of the front end of any current impulse of at least 40 per cent of its length and with the same adjustment a lengthening or shortening of the rear portion of any current impulse of at least 35 per cent with the start pulse undistorted. Since bias is nearly always present to some degree in the received signals, and since as interpreted by the receiving distributor it affects only the front end of the current impulses as illustrated in traces (b), (c), (d) and (e) of Fig. 5, the distributors are usually adjusted for maximum tolerance of front end distortions and then have ample tolerance for such distortions of the rear ends of the current impulses as are experienced under service conditions.

Regenerative Repeaters

As circuits become longer and more complex, eventually a point is reached beyond which signal distortion becomes so great that the signals cannot be reliably received without error. To overcome this limitation a device known as a regenerative repeater may be inserted in the line at this point. It has a receiving mechanism similar in principle to the receiving distributor of a teletypewriter and will accurately receive and interpret any signals which a teletypewriter would accurately record. This receiving mechanism is interconnected with a retransmitting mechanism, or sending distributor, which retransmits the signals reshaped and reformed so as to be substantially free from distortion. In its latest form a one-way regenerative repeater consists of a receiving magnet and a set of transmitting contacts interconnected by some relatively simple mechanical parts driven from

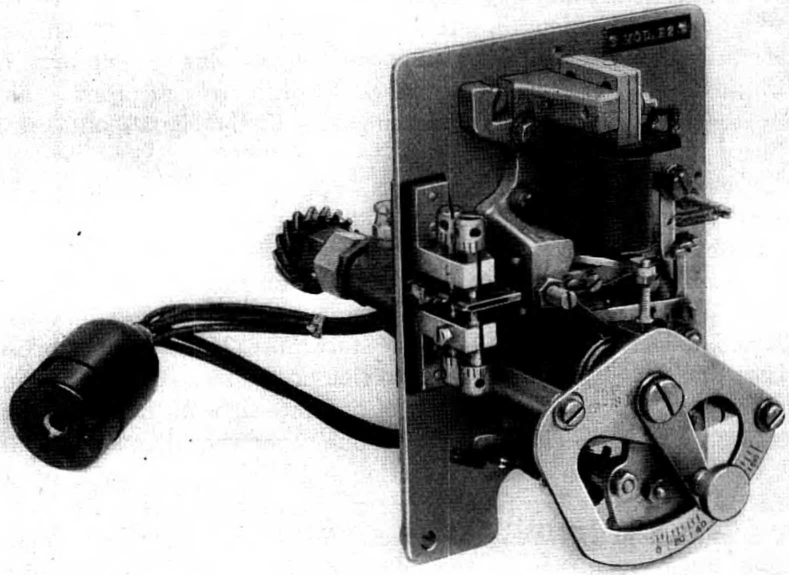


Fig. 6—Regenerator unit.

a motor. A photograph of such a recently developed regenerative repeater unit is reproduced in Fig. 6.

By means of these regenerative repeaters reliable teletypewriter service may be extended to any desired distances so long as the signals in any one regenerative repeater section are not too badly distorted to permit reliable operation of that section alone. Several regenerative repeaters may be operated in tandem on a single very long circuit if required and in fact a number of very difficult long circuits are operating satisfactorily under these conditions at the present time. A point worthy of note and which has not previously been mentioned is that the stop impulse in the code adopted for Bell System apparatus is slightly longer than the other impulses, which facilitates the use of regenerative repeaters in tandem without requiring complex speed control arrangements.

General Features of Teletypewriters

Teletypewriters are widely used for high speed written communications. Generally speaking, written communications are desired for purposes of accuracy. Therefore, high speed, accuracy and reliability are basic requirements for teletypewriter service.

In choosing an operating speed at which distributors of teletypewriters are to be set, several factors must be considered. These are the capabilities of the mechanisms of the machines, the average capabilities of operators for continuous sending at high speeds, the commercial need for high speeds, and the capabilities of the line circuits for transmitting the signals reliably over long periods without excessive distortions or excessive attention for maintenance and adjustment. A satisfactory compromise among these different factors seems at the present time to be about 60 words per minute, or 368 machine operations per minute, which is the speed usually employed in the Bell System. The machines themselves may be arranged and adjusted to be capable of higher speeds up to about 75 words per minute, and it may be that, in the future, service at these higher speeds will be justified under certain circumstances.

Accuracy and freedom from breakdown troubles are necessarily inter-related and both required to a very high degree for machines handling important written communications over long distances. To give some idea of the severity of these requirements, we have found from long experience that to produce a good machine we can not be satisfied in our laboratory tests unless the machine is capable of typing at least 1,000,000 consecutive words (6,000,000 operations) without error or trouble of any kind, and without requiring service attention of any kind other than normal replacement of paper and inking ribbons.

For rendering service economically with teletypewriters on subscribers' premises, an important requirement in order that the expense of maintenance be not prohibitive is that the machine should not require maintenance attention except at very infrequent intervals. Bell System machines are designed to require routine maintenance attention not oftener than once in two months where the machine is used continuously over periods of eight hours each day. To accomplish this the problem of lubrication has required very careful attention. It has necessitated the provision of oil reservoirs in certain places and the careful selection and specification of oils and greases. Another feature making for economical maintenance is interchangeable parts. In other words, if a part breaks or wears, it is replaceable by another part of the same type without requiring fitting and usually without readjustment.

At times customers wish to use teletypewriters on tables especially designed and arranged to suit the convenience of their offices. For this reason teletypewriters are designed as far as feasible to be self-contained units which can be mounted on any desk or table.

All present Bell System teletypewriters employ the start-stop system of synchronizing and are well adapted for the connection of any number of machines to one circuit with facilities for rapid to and fro intercommunication among the various stations. To permit optimum control of intercommunication and interruption of the sending station when desired, a device known as the "break lock" is incorporated in many machines. This device, together with a "break" key located on each machine, provides facilities whereby any station may interrupt a station which is sending, take control of the circuit and send. The operation of the "break" key opens the line transmitting a signal which causes the "break lock" device to function at the station which is sending and automatically stop any further sending from that station until the device is manually restored. This device is very important in the case of transmission from a perforated tape, which is described later.

Motor control devices are of importance for stations which are not in continuous use but which may wish to receive messages from time to time from distant stations without requiring an attendant to turn on the machine. Such devices are used both in private line and in TWX services. In the case of a private line it is often desired to have the machine normally idle with the motor stopped but so arranged that, when a distant station wishes to send a message, a signal may be sent which will automatically start the motor and condition the receiving machine so that it will properly record the message and then have its motor automatically stopped again at the end of the communication. Various devices are available for this purpose, some operating over the regular signaling circuit and others requiring a separate circuit. Similarly, in the case of TWX service, stations may, if desired, be equipped for unattended service so that, if the station is called and no attendant is present, the teletypewriter motor may be started remotely by the switchboard operator and the station conditioned to record the incoming message at the termination of which the motor can again be stopped by the switchboard operator.

Signal bells are usually provided on the machines so that, if it is desired to call an attendant to a working machine or to call attention to a specially important message being received, the bell can be rung by signals sent over the circuit.

A general feature incorporated in the design of all modern machines, and one which is not often appreciated, is the so-called "overlap." This feature makes high speed possible by overlapping the selecting and printing parts of the receiving operation. In other words it provides for the typing of one character to take place simultaneously with the reception of the selecting impulses for the next character.

Features of Page Teletypewriters

Page teletypewriters have been built in several different forms, notably with a moving paper carriage or a stationary paper carriage and with a typewheel or with type bars for printing. An early design employed a moving paper carriage and a typewheel, with an ink roller for inking the characters on the wheel. With this design it was impractical to make satisfactory carbon copies, the printed record was unevenly inked, and much trouble was experienced due to side printing, that is, unwanted printing of portions of letters adjacent to the desired letter on the typewheel. Furthermore, considerable trouble was had in properly feeding paper from a paper roll through the moving paper carriage.

To eliminate these limitations and troubles it was decided that for general service in the Bell System a new machine should be designed to be capable of making as many carbon copies as a typewriter and that it should use type bars and have a stationary paper carriage. This sort of machine was new in the art and required extensive development work to produce a satisfactory commercial design because of the inherent difficulties of moving an automatically operated basket of typebars back and forth in front of the stationary paper. The present standard No. 15 teletypewriter was the ultimate result of this work and has proved very satisfactory in general service over a number of years. It employs a typewriter ribbon for inking, has the paper roll inside the machine cover and makes very satisfactory carbon copies with various types of paper supply without being subject to the paper feed, inking and side print troubles previously experienced.

This machine has also lent itself to meeting later demands from business houses for typing either single or duplicate copies on special printed forms as commonly used in modern business practice. By equipping the platen with sprocket teeth and having feeding perforations along the edges of the forms, all copies of these forms are automatically held in perfect registration during typing at all stations connected to the circuit. In connection with the rapid handling of these forms a further requirement for automatic tabulation has been met by providing a tabulating device which on the transmission of a certain signal causes all carriages to move over rapidly to any predetermined position on the form and stop there for the typing of letters or figures in columns perfectly aligned. This device greatly facilitates the rapid transmission and reproduction of orders and the like on organized printed forms.

With the advent of TWX service a new situation arose in which many of the machines were only infrequently used and then for very short periods to make a single copy only. To render this service economically it seemed desirable to have a less expensive machine and since narrower capabilities were required this seemed entirely feasible. Accordingly a new machine known as the No. 26 teletypewriter has been developed primarily to print a single satisfactory copy although one carbon copy can be made if desired. To obtain low first cost this machine has a moving paper carriage and to secure a satisfactory printed record it employs ribbon inking and a typewheel arrangement which is a sort of cross between conventional typebar and typewheel designs. This typewheel is an assembly employing a small individual type pallet for each separate character. In the process of printing a character, a striking arm somewhat like the shank of a typebar comes forward and forces the individual type pallet against the ribbon to make an impression on the paper. The typewheel is rotated to different positions to select the different characters to be typed. In this way satisfactory inking and a clear cut impression without side print is obtained, which compares favorably with the record obtained on a typebar machine or typewriter. The entire machine costs appreciably less than the more comprehensive No. 15 machine. The No. 26 machine is illustrated in Fig. 7.

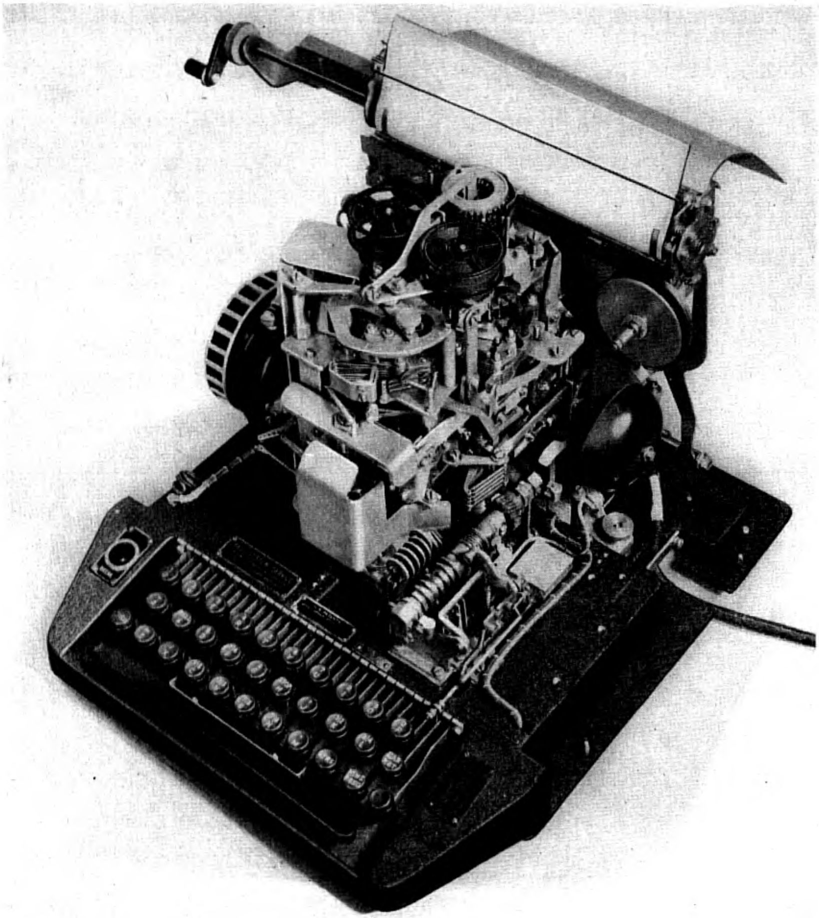


Fig. 7—No. 26 teletypewriter.

Features of Tape Teletypewriters

In the case of tape teletypewriters it is also necessary to have a clean printed record and occasionally there is a need for carbon copies. Accordingly, the tape machine standard for the Bell System is a typebar machine using an inking ribbon and known as the No. 14 teletypewriter. It is illustrated in Fig. 8.

A feature worthy of note is that with this machine typing always occurs at the same point introducing a problem in connection with platen wear. If the platen were fed by the usual ratchet in, say, 36 steps per revolution, there would be heavy wear concentrated at these 36 points and the platen would require frequent replacement to preserve good printing. To avoid this, the platen is fed through differential gearing so that on a second revolution the typing comes in a different spot from that of the first revolution; thus the wear is uniformly distributed over the entire circumference.

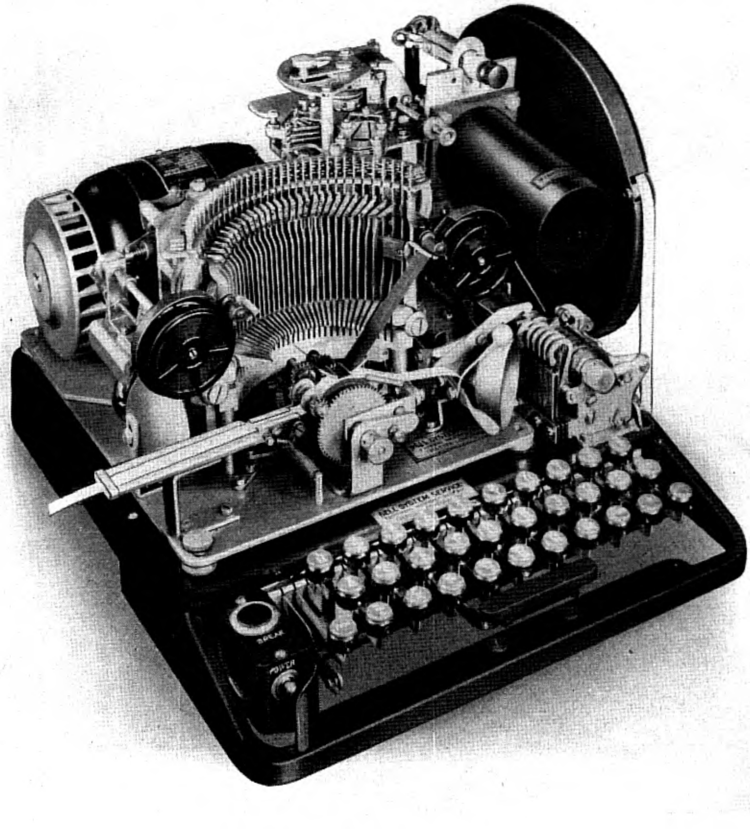


Fig. 8—No. 14 teletypewriter.

One carbon copy can be made by leading tapes through the machine from two rolls of record paper and one roll of carbon paper. Two carbon copies can be made in a similar way if desired.

The tapes employed may be either gummed on the back for convenient pasting on blanks for filing or may be plain paper tapes if the records are of temporary interest only. Also cellophane or similar transparent tape may be used if it is desired to project the record on a screen. A tape out signal is provided on the machine so that when a roll of tape becomes nearly exhausted a bell will ring continuously to give warning of this fact. Where a bell is not desired, the last few feet of tape on the roll are painted red to give similar warning.

If desired, this tape printing machine may be used on the same circuit with page printing machines such as the No. 15 teletypewriter, and when so employed is usually equipped with an "end of line indicator" to warn the operator of the approach of the end of the line in the page machine, so that suitable signals may be sent for starting a new line.

Features of TWX Switchboard Operators' Teletypewriters

Such machines must be small in size to permit their use in a switchboard position, quiet in operation to permit their use in the same room with a telephone switchboard and must be capable of working with any machine employed in the TWX system.

To meet these requirements the standard No. 14 tape teletypewriter has been modified in several important respects as follows:

1. It has been provided with a specially designed enclosing cover which reduces the machine noise radiated by at least 5 db more than standard covers.

2. The machine is tilted so as to raise the keyboard and permit the operator to assume a more elevated position nearer the switchboard jack field.

3. It is equippent with an end of line indicator mechanism and lamp to warn of the approach of the end of a line when sending to a page teletypewriter station so that the proper signals may be sent to start a new line.

4. The usual tape feeding mechanism which pulls the tape past the typing point and obscures some of the typed message is replaced by a so-called "push feed" mechanism which acts ahead of the typing point and makes the typed message more fully visible.

5. Many of the operators' machines are provided with specially arranged power supply and governing circuits so that their motors normally run from 115 volt a-c. commercial supply but in case of a power failure can be quickly switched to run from the 130 volt d-c. telegraph battery.

Features of Monitoring Teletypewriters

In connection with private wire teletypewriter service it has been found very desirable to have so-called monitoring teletypewriters in the repeater offices to facilitate testing between offices and with the subscriber stations. These machines must be adaptable to work with any subscriber's machine and to be usable for making test measurements on circuits.

The No. 14 tape teletypewriter has also been adapted to this service. It may be equipped with an end of line indicator to facilitate communication with a page teletypewriter. Also since commercial service is given at speeds of 40 and 60 words per minute, many of the monitoring machines are equipped with two-speed governors and a switch to provide for changing from one speed to the other. These machines are also usually arranged for normal operation from commercial power supply but emergency operation from the 130-volt telegraph battery.

For making test measurements over circuits a special orientation scale is provided together with a small crank extending through the cover for quickly shifting the orientation setting to any desired point. With the machine carefully adjusted to be practically free of harmful distorting effects on the signals, it may then be used for measuring distortions in received teletypewriter signals, the scale being arranged to read the total distortion directly in percent of a pulse length.

Tape Storage Transmission

A heavy volume of traffic may be transmitted rapidly and conveniently by the use of perforated tape. In this method a machine known as a perforator and having a keyboard like that of the teletypewriter is used for punching the code signals for the message in a strip of paper tape. This may be done with simultaneous typing of the message on the teletypewriter in which case the speed of perforating is limited to the speed for which the teletypewriter is set. If a typed record is not made simultaneously with the perforating, punched tape may be prepared at practically any speed within the capabilities of the operator. This punched tape may then be fed through a device known as a tape transmitter which automatically transmits the message signals from the tape at the maximum speed for which the teletypewriters connected to the circuit are set, which is usually 60 words per minute.

The method of transmitting from perforated tape has the distinct advantage of using the line at maximum efficiency at all times as compared with direct keyboard sending where pauses in operating the keys and interruptions to the operator result in direct losses of circuit time and effectively slower transmission.

Another important advantage of the perforated tape method is that errors may be corrected in the tape before transmission with the result that only errorless copy is transmitted on the circuit. This is done in the following manner and is illustrated in the section of perforated tape shown below. If the operator in attempting to write the word THE should strike the keys T and J (in error), realizing her error she back spaces the tape one division, strikes the "letters" key and then the correct keys H and E. The transmission of the "letters" signal will cause no operation in the recording teletypewriters since they are already in the "letters" case, and the word will be recorded correctly as though no error had been made. Similarly, entire words or groups of characters may be erased from the tape if desired.

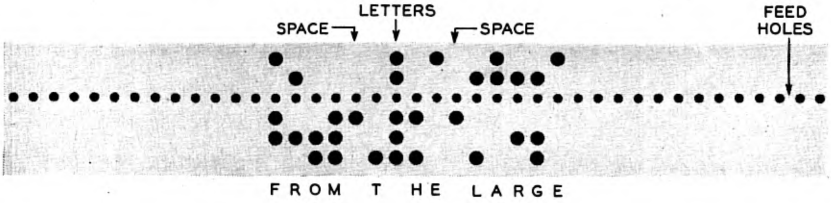


Fig. 9—Sample of strip of perforated tape.

For TWX service a further advantage of the perforated tape method is that the entire message may be punched in tape and checked by printing, if desired, before a call is placed and a connection established. Then, when the connection is established, the message can be automatically transmitted at maximum speed requiring a minimum time for the connection and giving a minimum charge for the call.

It is true, of course, that in this method there is some delay between perforation and transmission. For this reason short to and fro messages, as required in setting up a connection, may be better handled by direct keyboard. To facilitate such working, the perforator keyboard is normally arranged so that by throwing a switch this same keyboard may be used for direct keyboard sending without perforating. This switch also has an intermediate position in which the keyboard is connected for simultaneous direct sending and perforating. This provides for meeting the needs of certain TWX subscribers who wish to simultaneously type and punch the message so that the typed copy may be checked as it is perforated.

The complete page printing set arranged for tape transmission is known as the No. 19 teletypewriter set and is illustrated in Fig. 10. It employs a No. 15 teletypewriter as the page printing unit.

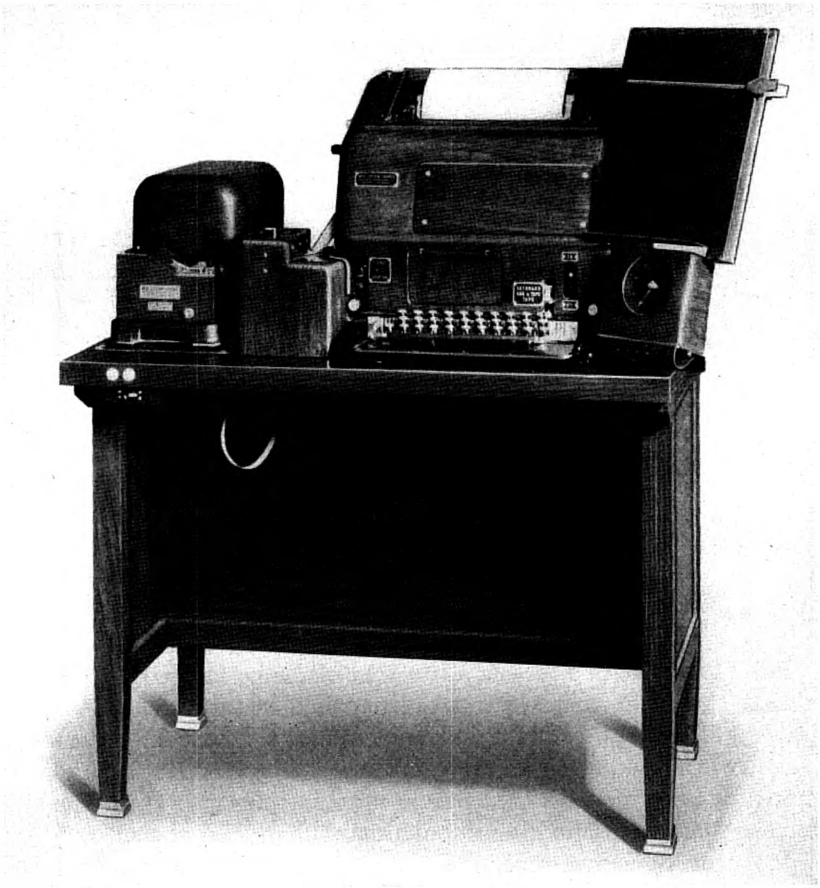


Fig. 10—No. 19 teletypewriter set.

Automatic Retransmission Using Reperforators

At times it is desirable to retransmit messages received from one circuit to some other machine or machines on a separate circuit. A unit known as a "reperforator" is often used to facilitate such retransmission. The reperforator now standard for the Bell System is a start-stop receiving device using the 5-unit permutation code. It is somewhat similar to the receiving-only tape teletypewriter except that the record produced consists of code perforations in a tape rather than typing on a tape. This perforated tape is the same as tape produced by a keyboard perforator as previously described, and may be used in an automatic transmitter for retransmitting the message on a separate circuit. The reperforator is usually associated with a receiving teletypewriter on a circuit and may be cut in or out manually or automatically from signals transmitted along with the message signals, so that it will automatically reproduce a code tape for use in automatically retransmitting such messages as desired on some new connection.

Conclusion

The fundamentals of teletypewriters, as described above, now seem to be fairly well established. The future should bring simpler and cheaper machines, especially where the more difficult requirements do not have to be met, and probably additional attachments and auxiliary features to extend the applications and convenience of operation.

The Dielectric Properties of Insulating Materials

By E. J. MURPHY and S. O. MORGAN

This article discusses the variation of dielectric constant and dielectric loss in the radio and power frequency range with the object of giving a simple picture of the type of mechanism which is able to produce anomalous dispersion in this range of frequencies. Some of the general characteristics of anomalous dispersion can be demonstrated as well on a simple and arbitrary model of the structure of dielectrics as on the more complex ones which correspond more closely to the actual structure of dielectrics. Such a derivation is given here in order to indicate the significance of the different factors which occur in the formulæ which have been proposed to account for the variation of dielectric constant and dielectric loss with frequency. This enables a distinction to be made conveniently between the general characteristics which are shared by several types of dielectric polarization and the special characteristics which are peculiar to a restricted class of polarizations or to a particular kind of polarization.

II. Dielectric Polarizability and Anomalous Dispersion

IN a previous paper[1] the general features of the dependence of dielectric constant on frequency were indicated schematically for the entire range extending from the frequencies used in power transmission to those of ultra-violet light. In the range of frequencies below the infra-red (that is, in the electrical range of frequencies) anomalous dispersion is the rule, normal dispersion not having been observed as yet, except for piezo-electric materials, whereas at high optical frequencies normal dispersion is the predominant feature. In the intermediate infra-red region it is not surprising to find a behavior which shows anomalous and normal dispersion in more nearly equal degrees of prominence.

It will be recalled that anomalous dispersion is the type of frequency-variation in which the dielectric constant decreases with increasing frequency, while normal dispersion is the reverse of this, the dielectric constant or refractive index increasing as the frequency increases. The use of the term *anomalous dispersion* to describe the dependence of dielectric constant on frequency in the radio and power frequency range is now widespread, and seems quite appropriate, for it brings out the point that the variation of dielectric constant with frequency in

[1] Murphy and Morgan, *B. S. T. J.*, *16*, 493 (1937).

the radio and power range is in certain respects the same type of phenomenon as optical anomalous dispersion.

Anomalous dispersion plays a very important part in the behavior of dielectrics in the electrical range of frequencies. It is seldom possible to interpret a set of measurements of dielectric constant or other dielectric properties without encountering some manifestation of anomalous dispersion or of the other characteristic types of behavior which follow as corollaries of it.

The two catagories, polarizability and dispersion, include a great deal of the dielectric behavior of insulating materials. This paper will deal primarily with anomalous dispersion, but the theory of anomalous dispersion is not entirely separable from that of the polarizations of which it is an attribute, so it will be necessary to discuss at least briefly the nature of dielectric polarization.

The Relation between Polarizability and Dielectric Constant

For our purposes a dielectric may be thought of as an assemblage of *bound charges*, where this term is intended to include the electrons and positive cores in atoms and molecules, the ions held at lattice points in ionic crystals and, in general, any assemblage of charged particles which are so bound together that they are not able to drift from one electrode to the other under the action of an applied electric field of uniform intensity. Actual dielectrics, of course, also contain some conduction electrons or ions which are free to drift through the material and discharge at the electrodes, producing a direct current conductivity. This conductivity is small at ordinary temperatures in materials classified as dielectrics.

The positions of these charged particles may be considered to be determined by an equilibrium of forces. When an electric field is applied this equilibrium is disturbed and the bound charges are displaced to new positions of equilibrium; then when the applied field is removed they revert to their initial positions. In the equilibrium positions which the charges occupy when a constant electric field has been impressed on the dielectric they have a larger potential energy than in their initial positions. Moreover, they do not revert instantly to their initial positions, and when the retardation is due to friction some of the potential energy of the bound charges is dissipated as heat in the dielectric.

When an alternating voltage is applied to the dielectric, we may think of the bound charges as moving back and forth with certain amplitudes, a different amplitude for each different type of bound charge. When the applied electric field is of unit intensity, the sum of the product of amplitude and charge extended over all of the *bound* charges in a unit volume of the material determines the *dielectric constant* of the material. The energy dissipated as heat by the motions of these bound charges in the applied electric field represents the *dielectric loss* per second, a quantity which is proportional to the a.-c. conductivity after the d.-c. conductivity has been subtracted from it. The imaginary part of the complex dielectric constant is proportional to the dielectric loss per cycle.

While the physical meaning of the dielectric constant and dielectric loss can be conveniently described, as above, in terms of the amplitudes and energy relationships of bound charges in their motions in an applied electric field, a more useful basis for the discussion is that provided by the concept of polarizability. In the present application the polarizability is equivalent to the product of charge and amplitude, but it has the advantage of being a quantity which is defined and discussed in the general theory of electricity as well as in that of dielectrics. The dielectric constant is then found to be related closely to the polarizabilities of the assemblages of charged particles which the dielectric contains.

The polarization of an assemblage of charges is a quantity defined in electrostatic theory as the vector sum

$$\mathbf{p} = \sum e_i \mathbf{s}_i, \tag{1}$$

where $\mathbf{s}_i$ is the distance of the i^{th} charge, e_i, from a point chosen as origin, and the summation is extended over all of the charges in the assemblage, for which e_i is a typical charge. (If the assemblage has no net charge ($\sum e_i = 0$), the origin may be arbitrarily located without affecting the value of $\mathbf{p}$.)

The polarization is a vector quantity. It can be written as the product of a scalar quantity p, which represents the magnitude or electric moment of the polarization and a unit vector $\mathbf{p}_1$ which gives the direction of the polarization; thus $\mathbf{p} = p\mathbf{p}_1$. As it will not be necessary to distinguish between the properties of isotropic and anisotropic materials in this article the direction of the polarization need not be emphasized. The notation will therefore be simplified, in general, by using the magnitude or scalar part of such vector quantities as the polarization, the electric field intensity and the displacement of charged particles.

To illustrate the application of equation (1) let us consider a very simple configuration consisting of two charges $+ e$ and $- e$ (see Fig. 1). The vector polarization of this configuration is $\mathbf{p} = e(\mathbf{s}_1 - \mathbf{s}_2) = p\mathbf{p}_1$, where p is the magnitude or electric moment of the polarization and

$\mathbf{p}_1$ is a unit vector in the direction of the vector $(\mathbf{s}_1 - \mathbf{s}_2)$. If now one of these charges is an electron ($e = 4.77 \times 10^{-10}$ e.s.u.) and the other a unit positive charge and they are separated by a distance of the order of magnitude of atomic distances (10^{-8} cm.), p has the value 4.77×10^{-18} e.s.u., or 4.77 Debye units. The permanent electric moments of molecules seldom exceed a few Debye units.

Let us now apply the definition contained in equation (1) to a dielectric material. In the first place it indicates that if we know the effective positions of the electrons and other charged particles which contribute to the structure of the material we can always, in principle, calculate the polarization of the body as a whole or any part of it. Actually the calculation of the polarization of a body as a whole or that of unit volume in it is in general a complicated matter involving statistical considerations, but there are special cases in which the result is rather obvious. For example, in a gas or liquid if all orientations of the molecules are equally probable in the absence of an applied field, the value obtained by taking the time-average of the summation indicated by (1) is zero. Equation (1) would also give the value zero when applied to all of the ions in a c.c. of a solution because any arbitrarily chosen small volume in the liquid would be as likely to contain a positive ion as a negative ion.

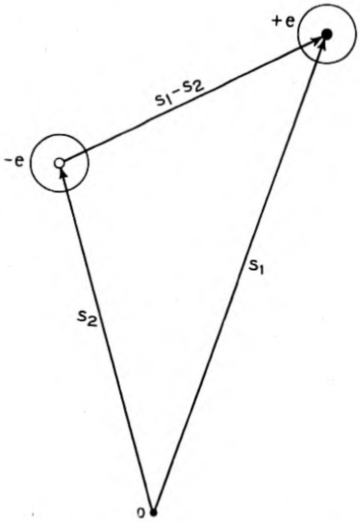


Fig. 1—The calculation of the polarization vector by the general method for a very simple configuration.

In some crystalline materials equation (1) gives the value zero because there is a suitable symmetry in the configuration of charged particles in the unit cell; for other solids equation (1) gives a finite value for the unit cell, but zero when applied to a volume of the material large enough to contain a great many crystallites with random orientations; however, there are some macroscopic crystals which have permanent polarizations. A solid material consisting of polar crystallites with random orientations is analogous, as far as equation (1) is concerned, to a liquid or gas containing polar molecules having random orientations; the polarization of the material as a whole is zero in either case.

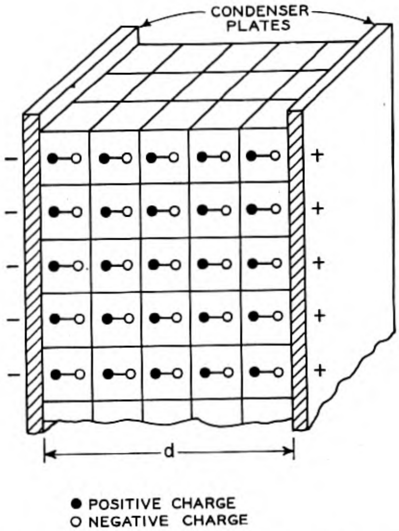


Fig. 2—A dielectric in a condenser. The circles joined by a bar represent "bound charges" of various kinds, including atoms and molecules.

Let us now consider a dielectric of any kind occupying the space between two plane, parallel condenser plates of great enough area and small enough separation that the electric field between the plates when they are charged may be considered to be directed normally to them (cf. Fig. 2). Consider the space between the plates of the condenser to be divided into small cubes of the same size, the purpose of this imaginary division of the dielectric being merely to obtain a representative specimen of the dielectric material. If the cube size is too small the instantaneous value of p obtained by applying equation (1) to all of the particles in a cube will vary appreciably from one cube to another;

but we can then increase the size of the cubes until p is the same for each cube to a close enough approximation. The polarization in each cube is then representative of that of the dielectric as a whole,[2] and by dividing $\sum e_i s_i$ for a typical cube by the volume of the cube we obtain the *polarization per unit volume*, which for the present will be designated as P. This quantity is a statistical mean value involving a summation over a large number of particles; its value depends not only on the structure of the material but upon the effect of thermal motions on the mean positions and orientations of the molecules or other elementary particles in the material. One of the most interesting points in dielectric theory is the consideration—pointed out by Debye and at the basis of his theory of polar molecules—that for some types of structure the mean positions of the particles from which P is calculated are unaffected by changes in the amplitude of thermal motions while for another type of structure (consisting of polar molecules free to assume many or at least several orientations) an increase of temperature decreases P, because the randomness of the orientations of the polar molecules is increased.

For many materials P is zero when no electric field is applied, and assumes a finite value only when an electric field is applied, though as has been indicated, some crystalline materials have a finite value of P even in the absence of an applied electric field. In either case, however, the application of an electric field causes the bound charges within the dielectric to be shifted in general to new equilibrium positions, corresponding to the slight change in the system of forces acting upon them, and if the material did not have a polarization before the application of the field, it assumes one; if it did, it assumes a different value of P. The value of P when an electric field E is applied will be designated as P_E, and that when no field is applied by P_0. Then $P_E - P_0$ is the polarization per unit volume induced by an applied field E. As the dielectric constant of a material depends upon the magnitude of the polarization induced in it by an applied field, and we are concerned here with dielectric constants, it will be desirable to simplify the notation by setting $P_E - P_0 = P$. This gives P a slightly different meaning than it had in the earlier part of the discussion, where it represented the total polarization per unit volume whatever its origin.

[2] A detailed consideration of the method of dividing a dielectric up into elementary volumes in order to compute the mean polarization encounters complications which need not be discussed here. A critical analysis of the method of computing the volume density of polarization of a dielectric is given by Mason and Weaver, "The Electromagnetic Field," Chicago (1929); Chapter III.

The relation between the applied electric field, E, and the polarization induced by it per unit volume is given by

$$P = \frac{\epsilon - 1}{4\pi} E \tag{2}$$

for isotropic materials. The constant $(\epsilon - 1)/4\pi$ is the *susceptibility* of the dielectric in e.s.u., and ϵ is the dielectric constant, which is defined as C/C_0, where C is the capacitance of the measuring condenser while it contains the dielectric and C_0 is its capacitance when empty.

For some purposes there are advantages in considering the actual polarization, which is produced by a discontinuous distribution of charged particles, to be replaced by a vector point function which gives equivalent external effects. Then a vector **P** may be considered to be associated with every point in the space occupied by the dielectric and the dielectric may be considered to have a continuous volume density of polarization,[3] **P**. In non-isotropic bodies the polarization vector **P** induced by an applied field **E** is not always in the same direction as **E**, but is assumed to be a linear vector function [4] of **E** (involving, in the general case, six independent constants), where both **E** and **P** are vector point functions.

In deriving the relationship between the dielectric constant and the molecular structure of a material it cannot be assumed in general that the local field which is impressed upon the elementary particles in the dielectric is simply the field E which can be computed by dividing the applied voltage V by the distance between the plates of the condenser, the intensity of the field being assumed to be uniform. For there is an interaction between the molecules of the dielectric such that each molecule exerts a force on every other molecule. In the absence of an applied electric field these forces combine with other influences to create a distribution for which the polarization per unit volume has the value P_0 (frequently zero, as has been mentioned). Then when a field is applied each element of volume in the dielectric is put into a polarized condition and in general the forces which it exerts upon the particles in other volume elements changes, because the charges in each volume element have been displaced to new positions. Consequently, the value assumed by P in a given cube of Fig. 2 will depend not only upon the direct action of the charges on the plates of the condenser—which determines the strength of the field E—but also

[3] Cf. Mason and Weaver, loc. cit. Chap. III.

[4] Cf. P. Debye, "Polar Molecules," Chemical Catalogue Co., New York (1929), pp. 32–35.

upon their indirect action through the polarization which they create in other elements of volume.

The contribution which the polarization of the dielectric makes to the force upon a charged particle in it has been calculated by Lorentz to be $(4\pi/3)P$, where P is the polarization per unit volume induced by the applied field. This calculation applies to an array of particles with cubic symmetry and to isotropic materials.[5] The *internal or local field* F is then given by

$$F = E + \frac{4\pi}{3} P. \tag{3}$$

E may be thought of as the force which has its origin in the direct interaction between the charges on the plates of the condenser and the charges in the polarizable complex on which attention has been fixed (such as one of the cubes of Fig. 2), while the term $(4\pi/3)P$ may be regarded as an indirect force coming from the other parts of the dielectric by virtue of their polarized state.

It is assumed in the theory of dielectrics that the structure of materials is such that P is a linear function of F (or a linear vector function in the case of anisotropic materials); then

$$P = kF, \tag{4}$$

where k is the polarizability per unit volume. It can be seen that

$$F = \frac{E}{1 - Ak}, \tag{4a}$$

where $A = 4\pi/3$, and consequently that the relation between the polarizability k and the susceptibility $(\epsilon - 1)/4\pi$ $(\equiv K)$ is

$$K \equiv \frac{\epsilon - 1}{4\pi} = \frac{k}{1 - Ak}, \tag{4b}$$

whenever (3) is a valid expression for the internal field.

The susceptibility can be calculated without presupposing the validity of equation (3) for the internal field, *while the value of k depends upon whether (3) or some other expression gives the strength of the internal field in the dielectric.*

If L is the number of molecules per cubic centimeter, $k/L(\equiv \alpha)$ is the polarizability per molecule. This molecular constant α is called the polarizability of the molecule. By multiplying α by Avogadro's number N, we obtain the polarizability per mole of the dielectric:

[5] H. A. Lorentz, "The Theory of Electrons," p. 138, and Notes 54 and 55.

$N\alpha = Nk/L$. And if m is the mass of a molecule, $Nm = M$, where M is the molecular weight, and $Lm = \rho$, where ρ is the density; so that

$$\frac{N}{L} = \frac{M}{\rho}$$

and the polarizability per mol may be written as Mk/ρ.

From equations (3) and (4) (or 4*b*) it can be shown that the polarizability is related to the dielectric constant by the familiar relation

$$k = \frac{3}{4\pi}\left[\frac{\epsilon - 1}{\epsilon + 2}\right], \tag{5}$$

which however is only valid when (3) is valid—and for some materials (3) is apparently not valid.

For gases the term $(4\pi/3)P$ in (3) is so small as compared with E that F is approximately equal to E and

$$k = K = \frac{\epsilon - 1}{4\pi}. \tag{6}$$

The polarizability and susceptibility are then equal. The physical reason for this is that the ratio of intermolecular space to the space occupied by molecules is much larger in a gas than in a solid or liquid and the direct force exerted by the charges on the condenser plates on a charged particle in the dielectric is then much greater than the indirect force which they exert through the polarization induced in other molecules.

It is customary to call the quantity $(4\pi/3)k$ the volume polarization, and it is often denoted by the letter p. The volume polarization may be thought of as $4\pi/3$ times the polarization induced in the dielectric per unit volume per unit applied field. The convenience of using $(4\pi/3)k$ instead of k comes from the occurrence of the factor $4\pi/3$ in the relation (5) between dielectric constant and polarizability.

On dividing equation (5) by the density we obtain a quantity which is called the mass polarization, as it is $4\pi/3$ times the polarizability per gram:

$$\frac{1}{\rho}\left[\frac{\epsilon - 1}{\epsilon + 2}\right] = \frac{4\pi}{3}\frac{k}{\rho}. \tag{7}$$

And on multiplying (7) by the molecular weight of the material we obtain

$$\frac{M}{\rho}\left[\frac{\epsilon - 1}{\epsilon + 2}\right] = \frac{4\pi}{3}\cdot\frac{M}{\rho}\cdot k = \frac{4\pi}{3}\frac{kN}{L} = \frac{4\pi}{3}N\alpha. \tag{8}$$

The quantity $(4\pi/3)N\alpha$ is the *molar polarization*, $N\alpha$ being the polarizability per mole.[6]

Equation (8), and also (7), expresses the Clausius-Mosotti relation when α is considered to be a constant characteristic of the individual molecule and independent of density. The function of ϵ on the left-hand side of (8) is independent of density whenever α is independent of density.

The following relation, analogous to that of Clausius and Mosotti but expressed in terms of the refractive index n, was derived by Lorentz and by Lorenz:

$$\frac{M}{\rho}\frac{n^2-1}{n^2+2}=\frac{4\pi}{3}N\alpha. \qquad (8a)$$

The left-hand member of this equation is called the *molar refraction*. Equations (8) and (8*a*) are equivalent because of the general relation between refractive index and dielectric constant ($n^2 = \epsilon$), but owing to the fact that refractive indices are measured at optical frequencies the molar refraction contains only the electronic part of the total molar polarization of the material. Subtracting the molar refraction from the total molar polarization, is one of the methods of determining the amount of polarization contributed by non-electronic polarizations.

It has been found that the Clausius-Mosotti relation is not equally satisfactory for all kinds of dielectric polarization. It gives good results when applied to electronic and atomic polarizations. For example, in an interesting paper on materials of high dielectric constant, Frank [7] has recently shown that the Clausius-Mosotti-Lorentz-Lorenz relationship aids materially in explaining the behavior of the dielectric constants of crystalline materials of high dielectric constant where the dielectric constant depends upon electronic polarizations. Where the polarizability of a molecule is the sum of the polarizabilities of the atoms of which it is composed it is to be expected that if the relation (5), or (8) or (8*a*) is valid the sum of the atomic polarizations would be equal to the molar polarization. Experimental agreement

[6] The polarizabilities of non-polar molecules and atoms are usually of the order of magnitude of 10^{-24} c.c., and the molar polarizations of such substances, consequently, are of the order of magnitude of a few c.c., since the molar polarization is $(4\pi/3) \times 6.06 \times 10^{23}$ times the polarizability of the individual molecule. The polarizability of a conducting sphere is equal to the cube of its radius. And, as atomic dimensions are of the order of magnitude of 10^{-8} cm., it is evident that the polarizabilities of atoms tend to be of a similar order of magnitude to the polarizabilities which would be expected if they behaved as conducting spheres, though there are large differences in the ratio of polarizability to volume for different atoms. The molar polarizations of polar molecules are in general larger than those of similar non-polar molecules and may be a few hundred c.c. (Cf. P. Debye, "Polar Molecules," pp. 12–19.)

[7] F. C. Frank, *Trans. Faraday Society*, *23*, (4), 513 (1937).

with this requirement has been found in optics where the refractive indices of molecules can be calculated approximately from the molar refraction (eq. 8*a*) obtained by adding the atomic refractions.

This additive property of electronic polarizations has been employed by Frank [8] to interpret the tendency of crystalline materials having high dielectric constants to be characterized by a high polarizability/volume ratio for the atoms or ions of which they are composed. This condition would tend to allow the largest number of highly polarizable particles to be concentrated in a given space, giving, on the additivity rule, a high molar polarization and a high dielectric constant.

On the other hand Wyman [9] has pointed out that the Clausius-Mosotti relation is not satisfactory when applied to *highly polar liquids*, such as water, and has found that for these substances it appears to be more satisfactory to consider that the polarization is related to the dielectric constant by the empirical relation

$$\frac{\epsilon + 1}{8.5} = \frac{4\pi}{3} k. \tag{9}$$

The calculation of the internal field by Lorentz, which provides the theoretical basis for equation (8), was made before the theory of polar molecules had been developed, but equation (8) has since been applied tentatively to polar molecules.[10] The problem of obtaining an improved relationship between polarizability and dielectric constant for materials having molecules with permanent electric moments has been studied in recent years by several investigators.[11] The calculation of the internal field usually involves the assumption that the effect of the molecules included in a small sphere surrounding the central molecule on which the force is being calculated is negligible on the average because of the random motions due to thermal agitation. On the supposition that such an assumption is not justified in a polar material because of the interactions of adjacent polar molecules, Onsager [12] has obtained a relation between polarizability and dielectric constant which for high dielectric constants is nearly the same as Wyman's empirical relation, equation (9). A comprehensive study of the effects of interaction between the dipoles of polar molecules has

[8] Loc. cit.

[9] Cf. Wyman, *Jour. Amer. Chem. Soc.*, *56*, 539 (1934); *58*, 1482 (1936).

[10] Cf. Debye, loc. cit., p. 13.

[11] Cf. Onsager, *Jour. Amer. Chem. Soc.*, *58*, 1486 (1936); Van Arkel and Snoek, *Trans. Faraday Soc.*, *30*, 707 (1934); Wyman, *Jour. Amer. Chem. Soc.*, *58*, 1482 (1936); Van Vleck, *Jour. Chem. Physics*, *5*, 320 (1937) and *5*, 556 (1937).

[12] Loc. cit.

been made by Van Vleck by the methods of statistical mechanics. He obtains an expression which agrees to a second approximation with that obtained by Onsager. Thus it seems that for highly polar liquids the relations between polarization and dielectric constant developed by Onsager, Wyman and Van Vleck may be more satisfactory than the Clausius-Mosotti relationship, though for many other materials the Clausius-Mosotti relationship is apparently valid or approximately valid.

In deriving expressions for the dependence of dielectric constant on frequency later in this article the formulæ obtained will naturally depend upon which of the equations, (5), (6) or (9), is taken as the relationship between polarizability and dielectric constant. The alternative expressions will be listed.

Derivation of a Dispersion Formula

The above-described relations between polarization and dielectric constant provide the means of obtaining expressions for the variation of dielectric constant with frequency when we have determined the dependence of polarizability on frequency. As our object is to exhibit the general features of anomalous dispersion shared by several particular types of polarization, it will be sufficient to derive dispersion formulæ containing constants the values of which are not specified, but which have a sufficiently obvious physical significance. The derivation given will parallel that of Lorentz in deriving a formula for optical dispersion,[13] and in fact is simply a special case of it in which certain terms are considered to be negligible by comparison with others.

An analogous procedure was used in one of the earliest attempts to explain anomalous dispersion in the electrical frequency range, the theory proposed by Drude [14] in 1898. This theory was based upon the hypothesis that anomalous dispersion in the electrical frequency range depends upon a mechanism similar to that to which optical dispersion was attributed, the difference being that the particles which produce anomalous dispersion in the electrical frequency range are so large that some of the terms in the optical dispersion formula can be neglected. The formulæ which Drude derived for electrical anomalous dispersion yield the same form of variation of dielectric constant with frequency as do the generally accepted theories of the present time, such as the Debye theory; the differences lie in the expressions given for the con-

[13] H. A. Lorentz, "The Theory of Electrons," Chapter IV. See also Korff and Breit, *Reviews of Modern Physics*, *4*, 471 (1932), where a review of the classical theory of optical dispersion is given.

[14] P. Drude, *Ann. d. Physik*, *64*, 131 (1898), "Zur Theorie der anomalien elektrischen Dispersion."

stants in the formulæ in terms of properties of the material. Another adaptation of optical dispersion theory to the explanation of dispersion in the electrical frequency range was proposed by Décombé [15] in 1912. He employed the Lorentz electron theory for the dispersion of light as a basis for the consideration that if the environment of some of the electrons in dielectrics is suitable their motions in an applied field could produce anomalous dispersion and dielectric loss in the electric frequency range. A similar simple and arbitrary assumption regarding the structure of dielectrics will also be employed here. However, it is not proposed as a theory of dielectric behavior but merely employed as a comparatively simple means of deriving and discussing relationships which can be demonstrated as well on a simple and arbitrary model as on the more complex ones which correspond more closely to the actual structure of dielectrics. The relation of the constants in the dispersion formulæ which will be derived here to the actual structure of dielectrics will only be indicated in a general qualitative way for the purpose of illustrating the physical nature of the processes involved; no attempt will be made to provide expressions for the dispersion constants in terms of other observable properties of the material.

In Fig. 2, let the applied potential be V, where V may vary in general in any way with the time, though in the present discussion it will be considered to vary sinusoidally with the time; the impressed field strength is then given by $E = V/d$. As in the more general discussion which preceded this, it will be assumed that the imaginary cells pictured in Fig. 2 contain large numbers of polarizable complexes consisting of positive and negative charges in equal numbers held in position by constitutive forces—the origin of which need not be specified for our present purposes—such that if they are displaced a distance s from their initial positions they will experience a force fs, where f is a constant, tending to restore them to their initial positions; and that while these charges are in motion as a result of the action of the impressed field they experience a frictional force $r\dot{s}$, where r is a constant and $\dot{s}$ is the velocity in the direction of the impressed field; and, finally that their motion is also retarded by an inertia reaction $m\ddot{s}$, proportional to the mass m and the acceleration $\ddot{s}$ of the particles.

The equation of motion for any typical charge e in a polarizable complex having the above-described specifications is

$$m\ddot{s} + r\dot{s} + fs = eF, \tag{10}$$

where F is given by equation (3) in materials to which the Lorentz calculation of the internal field applies, by $F \cong E$ in the case of gases

[15] L. Décombé, *Journal de Physique*, (5), *3*, 315 (1912).

and by other expressions—which in some cases may approximate either to $F = E$ or to $F = E + (4\pi/3)P$—for still other materials. The quantities F and s are vectors, but for isotropic materials s is in the same direction as F.

If, following the method employed by Lorentz, we write an equation of the form (10) for each charged particle in a physically small volume δ (such as the cubes of Fig. 2), multiply each equation by e, add the equations for all of the particles in δ, and divide by the volume δ, we obtain

$$m\ddot{P} + r\dot{P} + fP = ne^2F, \tag{11}$$

where $P \equiv (1/\delta)\sum es$ and n is the number of charged particles characterized by the constants m, r and f per unit volume. The volume δ may be considered to be that of one of the cubes in Fig. 2. As indicated earlier it should contain a sufficient number of molecules to give a good mean value for P, the polarization per unit volume, but at the same time it should be small enough not to mask significant spatial variations in P.

When the impressed field E is varying sinusoidally with the time at the frequency $\omega/2\pi$, the local or internal field F tending to displace each charged particle in the dielectric will also vary sinusoidally with the time, though in general out of phase with E, if F is given by equation (3), and can be considered to be given by the real part of $F_0e^{i\omega t}$. Under these conditions

$$P = kF_0e^{i\omega t}$$

is a solution of equation (10) for the steady state provided that

$$k = \frac{ne^2}{(ir\omega - m\omega^2 + f)}\cdot \tag{12}$$

k is the polarizability per unit volume and is a complex quantity, since the term $ir\omega$ in the denominator is an imaginary ($i = \sqrt{-1}$).

Equations (10), (11) and (12) apply to a dielectric having a single type of polarization characterized by the constants f, r, m, n and e. But in general an applied field induces several types of polarization simultaneously in a dielectric, and if we assume that it induces w types which are independent of each other, the total polarization per unit volume is given by

$$P = k_1F + k_2F + \cdots k_wF. \tag{13}$$

The total polarizability is then the sum of the individual polarizabilities, or

$$k = \sum_{j=1}^{w} k_j. \tag{14}$$

In this discussion it will be sufficient to consider that the different types of polarization designated by $k_1, k_2 \cdots k_w$ differ from one another only in having different sets of values for the constants of equation (12), designated by the subscripts $1,2,3 \cdots w$; for example, the character of the polarizability k_1 is specified by the set of constants m_1, r_1, f_1 and n_1.

In the first place it is evident that when the frequency of alternation of the voltage applied to the dielectric lies in the radio and power range it is possible to select any number of sets of values of m, r, f which will make the terms $m\omega^2$ and $r\omega$ negligible in comparison with f in the denominator of (12). Let m_1, r_1, f_1 be an example of such a set of constants and let there be n_1 particles per unit volume to which these constants apply. Then for this type of polarization equation (12) reduces to

$$k_1 = \frac{n_1 e^2}{f_1}. \tag{15}$$

This type of polarization is independent of frequency and will be referred to as an instantaneous polarization or an optical polarization. The main representatives of the instantaneous or optical polarizations are the electronic and atomic polarizations, which experience dispersion in the visible and infra-red but which are independent of frequency in the electrical range, and the contribution of this polarizability to the dielectric constant is therefore frequently calculated from refractive index measurements.

A second type of polarization results if we assume that the dielectric we are considering contains a class of particles for which $m\omega^2$ in equation (12) is negligible by comparison with $r\omega$ and with f, but in which $r\omega$ is of the same order of magnitude as f in the electrical range of frequencies. Let m_2, r_2, f_2 be a typical member of this class, the number of such particles per unit volume of the dielectric being n_2. Then for this class of particles equation (12) becomes

$$k_2 = \frac{n_2 e^2}{(ir_2\omega + f_2)}. \tag{16}$$

This expression represents the type of variation with frequency to which the name *anomalous dispersion* is given, and in the preceding paper the type of polarization which produces it was called an absorptive polarization.

It can readily be seen also that neglecting the $m\ddot{s}$ term in (10) or the $m\ddot{P}$ term in (11) leads to the same expression for k, i.e., equation (16), as does neglecting the $m\omega^2$ term in the denominator of (12). So for any member, (r_2, f_2, n_2), of the class of particles which produces anomalous dispersion, equation (10) reduces to

$$r_2\dot{s} + f_2 s = eF \tag{17}$$

and equation (11) becomes

$$r_2\dot{P} + f_2 P = n_2 e^2 F. \tag{18}$$

Décombé's theory, which has been mentioned earlier, was based upon an equation equivalent in most respects to (18), while Drude's expressions for dispersion were obtained by a method equivalent to neglecting $m\omega^2$ in (12).

Each term in equations (17) and (18) has an evident dynamical significance. Consequently, a physical picture of the essential nature of the anomalous dispersion process is given by equations (17) and (18) even though the values of constants r_2, f_2, n_2 and e are not specified in terms of independently measurable properties of the dielectric. Thus the term $f_2 s$ represents a restoring force tending to return the particles displaced by the impressed field to their initial positions, the constant f_2 acting as a stiffness coefficient; the term $r_2\dot{s}$ acts as a frictional force, r being a measure of the friction experienced by, for example, a moving ion or a rotating polar molecule; and, finally, eF is the driving force tending to displace a particle of charge e. Evidently conditions which are sufficient to produce anomalous dispersion exist whenever the motion of charged particles in an applied field is sufficiently specified by considering the effects of a restoring force proportional to the displacement of the typical particle and of a frictional force proportional to the velocity of the particle in the direction of applied field, as in equation (17). Or, putting it in more general terms, we may say that anomalous dispersion occurs whenever the relation between the polarization per unit volume and the force due to the internal electric field is given by an equation which can be reduced to (18). However, the possibility that anomalous dispersion may also occur under conditions which cannot be described by equation (18) is not excluded by the considerations given here.

A third type of polarization which can be obtained by selecting suitable sets of values for the constants of equation (12) is that in which none of the terms in the denominator of (12) can be neglected in the electrical range of frequencies. Let k_3 be the polarizability for this type of polarization which can then be represented by affixing the subscript 3 to the constants m, r, f and n of equation (12). This type of dispersion includes both the normal and the anomalous types but, as has already been indicated, in the radio and power ranges of frequency examples of a dispersion of this kind have not as yet been observed in dielectrics which are not piezo-electric.[16] It follows then that dielectrics behave as though the inertia of the particles which contribute to dielectric polarization is small enough that the inertia reaction $m\omega^2$ can be neglected in the electrical frequency range. This is an empirical result; the possibility of a polarization of the type k_3 occurring in the electrical frequency range is not excluded by the general theory of dispersion. The higher the frequency of an impressed field the greater should be the likelihood of encountering the type of frequency-variation described by k_3 (or equation (12)), because the prominence of the $m\omega^2$ term increases with the square of the frequency.

The preceding discussion shows that we can write equation (14) in the form

$$k = k_i + k_a, \tag{19}$$

where k is the total polarizability, k_i is the sum of the instantaneous polarizabilities and k_a the sum of the absorptive polarizabilities, that is, of the polarizabilities which vary with frequency according to equation (16). If for simplicity we take the case in which the dielectric has only one representative of k_i and one of k_a, we obtain by substituting the values of k_i and k_a given respectively in (15) and (16),

$$k = \frac{n_1e^2}{f_1} + \frac{n_2e^2}{(ir_2\omega + f_2)} \tag{20}$$

as an expression for the total polarizability.

Defining τ' by $\tau' \equiv r/f$, and dropping the subscripts in (20) to make the notation simpler, we obtain

$$k = k_i + \frac{ne^2}{f}\left[\frac{1}{1 + i\omega\tau'}\right], \tag{21}$$

which is the total polarizability per unit volume for a dielectric having two types of polarization, the one represented in (21) by the instantaneous polarizability k_i and the other by the absorptive polarizability

[16] Piezo-electric crystals such as quartz and Rochelle salt form exceptions, but for them dielectric polarization is coupled to macroscopic mechanical strains in the material and the mass reactance is due to the flexing or extension of the entire crystal. The dielectric constant of such a crystal as measured in almost any direction, shows an increase with increasing frequency, followed by anomalous dispersion. This is the behavior required by equation (12), or rather by an equation for the dielectric constant derivable from equation (12). This dispersion, however, depends upon the size and shape of the crystal, the nature of the electrodes and the manner of supporting the crystal during the measurements, and the exact interpretation of such measurements is a rather complex procedure. See, for example, W. P. Mason, *Proc. I. R. E.*, *23*, 1252–1263 (1935).

specified by the second term on the right. The quantity τ' is called the *relaxation-time.*

On multiplying the left-hand side of equation (21) by $(4\pi/3)(M/\rho)$ and the right-hand side by $(4\pi/3)(N/L)$ we obtain

$$\frac{4\pi}{3}\frac{Mk}{\rho} = \frac{4\pi}{3}N\left[\frac{k_i}{L} + \frac{ne^2}{fL}\left(\frac{1}{1 + i\omega\tau'}\right)\right], \tag{22}$$

which is the molar polarization.

For dielectrics to which the Clausius-Mosotti relation applies, equation (8) shows that

$$\frac{4\pi}{3}\frac{Mk}{\rho} = \frac{M}{\rho}\frac{\epsilon - 1}{\epsilon + 2} \tag{22a}$$

and in fact the expression on the right-hand side of (22*a*) is frequently called the molar polarization. Reference to equation (6) shows, however, that for gases (22*a*) reduces to the simpler relation.

$$\frac{4\pi}{3}\frac{Mk}{\rho} = \frac{M}{\rho}\frac{(\epsilon - 1)}{3}. \tag{22b}$$

And for Wyman's relation between dielectric constant and polarizability, which has been discussed earlier, the molar polarization becomes

$$\frac{4\pi}{3}\frac{Mk}{\rho} = \frac{M}{\rho}\frac{\epsilon + 1}{8.5}. \tag{22c}$$

Equations (22*a*), (22*b*) and (22*c*) are not the only relations between dielectric constant and molar polarization which have been proposed, but they apparently cover moderately well many of the conditions met in practice. For the right-hand member of equation (22) can be substituted whichever of the three expressions (22*a*), (22*b*), (22*c*) seems the most suitable for the type of dielectric under investigation.

If in equation (21) ω is set equal to zero we obtain the zero-frequency (or static) polarizability

$$k_0 = k_i + ne^2/f \tag{23}$$

and if ω is set equal to infinity we obtain

$$k_\infty = k_i. \tag{24}$$

Subtraction gives

$$k_0 - k_\infty = ne^2/f. \tag{25}$$

Substituting (24) and (25) in (21) gives

$$k = k_\infty + \left(\frac{k_0 - k_\infty}{1 + i\omega\tau'}\right). \tag{26}$$

The constants ne^2/f and k_i are not present in (26), being replaced by two special values of the polarizability, the zero-frequency value and the infinite-frequency value. However, it is not the polarizability but the dielectric constant which is directly observed in measurements on dielectrics, so it is desirable to replace k_0 and k_∞ by their equivalents in terms of the dielectric constant. But, as the earlier discussion has indicated, the relation between dielectric constant and polarizability is different for different types of dielectrics; three alternative expressions analogous to (22*a*), (22*b*) and (22*c*) will therefore be derived.

For materials to which equation (22*a*) (or the equivalent and simpler relation (5)) applies

$$k_0 - k_\infty = \frac{3}{4\pi}\left[\frac{\epsilon_0 - 1}{\epsilon_0 + 2} - \frac{\epsilon_\infty - 1}{\epsilon_\infty + 2}\right] = \frac{9(\epsilon_0 - \epsilon_\infty)}{4\pi(\epsilon_0 + 2)(\epsilon_\infty + 2)}, \quad (27)$$

where ϵ_0 is the zero-frequency dielectric constant and ϵ_∞ is the infinite-frequency dielectric constant. Then equation (26) can be replaced by

$$\frac{4\pi}{3}k = \frac{\epsilon_0 - 1}{\epsilon_\infty + 2} + \left[\frac{\epsilon_0 - 1}{\epsilon_0 + 2} - \frac{\epsilon_\infty - 1}{\epsilon_\infty + 2}\right]\frac{1}{1 + i\omega\tau'}. \quad (28)$$

By rationalizing and using the second expression given for $k_0 - k_\infty$ in equation (27) we can write equation (28) in the alternative form

$$\begin{aligned}\frac{4\pi k}{3} = \frac{\epsilon_\infty - 1}{\epsilon_\infty + 2} &+ \left[\frac{3(\epsilon_0 - \epsilon_\infty)}{(\epsilon_0 + 2)(\epsilon_\infty + 2)}\right]\cdot\frac{1}{1 + \omega^2\tau'^2} \\ &- i\left[\frac{3(\epsilon_0 - \epsilon_\infty)}{(\epsilon_0 + 2)(\epsilon_\infty + 2)}\right]\cdot\frac{\omega\tau'}{1 + \omega^2\tau'^2}. \quad (29)\end{aligned}$$

Equation (29) is the complex polarizability per unit volume multiplied by the factor $4\pi/3$ and expressed in terms of observable values of the dielectric constant and the relaxation-time τ'. The relaxation-time can also be expressed in terms of the reciprocal of a special value of the frequency; this permits all of the theoretical constants such as ne^2/f and τ' to be replaced by certain special values of the dielectric constant and a critical value of the frequency.

A simpler expression for the polarizability is obtained in the case of gases, or whenever equation (6) gives the relation between polarizability and dielectric constant. Equation (26) then gives

$$\frac{4\pi k}{3} = \frac{1}{3}\left(\epsilon_\infty - 1 + \frac{\epsilon_0 - \epsilon_\infty}{1 + i\omega\tau'}\right). \quad (30)$$

And for materials to which the relation (cf. equation (9)) proposed by

Wyman applies the procedure followed above yields

$$\frac{4\pi k}{3} = \frac{1}{8.5}\left(\epsilon_\infty + 1 + \frac{\epsilon_0 - \epsilon_\infty}{1 + i\omega\tau'}\right). \tag{31}$$

On multiplying equations (29), (30) and (31) by M/ρ three alternative formulæ for the molar polarization of a dielectric having polarizations of the type specified by equation (21) are obtained; the constants in these formulæ include only special values (ϵ_0 and ϵ_∞) of the dielectric constant and the relaxation-time, all of which can be obtained from dispersion curves.

The quantity $k_0 - k_\infty$ is a constant of the material, which, as equation (26) shows, represents the largest value which the absorptive part of the total polarizability, i.e., the k_a term in (19), can have for a given material; it may be described as the zero-frequency or static value of the absorptive part of the polarizability. Evidence as to the nature of a polarization can be obtained by investigating experimentally the dependence of $(k_0 - k_\infty)/\rho$ on temperature; for example, if the polarization is due to the changes of orientation of polar molecules according to the Debye theory this quantity should increase linearly with the reciprocal of the absolute temperature. It is useful, therefore, to express $(k_0 - k_\infty)/\rho$ in terms of observable values of the dielectric constant so that it may be plotted against temperature. In this connection there is, however, the same complication which has appeared in other places in this discussion regarding the relation between dielectric constant and polarizability. The three relations which have been discussed here yield for $(k_0 - k_\infty)/\rho$ the following expressions:

$$(k_0 - k_\infty)/\rho = \frac{3}{4\pi\rho}\left[\frac{\epsilon_0 - 1}{\epsilon_0 + 2} - \frac{\epsilon_\infty - 1}{\epsilon_\infty + 2}\right] \text{ (Clausius-Mosotti)} \tag{32a}$$

$$= \frac{3}{4\pi\rho}\left(\frac{\epsilon_0 - \epsilon_\infty}{3}\right) \qquad \text{(for gases)} \tag{32b}$$

$$= \frac{3}{4\pi\rho}\left(\frac{\epsilon_0 - \epsilon_\infty}{8.5}\right) \qquad \text{(Wyman).} \tag{32c}$$

The Complex Dielectric Constant

As the dielectric constant (ϵ) is the quantity directly measured in experimental investigations it is desirable to determine how it should vary with frequency for the type of dielectric polarization described in equation (21) or (29). Solving equation (5) for ϵ we obtain

$$\epsilon = \frac{1 + 8\frac{\pi}{3}k}{1 - 4\frac{\pi}{3}k}. \tag{33}$$

By substituting the expression for $4(\pi/3)k$ given in equation (28) into (33) we obtain

$$\epsilon = \frac{\dfrac{\epsilon_0}{\epsilon_0 + 2} + i\,\dfrac{\epsilon_\infty}{\epsilon_\infty + 2}\,\omega\tau'}{\dfrac{1}{\epsilon_0 + 2} + i\,\dfrac{1}{\epsilon_\infty + 2}\,\omega\tau'} \tag{34}$$

or

$$\epsilon = \frac{\dfrac{\epsilon_0}{\epsilon_0 + 2}\left(1 + i\,\dfrac{\epsilon_0 + 2}{\epsilon_\infty + 2}\cdot\dfrac{\epsilon_\infty}{\epsilon_0}\,\omega\tau'\right)}{\dfrac{1}{\epsilon_0 + 2}\left(1 + i\,\dfrac{\epsilon_0 + 2}{\epsilon_\infty + 2}\,\omega\tau'\right)}. \tag{34a}$$

Then, by setting

$$\frac{\epsilon_0 + 2}{\epsilon_\infty + 2}\,\tau' = \tau, \tag{35}$$

we obtain

$$\frac{\epsilon}{\epsilon_0} = \frac{\left(1 + i\,\dfrac{\epsilon_\infty}{\epsilon_0}\,\omega\tau\right)}{1 + i\omega\tau} \tag{36}$$

and transforming this into polar form to facilitate division gives

$$\frac{\epsilon}{\epsilon_0} = \frac{\rho_1 e^{i\varphi_1}}{\rho_2 e^{i\varphi_2}} = \frac{\rho_1}{\rho_2}\,e^{i(\varphi_1 - \varphi_2)} = \frac{\rho_1}{\rho_2}\left[\cos\,(\varphi_1 - \varphi_2) + i\sin\,(\varphi_1 - \varphi_2)\right], \tag{37}$$

where

$$\rho_1 = \left[1 + \left(\frac{\epsilon_\infty}{\epsilon_0}\right)^2 \omega^2\tau^2\right]^{\frac{1}{2}}, \qquad \rho_2 = [1 + \omega^2\tau^2]^{\frac{1}{2}},$$

$$\varphi_1 = \tan^{-1}\frac{\epsilon_\infty}{\epsilon_0}\,\omega\tau \qquad \text{and} \qquad \varphi_2 = \tan^{-1}\omega\tau.$$

Equation (37) then gives

$$\frac{\epsilon}{\epsilon_0} = \frac{1 + \dfrac{\epsilon_\infty}{\epsilon_0}\,\omega^2\tau^2}{1 + \omega^2\tau^2} + i\,\frac{\left(\dfrac{\epsilon_\infty}{\epsilon_0} - 1\right)\omega\tau}{1 + \omega^2\tau^2} \tag{38}$$

or

$$\frac{\epsilon}{\epsilon_0} = \frac{\epsilon_\infty}{\epsilon_0} + \frac{1 - \dfrac{\epsilon_\infty}{\epsilon_0}}{1 + \omega^2\tau^2} + \frac{i\left(\dfrac{\epsilon_\infty}{\epsilon_0} - 1\right)\omega\tau}{1 + \omega^2\tau^2}, \tag{38a}$$

from which we obtain

$$\epsilon = \epsilon_\infty + \frac{\epsilon_0 - \epsilon_\infty}{1 + \omega^2\tau^2} - \frac{i(\epsilon_0 - \epsilon_\infty)\omega\tau}{1 + \omega^2\tau^2}. \tag{39}$$

Equation (39) is *the complex dielectric constant* expressed in rectangular form for a dielectric having a polarizability (per unit volume) given by (21) and in which the internal field (F) is such that the Clausius-Mosotti relation (equation (8)) applies.

For gases the derivation of the expression for the complex dielectric constant from that for the polarizability is simpler, though the same in principle, as the above. From (22*b*) or (6) we see that $\epsilon = 1 + 4\pi k$; and on substituting (30) for $4\pi k$ and rationalizing we obtain

$$\epsilon = \epsilon_\infty + \frac{\epsilon_0 - \epsilon_\infty}{1 + \omega^2\tau'^2} - i\,\frac{(\epsilon_0 - \epsilon_\infty)\omega\tau'}{1 + \omega^2\tau'^2}. \tag{40}$$

And when the relation between polarizability and dielectric constant is that proposed by Wyman, cf. (9) or (22*c*), we again obtain (40) on substituting (31) for $(4\pi/3)k$ in (9) and rationalizing.

It will be noticed that the difference between (40) and (39) is that τ' appears in the former and τ in the latter, τ being given by (35). This shows that the factor $(\epsilon_0 + 2)/(\epsilon_\infty + 2)$ has its origin in the fact that for the conditions to which (39) applies $F = E + (4\pi/3)P$, while for the conditions to which (40) applies $F = E$, or is a linear function of E. τ is the relaxation-time for the dielectric constant, while τ' is the relaxation-time for the polarizable units in the material; when $F = E$ these two relaxation-times are equal.

For materials of high dielectric constant the factor $(\epsilon_0 + 2)/(\epsilon_\infty + 2)$ produces a considerable difference between τ and τ'; for example, for water or ice τ is about $23\tau'$. In a recent paper, R. H. Cole[17] has shown that when the volumes of certain polar molecules are calculated from τ by means of Debye's expression for the relaxation-time better agreement with the volume estimated from van der Waals' equation is obtained when Onsager's relation between polarization and dielectric constant is used instead of the Clausius-Mosotti relation. In particular, for water the van der Waals coefficient gives 13×10^{-24} c.c. for the volume of the molecule, while τ' gives 0.5×10^{-24} c.c. on the Clausius-Mosotti relation but 12×10^{-24} c.c. on the Onsager relation, and 4×10^{-24} c.c. for a modified Onsager relation. And if Wyman's relation is used, $\tau = 23\tau'$, and the volume should be 23 times that calculated on the basis of the Clausius-Mosotti relationship, or about 11×10^{-24} c.c.

Both equation (39) and equation (40) can be expressed in the form

$$\epsilon = \epsilon' - i\epsilon'', \tag{41}$$

[17] R. H. Cole, *Jour. Chem. Phys.*, *6*, 385 (1938).

where

$$\epsilon' = \epsilon_\infty + \frac{\epsilon_0 - \epsilon_\infty}{1 + \omega^2\tau^2}, \tag{41a}$$

$$\epsilon'' = \frac{(\epsilon_0 - \epsilon_\infty)\omega\tau}{1 + \omega^2\tau^2}, \tag{41b}$$

if τ is considered to be given by

$$\tau = \frac{\epsilon_0 + 2}{\epsilon_\infty + 2}\tau',$$

when the complex dielectric constant is given by (39), that is when the Clausius-Mosotti relation applies, and by

$$\tau = \tau',$$

when the material is a gas, or when Wyman's relation between polarizability and dielectric constant applies.

The real part ϵ' of the complex dielectric constant is usually referred to simply as the dielectric constant, while the imaginary part ϵ'' is frequently called the *loss factor*.[18] There are alternative ways of expressing the same property of the material; for example, the tangent of the loss angle, ϵ''/ϵ', is frequently used instead of ϵ''.

Comparison of Dispersion Formulæ

Comparison of (39) or (41), (41*a*), (41*b*) with equation (69), page 97, of Debye's "Polar Molecules" shows that the equation for the complex dielectric constant derived here is identical with that of the Debye theory (in the present notation τ' corresponds to τ in Debye's book). This means that any characteristics which can be derived from equations (41), (41*a*), (41*b*) without specifying the values of the constants ϵ_∞, ϵ_0 and τ are common to at least two types of polarization, that is, to the polarization due to the effect of an applied field on the orientation of polar molecules according to the Debye theory, and to the polarization described by equation (18) or equation (21).

The difference between the formulæ for dispersion derived here and those of the Debye theory are best seen by comparing the expressions which they yield for the molar polarization. On the Debye Theory:

$$\frac{M}{\rho}\frac{\epsilon - 1}{\epsilon + 2} = \frac{4\pi N}{3}\left[\alpha_0 + \frac{\mu^2}{3kT}\left(\frac{1}{1 + i\omega\tau'}\right)\right].$$

[18] After a suggestion made by E. T. Hoch, *B. S. T. J.*, November (1922), and by H. H. Race, *Jour. A. I. E. E.*, *51*, 354 (1932).

The present derivation gives:

$$\frac{M}{\rho}\frac{\epsilon - 1}{\epsilon + 2} = \frac{4\pi N}{3}\left[\frac{k_i}{L} + \frac{ne^2}{fL}\left(\frac{1}{1 + i\omega\tau'}\right)\right].$$

The constant α_0 has the same significance as k_i/L, only the notation being different; both represent the optical frequency, or "instantaneous," polarizability; μ is the permanent electric moment of the molecule, k Boltzmann's constant and T the absolute temperature. The quantity ne^2/fL which corresponds to $\mu^2/3kT$ in the Debye formula contains three constants n, e, and f whose physical significance is indicated only in a general way (see Appendix). τ' (or in Debye's notation τ) is equal to $4\pi\eta a^3/kT$ in the Debye theory while in the formula derived here $\tau' = r/f$ where r is a frictional coefficient whose physical origin is not specified. Thus though the formula for molar polarization derived here is not directly useful as a means of investigating the molecular (or other) origin of dielectric polarizations, it facilitates distinguishing those aspects of the Debye formula which are peculiar to a polarization depending upon changes in the orientation of polar molecules which are free to assume any (or at least more than one) orientation from the more general aspects shared by other types of polarization, such as the one specified by (18) and (21). Thus, the functions $\mu^2/3kT$ and $4\pi\eta a^3/kT$ are peculiar to the Debye theory, while the function $(1 + i\omega\tau)^{-1}$ also appears in the dispersion formula derived here, as well as in other formulæ to be discussed below.

We have seen that the viscous-elastic type of polarization specified by (18) and (21) produces a complex dielectric constant given by (39), or by the equivalent equations (41), (41a), (41b), where

$$\tau = \frac{\tau'(\epsilon_0 + 2)}{(\epsilon_\infty + 2)},$$

and that the same formulæ express the complex dielectric constant on the Debye theory of polar molecules when the constants τ' and $\epsilon_0 - \epsilon_\infty$ (or $k_0 - k_\infty$), are given the values derived for them on the Debye theory. Other theories have been proposed to explain the variation of dielectric constant and dielectric loss with frequency, but for the most part these have been derived for composite dielectrics, consisting of two or more layers of different materials, or of small spheres of one material dispersed or embedded in another material. These theories also yield formulæ (41), (41a) and (41b) for the complex dielectric constant, the expressions for τ' and $\epsilon_0 - \epsilon_\infty$ being, of course, different from those of the Debye theory.[19] These expressions are included in Table I.

TABLE I

CONSTANTS OF THE FORMULA FOR THE COMPLEX DIELECTRIC CONSTANT EQUATIONS (41), (41a) AND (41b)

Type of Polarization	$\epsilon_0 - \epsilon_\infty$	τ	$k = f(\epsilon)$
1. Orientation of polar molecules (The Debye theory).[1]	$\frac{4\pi}{3} \cdot \frac{(\epsilon_0+2)(\epsilon_\infty+2)}{3} \cdot \frac{L\mu^2}{3kT}$	$\frac{(\epsilon_0+2)}{(\epsilon_\infty+2)} \cdot \frac{4\pi\eta a^3}{kT}$	Eq. (5)
	$4\pi \cdot \frac{L\mu^2}{3kT}$	$\frac{4\pi\eta a^3}{kT}$	Eq. (6)
	$\frac{34\pi}{3} \cdot \frac{L\mu^2}{3kT}$	$\frac{4\pi\eta a^3}{kT}$	Eq. (9)
2. Displacement of changed particles against elastic restoring forces and viscous frictional forces of undetermined origin, as specified in equation (17). (Modification of Drude theory.)[2]	$\frac{4\pi}{3} \cdot \frac{(\epsilon_0+2)(\epsilon_\infty+2)}{3} \cdot \frac{ne^2}{f}$	$\frac{\epsilon_0+2}{\epsilon_\infty+2} \cdot \frac{r}{f}$	Eq. (5)
	$4\pi \cdot \frac{ne^2}{f}$	$\frac{r}{f}$	Eq. (6)
	$\frac{34\pi}{3} \cdot \frac{ne^2}{f}$	$\frac{r}{f}$	Eq. (9)
3. Interfacial or ionic polarizations:			
(a) Two-layer dielectric; layer (ϵ_1, γ_1) being of same thickness as layer (ϵ_2, γ_2), (Wagner).[3]	$\frac{(\epsilon_1\gamma_2-\epsilon_2\gamma_1)^2}{(\epsilon_1+\epsilon_2)(\gamma_1+\gamma_2)^2}$	$\frac{\epsilon_1+\epsilon_2}{\gamma_1+\gamma_2}$	
(b) Special case of (a); high-resistance blocking layer at electrode/dielectric boundary (Joffé).[4]	C_1/C_∞	RC_1	
(c) Suspension of spheres (ϵ_1, γ_1) in a medium (ϵ_2, γ_2), where $p \ll 1$ (Wagner),[5] (Gemant).[7]	$\frac{9p(\epsilon_1\gamma_2-\epsilon_2\gamma_1)^2}{(2\epsilon_2+\epsilon_1)(2\gamma_2+\gamma_1)^2}$	$\frac{2\epsilon_2+\epsilon_1}{2\gamma_2+\gamma_1}$	
(d) Special case of (c); conducting spheres in an insulating medium, where $\epsilon_1 = \epsilon_2$ and $\gamma_2 \ll \gamma_1$ (Wagner).[5]	$3p\epsilon_1$	$3\epsilon_1/\gamma_1$	
(e) Special case of (d); conducting shells (Miles and Robertson).[6]	$3p\epsilon_1$	$\frac{3\epsilon_1 b}{2\gamma_2 d}$	

[1] Reference numbers in this table refer to list of references at the end of this paper.

Table I contains a list of expressions which, when substituted for $(\epsilon_0 - \epsilon_\infty)$ and τ in (41a) and (41b), give several formulæ for the complex dielectric constant. Included in this list are most of the formulæ which have been proposed to explain the simplest type of variation of

[19] Cf. Gemant, "Elektrophysik der Isolierstoffe," Berlin (1930).

dielectric constant and loss factor with frequency which is observed in the different classes of materials to which the various items in the table refer. In some cases the original formulæ, as they appear in the literature, have been expressed in terms which do not show an obvious equivalence to (41), (41*a*) and (41*b*), but by re-expressing them in the form (41) and then determining ϵ_0 and ϵ_∞ by letting $\omega = 0$ and ∞, respectively, the list of expressions given in Table I is obtained. It is interesting that theories based on such widely dissimilar physical mechanisms as rotating polar molecules (Item 1, Table I) and a blocking-layer of high resistance at an electrode/dielectric interface (Item 3(*b*), Table I) should yield an identical form of variation with frequency.

For the first two polarizations listed in Table I, the alternative expressions obtained by assuming three alternative relationships between polarizability and dielectric constant are given. By means of Table II the quantities $(k_0 - k_\infty)$ and τ' can be obtained from $(\epsilon_0 - \epsilon_\infty)$ and τ in Table I. The resulting expressions can then be substituted for $(k_0 - k_\infty)$ and τ' in equation (26) yielding expressions for the polarizabilities of the different types of polarization listed in Table I. The molar polarization can then be obtained by multiplying (26) by $(4\pi/3)(M/\rho)$. However, in general it is not likely that any useful purpose is to be served by calculating the molar polarization for interfacial polarizations; a more significant quantity would be the polarization per conducting particle, when the polarization is of the type (3*d*), Table I, and the number of conducting particles per unit volume can be estimated.

TABLE II

THE RELATIONSHIP BETWEEN $(k_0 - k_\infty)$ AND $(\epsilon_0 - \epsilon_\infty)$ AND BETWEEN τ' AND τ

	$(k_0 - k_\infty)$	τ'
Clausius-Mosotti Relation........	$\frac{3}{4\pi} \cdot \frac{3(\epsilon_0 - \epsilon_\infty)}{(\epsilon_0 + 2)(\epsilon_\infty + 2)}$	$\frac{\epsilon_\infty + 2}{\epsilon_0 + 2}\tau$
Gases......................	$\frac{1}{4\pi} \cdot (\epsilon_0 - \epsilon_\infty)$	τ
Wyman's Empirical Relation	$\frac{3}{4\pi} \cdot \frac{\epsilon_0 - \epsilon_\infty}{8.5}$	τ

We have pointed out that a number of theories which have been proposed for the explanation of the variation of dielectric constant with frequency may be expressed in the forms (41*a*) and (41*b*) when the expressions listed in Table I are substituted for $(\epsilon_0 - \epsilon_\infty)$ and τ, but we have not yet indicated how these formulæ agree with experimental data. For such materials as ice (see Fig. 3, for example) and for certain alcohols and glycols, the experimental points agree fairly closely with the curves obtained by plotting equations (41*a*) and (41*b*) for a suitable choice of the values of the constants.

But for many other dielectrics, particularly non-homogeneous systems or disperse systems such as those listed under Item 3, Table I, the simple dispersion formulæ (41*a*) and (41*b*) often fall very far short of adequately representing the experimental data. Von Schweidler[20] and Wagner[21] have attempted to explain the form of dispersion curves obtained for such materials by postulating that the polarizations induced in the dielectric have a wide range of relaxation times at any given temperature, instead of a single relaxation time, as for the polarizations listed in Table I. A further contribution to the theory of the distribution of relaxation times has recently been made by Yager.[22] However, in spite of the existence of many materials which do not show the type of dispersion described by (41*a*) and (41*b*), the value of these formulæ in interpreting experimental data is considerable, particularly as applied to pure materials.

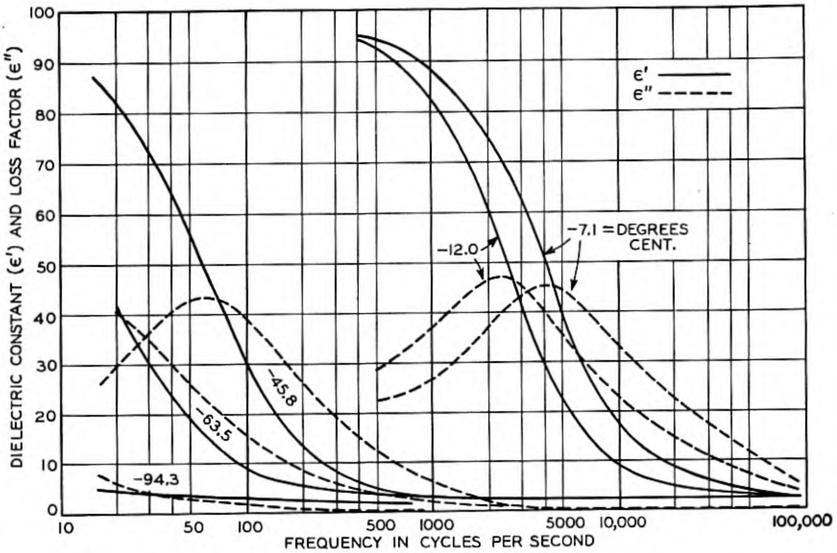


Fig. 3—Experimental dispersion curves for ice.

Table I emphasizes the point that mere agreement of experimental data for dielectric constant and dielectric loss with the theoretical

[20] E. v. Schweidler, *Ann. d. Phys.*, (4) (*24*), 711 (1907).
[21] K. W. Wagner, *Archiv f. Elektrotechnik*, *2*, 371 (1914).
[22] W. A. Yager, *Physics*, *7*, 434 (1936).

curves obtained by plotting (41*a*) and (41*b*) for suitably adjusted values of the constants only places the type of mechanism to which the observed dispersion can be attributed within the rather large catagory which includes at least the seven types of mechanism listed in the table. Data showing the dependence of $(k_0 - k_\infty)/\rho$ on temperature allows a further specialization of the processes which could account for the observed behavior; and of course a number of possibilities can be discarded on general grounds of physical improbability. And finally, agreement of the constants calculated from dielectric measurements with the values calculated from independent estimates of the sizes and other characteristics of the molecules or other elementary units which contribute to the polarization provides the most convincing evidence of the nature of the polarization. Such agreement is frequently obtained in the application of the Debye theory to gases and liquids.

The characteristics which can be deduced from equations (41*a*) and (41*b*) without substituting for the constants theoretical expressions, such as those given in Table I, are of considerable value in interpreting electrical measurements upon dielectrics. It may be convenient to describe these as the *general* characteristics of anomalous dispersion, distinguishing them thereby from the *special* characteristics peculiar to particular kinds of dielectric polarization which share the property of producing anomalous dispersion in the radio and power range of frequencies.

APPENDIX

The following list contains the definitions of the quantities which appear in Table I:

$\epsilon_1, \epsilon_2, \gamma_1, \gamma_2$ are respectively the dielectric constants and conductivities of two materials designated by subscripts 1 and 2, the unit of conductivity being such that $\gamma = 36\pi \times 10^{11}\,\lambda$, where λ is in $(\text{ohm}\cdot\text{cm})^{-1}$.

$\epsilon_0, \epsilon_\infty$ are respectively the dielectric constant at the lower and upper extremities of dispersion curves; they are called the zero-frequency (or static) dielectric constant and the infinite-frequency dielectric constant,

L is the number of molecules per unit volume,

η the viscosity of a liquid containing polar molecules,

k Boltzmann's constant,

T the absolute temperature,

μ the permanent electric moment of a polar molecule,

a the radius of a polar molecule, assumed to be spherical,
b the radius of a colloidal particle,
d the thickness of a conducting skin on the particle of radius b,
r a frictional resistance coefficient of unspecified origin,
f an elastic restoring force coefficient of unspecified origin,
n the number per unit volume of elementary charged particles subject to certain specified conditions,
p the ratio of the volume occupied by the spherical particles in ($3c$, d, e) to the total volume,
C_1 the capacity of the blocking layer of ($3b$), Table I,
R the resistance of the dielectric, exclusive of the blocking layer.

The following list contains the definitions of quantities which appear in other parts of the article.

ω is 2π times the frequency of alternation of the applied field,
V the applied voltage,
E the intensity of the applied field,
P the polarization per unit volume induced by a field E,
F the internal or local field,
ρ the density of the dielectric,
M the molecular weight of the material of which the dielectric is composed,
m the mass of a molecule; in another context, the mass of any charged particle considered in the discussion,
N is Avogadro's number, 6.06×10^{23} molecules per mole,
s the displacement of a charged particle from an equilibrium position by an applied field,
$\dot{s}$ the velocity of the charged particle in the applied field,
$\ddot{s}$ the acceleration of the particle in the applied field.

References Relating to Table I

1. P. Debye, "Polar Molecules," New York (1929).
2. P. Drude, *Ann. d. Physik*, *64*, 131 (1898); L. Décombé, *J. d. Physique* (5) *3*, 215 (1912); and the present article.
3. K. W. Wagner, Chap. I of Schering's "Die Isolierstoffe der Elektrotechnik," Springer, Berlin (1934).
4. A. Joffé, "The Physics of Crystals," New York (1928).
5. K. W. Wagner, *Arch. f. Elektrotechnik*, *2*, 371 (1914).
6. J. B. Miles and H. P. Robertson, *Phys. Rev.*, *40*, 583 (1932).
7. A. Gemant, "Die Elektrophysik der Isolierstoffe," Berlin (1930).

General reviews of the theory of dielectric behavior as it concerns dispersion for power and radio frequencies are included in the following places, among others:

1. E. Schrödinger, "Dielektrizität," Graetz, *Handb. d. Elek. u. d. Magn.*, Leipzig (1918), pp. 157–229.

2. E. v. Schweidler, "Die Anomalien der dielektrischen Erscheinungen," *ibid.*, p. 232; *Ann. d. Phys.* (4) *24*, 711 (1907).
3. J. B. Whitehead, "Lectures on Dielectric Theory and Insulation," McGraw-Hill (1927).
4. L. Hartshorn, *Jour. I. E. E.*, *64*, 1152 (1926).
5. P. Debye, "Polar Molecules," Chem. Cat. Co., New York (1929).
6. Schering's "Die Isolierstoffe der Elektrotechnik," Springer, Berlin (1924).
7. A. Gemant, "Die Elektrophysik der Isolierstoffe," Berlin (1930).

Abstracts of Technical Articles from Bell System Sources

Electron Microscope Studies of Thoriated Tungsten.[1] ARTHUR J. AHEARN and JOSEPH A. BECKER. Many past experiments have shown that the thermionic activity of a thoriated tungsten filament is determined by the concentration of thorium on its surface. This concentration is in turn determined by the rate of arrival and rate of evaporation of thorium. Typical published values of these rates are given in Fig. 1. An electron microscope used to obtain electron images of thoriated tungsten ribbons is described. Comparison with photomicrographs shows that the active and inactive patches composing an electron image agree in size, shape and number with the exposed grains of the tungsten. The electron microscope *shows that thorium comes to the surface in "eruptions"* at a relatively small number of randomly located points. From a comparison of photomicrographs showing thoria globules and electron images of thorium eruptions, it is deduced that *all the thorium in a globule comes to the surface when an eruption occurs.* Factors such as a high temperature flash and sudden heating and cooling of the filament affect the frequency of eruptions. Thorium eruptions are the only observed manner in which thorium arrives at the filament surface. They are repeatedly observed in the early stages of thoriation. Eruptions are not observed in the later stages of thoriation where conditions are unfavorable for their observance. Electron images of a Pintsch single crystal filament reveal alternate active and inactive bands parallel to the filament axis. Thorium eruptions occur only on the active bands. With a polycrystalline ribbon the surface migration of thorium from the eruption centers is isotropic; *with a single crystal ribbon there is a strongly preferred direction of migration.* X-ray analysis shows that the surface is a (211) plane and that *the preferred direction of migration agrees with the* (111) *direction in this plane.* During the process of thoriating a filament the relative emissions from different grains change by substantial amounts; in many cases the change is so great that the relative emissions are reversed. Measurements of work function differences between grains gave values ranging up to 0.6 volt.

The Mechanism of Hearing as Revealed through Experiment on the Masking Effect of Thermal Noise.[2] HARVEY FLETCHER. In an electri-

[1] *Phys. Rev.*, September 15, 1938.
[2] *Proc. Nat'l. Acad. Sci.*, July 1938.

cal conductor there is a statistical variation of the electrical potential difference between its two ends which is due to the thermal agitation of the atoms, including the electrons. This electrical noise is amplified by means of a vacuum tube amplifier and then converted into an acoustical noise by means of a telephone receiver held on the ear. When this noise is present it reduces the capability of the ear to hear other sounds. The intensity per cycle of the acoustical noise compared to the intensity of a pure tone which can just be perceived in the presence of a noise was determined experimentally using a group of observers. This relative intensity for a given frequency range was constant throughout a wide variation of intensity. However, its value does vary with the position in the frequency spectrum and it is the amount of this variation which enables one to calculate the relation between the frequency of the tone and its position of maximum stimulation along the basilar membrane. The results of such a calculation are given and shown to be in good agreement with determinations from animal experimentation.

Transcontinental Telephone Lines.[3] J. J. PILLIOD. A fourth transcontinental line has just been created by the completion of four pairs of open wire between Oklahoma City and Whitewater, California. This open-wire line connects at its eastern terminus with the already existing toll cables from the east, and at its western terminus with a toll cable running into Los Angeles.

In a cross-section of the United States just west of Denver, there are now 140 through telephone circuits and about the same number of telegraph circuits carried by four open-wire routes.

The four new pairs which constitute the transcontinental line carry, in addition to the usual voice frequency channels, three channels of carrier. But their design throughout has been such that twelve additional carrier circuits can be superimposed upon the four channels now provided by each wire pair.

The wires of each pair are spaced 8 inches apart with the nearest spacing between pairs being 26″ while crossarms are 36″ apart. New transposition systems have also been used to further reduce crosstalk.

Application of Statistical Methods to Manufacturing Problems.[4] W. A. SHEWHART. The application of statistical methods in mass production makes possible the most efficient use of raw materials and manufacturing processes, effects economies in production, and makes possible the highest economic standards of quality for the manufactured goods used by all of us. The story of the application, however,

[3] *Electrical Engineering,* October 1938.
[4] *Jour. Franklin Institute,* August 1938.

is of much broader interest. The economic control of quality of manufactured goods is perhaps the simplest type of scientific control. Recent studies in this field throw light on such broad questions as: How far can Man go in controlling his physical environment? How does this depend upon the human factor of intelligence and how upon the element of chance?

Observational Significance of Accuracy and Precision.[5] W. A. SHEWHART. Two of the most common terms used in pure and applied science are accuracy and precision. When such terms are used, as in the specification of quality of manufactured products, it is desirable that they have definite and, in so far as possible, experimentally verifiable meanings. It is, therefore, important to determine how far one can go towards attaining this end by applying with rigor the principle that *only that which is observable is significant.* In the application of the concepts of accuracy and precision, it is customarily assumed that the available data constitute a random sample. Hence, the first step in attaining experimentally verifiable meaning of these terms is to choose an operationally verifiable criterion of randomness. One such criterion is the quality control chart. In order to give experimental definiteness to any *measure* of either accuracy or precision derived from a random sample, it is also necessary to specify the way any statement involving the measure may be experimentally verified. To do this it is necessary to make at least four empirical choices as to the details of taking and analyzing the data in the process of verification. Hence, it appears that the meaning of either precision or accuracy is verifiable. Hence, it appears that the meaning of either precision or accuracy is verifiable only in a limited sense subject in any specific case to the choice of empirical criteria of verification.

The Time Lag in Gas-Filled Photoelectric Cells.[6] A. M. SKELLETT. In commercial gas-filled photoelectric cells there is a lag in response which becomes appreciable above frequencies in the neighborhood of 10,000 cycles. If this lag is due to the transit times of the ions across the cell, it should be possible to set up resonance conditions by varying the frequency of modulation of the incident light intensity. This has been accomplished in a cell of special design and the resonance conditions agree with the theory, thereby demonstrating that the transit time of the ions is the simple cause. The paper also discusses the flow of the ions and electrons across the cell and their impacts in relation to the flow of current in the external circuit.

[5] *Jour. Wash. Acad. Sciences,* August 15, 1938 (p. 381).
[6] *Internat'l. Projectionist,* September 1938; *Jour. Applied Physics,* October 1938.

Contributors to this Issue

Charles R. Burrows, B.S. in Electrical Engineering, University of Michigan, 1924; A.M., Columbia University, 1927; E.E., University of Michigan, 1935. Research Assistant, University of Michigan, 1922–23. Western Electric Company, Engineering Department, 1924–25; Bell Telephone Laboratories, Research Department, 1925–. Mr. Burrows has been associated continuously with radio research and is now in charge of a group investigating the propagation of ultra-short waves.

Arthur B. Crawford, B.S. in Electrical Engineering, Ohio State University, 1928. Member of Technical Staff, Bell Telephone Laboratories, 1928–. Mr. Crawford has been engaged chiefly in work relative to radio communication by ultra-short waves.

Carl R. Englund, B.S. in Chemical Engineering, University of South Dakota, 1909; University of Chicago, 1910–12; Professor of Physics and Geology, Western Maryland College, 1912–13; Laboratory Assistant, University of Michigan, 1913–14. Western Electric Company, 1914–25; Bell Telephone Laboratories, 1925–. As Radio Research Engineer Mr. Englund is engaged largely in experimental work in radio communication.

L. A. Meacham, B.S. in Electrical Engineering, University of Washington, 1929. Cambridge University, England, 1929–30. Bell Telephone Laboratories, 1930–. Mr. Meacham's work has been concerned with the generation and distribution of constant reference frequencies.

S. O. Morgan, B.S. in Chemistry, Union College, 1922; M.A., Princeton University, 1925; Ph.D., 1928. Western Electric Company, Engineering Department, 1922–24; Bell Telephone Laboratories, 1927–. Dr. Morgan's work has been on the relation between dielectric properties and chemical composition.

William W. Mumford, B.A., Willamette University, 1930. Bell Telephone Laboratories, 1930–. Mr. Mumford has been engaged in radio receiving work, chiefly on the problem of propagation and measurement in the ultra-short-wave region.

E. J. Murphy, B.S., University of Saskatchewan, Canada, 1918; McGill University, Montreal, 1919–20; Harvard University, 1922–23.

Western Electric Company, Engineering Department, 1923–25; Bell Telephone Laboratories, 1925–. Mr. Murphy's work is largely confined to the study of the electrical properties of dielectrics.

A. C. Norwine, A.B., University of Missouri, 1923; B.S. in Electrical Engineering, 1924; E.E., 1925. Bell Telephone Laboratories, 1925–. Mr. Norwine has been principally engaged in studies of the effects of transmission delay and voice operated devices on toll telephone circuits.

A. J. Rack, B.S. in Electrical Engineering, University of Illinois, 1930; M.A. in Physics, Columbia University, 1935. Bell Telephone Laboratories, 1930–. Starting with radio research, Mr. Rack has more recently been engaged in the analysis of special problems arising in amplifier circuits.

E. F. Watson, M.E., Cornell University, 1914. American Telephone and Telegraph Company, Engineering Department, 1914–19; Department of Development and Research, 1919–34. Bell Telephone Laboratories, 1934–. Mr. Watson has been concerned with the development of various types of telegraph equipment, particularly teletypewriters, telephotograph equipment, telegraph maintenance and testing equipment, grounded telegraph systems and regenerative telegraph repeaters. His present work as Teletypewriter Engineer is along these same lines.

S. B. Wright, M.E. in Electrical Engineering, Cornell University, 1919. Engineering Department and Department of Development and Research, American Telephone and Telegraph Company, 1919–34; Bell Telephone Laboratories, 1934–. Mr. Wright is engaged in transmission development of radio systems.