THE BELL SYSTEM

Technical Journal

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

VOLUME XXXVI JULY 1957 NUMBER 4

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published six times a year by the American Telephone and Telegraph Company, 195 Broadway, New York 7, N. Y. F. R. Kappel, President; S. Whitney Landon, Secretary; John J. Scanlon, Treasurer. Subscriptions are accepted at $5.00 per year. Single copies $1.25 each. Foreign postage is 65 cents per year or 11 cents per copy. Printed in U. S. A.

THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XXXVI JULY 1957 NUMBER 4

Noise Spectrum of Electron Beam in Longitudinal Magnetic Field

By W. W. Rigrod

(Manuscript received January 21, 1957)

Measurements of induced noise currents along drifting cylindrical electron beams have shown that noise fluctuations propagate as space-charge waves in the same fashion as RF signals of the same frequency. On many such beams, however, the regular standing-wave noise pattern is interrupted, after some drift distance, by a smooth steep increase in noise current, followed by slow, shallow undulations. This "growing noise" phenomenon, discovered by Smullin and his co-workers at M.I.T. several years ago, is the subject of study in this paper. Its importance is considerable, in a negative way, because it has hampered the development of medium-power traveling-wave-tube devices with acceptably low noise figures.

The experimental measurements show the growing noise pattern to be the result of a two-stage process. Its primary cause is rippled-beam amplification of noise fluctuations over a wide band of microwave frequencies, much higher than the usual observation frequency. This explains its elusiveness. In the second stage, noise energy is transferred to lower frequencies, due to intermodulation and other non-linear processes within the gain band. As the beat-frequency noise increments are excited by continuous arrays of frequency pairs, their standing-wave patterns overlap one another, resulting in a smooth growing-noise pattern.

In Part II of this paper, measurements of the noise spectrum of a rippled beam in the UHF region are described. These measurements reveal the presence of additional forms of instability. Calculations are made to account for some of these, and for aspects of rippled-beam amplification not previously understood.

Part I — The Growing Noise Phenomenon*

I INTRODUCTION

When an RF probe is moved along a magnetically-focused electron beam in a drift region, the noise power is at first found to vary periodically with distance from the electron gun.[1] For a sufficiently long beam, however, the periodic pattern is succeeded by an exponential rise, culminating in an irregular plateau. This so-called "growing noise" phenomenon has been extensively investigated by its discoverers, L. Smullin and his colleagues at the M.I.T. Research Laboratory of Electronics.[2, 3] They have established that this noise will begin to grow at a plane nearer the gun, and tend to grow at a faster rate, for electron beams (a) of higher perveance, (b) with less space-charge neutralization by positive ions, and (c) issuing from convergent, partly-shielded guns, rather than those immersed in the magnetic field.

The growth of microwave noise power in drifting beams has hampered the development of high-power, traveling-wave tubes with acceptably low noise figures, as such devices generally have convergent, partly-shielded electron guns. The problem has been evaded in the design of low-noise, low-power traveling-wave tubes, by resort to confined-flow, parallel beams.

Several theories have been proposed to explain the growing-noise wave:

(1) Excitation of higher-order modes with complex propagation constants, by electrons threading the beam transversely;[4]

(2) Slipping-stream amplification, due to either longitudinal or transverse velocity gradients;[5]

(3) Rippled-beam amplification;[6, 7, 8] and

(4) Electron-electron interactions leading eventually to equipartition of thermal energy, and thus an increase in longitudinal velocity fluctuations.

In Part I of this paper, measurements are presented which show that the principal cause of growing noise appears to be space-charge wave

* Presented at the I.R.E. Electron Tube Research Conference, Boulder, Colorado, June 27–29, 1956.

amplification due to beam rippling. The mechanism is studied in some detail, as its connection with the usually-observed exponential rise of noise is not immediately apparent. In Part II, the UHF noise spectrum and its spatial distribution in beams with large-amplitude, long wavelength ripples, are described. In addition, some of the underlying processes are analyzed.

II APPARATUS

As sketched in Fig. 1, the heart of the apparatus consists of an electron gun, drift tube, and movable probe, all enclosed in a demountable, continuously-pumped vacuum system. Outside of the vacuum envelope there is a shielded solenoid, extending the entire 18-inch length of the drift tube. The annular gap between the solenoid pole face and the magnetic shield about the gun is nearly all taken up by a soft-steel section of the vacuum envelope.

The electron gun is of the convergent Pierce type, with oxide-coated cathode and a coiled-coil filament heater producing negligible flux at the cathode surface. Surrounding the gun, and inside of the magnetic shield, is a small copper-wire coil that permits variation of this flux over a small range, either aiding or opposing the leakage flux due to the main focusing solenoid. The flux density at the cathode has been approximately calibrated in terms of currents in both coils. Throughout the experiments described below, the gun is pulsed with a 1,000 cps square wave of 2,200 volts on its anode, supplying 38 $\pm$ 1 ma peak current in space-charge-limited emission.

The novel feature of the probe is that its annular RF pickup gap couples to a 50-ohm coaxial line leading to the receiver, rather than to a resonant cavity. This permits RF power measurements over a wide range of frequencies. The inner conductor of the coaxial line serves as current-collector, being isolated and biased positively about 40 volts with respect to the outer conductor to prevent escape of slow secondaries. An adjustable vane can be locked in position in front of the probe (whose entrance aperture is 0.100 inch in diameter), so that circular apertures of various smaller sizes are fixed on the probe centerline, about 0.070 inch in front of the probe. With these apertures, measurements of collector-current variations along the beam furnish a rough picture of beam-ripple amplitudes and locations. In addition, the current-density variation across the beam can be estimated by moving a pinhole aperture in a broad arc through the beam centerline. Both the inner conductor of the probe and the intercepting vane are liquid-cooled.

The noise powers coupled to the coaxial probe are considerably smaller

than for a tuned coaxial cavity, because of the lower RF gap impedance of the former. To compensate for this drawback, a sensitive noise receiver is employed, similar in principle to the radiometer invented by R. H. Dicke.[9] The input noise power is replaced periodically by a matched load at room temperature by pulsing the beam on and off with a 1,000 cps square wave, and placing an isolator in front of the receiver. A synchronous detector eliminates gain-fluctuation noise and converts the receiver output to a dc voltage.

Noise power variations at various microwave frequencies are measured in terms of the changes of attenuation, between probe and receiver, required to keep the receiver output constant. These rapid adjustments in attenuation are performed by a servo amplifier-motor loop, and recorded on a chart, whose speed ($1\frac{1}{2}$ inches per minute) is synchronized with that of the moving probe. In the same way, records of collector current as a function of probe position can be obtained, and correlated with those of noise power. The probe can be moved a distance of about 17 inches, its position nearest the gun ($z = 0$ inches) corresponding to a distance of 0.95 inch between the anode and the input plane of the RF gap.

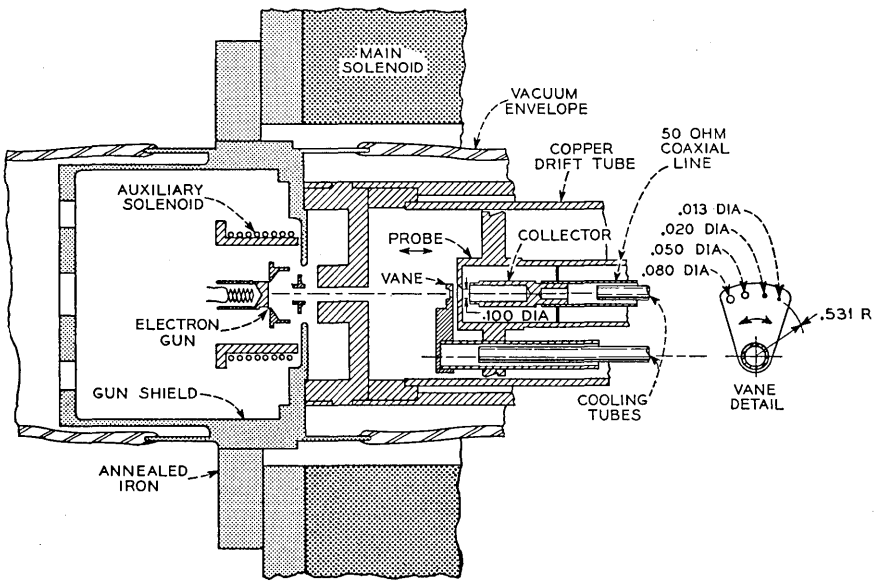

FIG. 1 — Cross-section of experimental tube, showing electron gun, probe, and two solenoids. The isolated current-collector electrode serves as inner conductor of a coaxial line. The induced RF power can be measured over a wide range of frequencies.

III EXPLORATORY MEASUREMENTS

The electron gun used in these experiments had been designed for use in a helix traveling-wave tube with a longitudinal focusing field of 600 gauss. Noise-power and collector-current curves, therefore, were first taken with 600 gauss to study a typical state of affairs in an operational beam. As seen in Fig. 2, the noise power at 3.9 kmc varies periodically with distance for about 4 inches from the gun, then climbs rather smoothly by nearly 23 db to an irregular plateau, where it undulates slowly, and finally levels off. The initial part of the growing noise curve at 10.7 kmc is missing because of inadequate receiver sensitivity, but its later portion is similar to that at 3.9 kmc, with about half the rate of noise climb. With the 0.020-inch aperture, the collector-current variations decrease in amplitude chiefly in the drift region preceding the noise climb; whereas those for the 0.100-inch aperture decrease afterwards. Both curves show a flattening in the growing-noise region itself, as well as a decrease in their *average* values after that region, signifying an increase in the average beam diameter.

A similar set of curves is shown in Fig. 3, for a focusing field of 279 gauss (about twice the nominal Brillouin field). Noise growth at 3.9 kmc starts later, and proceeds less steeply, than at 600 gauss. The noise-power curve for 10.7 kmc is much more articulated, with a semblance of periodicity, throughout the drift region. Collector-current curves for both 0.020- and 0.050-inch apertures show considerable reduction in current-ripple amplitude with distance; reaching virtually zero in the former case.

Another type of survey measurement is illustrated in Fig. 4. With the probe stationary at the far end of the drift space (about 18 inches from the gun anode), the main solenoid current is varied smoothly to change the focusing field from 0 to over 600 gauss, and synchronized records are made of collector current and noise power. (In this instance, the current in the auxiliary solenoid was +3.2 amperes.) At low magnetic fields, both the current and noise-power curves have large amplitude variations, which diminish as the field increase. At first glance, the noise peaks and valleys seem to coincide with those of collector current; certainly, some do. Closer inspection, however, reveals significant misalignments which cannot be accounted for by experimental error. When the three noise curves, at 3,050, 3,930, and 4,730 mc, respectively, are compared with each other, some characteristic features emerge:

(1) An average curve drawn through each pattern has one or two broad maxima, which tend to move toward higher field strengths with increasing frequency.

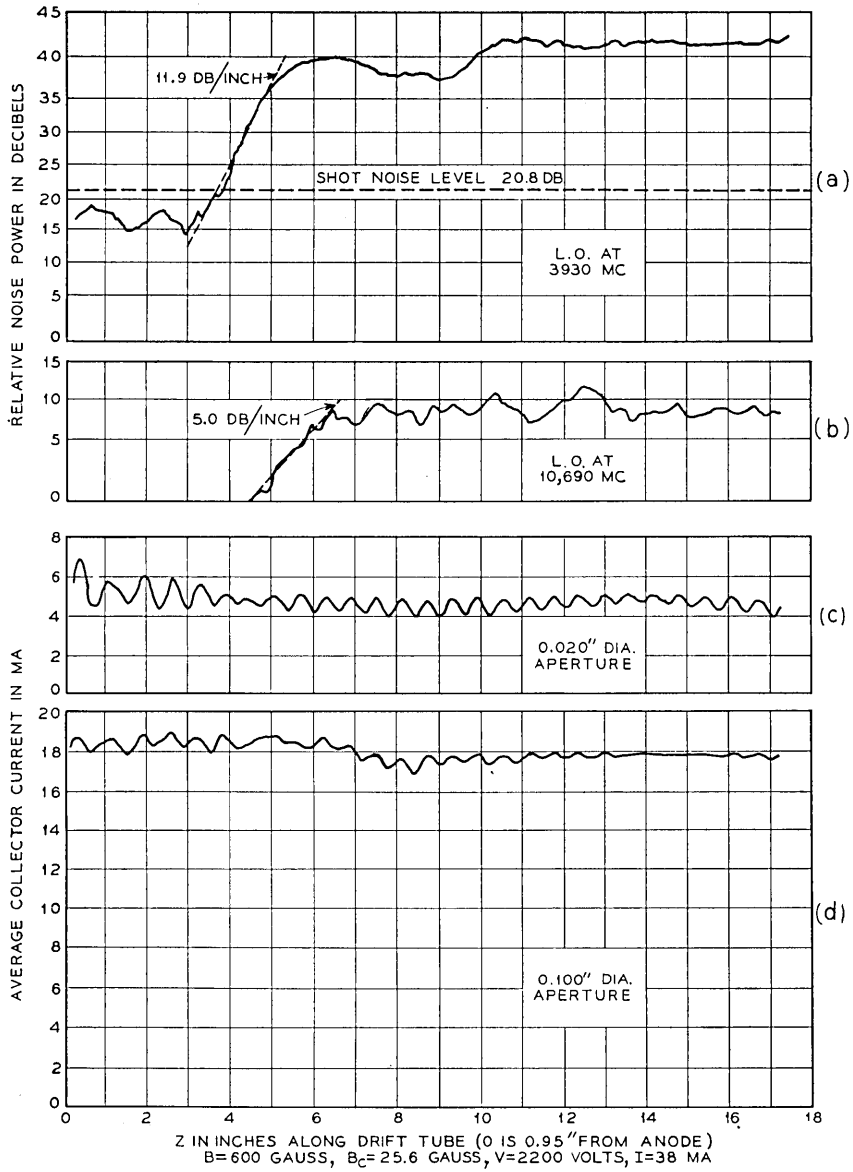

FIG. 2 — Typical smooth steep growing-noise patterns, near 4 and 10.7 kmc, respectively, with customary focusing field of about four times the Brillouin value. Collector-current traces through small and large apertures reveal decreases in ripple amplitude and increase in average beam diameter.

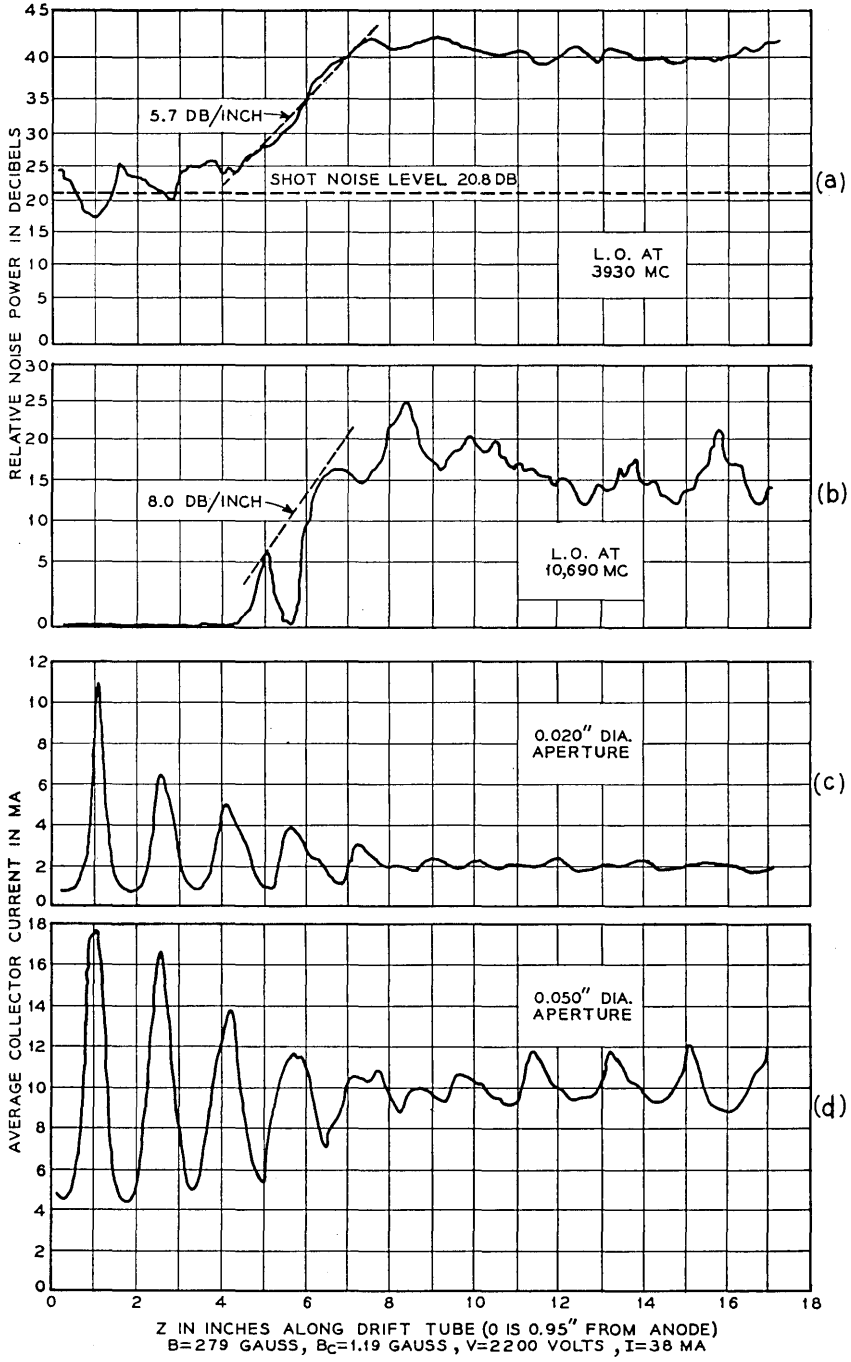

FIG. 3 — At about half the focusing fields used in Fig. 2, the growing-noise pattern is much the same at 4 kmc, but shows significant articulation at 10.7 kmc. The collector-current traces show pronounced decreases in ripple amplitude, and differences in ripple patterns obtained with different apertures.

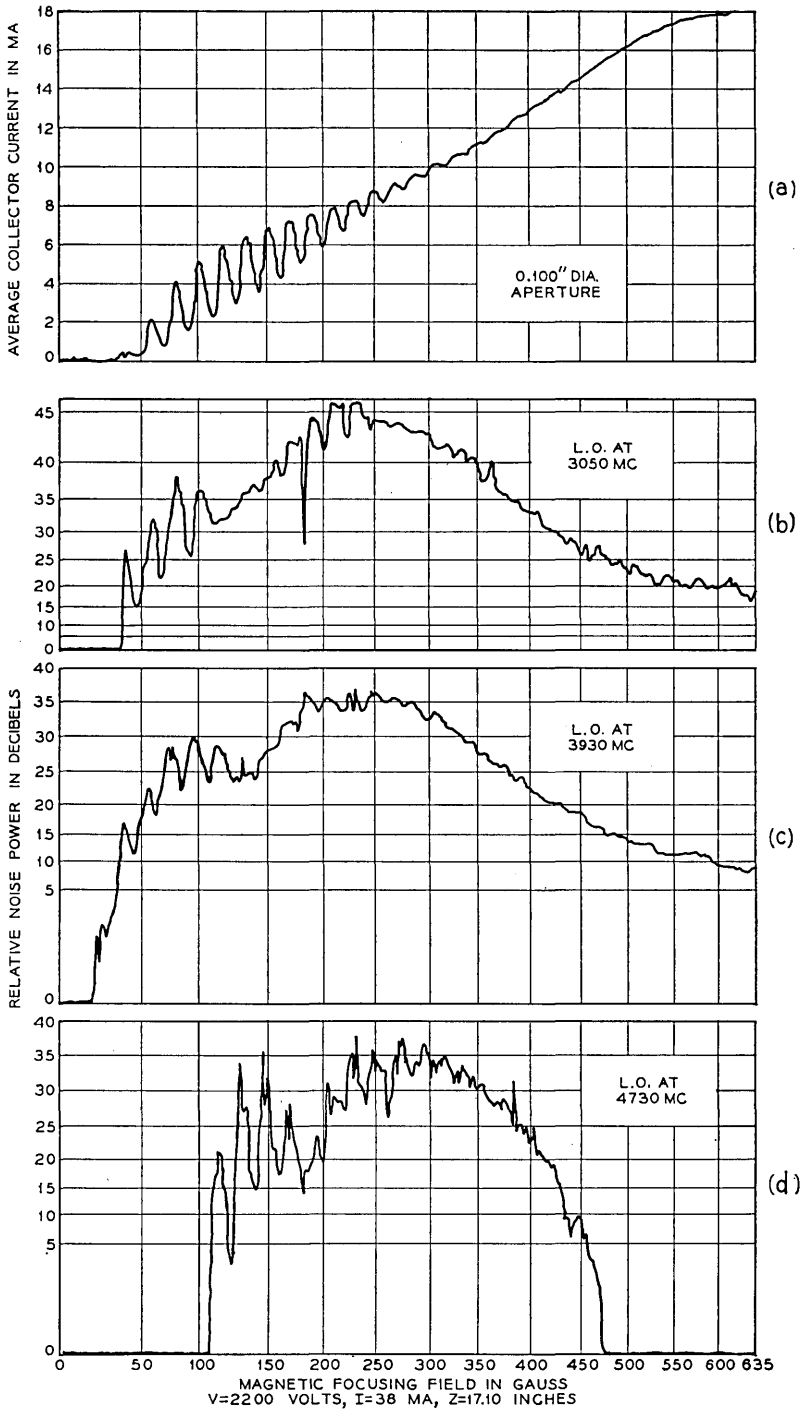

FIG. 4 — With the probe stationary at about 18 inches from gun anode, collector current and noise powers in bands about 1 kmc apart are recorded, as current in the main solenoid is varied from 0 to 1.5 amperes (0 to 635 gauss).

(2) The lower the frequency, the lower the field strengths at which noise amplitudes change most violently with field.

(3) The three noise curves resemble each other in small details.

The results of a great many records of the kind illustrated by Figs. 2 and 3 can be summarized as follows:

(1) There is always a decrease in beam-ripple amplitude associated with noise growth at any frequency. (Sometimes the ripple amplitude increases afterwards, as in Fig. 3.)

(2) The higher the frequency, for a given field, the more articulated or scalloped the noise pattern.

(3) No correlation can be found between rate of noise growth and either (a) distance from gun to take-off plane, or (b) net gain at the end of the drift region. The trends, as a function of magnetic field, are different at different frequencies.

(4) Greatest noise growth does not, as a rule, occur with zero flux threading the cathode. Sometimes two nearly equal peaks occur for two values of B_c, each of opposite polarity, referred to the sense of the main field.

(5) The noise-distance patterns change very slowly with frequency.

(6) No beam entirely ripple-free throughout its length has ever been observed by the writer.

IV ORIGIN OF GROWING NOISE

If noise growth is due to some amplification process, it should be possible to adjust the beam-focusing conditions so that the noise currents start increasing at the anode, and attain the greatest possible over-all gain at the end of the drift space. The enhanced activity of the unknown gain mechanism should presumably help identify it. The curves of Fig. 4 show that maximum noise occurs at different values of the focusing field, for different values of field at the cathode, and different probe positions. With the anode voltage and receiver frequency fixed, therefore, the conditions for greatest net noise growth can only be found by a series of trial settings of *both* magnetic fields, each followed by a recording of the noise-distance pattern. Eventually, a set of fields can be found for which the greatest total gain occurs; and such patterns are usually found to show fairly steady noise-amplitude increase, on the average, over the entire length of probe travel.

The results of this procedure for noise power near 4 kmc, as well as the patterns of collector-current versus distance with the same fields, using the 0.100-, 0.050-, and 0.020-inch apertures fixed at the probe centerline, are shown in Fig. 5. A similar set of records, for noise power

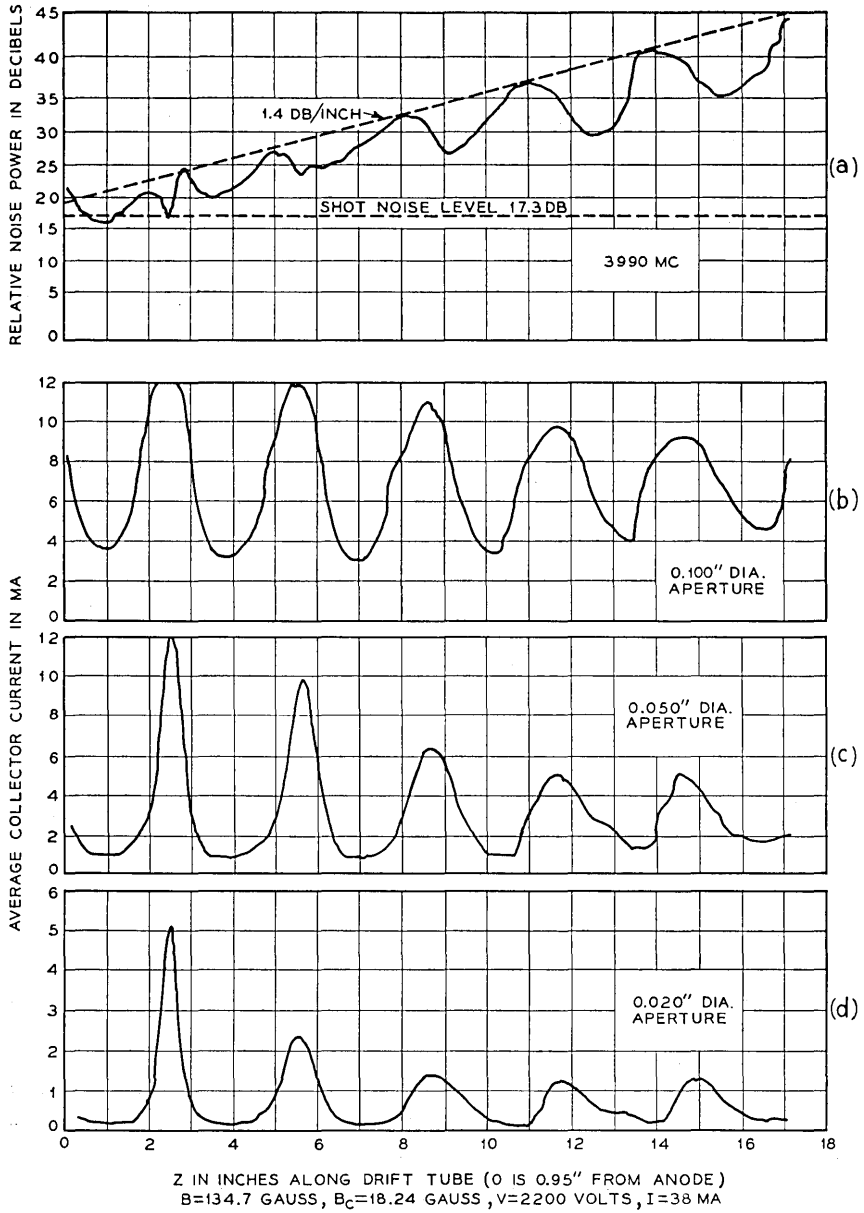

FIG. 5 — The magnetic fields in drift tube and at the cathode have been adjusted empirically to expand the growing-noise region over the entire drift region, with the L.O. at 3,990 mc. The field is slightly less than the Brillouin value, but the beam is strongly rippled because the gun was designed for best focusing at a much higher field. The noise-current maxima align with the average collector currents on their increasing slopes, for all three aperture sizes.

near 10.69 kmc, is shown in Fig. 6. The significant features of both sets of records can be summarized as follows:

(1) In both cases, the beam-ripple periods are equal to the *RF* scallop periods; i.e., the half-wavelengths of the space-charge standing waves. The noise minima tend to occur at planes where the collector currents are at their average values and decreasing; i.e., where the beam diameters are at their average values and increasing. The noise-current maxima occur where the beam diameters are about to decrease. These are the classical conditions for *rippled-beam amplification.*[6, 7, 8]

(2) In Fig. 5, the ripple amplitudes and peak values of all three collector-current curves decline appreciably with distance, the rate of decline being greatest for the smallest aperture. (Similar curves, not shown here, have displayed little or no such decline in the absence of noise growth.) This suggests that the RF noise power is amplified at the expense of dc energy associated with radially-directed electron velocities.

(3) In Fig. 6, the disparity among rates of decline of current-ripple amplitudes and their peak values, for the three aperture sizes, is even more pronounced. In addition, the ripple wavelength barely changes for the 0.100-inch aperture, but increases with drift distance for the smaller apertures, resulting in an increasing "phase shift" among them. Thus the current-density variations at different radii in the beam can contribute unequally to space-charge wave amplification, depending on their local ripple amplitude and phase. In this instance, the variations in current density along the beam are initially greatest near the axis, and suffer the greatest reduction there. It is worth noting that this "inner rippling" would be missed entirely in beam-size measurements with a large aperture.*

The decrease of beam ripple and the increase in average beam diameter, shown in Figs. 5 and 6, has been found to accompany rippled-beam amplification of impressed signals by T. G. Mihran.[8] Another corroboration of the identity of this gain mechanism can be obtained by comparing the measured noise gain per scallop with that predicted by theory for idealized conditions.[6, 7] For a beam with stepwise alternations of maximum and minimum beam diameters (ratio r_2/r_1), and with noise maxima and beam-diameter maxima coinciding, the gain per scallop is as follows:

$$G_m = \left(\frac{V_1}{V_2}\right)^{3/4} \left(\frac{r_2}{r_1}\right) \left(\frac{p_1}{p_2}\right). \tag{1}$$

Here, V is the beam potential, and p the reduction factor ω_q/ω_p. Although the actual rippled beam is far removed from either Brillouin or

* More information about "inner rippling" will be presented in Part II.

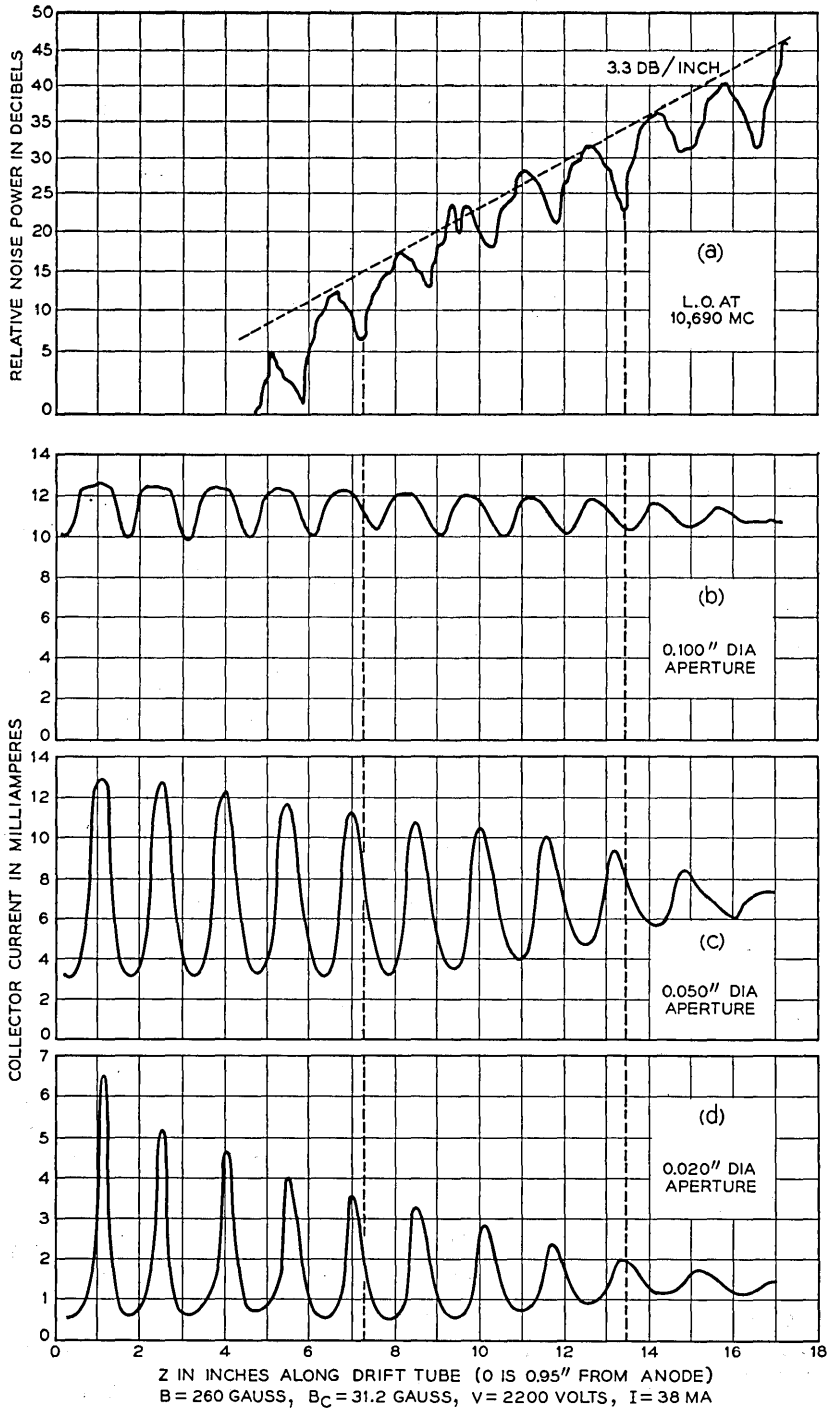

FIG. 6 — The fields have been adjusted as in Fig. 5, for maximum extension of the gain region, for noise power near 10,690 mc. Collector current measured with the smallest aperture shows the greatest decline in amplitude of variations, as well as advance in ripple phase relative to the current through the largest aperture.

TABLE I — MEASURED VERSUS CALCULATED MAXIMUM NOISE GAIN

Freq. mc.	Ripple Data		Gain in db per Scallop		
	Iris Dia. Inches	r_2/r_1	Brillouin Flow	Confined Flow	Measured
3,990	0.020	3.4	6.1	5.0	4.3
	0.050	3.2	6.3	5.4	
	0.100	1.9	4.0	3.3	
10,690	0.020	2.5	6.4	5.6	4.9
	0.050	1.8	4.8	4.4	
	0.100	1.1	<1.0	<1.0	

confined flow, the published values of reduction factor for both extremes can be used as first approximations.[10, 11] The ratio r_2/r_1 can be estimated by assuming the current density to be uniform over the beam cross-section near the middle of the drift region, for each of the three apertures used. The potential variations can be neglected. The results of such calculations are given in Table I.

As the computed gains are expected to be somewhat greater than those measured, because of the optimum conditions assumed, the best correspondence between measured and computed gain rates appears to be for the ripple data taken with the 0.050-inch iris at 3,990 mc, and that with the 0.020-inch iris at 10,690 mc. This distinction is in accord with previous qualitative comparison of Figs. 5 and 6, showing that most of the beam cooperates in the ripples of the former, but that "inner rippling" characterizes the latter.

Another calculation that reveals which part of the beam is interacting with the RF noise field in each case is that of the space-charge half-wavelength, as follows:

$$\frac{\lambda_s}{2} \cong \frac{\pi b}{p \cdot \beta_p b}, \tag{2}$$

where

$$\beta_p b = 174 I^{1/2}/V^{3/4}. \tag{3}$$

Here β_p is the plasma wave number, b the beam radius, and p the reduction factor, which can be evaluated as previously for the smooth beam in either ideal Brillouin or confined flow. For the gun used here, the square root of the perveance is

$$I^{1/2}/V^{3/4} = 0.606 \times 10^{-3} \text{ MKS units},$$

or

$$\lambda_s/2 \cong 29.8\, b/p. \tag{4}$$

TABLE II — MEASURED VERSUS CALCULATED SPACE-CHARGE HALF-WAVELENGTHS

Freq. mc.	Iris used, inches dia.	Avg. beam radius r_0 inches	γ rad./in.	$\lambda_s/2$, inches			Ripple Wavelength L, meas'd
				Brillouin Flow	Confined Flow	Meas'd	
3,990	0.050	0.057	22.9	3.2	2.7	3.0	3.06
10,690	0.020	0.033	61.2	1.6	1.3	1.47	1.52

Thus, agreement between this expression and the measured value requires the correct choice of the effective beam radius, b. It turns out that the suitable value for Fig. 5 (3,990 mc) is the average beam radius obtained from ripple data taken with the 0.050-inch iris, and that for Fig. 6 (10,690 mc) is obtained with data taken with the 0.020-inch iris. The results are summarized in Table II.

With this mechanism as the primary source of the noise gain, it becomes clear why nearly equal noise maxima were found, with some values of the main focusing field, B, for two values of cathode flux density B_c of opposite polarity. From approximate analyses of beam ripples when flux threads the cathode, such as those provided by McDowell[12] and others, it is found that the ripple wavelength depends nearly altogether on B. Its amplitude and spatial phase, however, depend on B_c, as this affects the beam geometry at the drift-space entrance. For a sufficiently wide range of variation of B_c, the spatial phase of the ripples can be varied from the proper relation with the space-charge standing wave for gain, through the positions for de-amplification, and back to gain again.

V THE GROWING-NOISE MECHANISM

Although many earlier noise records can be understood in the light of the rippled-beam amplification (RBA) process, this is not yet true of the smooth, steep noise growth usually observed, as in Figs. 2 and 3. The simple theory predicts that a space-charge wave will be amplified when, for small ripples, a "resonance" condition exists between the ripple wavelength, L, and the space-charge half-wavelength

$$L \cong n\lambda_s/2, \tag{5}$$

where n is an integer, usually unity. In addition, as mentioned earlier, there is an optimum phase relation between ripple and standing wave for maximum gain. These conditions are not satisfied by the records of

Figs. 2 and 3, except possibly for the 10.7 kmc noise current in Fig. 3 (at a relatively low magnetic field).

To establish a connection between the two types of noise growth, the noise record of Fig. 6 (for 10.7 kmc with greatly expanded gain region) is compared with that near 4 kmc under the same conditions, in Fig. 7. The growing noise region for 4 kmc does not start until at least four scallop wavelengths past the earliest observed 10.7 kmc noise growth. Moreover, the 4-kmc noise pattern resembles that for 10.7 kmc in many details. (The resemblance in details of noise patterns at nearby frequencies has been remarked before, in connection with Fig. 4.)

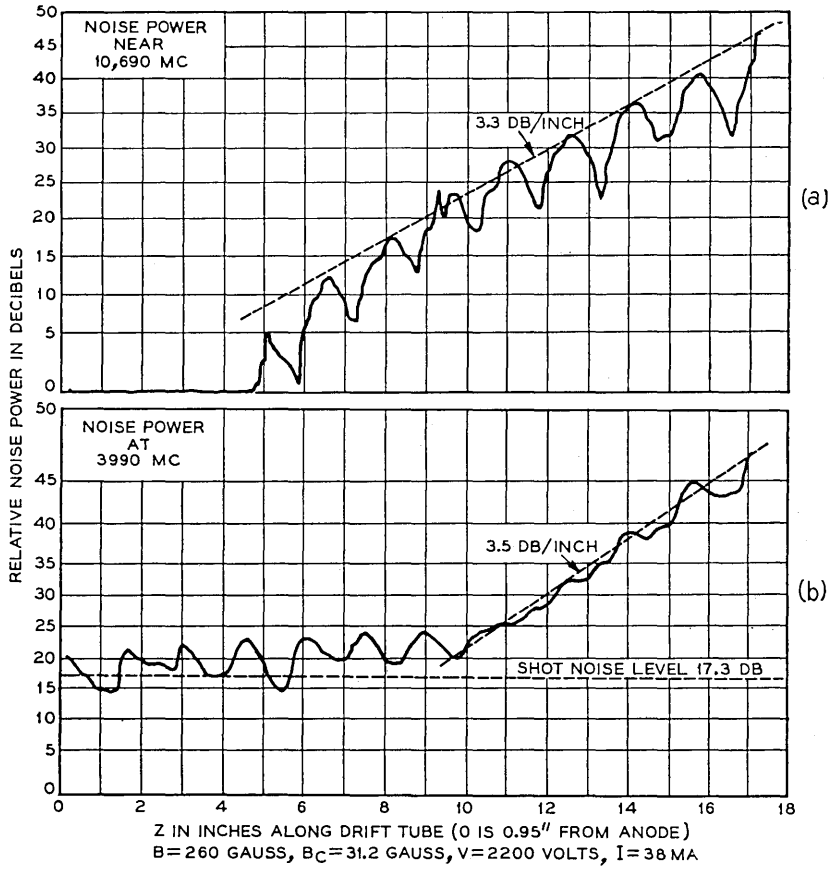

FIG. 7 — The pattern of growing noise in Fig. 6, near 10,690 mcs, is compared with that near 3990 mc for the same fields. The gain region of the latter curve starts much later and is much smoother than the former. The small irregularities on the 3,990 mc curve resemble the scallops of the 10,690 mc curve, in a blurred way.

This information suggests that the growing-noise mechanism is really a two-stage process: amplification of a broad band of microwave frequencies, located far above the observation frequency, followed by a transfer of noise energy to lower frequencies.

(*a*) *First stage*

At low magnetic fields, there are few ripples per unit length, but their amplitude is usually large; whereas at moderate to large magnetic fields, the ripple amplitude is small, but so is the ripple wavelength. In either case, the bandwidth of RBA is large, usually many thousands of megacycles.

The increase of both bandwidth and gain per scallop with ripple amplitude has been explained by Pierce[13] by analogy between the gain band of a rippled beam and the stop band of a transmission line filter: a sharply varying periodic disturbance on the latter will reflect short as well as long waves, whereas smoother perturbations will not reflect the shorter waves to any extent.

Another way to study the amplification bandwidth is to derive the equations for RF current in a one-dimensional beam with sinusoidal variation of the reduced plasma wave number, $\beta_q = p \cdot \beta_p$, as Heffner,[14] Bloom,[15] and others[16, 17] have done. This leads to a Mathieu equation, whose solutions may be studied on the Mathieu stability plot (A, q):

$$\frac{d^2I}{dx^2} + (A - 2q \cos 2x)I = 0. \tag{6}$$

Here I is the RF current, q a measure of the perturbation amplitude, $x = \pi z/L$, and $A = (2L/\lambda_q)^2$, where L is the ripple wavelength and λ_q the reduced plasma wavelength. Bloom has shown that, if n is the integral number of scallop wavelengths between initial and final planes, the greater the *product* nq, the greater the total amplification or deamplification, and the less critical the phase relation between RF standing wave and ripple for amplification. Ultimately, for very large nq, amplification will take place for all values of this phase angle.

At higher magnetic fields, both the ripple amplitude and ripple wavelength are decreased. This means that q is reduced, but n increased over any fixed span. This combination usually tends to increase the product nq up to some fairly high field, after which it may decline. More important, the reduction in ripple wavelength shifts the band of amplification to higher frequencies, and greatly increases the frequency band. This occurs because the "resonant" space-charge wavelength is shorter,

and short space-charge wavelengths correspond to high frequencies, where the former change very slowly with frequency.

(*b*) *Second stage*

When noise power over this large band has been amplified sufficiently, electron bunching becomes non-sinusoidal, and the beam becomes non-linear. Harmonics and beat-frequencies[18] of the fundamentals, and possibly sub-harmonics,[19] are excited. As the beat-frequencies are excited by a continuum of pairs of frequencies, their standing-wave patterns overlap one another, resulting in a "wash-out" of the noise minima, and a smooth growing-noise pattern. Eventually, the same non-linear processes take place within this subsidiary band, leading to a gradual leveling of the entire noise spectrum. The *initial* rates of rise of the intermodulation products, however, should take place closer to the gun and be greater, for a *lower* frequency. They will depend on both the spectrum of noise power in the primary band and, so to speak, the spectrum of "beam non-linearity" within that band.

To simulate this intermodulation process, two low-level klystron signals (9,050 and 12,275 mc, respectively) were simultaneously permitted to modulate the electron beam as it entered the drift tube, by means of a short length of lossy helix. The magnetic fields at the cathode and in the drift space were adjusted to produce a beam ripple which amplified both of these signals simultaneously over most of the drift space, as shown in Fig. 8 (a, b, c).* Noise-power records were then made at the difference-frequency, in the presence of the two modulation signals, Fig. 8(d), and in their absence, Fig. 8(e). The difference between the noise levels in the latter two records increases with distance, as both parent space-charge waves grow in amplitude, and the degree of beam non-linearity increases. Naturally, the contribution to 3,295-mc noise in the absence of modulating signals is far greater than that of the latter two alone, as the primary bandwidth of noise amplified is very great, and that of the signals very small.

There are several reasons why exponentially-growing noise should stop growing and level off, and sometimes even decrease slightly:

(1) depletion of dc kinetic energy in the beam ripples;

(2) de-amplification in the fundamental band, due to departure from the proper phase relation for gain between standing wave and ripple, if only over part of the band (Fig. 3);

* The fine ac detail superimposed on the pattern of Fig. 8(b) is due to interference between the waves traveling along the beam and that propagating as a waveguide mode in the drift space.

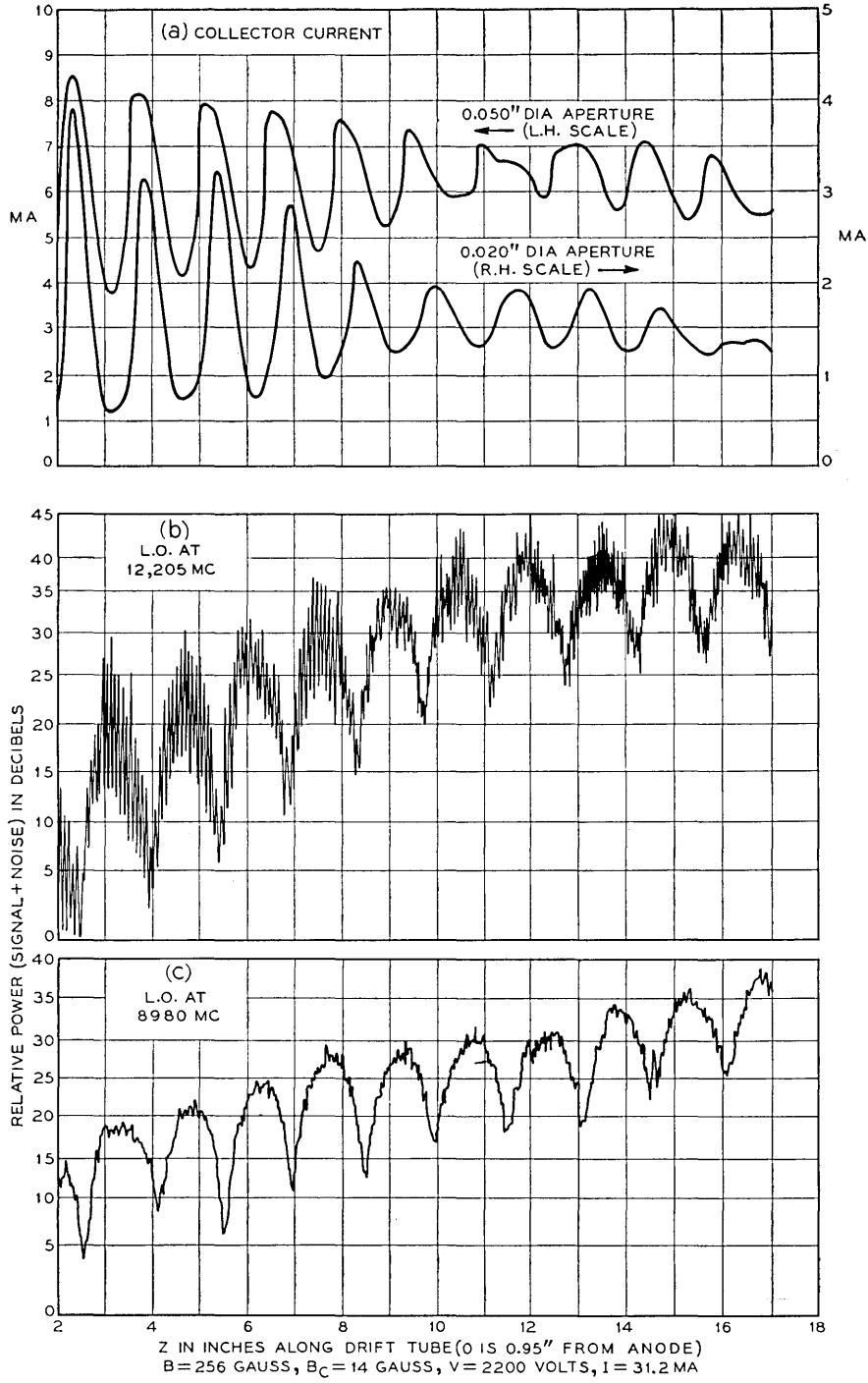

FIG. 8 (a), (b), (c) — The fields have been adjusted to give rippled-beam amplification of two weak klystron signals, 12,275 and 9,050 mc, simultaneously impressed on the beam at the entrance to the drift region.

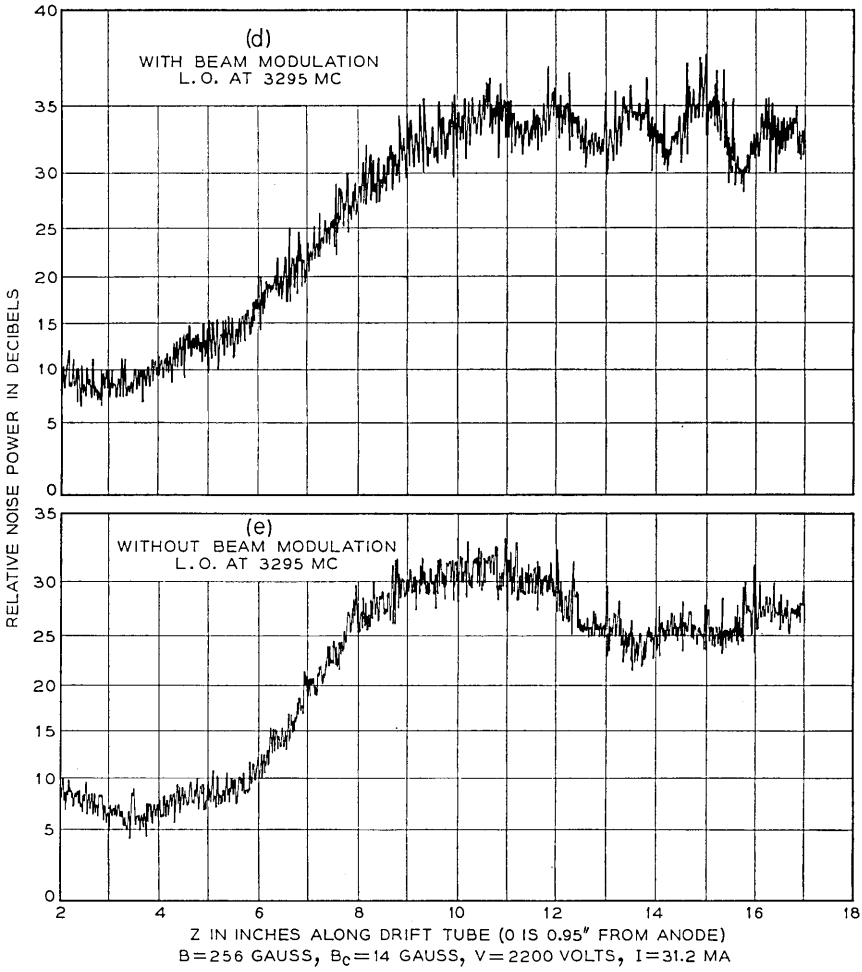

FIG. 8 (d), (e) — Noise power at the difference-frequency, 3,225 mc, is recorded, with and without the signals (b) and (c) present. The difference in ordinates of the two curves increases with distance from the gun, as the impressed signals grow and the beam becomes more non-linear.

(3) sufficient phase shift between inner and outer ripples in the beam for one to de-amplify as much as the other amplifies; and

(4) interference among the intermodulation products excited at different positions along the beam.*

The last effect is illustrated in Fig. 9, in exaggerated form. The beam

* Suggested by C. C. Cutler of Bell Telephone Laboratories.

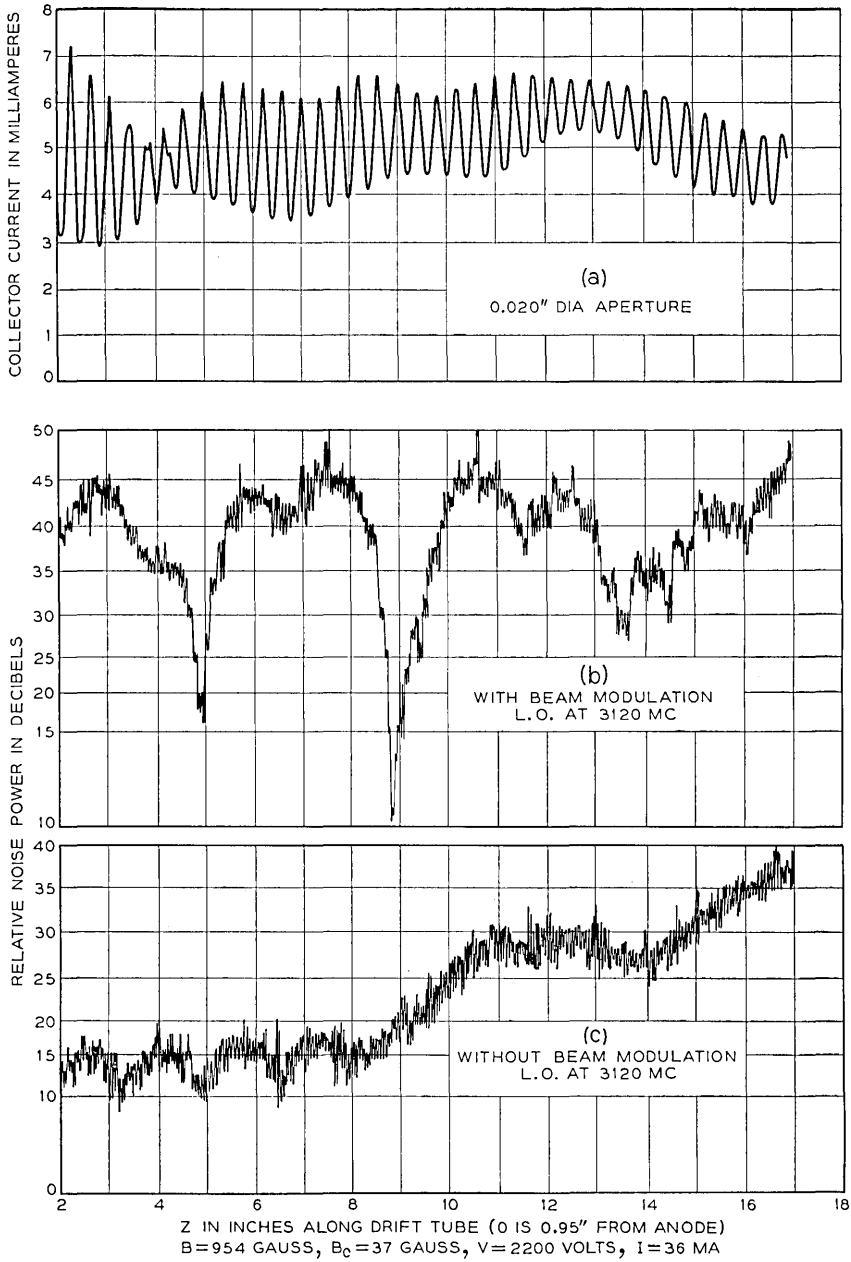

FIG. 9 — Two strong klystron signals are impressed on the beam as in Fig. 8, and noise power at their difference-frequency recorded, both with and without these signals present. The deep noise minima in (b) are due to destructive interference between trains of waves excited by intermodulation at different positions along the beam.

is simultaneously modulated as before with two klystron signals (8,400 and 11, 590 mc, respectively), but now at fairly high level; and the focusing field is made large. The interference dips in the pattern of 3,120-mc noise are quite deep, and are spaced irregularly and farther apart than the space-charge wavelength of any of the three frequencies involved. The third dip is shallower than the previous two because of the growth of 3,120-mc noise other than that due to the signals, as shown in Fig. 9(c). The latter pattern of noise in the absence of the two high-level, high-frequency signals suggests that the characteristic first gentle dip following the growing-noise region is indeed of the same nature as the artificially-produced interference dips, and has nearly the same quasi-period.

The pattern of dips agrees with simple calculations, based on this model, in which the amplitude of the difference-frequency intermodulation product, excited at any plane ζ, is assumed proportional to the product of the amplitudes of the two high-frequency space-charge standing waves, as follows:

$$| di_3 | \propto | i_1(\zeta) \cdot i_2(\zeta)\, d\zeta |, \tag{7}$$

where

$$i_n = I_n \sin p_n\beta_p\zeta \cdot \sin \omega_n(t - \zeta/u), \qquad (n = 1, 2).$$

The total current at $\zeta = z$ is the sum of contributions from all the standing waves excited to the left of it:

$$| i_3 | \propto \frac{1}{4} I_1 I_2 \int_0^z \cos p_3\beta_p(z - \zeta)[\cos (p_1 - p_2)\beta_p\zeta - \cos (p_1 + p_2)\beta_p\zeta]\, d\zeta. \tag{8}$$

This expression is readily integrated and evaluated.

VI CONCLUSIONS

Synchronized measurements of electron-current density and noise currents at several microwave frequencies have shown that the "growing noise" pattern in drifting cylindrical beams is the result of a two-stage process. In the first stage, rippled-beam amplification of noise fluctuations takes place over a very broad band of microwave frequencies, much higher than the usual observation frequency. In the second, noise energy is transferred to lower frequencies by intermodulation and other non-linear processes within this band. The element of non-linearity is supplied when primary noise gain is sufficient to make electron bunching

non-sinusoidal. Other sources of non-linearity are thermal velocities, non-laminar beam flow, etc. As the beat-frequency noise increments at any plane are produced by continuous arrays of frequency pairs, increasing in numbers and amplitude in various ways as primary amplification proceeds, the multiple standing-wave patterns at the observation frequency progressively overlap one another. This results in the smooth steep rise of noise power usually observed.

Phase correlation among the space-charge waves excited at successive planes on the beam by the same set of frequency pairs is indicated by gentle dips, due to their destructive interference, in the plateau following the initial noise rise.

Rippled-beam amplification occurs whenever the ripple wavelength and half the space-charge wavelength are nearly equal, and bear a favorable spatial relation to each other. However, this "phase" relation becomes less critical with an increase in either the number of ripple wavelengths over which synchronism persists, or the ripple amplitude, or both. Noise amplification by this mechanism, therefore, is probably present to some degree in all rippled streams, particularly at high fields. The extreme difficulty encountered in focusing ripple-free beams from convergent, shielded guns has to this date prevented the detection of any other primary gain mechanism, which may conceivably co-exist in such beams.

A conspicuous feature of rippled-beam amplification is the decrease in ripple amplitude due to conversion of dc into ac kinetic energy. Such changes in beam structure emphasize the inadequacy of beam-flow computations based entirely on dc force equations. A more detailed description of this dc-ac energy conversion is given in Part II.

ACKNOWLEDGMENTS

The experimental apparatus could not have been built without the combined efforts of many associates of the writer, principally A. R. Strnad, P. Hannes, J. S. Hasiak and J. M. Dziedzic. The author is also indebted to R. Kompfner, C. F. Hempstead and K. M. Poole for valuable suggestions; and above all to C. F. Quate for constant encouragement and advice.

REFERENCES

1. C. C. Cutler and C. F. Quate, Experimental Verification of Space-Charge and Transit Time Reduction of Noise in Electron Beams, Phys. Rev., **80**, p. 875, 1950.
2. L. D. Smullin and C. Fried, Microwave Noise Measurements on Electron Beams, Trans. I.R.E. **ED-1**, No. 4, p. 168, Dec., 1954.

3. C. Fried, Noise in Electron Beams, Tech. Rep. 294, Research Laboratory of Electronics, M.I.T., May 2, 1955.
4. J. R. Pierce and L. R. Walker, Growing Waves Due to Transverse Velocities, B.S.T.J., **35,** p. 109, Jan., 1956.
5. G. G. Macfarlane and H. G. Hay, Wave Propagation in a Slipping Stream of Electrons: Small Amplitude Theory, Proc. Royal Soc. (B) **63,** p. 409, 1950.
6. C. K. Birdsall, Rippled Wall and Rippled Stream Amplifiers, Proc. I.R.E., **42,** p. 1628, Nov., 1954.
7. R. W. Peter, S. Bloom, and J. A. Ruetz, Space-Charge-Wave Amplification along an Electron Beam by Periodic Change of the Beam Impedance, RCA Rev., **15,** p. 113, March, 1954.
8. T. G. Mihran, Scalloped Beam Amplification, Trans. I.R.E., **ED-3,** No. 1, p. 32, Jan., 1956.
9. R. H. Dicke, The Measurement of Thermal Radiation at Microwave Frequencies, Rev. Sci. Instr. **17,** p. 268, July, 1946.
10. W. W. Rigrod and J. A. Lewis, Wave Propagation Along a Magnetically-Focused Cylindrical Electron Beam, B.S.T.J., **33,** p. 399, March, 1954.
11. G. M. Branch and T. G. Mihran, Plasma Frequency Reduction Factors in Electron Beams, I.R.E. Trans., **ED-2,** No. 2, 3, April, 1955.
12. Informal communication from H. L. McDowell.
13. Informal communication from J. R. Pierce.
14. Informal communication from H. Heffner.
15. S. Bloom, Space-Charge Waves in a Drifting, Scalloped Beam, unpublished RCA Research Laboratories report.
16. O. E. H. Rydbeck and B. Agdur, Propagation of Space-Charge Waves in Guides and Tubes with Periodic Structure, L'Onde Electrique, **34,** p. 499, June, 1954.
17. P. V. Bliokh and Y. B. Feinberg, Space-Charge Waves in Electron Beams with Variable Velocity, Zhurnal Tekhn. Fiziki, **26,** p. 530, March, 1956.
18. C. C. Cutler, The Nature of Power Saturation in Traveling Wave Tubes, B.S.T.J., **35,** p. 841, July, 1956.
19. S. Lundquist, Subharmonic Oscillations in a Nonlinear System with Positive Damping, Quarterly of Appl. Math., **13,** No. 3, p. 305, Oct., 1955.

Noise Spectrum of Electron Beam in Longitudal Magnetic Field

Part II — The UHF Noise Spectrum

By W. W. Rigrod

(Manuscript received January 21, 1957)

Sharp peaks are found in the UHF spectrum (10 *to* 500 *mc*) *of an electron beam, emanating from a shielded diode. In the presence of a longitudinal magnetic field, the strongly rippled beam displays an additional set of peaks whose frequencies are proportional to the field strength. The largest of these, just above the cyclotron frequency, is connected with the overlap of a dense cluster of particle orbits, passing close to the beam axis. It can attain amplitudes of 65 db above background noise.*

The transverse distribution of UHF noise power is found to agree with that for ideal Brillouin flow, even in rippled beams. With long ripple wavelengths, two noise maxima are found to flank each beam waist. A small-signal wave analysis explains this pattern, and affords some insight into the energy-exchange processes in rippled-beam amplification. The reduction in "growing noise" due to positive ions is attributed to increased cancellation of net radial beam motion, due to overlap in particle orbits near the axis.

I INTRODUCTION

The reader is referred to Part I[1] for a description of the experimental apparatus and its operation. In this paper, measurements of noise power in the same electron beam are described, with frequencies chiefly in the 10- to 500-mc range, and relatively weak magnetic fields. For the UHF measurements, a calibrated coaxial step attenuator and a super-regenerative receiver (the Hewlett-Packard 417-A VHF Detector) are used. Relative noise-power amplitudes at fixed frequencies are measured as before, in terms of changes in attenuation between probe and receiver required to restore constant receiver output. To obtain qualitative information, however, such as the location of noise maxima along the beam, the series attenuation is fixed. The receiver output is amplified, rectified, and per-

mitted to register itself directly on the chart recorder, whose motion is synchronized with that of the probe. Very roughly, the detector output varies as the log of input power.

Measurements are described (a) of the UHF noise spectrum in the beam, just outside the gun anode; (b) of this spectrum at the end of the drift region, in a longitudinal magnetic field; (c) of the noise-power distribution along the axis; and (d) transverse to the axis of the rippled beam in the drift region. Two calculations are then outlined, one of wave propagation along the rippled beam (to explain the observed distribution patterns), and the other to account for some spectacular peaks in the beam spectrum (b).

II FIELD-INDEPENDENT PEAKS

When the noise spectrum of an electron beam is scanned by a tunable receiver, it is found that an irregular array of narrow-band peaks characterize the UHF region, below about 1000 mc. Of these peaks, some are due to spurious modulation effects,[2] and can be eliminated as follows:

(1) Transit-time oscillations due to positive ions, secondary electrons, or both. Such frequencies vary with probe (collector) position.

(2) Resonances in the probe and receiver, excited by the pulsed-voltage supply. These are unaffected by changes in collector current.

(3) Ion oscillations in the electron gun or beam. Their frequencies vary with anode voltage.

The remaining narrow-band peaks fall into two classes, depending on whether their frequencies vary with the magnetic field.

Well-defined peaks can be detected with the RF probe stationed one inch from the gun anode, with or without any focusing field. When the beam is focused by a longitudinal magnetic field, these disturbances propagate along the beam, and tend to increase in amplitude with distance, but not to change in frequency. A typical set of such frequencies, within the range of the tunable receiver is as follows: 15.9, 24.3, 31.2, 34.0, 48.5, 63.4, 77.0, 108, 151, 166, 270.5, 372 and 481 mc. (During this measurement, the anode voltage was 2,200, and the peak current about 40 ma.)

No consistent relation could be found between these frequencies and either the anode voltage or the cathode temperature, although unmistakable frequency changes did occur when these parameters were manipulated. Failure to establish such a relation may have been due to uncontrolled drift in cathode activity. In any case, the measurements did serve to narrow the field of possible mechanisms, by eliminating the following:

(1) Transverse positive-ion oscillations,[3] for which the frequencies vary as the square root of anode voltage.

(2) Transverse electron plasma oscillations (near or beyond the anode), for which the frequencies would be too high.

(3) Longitudinal electron plasma oscillations at the potential minimum, for the same reason (should be near 2,500 mc).

(4) Longitudinal diode oscillations.[4] When the electron transit angle through the diode is approximately $(n + \frac{1}{4})$ periods, where n is an integer, the real part of the diode conductance becomes negative, permitting oscillations to occur. Again the frequencies of such oscillations would be too high, (2,200 mc and higher) for the gun used, to conform to the observed values.

There is, however, one published theory for which an order-of-magnitude correspondence does exist between the measured and calculated frequencies. Klemperer[5, 6] has shown that a strip beam tends to break up into clusters of "pencils" at the cathode. He ascribes these to standing waves resulting from transverse oscillations in the space-charge cloud, and offers an expression for the wave velocity in this medium. Application of his formula to the cathode used in the present experiments results in a least frequency of 31.3 mc. Other observers, such as Smyth[7] and Veith,[8] have also reported evidence of interaction between electrons in a retarding-field region and RF fields, which may underlie these oscillations.

III FIELD-DEPENDENT PEAKS

With the RF probe stationed ten or more inches from the gun anode, narrow-band peaks can be found in the noise spectrum of the beam. The amplitudes of these peaks increase and their frequencies decrease with decreases in the magnetic field. For each probe position, the process of finding the peak of greatest amplitude involves repeated adjustments of the focusing field, the magnetic field at the cathode, and the receiver frequency.

When the fields have been so optimized, it is found that the probe is located at or near the first beam-diameter minimum, following that at the entrance to the drift space. When the field is doubled, and the "tuning" process repeated, the greatest peak is found to have about twice the frequency of the first, and the probe is found to be located at or near the second beam waist. It is convenient, therefore, to think of these peaks as "proper" frequencies of the $N = 1$, etc., modes of the rippled beam, where N is the number of ripple wavelengths between gun and probe.

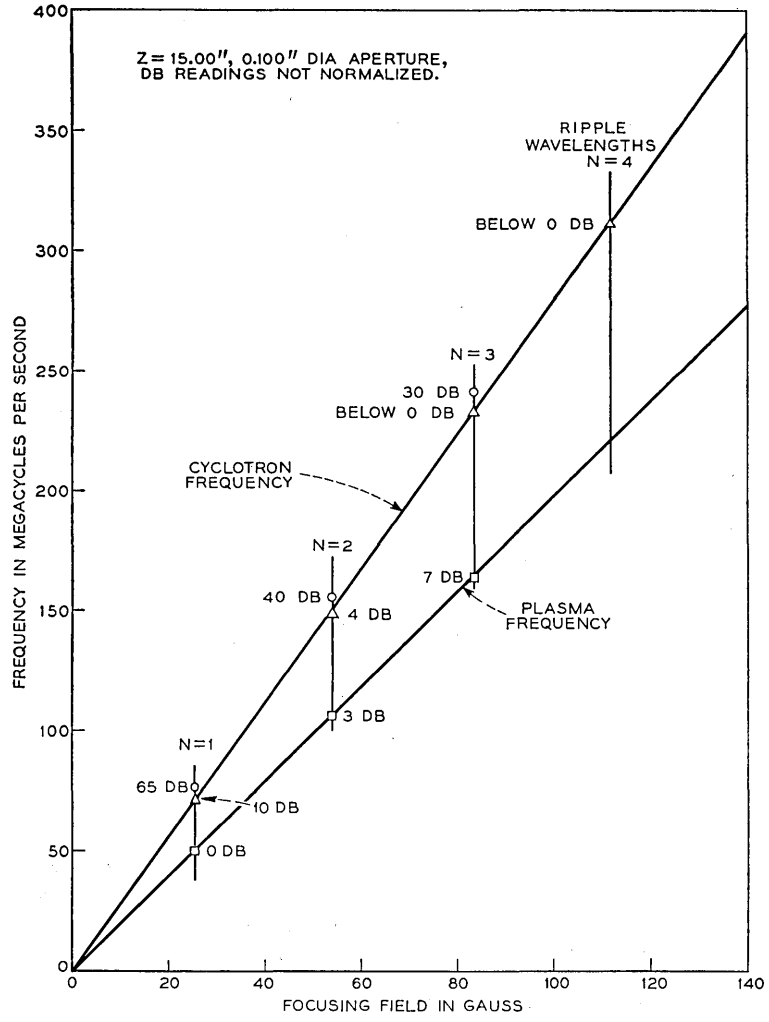

FIG. 1 — Frequencies and amplitudes of several narrow-band UHF peaks measured at a fixed probe position, about 16 inches from the gun anode. N is the number of beam-ripple wave-lengths between anode and probe. Other peaks have been observed at higher harmonics of the "proper" frequency (encircled points), and at about half that frequency.

As shown by the encircled points in Fig. 1, these frequencies range between 1.03 and 1.06 times the calculated cyclotron frequency, and have amplitudes as high as 65 db above the background noise. The amplitudes decrease with increasing N, falling off as the minus two-thirds power of the frequency.

At each of these optimum field settings, several weaker "satellite" peaks can also be detected, most readily those at the cyclotron frequency itself, and at 0.707 times the latter; i.e., the "plasma" frequency, as shown in Fig. 1. In addition, smaller peaks have been repeatedly observed at harmonics (up to the sixth) of the proper frequency, and one at slightly less than half of that frequency. (When a proper frequency was simulated by means of a signal generator, only its first harmonic could be detected in the receiver output.)

At the fields corresponding to $N = 4$ in Fig. 1, the cyclotron frequency (312 mc) was found, but not the proper frequency. The highest proper frequency observed was 240.5 mc, in the $N = 3$ mode. The proper-frequency peaks decrease with increasing focusing field, whereas the field-independent peaks excited in the electron gun tend to increase, at the far end of the drift region.

IV SPATIAL DISTRIBUTION OF UHF NOISE CURRENTS

In Figs. 2 to 5 are shown synchronized chart records of collector current, one or more UHF narrow-band peaks, and microwave noise power near 4,000 mc — all as functions of distance from the electron gun, for the $N = 1$ to 4 modes, respectively. In all runs, the beam was pulsed with a 1,000-cycle square wave, and the collector aperture set at a 0.100 inch diameter. The magnetic fields at the cathode and in the drift space were adjusted before each set of readings, with the probe at a common reference position, for greatest amplitude of some UHF peak. In Figs. 2 and 3, these were proper frequencies, whereas in Figs. 4 and 5 they were field-independent frequencies.

The content of these distribution curves can be summarized as follows:

(1) At the low fields employed (none quite equal to the nominal Brillouin value), the beam ripples are quite large, both in amplitude and wavelength.

(2) The proper-frequency traces have two or three maxima near each beam waist, and their amplitudes grow more rapidly with distance from the gun than any of the satellite frequencies.

(3) The patterns of the cyclotron and "plasma" frequencies do not differ significantly from those of the field-independent frequencies, and usually display two peaks near each beam waist.

(4) The collector-current maxima decrease with distance from the gun, although their minima change little. (The first maximum is sometimes flat-topped due to beam interception before it enters the drift space.) The rate of decrease of these maxima, and the rate of increase of proper-frequency amplitude, are greater, the longer the ripple wavelength.

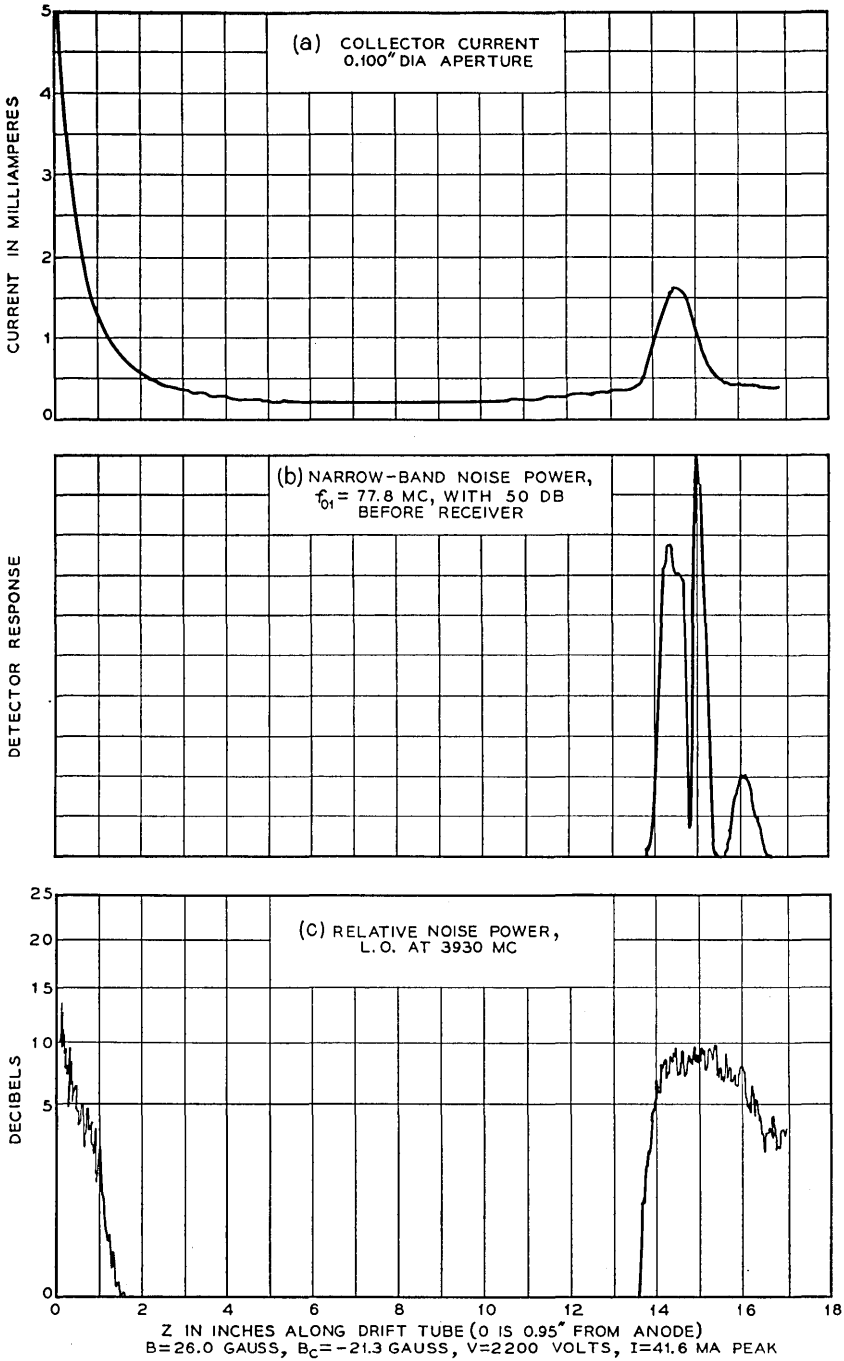

FIG. 2 — The fields have been adjusted for maximum amplitude of the $N = 1$ proper frequency, 77.8 mc, at a reference probe position ($z = 15$ inches). The synchronized probe records indicate three distinct maxima of this proper frequency near the beam waist.

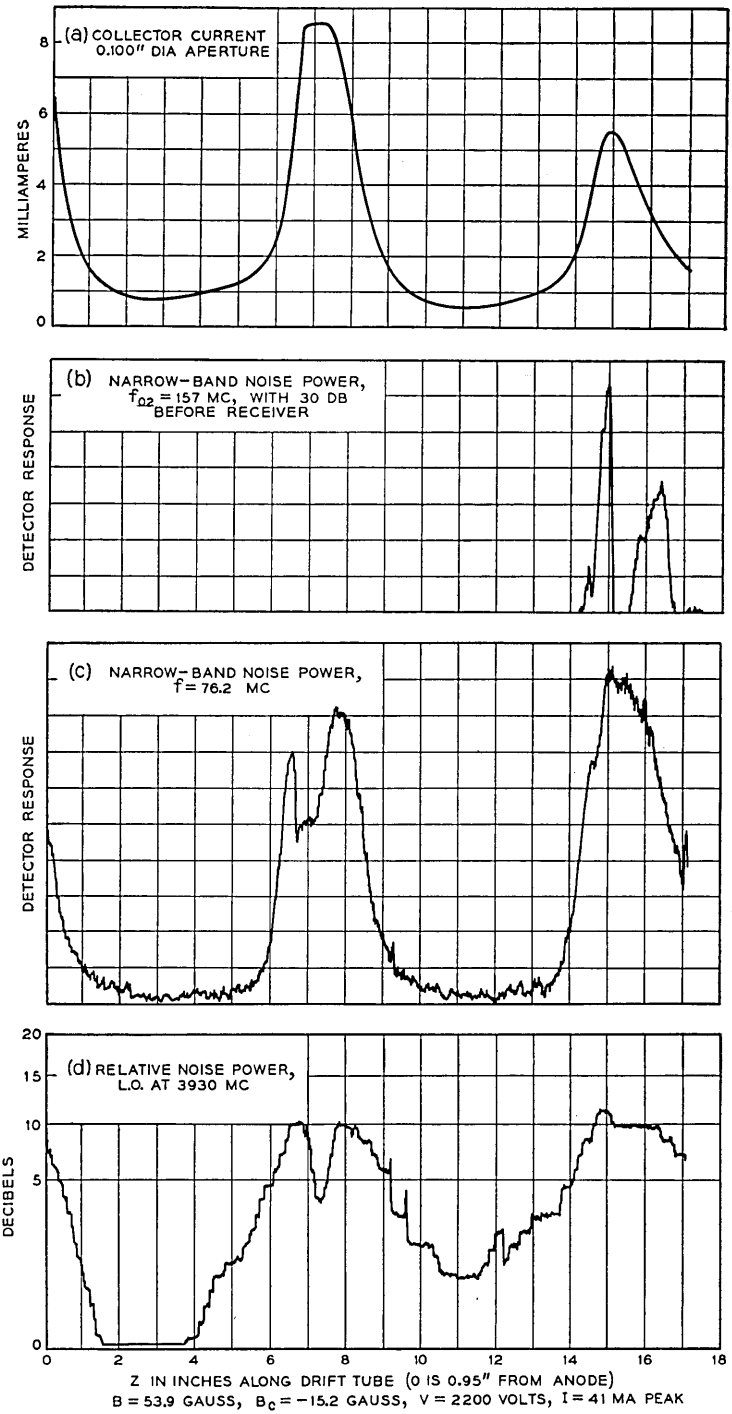

FIG. 3 — The longitudinal distributions of the $N = 2$ proper frequency, 157 mc, as well as its "satellite" 76.2 mc, and microwave noise power, are shown here, with fields adjusted for greatest amplitude of the proper frequency at $z = 15$ inches.

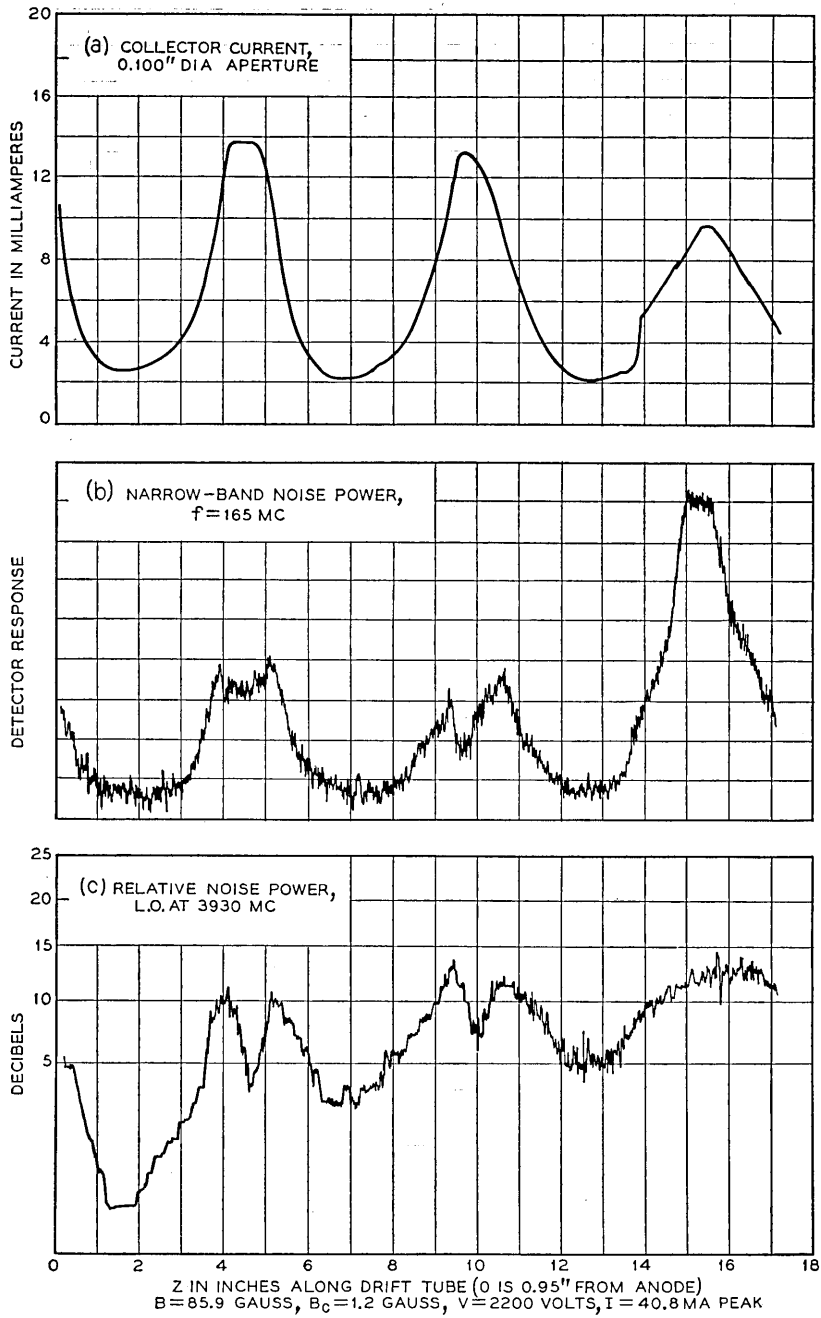

FIG. 4 — The fields have been adjusted for maximum amplitude, at the same reference probe position, of a wave excited in the diode, with frequency unaffected by the magnetic field, 165 mc. The cyclotron frequency for this field is 240.2 mc.

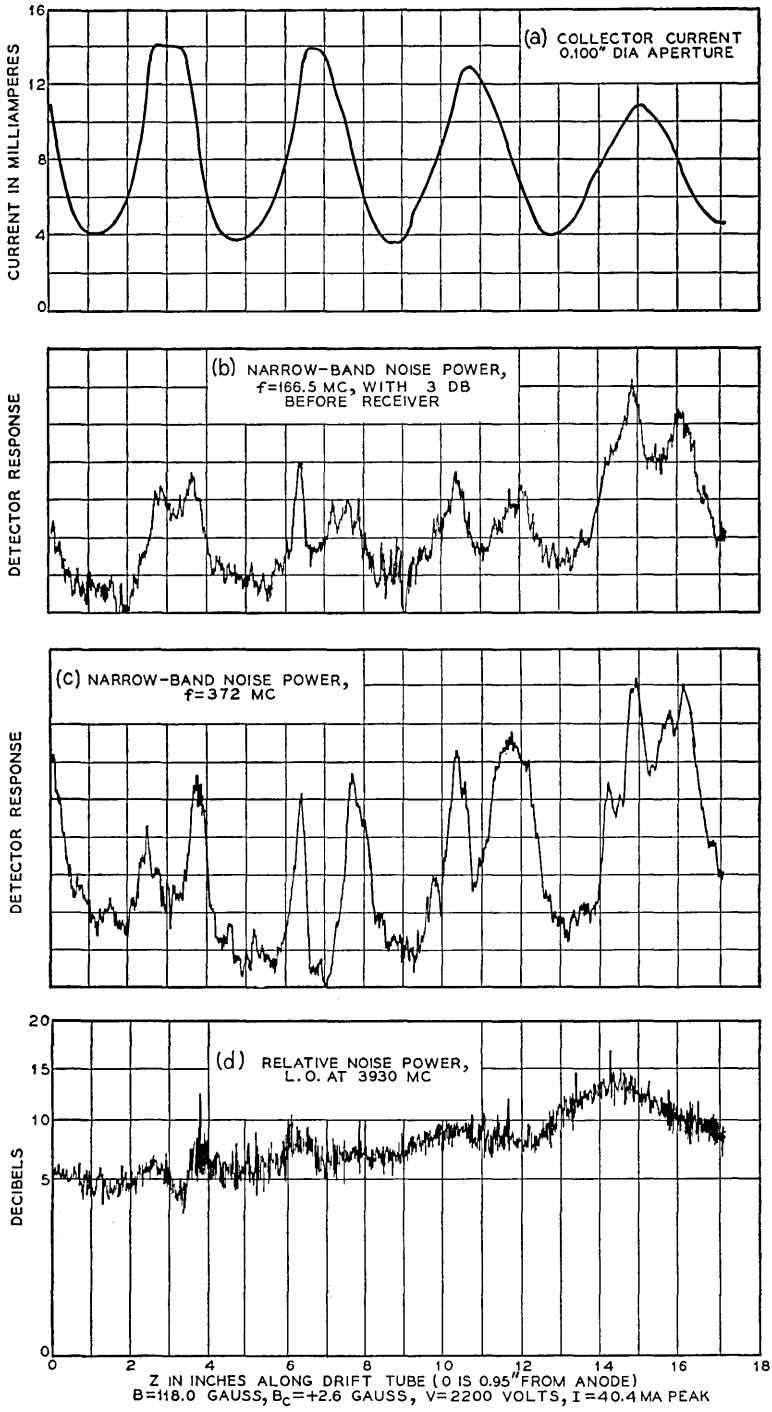

FIG. 5 — A procedure similar to that in Fig. 4 was followed, with four ripple wavelengths between anode and reference plane. The cyclotron frequency here is 329.5 mc.

(5) The patterns of microwave noise power resemble blurred envelopes of the UHF traces.

Some idea of the transverse distributions of UHF noise power and electron-current density, in a region of strong proper-frequency excitation, is given in Figs. 6 and 7. The measurements were taken by moving a small aperture in a broad arc through the probe centerline, just in front of the probe aperture. In both illustrations, the relative noise power has been "normalized" to compensate for variations in electron current traversing the RF gap.

The curves of Fig. 6 are typical of most such measurements. The beam-current density varies smoothly through a single broad maximum, and

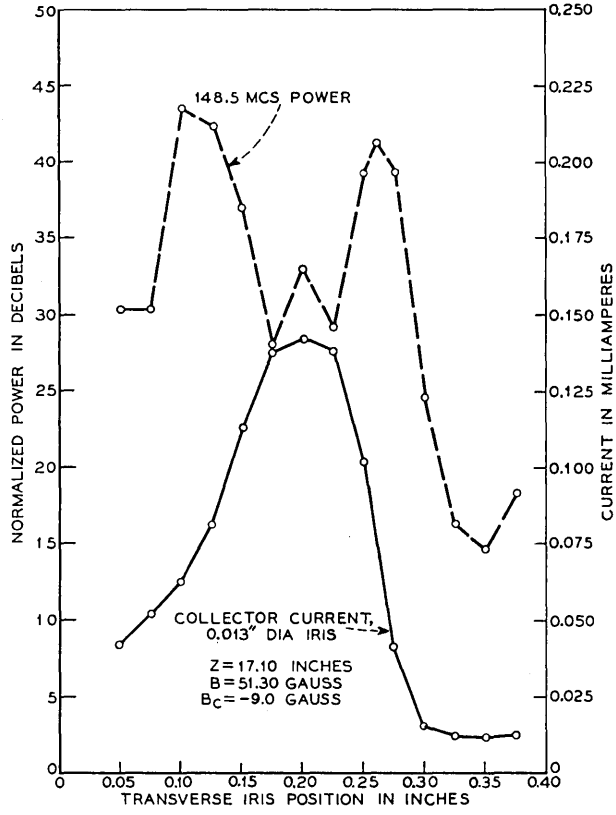

FIG. 6 — Simultaneous point-by-point measurements of collector current and relative noise power, obtained by moving an 0.013-inch diameter aperture in a broad arc through the probe centerline. The probe is stationary, about 18 inches from the gun anode, and the fields have been adjusted for maximum amplitude of the proper frequency, 148.5 mc. The cyclotron frequency is 143.8 mc.

the noise-power density is greatest at the rim of the beam so defined, and least near its center. No evidence of azimuthal periodicity was found. The curves of Fig. 7, which are less typical, indicate five distinct peaks of RF power, despite a nearly symmetrical pattern of collector current. At the time of this measurement, cathode emission may have been uneven, due to coating damage by ion bombardment.

In the rippled beam on which these measurements were made, the ratio of flux encircled at the cathode, to that in the drift space, was very small for most electrons. One would, therefore, expect the transverse noise-power distribution in this beam to resemble that in a smooth Brillouin beam.[9] The noise power expected when a pinhole aperture is located at the beam center can be compared with that when the aper-

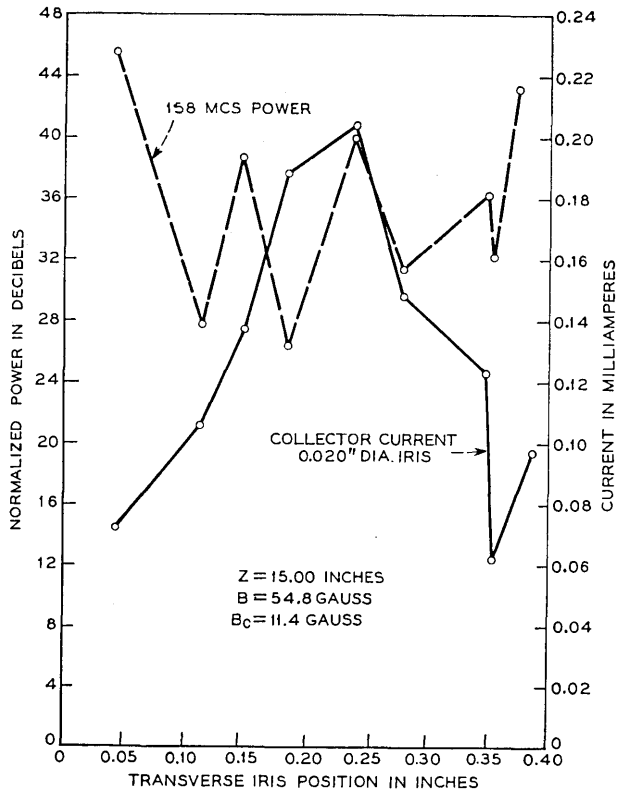

FIG. 7 — Transverse distribution measurements similar to those of Fig. 6. This pattern was obtained a week later than that of Fig. 6, and the cathode was operated at a higher temperature. The cyclotron frequency would be 153.2 mc for the field used.

ture straddles the beam rim, by taking the beam area exposed in the first case to be that of a sector of angle θ, and the length of beam surface in the second case to be that of the corresponding arc:

$$\frac{\text{Rf current sample inside of beam}}{\text{Rf current sample at rim of beam}} \cong \frac{\theta b^2 J_z}{2\theta b G_z} = \frac{b(\omega - \beta u)\cdot I_0(\beta b)}{2uI_1(\beta b)}.$$

Here b is the beam radius, and J_z, G_z the longitudinal components of volume and surface current densities, respectively. I_0 and I_1 are modified Bessel functions, β is the propagation constant, u the beam velocity, and ω the radian frequency. For the frequencies and beam radii employed in these measurements, this ratio is very much less than unity. Thus the pattern of Fig. 6 is in accord with this mode distribution. The multiple peaks of Fig. 7, however, do not conform to this picture, and are not understood at present.

As most of the RF power is concentrated near the rim of the beam, the question arises whether the double and triple peaks, in the longitudinal distribution patterns of Figs. 2 to 5, are not due to the probe aperture breaking through the beam rim. However, the dip between adjacent noise peaks is too great to be explained on the basis of reduced partition noise or weakened gap coupling, assuming the beam diameter there to be less than the gap diameter (0.100 inch). Moreover, double peaks occur even when the beam diameter exceeds the RF gap diameter; for instance, near the last three beam waists of Fig. 5. (When all of the beam is transmitted by the 0.100 inch aperture, the collector-current peak is flat-topped.) It seems likely, therefore, that the double and triple peaks correspond to peaks of amplitude over the entire beam cross-section.

V PROPAGATION ALONG THE RIPPLED BEAM

To find an explanation for the multiple peaks of space-charge current, a small-signal, slow-wave analysis of wave propagation along the rippled beam can be made, in which the special features of these experiments are exploited: long ripple wavelength, effectively no flux at the cathode, and low frequencies. The first of these features suggests that the propagation constants can be evaluated at each cross-section plane as though the beam were uniform, despite the presence of radial velocities. In addition, the space-charge density is assumed constant at each cross-section, and the electron flow laminar.

With these assumptions, the beam can be regarded as a fluid of moving charge, with a single-valued velocity at each point in space, as follows:

$$v_0 = (v_r , v_\theta , v_z) \tag{1}$$

where

$$v_r = r \cdot f(z), \qquad \text{or} \qquad \frac{\partial v_r}{\partial r} = \frac{v_r}{r}, \tag{2}$$

$$v_\theta = r\dot{\theta} = r\frac{\omega_c}{2}, \tag{3}$$

$$v_z = u. \tag{4}$$

Here, r, θ, z are the polar cylindrical coordinates, $\omega_c = \eta B$ the angular cyclotron frequency corresponding to the longitudinal focusing field B, and $f(z)$ a function describing the amplitude and spatial periodicity of the beam ripple. The experimental data indicates that the potential variations along the beam axis are negligible, permitting the assumption that the longitudinal velocity, u, is constant. MKS units are used.

Consistent with the distribution pattern of Fig. 6, the ac field can be represented by an axially-symmetric potential function, similar to that for the smooth Brillouin-flow beam:

$$V \sim I_0(\gamma r) \exp j(\omega t - \beta z), \tag{5}$$

$$\underline{E} = -\text{grad } V. \tag{6}$$

The ac equations of fluid motion are obtained by adding a small ac increment to each of the steady-state velocity components. In addition to the space-charge field, the ac electric field contributes forces acting on the charged medium; those contributed by the ac magnetic field are neglected:

$$\frac{d}{dt}(\underline{v}_0 + \tilde{\underline{v}}) = -\eta[-\text{grad } V - \text{grad } V_0 + (\underline{v}_0 + \tilde{\underline{v}}) \times \underline{B}]. \tag{7}$$

The ac velocity is distinguished by a tilde, and the dc velocity by a zero subscript. Here $\eta = e/m$ is the charge-mass ratio of the electron, a positive quantity. As all ac quantities are functions of spatial positions, their time differentiation (indicated by a dot) is equivalent to multiplication by $j(\omega - \beta u)$, written $j\omega_b$ for brevity.

The components of the force equation are expanded as follows:

$$\frac{\partial \tilde{v}_r}{\partial t} + (v_r + \tilde{v}_r)\frac{\partial}{\partial r}(v_r + \tilde{v}_r) + (u + \tilde{v}_z)\frac{\partial}{\partial z}(v_r + \tilde{v}_r) - \frac{(v_\theta + \tilde{v}_\theta)^2}{r} = \eta\frac{\partial V}{\partial r} + \eta\frac{\partial V_0}{\partial r} - \omega_c(v_\theta + \tilde{v}_\theta), \tag{8a}$$

$$\left[j\omega_b + v_r\left(\frac{\partial}{\partial r} + \frac{1}{r}\right)\right]\tilde{v}_r + \left(\frac{\partial v_r}{\partial z}\right)\tilde{v}_z = \eta\,\frac{\partial V}{\partial r}, \tag{8b}$$

$$\frac{\partial \tilde{v}_\theta}{\partial t} + (v_r + \tilde{v}_r)\left(\dot{\theta} + \frac{\partial \tilde{v}_\theta}{\partial r}\right) + u\,\frac{\partial \tilde{v}_\theta}{\partial z} + (v_r + \tilde{v}_r)\left(\dot{\theta} + \frac{\tilde{v}_\theta}{r}\right) = \eta\,\frac{\partial V}{\partial \theta} + \omega_e(v_r + \tilde{v}_r), \tag{9a}$$

$$\left[j\omega_b + v_r\left(\frac{\partial}{\partial r} + \frac{1}{r}\right)\right]\tilde{v}_\theta = 0, \tag{9b}$$

$$\frac{\partial \tilde{v}_z}{\partial t} + v_r\,\frac{\partial \tilde{v}_z}{\partial r} + u\,\frac{\partial \tilde{v}_z}{\partial z} = \eta\,\frac{\partial V}{\partial z}, \tag{10a}$$

$$\left[j\omega_b + v_r\,\frac{\partial}{\partial r}\right]\tilde{v}_z = -j\eta\beta V. \tag{10b}$$

An expression for the ac space-charge density, $\tilde{\rho}$, can be obtained in terms of its steady-state counterpart, ρ_0, by means of the charge-conservation equation:

$$\frac{\partial \tilde{\rho}}{\partial t} = -\underline{v}_0\cdot\text{grad}\,\tilde{\rho} - \underline{\tilde{v}}\cdot\text{grad}\,\rho_0 - \rho_0\,\text{div}\,\underline{\tilde{v}} - \tilde{\rho}\,\text{div}\,\underline{v}_0 - \underline{v}_0\cdot\text{grad}\,\rho_0 - \rho_0\,\text{div}\,\underline{v}_0\,. \tag{11}$$

As the beam diameter changes slowly, the dc space-charge density at each plane is taken to be inversely proportional to the square of the radius, b:

$$|\,\text{grad}\,\rho_0\,| = \frac{\partial \rho_0}{\partial z} = -\frac{2\rho_0}{b}\,\frac{\partial b}{\partial z} \cong -\frac{2v_r}{ur}\,\rho_0, \tag{12}$$

$$\text{div}\,\underline{v}_0 = -\frac{\underline{v}_0\cdot\text{grad}\,\rho_0}{\rho_0} \cong -\frac{2v_r}{r}, \tag{13}$$

$$\left[j\omega_b + v_r\left(\frac{\partial}{\partial r} + \frac{1}{r}\right)\right]\tilde{\rho} = -\rho_0\left[\frac{1}{r}\,\frac{\partial}{\partial r}(r\tilde{v}_r) - j\beta\tilde{v}_z\left(1 - j\,\frac{2}{\beta u}\,\frac{v_r}{r}\right)\right]. \tag{14}$$

At low frequencies (the UHF region),

$$\frac{\partial V}{\partial r} \ll \frac{V}{r}$$

as $(\gamma r)^2 \ll 1$. This inequality is also true of other ac quantities proportional to V, such as $\tilde{v}_z$ and $\tilde{\rho}$, and with a small error can be assumed to be true for $\tilde{v}_r$. When the operator $\partial/\partial r$ is omitted from (8), (9), (10), and (14), it is possible to solve explicitly for $\tilde{\rho}$ in terms of ρ_0 and V.

The laminar-flow rippled beam can be described by the particle trajectories, as follows:

$$r_i = r_{0i}(1 + \delta \cos \beta_e z), \tag{15}$$

where r_{0i} is the maximum radius for the particle considered, and $0 < \delta < 1$. For this model of the beam,

$$\frac{v_r}{r} = \frac{\dot{b}}{b} = \frac{-\delta\omega_c \sin \beta_e z}{1 + \delta \cos \beta_e z}, \tag{16}$$

$$\frac{1}{r}\frac{\partial v_r}{\partial z} = \frac{-\omega_c \beta_e \delta(\delta + \cos \beta_e z)}{(1 + \delta \cos \beta_e z)^2}. \tag{17}$$

The region of interest, judging from the observed peak locations, is not at the mid-plane of the beam waist, where $v_r = 0$, but on either side of that plane, where $| v_r/r |$ is greatest. It is readily found that, at these positions $(1/r)\ (\partial v_r/\partial z)$ is zero, and (14) can be written

$$\tilde{\rho} = \frac{\eta\rho_0 V}{\omega_b^2}\left[\frac{\gamma^2}{\left(1 - j\frac{v_r}{\omega_b r}\right)^2} - \beta^2 \frac{\left(1 - j\frac{2}{\beta u}\frac{v_r}{r}\right)}{\left(1 - j\frac{v_r}{\omega_b r}\right)}\right]. \tag{18}$$

This can be combined with Poisson's Equation,

$$\Delta V = (\gamma^2 - \beta^2)V = -\tilde{\rho}/\varepsilon, \tag{19}$$

to furnish a relation between γ and β:

$$\gamma^2\left[1 - \frac{R}{\left(1 - j\frac{v_r}{\omega_b r}\right)^2}\right] = \beta^2\left[1 - \frac{R\left(1 - j\frac{2}{\beta u}\frac{v_r}{r}\right)}{\left(1 - j\frac{v_r}{\omega_b r}\right)}\right] \tag{20}$$

where $R = \omega_p^2/\omega_b^2$ and $\omega_p^2 = -\eta\rho_0/\epsilon$, the square of the angular plasma frequency.

At the beam boundary, $r = b$, the continuity of the tangential field components and the change in radial electric displacement can be expressed in the form of an admittance equation:

$$\left[\frac{1}{V}\left(\frac{\partial V}{\partial r} - \frac{\tilde{\sigma}}{\epsilon}\right)\right]_b^{\mathrm{I}} = \left[\frac{1}{V}\frac{\partial V}{\partial r}\right]_b^{\mathrm{II}}. \tag{21}$$

Here I refers to the beam, $0 \leqq r \leqq b$, and II to the space between beam and the concentric conducting tube, $b \leqq r \leqq a$. The surface charge layer, $\tilde{\sigma}$, takes account of the surface ripple, of amplitude $\tilde{r}$:

$$\tilde{\sigma} = \rho_0 \tilde{r} = -\frac{j\rho_0 \tilde{v}_r}{\omega_b},$$

$$-\tilde{\sigma}/\epsilon = -\frac{R(\partial V/\partial r)}{1 - j\dfrac{v_r}{\omega_b r}}. \tag{22}$$

The appropriate potential functions in I and II are reduced by means of the low-frequency, or thin-beam, approximation, as follows:

$$\left[\frac{1}{V}\frac{\partial V}{\partial r}\right]_b^{\mathrm{I}} = \frac{\gamma I_1(\gamma b)}{I_0(\gamma b)} \cong \frac{\gamma^2 b}{2},$$

$$\left[\frac{1}{V}\frac{\partial V}{\partial r}\right]_b^{\mathrm{II}} = \beta\left[\frac{I_1(\beta b)K_0(\beta a) + I_0(\beta a)K_1(\beta b)}{I_0(\beta b)K_0(\beta a) - I_0(\beta a)K_0(\beta b)}\right] \tag{23}$$

$$\cong \frac{\beta^2 b}{2} - \frac{1}{b \ln a/b}$$

where the following small-argument approximations have been used:

$$K_0(x) \cong -\ln x,$$

$$K_1(x) \cong \frac{x}{2}\ln x + \frac{1}{x}. \tag{24}$$

The boundary equation thus provides a second relation between γ and β:

$$\frac{\gamma^2 b}{2}\left[1 - \frac{R}{\left(1 - j\dfrac{v_r}{\omega_b r}\right)}\right] = \frac{\beta^2 b}{2} - \frac{1}{b \ln \dfrac{a}{b}}. \tag{25}$$

For the smooth beam in Brillouin flow ($v_r = 0$), the boundary equation, to the same low-frequency approximation, is as follows:

$$R_0 \equiv \frac{\omega_p^2}{(\omega - \beta_0 u)^2} = \frac{2}{(\beta_0 b)^2 \ln \dfrac{a}{b}}. \tag{26}$$

To see how the beam ripple affects the propagation constant, it is sufficient to find its first-order effect; i.e., to assume relatively small radial velocities and find a solution for β which is not very different from its value, β_0, for the smooth beam:

$$\beta = \beta_0 + \delta = \beta_e \pm \beta_q + \delta, \tag{27}$$

where

$$|\delta| \ll \beta_0\ ; \qquad \beta_q \cong \frac{\omega_p}{u\sqrt{R_0}}; \qquad \beta_e = \frac{\omega}{u}.$$

In addition, $|\, v_r/\omega_b r \,|$ is less than unity, and $|\, 2v_r/\beta ur \,|$ can be neglected entirely. With these assumptions, the boundary and characteristic equations can be combined to solve for β:

$$\frac{\gamma^2}{\beta^2} = \frac{F(F-R)}{F^2-R} = \frac{F - \left(\frac{\beta_0}{\beta}\right)^2 R_0}{F-R}, \tag{28}$$

where

$$F = 1 - j\frac{v_r}{\omega_b r}.$$

Utilizing the low-frequency condition, $|\, R^2 \,| > |\, R \,| > 1$, this equation can be reduced and, after some algebra, solved:

$$\frac{\beta_0}{\beta}\left(\frac{R_0}{R}\right)^{1/2} = \frac{\beta_0(\delta \pm \beta_q)}{(\beta_0 + \delta)(\pm\beta_q)} \cong 1 - j\frac{v_r}{2\omega_b r},$$

$$\beta_s \cong \beta_e + \beta_q - j\frac{v_r}{2ur}, \tag{29a}$$

$$\beta_f \cong \beta_e - \beta_q + j\frac{v_r}{2ur}. \tag{29b}$$

These expressions show that the current in the slow wave (I_s) will grow when (v_r/r) is negative; i.e., when the beam is contracting, and decrease during its expansion. The fast wave (I_f) will do the opposite. In probe measurements along the beam, the detected ac power is proportional to the square of the total space-charge current, which has the following dependence on time and distance when the amplitudes of both waves are initially equal:

$$(I_s + I_f) = 2I_{\max} \cos(\omega t - \beta_e z) \cdot \cos(\beta_q z) \cdot \sinh\left(\frac{V_r z}{2ur}\right). \tag{30}$$

In UHF noise-power measurements along beams with long ripple wavelengths, the two planes of maximum $\pm (v_r/r)$ are separated by only a small fraction of a space-charge wavelength. Therefore, $\cos \beta_q z$ at the first of these planes is only slightly larger than at the second. Thus, two peaks of current are observed, in agreement with (30). By contrast, in rippled-beam amplification at microwave frequencies, shorter ripple wavelengths and smaller ripple amplitudes are employed. Then (v_r/r) varies nearly sinusoidally over the ripple wavelength. For maximum net gain per ripple, maximum negative (v_r/r) is adjusted to coincide with the plane of $\cos \beta_q z = 1$ (maximum current), and maximum positive (v_r/r) at the current minimum, half a wavelength beyond.

The gain constants in (29) are independent of frequency. The net gain per ripple wavelength, however, will vary with frequency, depending on how closely both the current maxima and minima coincide with the regions of maximum $\pm (v_r/r)$, respectively. This is a statement of the "resonance" condition between ripple wavelength and half the space-charge wavelength, which emerges from one-dimensional analyses[10] of this gain mechanism based on transmission-line analogies.

Such analyses generally assume small-amplitude sinusoidal variations of the reduced plasma wave number, β_q, along a one-dimensional beam in a longitudinal ac field with no losses. Periodic variations in either beam or wall diameters, or beam velocity, cause the beam "impedance" to vary periodically, imparting to it narrow-band filter-like properties equivalent to narrow-band signal gain. From another point of view,[11] these periodic impedance changes couple the fast and slow space-charge waves to each other intermittently, thereby effecting an energy transfer from the fast to the slow wave. As this coupling is lossless, I_s increases and I_f decreases with drift distance, in such a way as to keep their product constant. Then the product $I_{\max}I_{\min}$ increases, and the ratio $I_{\max}/_{\min}$ correspondingly decreases. In the case of *noise-power* amplification, two uncorrelated space-charge standing waves are present. Because the two slow waves cannot simultaneously be amplified at the expense of the two fast waves, the product $I_{\max}I_{\min}$ must remain constant.

The observed noise-current patterns in rippled-beam amplification,[1] however, are characterized by a *nearly constant ratio* $I_{\max}/I_{\min}$, and an *increase in the product* $I_{\max}I_{\min}$ along the beam, despite the fact that the beam voltage is fixed. This apparent contradiction can be resolved by a closer look at the energy-exchange processes.

Chu[12] has shown that the kinetic power flow in space-charge waves (the major part of the total power) is equal to the difference in powers carried by the fast and slow waves. This is equally true of beams with transverse motions and fields.[13] In rippled-beam amplification, whether analyzed as a modulated linear beam or at each beam cross-section separately, as here, the propagation constants are found to be complex conjugate quantities, whose real parts describe the ordinary fast and slow waves of a uniform beam. From either point of view, therefore, a decrease in I_f and an increase in I_s signifies an increase in the negative kinetic power flow carried by the waves, or a decrease in the total kinetic energy of the beam.

As shown in (29), the gain constants are proportional to v_r, indicating that the dc energy transferred to the waves when the beam contracts could only have come from the *radial* kinetic energy, not the longitudinal.

The direction of energy transfer is reversed during the subsequent beam expansion. If the ripple were perfectly symmetrical, therefore, and the dc-ac energy exchange perfectly reversible, the net effect of a beam ripple would be zero. Neither of these conditions is quite true in actual beams. Rippled flow is never truly laminar, and $|v_r/r|$ usually decreases with drift distance as the flow loses coherence; i.e., it is greater in beam contraction than in the next expansion. This by itself would produce a net gain per ripple in I_s, and a net loss in I_f, of equal amounts. In addition, however, unavoidable small non-linearities in electron motions prevent all of the ac energy in a de-amplified wave from being converted back to dc kinetic energy. Thus it is possible for *both* the fast and slow waves to increase in a ripple wavelength, the latter always more than the former.

The greater gain of the slow wave entails a loss of radial kinetic energy, in agreement with the observation that the ripple amplitude always decays more rapidly when rippled-beam amplification takes place. The incomplete reversibility of the ac-dc energy exchange probably accounts for the observed increase in $I_{max}I_{min}$ for noise currents. Finally, the net amplification of all of the space-charge waves, fast as well as slow, is in accord with the observed near-constancy of the ratio I_{max}/I_{min} for microwave-frequency noise, despite increases in the product $I_{max}I_{min}$ of 30 db and more.

VI ORIGIN OF THE PROPER-FREQUENCY PEAKS

Of the various peaks in the beam's noise spectrum, described in Section III and Fig. 1, those with "proper frequencies," slightly above the cyclotron value, are so large in amplitude that even an approximate analysis should be able to account for them. To do so, a "working model" of the beam is needed, which conforms to the experimental conditions which existed during the observations:

(1) The peak intensities were greatest near the middle of each beam waist, and decreased with decrease in ripple amplitude.

(2) The focusing field was below the nominal Brillouin value. The field at the cathode, B_c, was finite and opposed to the main field, B.

(3) Collector-current measurements along the beam axis showed the ratio of maximum to minimum current to be greater, the smaller the aperture.

(4) The gas pressure was about 10^{-7} mm Hg. The beam was pulsed with a 1,000-cycle square wave.

Item (3) indicates that the flow was non-laminar; and Item (4) indicates the presence of positive ions. All the items are consistent with the following picture:

In a beam with large ripples, nearly all electrons have their maximum radii and zero radial velocity at the same z-plane. Those with sufficiently large maximum radius will have enough transverse kinetic energy to surmount the space-charge forces at the beam waist, and pass through or close to the axis. Others, with smaller maximum radii, will spiral about that axis. Dolder and Klemperer[14] have observed a similar division of electrons into "crossovers" and non-crossovers, in electron-optical systems without magnetic fields.

Positive ions tend to neutralize the electronic space charge at the beam waists, broadening the region in which crossover occurs. The crossover trajectories thereupon overlap one another, resulting in multivalued transverse particle velocities in this region. In a first-order (linearized) study of wave propagation along the beam, one must replace the actual multivelocity charge motions with a single "fluid" of charge, whose velocity at any point is the average of the particle velocities there. It is clear that the z-velocity of the stream is u, and the radial velocity zero. The tangential velocity, $v_\theta = (\dot{\theta}r)_{\text{av}}$, however, is more complicated.

Owing to the partial or total neutralization of electronic space charge at the beam waists, and their large radii elsewhere, the crossover electrons will encounter virtually no space-charge forces in their paths. Their transverse paths will consequently be circles about fixed centers, described with angular velocity equal to the cyclotron frequency. Their angular velocity about the beam axis is given by Busch's Theorem:

$$\dot{\theta} = \frac{\omega_c}{2}\left[1 + \frac{K}{r^2}\right],$$

where

$$K = -r_c^2\left(\frac{B_c}{B}\right) = r_{\max}r_{\min} \tag{31}$$

is a positive quantity, as B_c/B is negative. Here, r_c is the radius at which a particular electron left the cathode, and r is its radius in the drift region. The angular velocity, $\dot{\theta}$, is greater than $\omega_c/2$ at all times, and exceeds ω_c in the waist region of the beam. The average value of v_θ at any point here, therefore, is greater than $\omega_c r$ and presumably varies from point to point in some unknown way.

If v_θ is left unspecified, and the assumptions adopted of zero space-charge forces and radial velocity over a finite length of beam:

$$\eta\frac{\partial V_0}{\partial r} = 0, \qquad v_r = 0, \qquad \frac{dv_r}{dt} = 0, \tag{32}$$

the radial component of the force equation (7) in Euler coordinates can

be written as follows:

$$\left(\frac{dv_0}{dt}\right)_r = -\frac{v_\theta^2}{r} = -\omega_c v_\theta , \tag{33}$$

$$v_\theta = 0 \quad \text{and} \quad \omega_c r. \tag{34}$$

Thus, the radial "balance" conditions (32) are consistent with either of two values for v_θ, of the equivalent stream with single-valued velocities. As it develops that either of these values leads to the same result, the first one will be used here for simplicity, $v_\theta = 0$.

An ac traveling wave along this beam cannot have any θ-dependency, because the beam has no single value of angular velocity $\dot{\theta}$, which might remain in synchronism with that of the wave. Thus, the perturbed dynamics equation (7) can be expanded, with the assumptions of an axial-symmetric ac field given by (5) and (6), a stream with steady-state velocity (0, 0, u), constant space-charge density ρ_0, and no space-charge forces, as follows:

$$\frac{\partial \tilde{v}_r}{\partial t} + u\frac{\partial \tilde{v}_r}{\partial z} = \eta\frac{\partial V}{\partial r} - \omega_c \tilde{v}_\theta ,$$

$$\frac{\partial \tilde{v}_\theta}{\partial t} + u\frac{\partial \tilde{v}_\theta}{\partial z} = \omega_c \tilde{v}_r ,$$

$$\frac{\partial \tilde{v}_z}{\partial t} + u\frac{\partial \tilde{v}_z}{\partial z} = \eta\frac{\partial V}{\partial z} .$$

These are solved for the ac velocity components:

$$\tilde{v}_r = \frac{-j\eta\omega_b}{\omega_b^2 - \omega_c^2}\frac{\partial V}{\partial r} , \tag{35}$$

$$\tilde{v}_\theta = -\frac{j\omega_c}{\omega_b}\tilde{v}_r , \tag{36}$$

$$v_z = -\frac{\eta\beta}{\omega_b} V. \tag{37}$$

With grad ρ_0 = div $\underline{v_0}$ = 0, the charge-conservation equation (11) can be solved for $\tilde{\rho}$:

$$\frac{\partial \tilde{\rho}}{\partial t} = -\underline{v_0}\cdot\text{grad}\ \tilde{\rho} - \rho_0 \text{ div } \underline{\tilde{v}},$$

$$\tilde{\rho} = \frac{j\rho_0}{\omega_b}\text{ div } \underline{\tilde{v}} = \eta\rho_0\left[\frac{\gamma^2}{\omega_b^2 - \omega_c^2} - \frac{\beta^2}{\omega_b^2}\right]. \tag{38}$$

At very large ripple amplitudes, it is a fair assumption that the density of non-crossover electrons is negligible relative to that of crossovers in this region. Poisson's equation (19) can then be combined with the

above expression to obtain the characteristic equation:

$$\left(\frac{\beta}{\gamma}\right)^2 = \frac{1 - \dfrac{\omega_p^2}{\omega_b^2 - \omega_c^2}}{1 - \dfrac{\omega_p^2}{\omega_b^2}}. \tag{39}$$

In a frame of reference moving with the stream, u' is 0, $\beta' u'$ is 0, and $\omega_b' = \omega$. Then,

$$\left(\frac{\beta'}{\gamma'}\right)^2 = \frac{(\omega^2 - \omega_c^2 - \omega_p^2)\omega^2}{(\omega^2 - \omega_p^2)(\omega^2 - \omega_c^2)} \tag{40}$$

and β' becomes an infinite imaginary quantity when ω is ω_c. The phase velocity in the moving frame is infinite, as the real part of β' is zero; therefore the phase velocity v_p in the rest frame is also infinite. Thus, there is no Doppler shift in the "resonant" frequency observed in the rest frame:

$$\omega_{\text{observed}} = \frac{\omega_c}{1 - \dfrac{u}{v_p}} = \omega_c. \tag{41}$$

As the actual beam has a z-velocity spread, the field is never perfectly uniform, and as the calculation is valid for small ac quantities only, the discrepancy between this result and the observed "proper" frequencies, which were 1.03 to 1.06 times the cyclotron value, is not unexpected.

The singularity in (40) is seen to disappear when $\omega_p^2 = 0$. This indicates that an exact calculation would show the gain constant $(-j\beta)$ to increase with ρ_0, the density of the crossover electrons. Their trajectories, described by (31), and the absence of space-charge forces are such that $K = r_{\max} r_{\min}$; that is, the greater $r_{\max}$, the smaller $r_{\min}$, the distance of closest approach to the axis. Thus, a larger ripple amplitude (permitted by a lower magnetic field) produces a greater electron density in the waist region, and accordingly a greater oscillation amplitude at the resonant frequency, as observed.

The foregoing mathematics describes a form of resonance, the infinite phase velocity corresponding to longitudinal "cutoff" in a waveguide. Unlike a waveguide, however, the disturbance increases rather than attenuates along the axis, due to the transfer of dc kinetic energy (represented by v_θ^2/r) to the ac fields (excited by noise fluctuations at the cathode), at the cyclotron frequency ω_c.

Except for the direction of energy transfer, the situation is analogous to that of a low-pressure gas in a uniform magnetic field, when stressed by an impressed ac field of varying frequency. It has been found that the breakdown field at the cyclotron frequency is very much less than

at other frequencies.[15] Here the energy supplied by the ac field is coupled most effectively to the free electrons at the resonant frequency, increasing their dc kinetic energy until the gas breaks down. The circular ac charge motions due to the dc magnetic and the ac electric fields are superimposed on high-velocity random motions, similar to the radial motions in the drifting beam.

The UHF peaks observed at harmonics of the proper frequency may simply be due to the non-linear character of the beam, when excited by the high-level fundamental oscillations. The other faint satellite peaks, near 0.5 ω_c and 0.707 ω_c, seem to be associated with the unneutralized space-charge density at the beam waist.

The conspicuous role played by crossover electrons in the waist region of rippled beams, due to the tendency of their orbits to overlap there, leads one to re-examine their influence on rippled-beam amplification. As seen in the previous section, this gain process depends on the average value of (v_r/r) at each cross-section plane of the beam. The fraction of all electrons which penetrate to the beam axis depends on competition between the unneutralized space-charge forces and the particle's transverse kinetic energy. An increase in positive ion density tends to make the potential depressions at beam waists broader and shallower, and thereby increase the number of crossover electrons as well as the axial distance over which they reach the axis. The net effect is to reduce the average value of $| v_r |$ over a greater portion of the ripple wavelength, and thus reduce the net gain of the space-charge wave. This may explain why the "growing noise" phenomenon tends to be inhibited by an increase in positive ion density.

VII CONCLUSIONS

Evidence is found of oscillations with frequencies in the 10- to 500-mc region inside of an electron-gun diode. There is some basis for associating them with electron-field interaction in the retarding region of the diode. Another type of narrow-band noise peak is found near the waists of a strongly rippled beam in a longitudinal magnetic field, with frequencies proportional to the field strength. The strongest of these, at about 1.05 times the angular cyclotron frequency, ω_c, as well as its harmonics, can be explained by the resonant behavior of a short section of the beam, in which the average transverse velocity is nullified by overlap in particle orbits. Fainter satellite peaks, near 0.5 ω_c, 0.707 ω_c, and ω_c, respectively, accompany the dominant frequency.

In a drifting beam launched from a shielded electron gun and focused by an axial field, the transverse distribution of noise (or signal) intensity is found to agree with that predicted for ideal Brillouin flow. Despite

the presence of thermal motions and beam ripples, the ac power is found to be concentrated chiefly at the rim of the beam. Occasionally, several concentric rings of noise maxima are found within the beam, possibly due to unusual cathode conditions.

When the ripple wavelength is very long, two maxima of noise power are observed to flank each beam waist. A first-order calculation of wave propagation along a rippled laminar-flow beam accounts for this pattern by showing that space-charge waves grow at the expense of dc kinetic energy in the radial charge motion. In rippled-beam amplification of noise, the product $I_{max}I_{min}$ has been found to increase, and the the ratio I_{max}/I_{min} remain nearly constant, because both fast and slow waves are amplified, the former less than the latter, and because the wave coupling is not lossless.

Positive ions tend to collect at the waist of rippled beams, thereby extending the region in which electrons pass close to the axis, instead of circling about it. The overlap of their orbits leads to net cancellation of radial charge motion, and hence a reduction in rippled-beam amplification. This may explain why positive ions tend to inhibit the "growing noise" phenomenon.

REFERENCES

1. W. W. Rigrod, Noise Spectrum of Electron Beam in Longitudinal Magnetic Field. Part I — The Growing Noise Phenomenon, p. 831 of this issue.
2. C. C. Cutler, Spurious Modulation of Electron Beams, Proc. I.R.E., **44**, p. 61, Jan., 1956.
3. K. G. Hernquist, Plasma Ion Oscillations in Electron Beams, J. Appl. Phys. **26**, p. 544, May, 1955.
4. F. B. Llewellyn and A. E. Bowen, The Production of UHF Oscillations by Diodes, B.S.T.J., **18**, p. 280, April, 1939.
5. O. Klemperer, Influence of Space Charge on Thermionic Emission Velocities, Proc. Royal Soc. (London) (A) **190**, p. 376, 1947.
6. K. T. Dolder and O. Klemperer, High Frequency Oscillations in the Space Charge of some Electron Emission Systems, Journal of Electronics, **1**, p. 601 May, 1956.
7. C. N. Smyth, Total Emission Damping with Space-Charge-Limited Cathodes, Nature, **157**, p. 841, June 22, 1946.
8. W. Veith, Electron Energy Distribution in Space-Charge-Limited Electron Streams, Zeit. f. angew. Physik, **7**, No. 9, p. 437, 1955.
9. W. W. Rigrod and J. A. Lewis, Wave Propagation Along a Magnetically-Focused Cylindrical Electron Beam, B.S.T.J., **33**, p. 399, March, 1954.
10. R. W. Peter, S. Bloom, and J. A. Ruetz, Space-Charge-Wave Amplification Along an Electron Beam by Periodic Change of the Beam Impedance, RCA Rev., **15**, p. 113, March, 1954.
11. J. R. Pierce, The Wave Picture of Microwave Tubes, B.S.T.J., **33**, p. 1343, Nov., 1954.
12. L. J. Chu, 1951 I.R.E. Electron Tube Conference on Electron Devices.
13. H. A. Haus and D. L. Bobroff, Small Signal Power Theorem for Electron Beams (to be published).
14. K. T. Dolder and O. Klemperer, Space-Charge Effects in Electron Optical Systems, J. App. Phys., **26**, p. 1461, Dec., 1955.
15. S. J. Buchsbaum and E. Gordon, Highly Ionized Microwave Plasma, M.I.T. R.L.E. Quarterly Prog. Rep., p. 11, Oct. 15, 1956.

Distortion Produced in a Noise Modulated FM Signal by Nonlinear Attenuation and Phase Shift

By S. O. Rice

(Manuscript received December 6, 1956)

An expression is given for the FM distortion introduced by a transducer whose attenuation and phase shift depend upon the frequency in an arbitrary way. This expression appears to be difficult to evaluate, but it yields useful approximations for the second and third order modulation terms. In all of the work, it is assumed that the distortion is small compared to the signal, and that the signal can be represented by a random noise having the same power spectrum.

INTRODUCTION

A number of workers have been concerned with the problem of computing the distortion introduced by a transducer when an FM wave passes through it. Some of the earliest results were published by Carson and Fry[1] and by van der Pol.[2] Several contributions to the subject have been made recently in connection with studies of microwave radio systems.

An excellent paper on this subject has been published recently by R. G. Medhurst and G. F. Small.[3] Although their results differ considerably in form from those given here, they are nevertheless closely related to ours — their "sinusoidal variations of transmission characteristics" being special cases of our "nonlinear attenuation and phase shift."

Here we treat the problem by applying a method used in a recent paper[4] to study the distortion produced by an echo. Two assumptions are made, (1) that the distortion is small compared to the signal, and (2) that the signal can be represented by a random noise which has the same power spectrum as the signal. In Section I, we review some known results and put them in a form suited to our needs. Sections II and III are devoted to the derivation of our main formulas. The principal result is given by the triple integral (3.2) for the power spectrum of the dis-

tortion. Unfortunately, the integrals are difficult to evaluate. However, it is possible to obtain approximations for the second and third order modulation terms. These are given in Section IV. Some miscellaneous comments are made in Section V.

I APPROXIMATE EXPRESSION FOR THE DISTORTION $\theta(t)$

Let the FM signal be $\varphi'(t) = d\varphi/dt$ (for phase modulation the signal would be $\varphi(t)$). Then the FM wave is the real part of

$$v_i(t) = e^{ipt+i\varphi(t)} \tag{1.1}$$

where $p = 2\pi f_p$ is the carrier frequency. Let this wave pass through a transducer having attenuation α and phase shift β, where α and β are even and odd functions, respectively, of the frequency f. When a unit impulse of voltage $\delta(t)$ is applied to the transducer input, the output is

$$g(t) = \int_{-\infty}^{\infty} e^{-\alpha-i\beta+2\pi ift}\, df. \tag{1.2}$$

For physical systems, $g(t)$ is zero for negative t.

When $v_i(t)$ is applied to the transducer input, the output is

$$v_0(t) = \int_{-\infty}^{\infty} v_i(t')g(t - t')\, dt'. \tag{1.3}$$

When $v_0(t)$ is applied to an FM receiver, the detector output consists of the original signal $\varphi'(t)$ plus the distortion $\theta'(t)$ introduced by the transducer. Comparison with (1.1) shows that $\theta(t)$ may be obtained by solving

$$V(t)e^{ipt+i\varphi(t)+i\theta(t)} = v_0(t) \tag{1.4}$$

when p, $\varphi(t)$, $v_0(t)$ are assumed to be known, and $V(t)$, $\theta(t)$ unknown. When $V(t)$ is taken to be positive, (1.4) determines $\theta(t)$ except for an additive term of $2\pi n$ where n is an integer.

We now assume that the transducer acts like a good transmission medium in that the output differs but little from the input. More precisely, we assume

$$| v_0(t) - v_i(t) | \ll 1. \tag{1.5}$$

Since $| v_i(t) | = 1$, it follows that $| v_0(t) | \approx 1$. Transducers having appreciable attenuation and delay may be regarded as two transducers in tandem, one with constant (independent of f) values of α and β/f which are roughly equal to those of the original transducer, and the second

with variable α and β/f. The first transducer produces no distortion of the signal, and if condition (1.5) is satisfied by the second, the considerations of this paper will apply.

Equation (1.4) may be written as

$$V(t)e^{i\theta(t)} = v_0(t)/v_i(t)$$

so that

$$\theta(t) = \text{Im} \log \frac{v_0(t)}{v_i(t)}. \tag{1.6}$$

When we write

$$v_0(t)/v_i(t) = 1 + [v_0(t) - v_i(t)]/v_i(t),$$

expand the logarithm in (1.6), and use (1.5), we obtain our approximate expression for $\theta(t)$:

$$\begin{aligned} \theta(t) &= \text{Im}\,[v_0(t) - v_i(t)]/v_i(t) = \text{Im}\, v_0(t)/v_i(t) \\ &= \text{Im}\,[v_i(t)]^{-1} \int_{-\infty}^{\infty} v_i(t')\, g\,(t - t')\, dt' \\ &= \text{Im} \int_{-\infty}^{\infty} \exp\,[ip(t' - t) + i\varphi(t') - i\varphi(t)]g(t - t')\, dt'. \end{aligned} \tag{1.7}$$

So far there is nothing essentially new in our work.[5]

II AUTOCORRELATION FUNCTION OF $\theta(t)$

In Section I, $\varphi'(t)$ could be any reasonable sort of signal. In the following work we assume that it is a Gaussian noise whose power spectrum, $w_{\varphi'}(f)$, is given to us. The power spectrum of $\varphi(t)$ is

$$w_\varphi(f) = w_{\varphi'}(f)/(2\pi f)^2, \tag{2.1}$$

and its autocorrelation function is

$$\psi_\tau = \int_0^\infty w_\varphi(f) \cos 2\pi f\tau\, df. \tag{2.2}$$

We have written ψ_τ instead of $\psi(\tau)$ or $R_\varphi(\tau)$ to simplify the appearance of the formulas which occur in our work.

Our problem is to find the power spectrum, $w_\theta(f)$, of the distortion $\theta(t)$, given $w_\varphi(f)$. The method of solution is much the same as that used in Reference 4. We first find the autocorrelation function $R_\theta(\tau)$ of $\theta(t)$ and then obtain $w_\theta(f)$ by taking the Fourier cosine transform of $R_\theta(\tau)$.

Let the last integral in (1.7) be $F(t)$ so that $\theta(t) = \text{Im } F(t)$. Then

$$\theta(t)\theta(t+\tau) = \tfrac{1}{2} \text{Re } \{F(t)F^*(t+\tau) - F(t)F(t+\tau)\} \tag{2.3}$$

where $F^*(t+\tau)$ is the complex conjugate of $F(t+\tau)$. The autocorrelation function of $\theta(t)$ is obtained by averaging over the ensemble of the noise functions $\varphi(t)$:

$$\begin{aligned} R_\theta(\tau) &= \text{av } \theta(t)\theta(t+\tau) \\ &= \text{av } \frac{1}{2} \text{Re} \left\{ \int_{-\infty}^{\infty} dt' \int_{-\infty}^{\infty} dt'' \exp [ip(t'-t) + i\varphi(t') - i\varphi(t)] \right. \\ &\quad \cdot g(t-t')g(t+\tau-t'')[\exp [-ip(t''-t-\tau) \\ &\quad - i\varphi(t'') + i\varphi(t+\tau)] - \exp [ip(t''-t-\tau) + i\varphi(t'') \\ &\quad \left. - i\varphi(t+\tau)] \right\}. \end{aligned} \tag{2.4}$$

Since $g(t)$ is real, $g^*(t) = g(t)$. The averaging process may be carried out by a method analogous to that used in Reference 4. The formula to be used is

$$\begin{aligned} &\text{av exp } [i\varphi(t') - i\varphi(t) + ia\varphi(t'') - ia\varphi(t+\tau)] \\ &\quad = \exp [-\psi_0(1+a^2) + \psi_{t'-t} - a\psi_{t'-t''} + a\psi_{t'-t-\tau} \\ &\qquad + a\psi_{t-t''} - a\psi_\tau + a^2\psi_{t''-t-\tau}] \end{aligned} \tag{2.5}$$

where a is either -1 or $+1$, and ψ_τ is an even function of τ. When (2.5) is used in (2.4) a double integral for $R_\theta(\tau)$ is obtained. The substitutions

$$\begin{aligned} x &= t - t', \\ y &= t + \tau - t'', \\ R_v &= \psi_{\tau+x-y} - \psi_{\tau+x} - \psi_{\tau+x} - \psi_{\tau-y} + \psi_\tau \end{aligned} \tag{2.6}$$

convert the double integral into

$$\begin{aligned} R_\theta(\tau) = \frac{1}{2} \text{Re} \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, g(x) e^{-ipx-2\psi_0+\psi_x+\psi_y} \\ \cdot g(y)[e^{ipy+R_v} - e^{-ipy-R_v}]. \end{aligned} \tag{2.7}$$

The symbol R_v is chosen to agree as closely as possible with the notation of Reference 4. There R_v was the autocorrelation of the random function, $v(t)$, where $v(t+T) = \varphi(t) - \varphi(t+T)$, T being the echo delay. Here, R_v is the average value of the product,

$$[\varphi(t) - \varphi(t+y)]\,[\varphi(t+\tau) - \varphi(t+\tau+x)]$$

which becomes the autocorrelation function of $v(t)$ when $y = x = T$.

It may be verified that the expression (2.7) for $R_\theta(\tau)$ is an even function of τ, as it should be. Expression (2.7) is the autocorrelation function we set out to find.

The distortion $\theta(t)$ has an average value, $\bar{\theta}$, whose square is $R_\theta(\infty)$. Since $\varphi(t)$ is a noise function, its autocorrelation function ψ_τ goes to zero as τ approaches ∞. Hence, $R_\theta(\infty)$ is given by the expression obtained from (2.7) by setting $R_v = 0$. The autocorrelation function of $\theta(t) - \bar{\theta}$ is

$$R_{\theta-\bar{\theta}}(\tau) = R_\theta(\tau) - R_\theta(\infty)$$

$$= \frac{1}{2}\,\mathrm{Re}\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\; g(x)e^{-ipx-2\psi_0+\psi_x+\psi_y} \cdot g(y)[e^{ipy}(e^{R_v} - 1) - e^{-ipy}(e^{-R_v} - 1)]. \tag{2.8}$$

III POWER SPECTRUM OF THE DISTORTION

Since $\theta(t)$ has an average value which is generally not zero, its power spectrum, $w_\theta(f)$, has a spike of infinite height at $f = 0$ corresponding to the power in the dc component $\bar{\theta}$. When this spike is subtracted from $w_\theta(f)$ the remainder is the power spectrum of $\theta(t) - \bar{\theta}$ given by

$$w_{\theta-\theta}(f) = 4\int_0^\infty R_{\theta-\bar{\theta}}(\tau)\cos 2\pi f\tau\, d\tau. \tag{3.1}$$

When we use (2.8) and note that $R_{\theta-\bar{\theta}}(\tau)$ is an even function of τ, we obtain

$$w_{\theta-\bar{\theta}}(f) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy g(x)g(y)e^{-2\psi_0+\psi_x+\psi_y} \int_{-\infty}^{\infty} [\cos(px - py) \cdot (e^{R_v} - 1) - \cos(px + py)(e^{-R_v} - 1)] \cos 2\pi f\tau\, d\tau. \tag{3.2}$$

Reasoning similar to that given in Reference 4 shows that the interchannel interference spectrum, $w_c(f)$, (i.e., $w_c(f)\Delta f$ is the average amount of distortion power received in an idle channel of width Δf centered on the frequency f, all other channels being busy) may be obtained from (3.2) by replacing $(e^{\pm R_v} - 1)$ by $(e^{\pm R_v} \mp R_v - 1)$.

The power spectrum of $\theta(t) - \bar{\theta}$ may be regarded as made up of modulation products of all orders. It turns out that the contribution of n^{th} order products is given by the integral of the $R_v{}^n$ terms obtained from the power series expansions of $\exp[\pm R_v]$.

IV FIRST AND SECOND ORDER MODULATION TERMS

Here we shall study the first and second order modulation terms. These arise from the first and second powers of R_v in the expansion of

the quantity within the square brackets in (3.2):

$$2R_v \cos px \cos py + R_v^2 \sin px \sin py. \tag{4.1}$$

The integrations with respect to τ may be performed with the help of

$$\int_{-\infty}^{\infty} \psi_{\tau+b} \cos 2\pi f\tau \, d\tau = \frac{w_\varphi(f)}{2} \operatorname{Re} e^{-i2\pi fb}, \tag{4.2}$$

$$\int_{-\infty}^{\infty} \psi_{\tau+b}\psi_{\tau+c} \cos 2\pi f\tau \, d\tau$$

$$= \operatorname{Re} \frac{1}{4} \int_{-\infty}^{\infty} du \, w_\varphi(u) w_\varphi(f-u) \exp\{-i2\pi[bu + c(f-u)]\} \tag{4.3}$$

which follow from (2.2) and the fact that we have defined $w(-f)$ to be equal to $w(f)$. In our notation the total power in a random noise function is the integral of $w(f)$ from $f = 0$ to $f = \infty$.

The first order modulation term is obtained from (3.2) by replacing the term within the square bracket by $2R_v \cos px \cos py$. When the expression (2.6) for R_v is used, the integration with respect to τ may be performed with the help of (4.2):

$$\int_{-\infty}^{\infty} R_v \cos 2\pi f\tau \, d\tau = \frac{w_\varphi(f)}{2} \operatorname{Re}\left[(e^{-2\pi ixf} - 1)(e^{i2\pi yf} - 1)\right]. \tag{4.4}$$

This leads to the following expression for the first order modulation term in (3.2)

$$w_\varphi(f) \left| \int_{-\infty}^{\infty} dx g(x) e^{-\psi_0+\psi_x} \cos px (e^{-2\pi ixf} - 1) \right|^2. \tag{4.5}$$

This is the quantity which is to be subtracted from $w_{\theta-\bar{\theta}}(f)$ to obtain the interchannel interference spectrum $w_c(f)$.

The second order modulation term is handled in much the same manner. With the help of (4.3) it may be shown that

$$\begin{aligned} \int_{-\infty}^{\infty} R_v^2 \cos 2\pi f\tau \, d\tau = \operatorname{Re} \frac{1}{4} \int_{-\infty}^{\infty} du \, w_\varphi(u) w_\varphi(f-u) \\ \cdot (e^{-2\pi ixu} - 1)\,(e^{-2\pi ix(f-u)} - 1) \\ \cdot (e^{2\pi iyu} - 1)\,(e^{2\pi iy(f-u)} - 1). \end{aligned} \tag{4.6}$$

From this it follows that the second order modulation term in (3.2) is

$$\begin{aligned} \frac{1}{2!\,2} \int_{-\infty}^{\infty} du \, w_\varphi(u) w_\varphi(f-u) \left| \int_{-\infty}^{\infty} dx \, g(x) e^{-\psi_0+\psi_x} \sin px \right. \\ \left. \cdot (e^{-2\pi ixu} - 1)(e^{-2\pi ix(f-u)} - 1) \right|^2. \end{aligned} \tag{4.7}$$

When $\psi_0 - \psi_x$ is so small that $\exp(-\psi_0 + \psi_x)$ may be replaced by unity, as it is in some important practical cases, approximations may be obtained for (4.5) and (4.7). The integral in x may be expressed as the sum of integrals of the type

$$\begin{aligned} \int_{-\infty}^{\infty} g(x)e^{-ipx-2\pi iax}\,dx &= [e^{-\alpha-i\beta}]_{f=a+f_p} \\ &= G_a + iB_a, \\ \int_{-\infty}^{\infty} g(x)e^{ipx-2\pi iax}\,dx &= G_{-a} - iB_{-a}. \end{aligned} \tag{4.8}$$

The values of the integrals follow from (1.2) and the Fourier integral theorem. G and B are, respectively, even and odd functions of frequency, and G_a, B_a are their values at the frequency $f = f_p + a$ where $f_p = p/2\pi$ is the carrier frequency:

$$\begin{aligned} G \text{ at frequency } f_p + a &= G_a, \\ B \text{ at frequency } f_p + a &= B_a. \end{aligned}$$

In this way we get the approximation

$$4^{-1}w_\varphi(f)\,[(G_f - 2G_0 + G_{-f})^2 + (B_f - B_{-f})^2] \tag{4.9}$$

for the first order modulation term, and

$$\begin{aligned} &\frac{1}{2!8}\int_{-\infty}^{\infty} du\, w_\varphi(u)w_\varphi(f-u)[(G_u - G_{-u} + G_{f-u} - G_{-f+u} - G_f \\ &+ G_{-f})^2 + (B_u + B_{-u} + B_{f-u} + B_{-f+u} - B_f - B_{-f} - 2B_0)^2] \end{aligned} \tag{4.10}$$

for the second order modulation term.

Expression (4.10) is an approximation to the second order modulation term (4.7). When most of the interchannel interference is due to second order modulation products, (4.10) is also an approximation to $w_c(f)$, the interchannel interference spectrum. The following remarks may be of some help in deciding whether (4.10) may be used.

1. For the case of phase modulation and a "flat" signal band, the first of equations (5.3) shows that ψ_0 and ψ_τ may be made as small as we please by choosing the signal power (as measured by P_0) small enough. Since R_v is proportional to P_0, P_0 may be chosen small enough to make R_v^3 and higher order terms negligible in the expansion of the integrand of (3.2) (unless there is some sort of symmetry which causes the second order terms to vanish). In this case the interference is mostly second order modulation and (4.7) is a good approximation to $w_c(f)$. Furthermore, as P_0 approaches zero, $\exp(-\psi_0 + \psi_x)$ approaches unity

and (4.10) becomes a good approximation to (4.7). Just how small P_0 has to be depends upon the signal bandwidth, f_b, and the characteristics of the transducer.

2. For the case of FM and a flat signal band, the second of equations (5.3) shows that even if P_0 is small, the difference $\psi_0 - \psi_\tau$ approaches ∞ as $|\tau|$ approaches ∞. To justify the use of (4.10) in this case it is necessary to take into account the behavior of $g(t)$, the response of the transducer to the unit impulse $\delta(t)$. For example, if the duration of $g(x)$ in (4.7) is so brief that $g(x)$ becomes negligibly small before $-\psi_0 + \psi_x$ becomes appreciably different from zero (which may be achieved by making P_0 small enough) then (4.10) is a good approximation to (4.7).

3. When the attenuation, α, and phase shift, β, are given for any particular transducer, the corresponding $g(t)$ may be obtained from (1.2). Once $g(t)$ and $\psi_0 - \psi_\tau$ are known, the conditions under which $\exp(-\psi_0 + \psi_x)$ may be replaced by unity in (4.7) and $O(R_v^3)$ terms neglected in (3.2) may be determined by direct examination of the integrals.

As might be expected, the third order modulation results are quite complicated. The third order modulation term in (3.2) is

$$\frac{1}{3!4}\int_{-\infty}^{\infty} df' \int_{-\infty}^{\infty} df'' w_\varphi(f') w_\varphi(f'') w_\varphi(f''') \left| \int_{-\infty}^{\infty} dx\, g(x) \cos px\, e^{-\psi_0+\psi_x}(z^{f'}-1)(z^{f''}-1)(z^{f'''}-1) \right|^2 \tag{4.11}$$

where $f''' = f - f' - f''$ and $z = \exp(-i2\pi x)$. When ψ_0 is small this is approximately

$$\frac{1}{3!16}\int_{-\infty}^{\infty} df' \int_{-\infty}^{\infty} df'' w_\varphi(f') w_\varphi(f'') w_\varphi(f''')[H^2 + K^2] \tag{4.12}$$

where

$$\begin{aligned} H = {} & m(f') + m(f'') + m(f) + m(f - f' - f'') \\ & - m(f - f') - m(f - f'') - m(0) - m(f' + f''), \end{aligned} \tag{4.13}$$

$$m(f) = G_f + G_{-f}, \qquad n(f) = B_f - B_{-f},$$

and K is an expression obtained from H by replacing n by m.

V MISCELLANEOUS COMMENTS

Here we make some miscellaneous comments related to the foregoing results.

If the transducer is perfect except for an echo, its response to a unit impulse $\delta(t)$ is

$$g(t) = \delta(t) + r\delta(t - T) \tag{5.1}$$

where r and T are the amplitude and the delay of the echo. The results obtained using (5.1) agree, as they should, with the results obtained in Reference 4. Of course, r must be assumed small compared to unity in order that condition (1.5) may hold.

When the power spectrum of the signal is equal to a constant P_0 over the band (f_a, f_b) and zero elsewhere we have for phase and frequency modulation, respectively,

$$\begin{aligned} &\text{PM:} \quad w_\varphi(f) = P_0, && f_a < f < f_b, \\ &\text{FM:} \quad w_\varphi(f) = P_0/(2\pi f)^2, && f_a < f < f_b. \end{aligned} \tag{5.2}$$

When $f_a = 0$ the autocorrelation functions are

$$\begin{aligned} &\text{PM:} \quad \psi_\tau = P_0 f_b (\sin v)/v, \\ &\text{FM:} \quad \psi_0 - \psi_\tau = \mathrm{A}[-1 + \cos v + vSi(v)], \\ &v = 2\pi f_b \tau, \qquad A = P_0 f_b (2\pi f_b)^{-2} = (\sigma/f_b)^2. \end{aligned} \tag{5.3}$$

The mean square values of the signals are $P_0 f_b$ (radians)^2 for PM and $P_0 f_b$ (radians/sec)^2 for FM. If, for FM, σ is the rms frequency deviation in cps (so that the "peak" deviation is, say, 4σ cps) then $(2\pi\sigma)^2 = P_0 f_b$. The difference $\psi_0 - \psi_\tau$ is used in the FM case to avoid difficulty at $f = 0$. It will be noticed that our formulas are such that the ψ's may be replaced by $(\psi - \psi_0)$'s without altering the values of the various exponents, etc. In microwave systems the quantity A is often small in comparison with unity.

As an example of the use of the second order modulation approximation (4.10) consider the case where the attenuation, α, is zero and the phase shift $\beta = a_2(f - f_p)^2/2$ radians, a_2 being small. Then, since $G \approx 1 - \alpha$, $\beta \approx -\beta$, we have $G_u \approx 1$ and

$$\begin{aligned} B_u &\approx -[\beta \text{ for } f = f_p + u] \\ &= -a_2 u^2/2. \end{aligned} \tag{5.4}$$

When we take the FM case of (5.2) and substitute in the approximation (4.10), the interchannel interference power spectrum is found to be

$$\begin{aligned} \frac{1}{2!\,8} \int_{f-f_b}^{f_b} \frac{P_0}{(2\pi u)^2} \frac{P_0}{(2\pi)^2 (f - u)^2} [0 + (2a_2 u(f - u))^2]\, du \\ = (2\pi)^{-4} (a_2 P_0/2)^2 (2f_b - f). \end{aligned} \tag{5.5}$$

Dividing by $w_\varphi(f) = P_0/(2\pi f)^2$ gives the ratio of the interference power to the signal power

$$(a_2\sigma f/2)^2(2 - f/f_b) \tag{5.6}$$

where the relation $P_0 = (2\pi\sigma)^2/f_b$ has been used to eliminate P_0. Here σ is the rms frequency deviation of the FM signal in cps. The expression (5.6) agrees with results of some earlier work done at Bell Telephone Laboratories. In that work the second order modulation products were summed directly.

It is interesting to apply the formulas given here to some of the cases considered by Medhurst and Small.[3] They have shown that when (in our notation) $\alpha = -r \cos 2\pi fT$ and $\beta = 0$ the power spectrum of the distortion is

$$w_{\theta-\bar{\theta}}(f) = \sin^2 \pi fT[w_{\theta-\bar{\theta}}(f)]_{\text{echo}}\,, \tag{5.7}$$

and when $\alpha = 0$ and $\beta = r \sin 2\pi fT$,

$$w_{\theta-\bar{\theta}}(f) = \cos^2 \pi fT[w_{\theta-\bar{\theta}}(f)]_{\text{echo}}\,. \tag{5.8}$$

Here $[w_{\theta-\bar{\theta}}(f)]_{\text{echo}}$ is the power spectrum of the distortion due to a simple echo of amplitude r and delay T (corresponding to $\alpha = -r \cos 2\pi fT$ and $\beta = r \sin 2\pi fT$). Expressions (5.7) and (5.8) may also be obtained by setting the impulse response $g(t)$ equal to

$$\delta(t) + \frac{r}{2}\,\delta(t - T) \pm \frac{r}{2}\,\delta(t + T)$$

in (3.2).

The second order modulation approximation for the $\alpha = -r \cos 2\pi fT$, $\beta = 0$ case may be obtained from (4.10) and turns out to be

$$\int_{-\infty}^{\infty} w_\varphi(u)w_\varphi(f - u)[2r \sin pT \sin \pi fT \sin \pi uT \sin \pi(f - u)T]^2\, du. \tag{5.9}$$

It is seen that this contains the factor $\sin^2 \pi fT$ predicted by (5.7). When (5.9) is applied to the FM case of (5.2) an integral something like (5.5) (but more complicated) is obtained. The ratio of the second order modulation interference power to the signal power is found to be

$$2[r \sin pT \sin \pi fT]^2(\sigma/f_b)^2UK \tag{5.10}$$

where K is the quantity

$$K = 2\alpha^2 \int_{\alpha-U}^{U} \left[\frac{\sin (y/2) \sin (\alpha - y)/2}{y(\alpha - y)}\right]^2 dy \tag{5.11}$$

tabulated in Table 4.2 of Reference 4 and

$$\alpha = 2\pi fT, \qquad U = 2\pi f_b T. \tag{5.12}$$

The parameters a and k that appear in the table are defined by

$$a = f/f_b \quad \text{and} \quad k = 8f_b T.$$

These formulas serve to supplement the formulas and curves given by Medhurst and Small.

ACKNOWLEDMENT

I wish to express my thanks to H. E. Curtis who has furnished some of the examples given in this paper and to E. D. Sunde, S. Doba, and others for their helpful comments.

REFERENCES

1. J. R. Carson and T. C. Fry, Variable Frequency Electric Circuit Theory With Application to the Theory of Frequency Modulation, B.S.T.J., **16,** 510–540, Oct., 1937.
2. B. van der Pol, The Fundamental Principles of Frequency Modulation, Jl. I.E.E., **93,** pp. 153–158, 1946.
3. R. G. Medhurst and G. F. Small, Distortion in Frequency-Modulation Systems Due to Small Sinusoidal Variations of Transmission Characteristics, Proc. I.R.E., **44,** pp. 1608–1612, Nov., 1956.
4. W. R. Bennett, H. E. Curtis, and S. O. Rice, Interchannel Interference in FM and PM Systems Under Noise Loading Conditions, B.S.T.J., **34,** pp. 601–636, May, 1955. The same problem has been treated independently and in much the same way by S. V. Borodich, On the Nonlinear Distortions Caused by Variations of the Antenna Feeder in Multichannel Frequency Modulation Systems, Radiotekhnika, Moscow, **10,** pp. 3–14, 1955.
5. These results are closely associated with some given by M. K. Zinn, Transient Response of an FM Receiver, B.S.T.J., **29,** pp. 714–731, 1948.

Self-Timing Regenerative Repeaters

By E. D. SUNDE

(Manuscript received March 29, 1956)

In self-timing regenerative repeaters, a timing wave for control in pulse regeneration is derived from the binary pulse train at each repeater with the aid of a resonant circuit tuned to the pulse repetition frequency. The timing wave can be made to exercise complete control in retiming of pulses independent of the received pulse train, or it can be combined with the received pulse train to provide partial retiming. The timing principles are discussed here for a particular type of self-timed regenerative repeater invented by Wrathall, in which a timing wave derived from either the received or the regenerated pulse train is combined in a particular way with the received pulse train. The regeneration characteristics of such repeaters as determined by various design parameters are investigated, together with the cumulation of timing deviations in repeater chains and the circuit requirements that must be met to insure satisfactory performance.

INTRODUCTION

Pulse transmission systems employing binary codes, such as PCM, have two inherent properties that are desirable from the standpoint of avoiding excessive transmission impairments by noise and other imperfections in the transmission medium. For one thing binary pulse codes permit substantial transmission distortion of pulses within certain tolerable limits with negligible degradation of received signals. For another, regenerative repeaters can be used at intervals along a route to prevent accumulation of transmission distortion of pulses from various sources, so that virtually the entire allowable distortion can be permitted in each link or repeater section.

The above desirable properties are secured in exchange for increased channel bandwidth, and can be used to full advantage in applications of binary pulse systems to such transmission media as radio and wave guides, where transmission is at such high frequencies that increased channel bandwidth does not entail increased attenuation. In wire circuits, however, where baseband transmission is the more economical

method, attenuation increases nearly in proportion to the square root of the channel bandwidth. For this reason, rather short repeater spacings may be required for binary pulse systems, so that for economical applications to wire circuits it is imperative to have reliable regenerative repeaters of simple design.

In their principle of operation regenerative repeaters are by nature more complicated than ordinary repeaters. In addition to providing gain to off-set attenuation in the transmission medium, as in ordinary repeaters, they must also perform gating operations for sampling and regenerating the received pulse train. This, however, does not preclude the possibility that these operational principles can be implemented in repeater design by instrumentation that is simpler than required for ordinary repeaters.

The possibility of simple instrumentation resides partly in the circumstance that equalization circuitry for regenerative repeaters can be substantially simpler than for ordinary repeaters, owing to less exacting requirements on equalization. Furthermore, satisfactory performance in pulse regeneration can be achieved without very precise timing in sampling and regeneration of pulse trains. It is thus possible to secure nearly the same performance as for ideal regenerative repeaters by partial rather than complete exact retiming of pulse trains at each repeater. This facilitates simple gating arrangements for regeneration of pulses. Moreover, it permits a timing wave for control of gating operations to be derived from either the received or regenerating pulse trains with the aid of a simple resonant of circuit.

The simplicity of instrumentation permitted by these considerations is exemplified in a self-timed regenerative repeater for baseband pulses invented by L. R. Wrathall of Bell Telephone Laboratories. The circuitry of the repeater together with the results of tests on laboratory models are dealt with elsewhere[1] and not considered here. The purpose of this paper is an analysis of the timing principles underlying this type of repeater together with its regeneration characteristics as determined by various basic design parameters, on the assumption of ideal implementation of the timing principles by appropriate instrumentation. In the Wrathall repeater "quantized feed-back" is employed as a means of reducing the effect of low-frequency cut-off in transformers. Since this is not an essential feature of self-timing repeaters and has no direct bearing on the timing principles, it is disregarded herein.

---

[1] L. R. Wrathall, Transistorized Binary Pulse Regenerator, B.S.T.J., **35**, pp. 1059–1084, Sept., 1956.

I REGENERATION AND RETIMING

1.0 General

In an ideal regenerative repeater the received pulse train is sampled at proper fixed intervals, to determine whether a pulse is present. The regenerated pulses transmitted into the next repeater section are all of the same shape and amplitude, independent of the shape of the input pulses. Thus pulse distortion from noise and other system imperfections is removed, provided the maximum distortion is held within proper limits. Errors in the form of pulses in place of spaces, or conversely, are encountered when these limits are exceeded. In a repeater chain there will be cumulation of errors in proportion to the number of repeater sections in tandem. However, the rate of errors in each section and thus in the whole chain can be limited by a relatively small increase in the signal-to-noise ratio of each section as the number of repeaters in tandem is increased. This increase in signal-to-noise ratio with increasing length of the repeater chain is much less than with ordinary nonregenerative repeaters. For this reason regenerative rather than ordinary repeaters are desirable, though not essential for systems employing binary codes.

An ideal regenerative repeater with the above features would entail rather complicated instrumentation for precise timing, sampling and pulse regeneration. With partial rather than complete exact retiming the repeaters can be simplified, in exchange for some sacrifice in performance, as shown later.

1.1 Regeneration Without Retiming

It would be possible to have a repeater in which pulses would be regenerated in amplitude and shape, but without retiming. Pulses would in this case be regenerated when the amplitude of the pulse exceeded a certain triggering level L. If the pulse shape is given by $P(t)$, this would occur at a time t_0 such that

$$P(t_0) = L. \tag{1.1}$$

This would permit simple instrumentation, since regenerated pulses would be triggered without separate sampling of the received pulse train. With this method, however, timing deviations in the regenerated pulses would result from transmission distortion of the received pulses by noise and other system imperfections. These timing deviations would cumulate in a repeater chain and cause a reduction in the tolerance of the repeaters to noise, such that the signal-to-noise ratio would have to

be increased with the number of repeaters in tandem in the same way as for ordinary repeaters.

1.2 Regeneration with Complete Retiming

With complete retiming, the instants of pulse regeneration would be controlled by a periodic retiming wave, $R(t)$, with a fundamental period equal to the interval between pulses. The received pulse train would be sampled at instants when the retiming wave had a certain level L_s . The sampling instants t_0 would thus be given by

$$R(t_0) = L_s . \tag{1.2}$$

$R(t_0)$ would satisfy this equation for $t_0 = nT \pm \Delta T$, where T is the nominal interval between pulses, n is an integer and ΔT is a certain tolerable deviation from the desired sampling instants. Pulses would be regenerated provided $P(t_0) > L$ and would be omitted if $P(t_0) < L$.

With this method the timing deviations in regenerated pulses would be limited to $\pm\Delta T$, regardless of the timing deviations in received pulses. There would be no cumulation of timing deviations in a repeater chain. However, the tolerance of the repeaters to noise would be somewhat reduced by the timing deviations $\pm\Delta T$.

1.3 Regeneration with Partial Retiming

Partial retiming is obtained by a combination of the above two methods, by triggering regenerated pulses without sampling at instants t_0 determined by

$$P(t_0) + R(t_0) = L. \tag{1.3}$$

To permit regeneration without sampling and without a marked reduction in the tolerance of the repeaters to noise, the timing wave $R(t)$ must meet certain conditions illustrated in Fig. 1. One is that it must be a nearly periodic function as for complete retiming. The second condition is that $R(t)$ must be zero near the sampling points to obtain substantially the same tolerance to noise in the presence of a pulse as in the absence of a pulse. A third condition is that $R(t)$ must have substantial negative values between sampling points in order that the repeater be rather insensitive to noise between sampling points, as with complete retiming. It will be recognized that, in general, the maximum value of $R(t)$ need not necessarily be zero, as in the above illustration. It can be greater or smaller than zero, provided the triggering level is

modified accordingly. A maximum value of zero is, however, convenient from the standpoint of instrumentation.

A limiting shape of retiming wave that would result in complete retiming, but without the need for special sampling is also illustrated in Fig. 1.

1.4 Derivation of Timing Wave from Pulse Train

As shown above, the retiming wave must be essentially periodic, with a fundamental frequency equal to the pulse repetition frequency $f = 1/T$, where T is the interval between pulses. The simplest form is a sinusoidal wave, which can be derived from the pulse train at repeaters with the aid of a narrow band-pass filter, such as a simple resonant circuit cen-

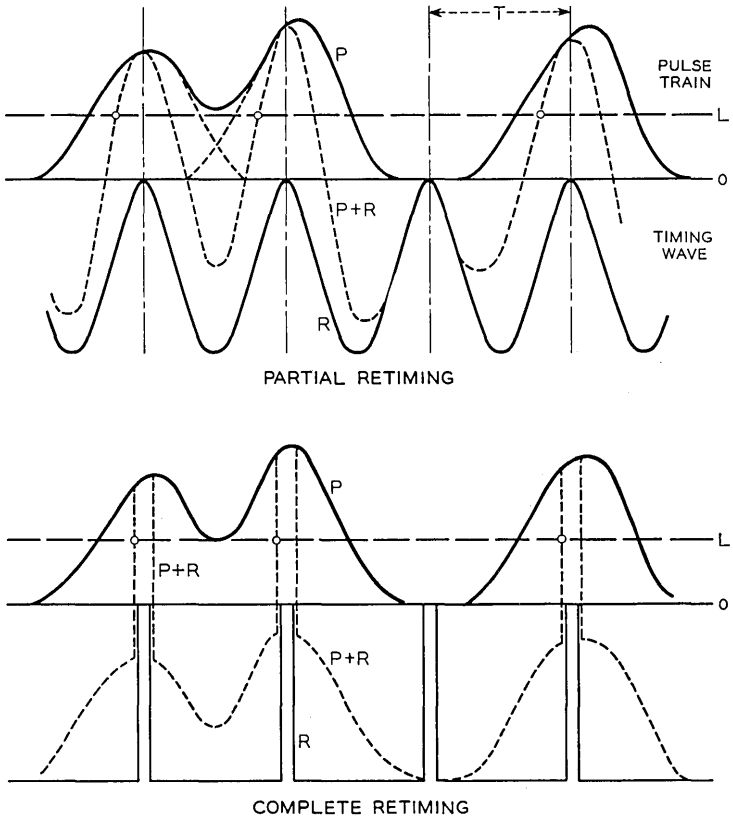

Fig. 1 — Principle of partial retiming method.

tered on the pulse repetition frequency. This possibility resides in the circumstance that a random "on-off" pulse train can be resolved into two components. One is an infinite sequence of pulses of the same polarity and equal amplitude, the other a sequence of randomly positive and negative polarity. The response of a resonant circuit to the first component is a steady state sinusoidal wave of the pulse repetition frequency. The second component gives rise to random variations in amplitude and phase, which in principle can be limited to any desired extent by limiting the band of the resonant circuit and the deviation in the resonant frequency from the pulse repetition frequency.

A principal feature of this method of "self-timing", aside from its simplicity, is that the timing wave becomes a slave of the pulse train. Thus, if there is a fixed delay in pulse regeneration at a repeater, the same delay is imparted to the timing wave derived from the pulse train at the next repeater. This prevents a cumulation of such fixed delays with respect to the timing wave, but not with respect to an absolute time scale; i.e., with respect to an ideal timing wave transmitted along the repeater chain and independent of the pulse train.

1.5 Self-Timed Repeaters with Partial Retiming

A timing wave derived from the pulse train with the aid of a resonant circuit can be used in conjunction with complete or partial retiming. With complete retiming, pulses could be regenerated at the zero points in the timing wave, and the effects of amplitude variations in the timing wave can thus be avoided. Timing deviations in the regenerated pulses would in this case depend only on phase deviations in the timing wave, caused partly by the component of randomly positive and negative polarity in the pulse train and partly by timing deviations in the pulse train from which the timing wave is derived.

With partial retiming the situation is more complex. Timing deviations in regenerated pulses in this case depend not only on amplitude and phase variations in the timing wave, but also on the regeneration characteristics of the repeaters.

1.6 Types of Timing Deviations

In a regenerated pulse train there will be fixed and random timing deviations. Of the latter there are three types. One is the timing deviation taken in relation to an exact timing wave with a period T equal to the nominal pulse interval. The second is the timing deviation taken in relation to the timing wave derived from the pulse train, which in itself

will contain random deviations. The third type is random deviations in the interval of adjacent pulses. If the first type is held within tolerable limits, this will also be the case for the second and third types. For this reason only the first type is considered herein.

II REGENERATION CHARACTERISTICS WITH PARTIAL RETIMING

2.0 General

With partial retiming, there will be timing deviations in the regenerated pulses as a result of timing deviations, amplitude variations and distortion by noise of both the received pulses and the timing wave.

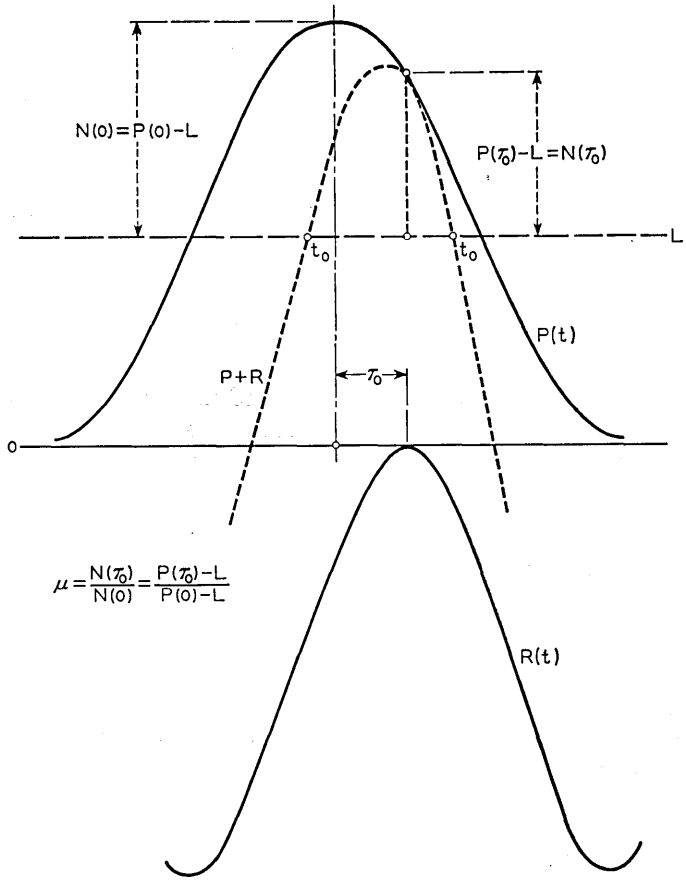

Fig. 2 — Reduction in tolerance to noise by displacement in timing wave.

The conversion of these variations into timing deviations in the regenerated pulses depends on certain relationships between the pulse train and the timing wave, discussed in the following sections.

2.1 Tolerance to Noise

From Fig. 2 it can be seen that if the timing wave is displaced by τ_0, the value of $P(t) + R(t - \tau_0)$ in the presence of a pulse exceeds the triggering level by a maximum amount

$$[P(t) + R(t - \tau_0) - L]_{\max} \cong [P(\tau_0) - L]. \tag{2.1}$$

It will be recognized that the right-hand side of this equation represents the tolerance to noise of negative amplitudes with instantaneous sampling at $t = \tau_0$, as in an ideal repeater with complete retiming.

With partial retiming, the tolerance to noise will be less than the above maximum value. However, it will be greater than the average of $P(t) + R(t - \tau_0) - L$ in the range where the latter difference is positive. Let it be assumed that it is smaller than the maximum by a factor k somewhat smaller than unity. The tolerance to noise with a displacement τ_0 in the timing wave is then smaller than without a displacement (i.e., $\tau_0 = 0$) by the factor

$$\mu = \frac{k[P(\tau_0) - L]}{k[P(0) - L]} = \frac{P(\tau_0) - L}{P(0) - L}. \tag{2.2}$$

The tolerance to noise will thus be reduced in a way similar to that for an ideal repeater with complete retiming. The absolute tolerance to noise will be less than for a repeater with complete retiming by a factor k somewhat smaller than unity, say in the order 0.8, corresponding to about 2 db.

2.2 Conversion of Timing Deviations

With partial retiming, timing deviations in received pulses and in the timing wave are converted into smaller deviations in regenerated pulses.

Let τ_p be a time displacement in a received pulse and τ_r in the timing wave, both in the positive direction. Pulses will then be regenerated at a time t_0' given by

$$P(t_0' - \tau_p) + R(t_0' - \tau_r) = L \tag{2.3}$$

where the minus signs are used since this corresponds to a displacement of P and R in the positive direction. Subtracting (1.3) from (2.3),

$$P(t_0' - \tau_p) - P(t_0) + R(t_0' - \tau_r) - R(t_0) = 0. \tag{2.4}$$

By adding and subtracting $P(t_0') + R(t_0')$ and rearranging terms, (2.4) can also be written

$$[P(t_0') - P(t_0)] + [R(t_0') - R(t_0)] = [P(t_0') - P(t_0' - \tau_p)] + [R(t_0') - R(t_0' - \tau_r)]. \tag{2.5}$$

For small values of τ_p and τ_r, such that $\delta_\tau = t_0' - t_0$ is sufficiently small, both sides of (2.8) can be represented in differential form as

$$\delta_\tau[P'(t_0) + R'(t_0)] = \tau_p P'(t_0) + \tau_r R'(t_0) \tag{2.6}$$

where $P'(t_0) = dP_0(t)/dt$ at $t = t_0$, and R' is correspondingly defined.

Equation (2.9) can be written in the form

$$\delta_\tau = p_\tau \tau_p + r_\tau \tau_r \tag{2.7}$$

where

$$p_t = \frac{P'(t_0)}{P'(t_0) + R'(t_0)}, \qquad r_\tau = \frac{R'(t_0)}{P'(t_0) + R'(t_0)}, \tag{2.8}$$

and

$$p_\tau + r_\tau = 1. \tag{2.9}$$

With random uncorrelated displacements of rms values $\bar{\tau}_p$ and $\bar{\tau}_r$, the rms value of δ_τ is

$$\bar{\delta}_\tau = (p_\tau^2 \bar{\tau}_p^2 + r_\tau^2 \bar{\tau}_r^2)^{1/2} \tag{2.10}$$

Equation (2.9) and (2.10) give the timing deviations in regenerated pulses in terms of the deviations τ_p and τ_r in the received pulses and in the timing wave. To limit timing deviations in the regenerated pulses, it is necessary to make p_τ and the product $r_\tau \tau_r$ small. This will entail the use of a timing wave comparable in amplitude to that of the pulses, or greater, in conjunction with a small timing deviation τ_r in the timing wave.

2.3 Conversion of Amplitude Variations Into Timing Deviations

With partial retiming there is a conversion of amplitude variations in the received pulses and in the timing wave into timing deviations in the regenerated pulses.

Let the pulses have an amplitude variation a_p and the timing wave a_r expressed as fractions of the normal values. Pulses will then be regenerated at a time t_0' given by

$$(1 + a_p)P(t_0') + (1 + a_r)R(t_0') = L. \tag{2.11}$$

Subtracting (1.3) from (2.11),

$$[P(t_0') - P(t_0)] + [R(t_0') - R(t_0)] = -a_p P(t_0') - a_r P(t_0').$$

For small values of a_p and a_r, such that $\delta_a = t_0' - t_0$ is sufficiently small, the same procedure as in Section 2.2 gives

$$\delta_a = (p_a a_p + r_a a_r), \tag{2.12}$$

and

$$p_a = \frac{-P(t_0)}{P'(t_0) + R'(t_0)}, \qquad r_a = \frac{-R(t_0)}{P'(t_0) + R'(t_0)}. \tag{2.13}$$

For uncorrelated variations of rms amplitude $\underline{a}_p$ and $\underline{a}_r$ the corresponding rms timing deviation is

$$\underline{\delta}_a = (p_a^2 \underline{a}_p^2 + r_a^2 \underline{a}_r^2)^{1/2}. \tag{2.14}$$

Equations (2.12) and (2.14) give the timing deviations in regenerated pulses resulting from amplitude variations in the pulses and in the timing wave.

2.4 Resultant Timing Deviations in Regenerated Pulses

For small variations in the pulses and in the timing wave as considered previously, the resultant timing deviation in a particular regenerated pulse is

$$\Delta = \delta_\tau + \delta_a. \tag{2.15}$$

Considering a large number of pulses, the resultant rms timing deviation in terms of the rms deviation in the received pulses and in timing wave is

$$\underline{\Delta} = (\underline{\delta}_\tau^2 + \underline{\delta}_a^2)^{1/2}. \tag{2.16}$$

These expressions can also be written

$$\Delta = \Delta_p + \Delta_r, \tag{2.17}$$

$$\underline{\Delta} = (\underline{\Delta}_p^2 + \underline{\Delta}_r^2)^{1/2}, \tag{2.18}$$

$$\Delta_p = p_\tau \tau_p + p_a a_p,$$

$$\underline{\Delta}_p^2 = p_\tau^2 \bar{\tau}_p^2 + p_a^2 \underline{a}_p^2, \tag{2.19}$$

$$\Delta_r = r_\tau \tau_r + r_a a_r,$$

$$\underline{\Delta}_r^2 = r_\tau^2 \bar{\tau}_r^2 + r_a^2 \underline{a}_r^2. \tag{2.20}$$

III ILLUSTRATIVE REGENERATION CHARACTERISTICS

3.0 General

In this section the general equations given in the preceding sections are applied to a particular case, in order to obtain specific expressions for the regeneration characteristics and illustrative curves, as an aid to further analysis. The particular case selected for illustration approximates the conditions in experimental Wrathall repeaters, and may be regarded as an idealized model of such a repeater, in which certain effects to be discussed later are ignored.

3.1 Pulse Shape

It will be assumed that the pulses are transmitted at intervals T and that the shape of the received pulses after equalization is given by:

$$P(t) = \frac{1}{2}\left[1 + \cos\frac{\pi}{\eta}\frac{t}{T}\right]. \tag{3.1}$$

This is the familiar "raised cosine" type of pulse. With $\eta = 1$ the pulse width is the maximum that can be tolerated without intersymbol interference. With $\eta = \frac{3}{4}$, the amplitude of a pulse train at a point midway between two success pulses is equal to half the peak amplitude of a pulse. The latter assumption will be made here, for reasons discussed later.

3.2 Retiming Wave

The retiming wave is assumed to be given by

$$R(t) = -\frac{1}{2}\cos\psi\left[1 - \cos\left(2\pi\frac{t}{T} - \psi\right)\right]. \tag{3.2}$$

This type of retiming wave can be obtained if a sinusoidal wave of the pulse repetition frequency $f = 1/T$ is applied to a resonant circuit to reduce distortion of the timing wave by noise. The resonant circuit would have a nominal resonant frequency $f = 1/T$, but because of mistuning it would actually be f_0. The output of the resonant circuit after appropriate adjustment of amplitude would be of the form [Appendix I, equation (2)]:

$$R_0(t) = \frac{1}{2}\cos\psi\cos\left(2\pi\frac{t}{T} - \psi\right), \tag{3.3}$$

where ψ is the phase shift of the resonant circuit at the frequency f,

given by:

$$\tan \psi = Q\left(\frac{f}{f_0} - \frac{f_0}{f}\right), \tag{3.4}$$

and Q is the loss constant of the resonant circuit. If the peaks of the wave given by (3.3) are held at zero potential, a retiming wave as given by (3.2) is obtained. This type of retiming wave can also be obtained by applying an infinite sequence of rectangular pulses of equal amplitudes with spacing T to a resonant circuit.

3.3 Triggering Instants

With a pulse shape and retiming wave as assumed above, the resultant wave is given by

$$P(t) + R(t) = \frac{1}{2}\left[1 + \cos\frac{\pi}{\eta}\frac{t}{T}\right] - \frac{\cos\psi}{2}\left[1 - \cos\left(2\pi\frac{t}{T} - \psi\right)\right]. \tag{3.5}$$

This wave is shown in Fig. 3 for $\psi = 0$ and $\pm 60°$. For $\psi = \pm 90°$ the retiming wave disappears, so that the combined wave is $P(t)$.

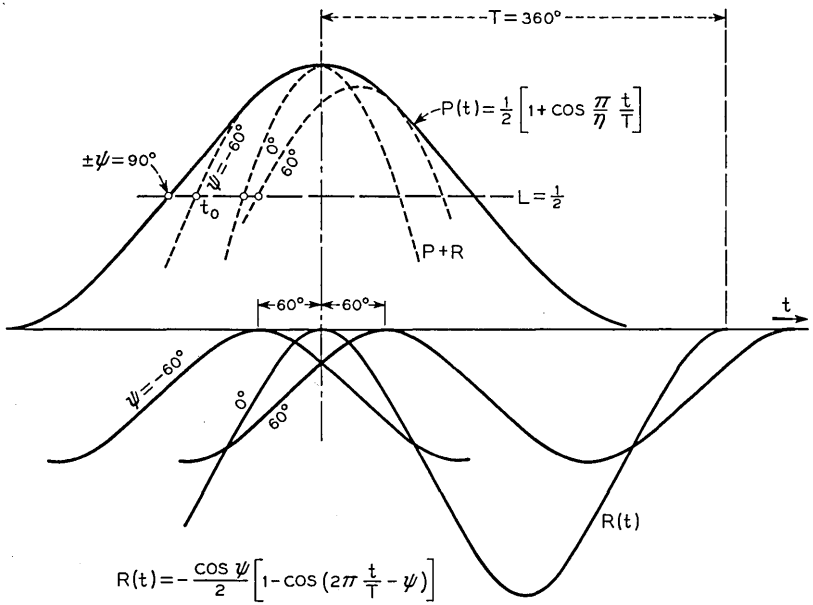

Fig. 3 — Illustrative example of pulse shape and retiming wave.

The triggering instants t_0 are obtained from the relation

$$P(t_0) + R(t_0) = L. \tag{3.6}$$

With complete retiming the optimum performance, with positive and negative noise amplitudes of equal probabilities, is obtained with a triggering level $\frac{1}{2}$. With partial retiming, optimum performance is obtained with a somewhat lower triggering level, but this is of secondary importance in connection with the present analysis. For this reason $L = \frac{1}{2}$ is assumed, in which case the following equation is obtained for determination of t_0 :

$$\cos \frac{\pi}{\eta} \frac{t_0}{T} - \cos \psi \left[1 - \cos \left(2\pi \frac{t_0}{T} - \psi \right) \right] = 0 . \tag{3.7}$$

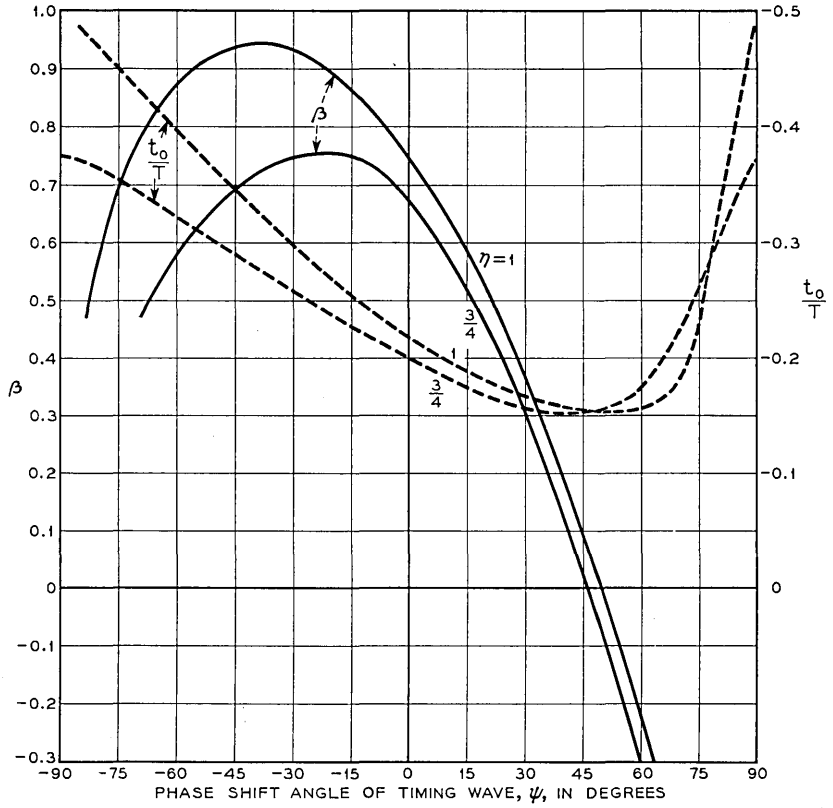

Fig. 4 — Triggering times versus phase shifts in timing wave.

TABLE I — VALUES OF t_0/T FOR $\eta = \frac{3}{4}$ AND $\eta = 1$

ψ	−90°	−60°	−30°	0	30°	60°	90°
$\eta = \frac{3}{4}$	−0.375	−0.322	−0.258	−0.198	−0.156	−0.17	−0.375
$\eta = 1$	−0.50	−0.391	−0.293	−0.215	−0.170	−0.15	−0.50

TABLE II — VALUES OF p_τ AND r_τ FOR $\eta = \frac{3}{4}$

ψ	−90°	−60°	−30°	0	30°	60°	90°
p_τ	1	0.61	0.43	0.32	0.32	0.50	1
r_τ	0	0.39	0.57	0.68	0.68	0.50	0

This equation is satisfied for the values of t_0/T given in Table I. The values of t_0/T are also shown in Fig. 4 as a function of ψ.

3.4 Conversion Factors for Time Deviations

The conversion factors defined by (2.8) become:

$$p_\tau = \frac{1}{D} \sin \frac{\pi}{\eta} \frac{t_0}{T} = 1 - r_\tau , \tag{3.8}$$

$$r_\tau = \frac{1}{D} 2\eta \cos \psi \sin \left(2\pi \frac{t_0}{T} - \psi \right), \tag{3.9}$$

and

$$D = \sin \frac{\pi}{\eta} \frac{t_0}{T} + 2\eta \cos \psi \sin \left(2\pi \frac{t_0}{T} - \psi \right), \tag{3.10}$$

where t_0/T has the values given previously as a function of ψ.

For various values of ψ, the factors for $\eta = \frac{3}{4}$ are given in Table II and in Fig. 5.

3.5 Conversion Factors for Amplitude Into Time Deviations

The conversion factors defined by (2.13) become

$$p_a = -\frac{T\eta}{\pi} \frac{1}{D} \left[1 + \cos \frac{\pi}{\eta} \frac{t_0}{T} \right], \tag{3.11}$$

and

$$r_a = \frac{T\eta}{\pi} \frac{1}{D} \cos \psi \left[1 - \cos \left(2\pi \frac{t_0}{T} - \psi \right) \right], \tag{3.12}$$

where D and t_0/T are defined as before.

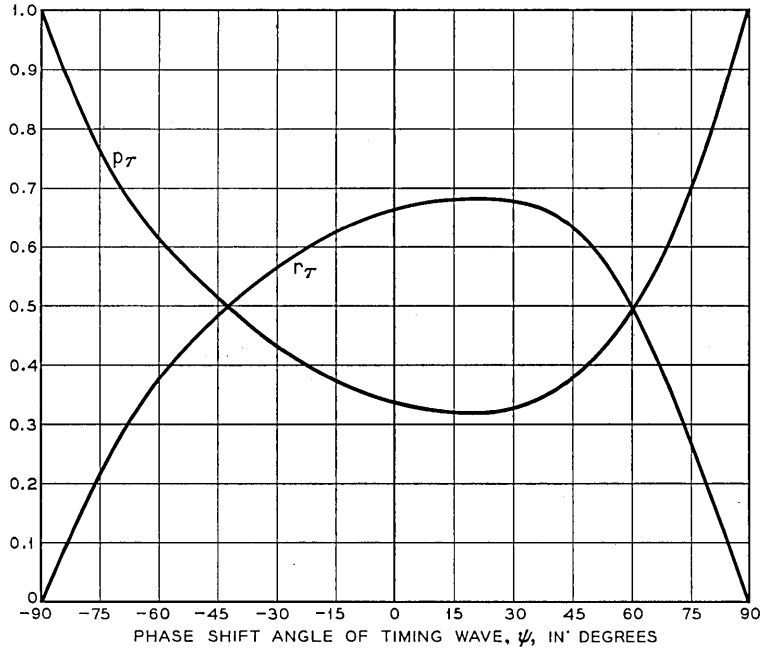

Fig. 5 — Conversion of timing deviations in received pulses and in timing wave into timing deviations in regenerated pulses, for pulse shapes and timing waves shown in Fig. 3. Timing deviations in regenerated pulses in relation to timing deviation t_p in pulses and t_r in retiming wave is $p_\tau t_p + r_\tau t_r$.

For various values of ψ the factors for $\eta = \frac{3}{4}$ are given in Table III and in Fig. 6.

3.6 Correlated Amplitude and Time Deviations

The amplitude and time deviations in the pulses are generally uncorrelated, but this does not always apply to the timing wave. In particular, if a deviation τ_r in the timing wave is the result of a change in the phase ψ, it will be accompanied by a given amplitude variation. A change in phase by $\Delta\psi$ is related to the corresponding time deviation τ_r by

$$\Delta\psi = \frac{2\pi}{T}\tau_r . \tag{3.13}$$

TABLE III — VALUES OF p_a/T AND r_a/T FOR $\eta = \frac{3}{4}$

ψ	−90°	−60°	−30°	0	30°	60°	90°
p_a/T	−0.24	−0.185	−0.175	−0.19	−0.22	−0.325	−0.24
r_a/T	0	0.035	0.055	0.072	0.106	0.14	0

With this change in phase, the factor $\cos \psi$ of (3.2) is modified to

$$\begin{aligned} \cos (\psi + \Delta\psi) &= \cos \psi \cos \Delta\psi - \sin \psi \sin \Delta\psi, \\ &\cong \cos \psi - \frac{2\pi}{T} \tau_r \sin \psi \end{aligned} \tag{3.14}$$

where the approximation applies for small values of ψ. The amplitude variation resulting from the above change in phase is accordingly

$$a_r = -\tau_r \frac{2\pi}{T} \sin \psi. \tag{3.15}$$

Considering both the time deviation τ_r and the corresponding amplitude variation a_r, the resultant time deviation in regenerated pulses is in accordance with (2.20)

$$\Delta_r = r_\tau \tau_r + r_a a_r . \tag{3.16}$$

The resultant equation can be written

$$\Delta_r = \beta \tau_r \tag{3.17}$$

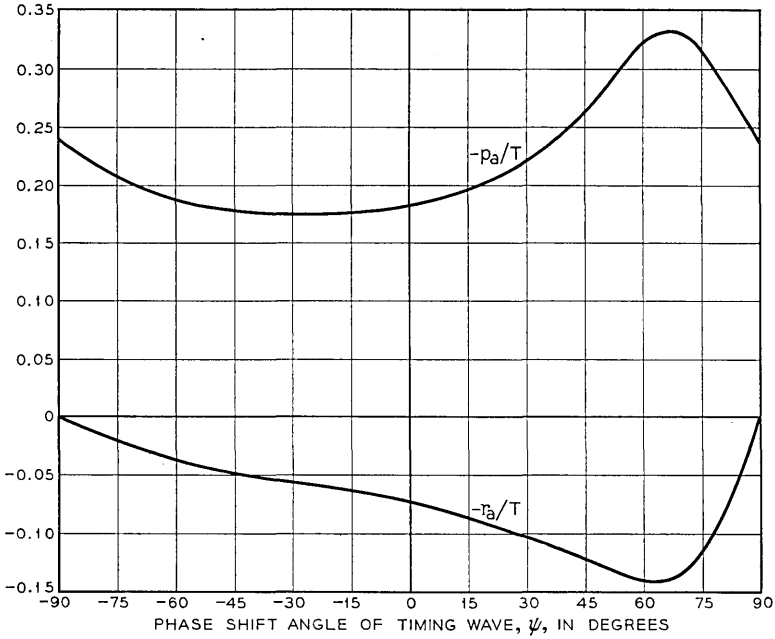

Fig. 6 — Conversion of amplitude variations in received pulses and in timing wave into timing deviations in regenerated pulses, for pulse shapes and timing waves shown in Fig. 3. Timing deviations in regenerated pulses for amplitude variations a_p and a_r in received pulses and in timing wave is $p_a a_p + r_a a_r$.

where

$$\beta = \frac{2\eta \cos\psi}{D}\left\{\sin\left(2\pi\frac{t_0}{T} - \psi\right) + \sin\psi\left[1 - \cos\left(2\pi\frac{t_0}{T} - \psi\right)\right]\right\} \tag{3.18}$$

and D and t_0/T are defined as before.

The factor β indicates the time deviation in regenerated pulses in relation to the time deviation τ_r in the timing wave which results from a phase shift $\Delta\psi$ as given by (3.13). It may be regarded as a timing feedback factor that is of interest in connection with timing from regenerated pulses as discussed later. The factor β is shown in Fig. 4 for $\eta = \frac{3}{4}$ and $\eta = 1$.

3.7 Reduction in Tolerance to Noise by Timing Deviations

When the pulse shape is given by (3.1) and the timing wave is displaced by τ_0, the tolerance to noise is in accordance with (2.2) reduced by the factor

$$\mu = \frac{\frac{1}{2}\left(1 + \cos\frac{\pi}{\eta}\frac{\tau_0}{T}\right) - \frac{1}{2}}{\frac{1}{2}\left(1 + \cos\frac{\pi}{\eta}\frac{0}{T}\right) - \frac{1}{2}} = \cos\frac{\pi}{\eta}\frac{\tau_0}{T}. \tag{3.19}$$

For a phase displacement ψ,

$$\tau_0 = T\psi/2\pi, \tag{3.20}$$

and

$$\mu = \cos\frac{\psi}{2\eta}. \tag{3.21}$$

For $\eta = \frac{3}{4}$, the factor μ and the corresponding reduction in the tolerance to noise in db are as follows:

ψ =	0	±30°	±45°	±60°	±90°
μ =	1	0.94	0.866	0.766	0.5
μ_{db} =	0	0.5	1.2	2.3	6

IV DERIVATION OF TIMING WAVE FROM PULSE TRAIN

4.0 General

The retiming wave $R(t)$ must have a fixed relation to the received pulses, with certain tolerable fixed and random deviations to be considered later. Such a timing wave can be derived from the pulse train with the aid of a sufficiently narrow band-pass filter, the simplest form of which is a resonant circuit consisting of a coil and capacitor in series or in parallel.

A train of rectangular "on-off" pulses is shown in Fig. 7 as it would appear at the output of a regenerative repeater and at the input of the next repeater, (dotted) with uniform intervals T between sampling points.

As indicated in Fig. 7, the pulse train can be regarded as being made up of two components. One of these is an infinite sequence of pulses of one polarity, the other an infinite sequence of randomly positive and negative polarity.

It will be recognized that the first of the above components at the output has a fundamental frequency equal to the pulse repetition frequency, $f = 1/T$, and the forced response of a resonant circuit to this component will be the pulse repetition frequency, regardless of any imperfections in tuning. In order that this frequency be present in the received pulse train, it is necessary that the spectrum of the received pulses extend beyond the pulse repetition frequency, so that there will be a ripple in a long sequence of received pulses of one polarity, as indicated in the illustration.

The second random component of the pulse train will have a frequency spectrum that is nearly uniform over the band of the tuned circuit, and which will vary in amplitude depending on the composition of the pulse train. The response of the tuned circuit to this component is thus rather complex, and must be treated on an approximate statistical basis. It will consist of an almost periodic wave with random amplitude and phase modulation, and with mean frequency equal to the resonant frequency.

Owing to the presence of the second component, there will be a variation with time in the amplitude and phase of the response of a mistuned resonant circuit, and resultant deviations in timing. The regenerated pulses will thus not be uniformly spaced, but will in general have random deviations from the desired exact positions. Such deviations can be created by superposing on a train with uniform spacing a random dipulse train, as indicated in Fig. 7. The resonant circuit response to this

dipulse train would be expected to be smaller than to the random amplitude component of the pulse train. It may be regarded as a third component representing a second order effect resulting from the second component.

In the Appendix, this method of superposition has been used as a basis of an analysis of a resonant circuit response to a random binary pulse train. This problem has also been dealt with by somewhat different methods in prior unpublished work by W. R. Bennett and J. R. Pierce, both of Bell Telephone Laboratories.

In this analysis it is assumed that the regenerated pulses are of sufficiently short duration to be regarded as impulses. The response of the resonant circuit to the second and third components above, when taken in relation to that for the first component will, however, remain very

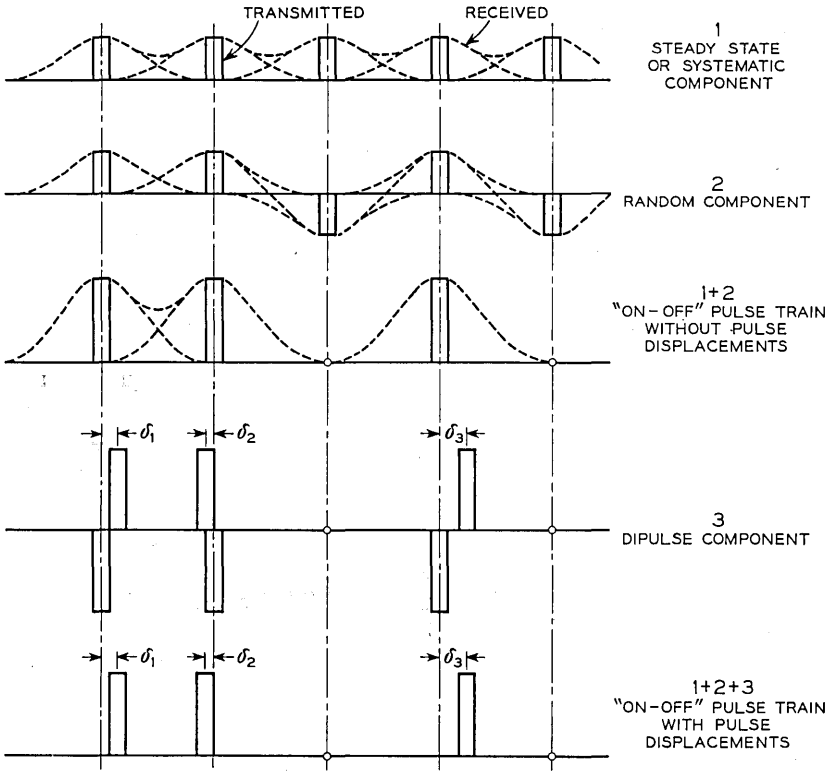

Fig. 7 — Resolution of "on-off" pulse train with timing deviations into systematic component (1), random component (2), and time displacement component (3).

nearly the same for other pulse shapes, provided the frequency spectrum of the pulses can be regarded as approximately constant over the important portion of the band of the resonant circuit. This approximation is legitimate for resonant circuits with a loss constant Q and pulse shapes at the input of repeaters as considered here.

4.1 Resonant Circuit Response to Steady State Component

The first component consists of an infinite sequence of impulses of amplitude $\frac{1}{2}$ and all of the same polarity, at intervals T. This sequence has a fundamental frequency $f = 1/T$. When it impinges on a resonant circuit with resonant frequency $f_0 = f - \Delta f$ and loss constant Q, the response is of the form

$$A_s(t) = \cos \psi \cos (\omega t - \psi), \tag{4.1}$$

and

$$\tan \psi = Q(f/f_0 - f_0/f) \cong 2Q \frac{\Delta f}{f}. \tag{4.2}$$

The response is thus a steady state sinusoidal wave of frequency f displaced from the fundamental component of the input wave by the phase shift ψ and reduced in amplitude by $\cos \psi$. This is the phase shift and amplitude reduction of the resonant circuit at the frequency f when the resonant frequency is f_0.

4.2 Resonant Circuit Response to Random Signal Component

The second component consists of an infinite random sequence of impulses of amplitude $\pm\frac{1}{2}$, at intervals T. The response of the resonant circuit to this component will be a randomly fluctuating wave $A_r(t)$ of mean value 0. The maximum positive amplitude is obtained when all impulses of the second component are positive and is $A_r(t) = A_s$. The maximum negative amplitude is $A_r(t) = -A_s$. Owing to the presence of this component the total output of the resonant circuit $A_s + A_r(t)$ can thus fluctuate between the limits 0 and $2A_s$, but the actual fluctuations of significant probability will be smaller.

The above fluctuations can be resolved into a component in phase with the steady state response given by (4.1) and another component at quadrature with the steady state timing wave. The rms values of these components taken in relation to the amplitude of the steady state wave are

$$\underline{a_r}' = \underline{A_r}'/A_s = \left(\frac{\pi}{2Q}\right)^{1/2} [1 - \psi^2/2]^{1/2} \frac{1}{\cos \psi}, \tag{4.3}$$

and

$$\underline{a}_r'' = \underline{A}_r''/A_s = \left(\frac{\pi}{4Q}\right)^{1/2} \frac{|\psi|}{\cos\psi}. \tag{4.4}$$

These relations apply for small values of ψ and for $\pi/Q \ll 1$.

The resultant rms amplitude variation in the timing wave is $\underline{a}_r = \underline{a}_r'$ as given by (4.3).

The rms phase error $\bar{\varphi}_r$ resulting from the quadrature component $\underline{a}_r''$ is given by

$$\tan\bar{\varphi}_r \cong \bar{\varphi}_r = a_r''. \tag{4.5}$$

The corresponding rms time deviation is $(T/2\pi)\bar{\varphi}_r$ or

$$\bar{\delta}_r = \frac{T}{2\pi}\left(\frac{\pi}{4Q}\right)^{1/2} \frac{|\psi|}{\cos\psi}. \tag{4.6}$$

With regard to the probability of exceeding the above rms values by various factors the normal law can probably be invoked with reasonable accuracy. As mentioned before, the maximum possible amplitudes are $\hat{A}_r(t) = \pm A_s$ which would correspond to a peak factor $(2Q/\pi)^{1/2}$. With $Q = 100$, the factor is about 8, while with $Q = 1000$ it is about 25. Based on the normal law the probability of exceeding the rms value by a factor of 4 is about 5×10^{-5}, and by a factor of 5, about 10^{-7}. The normal law would be expected to apply, since the limiting peak values are substantially greater than the peak values expected with significant probabilities.

4.3 Resonant Circuit Response to Pulse Displacements

Because of the random components given by (4.3) and (4.4), the timing wave will contain small random amplitude and phase deviations from a sinusoidal wave represented by (4.1). This will result in small random deviations in the positions of regenerated pulses triggered from the timing wave, which is represented by the third component shown in Fig. 7. When the rms deviation in the pulse positions is $\bar{\delta}$, there will be an additional random quadrature component in the timing wave which, when taken in relation to the steady state component, is given by

$$\underline{a}_\delta'' = \underline{A}_\delta''/A_s = \omega\bar{\delta}\left(\frac{\pi}{Q}\right)^{1/2}. \tag{4.7}$$

The corresponding rms phase deviation is given by

$$\bar{\varphi}_\delta \cong \underline{a}_\delta''. \tag{4.8}$$

The resultant rms time deviation is $(T/2\pi)\bar{\varphi}_\delta$ or

$$\delta_\delta = \delta\alpha, \tag{4.9}$$

and

$$\alpha = (\pi/Q)^{1/2}. \tag{4.10}$$

The above factor α applies to a single resonant circuit. When the rms timing deviations represented by (4.9) are present in the regenerated pulse train, the rms deviation at the output of the second resonant circuit is

$$\delta_{\delta,2} = \delta\alpha_1\alpha_2 ,$$

where

$$\alpha_1 = \alpha.$$

With n resonant circuits in tandem,

$$\delta_{\delta,n} = \delta\alpha_1\alpha_2\alpha_3 \cdots \alpha_n . \tag{4.11}$$

The factors α_n are given by

$$\alpha_1 = \alpha = (\pi/Q)^{1/2}, \tag{4.12}$$

$$\alpha_j = \left(1 - \frac{1}{2(j-1)}\right)^{1/2} \qquad j \geqq 2, \tag{4.13}$$

$$\alpha_2 = (1 - \tfrac{1}{2})^{1/2},$$

$$\alpha_3 = (1 - \tfrac{1}{4})^{1/2},$$

$$\alpha_4 = (1 - \tfrac{1}{6})^{1/2}, \text{ etc.}$$

$$[\alpha_1 \cdot \alpha_2 \cdot \alpha_3 \cdots \alpha_n]^2 = \alpha^2 \frac{1 \cdot 3 \cdot 5 \cdots [2(n-1)-1]}{2 \cdot 4 \cdot 6 \cdots 2(n-1)} \tag{4.14}$$

$$= \alpha^2 \frac{(2n)!}{2^{2n}(n!)^2} \tag{4.15}$$

$$= \alpha^2 \left(\frac{1}{\pi n}\right)^{1/2} \qquad \text{when} \qquad n \gg 1. \tag{4.16}$$

The factors α_j for $j \geqq 2$ represent the reduction in timing deviations resulting from the reduction in bandwidth as resonant circuits are added in tandem. If resonant circuits with a narrow flat pass-band were used, the bandwidth of any number of resonant circuits in tandem would be the same as for a single resonant circuit. In this case $\alpha_2 = \alpha_3 = \alpha_n = 1$.

4.4 Deviations in Timing Wave

The timing wave derived from an "on-off" pulse train with the aid of a resonant circuit will in accordance with the expressions given in the previous sections contain three types of amplitude and timing deviations.

The first type is a fixed amplitude reduction by a factor a_0 and a fixed time deviation τ_0 given by

$$a_0 = \cos \psi, \tag{4.17}$$

and

$$\tau_0 = \frac{T}{2\pi} \psi, \tag{4.18}$$

where ψ is given by (4.2).

The second type is a random amplitude and time deviation resulting from the random amplitude component of the pulse train, which have rms values

$$a_r \cong \left(\frac{\pi}{2Q}\right)^{1/2} [1 - \psi^2/2]^{1/2} \frac{1}{\cos \psi}, \tag{4.19}$$

and

$$\underline{\delta}_r \cong \frac{T}{2\pi} \left(\frac{\pi}{4Q}\right)^{1/2} \frac{|\psi|}{\cos \psi}. \tag{4.20}$$

The third type is a random amplitude and time deviation resulting from random timing deviations $\bar{\tau}_p = \underline{\delta}$ in the pulse train. The amplitude variation can be disregarded and the rms time deviation is

$$\underline{\delta}_\delta = \underline{\alpha} \bar{\tau}_p, \qquad \underline{\alpha} = \left(\frac{\pi}{Q}\right)^{1/2}. \tag{4.21}$$

The total rms amplitude variation is accordingly given by (4.19). The total rms timing deviation obtained by combining (4.20) and (4.21) is

$$\bar{\tau}_r = \underline{\delta}_r^2 + \underline{\alpha}^2 \bar{\tau}_p^2)^{1/2}. \tag{4.22}$$

The expressions for $\underline{\delta}_r$ and $\bar{\tau}_r$ are the quantities appearing in (2.20) for $\underline{\Delta}_r$, the total rms timing deviation in regenerated pulses resulting from random amplitude and timing deviations in the timing wave.

V SELF-TIMED REPEATERS WITH PARTIAL RETIMING

5.0 General

As shown in the preceding section, timing for pulse regeneration can be derived from the pulse trains, with certain random phase and amplitude variations in the timing wave that can be reduced by increasing the loss constant Q of the resonant circuit. This method of "self-timing" can be combined with partial retiming, and the regeneration characteristics of this type of repeater will be discussed in the following sections.

For purposes of numerical illustration, the same type of pulse shape and timing wave will be assumed as in the previous numerical illustration in Section III. This pulse shape and timing wave closely approximates those in experimental Wrathall repeaters, in which timing is derived from the regenerated pulse train. In the following discussion timing from the received pulse train will also be considered.

5.1 Timing from Received Pulse Train

It will be assumed that the timing wave is derived from the received pulse train with the aid of a resonant circuit and that random timing deviations are absent. The response of the resonant circuit is then a sinusoidal wave as given by (4.1). From this wave it is possible to obtain a retiming wave of the form

$$R(t) = -\cos\psi\left[1 - \cos\left(2\pi\frac{t}{T} - \psi\right)\right]. \tag{5.1}$$

This can be accomplished by holding the peaks of the timing wave from the resonant circuit at zero potential with a diode. This is the form of retiming wave previously considered in Section III, in conjunction with a pulse shape given by (3.1).

As shown in Section 3.7, the tolerance to noise will vary with the phase shift ψ of the resonant circuit, in accordance with (3.21). If a reduction in the tolerance to noise of about 2 db is allowed, the maximum permissible phase shift would be about $\psi = 1$ radian (57.6°). On this basis the maximum permissible deviation $\Delta f_{\max}$ in the resonant frequency from the pulse repetition frequency f as obtained from (4.2) with $\psi = 1$ radian becomes

$$\frac{\Delta f_{\max}}{f} = \frac{\tan\psi}{2Q} = \frac{1.58}{2Q}. \tag{5.2}$$

For various values of Q in the range that can be realized by simple

resonant circuits, the permissible deviations are as follows:

Q	10	25	50	100	200
$\Delta f_{max}/f$	0.08	0.030	0.016	0.008	0.004

This assumes that there are no random timing deviations and that the tolerance to noise is reduced by not more than 2 db.

5.2 Timing from Regenerated Pulse Train

It will again be assumed that there are no random timing deviations. Without a phase shift in the resonant circuit, let the regenerated pulses be triggered at a time t_0. When there is a phase shift ψ', the pulses will be triggered at a time t_0'. The timing wave derived from the regenerated pulses will then have a time shift

$$\Delta = t_0' - t_0 + \frac{T}{2\pi}\psi'.$$

This time shift will cause pulses to be regenerated with a time shift $\beta'\Delta$, which must equal $t_0' - t_0$. Accordingly,

$$t_0' - t_0 = \beta'\left(t_0' - t_0 + \frac{T}{2\pi}\psi'\right),$$

and

$$t_0' - t_0 = \frac{T}{2\pi}\frac{\beta'\psi'}{1-\beta'}. \tag{5.3}$$

With timing from the received pulse train with a phase shift ψ in the resonant circuit, the following relation applies:

$$t_0' - t_0 = \frac{T}{2\pi}\beta\psi. \tag{5.4}$$

If $t_0' - t_0$ is to be the same in both cases, so that the timing wave and tolerance to noise is the same, the following relation must exist between the phase shifts in the resonant circuit:

$$\psi' = \psi(1-\beta')\frac{\beta}{\beta'}. \tag{5.5}$$

In this expression, β and β' are the factors shown in Fig. 4. It will be recognized from (5.5) that the smallest permissible phase shifts are obtained for large values of β'. From Fig. 4, it is seen that the largest

values of β are for phase shifts between 0 and $-60°$. For $\eta = \frac{3}{4}$, $\beta \cong 0.7$ and for $\eta = 1$, $\beta \cong 0.9$.

For $\eta = \frac{3}{4}$ and $\eta = 1$ the tolerable maximum phase shifts ψ' in the resonant circuit with timing from the regenerated pulse train, in relation to the maximum tolerable ψ with timing from the input, are

$$\psi' \cong 0.3\psi \quad \text{for} \quad \eta = \tfrac{3}{4}, \tag{5.6}$$

and

$$\psi' \cong 0.1\psi \quad \text{for} \quad \eta = 1.$$

Although greater phase shifts can be tolerated when ψ is positive, and β' is smaller than above, the requirements on the resonant circuit must be based on the worst condition that can be encountered, as above.

From (5.6) it follows that for $\eta = 1$ the requirements on the permissible phase shift in the resonant circuit are much more severe than for $\eta = \frac{3}{4}$. For this reason the latter value of η is decidedly preferable for the particular case in which the peak amplitudes of the pulse train and the timing waves are equal, as assumed here. A value $\eta = \frac{3}{4}$ is also desirable from the standpoints of avoiding intersymbol interference between adjacent pulses at the triggering instants, to permit the timing wave to be derived from the pulse train and to permit self-starting of the repeaters, as discussed later.

In accordance with (5.6) the maximum tolerable frequency deviation for $\eta = \frac{3}{4}$ will be less than with timing from the received pulse train by a factor of about 0.3. The maximum permissible frequency deviation for a phase shift of about one radian in the timing wave and 0.3 radian in the resonant circuit, will accordingly be about as follows:

Q	10	25	50	100	200
$\Delta f_{max}/f$	0.025	0.009	0.005	0.0025	0.0012

For a repeater with complete rather than partial retiming, the factor β would be unity, and timing from the regenerated pulse train would not be possible.

5.3 Random Timing Deviations

In combining random timing deviations from various sources at a particular repeater, it will be assumed that there is no correlation between the various deviations, so that they will combine on a root-sum-square basis.

In accordance with (2.21) the rms timing deviation at the output is then:

$$\underline{\Delta}^2 = (p_\tau^2 \bar{\tau}_p^2 + p_a^2 \underline{a}_p^2) + (r_\tau^2 \bar{\tau}_r^2 + r_a^2 \underline{a}_r^2), \tag{5.7}$$

where in accordance with (4.13) and (4.16)

$$\underline{a}_r = \left[\frac{\pi}{2Q}(1 - \psi^2/2)\right]^{1/2} \frac{1}{\cos \psi}, \tag{5.8}$$

$$\bar{\tau}_r = (\underline{\delta}_r^2 + \underline{\alpha}^2 \bar{\tau}_p^2)^{1/2}, \tag{5.9}$$

$$\underline{\alpha} = \left(\frac{\pi}{Q}\right)^{1/2}, \tag{5.10}$$

$$\underline{\delta}_r = \frac{T}{2\pi}\left(\frac{\pi}{4Q}\right)^{1/2} \frac{\psi}{\cos \psi}. \tag{5.11}$$

When (5.9) is inserted in (5.7)

$$\underline{\Delta}^2 = (p_\tau^2 + \underline{\alpha}^2 r_\tau^2)\bar{\tau}_p^2 + p_a^2 \underline{a}_p^2 + r_\tau^2 \underline{\delta}_r^2 + r_a^2 \underline{a}_r^2. \tag{5.12}$$

This expression gives the rms timing deviation at the output in terms of the rms deviation $\bar{\tau}_p$ at the input and the various repeater parameters.

With timing from the output, rather than the input as assumed above, $\bar{\tau}_p$ is replaced by $\underline{\Delta}$ in (5.9), and the following relation is obtained:

$$\underline{\Delta}^2(1 - \underline{\alpha}^2 r_\tau^2) = p_\tau^2 \bar{\tau}_p^2 + p_a^2 \underline{a}_p^2 + r_\tau^2 \underline{\delta}_r^2 + r_a^2 \underline{a}_r^2. \tag{5.13}$$

In the above expressions $p_\tau^2 \cong 0.15$, $r_\tau^2 \cong 0.4$ and $\underline{\alpha}^2 \cong 0.03$ ($Q = 100$). The term $\underline{\alpha}^2 r_\tau^2$ can thus be neglected in comparison with p_τ^2 in (5.12) and in comparison with 1 in (5.13).

The following expression is thus obtained with timing from either the input or the output:

$$\begin{aligned} \underline{\Delta}^2 &= (p_\tau^2 \bar{\tau}_p^2 + p_a^2 \underline{a}_p^2) + (r_\tau^2 \underline{\delta}_r^2 + r_a^2 \underline{a}_r^2) \\ &= \underline{\Delta}_p^2 + \underline{\Delta}_r^2. \end{aligned} \tag{5.14}$$

5.4 Magnitude of Random Timing Deviations

The first two terms of (5.14) represents the rms timing deviations in the regenerated pulses resulting from timing deviations and amplitude variations in the received pulses. The last two terms represent the timing deviations resulting from timing deviations and amplitude variations in the timing wave. The conversion factors p_τ, p_a, r_τ and r_a are discussed in Section II and representative values given in Figs. 5 and 6. The values of $\underline{a}_r$ and $\underline{\delta}_r$ are obtained from (5.8).

TABLE IV — RMS DEVIATIONS FROM TIMING WAVE DISTORTION FOR $Q = 100$

ψ	−60°	−30°	0°	30°	60°
$r_\tau \underline{\delta}_r/T$	0.011	0.005	0	0.006	0.015
$r_a \underline{a}_r/T$	0.006	0.007	0.009	0.014	0.024
$\underline{\Delta}_r/T$	0.0126	0.009	0.009	0.015	0.028
$\bar{\varphi}_r$	4.5°	3.2°	3.2°	5.4°	10°

In Table IV are given the values of the two last terms in (5.14), which represents the rms deviations $\underline{\Delta}_r$ owing to random deviations in the timing wave. The results are given for the particular case in which $Q = 100$, and for other values of Q are inversely proportional to $Q^{1/2}$. The table shows the deviations as a fraction of the interval T between pulses, and also as the corresponding rms phase deviation $\bar{\varphi}_r$.

In Table V are given the values of the first two terms in (5.14), which represents the rms deviation $\underline{\Delta}_p$ in the regenerated pulses resulting from random amplitude and timing deviation in the received pulses. In binary systems it is customary to limit the rms pulse distortion to $\underline{a}_p = \frac{1}{10}$, corresponding to $\frac{1}{10}$ the peak amplitude of the received pulses, or $\frac{1}{5}$ the triggering level (17 db signal-to-noise ratio). The corresponding rms phase deviation would be about $\frac{1}{10}$ radian, corresponding to an rms deviation $\bar{\tau}_p$ in the pulses of 0.016 the pulse spacing, or $\bar{\tau}_p/T \cong 0.016$. The total rms timing deviation obtained from (5.14) and the corresponding rms phase deviation are given in Table VI.

TABLE V — RMS DEVIATIONS RESULTING FROM PULSE DISTORTION

ψ	−60°	−30°	0°	30°	60°
$p_a \underline{a}_p/T$	0.019	0.018	0.019	0.022	0.032
$p_\tau \bar{\tau}_p/T$	0.010	0.007	0.005	0.005	0.008
$\underline{\Delta}_p/T$	0.021	0.020	0.020	0.023	0.033
$\bar{\varphi}_p$	7.5°	7.2°	7.2°	8.2°	12°

TABLE VI — TOTAL RMS DEVIATIONS FROM TIMING WAVE AND PULSE DISTORTION

ψ	−60°	−30°	0°	30°	60°
$\underline{\Delta}/T$	0.025	0.022	0.022	0.028	0.043
$\bar{\varphi}$	9°	8°	8°	10°	16°

The probability that random phase deviations will exceed the above rms values by a factor of more than 4 is small enough to be ignored. On this basis the sum of the fixed and random deviations would be limited to about 70°, if the fixed phase shift ψ is less than $\pm 30°$. With this requirement on the fixed phase shift for satisfactory performance, the values of Δf_{max} would be about half as great as previously given in Sections 5.1 and 5.2, for a single repeater as considered here.

VI REPEATER CHAINS

6.0 General

In the previous section, a single self-timed repeater was considered, from the standpoint of fixed and random timing deviations, as determined by various repeater design parameters. In a repeater chain there will be some cumulation of random timing deviations as the number of repeaters in tandem is increased, and a resultant reduction in the tolerance to noise of repeaters toward the end of the chain. Exact evaluation of such cumulation is rendered difficult by the circumstance that timing deviations from various sources may not follow the same law of combination along the repeater chain. In the following, expressions are given based both on root-sum-square and direct addition of random timing deviations, which can be regarded as lower and upper limits.

6.1 Combination of Random Timing Deviations

To determine the rms value of random timing deviations at the end of a repeater chain, it is necessary to combine random deviations from various repeaters. Random deviations from various sources at a repeater do not necessarily follow the same law of cumulation along a repeater chain. Since there is no correlation between timing deviations caused by noise in various repeater sections, these can be combined on a root-sum-square basis. This, however, may not be appropriate as regards the combination of timing deviations resulting from imperfections in the timing wave. Thus, with perfect tuning of all resonant circuits, the timing waves at various repeaters would have virtually identical amplitude variations, but no phase deviations. While in this case there would be complete correlation between the timing wave variations at the repeaters, it does not follow that the resultant timing deviations should be combined directly rather than on a root-sum-square basis along the repeater chain. The timing deviations at the end of a chain of N repeaters resulting from amplitude variations in the timing wave of the first repeater will be modified by N intermediate resonant circuits. Those

resulting from amplitude variations at subsequent repeaters will be modified by N-1, N-2 etc. intermediate resonant circuits. The situation is similar to that of applying identical noise waves at the input of each of N resonant circuits in tandem. At the output the N noise waves will have different shapes owing to restriction of the band and increasing phase distortion as the number of resonant circuits in tandem increases. For this reason combination on a root-sum-square basis appears justified also in this case, particularly with various degrees of mistuning of the resonant circuits, so that the amplitude variations in the timing waves will differ in phase among repeaters.

6.2 Propagation of Timing Deviations

To determine the cumulation of timing deviations along a repeater chain, it is convenient to first consider a single repeater as a source of timing deviations, and to determine the propagation of these timing deviations along a repeater chain. In the following, γ_n will designate the rms propagation factor for n repeaters in tandem; i.e., the factor by which the rms timing deviations at the end of a chain of n repeaters is smaller than at the first repeater, with timing deviations originating at the first repeater only.

Let the rms timing deviation at the output of the first repeater as given by (5.14) for convenience be taken as unity. At the output of the second repeater the squared rms timing deviation is then reduced by the factor

$$\gamma_2^2 = p_\tau^2 + \underline{\alpha}_1^2 r_\tau^2, \qquad \underline{\alpha}_1 = \underline{\alpha}. \tag{6.1}$$

As indicated symbolically in Fig. 8, the first term represents the reduction owing to partial retiming. The second term is the additional devia-

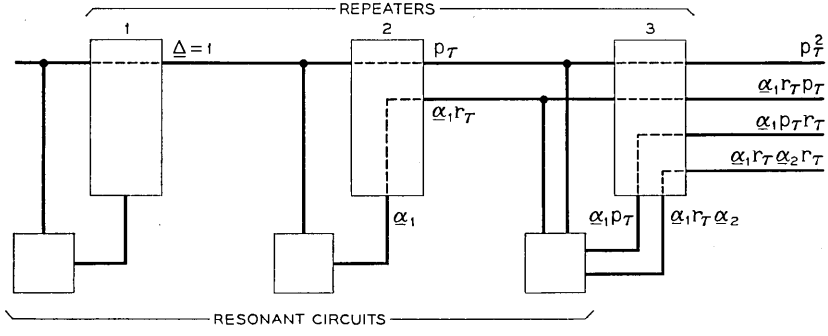

Fig. 8 — Propagation of random timing deviations along repeater chain.

tion resulting from the effect on the timing wave of unit rms deviation in the received pulse train at the second repeater.

At the output of the third repeater, the squared rms deviation is smaller than at the output of the first repeater by the factor

$$\gamma_3^2 = (p_\tau^2 + \alpha_1^2 r_\tau^2) p_\tau^2 + p_\tau^2 \alpha_1^2 r_\tau^2 + \alpha_1^2 r_\tau^2 \alpha_2^2 r_\tau^2 \tag{6.2}$$

$$= p_\tau^4 + 2\alpha_1^2 p_\tau^2 r_\tau^2 + \alpha_1^2 \alpha_2^2 r_\tau^4. \tag{6.3}$$

As indicated in Fig. 8, the first term in (6.2) represents the reduction owing to partial retiming of the received pulse train at the third repeater. The second term, $(p_\tau \alpha_1 r_\tau)^2$, is the additional rms deviation resulting from the effect on the timing wave at the third repeater of an rms deviation p_τ in the received pulse train. The third term $(\alpha_1 r_\tau \cdot \alpha_2 r_\tau)^2$ is the additional deviation caused by the effect on the timing wave of an rms deviation $\alpha_1 r_\tau$ in the received pulse train. The factor $\alpha_2 r_\tau$ represents the modification in the rms deviation $\alpha_1 r_\tau$ by a second resonant circuit, with α_2 defined as in Section 4.3.

At the output of the fourth repeater, the rms timing deviation is reduced by the following factor, obtained in the same manner:

$$\gamma_4^2 = p_\tau^6 + 3\alpha_1^2 p_\tau^4 r_\tau^2 + 3\alpha_1^2 \alpha_2^2 p_\tau^2 r_\tau^4 + \alpha_1^2 \alpha_2^2 \alpha_3^2 r_\tau^6. \tag{6.4}$$

At the output of repeater n, the squared rms timing deviation is smaller than at the output of the first repeater by the propagation factor

$$\begin{aligned} \gamma_n^2 = {} & p_\tau^{2(n-1)} + \frac{(n-1)}{1!} p_\tau^{2(n-2)} r_\tau^2 \alpha_1^2 \\ & + \frac{(n-1)(n-2)}{2!} p_\tau^{2(n-3)} r_\tau^4 \alpha_1^2 \alpha_2^2 \\ & + \frac{(n-1)(n-2)(n-3)}{3!} p_\tau^{2(n-4)} r_\tau^1 \alpha_1^2 \alpha_2^2 \alpha_3^2 \\ & + \cdots + r_\tau^{2(n-1)} \alpha_1^2 \alpha_2^2 \alpha_3^2 \cdots \alpha_{n-1}^2, \end{aligned} \tag{6.5}$$

where p_τ and r_τ are defined as in Section 2.2, and α_1, $\alpha_2 \cdots \alpha_n$ as in Section 4.3.

In the above formulation the rms deviation at the output of the first repeater was assumed given by (5.14), which is an approximation of (5.12) in which the term $\alpha^2 r_\tau^2 \bar{\tau}_p^2$ was neglected. This term will have a different propagation factor ρ_n, the expression for which differs from that for γ_n as given by (6.5) in that the subscripts of the factors α_j are raised by one unit. Thus,

$$\rho_n^2 = p_r^{2(n-1)} + \frac{(n-1)}{1!} p_r^{2(n-2)} r_r^2 \underline{\alpha}_2^2 + \cdots + r_r^{2(n-1)} \underline{\alpha}_2^2 \underline{\alpha}_3^2 \cdots \underline{\alpha}_n^2. \tag{6.6}$$

The rms deviation at the output of repeater n thus becomes

$$\underline{\Delta}_n^2 = (\underline{\Delta}_p^2 + \underline{\Delta}_r^2)\gamma_n^2 + \underline{\alpha}^2 r_r^2 \bar{\tau}_p^2 \rho_n^2. \tag{6.7}$$

In the case of repeaters with partial retiming the last term in (6.7) can be neglected, in which case the cumulation of timing deviation will be virtually the same when the timing wave is derived from the regenerated as when it is derived from the received pulse train.

The above expressions apply for resonant circuits consisting of a coil and capacitor which have a gradual cut-off. If resonant circuits with a flat pass-band and sharp cut-offs were used, $\alpha_2 = \alpha_3 = \alpha_n$ and (6.5) can be simplified to

$$\gamma_n^2 = (1 - \underline{\alpha}_1^2) p_r^{2(n-1)} + \underline{\alpha}_1^2 (p_r^2 + r_r^2)^{(n-1)}. \tag{6.8}$$

6.3 Cumulation of Timing Deviations

The cumulation of random timing deviations from various repeaters in a chain can be determined from the propagation constant given above for any prescribed law of combination of timing deviations from various repeaters. When equal rms deviations are contributed by each of N repeaters, and they are combined on a root-sum-square basis, the rms deviation at the end of a repeater chain is greater than for a single repeater by the cumulation factor

$$C = \left(\sum_{n=1}^{N} \gamma_n^2 \right)^{1/2}. \tag{6.9}$$

An upper limit to C is obtained by taking $\underline{\alpha}_2 = \underline{\alpha}_3 = \underline{\alpha}_n = 1$ in (6.5) in which case γ_n^2 is given by (6.8); (6.9) then becomes for $N = \infty$

$$C = \left[(1 - \underline{\alpha}_1^2) \frac{1}{1 - p_r^2} + \underline{\alpha}_1^2 \frac{1}{1 - p_r^2 - r_r^2} \right]^{1/2} \tag{6.10}$$

$$\cong \left(\frac{1}{1 - p_r^2} \right)^{1/2}, \tag{6.11}$$

where the terms in $\underline{\alpha}_1^2$ have been neglected in (6.11), since $\underline{\alpha}_1^2 = \underline{\alpha}^2 \ll 1$, about 0.03 for $Q = 100$.

From Fig. 5 it will be seen that when $\psi < \pm 60°$, $p_r < 0.6$. Hence $C < 1.25$. Cumulation of random timing deviations can thus for practical

purposes be disregarded, with root-sum-square combination as assumed above. The value of C obtained from (6.11) will differ from that obtained from (6.9) when γ_n is given by (6.5), by a small fraction of one per cent.

Although root-sum-square combination appears justified for reasons given before, it is of interest to determine an upper limit to the cumulation based on direct addition of random timing deviations. The maximum cumulation factor thus obtained is

$$C_{\max} = \sum_{n=1}^{N} \gamma_n . \tag{6.12}$$

Employing (6.8) for γ_n and neglecting the terms in α_1^2, the upper limit to the cumulation factor for $N = \infty$ becomes

$$C_{\max} = \frac{1}{1 - p_r}. \tag{6.13}$$

With $p_r < 0.6$ for $\psi < \pm 60°$, $C_{\max} < 2.5$.

If the above maximum cumulation factor is applied to random timing deviations resulting from amplitude variations in the timing wave, as given in Table IV of Section 5.4, the resultant rms phase deviation at the end of a long repeater chain could be as great as 25°, rather than 10° for a single repeater, when $\psi = 60°$ and $Q = 100$. To attain satisfactory performance it would in this case be necessary to limit the maximum fixed phase shift to substantially less than $\pm 60°$, which would entail greater frequency precision than indicated in Sections 5.1 and 5.2.

If $\psi < \pm 15°$, $p_r < 0.40$ and $C_{\max} < 1.7$. In this case the rms phase deviation as given in Table I for a single repeater is $\bar{\varphi} \cong 4°$, and the rms phase deviation in a long repeater chain would be less than 7°. In a long repeater chain the rms phase deviation resulting from pulse distortion would be greater than given in Table II by an rms cumulation factor $C = 1.08$ for $p_r = 0.4$, and would thus be about 8° when $\psi < \pm 15°$. The total rms phase deviation would thus be about $(7^2 + 8^2)^{1/2} \cong 11°$. Random phase deviations exceeding 4 times the latter value, or about 45°, would be rather unlikely. The sum of the fixed and random phase deviations would thus be limited to about 60°, so that satisfactory performance would be expected when the fixed phase deviation is limited to about $\pm 15°$.

With the approximations for γ_n employed above, the rms cumulation factor for a chain of N repeaters as obtained from (6.9) is less than for $N = \infty$ by the factor $(1 - p_r^{2N})^{1/2} \cong 0.99$ for $p_r = 0.5$ and $N = 3$. The maximum cumulation factor obtained from (6.12) is less than for $N = \infty$ by the factor $1 - p_r^N \cong 0.99$ for $N = 6$. Thus, cumulation of random

timing deviations is virtually completed in a chain of 3 to 6 repeaters, so that for experimental determinations of the degree of cumulation it suffices to operate a few repeaters in tandem.

6.4 Repeaters with Complete Retiming

In the particular case of complete retiming, $p_\tau = 0$ and $r_\tau = 1$ in (6.5) and (6.6) so that

$$\gamma_n = \underline{\alpha}_1\underline{\alpha}_2\underline{\alpha}_3 \cdots \underline{\alpha}_{n-1}\,, \tag{6.14}$$

$$\rho_n = \underline{\alpha}_2\underline{\alpha}_3 \cdots \underline{\alpha}_n\,. \tag{6.15}$$

For $n \gg 1$, approximation (4.16) can be employed, in which case

$$\gamma_n = \underline{\alpha}\left(\frac{1}{\pi n}\right)^{1/4}, \qquad \rho_n = \left(\frac{1}{\pi n}\right)^{1/4}. \tag{6.16}$$

In this case (5.14) simplifies to

$$\underline{\Delta}_p^{\,2} + \underline{\Delta}_r^{\,2} = \underline{\delta}_r^{\,2}, \tag{6.17}$$

since $p_a = 0$, $r_a = 0$, $p_\tau = 0$ and $r_\tau = 1$.

Hence (6.7) becomes

$$\underline{\Delta}_n^{\,2} = \underline{\delta}_r^{\,2}\gamma_n^{\,2} + \bar{\tau}_p^{\,2}\underline{\alpha}^2\rho_n^{\,2}. \tag{6.18}$$

With approximations (6.16),

$$\underline{\Delta}_n^{\,2} = (\underline{\delta}_r^{\,2} + \bar{\tau}_p^{\,2})\underline{\alpha}^2\left(\frac{1}{\pi n}\right)^{1/2}. \tag{6.19}$$

At the output of the first repeater,

$$\underline{\Delta}_1^{\,2} = \underline{\delta}_r^{\,2} + \underline{\alpha}^2\bar{\tau}_p^{\,2}. \tag{6.20}$$

For $n \gg 1$ the squared propagation factor is accordingly

$$\underline{\Delta}_n^{\,2}/\underline{\Delta}_1^{\,2} = \underline{\alpha}^2\,\frac{\underline{\delta}_r^{\,2} + \bar{\tau}_p^{\,2}}{\underline{\delta}_r^{\,2} + \underline{\alpha}^2\bar{\tau}_p^{\,2}}\left(\frac{1}{\pi n}\right)^{1/2}. \tag{6.21}$$

The squared rms cumulation factor for $N \gg 2$ repeaters becomes

$$\begin{aligned} C^2 &\cong 1 + \underline{\alpha}^2\,\frac{\underline{\delta}_r^{\,2} + \bar{\tau}_p^{\,2}}{\underline{\delta}_r^{\,2} + \underline{\alpha}^2\bar{\tau}_p^{\,2}}\int_2^N \left(\frac{1}{\pi n}\right)^{1/2} dn \\ &\cong 1 + \underline{\alpha}^2\,\frac{\underline{\delta}_r^{\,2} + \bar{\tau}_p^{\,2}}{\underline{\delta}_r^{\,2} + \underline{\alpha}^2\bar{\tau}_p^{\,2}}\left[\left(\frac{4N}{\pi}\right)^{1/2} - \left(\frac{8}{\pi}\right)^{1/2}\right]. \end{aligned} \tag{6.22}$$

In the particular case of perfect tuning of all resonant circuits $\underline{\delta}_r = 0$ and

$$(\Delta_n/\Delta_1)^2 \cong \left(\frac{1}{\pi n}\right)^{1/2}, \tag{6.23}$$

$$C^2 \cong 1 + \left(\frac{4N}{\pi}\right)^{1/2} - \left(\frac{8}{\pi}\right)^{1/2},$$

$$C \cong \left(\frac{4N}{\pi}\right)^{1/4}. \tag{6.24}$$

The last expression gives the factor by which the rms timing deviation at the output of repeater N is greater than at the output of the first repeater. The rms deviation at the output of the first repeater is greater than at the input by the factor α. The rms deviation at the output of repeater N is thus greater than at the input of the first repeater by the factor,

$$C_1 = \alpha \left(\frac{4N}{\pi}\right)^{1/4}. \tag{6.25}$$

For this particular case ($\delta_r = 0$) expressions equivalent to those above have been derived in unpublished work by H. E. Rowe of Bell Telephone Laboratories.

In accordance with (6.22) and (6.24) the cumulation of random timing deviations increases indefinitely with N when retiming is complete. The cumulation factor as given by (6.24) is in fact the same as would be obtained if a timing wave were transmitted on a separate pair, with a resonant circuit at each repeater to limit noise and with amplification of the timing wave at each repeater to obtain the same amplitude of the timing wave as when it is derived from the pulse train. With partial retiming cumulation is limited, for the reason that there is partial regeneration of both the pulse train and the timing wave.

Although with complete retiming the cumulation factor increases indefinitely with N, this is of but little practical significance, because of the slow rate of cumulation. At the output of a chain of N repeaters an rms deviation approximately equal to that at the input of the first repeater could be tolerated, in which case $C_1 \cong 1$. On this basis the permissible number of repeaters would be

$$N \cong \frac{\pi}{4}\frac{1}{\alpha^4} = \frac{\pi}{4}\left(\frac{Q}{\pi}\right)^2, \tag{6.26}$$

$$\cong 800 \quad \text{when} \quad Q = 100.$$

This assumes exact tuning of all resonant circuits. With mistuning of the resonant circuits, the permissible number of repeaters in tandem for a specified rms deviation at the output of the final repeater can be

determined with the aid of the cumulation factor given by (6.22). For example, if the rms deviation at the output of repeater N is assumed the same as at the input of the first repeater, the permissible number of repeaters in tandem is less than given by (6.26) by the factor $[(1 - m^2)/(1 + m^2)]^2$, $m = \delta_r/\bar{\tau}_p$. When the fixed phase shift is 30°, $m \cong 0.5$ and $N \cong 300$.

6.5 Self-Starting of Self-Timed Repeaters

With self-timing it is necessary that repeaters be self-starting if the timing wave should be absent for any reason. If each repeater is self-starting, this will also be the case for a repeater chain, since starting will be progressive along the chain. Initially, before the timing wave has reached the appropriate amplitude at all repeaters, there will be a high rate of digital errors.

With timing from the received pulse train, the resonant circuit will be excited by every pulse and the timing wave will reach its normal amplitude in about $n \cong Q$ pulses. With timing from the regenerated pulse

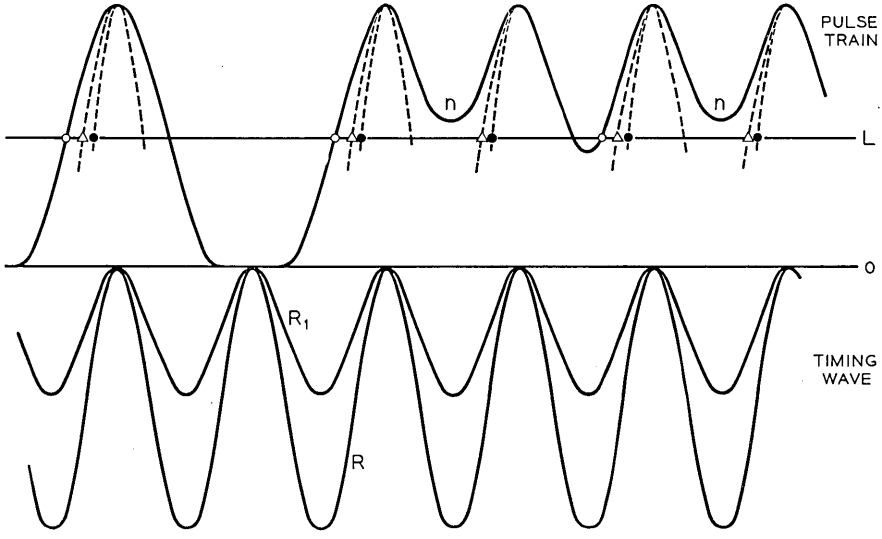

Fig. 9 — Progression of repeater starting in absence of timing wave when timing is derived from regenerated pulse train.
○ Triggering points with timing wave absent. Noise prevents triggering at certain points, n. Timing wave reaches fraction of normal value, R_1.
△ Triggering points with timing wave R_1. Timing wave increases to normal amplitude R.
● Triggering points with normal timing wave.

train the resonant circuit will not be excited by every pulse, unless the shape of the received pulses is such that there are virtually no overlaps between pulses so that the triggering level will be penetrated by each pulse.

With a pulse shape as assumed in the previous analysis, the amplitude of a pulse train midway between pulses is half the peak amplitude of the pulses, as indicated in Fig. 9. In the presence of noise, triggering will in this case occur on the average for every second pulse, as indicated in the above figure. If it is assumed that the resonant circuit has the maximum permissible phase shift of about 20° allowed with timing from the output, the amplitude of the timing wave with excitation from every pulse will be virtually equal to the peak pulse amplitude. With excitation from half the pulses, the amplitude of the timing wave will rapidly reach half the peak amplitude of the pulses. When this initial timing wave is combined with the pulse train, triggering will occur for virtually all pulses, as indicated in Fig. 9. It will thus reach its normal value. If the phase shift is greater than 20° as assumed above, say 60°, the initial amplitude of the timing wave will be $\frac{1}{4}$ the peak pulse amplitude. Combination of this initial timing wave with the pulse train will increase the number of pulses exciting the resonant circuit, which in turn increases the amplitude of the timing waves, etc.

Self-starting with a pulse shape as assumed in this analysis is thus insured.

VII SUMMARY

In self-timing regenerative repeaters as considered here, a timing wave is derived from either the received or regenerated pulse train with the aid of a simple resonant circuit tuned to the pulse repetition frequency. This timing wave is combined linearly with received pulse trains as indicated in Fig. 1, and pulses are regenerated when the combined wave penetrates a certain triggering level.

It is concluded that if these timing principles are implemented by appropriate repeater instrumentation, a performance can be realized that approaches that of ideal regenerative repeaters. To this end it is necessary to meet certain requirements with regard to the loss constant Q of the resonant circuit, its frequency precision, the shape of received pulses and the amplitude of the timing wave in relation to that of received pulses.

Equalization of each repeater section should preferably be such that the received pulses have a shape and duration in relation to the pulse

interval as indicated in Fig. 3, and the peak amplitude of the timing wave should be about equal to that of the received pulses. Under these conditions the pulse repetition frequency will be present in the received pulse train in sufficient amplitude to permit derivation of the timing wave from the received pulse train, and to permit rapid self-starting in the absence of a timing wave if it is derived from the regenerated pulse train.

A loss constant of the resonant circuit $Q \cong 100$ appears desirable. This value is sufficiently low to be readily realized with simple resonant circuits consisting of a coil and capacitor in series or parallel, without unduly severe requirements on its frequency precision. It is also adequately high from the standpoint of avoiding excessive random timing deviations in regenerated pulses from amplitude and phase deviations in the timing wave.

The tolerable deviation in the resonant frequency from the pulse repetition frequency with $Q = 100$ is about 0.2 per cent when the timing wave is derived from the received pulse train, and about 0.06 per cent when it is derived from the regenerated pulse train. These frequency precisions correspond to a maximum fixed phase shift of 15° in the timing wave, and allow for the possibility that random timing deviations resulting from amplitude variations in the timing wave may cumulate directly along a repeater chain, rather than on a root-sum-square basis. With root-sum-square cumulation of timing deviations from all sources, the frequency deviations could be about twice as great.

When the above requirements are met the reduction in the tolerance to noise owing to timing deviations in a repeater chain is limited to about 2 db. If the requirements on frequency precision of the resonant circuit are met, substantial degradation or improvement in performance would not be expected as a result of moderate changes in the other design parameters.

VIII ACKNOWLEDGMENTS

In this presentation the writer had the benefit of unpublished work, referred to previously, by W. R. Bennett and J. R. Pierce on the derivation of a timing wave from a pulse train with the aid of a resonant circuit, and by H. E. Rowe on the cumulation of timing deviations in a chain of repeaters with complete retiming. Bennett, Pierce and Rowe are at Bell Telephone Laboratories. He is also indebted to H. E. Rowe for a critical review that resulted in several improvements in the analysis.

Appendix

IX RESONANT CIRCUIT RESPONSE TO RANDOM BINARY PULSE TRAINS

1 General

In the following analysis of the response of a resonant circuit to a binary "on-off" pulse train, the pulses are assumed of sufficiently short duration to be regarded as impulses. This is a legitimate approximation when the duration does not exceed about half the interval between pulses.

The pulse train is regarded as made up of three components, as indicated in Fig. 10. The first is a systematic component consisting of pulses of amplitude $\frac{1}{2}$. This component gives rise to a steady state response at the fundamental frequency of the pulse sequence. The second component consists of pulses of amplitude $\pm\frac{1}{2}$, with random $\pm$ polarity. This component gives rise to a random component in the resonant cir-

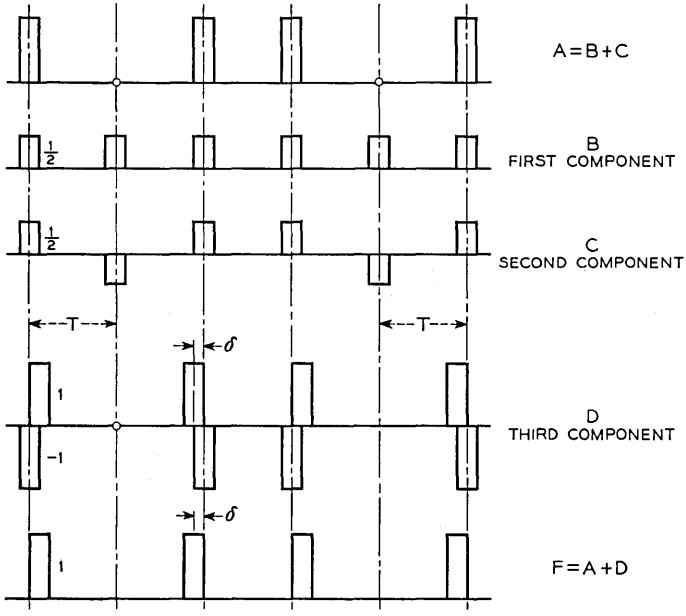

Fig. 10 — Components of random binary on-off pulse train. A. — Transmitted "on-off" pulses. B. — Steady state pulse train of fundamental frequency $f = 1/T$. C. — Random pulse train with zero mean value. D. — Random pulse train with displacements $\pm\delta$. F. — "On-off" pulse train with displacements $\pm\delta$ from average pulse interval T.

cuit response; i.e., a fluctuation about the steady state value derived from the first component.

The third component consists of a train of dipulses. Each dipulse consists of a pair of pulses of amplitude 1 and -1, displaced by an interval $\pm\delta$. The response of the resonant circuit to this component gives the effect of random displacements $\pm\delta$ in the original "on-off" pulse train.

2 Impedance of Resonant Circuit

The impedance of a resonant circuit consisting of R, L and C in parallel is

$$Z(i\omega) = \bar{Z}(i\omega)e^{-i\psi}, \tag{1}$$

$$\bar{Z}(i\omega) = \frac{R}{[1 + Q^2(\omega/\omega_0 - \omega_0/\omega)^2]^{1/2}} = R\cos\psi, \tag{2}$$

$$\tan\psi = Q(\omega/\omega_0 - \omega_0/\omega), \tag{3}$$

where

$$Q = \omega_0 RC = \text{Loss constant}, \tag{4}$$

and

$$\omega_0 = (1/LC)^{1/2} = \text{Resonant frequency}. \tag{5}$$

The above expressions also apply for the admittance of a resonant circuit consisting of R, L and C in series, except that in this case $Q = \omega_0 L/R$.

3 Impulse Response of Resonant Circuit

When a rectangular pulse of unit amplitude and sufficiently short duration δ_0 is applied to a resonant circuit, the impulse response for $Q \gg 1$ is of the form

$$P(t) = P(0)\cos\omega_0 t e^{-\omega_0 t/2Q}, \tag{6}$$

where

$$P(0) = \omega_0\delta_0 R/Q. \tag{7}$$

$P(t)$ designates voltage in response to an impulse current in the case of a parallel resonant circuit, or the current in response to an impulse voltage in the case of a series resonant circuit.

4 Response to Steady State Impulse Train

Let a long sequence of impulses of amplitude $\frac{1}{2}$ and the same polarity impinge on a resonant circuit at uniform intervals T. The response after N impulses is then

$$A_s(t) = \frac{1}{2}\sum_{n=0}^{N} P(t - nT) \tag{8}$$

$$= \frac{P(0)}{2}\sum_{n=0}^{N} \cos \omega_0(t - nT)e^{-\omega_0(t-nT)/2Q}. \tag{9}$$

The subscript s indicates a systematic component.

The above series is conveniently summed by taking the real part of the series

$$A_s'(t) = \frac{P(0)}{2}\sum_{n=0}^{N} e^{i\omega_0(t-nT)}e^{-\omega_0(t-nT)/2Q}. \tag{10}$$

With $t = NT + t_0$, $0 < t_0 < T$:

$$A_s'(t) = \frac{P(0)}{2}e^{i\omega_0 t_0}e^{-\omega_0 t_0/2Q}\sum_{n=0}^{N} e^{i\omega_0 t(N-n)}e^{-\omega_0 t(N-n)/2Q}. \tag{11}$$

When $N \to \infty$, the steady state responses becomes

$$A_s'(t) = \frac{P(0)}{2}e^{i\omega_0 t_0}e^{-\omega_0 t_0/2Q}\frac{1}{1 - e^{i\omega_0 t}e^{-\omega_0 t/2Q}}. \tag{12}$$

The interval between pulses can be written

$$T = 2\pi/\omega, \tag{13}$$

where ω is the fundamental frequency of the impulse train, or the pulse repetition frequency.

Let

$$\omega_0 = \omega - \Delta\omega,$$

so that

$$\omega_0 = \frac{2\pi}{T}(1 - \Delta\omega/\omega). \tag{14}$$

The following approximations then apply:

$$e^{+i\omega_0 T} = e^{2\pi i}e^{-2\pi i\Delta\omega/\omega} = e^{-2\pi i\Delta\omega/\omega},$$
$$\cong 1 - 2\pi i\Delta\omega/\omega; \tag{15}$$

$$e^{-\omega_0 T/2Q} = e^{-\pi/Q}e^{+(\pi/Q)\Delta\omega/\omega},$$
$$\cong 1 - \pi/Q \text{ when } \pi/Q \ll 1. \tag{16}$$

With these approximations

$$1 - e^{i\omega_0 T} e^{-\omega_0 T/2Q} \cong \frac{\pi}{Q}[1 + i\psi], \tag{17}$$

where

$$\psi = \frac{2\Delta\omega}{\omega} Q, \tag{18}$$

will be recognized as the phase shift of the resonant current at the frequency ω, as obtained from (3) with $\omega = \omega_0 + \Delta\omega$.

A further approximation that can be introduced in (12) is

$$\begin{aligned} e^{i\omega_0 t_0} e^{-\omega_0 t_0/2Q} &= e^{i\omega t_0} e^{-i\Delta\omega t_0} e^{-\omega_0 t_0/2Q}, \\ &\cong e^{i\omega t_0}, \end{aligned} \tag{19}$$

since $t_0 < T$ and $\Delta\omega t_0$ and $\omega_0 t_0/2Q \ll 1$.

With the above approximations (12) becomes

$$A_s' = \frac{P(0)}{2} \frac{Q}{\pi} e^{i[\omega t_0 - \psi]} \cos\psi. \tag{20}$$

The real part of this expression is

$$A_s = \frac{P(0)}{2} \frac{Q}{\pi} \cos(\omega t_0 - \psi) \cos\psi, \tag{21}$$

which is the response to the steady state component of the pulse train.

5 Response to Random Component of Impulse Train

Let a sequence of impulses of amplitude $\frac{1}{2}$ and randomly positive and negative polarity impinge on the resonant circuit at intervals T. The response is then,

$$A_r(t) = \frac{P(0)}{2} \sum_{n=0}^{N} \pm \cos\omega_0(t - nT) e^{-\omega_0(t-nT)/2Q}. \tag{22}$$

This expression for the random component differs from (9) for the systematic component in that the impulses have random $\pm$ polarity. If all signs are chosen the same, the values of $A_r(t)$ will be either $-A_s(t)$ or $+A_s(t)$. The resultant response of the resonant circuit, i.e. $A_s(t) + A_r(t)$, can thus vary between the limit 0 and $2A_s(t)$. $A_r(t)$ represents a random fluctuation about $A_s(t)$ as a mean value. In the following the rms value of this fluctuation is evaluated.

In order to determine the components of $A_r(t)$ in phase and at quadra-

ture with the steady state response as given by (21), it is convenient to write

$$\omega_0 = \omega - \Delta\omega,$$

$$\begin{aligned}\cos \omega_0(t - nT) &= \cos [\omega(t - nT) - \psi + \psi - \Delta\omega(t - nT)] \\ &= \cos [\omega(t - nT) - \psi] \cos [\psi - \Delta\omega(t - nT)] \\ &\quad - \sin [\omega(t - nT) - \psi] \sin [\psi - \Delta\omega(t - nT)].\end{aligned} \tag{23}$$

With $t = NT + t_0$, and $\omega T = \pi$, (22) can be written:

$$\begin{aligned}A_r(t) &= \cos (\omega t_0 - \psi) \sum_{n=0}^{N} \pm \cos [\psi_1 - \Delta\omega T(N - n)] e^{-\omega_0(T)(N-n)/2Q} \\ &\quad - \sin (\omega t_0 - \psi) \sum_{n=0}^{N} \pm \sin [\psi_1 - \Delta\omega T(N - n)] e^{-\omega_0(T)(N-n)/2Q}\end{aligned} \tag{24}$$

where $\psi_1 = \psi - \Delta\omega t_0 = \psi\left(1 - \dfrac{\omega t_0}{2Q}\right) \cong \psi$, since $\omega t_0/2Q \leq \pi/2Q \ll 1$.

With equal probabilities of a plus or a minus sign in the summations, the rms value of the in-phase component becomes

$$\begin{aligned}\underline{A_r}' &= \left[\sum_{n=0}^{N} \cos^2 [\psi - \Delta\omega T(N - n)] e^{-\omega_0 T(N-n)/Q}\right]^{1/2} \\ &= \left[\sum_{n=0}^{N} \frac{1}{2} (1 + \cos 2[\psi - \Delta\omega T(N - n)] e^{-\omega_0 T(N-n)/Q})\right]^{1/2}.\end{aligned} \tag{25}$$

The rms value of the quadrature component becomes

$$\begin{aligned}\underline{A_r}'' &= \left[\sum_{n=0}^{N} \sin^2 [\psi - \Delta\omega T(N - n)] e^{-\omega_0 T(N-n)/Q}\right]^{1/2} \\ &= \left[\sum_{n=0}^{N} \frac{1}{2} (1 - \cos 2[\psi - \Delta\omega T(N - n)] e^{-\omega_0 T(N-n)/Q})\right]^{1/2}.\end{aligned} \tag{26}$$

These expressions can be transformed into sums of geometric series by writing

$$\cos x = \tfrac{1}{2}(e^{ix} + e^{-ix}), \qquad x = 2[\psi - \Delta\omega T(N - n)].$$

Evaluation of (25) and (26) by this method gives

$$\underline{A_r}' = \frac{P(0)}{2} \frac{1}{2^{1/2}} \left[\frac{1}{1 - e^{-\omega_0 T/Q}} + \frac{N}{D}\right]^{1/2}, \tag{27}$$

$$\underline{A_r}'' = \frac{P(0)}{(2)} \frac{1}{2^{1/2}} \left[\frac{1}{1 - e^{-\omega_0 T/Q}} - \frac{N}{D}\right]^{1/2}, \tag{28}$$

where

$$N = \cos 2\psi(1 - \cos 2\Delta\omega T e^{-\omega_0 T/2Q}) + \sin 2\psi \sin 2\Delta\omega T e^{-\omega_0 T/2Q}, \quad (29)$$

$$D = 1 + e^{-2\omega_0 T/Q} - 2e^{-\omega_0 T/Q} \cos 2\Delta\omega T. \quad (30)$$

With the same approximations as used previously in connection with (12) and with

$$\cos 2\Delta\omega T \cong 1 - 2(\Delta\omega T)^2, \qquad \sin 2\Delta\omega T \cong 2\Delta\omega T,$$

$$N \cong \frac{2\pi}{Q}, \quad (31)$$

$$D \cong \left(\frac{2\pi}{Q}\right)^2 [1 + \psi^2], \quad (32)$$

$$1 - e^{-\omega_0 T/Q} = 2\pi/Q. \quad (33)$$

With these approximations in (27) and (28),

$$\underline{A_r}' \cong \frac{P(0)}{2}\left(\frac{Q}{2\pi}\right)^{1/2} [1 - \psi^2/2]^{1/2}, \quad (34)$$

$$\underline{A_r}'' \cong \frac{P(0)}{2}\left(\frac{Q}{2\pi}\right)^{1/2} \frac{|\psi|}{2^{1/2}}, \quad (35)$$

which apply when ψ is small and $(2\pi/Q) \ll 1$.

6. Response to Random Dipulse Train

Each dipulse is assumed to consist of two impulses of unit amplitude and opposite polarity, displaced by an interval δ, which in general will be a function of the pulse position; i.e., $\delta = \delta(n)$. The response of the resonant circuit to a train of such dipulses, obtained by taking the difference in response to the two impulses, is given by

$$A_\delta(t) = P(0)\left[\sum_{n=0}^{N} \cos \omega_0(t - nT)e^{-\omega_0(t-nT)/2Q} - \cos \omega_0[t - nT + \delta(n)]e^{-\omega_0[t-nT+\delta(n)]/2Q}\right]. \quad (36)$$

In determining the response, mistuning of the resonant current can be disregarded; i.e., $\omega_0 = \omega$. Furthermore, in the second term of (36) it is permissible to take

$$\exp[-\omega_0\delta(n)/2Q] \cong 1.$$

With the following further approximations

$$\begin{aligned}\cos \omega_0(t - nT) - \cos \omega_0[t - nT + \delta(n)] \\ = \sin \omega_0[t + \delta(n)/2]\, 2 \sin [\omega_0\delta(n)/2], \\ \cong \omega_0\delta(n) \sin \omega_0 t,\end{aligned} \tag{37}$$

expression (36) becomes:

$$\begin{aligned}A_\delta(t) &= P(0)\omega_0 \sin \omega_0 t \sum_{n=0}^{N} \delta(n) e^{-\omega_0(t-nT)/2Q} \\ &= P(0)\omega_0 \sin \omega_0 t_0 \sum_{n=0}^{N} \delta(n) e^{-\omega_0 T(N-n)/2Q},\end{aligned} \tag{38}$$

where the substitution $t = NT + t_0$ has been made as in previous expressions.

The above expression shows that the resonant circuit response will be at quadrature with the steady state timing wave $\cos \omega t_0$.

In the above expressions, the dipulses are assumed to be present at intervals T, whereas in a random pulse train they will be present at average intervals $2T$. The rms value of the quadrature component with randomly positive and negative dipulses at intervals $2T$, with an rms displacement $\underline{\delta}$, is

$$\begin{aligned}\underline{A}''_\delta &= P(0)\omega_0\underline{\delta} \left[\sum_{n=0}^{N} e^{-2\omega_0 T(N-n)/Q}\right]^{1/2} \\ &= \frac{P(0)}{2}\, \omega_0\underline{\delta} \left(\frac{Q}{\pi}\right)^{1/2}.\end{aligned} \tag{39}$$

In (38) the function $e^{-\omega_0 t/2Q}$ will be recognized as the impulse response function of a circuit with impedance

$$Z(i\omega) = \frac{1}{\beta + i\omega}, \qquad \beta = \omega_0/2Q, \tag{40}$$

$$= \overline{Z}(i\omega)e^{-i\psi}, \tag{41}$$

$$\overline{Z}(i\omega) = \frac{1}{\beta}\left[\frac{1}{1 + \omega^2/\beta^2}\right]^{1/2}, \tag{42}$$

$$\tan \psi = \omega/\beta. \tag{43}$$

It will also be recognized that (39) corresponds to the rms response of such a circuit, when impulses $\delta(n)$ of random amplitude with an rms value $\underline{\delta}$ are applied to average intervals $2T$. Thus (39) can alternately be obtained from

$$\underline{A}_\delta'' = P(0)\omega_0\delta\left[\frac{1}{2T}\frac{1}{2\pi}\int_{-\infty}^{\infty}[\bar{Z}(i\omega)]^2\,d\omega\right]^{1/2}$$

$$= P(0)\omega_0\delta\left[\frac{1}{4\pi T\beta^2}\beta(\tan^{-1}\omega/\beta)_{-\infty}^{\infty}\right]^{1/2} \tag{44}$$

$$= P(0)\omega_0\delta\left(\frac{1}{4T\beta}\right)^{1/2}$$

$$= \frac{P(0)}{2}\,\omega_0\delta\left(\frac{Q}{\pi}\right)^{1/2}. \tag{45}$$

Let the output of the first resonant circuit be applied to a second resonant circuit, and in turn to n successive resonant circuits, with an amplitude amplification β between successive resonant circuits. At the output of the n^{th} resonant circuit, the rms amplitude of the response is then obtained from

$$\underline{A}_{\delta,n}'' = P(0)\omega_0\delta\left[\frac{\beta^{2(n-1)}}{4\pi T}\int_{-\infty}^{\infty}[\bar{Z}^2(i\omega)]^n\,d\omega\right]^{1/2}$$

$$= P(0)\omega_0\delta\left[\frac{1}{4\pi T\beta^2}\int_{-\infty}^{\infty}\frac{d\omega}{(1+\omega^2/\beta^2)^n}\right]^{1/2} \tag{46}$$

$$= \frac{P(0)}{2}\,\omega_0\delta\left(\frac{Q}{\pi}\right)^{1/2}, \qquad I_n = \underline{A}_\delta'' I_n,$$

where

$$I_n^2 = \frac{1}{\pi\beta}\int_{-\infty}^{\infty}\frac{d\omega}{(1+\omega^2/\beta^2)^n}, \tag{47}$$

$$= 1, \qquad n = 1,$$

$$= \frac{2n-3}{2(n-1)}\,I_{n-1}^2, \qquad n \geq 2, \tag{48}$$

$$= I_{n-1}^2\left(1 - \frac{1}{2(n-1)}\right),$$

$$I_2^2 = (1 - \tfrac{1}{2}), \qquad I_3^2 = (1 - \tfrac{1}{4})I_2^2, \qquad I_4^2 = (1 - \tfrac{1}{6})I_3^2.$$

Thus (46) can be written:

$$\underline{A}_{\delta,n}'' = \underline{A}_\delta''\underline{\alpha}_2\underline{\alpha}_3\cdots\underline{\alpha}_n, \tag{49}$$

where

$$\underline{\alpha}_j^2 = 1 - \frac{1}{2(j-1)}, \tag{50}$$

$$\alpha_2^2\alpha_3^2 \cdots \alpha_n^2 = \left(1 - \frac{1}{2}\right)\left(1 - \frac{1}{4}\right)\left(1 - \frac{1}{6}\right) \cdots \left(1 - \frac{1}{2(n-1)}\right) \quad (51)$$

$$= \frac{1 \cdot 3 \cdot 5 \cdot 7 \cdots [2(n-1) - 1]}{2 \cdot 4 \cdot 6 \cdot 8\, 2(n-1)} \quad (52)$$

$$= \frac{(2n)!}{2^{2n}(n!)^2}. \quad (53)$$

When $n \gg 1$, (51) approaches the value

$$\alpha_2^2\alpha_3^2 \cdots \alpha_n^2 \cong \left(\frac{1}{\pi n}\right)^{1/2}. \quad (54)$$

The latter approximation is based on the following expression, for $x = -\frac{1}{2}$, given in Whittaker and Watson's: "Modern Analysis" page 259:

$$\lim_{n\to\infty} (1 + x)(1 + x/2)(1 + x/3) \cdots (1 + x/n) = \frac{n^x}{\Gamma(1 + x)}, \quad (55)$$

where Γ is the gamma function, $\Gamma(-\frac{1}{2} + 1) = \pi^{1/2}$.

The above analysis assumes that the timing wave at each resonant circuit is applied directly to the next resonant circuit, except for the amplification between resonant circuits. This would be the case if the timing wave were transmitted on a separate pair, in which case $\underline{A}''_{\delta,n}$ would be the rms quadrature component owing to noise in the timing circuit.

In regenerative repeaters, deviations in the timing wave resulting from the quadrature component are imparted at intervals T into the next repeater section as deviations in the spacing of pulses. These timing deviations occurring at intervals T will have a certain random amplitude distribution, which can be regarded as having a certain frequency spectrum. When the deviations are discrete and occur at intervals T, the spectrum will extend to a maximum frequency $f_{max} = 1/2T$, or $\omega_{max} = \pi/T = \omega_0/2$. In this case the upper and lower limits of the integrals above would be replaced by $\pm\omega_0/2$, except for the first repeater section. The recurrence relation (48) is then no longer exact, but the resultant modification is insignificant and can be disregarded. This will be seen when the value $\omega_0/2$ is inserted for ω in the integrand of (47), which then becomes $1/(1 + Q^2)$, as compared with 1 for $\omega = 0$. Thus the contribution to the integrals for $\omega > \omega_0/2$ can for practical purposes be disregarded.

A Sufficient Set of Statistics for a Simple Telephone Exchange Model

By V. E. BENEŠ

(Manuscript received October 17, 1956)

This paper considers a simple telephone exchange model which has an infinite number of trunks and in which the traffic depends on two parameters, the calling-rate and the mean holding-time. It is desired to estimate these parameters by observing the model continuously during a finite interval, and noting the calling-time and hang-up time of each call, insofar as these times fall within the interval. It is shown that the resulting information may, for the purpose of this estimate, be reduced without loss to four statistics. These statistics are the number of calls found at the start of observation, the number of calls arriving during observation, the number of calls terminated during observation, and the average number of calls existing during the interval of observation. The joint distribution of these sufficient statistics is determined, in principle, by deriving a generating function for it. From this generating function the means, variances, covariances, and correlation coefficients are obtained. Various estimators for the parameters of the model are compared, and some of their distributions, means, and variances presented.

I THEORETICAL PROBLEMS AND METHODS OF TRAFFIC MEASUREMENT

Four important kinds of theoretical problems arise in the measurement of telephone traffic. These are: (1) the choice of a mathematical model, containing parameters characteristic of the traffic, to serve as a description; (2) the devising of efficient methods of estimating the parameters; (3) the determination of the anticipated accuracy of measurements; and (4) the assessment of actual accuracy, after measurements have been made.

The present paper deals with aspects of the second and third kinds of problem, for the simplest and least realistic mathematical model of telephone traffic. Specifically, for this model, we treat the problems of (i) complete extraction of the information from a given observation period,

without regard to costs of observation, and (ii) determination of the anticipated accuracy of certain methods of estimation which arise naturally from the discussion of complete extraction.

The method by which we attack problems (i) and (ii) in this paper has three stages. First we choose a small number of significant properties of, or factors in, the physical system we are studying. Then we abstract these properties into a mathematical model of the physical system. Finally, from the properties of the model, we derive results which may be interpreted as answers to the two problems treated. The advantage of this method is that we can use the precise, powerful apparatus of mathematics in studying the model; its limitation is that it yields results which are only as accurate as the model in describing reality.

A method similar to the above forms the theoretical underpinning of telephone traffic engineering itself. To design equipment effectively, the traffic engineer needs a description of the traffic that is handled by central offices. He decides what properties of the entire system of telephone equipment and customers will be most useful to him in describing the traffic. He then designates certain parameters to serve as mathematically precise idealizations of these properties, and in terms of these parameters constructs a model of the traffic, upon which he bases much of his engineering.

In choosing a mathematical model for a physical system, one is confronted with two generally opposed desiderata: fidelity to the system described, and mathematical simplicity. The model may involve important departures from physical reality; a model that is sufficiently amenable to mathematical analysis often results only after one has introduced admittedly false assumptions, ignored certain effects and correlations, and generally oversimplified the system to be studied. However, the abstract model will be an exact and simple tool for analysis.

We can construct a simple mathematical model for the operation of a telephone central office by leaving out of consideration many important facts about such systems, and by concentrating on factors most relevant to operation. Since we are interested in telephone traffic and in the availability of plant, it seems natural to require that a realistic model take account of at least the following five significant factors: (1) the demand for telephone service; (2) the rate at which requests for service can be processed and connections established; (3) the lengths of conversations; (4) the supply of central office equipment; and (5) the manner in which the first four factors are interrelated. Unfortunately, the mathematical complexity of such a realistic model precludes easy investigation. Therefore, the model used in this paper is based only on factors (1) and (3).

The demand for telephone traffic is usually made precise by describing a stochastic process which represents the way in which requests for telephone service occur in time. A realistic description will take account of the facts that, the demand is not constant, but has daily extremes, and that in small systems, the demand may be materially lessened when many conversations are in progress. Since taking account of the first fact leads to a more complicated model in which our investigations are more difficult, we ignore it, with the proviso that the results we derive are only applicable to systems and observations for which the demand is nearly constant. The second kind of variation in demand becomes insignificant as the number of subscribers increases and the traffic remains constant. Hence, we further confine the applicability of our results to systems with large numbers of subscribers, and we assume that the demand does not depend on the number of conversations in existence.

With these assumptions, a mathematically convenient description of the demand is specified by the condition that the time-intervals between requests for service have lengths which are mutually independent positive random variables, with a negative exponential distribution.

A telephone central office contains two kinds of equipment: control circuits which establish a desired connection, and talking paths over which a conversation takes place. The time that a request for service occupies a unit of equipment, be the unit a control circuit or a talking path, is called the holding-time of the unit. A request for service affects the availability of both kinds of equipment but, except for special cases, the holding-times of talking paths are usually much longer than the holding-times of control units such as markers, connectors, or registers. In view of this disparity, we assume that the only holding-times of consequence are the lengths of conversations; i.e., the holding-times of talking paths. We assume also that these lengths are mutually independent positive random variables, with a negative exponential distribution.

For the simplest mathematical model of telephone traffic, we may consider the arrangement of switches and transmission lines which constitutes a talking path in the physical office to be replaced by an abstract unit called a "trunk". A trunk is then an abstraction of the equipment made unavailable by one conversation, and we may measure the supply of talking paths in the office by the number of trunks in a model. The word "trunk" is also used to mean a transmission line linking two central offices, but as long as we have explained our use of the word there need be no confusion. Often the number of transmission lines leading out of an office is a major limitation on its capacity to carry conversations, and in this case the two uses of the word "trunk" are very similar. Un-

fortunately, we do not take advantage of this similarity, since we make the mathematically convenient but wholly unrealistic assumption that the number of trunks in the model is infinite.

The model we investigate thus depends on only two of the factors previously listed as essential to a realistic model: namely, (1) the demand for service, and (3) the lengths of conversations. In view of the simplicity and inaccuracy of this model, the question arises whether much is gained from a detailed analysis. Such scrutiny may indeed reveal little that is of great practical value to traffic engineers. It is important methodologically, however, to have a detailed treatment of at least one approximate case. We undertake this detailed treatment largely for the insight that it may give into methods which could be useful in dealing with more complex and more accurate models.

Once a designer has chosen a model and has specified the parameters he would like to have measured, it is up to the statistician to invent efficient means of measurement, by choosing, for each parameter, some function of possible observations to serve as an estimate of that parameter. One measure of efficiency that is of mostly theoretical interest is the observation time required to achieve a given degree of anticipated accuracy; the most realistic measure of efficiency is in terms of dollars and man-hours. It may often be more efficient, in the sense of the latter measure, to spread observation over enough more time to compensate for the inability of an intrinsically cheaper method of measurement to extract all of the information present in a fixed time of observation. For example, periodic scanning of switches in a telephone exchange is usually less costly than continuous observation. As a result, telephone traffic measurement is usually carried out by averaging sequences of instantaneous periodic observations of the number of calls present, rather than by continuous time averaging, although it can be shown that continuous observation is more efficient at extracting information. Thus statistical efficiency, which may be expensive in terms of measuring equipment, can be exchanged for observation time, which may be less costly. This exchange brings about a reduction in cost without impairing accuracy.

Our concern in this paper is with the less practical problems of complete extraction, and of the anticipated accuracy of estimation methods based on complete extraction. Let us consider how our mathematical model can shed light on these problems. A mathematical model may or may not be a faithful description of the behavior of real telephone systems. Nevertheless random numbers, with or without modern computing machines, enable one to make experiments and observations on physical situations which approximate, arbitrarily closely, any mathematical model. Thus we can speak meaningfully of events in the model, and of

making measurements and observations on the model. The mathematical model elucidates our problems in the following ways: (1) it enables us to state precisely what information is provided by observation; (2) it enables us to explain what we mean by complete extraction of information; and (3) it enables us to derive results about the anticipated accuracy of measurements in the model. These results will have approximately true analogues in physical situations to which the model is applicable.

The calls existing during the observation interval (O, T) fall into four categories: (i) those which exist at O, and terminate before T; (ii) those which fall entirely within (O, T); (iii) those which exist at O and last beyond T; and (iv) those which begin within (O, T) and last beyond T. For calls of category (i), we assume that we observe the hang-up time of each call; for category (ii), we observe the matching calling-time and hang-up time of each conversation; for category (iii), we observe simply the number of such calls; and for category (iv), we observe the calling-times. Table I summarizes the kinds of calls and the information observed about each.

What we mean by the complete extraction of information is made precise by the statistical concept of *sufficiency*. By a statistic we mean any function of the observations, and by an estimator we mean a statistic which has been chosen to serve as an estimate of a particular parameter. Roughly and generally, a set S of statistics is sufficient for a set P of parameters when S contains all the information in the original data that was relevant to parameters in P. If S is sufficient for P, there is a set E of estimators for parameters in P, such that the estimators in E depend only on statistics from S, and such that an estimator from E does at least as well as any other estimator we might choose for the same parameter. Thus we incur no loss in reducing the original data (of specified form) to the set S of statistics. It remains to state what it means for S to contain all the relevant information. We do this in terms of our model.

The mathematical model we are adopting contains two distribution

TABLE I — INFORMATION OBSERVED

Types of Calls	Start in (O, T)	Start before O
End in (O, T)	(ii), matching calling-times and hang-up times known, number of calls known	(i), hang-up times known, number of calls known
End after T	(iv), calling-times known, number of calls known	(iii), number of calls known

functions, that of the intervals between demands for service, and that of the lengths of conversations. We have supposed that these distributions are both of negative exponential type, each depending on a single parameter. Thus we know the functional form of each distribution, and each such form has one unknown constant in it. Since the mathematical structure of the model is fully specified except for the values of the two unknown constants, we can assign a likelihood or a probability density to any sequence Σ of events in the model during the interval (O, T). This likelihood will depend on the parameters, on Σ, and on the number of calls in existence at the start O of the interval. If the likelihood $L(\Sigma)$ can be factored into the form $L = F \cdot H$, where F depends on the parameters and on statistics from the set S only, and H is independent of the parameters, then the set S of statistics may be said to summarize all the information (in a sequence Σ) relevant to the parameters. If L can be so factored, then S is sufficient for the estimation of the parameters.

The mathematical model to be used in this paper is described and discussed in Sections II and III, respectively. Section IV contains a summary of notations and abbreviations which have been used to simplify formulas.

In Appendix A we show that the original data we have allowed ourselves can be replaced by four statistics, which are sufficient for estimation. In Appendix B and Sections V–VIII we discuss various estimators (for parameters of the model) based on these four statistics. To determine the anticipated accuracy of these methods of measurement, we consider the statistics themselves as random variables whose distributions are to be deduced from the structure of the model.

A primary task is the determination of the joint distribution of the sufficient statistics. In view of the sufficiency, this joint distribution tells us, in principle, just what it is possible to learn from a sample of length T in this simple model. By analyzing this distribution we can derive results about the anticipated accuracy of measurements in the model.

The joint distribution of the sufficient statistics is obtainable in principle from a generating function computed in Appendix C, using methods exemplified in Section X. This generating function is the basic result of this paper. The implications of this result are summarized in Section IX, which quotes the generating function itself, and presents some statistical properties of the sufficient statistics in the form of four tables: (i) a table of generating functions obtainable from the basic one; (ii) a table of mean values; (iii) a table of variances and covariances; and (iv), a table of squared correlation coefficients. (The coefficients are all non-negative.)

II DESCRIPTION OF THE MATHEMATICAL MODEL

Throughout the rest of the paper we follow a simplified form of the notational conventions of J. Riordan's paper[11] wherever possible. A summary of notations is given in Section IV. The model we study has the following properties:

(i) Demands for service arise individually and collectively at random at the rate of a calls per second. Thus the chance of one or more demands in a small time-interval Δt is

$$a\Delta t + o(\Delta t),$$

where $o(\Delta t)$ denotes a quantity of order smaller than Δt. The chance of more than one demand in Δt is of order smaller than Δt. It can be shown (Feller,[2] p. 364 et seq.) that this description of the demand is equivalent to saying that the intervals between successive demands for service are all independent, with the negative exponential distribution

$$1 - e^{-at}.$$

This again is equivalent to saying that the call arrivals form a Poisson process;[2] i.e., that for any time interval, t, the probability that exactly n demands are registered in t is

$$\frac{e^{-at}(at)^n}{n!}.$$

Thus the number of demands in t has a Poisson distribution with mean at.

(ii) The holding-times of distinct conversations are independent variates having the negative exponential distribution

$$1 - e^{-\gamma t},$$

where γ is the reciprocal of the mean holding-time h. This description of the holding-time distribution is the same as saying that the probability that a conversation, which is in progress, ends during a small time-interval Δt is

$$\gamma\Delta t + o(\Delta t),$$

without regard to the length of time that the conversation has lasted Feller, p. 375).

(iii) The model contains an infinite number of trunks. Thus, at no time will there be insufficient central office equipment to handle a demand for service, and no provision need be made for dealing with demands that cannot be satisfied.

The original work on this particular model for telephone traffic is in Palm,[9] and Palm's results have been reported by Feller[3] and Jensen.[4] The results have been extended heuristically to arbitrary absolutely continuous holding-time distributions by Riordan,[11] following some ideas of Newland[8] suggested by S. O. Rice.

Let $P_{ij}(t)$ be the probability that there are j trunks busy at t if there were i busy at 0. And let $P_i(t, x)$ be the generating function of these probabilities, defined by

$$P_i(t, x) = \sum_{j=0}^{\infty} x^j P_{ij}(t).$$

Then Palm[9] has shown (pp. 56 et seq.) that

$$P_i(t, x) = [1 + (x - 1)\, e^{-\gamma t}]^i \exp\{(x - 1)ah\,(1 - e^{-\gamma t})\}.$$

This is formula (12) of Riordan[11] with his g replaced by $e^{-\gamma t}$. It can be verified that the random variable $N(t)$ is Markovian; the limit of $P_i(t, x)$ as $t \to \infty$ is

$$\exp\{(x - 1)\, ah\},$$

so that the equilibrium distribution of the number of trunks in use is a Poisson distribution with mean $b = ah$. The shifted random variable $[N(t) - b]$ then has mean zero, and covariance function $be^{-\gamma t}$.

For additional work on this model the reader is referred to F. W. Rabe,[10] and to H. Stormer.[12]

III DISCUSSION OF THE MODEL

Let us envisage the operation of the model we have described by considering the random variable $N(t)$ equal to the number of trunks busy at time t. As a random function of time, $N(t)$ jumps up one unit step each time a demand for service occurs, and it jumps down one unit step each time a conversation ends. If $N(t)$ reaches zero, it stays there until there is another demand for service. If $N(t) = n$, the probability that a conversation ends in the next small time-interval Δt is

$$n\gamma\Delta t + o(\Delta t),$$

because the n conversations are mutually independent. A graph of a sample of $N(t)$ is shown in Fig. 1.

The model we described departs from reality in several important ways, which it is well to discuss. First, the assumption that the number of trunks is infinite is not realistic, and is justified only by the mathematical complication which results when we assume the number of trunks

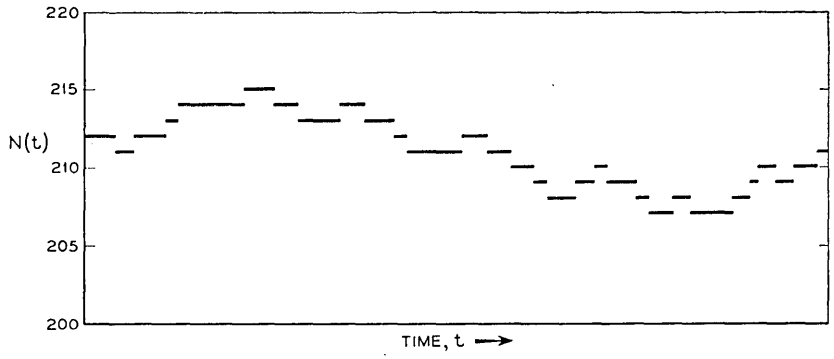

Fig. 1 — A graph of $N(t)$.

to be finite. It can also be argued that unlimited office capacity is approached by offices with adequate facilities and low calling rates, and therefore, in some practical cases at least, the model is not flagrantly inaccurate.

Second, the choice of a constant calling rate for the model ignores the fact that in most offices the calling rate is periodic. Thus, the applicability of our results to offices whose calling rates undergo drastic changes in time is restricted to intervals during which the normally variable calling rate is nearly constant. Finally, although the assumption of a negative exponential distribution of holding-time affords the model great mathematical convenience, it is doubtful whether in a realistic model the most likely holding-time would have length zero, as it does in the present one.

IV SUMMARY OF NOTATIONS

a = Poisson calling rate

h = mean holding-time

$\gamma = h^{-1}$ = hang-up rate per talking subscriber

$b = ah$ = average number of busy trunks

$N(t)$ = number of trunks in use at t

(O, T) = interval of observation

$n = N(O)$ = number of trunks in use at the start of observation

A = number of calls arriving in (O, T)

H = number of hang-ups in (O, T)

$K = A + H$

$$Z = \int_0^T N(t)\, dt$$

$M = Z/T$ = average of $N(t)$ over (O, T)

$\{p_n\}$ = the (discrete) probability distribution of n, the number of trunks found busy at the start of observation

An estimator for a parameter is denoted by adding a cap ($\hat{}$) and a subscript. The subscripts differentiate among various estimators for the same parameter. We use $\hat{a}_c = A/T$, $\hat{\gamma}_c = H/Z$, $\hat{a}_1 = K/2T$, $\hat{\gamma}_1 = K/2Z$, and $\hat{\gamma}_2 = A/Z$.

Also, it is convenient to use the following abbreviations: r for γT, and C for $(1 - e^{-r})/r$, where r is the dimensionless ratio of observation-time to mean holding-time. The symbol E is used throughout to mean mathematical expectation.

V THE AVERAGE TRAFFIC

We have adopted a model which depends on two parameters, the calling rate a, and the mean holding-time h, or its reciprocal γ. Before searching for a set of statistics that is sufficient for the estimation of these parameters, let us consider the product $ah = b$. This product is important because, as we saw in Section II, the equilibrium distribution of the number of trunks in use depends only on b, and not on a and h individually. Indeed, the equilibrium probability that n trunks are busy is

$$\frac{e^{-b}b^n}{n!},$$

and the average number of busy trunks in equilibrium is just b.

The average number of trunks busy during a time interval T is

$$M = \frac{1}{T}\int_0^T N(t)\, dt;$$

i.e., the integral of the random function $N(t)$ over the interval T, divided by T. This suggests that for large time intervals T, M will come close to the value of b, and can be used as an estimator of b. Since M is a random variable, the question arises, what are the statistical properties of M? This question has been considered in the literature, the principal references being to F. W. Rabe[10] and to J. Riordan.[11] Riordan's paper is a determination of the first four semi-invariants of the distribution of M during a period of statistical equilibrium, but without restriction on the

assumed frequency distribution of holding-time. It follows from Riordan's results that M converges to b in the mean, which is to say that

$$\lim_{T\to\infty} E\{|M - b|^2\} = 0.$$

It also follows that M is an unbiased estimator of b; i.e., that $E\{M\} = b$, and that M is a consistent estimator of b, which means that

$$\lim_{T\to\infty} pr\{|M - b| > \varepsilon\} = 0$$

for each $\varepsilon > 0$.

VI MAXIMUM CONDITIONAL LIKELIHOOD ESTIMATORS

As shown in Appendix A, the likelihood L_c of an observed sequence, conditional on $N(O)$, is defined by

$$\ln L_c = A \ln a + H \ln \gamma - \gamma Z - aT.$$

According to the method of maximum likelihood, we should select, as estimators of a and γ respectively, quantities $\hat{a}_c$ and $\hat{\gamma}_c$ which maximize the likelihood L_c. Now a maximum of L_c is also one of $\ln L_c$, and vice versa. Therefore a_c and γ_c are determined as roots of the following two equations, called the likelihood equations:

$$\frac{\partial}{\partial a} \ln L_c = 0; \qquad \frac{\partial}{\partial \gamma} \ln L_c = 0.$$

The solutions to the likelihood equations are

$$\hat{a}_c = \frac{A}{T}, \qquad \hat{\gamma}_c = \frac{H}{Z}.$$

These are the maximum conditional likelihood estimators of a and γ. The estimator $\hat{a}_c$ is the number of requests for service in T divided by T; this is intuitively satisfactory, since $\hat{a}_c$ estimates a calling rate.

Since maximum likelihood estimators of functions of parameters are generally the same functions of maximum likelihood estimators of the parameters, we see that AZ/HT is a maximum likelihood estimator of b.

VII PRACTICAL ESTIMATORS SUGGESTED BY MAXIMIZING THE LIKELIHOOD L, DEFINED IN APPENDIX A

We obtain as likelihood equations

$$\frac{\partial}{\partial a} \ln L = 0, \qquad \frac{\partial}{\partial \gamma} \ln L = 0.$$

These may be written as

$$a = \frac{n + A}{h + T},$$

and

$$\gamma = \frac{H + \frac{a}{\gamma}}{Z + \frac{n}{\gamma}}.$$

The first of these shows the estimated calling rate as a pooled combination of the conditional estimate A/T, considered in the last section, and an estimate n/h based on the initial state. This latter estimate has the form

$$\frac{\text{calls in progress}}{\text{mean holding time}},$$

and so is intuitively reasonable, since $b/h = a$. The second equation exhibits our estimate of γ as a pooled combination of the conditional estimate H/Z and the ratio a/n. This ratio is acceptable as an estimate of γ, since $a/b = \gamma$ and $b = E\{n\}$ is the average value of n.

If we substitute, in the right-hand sides of these equations, the conditional estimators A/T, H/Z, and Z/H for a, γ, and h, respectively, we obtain simple, intuitive estimators which include the influence of the initial state n, and show how it decreases with increasing T. Thus

$$\frac{n + A}{\frac{Z}{H} + T} \quad \text{estimates } a,$$

$$\frac{H + \frac{AZ}{TH}}{Z + \frac{nZ}{H}} \quad \text{estimates } \gamma.$$

VIII OTHER ESTIMATORS

Additional estimators may be arrived at by intuitive considerations, or by modifying certain maximum likelihood estimators. Some estimators so obtained are important because they use more of the information available in an observation than do the conditional estimators $\hat{a}_c$ and $\hat{\gamma}_c$, without being so complicated functionally that we cannot easily study their statistical properties.

It seems reasonable, and can be shown rigorously (Appendix C), that for an interval (O, T) of statistical equilibrium, the distribution of A and that of H are the same. Thus we can argue that, for long time intervals, A and H will not be very different. This suggests using

$$\hat{a}_1 = \frac{A + H}{2T} = \frac{K}{2T}$$

as an estimator of a. This estimator does not involve γ, and it uses not only information given by A, but also information supplied by arrivals occurring possibly before the start of observation.

Similarly, since $b = a/\gamma$, and M is a consistent and unbiased estimator of b, we may use

$$\hat{\gamma}_1 = \frac{K}{2Z} = \frac{1}{\hat{h}_1}$$

to estimate γ, and its reciprocal to estimate h. Finally, since for long (O,T) we have $A \sim H$, we may try

$$\hat{\gamma}_2 = \frac{A}{Z} = \frac{1}{\hat{h}_2}$$

as an estimator of γ, and its reciprocal as an estimator of h.

IX THE JOINT DISTRIBUTION OF THE SUFFICIENT STATISTICS

The basic result of this paper is a formula for the generating function

$$E\{z^n x^{N(T)} w^A u^H e^{-\zeta Z}\} \tag{9.1}$$

for the joint distribution of the random variables n, $N(T)$, A, H, and Z. This formula is derived in Appendix C, by methods illustrated in Section X. For an initial n distribution $\{p_n\}$, the generating function is

$$\sum_{n \geqq 0} p_n z^n \left[\frac{(\zeta x + \gamma x - \gamma u)e^{-(\zeta+\gamma)T} + \gamma u}{\zeta + \gamma}\right]^n \cdot \exp\left\{\frac{aw(\zeta x + \gamma x - \gamma u)[1 - e^{-(\zeta+\gamma)T}]}{(\zeta + \gamma)^2} + \frac{a\gamma wuT}{\zeta + \gamma} - aT\right\}. \tag{9.2}$$

It is proved in Appendix A that the set of statistics $\{n, A, H, Z\}$ is sufficient for estimation on the basis of the information assumed, which was described in Section I. Thus the generating function (9.2) specifies, at least in principle, what can be discovered about the process from an observation interval (O, T), for which $N(O)$ has the distribution $\{p_n\}$. All the results summarized in this section are consequences of (9.2).

TABLE II

X	$\ln E\{X\}$
1. $e^{-\zeta Z}$	$b\left[-\zeta T + \frac{\zeta^2 T}{\zeta+\gamma} - \frac{\zeta^2(1-e^{-(\zeta+\gamma)T})}{(\zeta+\gamma)^2}\right]$
2. $e^{-\zeta M}$	$b\left[-\zeta + \frac{\zeta^2}{\zeta+r} - \frac{\zeta^2(1-e^{-(\zeta+r)})}{(\zeta+r)^2}\right]$
3. y^K	$2aTC(y-1) + aT(1-C)(y^2-1)$
4. $e^{-\zeta \hat{a}_1}$	$2aTC(e^{-\zeta/2T}-1) + aT(1-C)(e^{-\zeta/T}-1)$
5. $y^K e^{-\zeta M}$	$b\left[\left(1-\frac{ry}{\zeta+r}\right)^2[e^{-\zeta(+r)}-1] - r\left(1-\frac{ry^2}{\zeta+r}\right)\right]$

By substitution, and by either letting the appropriate power series variables $\longrightarrow 1$, or letting $\zeta \longrightarrow 0$, or both, we can obtain from (9.2) the generating function of any combination of linear functions of the basic random variables n, $N(T)$, A, H, and Z. Some of the generating functions thereby obtained are listed in Table II, in which the entries all refer to an interval (O, T) of equilibrium.

Since, for equilibrium (O, T), the generating functions are all exponentials, it has been convenient to make Table II a table of logarithms of expectations, with random variables X on the left, and functions ln $E\{X\}$ on the right. C as a function of r is plotted in Fig. 2.

Entry 1 of Table II is actually the cumulant generating function of Z for equilibrium (O, T); similarly, Entry 2 is that of M, and depends only on the average traffic b and the ratio r. The form of the general cumulant of M is

$$k_n = b\frac{n(n-1)}{T^n}\int_0^T (T-x)x^{n-2}e^{-\gamma x}\,dx.$$

This result coincides with a special case (exponential holding-time) of a conjecture of Riordan.[11] This conjecture was first established (for a general holding-time distribution) in unpublished work of S. P. Lloyd. The cumulant generating function permits investigation of asymptotic properties. We prove in Section X that the standardized variable

$$\begin{aligned} v &= (\gamma T/2b)^{1/2}(M-b) \\ &= (r/2b)^{1/2}(M-b) \end{aligned}$$

is asymptotically normally distributed with mean 0 and variance 1.

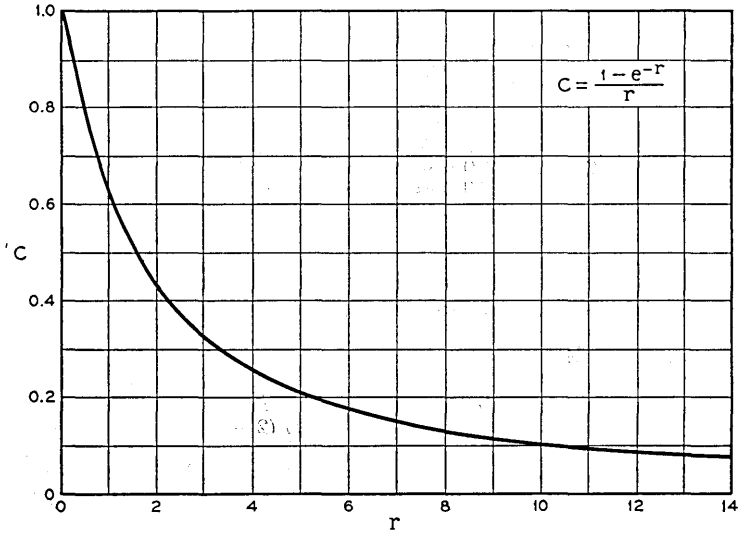

Fig. 2 — C as a function of r.

From Entry 3 of Table II it can be seen that K is distributed as $2u + v$, where u and v follow independent Poisson distributions with the respective parameters $aT(1 - C)$ and $2aTC$. The probability that $K = n$ for an interval of equilibrium is

$$r_n = \exp\{aT(C - 1)\} \sum \frac{(2aTC)^{n-2j}}{(n - 2j)!} \frac{(aT - aTC)^j}{j!},$$

where the sum is over j's for which $0 \leqq 2j \leqq n$.

The estimator $\hat{a}_1$ for a is equal to $K/2T$, and has mean and variance given by

$$E\{\hat{a}_1\} = a,$$

$$\text{var}\{\hat{a}_1\} = \frac{a}{2T}(2 - C).$$

The distribution of $\hat{a}_1$ is given by

$$\text{pr}\{\hat{a}_1 \leqq x\} = \sum r_n,$$

the summation being over $n \leqq 2Tx$.

From (9.2) one can obtain, by substitution of the stationary n distribution for $\{p_n\}$, and subsequent differentiation, the means, variances, covariances, and correlation coefficients of the sufficient statistics, for

TABLE III — $E\{X, Y\}$

	1	n	A	H	K	Z
1	1	b	aT	aT	$2aT$	bT
n		$b(1+b)$	baT	$aT(C+b)$	$aT(C+2b)$	$bT(C+b)$
A			$aT(1+aT)$	$aT(1-C+aT)$	$aT(2-C+aT)$	$bT(1-C+aT)$
H				$aT(1+aT)$	$aT(2-C+aT)$	$bT(1-C+aT)$
K					$2aT(2-C+2aT)$	$2bT(1-C+aT)$
Z						$bTh(2-2C+aT)$

TABLE IV — $\text{cov}\{X, Y\}$

	n	A	H	K	Z
n	b	0	aTC	aTC	bTC
A		aT	$aT(1-C)$	$aT(2-C)$	$bT(1-C)$
H			aT	$aT(2-C)$	$bT(1-C)$
K				$2aT(2-C)$	$2bT(1-C)$
Z					$2bTh(1-C)$

equilibrium intervals (O, T). It has been convenient to display these in three triangular arrays, the first consisting of expectations of products, the second comprising the variances and covariances, and the third exhibiting, for simplicity, the squared correlation coefficients, since the correlation coefficients are never negative for these random variables.

In Table III, the entry with coordinates (X, Y) is $E\{XY\}$ for equilibrium (O, T). All three tables are expressed in terms of a, b, T, h, r, and C, the last of which is plotted in Fig. 2.

The variances and covariances of the sufficient statistics are listed in Table IV; the entries are of the form:

$$\text{cov}\ \{X, Y\} = E\{XY\} - E\{X\}E\{Y\}.$$

Table V, finally, lists the squared correlation coefficients; i.e., the quantities

$$\rho^2(X, Y) = \frac{\text{cov}^2\ \{X, Y\}}{\text{var}\ \{X\}\ \text{var}\ \{Y\}}.$$

For any time interval (O, T), A has a Poisson distribution with parameter aT, so that $T\hat{a}_c$ does also. Therefore the distribution of $\hat{a}_c$ is given by

$$\text{pr}\ \{\hat{a}_c \leqq x\} = \sum \frac{e^{-aT}(aT)^n}{n!},$$

where the summation is over $n \leqq xT$. Evidently

$$E\{\hat{a}_c\} = a,$$

and

$$\text{var } \{\hat{a}_c\} = \frac{a}{T},$$

so that $\hat{a}_c$ is an unbiased and consistent estimator of a. We now compare the variances of estimators $\hat{a}_c$ and $\hat{a}_1$. From Table IV we have

$$\text{var } \{\hat{a}_1\} = \frac{a}{T}\left(1 - \frac{C}{2}\right) < \frac{a}{T} = \text{var } \{\hat{a}_c\},$$

so that $\hat{a}_1$ is a better estimator of a for any $T > 0$, in the sense that its variance is less.

X THE DISTRIBUTIONS of Z AND M

Since we have defined

$$Z = \int_0^T N(t)\, dt,$$

we can regard Z as the result of growth whose rate is given by the random step-function $N(t)$; when $N(t) = n$, Z is growing at rate n. An idea similar to this is used by Kosten, Manning, and Garwood[6], and by Kosten alone.[5] Now the $Z(T)$ process by itself is not Markovian, but it can be seen that the two-dimensional variable $\{N(t), Z(t)\}$ itself is Markovian. Let $F_n(z, t)$ be the probability that $N(t) = n$ and $Z(t) \leqq z$. Since the two-dimensional process is Markovian, we can derive infinitesimal relations for $F_n(z, t)$ by considering the possible changes in the system during a small interval of time Δt.

TABLE V — $\rho^2(X, Y)$

	n	A	H	K	Z
n	1	0	$1 - e^{-r}$	$\frac{rC^2}{2 - C}$	$\frac{rC^2}{2(1 - C)}$
A		1	$1 - C$	$\frac{2 - C}{2}$	$\frac{1 - C}{2}$
H			1	$\frac{2 - C}{2}$	$\frac{1 - C}{2}$
K				1	$\frac{1 - C}{2 - C}$
Z					1

If $N(t) = n$, then the probability is $[1 - \gamma n\Delta t - a\Delta t - o(\Delta t)]$ that there is neither a request for service nor a hang-up during Δt following t, and that $Z(t + \Delta t) = Z(t) + n\Delta t$. Therefore the conditional probability that $N(t + \Delta t) = n$ and $Z(t + \Delta t) \leqq z$, given that no changes occurred in Δt, is

$$F_n(z - n\Delta t, t).$$

For $N(t) = (n + 1)$, the probability is $\gamma(n + 1)\Delta t + o(\Delta t)$ that one conversation will end during Δt following t. The increment to $Z(t)$ during Δt will depend on the length x of the interval from t to the point within Δt at which the conversation ended. The increment has magnitude $(n + 1)x + n(\Delta t - x) = x + n\Delta t$, as can be verified from Fig. 3, in which the shaded area is the increment. Since x is distributed uniformly between 0 and Δt, the increment $x + n\Delta t$ is distributed uniformly between $n\Delta t$ and $(n + 1)\Delta t$. Therefore the conditional probability that $N(t + \Delta t) = n$ and $Z(t + \Delta t) \leqq z$, given that one conversation ended in Δt, is

$$\frac{1}{\Delta t}\int_{n\Delta t}^{(n+1)\Delta t} F_{n+1}(z - u, t)\, du.$$

By a similar argument it can be shown that the probability that one request for service arrives in Δt is $a\Delta t + o(\Delta t)$, and that the conditional probability that $N(t + \Delta t) = n$ and $Z(t + \Delta t) \leqq z$, given that one request arrived during Δt, is

$$\frac{1}{\Delta t}\int_{(n-1)\Delta t}^{n\Delta t} F_{n-1}(z - u, t)\, du.$$

Define $F_n(z, t)$ to be identically 0 for negative n. Adding up the probabil-

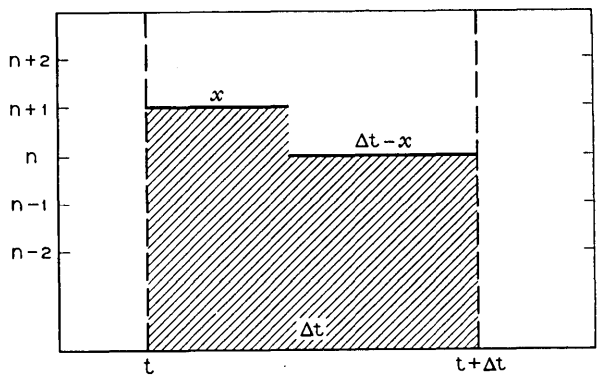

Fig. 3 — Increment to Z in Δt.

ities of mutually exclusive events, we obtain the following infinitesimal relations for $F_n(z, t)$:

$$F_n(z, t + \Delta t) = \gamma(n + 1) \int_{n\Delta t}^{(n+1)\Delta t} F_{n+1}(z - u, t)\, du + a \int_{(n-1)\Delta t}^{n\Delta t} F_{n-1}(z - u, t)\, du + F_n(z - n\Delta t, t) \cdot [1 - \Delta t(\gamma n + a)] + o(\Delta t), \quad \text{for any } n.$$

Expanding the penultimate term of the right side in powers of $n\Delta t$, and the left side in powers of Δt, we divide by Δt, and take the limit as Δt approaches 0. Now

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{n\Delta t}^{(n+1)\Delta t} F_{n+1}(z - u, t)\, du = F_{n+1}(z, t).$$

Thus, omitting functional dependence on z and t for convenience, we reach the following partial differential equations for $F_n(z, t)$:

$$\frac{\partial}{\partial t} F_n = \gamma(n + 1)F_{n+1} + aF_{n-1} - n \frac{\partial}{\partial z} F_n - [\gamma n + a]F_n, \quad \text{for any } n. \tag{10.1}$$

Since $Z(0) = 0$, we impose the following boundary conditions:

$$\begin{aligned} F_n(0, t) &= 0 \quad \text{for } n > 0 \text{ and } t > 0, \\ F_n(z, 0) &= p_n \quad \text{for } z \geqq 0, \\ F_n(z, 0) &= 0 \quad \text{for } z < 0, \end{aligned} \tag{10.2}$$

where the sequence $\{p_n\}$ forms an arbitrary $N(0)$ distribution that is zero for negative n.

To transform the equations, we introduce the Laplace-Stieltjes integrals

$$\varphi_n(\zeta, t) = \int_{0-}^{\infty} e^{-\zeta z}\, dF_n(z, t), \qquad t \geqq 0, \qquad \text{Re } (\zeta) > 0,$$

in which the Stieltjes integration is understood always to be on the variable z. We note that

$$\int_{0-}^{\infty} e^{-\zeta z} F_n(z, t)\, dz = \frac{1}{\zeta} \varphi_n(\zeta, t),$$

and that

$$\varphi_n(\zeta, t) = F_n(0, t) + \int_{0}^{\infty} e^{-\zeta z} \frac{\partial}{\partial z} F_n\,(z, t)\, dz.$$

Applying now the Laplace-Stieltjes transformation to (10.1), we obtain

$$\frac{\partial \varphi_n}{\partial t} = \gamma(n+1)\varphi_{n+1} + a\varphi_{n-1} - n\zeta\varphi_n + n\zeta F_n(0,t) - [\gamma n + a]\varphi_n , \tag{10.3}$$

in which we have left out functional dependence on ζ and t where it is unnecessary. By the boundary conditions (10.2), $n\zeta F_n(0, t) = 0$ for $n \geqq 0$ and $t > 0$; in (10.3) we may therefore omit this term in the region $t > 0$. Let φ be defined by

$$\varphi(x, \zeta, t) = \sum_{n=0}^{\infty} x^n \varphi_n(\zeta, t).$$

The series is absolutely convergent for $| x | < 1$, since

$$| \varphi_n(\zeta, t) | \leqq 1, \text{ for all } n.$$

The following partial differential equation for φ is obtained from (10.3):

$$\frac{\partial \varphi}{\partial t} + [\zeta x + \gamma(x-1)]\frac{\partial \varphi}{\partial x} = a(x-1)\varphi. \tag{10.4}$$

If we integrate out the information about Z by letting ξ approach 0 in this equation, we obtain the equation derived by Palm (loc. cit.) for the generating function of $N(t)$. Therefore our equation has a solution of the same form as Palm's. For the boundary conditions (10.2), this solution is

$$\varphi = \exp\left\{\frac{a[1 - e^{-(\zeta+\gamma)t}]}{(\zeta+\gamma)^2}[\zeta x + \gamma(x-1)] - \frac{a\zeta t}{\zeta+\gamma}\right\} \cdot \sum_{n=0}^{\infty} p_n \left[\frac{[\zeta x + \gamma(x-1)]e^{-(\zeta+\gamma)t} + \gamma}{\zeta+\gamma}\right]^n. \tag{10.5}$$

Actually φ contains more information than we want since it yields the joint distribution of N and Z. We may integrate out the former variable by letting x approach 1 in 10.5. Then,

$$E\{\exp(-\zeta Z)\} = \exp\left\{\frac{a\zeta(1 - e^{-(\zeta+\gamma)T})}{(\zeta+\gamma)^2} - \frac{a\zeta T}{\zeta+\gamma}\right\} \cdot \sum_{n=0}^{\infty} p_n \left[\frac{\zeta e^{-(\zeta+\gamma)T} + \gamma}{\zeta+\gamma}\right]^n$$

is the Laplace transform of the distribution of Z for an arbitrary $N(O)$ distribution $\{p_n\}$. This result is not restricted to an interval (O, T)

of statistical equilibrium; however, if the sequence $\{p_n\}$ does form the stationary distribution discussed in Section II, then

$$\sum x^n p_n = \exp\{b(x-1)\}, \tag{10.6}$$

and

$$\psi = \exp\left\{\frac{b\zeta^2(e^{-(\zeta+\gamma)T}-1)}{(\zeta+\gamma)^2} - \frac{a\zeta T}{\zeta-\gamma}\right\} \tag{10.7}$$

is the Laplace transform of the distribution of Z for an interval (O, T) of statistical equilibrium.

The Laplace transform is a moment generating function expressible as

$$\psi = \sum_{n=0}^{\infty} \frac{(-\zeta)^n m_n}{n!},$$

where m_n is the n^{th} ordinary moment of Z. Differentiation of 10.7 then gives a recurrence relation for the moments upon equating powers of $(-\zeta)$. Thus,

$$(\zeta+\gamma)^3\left(-\frac{\partial\psi}{\partial\zeta}\right) = \psi\cdot\{2a\zeta(1-e^{-(\zeta+\gamma)T}) + (\zeta+\gamma)[\gamma aT + bT\zeta^2 e^{-(\zeta+\gamma)T}]\},$$

and

$$\begin{aligned}\gamma^3 m_{n+1} - 3\gamma^2 n m_n + 3\gamma n(n-1)m_{n-1} - n(n-1)(n-2)m_{n-2} \\ = a\gamma^2 T m_n - (2a + a\gamma T)n m_{n-1} + 2ane^{-\gamma T}(m+T)^{n-1} + n \\ \cdot(n-1)aTe^{-\gamma T}(m+T)^{n-2} - n(n-1)(n-2)bTe^{-\gamma T}(m+T)^{n-3},\end{aligned} \tag{10.8}$$

where $(m+T)^n$ is the usual symbolic abbreviation of

$$\sum_{j=0}^{n}\binom{n}{j} T^j m_{n-j}.$$

From the recurrence (10.8) it is easily verified that

$$m_1 = bT,$$

$$m_2 = (bT)^2 + \frac{2bT}{\gamma}[1-C],$$

from which it follows that the variance of Z is

$$\text{var}\{Z\} = \frac{2bT}{\gamma}[1-C].$$

Since ψ is the Laplace-Stieltjes transform of the distribution of Z over an interval of equilibrium, $\ln \psi$ is the cumulant generating function, and has the following simple form:

$$\begin{aligned} \ln \psi &= b\left[\frac{\zeta^2(e^{-(\zeta+\gamma)T} - 1)}{(\zeta+\gamma)^2} - \frac{\gamma\zeta T}{\zeta+\gamma}\right] \\ &= b\left[-\zeta T + \frac{\zeta^2 T}{\zeta+\gamma} - \frac{\zeta^2(1 - e^{-(\zeta+\gamma)T})}{(\zeta+\gamma)^2}\right]. \end{aligned} \tag{10.9}$$

M is a linear function of Z, so we may obtain the cumulant generating function of M in accordance with Cramér[1] (p. 187). This function is

$$b\left[-\zeta + \frac{\zeta^2}{r+\zeta} - \frac{\zeta^2(1 - e^{-r}e^{-\zeta})}{(r+\zeta)^2}\right], \tag{10.10}$$

and depends only on b and r.

The mean and variance of M for an interval of equilibrium are respectively given by

$$E\{M\} = b,$$

$$\text{var}\{M\} = \frac{2b}{r}[1 - C], \qquad \text{with} \quad C = \frac{1 - e^{-r}}{r},$$

results which were first proved in Riordan.[11] A normal distribution having the mean and variance of M has the cumulant generating function

$$b\left[-\zeta + \frac{\zeta^2}{r} + \frac{\zeta^2(e^{-r} - 1)}{r^2}\right], \tag{10.11}$$

which is to be compared to (10.10). Since var $\{M\}$ goes to 0 as T approaches ∞, we may expect that a suitably normalized version of Z will be asymptotically normally distributed as T approaches ∞. The cumulant generating function of the normalized variable $(2bhT)^{-1/2}(Z - bT)$ is

$$\frac{\zeta^2}{\zeta\left(\frac{2}{aT}\right)^{1/2} + 2\left[1 + \frac{\exp\left\{-\zeta\left(\frac{2b}{r}\right)^{-1/2} - r\right\} - 1}{\zeta\left(\frac{2b}{r}\right)^{-1/2} + r}\right]},$$

which approaches $\zeta^2/2$ as $T \to \infty$. It follows that the normalized variable is asymptotically normal with mean 0 and variance 1, and that $(r/2b)^{\frac{1}{2}}(M - b)$ is also asymptotically normal (0, 1).

APPENDIX A

PROOF THAT $\{n, A, H, Z\}$ IS SUFFICIENT.

We observe the system during the interval (O, T), and gather the information specified in Section I, and summarized in Table I. From this information we can extract four sets of numbers, described as follows:

- S_a the set of complete observed inter-arrival times, not counting the interval from the last arrival until T
- S_h the set of complete observed holding times
- S_1 the set of hang-up times for calls of category (i)
- S_4 the set of calling-times for calls of category (iv)

In addition, our data enable us to determine the following numbers:

- n the number $N(O)$ of calls found at the start of observation
- k the number of calls of category (iii); i.e., of calls which last throughout the interval (O, T)
- x the length of the time-interval between the last observed arrival and T

In view of the negative exponential distributions which have been assumed for the inter-arrival times and the holding-times, and in view of the assumptions of independence, we can write the likelihood of an observed sequence of events as

$$L = e^{-k\gamma T - ax} p_n \cdot \prod_{u \epsilon S_a} a e^{-au} \cdot \prod_{z \epsilon S_h} \gamma e^{-\gamma z} \cdot \prod_{w \epsilon S_1} e^{-\gamma w} \cdot \prod_{y \epsilon S_4} e^{-\gamma(T-y)},$$

so that

$$\ln L = -\gamma k T - ax + \ln p_n + A \ln a - \sum_{u \epsilon S_a} au + H \ln \gamma - \sum_{z \epsilon S_h} \gamma z - \sum_{w \epsilon S_1} \gamma w - \sum_{y \epsilon S_4} \gamma(T - y)$$

It is easily seen that the summations and the two initial terms can be combined into a single term, so that we obtain

$$\ln L = \ln p_n + A \ln a + H \ln \gamma - \gamma Z - aT.$$

This shows that L depends only on the statistics n, A, H, and Z; it follows that the information we have assumed can be replaced by the set of statistics $\{n, A, H, Z\}$, and that these are sufficient for estimation based on that information.

The likelihood is sometimes defined without reference to the initial state, by leaving the factor p_n out of the expression for L. Strictly speaking, this omission defines the conditional likelihood for the observed

sequence, conditional on starting at n. We use the notation:

$$L_c = \frac{L}{p_n}.$$

A definition of likelihood as L_c has been used by Moran.[7] Clearly

$$\ln L_c = A \ln a + H \ln \gamma - \gamma Z - aT.$$

Appendix B

Unconditional Maximum Likelihood Estimates

The definition of likelihood as L leads to complicated results which are of theoretical rather than practical interest. For this reason these results have been relegated to an appendix.

The results of setting $\partial/\partial\gamma \ln L$ and $\partial/\partial a \ln L$ equal to zero lead, respectively, to the likelihood equations

$$a - \gamma(n - H) - \gamma^2 Z = 0,$$

$$\gamma n - a + \gamma A - a\gamma T = 0.$$

Considered as a system of equations for γ and a, this pair has the non-negative roots

$$\hat{\gamma} = \frac{H - n - M + \{(H - n - M)^2 + 4MK\}^{1/2}}{2Z},$$

$$\hat{a} = \frac{K}{T} - \hat{\gamma}M.$$

These are the unconditional maximum likelihood estimators for γ and a. Although $\hat{a}_c$ depended only on A and T, and $\hat{\gamma}_c$ only on H and Z, the unconditional estimators depend on all of n, A, H, Z, and T. We may obtain a maximum unconditional likelihood estimator for b as well, either by considering L to be a function of b and γ, or from general properties of maximum likelihood estimators. Since $b = a/\gamma$, we expect that $\hat{b} = \hat{a}/\hat{\gamma}$, as can be verified by an argument similar to that used above for $\hat{a}$ and $\hat{\gamma}$.

The estimators $\hat{a}$, b, and $\hat{\gamma}$ obtained in this Appendix may turn out to be useful in practice, but their complicated dependence on the sufficient statistics n, A, H, and Z makes a study of their statistical properties difficult. As a first step along such a study, we have derived the generating function of the joint distribution of the sufficient statistics in Appendix C. The greater simplicity of the conditional estimators of Section VI makes it possible to study their statistical properties. This

fact gives them a practical ascendancy over the unconditional estimators, even though the latter may be more efficient statistically by dint of using all the information available in an observation.

Appendix C

The joint distribution of $N(t)$, n, A, H, and Z

By methods already used in Section X one can obtain a generating function for the joint distribution of all the random variables n, $N(t)$, A, H, and Z. Let

$$\Phi = E\{x^{N(t)}w^A u^H e^{-\zeta Z}\}.$$

Then Φ satisfies the differential equation

$$\frac{\partial \Phi}{\partial t} + [\zeta x + \gamma x - \gamma u]\frac{\partial \Phi}{\partial x} = a(wx - 1)\Phi,$$

whose solution has the form

$$\Phi = R\{[\zeta x + \gamma x - \gamma u]e^{-(\zeta+\gamma)t}\} \cdot \exp\left(\frac{aw[\zeta x + \gamma x - \gamma u][1 - e^{-(\zeta+\gamma)t}]}{(\zeta + \gamma)^2} + \frac{a\gamma wut}{\zeta + \gamma} - at\right),$$

where the function R is determined by the initial distribution $\{p_n\}$ through the relation

$$R\{\xi\} = \sum_{n \geqq 0} p_n \left[\frac{\xi + \gamma u}{\zeta + \gamma}\right]^n.$$

From these results it follows that the generating function

$$E\{z^n x^{N(T)} w^A u^H e^{-\zeta Z}\}$$

is given by

$$\sum_{n \geqq 0} p_n z^n \left(\frac{(\zeta x + \gamma x - \gamma u)e^{-(\zeta+\gamma)T} + \gamma u}{\zeta + \gamma}\right)^n \cdot \exp\left(\frac{aw(\zeta x + \gamma x - \gamma u)[1 - e^{-(\zeta+\gamma)T}]}{(\zeta + \gamma)^2} + \frac{a\gamma wuT}{\zeta + \gamma} - aT\right).$$

If $\{p_n\}$ forms the stationary distribution, this reduces to

$$\exp\left[b\left(\frac{z(\zeta x + \gamma x - \gamma u)e^{-(\zeta+\gamma)T} + \gamma uz}{\zeta + \gamma} - 1\right) + \frac{aw(\zeta x + \gamma x - \gamma u)[1 - e^{-(\zeta+\gamma)T}]}{(\zeta + \gamma)^2} + \frac{a\gamma wuT}{\zeta + \gamma} - aT\right].$$

If, in this last expression, we let x approach 1, z approach 1, and u approach 1, we obtain

$$\exp\left[\left(1 - \frac{\gamma w}{\zeta + \gamma}\right)\left(\frac{b\zeta(e^{-(\zeta+\gamma)T} - 1)}{\zeta + \gamma} - aT\right)\right] \qquad \text{(C)}$$

as the generating function $E\{w^A e^{-\zeta Z}\}$ for an interval of equilibrium. Alternately, if instead we let x approach 1, z approach 1, and w approach 1, we obtain (C) with u substituted for w; this implies the not-surprising result that for an interval of equilibrium, the two-dimensional variables $\{A, Z\}$ and $\{H, Z\}$ have the same distribution. From this and (C) it follows that for equilibrium (O, T), (i) A and H both have a Poisson distribution with mean aT, and (ii) the estimators $\hat{h}_c$ and $\hat{h}_2$ have the same distribution.

ACKNOWLEDGEMENT

The author would like to express his gratitude for the helpful comments and suggestions of E. N. Gilbert, S. P. Lloyd, J. Riordan, J. Tukey, and P. J. Burke.

REFERENCES

1. H. Cramér, Mathematical Methods of Statistics, Princeton, 1946.
2. W. Feller, An Introduction to Probability Theory and its Applications, John Wiley and Sons, New York, 1950.
3. W. Feller, On the Theory of Stochastic Processes with Particular Reference to Applications, Proceedings of the Berkeley Symposium on Math. Statistics and Probability, Univ. of California Press, 1949.
4. A. Jensen, An Elucidation of Erlang's Statistical Works Through the Theory of Stochastic Processes, in The Life and Works of A. K. Erlang, Copenhagen, 1948.
5. L. Kosten, On the Accuracy of Measurements of Probabilities of Delay and of Expected Times of Delay in Telecommunication Systems, App. Sci. Res., **B2,** pp. 108–130 and pp. 401–415, 1952.
6. L. Kosten, J. R. Manning, F. Garwood, On the Accuracy of Measurements of Probabilities of Loss in Telephone Systems, Journal of the Royal Statistical Society (B), **11,** pp. 54–67, 1949.
7. P. A. P. Moran, Estimation Methods for Evolutive Processes, Journal of the Royal Statistical Society (B), **13,** pp. 141–146, 1951.
8. W. F. Newland, A Method of Approach and Solution to Some Fundamental Traffic Problems, P.O.E.E. Journal, **25,** pp. 119–131, 1932–1933.
9. C. Palm, Intensitätsschwankungen im Fernsprechverkehr, Ericsson Technics, **44,** 1943.
10. F. W. Rabe, Variations of Telephone Traffic, Elec. Comm., **26,** 243–248, 1949.
11. J. Riordan, Telephone Traffic Time Averages, B.S.T.J., **30,** 1129–1144, 1951.
12. H. Stormer, Anwendung des Stichprobenverfahrens beim Beurteilen von Fernsprechverkehrsmessungen, Archiv der Elektrischen Übertragung, **8,** pp. 439–436, 1954.

Fluctuations of Telephone Traffic

By V. E. BENEŠ

(Manuscript received November 9, 1956)

The number of calls in progress in a simple telephone exchange model characterized by unlimited call capacity, a general probability density of holding-time, and randomly arriving calls is defined as $N(t)$. A formula, due to Riordan, for the generating function of the transition probabilities of $N(t)$ is proved. From this generating function, expressions for the covariance function of $N(t)$ and for the spectral density of $N(t)$ are determined. It is noted that the distributions of $N(t)$ are completely specified by the covariance function.

I INTRODUCTION

The aim of this paper is to study the average fluctuations of telephone traffic in an exchange, by means of a simple mathematical model to which we apply concepts used in the theory of stochastic processes and in the analysis of noise.

The mathematical model we use is based on the following assumptions: (1) requests for telephone service arise individually and collectively at random at an average rate of a per second; (2) the holding-times of calls are mutually independent random variables having the common probability density function $h(u)$; and (3) the capacity of the exchange is effectively unlimited, and no call is blocked or delayed by lack of equipment. This telephone exchange model has been described by J. Riordan.[5]

As a measure of traffic, it is natural to use the number of calls in progress in the exchange. We are thus led to consider a random step-function of time $N(t)$, defined as the number of calls in progress at time t. $N(t)$ fluctuates about an average in a manner depending on the calling-rate, a, and the holding-time density, $h(u)$.

II PROOF OF RIORDAN'S FORMULA FOR TRANSITION PROBABILITIES

Let $P_{m,n}(t)$ be the probability that n calls are in progress at t if m calls were in progress at 0. Define the generating function of these prob-

abilities as

$$P_m(t, x) = \sum_{n \geqq 0} P_{m,n}(t)x^n,$$

and let

$$f(u) = \int_u^\infty h(x)\, dx,$$

so that the average holding-time, h, is given by

$$h = \int_0^\infty f(u)\, du.$$

Riordan[5] has given the following formula for $P_m(t, x)$:

$$P_m(t, x) = [1 + (x - 1)g(t)]^m \exp\{(x - 1)ah[1 - g(t)]\}, \qquad (1)$$

with

$$g(t) = \frac{1}{h}\int_t^\infty f(u)\, du.$$

For exponential holding-time density, this formula had already been derived (as the solution of a differential equation) by Palm.[2]

In private communication, J. Riordan has suggested that his proof of (1) is incomplete. We therefore give a new proof of (1).

We seek the generating function of $N(t)$, conditional on the event $N(0) = m$. We obtain it by first computing the joint generating function of $N(0)$ and $N(t)$; that is,

$$E\{y^{N(0)}x^{N(t)}\}. \qquad (2)$$

The desired conditional generating function is then the coefficient of y^m in (2), divided by the probability that $N(0) = m$.

To obtain a formula for (2), we exhaust the interval $(-\infty, 0)$ by division into a countable set of disjoint intervals, I_n, the n^{th} having length $T_n > 0$. Let S_n be the sum of the first n lengths, T_j. Let $\xi_n(t)$, for $t > -S_{n-1}$, be the number of those calls which arrive in I_n and are still in progress at t. And let $\eta(t)$ be the number of calls arriving during $(0, t)$, $t > 0$, and still in existence at t. Then

$$N(0) = \sum_{n \geqq 1} \xi_n(0), \qquad (3)$$

$$N(t) = \eta(t) + \sum_{n \geqq 1} \xi_n(t), \qquad t > 0. \qquad (4)$$

Since calls arriving during disjoint intervals are independent, we know

that $\eta(t)$ is independent of all the ξ's, and that $\xi_n(t)$ is independent of $\xi_j(\tau)$ if $n \neq j$. Of course, $\xi_n(t)$ and $\xi_n(\tau)$ are not independent. It follows that if the infinite product converges, then for $t > 0$

$$E\{y^{N(0)}x^{N(t)}\} = E\{x^{\eta(t)}\} \prod_{n=1}^{\infty} E\{y^{\xi_n(0)}x^{\xi_n(t)}\}. \tag{5}$$

We now compute the terms of the product. If a call originates in interval I_n, it still exists at 0 with probability

$$Q_n = \frac{1}{T_n}\int_0^{T_n} f(u + S_{n-1})\,du = \frac{1}{T_n}\int_{S_{n-1}}^{S_n} f(u)\,du.$$

Hence if k calls arrived in I_n, the probability that m of them are still in progress at 0 is

$$\begin{aligned}\operatorname{pr}\{\xi_n(0) &= m \mid k \text{ calls arrive in } I_n\}\\ &= \binom{k}{m} Q_n^m(1 - Q_n)^{k-m}, \qquad m \leqq k.\end{aligned}$$

Similarly, if a call originates in I_n and exists at 0, it also exists at $t > 0$ with probability

$$K_n = (Q_nT_n)^{-1}\int_0^{T_n} f(u + t + S_{n-1})\,du.$$

Therefore

$$\begin{aligned}E\{x^{\xi_n(t)} \mid \xi_n(0) &= m \text{ and } k \text{ calls arrive in } I_n\}\\ &= [1 + (x - 1)K_n]^m,\end{aligned}$$

and so

$$\begin{aligned}E\{y^{\xi_n(0)}x^{\xi_n(t)} \mid k \text{ calls arrive in } I_n\}&\\ &= \{1 + \langle y[1 + (x - 1)K_n] - 1\rangle Q_n\}^k\\ &= \alpha^k.\end{aligned}$$

The number of calls arriving during I_n has a Poisson distribution with mean aT_n; hence

$$\begin{aligned}E\{y^{\xi_n(0)}x^{\xi_n(t)}\} &= \exp\{aT_n(\alpha - 1)\}\\ &= \exp\{aT_nQ_n\langle y[1 + (x - 1)K_n] - 1\rangle\}.\end{aligned} \tag{6}$$

By reasoning like that leading to (6), it can be shown that

$$E\{x^{\eta(t)}\} = \exp\left\{at(x-1)\frac{1}{t}\int_0^t f(u)\,du\right\} = \exp\{ah(x-1)[1-g(t)]\}. \tag{7}$$

Now

$$\sum_{n\geqq 1} aT_nQ_n = a\int_0^\infty f(u)\,du = ah,$$

$$\sum_{n\geqq 1} aT_nQ_nK_n = a\sum_{n\geqq 1}\int_0^{T_n} f(u+t+S_{n-1})\,du,$$

$$= a\int_t^\infty f(u)\,du = ahg(t).$$

Therefore the infinite product is convergent, and

$$E\{y^{N(0)}x^{N(t)}\} = \exp\{ah(x-1)[1-g(t)] + \sum_{n\geqq 1} aT_nQ_n\langle y[1+(x-1)K_n]-1\rangle\} \tag{8}$$

$$= \exp\{ah\langle(x-1)[1-g(t)]+(y-1)+y(x-1)g(t)\rangle\}.$$

Thus the generating function of the joint distribution of $N(0)$ and $N(t)$ is independent of the division of $(-\infty, 0)$ into intervals I_n. By letting x approach 1 in (8) and finding the coefficient of y^m in the resulting limit, we find that

$$\text{pr}\{N(0) = m\} = \frac{e^{-ah}(ah)^m}{m!}. \tag{9}$$

The coefficient of y^m in (8) itself is

$$\frac{e^{-ah}(ah)^m}{m!}[1+(x-1)g(t)]^m \exp\{(x-1)ah[1-g(t)]\},$$

and so using (9) we find that the required conditional generating function of $N(t)$, given $N(0) = m$, is given by Riordan's formula (1).

III THE AUTOCORRELATION

In terms of $N(t)$ one can define various stochastic integrals which will be characteristic of the process. A simple one which has been extensively treated in connection with estimating the average traffic is

$$M = \frac{1}{T}\int_0^T N(t)\,dt,$$

the average of $N(t)$ over an interval $(0, T)$. The chief references in the literature on M are References 3 and 5. If we consider $N(t)$ during an interval $(0, T + \tau)$, a measure of the coherence of $N(t)$ during this interval, i.e., of the extent to which $N(t)$ hangs together, is given by the integral

$$U(T, \tau) = \frac{1}{T} \int_0^T N(t)\, N(t + \tau)\, dt,$$

depending on values of $N(t)$ taken τ apart. When the limit $\psi(\tau)$ of u as T approaches ∞ exists, it is usually called the autocorrelation function; most statisticians, however, reserve the term "correlation" for suitably normalized, dimensionless quantities. It can be shown that this limit exists and is the same for almost all $N(t)$ in the ensemble. It then coincides with the ensemble average, i.e.,

$$\begin{aligned} \psi(\tau) &= \lim_{T \to \infty} U(T, \tau), \text{ almost all } N(t), \\ &= E\{N(t)N(t + \tau)\}. \end{aligned}$$

The function, ψ, for the system we are discussing is derived by Riordan,[5] and we reproduce his argument for ease of understanding. For equilibrium, and $b = ah$, we have

$$E\{N(t)N(t + \tau)\} = \sum_{m=0}^{\infty} \frac{e^{-b}(b)^m}{m!}\, m \frac{\partial}{\partial x} P_m(\tau, x) \Big]_{x=1}.$$

Now

$$\frac{\partial}{\partial x} P_m(\tau, x) \Big]_{x=1} = mg(\tau) + b[1 - g(\tau)],$$

so that

$$\begin{aligned} \psi(\tau) &= \sum_{m=0}^{\infty} \frac{e^{-b}b^m}{m!}\, m\{mg(\tau) + b[1 - g(\tau)]\}, \\ &= b^2 + bg(\tau). \end{aligned} \tag{10}$$

(Cf.,[5] p. 1136)

The limiting value of $\psi(\tau)$ for τ approaching ∞ is the square of the mean occupancy, b, and the limiting value of $\psi(\tau)$ for τ approaching 0 is the mean square occupancy, $b^2 + b$, the second moment of the Poisson distribution with mean b.

IV. THE COVARIANCE AND SPECTRAL DENSITY

The average value of $N(t)$ is $b = ah$. One way to study the fluctuations of $N(t)$ about its average is by means of the power spectrum used in the analysis of noise. (Cf. Rice.[4]) We resolve the difference $[N(t) - b]$ into sinusoidal components of non-negative frequency, and postulate a noise current proportional to this difference dissipating power through a unit resistance. The spectrum $w(f)$ is then the average power due to frequencies in the interval $(f, f + df)$.

More formally, we consider the Fourier integral

$$S(f, T) = \int_0^T [N(t) - b]e^{-2\pi ift}\, dt,$$

and we recall, for completeness, the relationship between $S(f, T)$ and the covariance function, $R(\tau)$, of $[N(t) - b]$. If

$$w(f) = \lim_{T\to\infty} \frac{2 \mid S(f, T)\mid^2}{T},$$

then

$$\begin{aligned} w(f) &= 4\int_0^\infty R(\tau) \cos 2\pi f\tau\, d\tau, \\ R(\tau) &= \int_0^\infty w(f) \cos 2\pi f\tau\, df. \end{aligned} \tag{11}$$

(Cf. Rice,[4] p. 312 ff.)

At the same time, we have

$$\begin{aligned} R(\tau) &= E\{[N(t) - b][N(t + \tau) - b]\} \\ &= \psi(\tau) - b^2 \\ &= bg(\tau). \end{aligned}$$

Let $X(t)$ be any stochastic process which is known to be the occupancy of a telephone exchange of unlimited capacity, having a probability density of holding-time, and subject to Poisson traffic. From the preceding result it can be seen that the covariance function of $X(t)$ determines the distributions of the $X(t)$ process completely, since

$$a = -\frac{dR}{d\tau}\Big]_{t=0},$$

$$f(\tau) = \int_\tau^\infty h(u)\, du = -a^{-1}\frac{dR}{d\tau}.$$

If the holding-times are bounded by a constant, k, then readings of $N(t)$ taken further apart than k are uncorrelated. In fact, such values

are independent, because no call which contributes to $N(t)$ can survive until $(t + k)$, with probability 1.

Using (11), we see that

$$\begin{aligned} w(f) &= 4\int_0^\infty \cos 2\pi f\tau R(\tau)\, d\tau \\ &= 4b\int_0^\infty \cos 2\pi f\tau g(\tau)\, d\tau \\ &= 4a\int_0^\infty \cos 2\pi f\tau \int_\tau^\infty \int_y^\infty h(u)\, du\, dy\, d\tau \qquad (12) \\ &= \frac{2a}{\pi f}\int_0^\infty \sin 2\pi f\tau \int_\tau^\infty h(u)\, du\, d\tau \\ &= \frac{a}{\pi^2 f^2}\left[1 - \int_0^\infty \cos 2\pi f\tau h(\tau)\, d\tau\right]. \end{aligned}$$

Equation (12) expresses the mean square of the frequency spectrum of the fluctuations of the traffic away from the average in terms of the calling-rate and the cosine transform of the holding-time density, $h(u)$. The calling-rate appears only as a factor, and so does not affect the shape of $w(f)$. The function $w(f)$ is what Doob[1] (p. 522) calls the "spectral density function (real form)."

V EXAMPLE 1. $N(t)$ MARKOVIAN

Let the frequency $h(u)$ be negative exponential, so that

$$h(u) = \frac{1}{h}\, e^{-u/h}, \qquad (13)$$

where h is the mean holding-time. It is shown in Riordan[5] p. 1134, that $N(t)$ is Markovian if and only if $h(u)$ has the form (13). From page 523 of Doob[1] we know that the covariance function of a real, stationary Markov process (wide sense) has the form

$$R(\tau) = R(0)e^{-\alpha\tau}, \qquad \alpha \text{ constant}. \qquad (14)$$

Under the assumption (13), the covariance of $N(t)$ is

$$\begin{aligned} R(\tau) = bg(\tau) &= \frac{b}{h}\int_\tau^\infty \int_y^\infty h(u)\, du\, dy \\ &= b\int_\tau^\infty \int_y^\infty \frac{1}{h^2}\, e^{-u/h}\, du\, dy \\ &= be^{-\tau/h}, \end{aligned}$$

in agreement with (14). The spectral density can now be obtained from (11) or (12); it is

$$w(f) = \frac{4bh}{1 + 4\pi^2 f^2 h^2}.$$

This is the same as would be obtained for a Markov process that alternately assumed the values $+\sqrt{ah}$, $-\sqrt{ah}$ at the Poisson rate of $(2h)^{-1}$ changes of sign per sec. (Cf. Rice[4] p. 325.)

VI EXAMPLE 2. HOLDING-TIME DISTRIBUTED UNIFORMLY IN (α,β)

Let $h(u)$ be constantly equal to $(\beta - \alpha)^{-1}$ in the interval (α, β), and constantly 0 elsewhere. Then by (12),

$$w(f) = \frac{a}{\pi^2 f^2}\left[1 - \frac{1}{\beta - \alpha}\int_\alpha^\beta \cos 2\pi f t\, dt\right]$$
$$= \frac{a}{\pi^2 f^2}\left[1 - \frac{\sin 2\pi f\beta - \sin 2\pi f\alpha}{2\pi f(\beta - \alpha)}\right].$$

Now we see that

$$f(y) = \int_y^\infty h(u)\, du = \begin{cases} 1 & \text{for } y \leqq \alpha \\ \dfrac{\beta - y}{\beta - \alpha} & \text{for } \alpha \leqq y \leqq \beta \\ 0 & \text{for } y \geqq \beta \end{cases}$$

so that

$$R(\tau) = \begin{cases} a\left[\alpha - \tau + \dfrac{\beta - \alpha}{2}\right] & 0 \leqq \tau \leqq \alpha \\ \dfrac{a}{2}\dfrac{(\beta - \tau)^2}{\beta - \alpha} & \alpha \leqq \tau \leqq \beta \\ 0 & \tau \geqq \beta \end{cases} \tag{15}$$

is the covariance function of the process $N(t)$ when holding-time is distributed uniformly in (α, β).

If, formally, we let $(\beta - \alpha)$ approach 0 while keeping $\frac{1}{2}(\alpha + \beta)$ fixed, then the holding-times become concentrated in the neighborhood of the mean, h; in the limit, as $h(u)$ tends to a singular normal distribution with mean, h, and variance zero, we obtain

$$w(f) = \frac{a}{\pi^2 f^2}[1 - \cos 2\pi f h] \tag{16}$$

as the spectral density function for the $N(t)$ process with constant holding-time, $h = \frac{1}{2}(\alpha + \beta)$. Similarly, from (15), we note that as the holding-times become singularly normal with mean, h, and variance zero, the covariance function becomes

$$R(\tau) = \begin{cases} = a(h - \tau) & 0 \leqq \tau \leqq h \\ = 0 & \tau \geqq h. \end{cases}$$

We can express (16) as

$$w(f) = 2ah^2 \left(\frac{\sin \pi fh}{\pi fh}\right)^2,$$

and note that this is exactly like the power spectrum of a random telegraph wave constructed by choosing values $+\sqrt{ah}$, $-\sqrt{ah}$ with equal probability and independently for each interval of length, h. (Cf. Rice,[4] page 327.)

REFERENCES

1. J. L. Doob, Stochastic Processes, John Wiley and Sons, New York, 1953.
2. C. Palm, Intensitätsschwankungen im Fernsprechverkehr, Ericsson Technics, **44,** pp. 1–189, 1943.
3. F. W. Rabe, Variations of Telephone Traffic, Elec. Commun., **26,** pp. 243–248, 1949.
4. S. O. Rice, Mathematical Analysis of Random Noise, B.S.T.J., **23,** pp. 282–332, 1944, and **24,** pp. 46–156, 1945.
5. J. Riordan, Telephone Traffic Time Averages, B.S.T.J., **30,** pp. 1129–1144, 1951.

High-Voltage Conductivity-Modulated Silicon Rectifier

By H. S. VELORIC and M. B. PRINCE

(Manuscript received May 1, 1957)

Silicon power rectifiers have been made which have reverse breakdown voltages as high as 2,000 volts and forward characteristics comparable to those obtained in much lower voltage devices. It is shown that the magnitude and temperature dependence of the currents can be explained on the basis of space-charge generated current with a trapping level 0.5 eV below the conduction band or above the valence band. The breakdown voltage of a P^+IN^+ diode is computed from avalanche multiplication theory and is shown to be a function of the width of the nearly intrinsic region. A simple diffusion process is evaluated and shown to be adequate for diode fabrication. The characteristics of devices fabricated from high-resistivity compensated, floating-zone refined, and gold-doped silicon are presented. The surface limitation to high inverse voltage rectifiers is discussed.

I INTRODUCTION

The desire for high voltage rectifiers in the electronic industry has pushed the peak inverse voltage of solid state rectifiers to higher and higher values. The purpose of this paper is to present some of the considerations necessary in designing a device with a high inverse voltage and an excellent forward characteristic. In many cases the device characteristics are predictable. Conversely, high voltage diodes are excellent tools for studying many solid state phenomena.

It has been shown[1] that it is possible by the use of the conductivity modulation principle to separate the design of the forward current-voltage characteristic from the reverse current-voltage characteristic of a silicon *p-n* junction rectifier. Units have been fabricated by the diffusion of boron and phosphorus into high resistivity material, that have reverse breakdown voltages in the range of 1,000 to 2,000 volts.

The reverse currents are of the order of a microampere per square centimeter at room temperature and increase approximately as the square

root of the applied voltage. The magnitude, voltage dependence, and temperature dependence of the reverse currents can be explained as due to space-charge generated current[2] with a trapping level 0.5 eV from either the conduction or valence band. These effects will be discussed in Section II.

In Section III the breakdown voltage and its dependence on the resistivity and width of the high resistivity region of the rectifier will be considered.

In the next section the forward current is discussed and explained by considering both a space-charge region generated current and a diffusion current that takes into account high levels of minority carrier injection.[3]

Device processing information is given in Section V, together with an evaluation of different sources of high resistivity silicon. The devices to be discussed in this paper have been processed with high resistivity p-type material, although some devices have been made with n-type material.

Finally, a discussion of some surface problems associated with high voltage rectifiers is given in Section VI.

Although this paper is entitled "High-Voltage Conductivity-Modulated Silicon Rectifier", the theoretical arguments are applicable to all semiconductor diodes. However, the experimental results have been limited by considering only high voltage diodes.

II REVERSE CURRENT-VOLTAGE CHARACTERISTIC

2.1 Theory

The simple theory[4] for a p-n junction yields an expression for the reverse saturation current density (I_0) which is:

$$I_0 = q\left[n_p\left(\frac{D_n}{\tau_n}\right)^{1/2} + p_n\left(\frac{D_p}{\tau_p}\right)^{1/2}\right], \tag{2-1}$$

where q is the electron charge, n_p is the equilibrium electron density in p-type material, p_n is the equilibrium hole density in n-type material, D_n and D_p are the diffusion constants for electrons and holes, and τ_n and τ_p are the minority carrier lifetimes for electrons and holes.

When reasonable numbers are substituted into (2-1), I_0 at room temperature is of the order of 10^{-10} amperes per square centimeter. This quantity doubles with every increase of 4° C. The theory also contains no voltage dependence of this current. Even when breakdown multiplication[5] is taken into account, there is essentially no voltage dependence

at voltages less than half the breakdown voltage. The magnitude and temperature and voltage dependences of measured diodes do not agree with these theoretical values at room temperatures.

Recently, Pell[6] has shown that the reverse currents at low temperatures in germanium, and at room temperatures in silicon, are dominated by space-charge generated current. The space-charge generated current density (I_{sc}) is given by

$$I_{sc} = q\, W\, G\, M, \tag{2-2}$$

where W is the width of space-charge region, G is the generation rate of hole-electron pairs in the space-charge region, and M is the breakdown multiplication ($M \sim 1$ except near the breakdown voltage). G is given by[2]

$$G = \frac{n_1 p_1}{\tau_{p0} n_1 + \tau_{n0} p_1}, \tag{2-3}$$

where n_1 and p_1 are the densities of electrons and holes respectively if the Fermi levels were at the energy level of the recombination centers, and τ_{n0} and τ_{p0} are the minority carrier lifetimes of electrons and holes respectively in heavily doped p-type and n-type silicon. This expression assumes constant generation over the space-charge region. Thus,

$$n_1 = N_c \exp \frac{q}{kT} (V_r - V_c) = n_i \exp \beta(V_r - V_i), \tag{2-4a}$$

and

$$p_1 = N_v \exp \frac{q}{kT} (V_v - V_r) = n_i \exp - \beta(V_r - V_i), \tag{2-4b}$$

where V_r is the recombination level above the valence band edge V_v, V_i is the midband intrinsic level, $\beta = q/kT$, N_c and N_v are the effective densities of states in the conduction and valence bands $\cong 2.4 \times 10^{19} (T/300)^3$, V_c is the conduction band edge, k is the Boltzmann's constant, and T is the absolute temperature.

Substituting (2-4) into (2-3), one obtains:

$$G = \frac{n_1}{2\sqrt{\tau_{n0}\tau_{p0}}} \frac{1}{\cosh \left[\beta(V_r - V_i) + \frac{1}{2} \ln \frac{\tau_{p0}}{\tau_{n0}} \right]}. \tag{2-5}$$

For the diffused silicon junctions under consideration, it has been found[7] that τ_{n0} equals 1.2×10^{-6} seconds and τ_{p0} equals 0.4×10^{-6} seconds. Also, $n_i = 3.74 \times 10^{15}\, T^{3/2} e^{-6250/T}$ and $V_i = 0.54$ volts. Using these

numbers, (2-6) becomes:

$$G = \frac{1.25 \times 10^{16} \left(\frac{T}{300}\right)^{3/2} e^{20.8(1-300/T)}}{\cosh\left[38.62\left(\frac{300}{T}\right)(V_r - 0.54) - 0.55\right]}, \tag{2-6a}$$

and

$$G = G_{300}(V_r) f(T, V_r), \tag{2-6b}$$

where

$$G_{300}(V_r) = \frac{1.25 \times 10^{16}}{\cosh\,[38.62(V_r - .54) - 0.55]}, \tag{2-7a}$$

and (2-7b)

$$f(T, V_r) = \left(\frac{T}{300}\right)^{3/2} e^{20.8(1-300/T)} \frac{\cosh\,[38.62(V_r - 0.54) - 0.55]}{\cosh\left[38.62\left(\frac{300}{T}\right)(V_r - 0.54) - 0.55\right]}.$$

In (2-6b), $G_{300}(V_r)$ is the generation rate for a recombination level at V_r equal to 300° K, and $f(T, V_r)$ is the temperature variation of G for a recombination level at V_r normalized to 300° K.

Curves of $f(T, V_r)$ are given in Fig. 1 for several values of V_r with a curve $g(T)$ which is the temperature variation of the reverse saturation current (I_0). Table I gives values for G_{300} for various V_r.

In the reverse biased diffused junctions made with high resistivity material, the junction may be considered abrupt. Therefore, the width (W) of the space-charge region, when the junction is reverse biased to a voltage V, is given by

$$W = \left(\frac{\kappa V}{2qN_A}\right)^{1/2} = 3.14 \times 10^{-5}(V\rho_p)^{1/2} \text{ cm} \tag{2-8}$$

where the units after the first equal sign are electrostatic, and κ is the dielectric constant. In the second expression, V is in volts, and ρ_p, the base material resistivity, expressed in ohm-centimeters. Thus,

$$I_{sc} = 4 \times 10^{-24} G_{300}(V_r) f(T, V_r)[V\rho_p]^{1/2} \text{ amperes-cm}^{-2}. \tag{2-9}$$

It is seen that I_{sc} varies theoretically as the square root of the reverse voltage for values of V less than $\frac{1}{2} V_B$, the breakdown voltage, in which range avalanche multiplication is negligible. The quantity I_{sc} varies inversely with $N_A^{1/2}$ and will be large for high voltage devices with small N_A. The I_{sc} at 300° K for a rectifier with 40 ohm-centimeter base ma-

terial and a reverse bias of 100 volts is given in Table I as a function of V_r. The numbers compare with 8×10^{-10} ampere per square centimeter for I_0. Thus, from diode measurements at room temperature and above, one could not observe V_T less than 0.3 eV from either the conduction or valence band. In fact, from a measurement of the temperature dependence of the reverse currents, one can determine only the recombination level lying closest to the center of the forbidden band. This can be seen more clearly from the following argument: There will be a contribution to the reverse current from the diffusion current $I_{o300}g(T)$ which varies with temperature as $g(T)$. There will be contributions to the reverse cur-

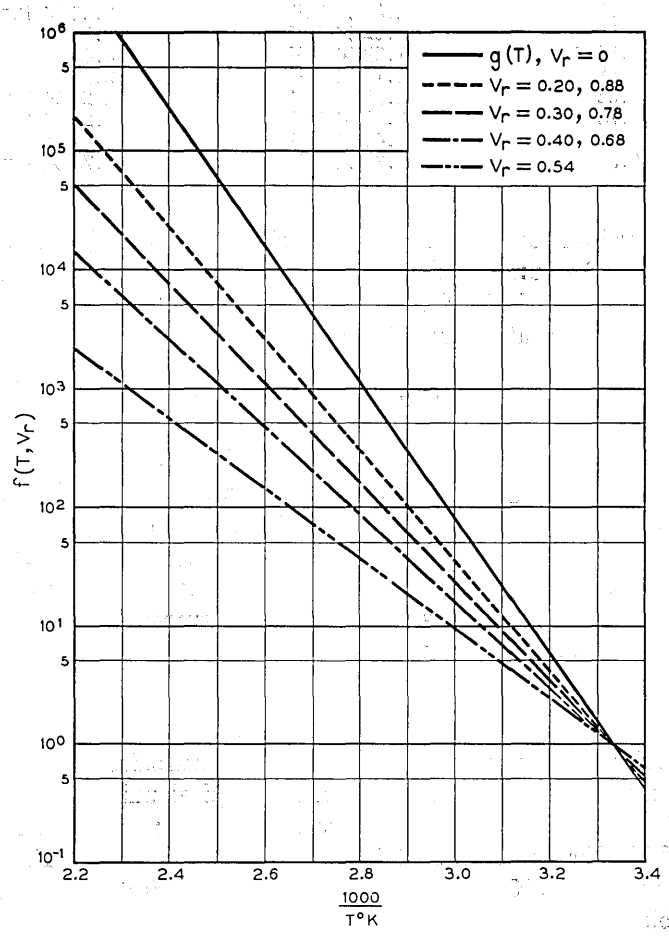

Fig. 1 — The temperature variation of the generation rate, $f(T, V_r)$, for several values of the recombination level, V_r.

TABLE I — VALUES OF G AND SPACE CHARGE GENERATED CURRENT AT 300° K FOR VARIOUS VALUES OF THE TRAPPING LEVEL V_r

$T = 300°$ K $\quad E_g = 1.08eV \quad np = 3 \times 10^{20}$ cm^{-6}

$\tau_{n0} = 1.2 \times 10^{-6}$ sec $\quad \tau_{p0} = 0.4 \times 10^{-6}$ sec

V_r Volts above Valence Band	G_{300} cm^{-3} sec^{-1}	I_{sc} ($V = -100$ volts, $\rho = 40$ ohm-cm) microamperes/cm^2
0.10	5.92×10^{8}	1.88×10^{-7}
0.20	2.84×10^{10}	9.03×10^{-6}
0.30	1.35×10^{12}	4.29×10^{-4}
0.40	6.5×10^{13}	2.06×10^{-2}
0.50	3.02×10^{15}	0.96
0.54	1.08×10^{16}	3.43
0.58	8.1×10^{15}	2.58
0.68	1.95×10^{14}	6.2×10^{-2}
0.78	3.96×10^{12}	1.26×10^{-3}
0.88	8.45×10^{10}	2.69×10^{-5}
0.98	1.78×10^{9}	3.75×10^{-7}

rent by the individual trapping centers given by $I_{sc300}(V_r)f(T, V_r)$, where $I_{sc300}(V_r)$ is the current at 300° K due to generation at recombination centers located at the level V_r, and $f(T, V_r)$ is the temperature variation of the generation rate. Thus, the total reverse current is given by

$$I_{\text{reverse}} = I_{o300}g(T) + \sum_{V_r} I_{sc300}(V_r)f(T, V_r), \tag{2-10}$$

where the summation is over all recombination levels. The relative currents at 300° K are given in Table I. The greatest contribution at 300° K is due to the level nearest the center of the forbidden band. As the temperature increases, all the terms under the summation sign approach each other. Before a second recombination level contributes significantly to the reverse current, however, the saturation current will have become the most important component.

2.2 Experimental Results

To evaluate the theory for the reverse currents in silicon N^+P junctions, careful measurements were made on five typical units for the reverse current-voltage characteristics at various temperatures from 300° K to 435° K. The curves were taken with a X-Y recorder. The voltage ranged from 0 to 200 volts so that multiplication effects were completely negligible.

From the recorded data, curves of I_r versus $1{,}000/T°$ K were plotted for $V = -10$, -40, and -160 volts. The set of curves for diode No. 3

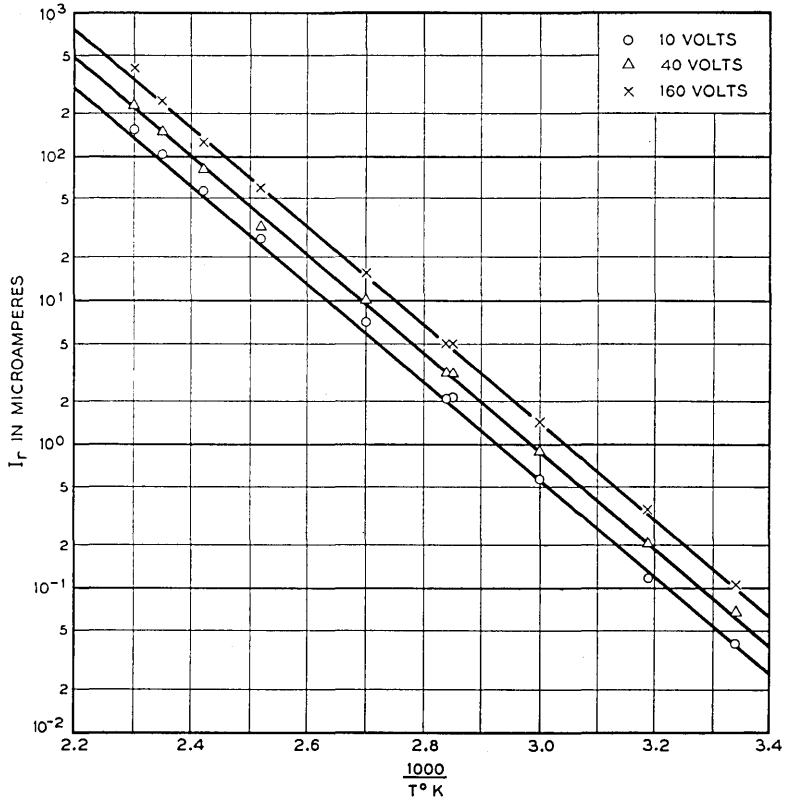

Fig. 2 — The temperature variation of "reverse current" for a typical diode at −10, −40, and −160 volts.

is given in Fig. 2. The slope of these curves indicates that the recombination level lies near 0.5 eV below the conduction band or above the valence band. The junction area of this device is 0.015 cm^2; thus, the current density at 300° K and at −100 volts is 4.4 microamperes per square centimeter. This compares with the order of one microampere as listed in Table I. This suggests that the $(\tau_{n0}\tau_{n0})^{\frac{1}{2}}$ is overestimated. The agreement of this measurement with theory is reasonable.

The voltage variation of the reverse currents does not agree with theory as well as the magnitude and temperature dependence. The experimental results give, as the voltage dependence, an expression:

$$I_r \sim V^{1/N},$$

where N equals 2.9. This compares with the theoretical value of $N = 2$.

Some of this discrepancy can be attributed to the fact that the junction is not truly an abrupt junction. A "graded" junction would yield N equals 3. Measurements of capacitance versus voltage, which essentially measure the width of the space-charge region, yield N equals 2.4. Thus, these devices in the relatively low voltage range still have some contribution from the gradation of the diffused junction.

The highest temperature points in Fig. 2 deviate above the straight lines. This deviation can be attributed to the onset of the contribution from the I_0 component. Calculations indicate that, by $T = 220°$ C, the contribution to the reverse current by the space-charge current is equaled by the saturation current and that, by $T = 320°$ C, the space-charge-generated current is negligible compared to the saturation current.

III BREAKDOWN VOLTAGE OF PN AND PIN JUNCTIONS

3.1 Theory

It has been demonstrated that, in germanium[8, 9] and silicon, reverse biased junctions breakdown as a result of a solid state analogue of the Townsend β Avalanche Theory. Multiplication and breakdown occur when electrons or holes are accelerated to energies sufficient to create hole-electron pairs by collisions with valence electrons. The breakdown phenomena in silicon for graded and step junctions has been previously considered.[8, 10] Depending on the impurity distribution, the field in the junction will be a function of distance and will have a maximum value in the region of zero net impurity concentration. The breakdown voltage is a critical function of the space-charge distribution.

In this section the existing multiplication theory is extended to the case of PIN junctions. It is shown that relatively wide intrinsic regions are required to obtain breakdown voltages greater than 1000 volts.

Fig. 3 is a plot of the impurity, charge, and field distributions in PIN and PπN junctions. Fig. 3(a) schematically illustrates the geometry of the three region devices considered, and Fig. 3(b) is a plot of the impurity distribution. In this analysis step junctions will be assumed. For the PIN junction there are no uncompensated impurities in the intrinsic region, and no net charge. At low reverse voltage, the field will sweep through the intrinsic layer and will increase with increasing reverse bias until the breakdown field is reached.

Absolutely intrinsic material is not yet available, and devices are made from high resistivity π-type material. In this class of devices there is some uncompensated impurity and charge in the center region. The field will have a maximum value at the $N^+ \pi$ junction and will decrease

with increasing distance into the π region. At sufficiently high reverse bias the field may sweep into the P^+ region.

Breakdown in silicon[8] is a multiplicative process described by

$$1 - \frac{1}{M}\int_0^W \alpha_1 \, dx, \tag{3-1}$$

where M is the multiplication factor, W is the space-charge width, and α_1 is the rate of ionization which is a strong function of the field in the junction. For a PIN structure, the field is constant, at breakdown M approaches ∞, and

$$\alpha_1 W = 1. \tag{3-2}$$

The ionization rate at breakdown is then a simple function of the width of the intrinsic region. McKay[8] and Wolf[11] have considered α_i as a func-

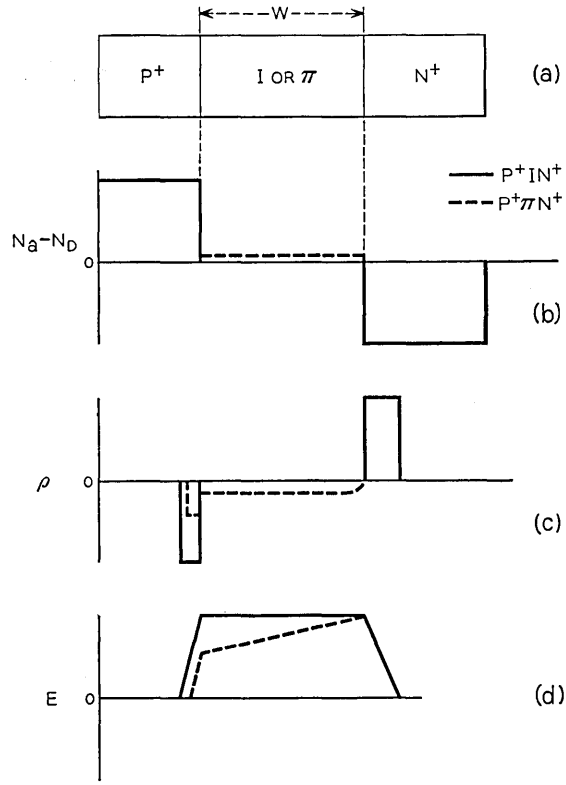

Fig. 3 — Impurity, charge and field distribution in PIN and PπN junctions.

tion of the field. If α_1 is fixed, the field at breakdown can be determined. The breakdown voltage is $\int_0^W Edx$.

Fig. 4 is a plot of breakdown voltage as a function of space-charge width for PN and PIN diodes. The PIN values are calculated; the PN data is previously unpublished data supplied by K. G. McKay.

Some interesting observations can be made from Fig. 4:

1. The plot of breakdown voltage versus barrier width for a PN step junction assumes that the space-charge region does not extend through the high resistivity side of the junction. For this class of junctions the breakdown voltage is determined by the impurity concentration as shown in Fig. 5. The plot of breakdown versus space-charge width for a PIN diode assumes that the space-charge region extends from the P to N region at very low bias, and that it is limited by the width of the I region. If a constant field is assumed in the I region, the breakdown voltage is a function of the barrier width.

2. Although the space-charge region can reach through the I region at low bias, the avalanche breakdown voltage is a function of the width of the I region.

3. For the devices considered here with π regions in the order of 10^{-2} cm, the maximum breakdown voltage is in the order of 2,000 volts.

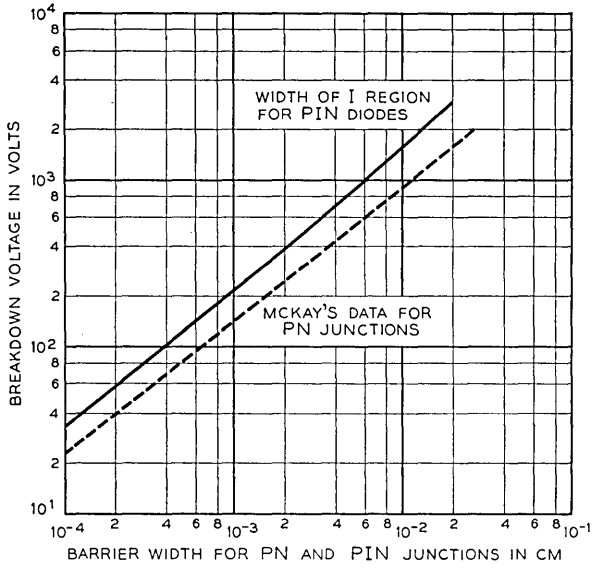

Fig. 4 — Breakdown voltage as a function of barrier width for PN and PIN junctions.

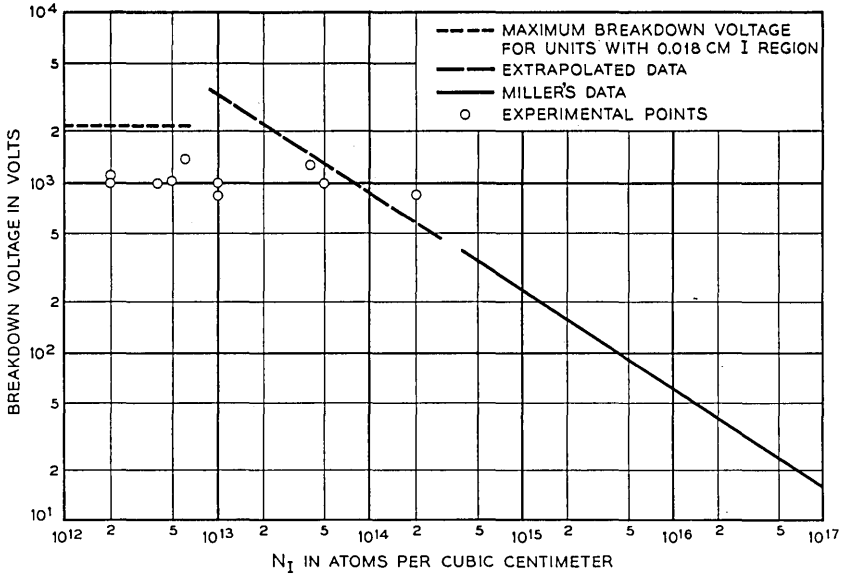

Fig. 5 — Breakdown voltage versus impurity concentration for silicon step junctions.

3.2 Experiment

Fig. 5 is a plot of breakdown voltage versus impurity concentration for silicon step junctions. The plot above 300 volts is extrapolated from the data of Miller[10] and Wilson.[9]

Capacity data, discussed in Section V, indicates that many devices show body breakdown. A few rectifiers break down at voltages as high as 2,000 volts. In many high voltage devices the breakdown voltage is not limited by geometry but by surface problems.

IV FORWARD CURRENT-VOLTAGE CHARACTERISTIC

4.1 Theory

It will be shown in this section that the forward current-voltage characteristic as well as the reverse characteristic can be completely explained by considering both a space-charge region generated current and a diffusion current. The diffusion current component must also take into consideration the effect of high injection levels of minority carriers.

According to the Shockley-Read[2] theory, the rate of recombination, U, of holes and electrons in a semiconductor is given by:

$$U = -G = \frac{pn - n_i^2}{\tau_{p0}(n + n_1) + \tau_{n0}(p + p_1)} \tag{4-1}$$

where p, and n are the instantaneous concentrations of holes and electrons, respectively. When a PN junction is forward biased, holes and electrons are injected into the space-charge region which has been reduced in width. Some of these carriers diffuse through the space charge region and give rise to the normal diffusion current when the excess minority carriers recombine with majority carriers in field free regions. The other carriers recombine according to (4-1) in the space-charge region giving rise to what is called the space-charge generated current. In the reverse biased junction, the current is due to carriers generated in the space-charge region; whereas, in the forward biased junction, the current is due to recombination of carriers. The quantity U is large in the space-charge region since both p and n are large in this region. In the field free regions, however, one of these quantities is usually small and the product deviates only slightly from n_i^2.

The space-charge generated current, I_{sc}, is given approximately by:[12]

$$I_{sc} = \frac{2qWn_i}{(\tau_{n0}\tau_{p0})^{1/2}} \frac{\sinh \beta \frac{V}{2}}{\beta(V_B - V)} f(b), \tag{4-2}$$

where V_B is the built-in potential of the junction, and $f(b)$ is discussed in Reference 12 and is approximately 1.5 for recombination centers near the intrinsic level as is the case for the diodes under consideration. For shallower recombination levels the function $f(b)$ is much smaller and depends strongly upon the forward applied voltage.

For the forward-biased junction, the space-charge region is narrow, the concentration gradient can be considered linear and W is given by the following expression:[4]

$$W = 4.35 \times 10^2 \left(\frac{V_{\text{junction}}}{\alpha}\right)^{1/3} \text{cm}, \tag{4-3}$$

where V_{junction} is the total potential across the junction in volts and α is the concentration gradient at the junction in cm^{-4}. These are given by:

$$\begin{aligned} V_{\text{junction}} &= V_{\text{built-in}} - V \\ &= kT/q \ln (N_A N_D/n_i^2) - V \\ &= 0.792 - V. \end{aligned} \tag{4-4}$$

$$\text{Also, } \alpha = \frac{C_0}{\sqrt{\pi Dt}} e^{-x_j^2/4Dt} \text{ for diffused junctions,} \tag{4-5}$$

where C_0 = surface concentration of diffusant = 3×10^{19} cm^{-3},

D = diffusion constant = 3×10^{-12} cm^2/sec,

t = diffusion time = 5.7×10^4 sec,

x_j = junction depth below surface = 0.003 cm.

When these numbers are substituted into the equations, at 300° K:

$$W = 9.25 \times 10^{-4}(0.792 - V)^{1/3} \text{ cm}. \tag{4-6}$$

For the diodes under consideration:

$$\tau_{n0} = 1.2 \times 10^{-6} \text{ sec}, \qquad \tau_{p0} = 0.4 \times 10^{-6} \text{ sec}.$$

When these expressions are substituted into (4-2), one obtains at 300° C:

$$I_{sc} = 2.8 \times 10^{-7} \frac{\sinh 19.31V}{(0.792 - V)^{2/3}} \text{ amp/cm}^2. \tag{4-7}$$

In order to fit the experimental data, it is necessary to multiply (4-7) by a factor of 5. This may be due to an overestimation of $(\tau_{n0}\tau_{p0})^{\frac{1}{2}}$. Therefore, the equation which shall be used in the remainder of this section will be:

$$I_{sc} = 1.4 \times 10^{-6} \frac{\sinh 19.31V}{(0.792 - V)^{2/3}} \text{ amp/cm}^2. \tag{4-8}$$

A plot of this expression is given in Fig. 6.

The normal diffusion current for low level diffusion,[4] I_{DL}, is given by

$$I_{DL} = I_0(e^{qV/kT} - 1) \tag{4-9}$$

where I_0 is given by (2-1). I_0 for the diodes under discussion is approximately 8×10^{-10} ampere/cm^2 at 300° K. When the injected minority carrier density approaches the equilibrium majority carrier density, the form of (4-9) changes. The high injection level diffusion current, I_{DH}, is then given by[3]

$$I_{DH} = I_{DH0}(e^{qV/2kT} - 1), \tag{4-10}$$

where I_{DH0} equals qn_is/τ, and s equals the width of the high resistivity region. For the diodes under discussion, I_{DH0} is approximately 2×10^{-6} amperes/cm^2 at 300° K. A current-voltage plot of these currents at 300° K for $V_r = 0.50$ is given in Fig. 6 together with their sum. It can be observed that the resulting characteristic starts with slope of qV/kT and bends over to a slope of $qV/2kT$ near 0.10 volt. The slope increases again to near qV/kT at 0.35 volts and decreases once more to $qV/2kT$ above 0.40 volts giving a bump to the over-all characteristic.

Next, consider the temperature dependence of the coefficients of the forward current components

$$I_{sc0} \sim n_i(T), \tag{4-11a}$$

$$I_0 \sim n_i^2 \frac{D_n(T)^{1/2}}{\tau_n(T)} \sim n_i(T)^2, \tag{4-11b}$$

and

$$I_{DH0} \sim \frac{n_i(T)}{\tau(T)} \sim n_i(T). \tag{4-11c}$$

The largest variation of these coefficients is due to the variation of

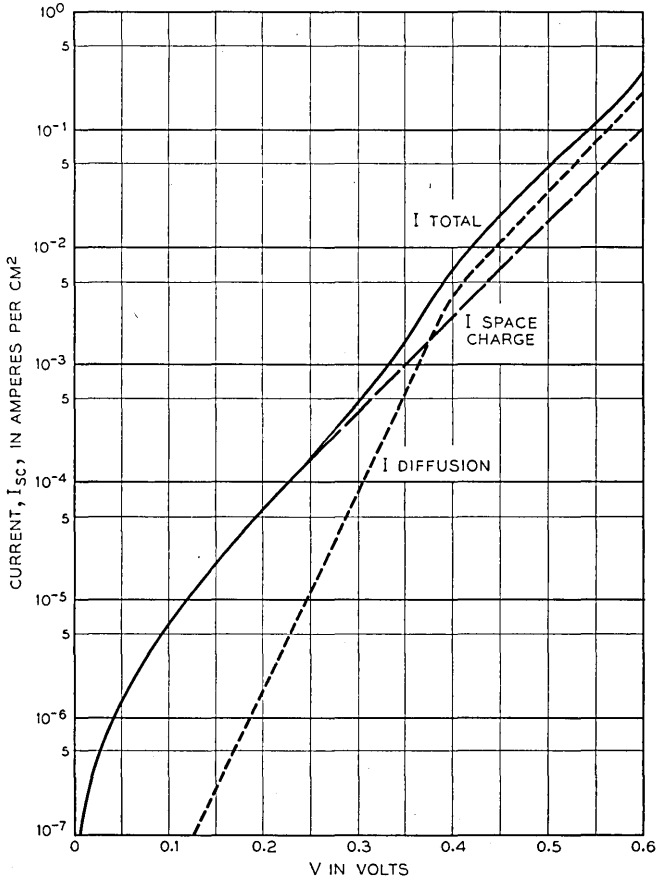

Fig. 6 — The two components of current for a forward biased junction.

n_i with temperature. Fig. 7 gives a plot of this variation. The temperature variations of the other parameters are all small compared to that of n_i. Thus, as in the case of the reverse currents, at sufficiently high temperatures, the diffusion current makes the more important contribution.

In the case of the forward current, I_{sc} is relatively insensitive to the distribution of impurities; therefore, the results of this section are important for all forward-biased diodes. In high-voltage diodes, to keep the resistive voltage drop small, it is necessary to maintain high minority carrier lifetime in the center region. The diffusion length of injected

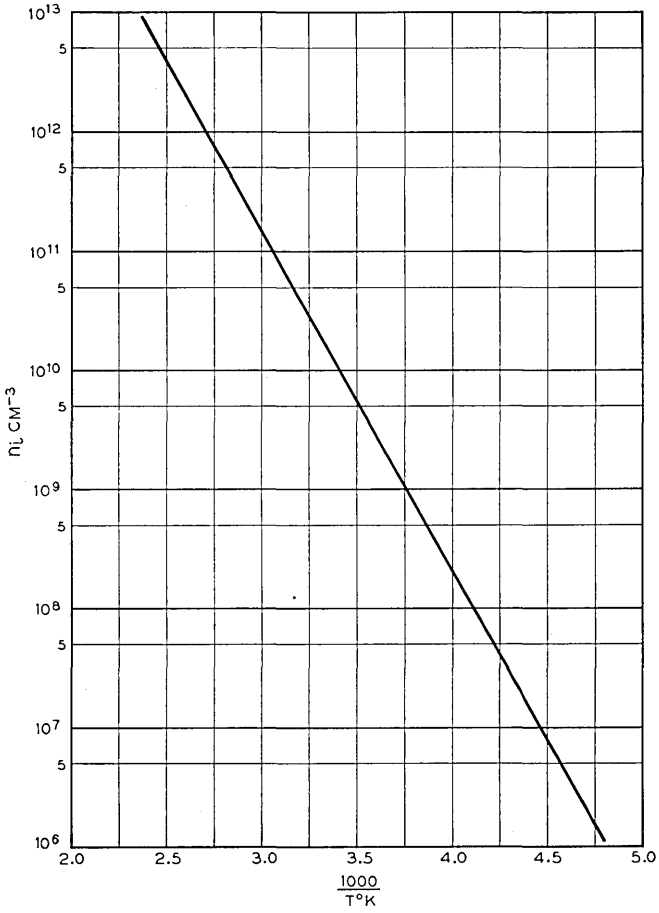

Fig. 7 — The variation of n_i with temperature.

minority carriers should be the order of or larger than the center region width.

4.2 Experimental Results

The forward characteristics of the five typical high voltage rectifiers mentioned in Section 2.2 were measured with a *X-Y* recorder, and all showed similar shapes. Diode No. 3 will be discussed in detail in this section.

The forward current-voltage characteristic was measured at three temperatures: 220° K, 300° K, and 375° K. Measurements below cur-

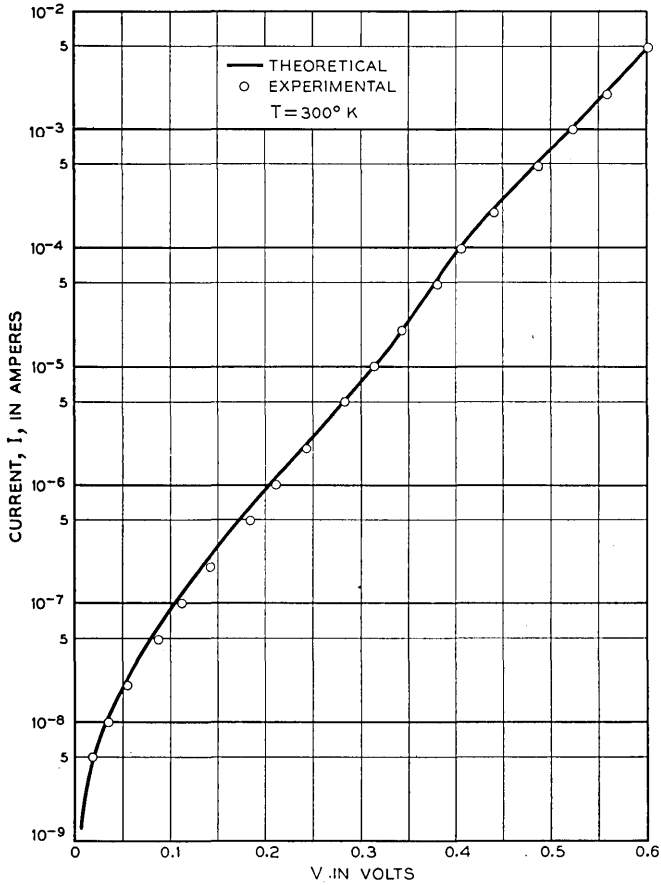

Fig. 8 — The calculated and observed current-voltage characteristic of a forward biased junction at 300° K.

rents of one microampere were made only at 300° K. Currents above 10 milliamperes were not measured since internal power losses would cause temperature variations. The unit has a junction area of 0.015 cm^2, junction lifetime of 4 microseconds at 300° K, S equal to 0.008 cm, and N_A equal to 3×10^{14} cm^{-3}. When these numbers are substituted into the expressions for the coefficients, one obtains, at 300° K,

$$I_{sc0} = 1.4 \times 10^{-8} \text{ amperes},$$

$$I_0 = 1.2 \times 10^{-11} \text{ amperes},$$

$$I_{DH0} = 3.0 \times 10^{-8} \text{ amperes}.$$

Fig. 8 shows a semilogarithmic plot of the current-voltage characteristic at 300° K over a range of $6\frac{1}{2}$ decades. The circles represent measured points and the solid line is the theoretical curve.

Using the variation of n_i given in Fig. 8 and the temperature variations as given in (4-11), one can obtain the coefficients for any temperature. This has been done for two temperatures, 220° K and 375° K. Figs. 9 and 10 show the theoretical and experimental plots at 375° K and 220° K

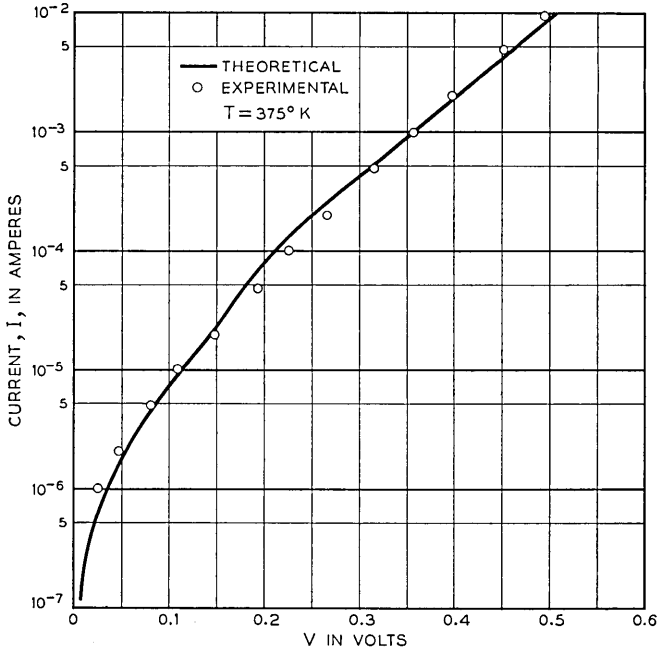

Fig. 9 — The calculated and observed current-voltage characteristic of a forward biased junction at 375° K.

respectively. The circles represent measured points and the solid lines are the calculated theoretical curves. It is observed that the fit in Fig. 9 is quite good; whereas, the fit in Fig. 10 is not as good as at the other temperatures. However, even this figure shows good qualitative agreement of the deviation from a straight line. Some of the factor of two discrepancy in Fig. 10 can be ascribed to the temperature variation of the other parameters, and some to a possible error in the measurement of temperature which would be reflected in the value of n_i.

It should be noted that at all temperatures the IR drop in the high resistivity region is not observable to the limits of the experimental measurements of forward current, 10 milliamperes. This is due to the fact that the region has been conductivity modulated by the forward current. This requires a sufficient minority carrier lifetime in the region so that most of the injected carriers diffuse across the region before recombining. Such lifetimes can be maintained in diffused junctions[7]

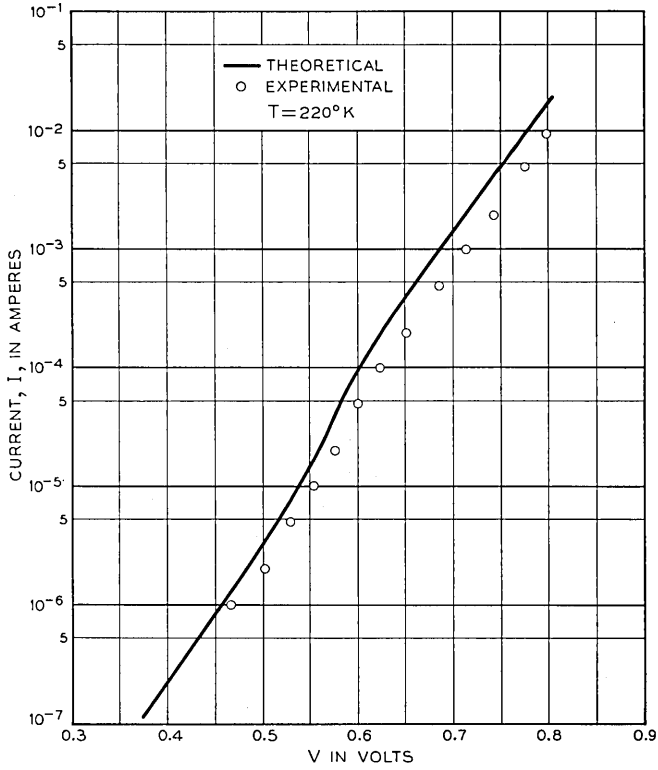

Fig. 10 — The calculated and observed current-voltage characteristic of a forward biased junction at 220° K.

to permit the high resistivity region to be at least as wide as 0.025 centimeters. Thus even in high voltage rectifiers it is still possible to design the forward and reverse current voltage characteristics independently.

V DEVICE PROCESSING

5.1 Silicon Material

Fig. 5 shows that step junctions which break down at over a thousand volts must have a background impurity concentration $\leqq 10^{14}$ atoms/cm^3. The highest grade commercial semiconductor silicon has 5×10^{14} impurities/cm^3 (20–50 Ω cm P type). This material must be processed to reduce the impurity level. To date, high voltage devices have been processed from four types of high resistivity material: floating zone refined, compensated, gold diffused, and horizontal zone refined silicon.

Some silicon was prepared by adding N-type impurities to reduce $| N_D - N_A | < 10^{14}$. Maintaining this delicate balance in material where $N_D \simeq N_A$ is difficult. The boron is relatively uniformly distributed since the distribution constant is close to unity. N-type impurities are less uniformly distributed in the crystal since the distribution constants are considerably less than unity. High resistivity compensated silicon is full of N- and P-region striations. The units processed from this material generally had poor electrical characteristics.

Table II is a typical contour of a compensated crystal. The resistivity varies around the crystal and changes along the length of the crystal. At the bottom of the crystal the resistivity goes through a maximum. The tail end is converted from P to N type.

A number of devices have been fabricated from silicon processed with

TABLE II — A TYPICAL CONTOUR OF A HIGH RESISTIVITY COMPENSATED CRYSTAL

Crystal A-161, Oriented 111, Rotated $\frac{1}{2}$ RPM

Distance from seed (inches)	Resistivity (Ω cm) at Angle				Impurity Type
	0°	90°	180°	270°	
$\frac{1}{2}$	28	33	23	30	P
$\frac{3}{4}$	25	31	31	32	P
$1\frac{1}{2}$	41	22	27	34	P
$1\frac{3}{4}$	57	51	63	37	P
2	160	160	87	200	P
$2\frac{1}{4}$	510	520	—	—	—
$2\frac{1}{2}$	—	—	1200	—	—
$2\frac{3}{4}$	2.9	1.2	0.8	0.8	N

TABLE III — THE CHARACTERISTICS OF SOME HIGH VOLTAGE RECTIFIERS PROCESSED FROM GOLD DIFFUSED AND ZONE REFINED SILICON

Units[1]	E_R (volts)[2]			E_F (volts)[3]			τ (μsec)[4]	Silicon-Type
	10μa	100μa	1 ma	10 ma	100 ma	1A		
Me-512	30	120	500	0.8	1	1.3	1.6	Gold diffusion $\rho \sim 16{,}000\ \Omega$ cm
513	22	200	600	1.0	1.2	1.5	0.6	
514	300	600	1000	0.8	1	1.5	2.1	
515	300	500	1000	0.8	1	1.4	1.2	
Me-375	1200	1500		2.5	3.5		<1	Floating zone refined $\rho \sim 6{,}000\ \Omega$ cm
376	70	300	800	2.5	4.0		<1	
377	16	120	700	3.5	7			
378	320	400	800	2.5	3.5			

[1] These units have an area $\sim 10^{-3}$ cm^2.
[2] This is the reverse voltage at which these units pass the indicated current.
[3] This is the forward bias at which the units pass the indicated current.
[4] This is the lifetime measured at 30 milliamperes forward current by the pulse injected technique. The lifetime did not seem very sensitive to small variations in injected current.

the floating zone apparatus.[13] This technique removes impurities from molten silicon by treatment with hydrogen containing water vapor. The material obtained from this process has an impurity level in the range of 10^{12} to 5×10^{13} acceptors/cm^3 (2,000 to 16,000 Ω cm P type).

Table III gives the characteristics of some of the better diodes made from such floating zone silicon. The reverse currents are larger than that predicted by (2.10). The lifetime at high injection is in the order of 1 μsec.

N-type silicon with a resistivity range of 10 to 30 Ω cm was diffused with gold at 1,200° for sixteen hours. With this diffusion program the gold is uniformly distributed in the material.[14] The resistivity after gold diffusion was in the range of 2,000 to 15,000 Ω cm. The characteristics of several devices processed from this material are given in Table III. This technique has many attractive features; however, additional work was not done because the lifetime in the diffused material was consistently lower than that required for conductivity modulation.

One successful purification technique is horizontal zone refining of silicon in a quartz boat. With the number of passes used, the background acceptor concentration is observed to be in the order of 5×10^{13} to 10^{14} (100–1,000 Ω cm P type). Most of the devices reported in this paper are fabricated from this material.

Capacity data for devices fabricated from various types of high resistivity silicon is shown in Fig. 11. The plot shows that the high resistivi-

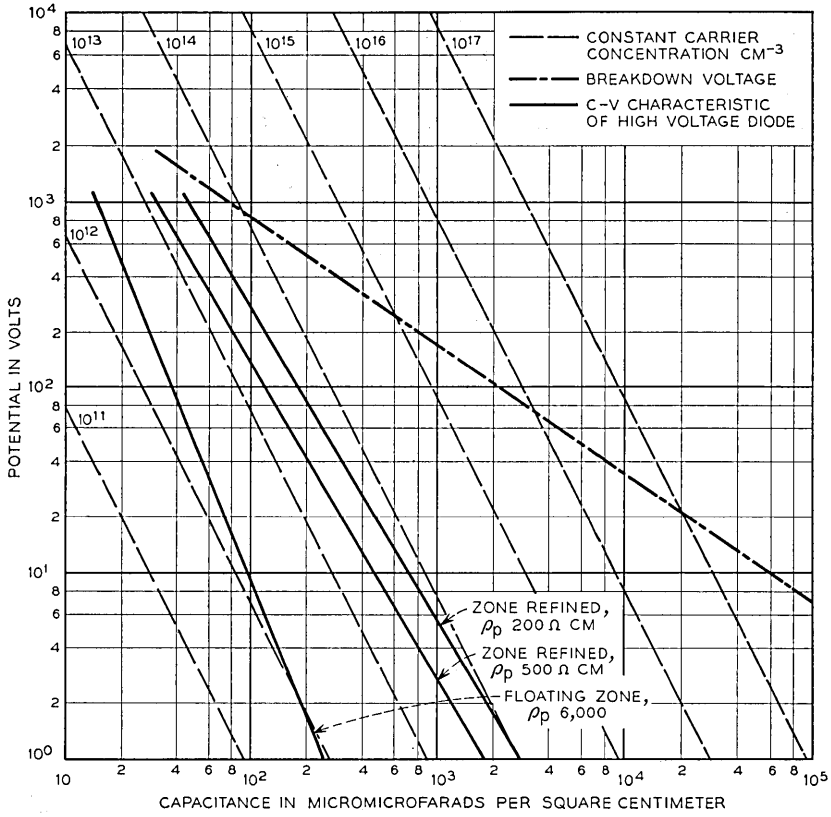

Fig. 11 — Capacity/cm² for high voltage rectifiers processed from various types of high resistivity material.

ties measured by the four point probe before diffusion are indicative of the impurity level after processing. The water vapor floating zone refined material has an impurity level of 10^{12} acceptors/cm^3; the other material is in the range of 10^{13} to 10^{14}. The breakdown voltage line was calculated from the data in Fig. 5.

5.2 Diffusion

In this section some of the practical difficulties observed in utilizing the diffusion technique will be considered. In the fabrication of transistors close geometry control is necessary in order to obtain the desired device characteristic. It has been shown[1] that in conductivity modulated rectifiers the only geometry requirement is that the width of the center

region be less than the diffusion length of the minority carrier. High surface concentration of diffusant is desirable since this facilitates the contact problem. This suggests that the diffusion system can be much less involved than that required for diffused transistors.[15] Some of the data presented in this section will show that the open tube diffusion technique[16] can lead to variations in diffusion parameters.

The diffusion of impurities into silicon is complicated by variations in the boundary conditions at the surface. Frosch[15] has shown that surface concentrations can be varied over six decades.

5.2.1 Device Diffusion Theory

Several important impurity distributions have been considered[17]. Two distributions are important in the open tube process:

1. Error Function Complement, ERFC, Distribution or Infinite Diffusant Source. If the diffusant is deposited on the silicon and serves as an infinite source, the added impurities will have an *erfc* distribution. For one diffusant and a fixed diffusion program this distribution will result in the deepest penetrations and smallest sheet resistances of all possible distributions. The sheet resistance is a measure of the total number of added impurities. The data presented later indicates that the added impurities frequently have an *erfc* distribution.

2. Gaussian and Modified Gaussian Distribution. A number of impurity atoms enters the solid, and a surface barrier builds up with time which prevents additional atoms from entering.[17] Initially, the diffusant is assumed to be present in an infinitely thin layer at the surface with diffusion into or out of the material possible. In the range of silicon doping levels and surface concentrations used, a Gaussian, modified Gaussian or *erfc* distribution for a given diffusion program lead to approximately equal junction depths.

The sheet resistance and the diffusion depth have been related[18] to the surface concentration for an *erfc* distribution. If the distribution is Gaussian instead of *erfc*, then for the same value of sheet resistance and diffusion depth the surface concentration should be reduced by one-third. Since the sheet resistance is related to the total number of impurities through a mobility term, quantitative interpretation of the data for any case other than *erfc* or Gaussian distributions would be difficult.

5.2.2 Experimental Results

The sheet resistance was measured by the four-point probe method, and the diffusion depths, by angle lapping and staining.[19] Surface con-

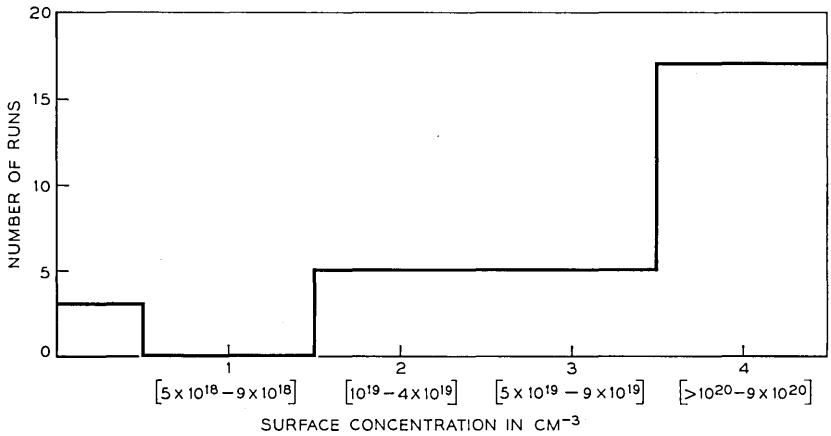

Fig. 12 — Distribution of surface concentration of 28 P_2O_5 diffusions by the open tube deposition technique.

centrations were calculated assuming an *erfc* distribution.[18] All the diffusions are on lapped silicon surfaces in the temperature range of 1,200 to 1,300° C.

Fig. 12 shows the distribution of surface concentration of 28 P_2O_5 diffusions by the open tube process.[16] The surface concentrations vary from 10^{19} to 5×10^{20} atoms/cm^3. These values are about a decade lower than the closed tube values of surface concentrations reported by Fuller.[19]

The measured diffusion depths were in the order of 2×10^{-3} to 5×10^{-3} cm. Fig. 13 shows the distribution of diffusion depths normalized with the calculated diffusion depth as unity. The diffusion depths were calculated from the measured surface concentration assuming an *erfc*[19] distribution.

The observed variation in diffusion depth is difficult to explain. Some of the possibilities which have been considered are:

1. The diffusion temperature from lot to lot would have to be from 0 to 50 degrees below the expected value to explain the variations. Discrepancies this large have not been observed.

2. One impurity distribution which may explain some of the results is a modified Gaussian with considerable out diffusion. There are some runs with high sheet resistance and diffusion depths which are consistent with this picture. Generally the sheet resistances are so small that there could not be much out diffusion.

3. Some workers have suggested the possibility of the diffusion constant being a function of the surface concentration. Fig. 13 does not

indicate any correlation between surface concentration and diffusion constant.

The variations in diffusion process control have not been observed to effect the production of rectifiers. If better geometry control is necessary, more sophisticated diffusion techniques are required.

VI PULSE PROPERTIES AND RELIABILITY

Important considerations in all diode applications are the pulse properties and reliability in operation. In this section some problems which are associated with avalanche breakdown are described and the results related to recent work on surface and body breakdown.

6.1 Theory

Several workers[20] have considered the possibility of a negative resistance in the avalanche region for reverse biased junctions in which one side is either intrinsic or so weakly doped that the space charge of the carriers cannot be neglected. A negative resistance might be observed at very high current densities in an $N^{+}\pi$ junction.

One possible source of a negative resistance would be a large temperature rise due to current concentration at a few points instead of a uni-

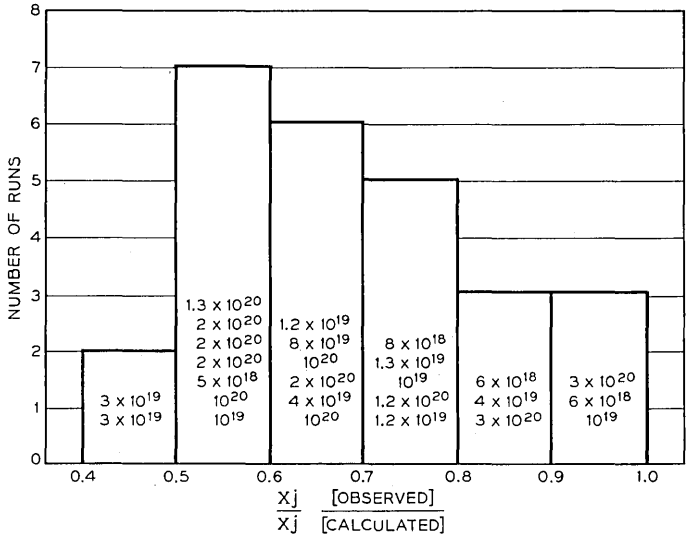

Fig. 13 — Distribution of diffusion depths for diffusion by the open tube deposition technique.

form flow through the junctions. This case is of particular significance in high voltage rectifiers where small reverse currents result in relatively large power. It has been pointed out in Sections 3.2 and 5.1 that body avalanche breakdown is frequently not observed in these devices.

Avalanche breakdown current in silicon[8] is carried by discrete pulses of about 50 μa at their onset and increasing with increasing current to about 100 μa. Approximate calculations[21] show that the ionizing regions of these microplasmas are about 500 Å in extent, have a current density $\approx 2 \times 10^6$ amp/cm^2, and have a net space-charge density $\approx 10^{18}$/cm^3. These pulses for junctions with E_{max} less than 500 kv/cm appear to be independent of junction width and built-in space-charge. Rose considers the statistical problem associated with a large number of pulses and presents a picture which is consistent with most of the experimental data. He calculates the temperature rise, assuming the avalanche power is 1×10^{-2} watts and is dissipated uniformly in a sphere. The maximum temperature rise for a cluster of two or three pulses is in the order of 25° C. For the picture Rose presents, the temperature rise due to the microplasma should be relatively insensitive to the breakdown voltage. Thermal collapse of rectification, i.e., increase of temperature until the silicon is intrinsic, will probably not occur in the region of avalanche multiplication. Two important conclusions can be obtained:

1. Avalanche breakdown should occur as a random process with a uniform probability over the junction. Large temperature rises due to a breakdown of microplasma will probably not occur since the resulting temperature rise would cause the breakdown voltage in that spot to increase. The power is dissipated throughout the path of the current pulse in the space-charge region.

2. A thermal effect in silicon due to heating by the small plasma has a very short time constant of the order of 10^{-10} seconds.[21] It is not possible to separate a thermal effect of this type by reducing the pulse width. The heating and cooling time is short compared to the pulse time in these experiments.

The pulse properties of a junction would be quite different if the breakdown occurred at one spot instead of many spots distributed over the junction. Breakdown at a single spot on the surface has been observed.[22]

6.2 Experimental Results

Many rectifiers were given a voltage pulse which carried them into breakdown. There was a wide distribution of V-I characteristics. Many diodes did not show a negative resistance up to the maximum instan-

taneous power the pulser could deliver, 5kw. These diodes are not considered in the subsequent analysis.

The diodes were subjected to 50 μsec triangular voltage pulses which would send them into breakdown. Variations in pulse conditions did not effect the I-V characteristic until large pulses destructively damaged the unit.

Fig. 14 is a sketch of a typical V-I characteristic and Fig. 15, shows the voltage-versus-time characteristic for a diode with a negative resistance. The V-I curve can be broken into four regions:

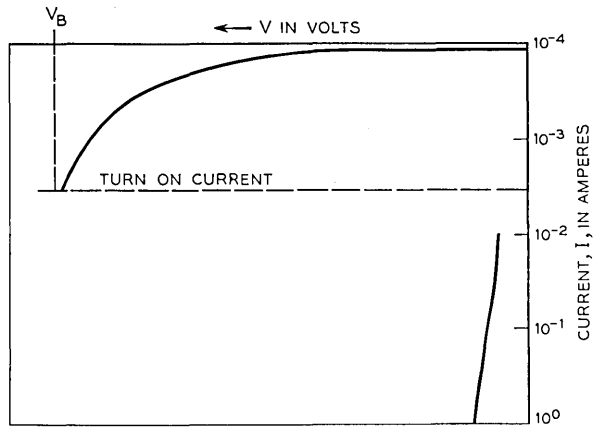

Fig. 14 — A typical V-I characteristic for a diode in which a negative resistance is observed.

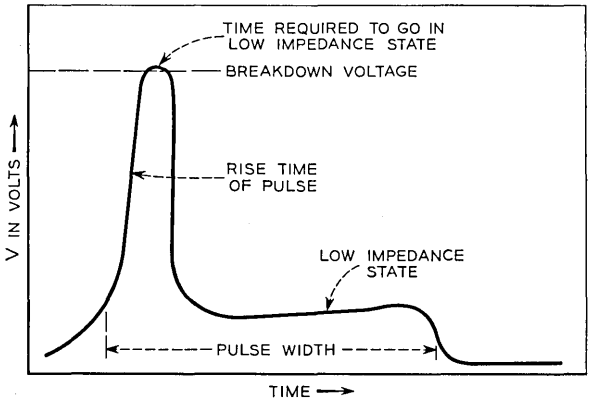

Fig. 15 — A typical V-T characteristic for a diode in which a negative resistance is observed.

1. A high impedance state before the breakdown voltage is reached.
2. A current required to turn on the negative resistance; this current varies from 10^{-3} to 1 amp.
3. The transition to a low impedance state.
4. The low impedance region in which the current is probably limited by the circuit impedance.

The V-T curve can be broken in four regions:

1. The time it takes the pulse to reach the breakdown voltage.
2. The time the diode can maintain the breakdown voltage less than 1 μsec. This is beyond the resolution of the oscilloscope.
3. The time required to fall to the low voltage (low impedance) state, is less than 1 μsec.
4. The remainder of the pulse in the low voltage state.

Fig. 16 is a plot showing the current and voltage required to turn on a negative resistance in several power rectifiers (area $\sim 10^{-2}$ cm^2). To

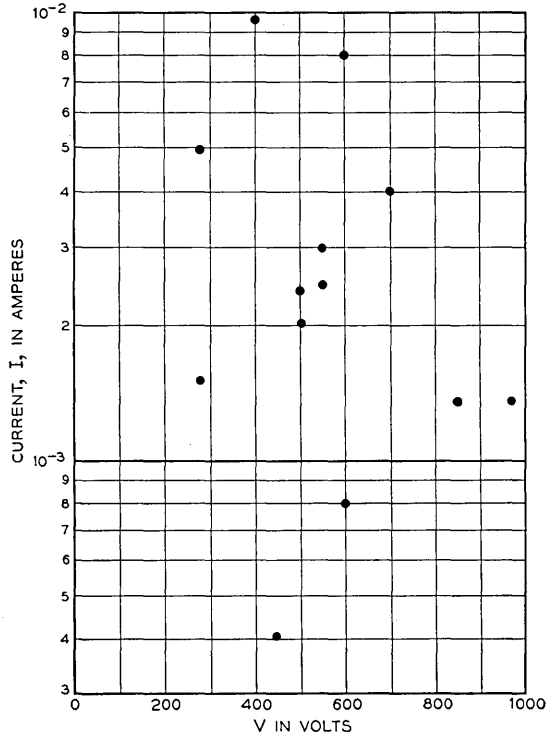

Fig. 16 — Current and voltage required to turn on a negative resistance in several power rectifiers ($A \sim 10^{-2}$ cm^2).

consider the spread of breakdown voltage, the data was normalized to the instantaneous power required to turn on a negative resistance. This turn-on power was the turn-on power multiplied by the voltage.

Fig. 17 is a plot of the distributions of turn-on power for the rectifiers which had a negative resistance plotted as a log normal distribution on probability paper. The median value for the turn-on power is 1.2 watts. Eighty per cent of these diodes went into a negative resistance condition at powers between 0.1 and 10 watts. Many diodes could dissipate several kilowatts with no negative resistance. These were not included.

Experiments show that devices which show surface breakdown will collapse at power levels which are orders of magnitude below that observed for devices in which body breakdown is observed.

The picture is more cloudy with smaller area rectifiers (area $\sim 10^{-3}$ cm^2). In these devices it was not possible to predict the pulse properties

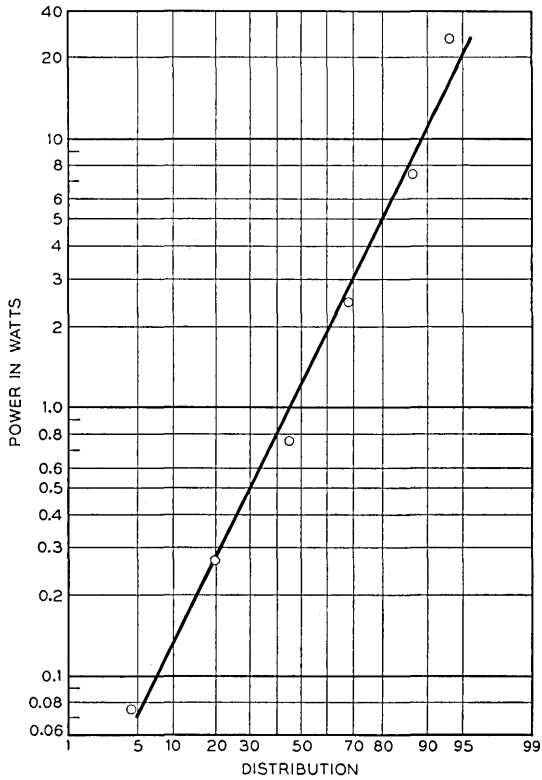

Fig. 17 — The distribution of turn-on power for rectifiers ($A \sim 10^{-2}$ cm^2) in which a negative resistance is observed plotted as a log normal distribution on probability paper.

of the device from the reverse I-V characteristic. This may be attributed to the decrease in power capabilities of the body breakdown process in the smaller devices. This also suggests that the smaller devices have a less severe surface problem.

The distribution of turn-on power for a few hundred small area rectifiers ($A \sim 10^{-3}$ cm^2) is shown in Fig. 18. The median of the distribution occurs at 40 watts. Eighty percent of the units will show a negative resistance when pulsed at power levels between 3 and 500 watts.

VII CONCLUSION

High voltage rectifiers have been fabricated using several sources of high resistivity material employing an uncomplicated diffusion process.

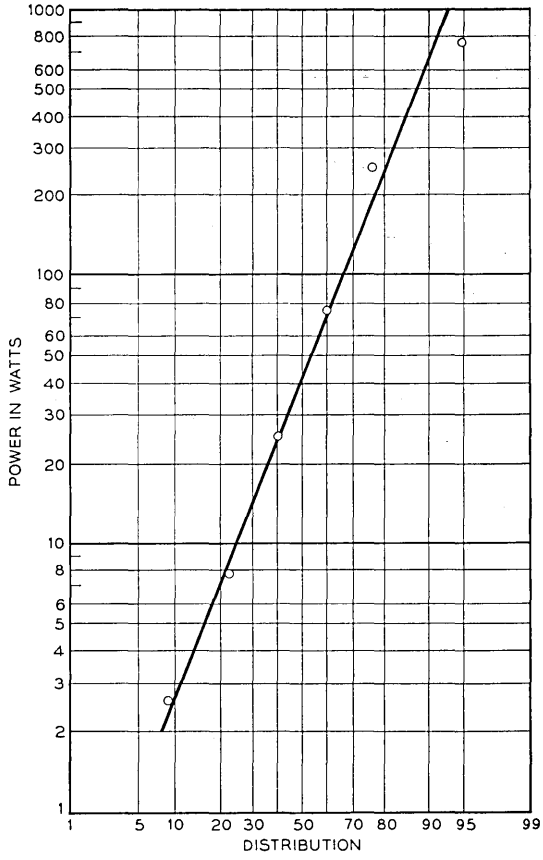

Fig. 18 — The distribution of turn-on power for small area rectifiers ($A \sim 10^{-3}$ cm^2) plotted as a log normal distribution on probability paper.

The most consistent results were obtained using horizontal zone refined silicon. The open tube diffusion technique has sufficient control to satisfy the fabrication requirements.

The magnitudes, voltage and temperature dependences of both the forward and reverse currents of silicon rectifiers can be explained by including a recombination level near the middle of the forbidden energy gap. Design equations for the forward and reverse characteristic of a diode are presented for several important cases. The breakdown voltage of the high voltage devices was shown to be a function of the width of the high resistivity region.

One unsolved problem is the surface limitation of breakdown voltage and reverse currents. This has been observed to decrease the breakdown voltages and increase the reverse currents to undesirable levels.

ACKNOWLEDGEMENT

The authors wish to thank their colleagues for many helpful discussions. Thanks are due Dr. R. N. Noyce of the Shockley Semiconductor Laboratory for making his paper available before publication. The work on zone refined and compensated silicon was done by S. J. Silverman. The floating zone material was supplied by H. E. Bridgers. Much of the experimental work was done by A. R. Tretola, T. J. Vasko and F. R. Lutchko. G. J. Levenbach assisted with the statistical aspects.

REFERENCES

1. M. B. Prince, B.S.T.J., **35**, p. 661, 1956. R. N. Hall, Proc. I.R.E., **40**, p. 1512, 1952.
2. W. Shockley, and W. T. Read, Jr., Phys. Rev., **87**, p. 835, 1952.
3. R. N. Hall, Proc. I.R.E., **40**, p. 1512, 1952. J. S. Saby, Proc. Rugby Conference on Semiconductors, 1956.
4. W. Shockley, B.S.T.J., **28**, p. 435, 1949.
5. K. G. McKay, Phys. Rev., **94**, p. 877, 1954.
6. E. M. Pell, J. Appl. Phys., **26**, p. 658, 1955. E. M. Pell, and G. M. Roe, J. Appl. Phys., **27**, p. 768, 1956.
7. B. Ross, and J. R. Madigan, Bull. A.P.S., **2**, p. 65, 1957.
8. K. G. McKay, Phys. Rev., **94**, p. 877, 1954.
9. S. L. Miller, Phys. Rev., **99**, p. 1234, 1955.
10. S. L. Miller, Phys. Rev., **105**, p. 1246, 1957.
11. P. A. Wolff, Phys. Rev., **95**, p. 1415, 1954.
12. C. T. Sah, R. N. Noyce, and W. Shockley, "Carrier Generation and Recombination in *p-n* Junctions and *p-n* Junction Characteristics", to be published in the Proc. I.R.E.
13. H. C. Theuerer, J. Metals, p. 1316–1319, Oct. 1956.
14. J. A. Burton, Physica, **20**, p. 845–854, 1954.
15. C. J. Frosch and L. Derick, J. Elec. Chem. Soc., to be published.
16. K. D. Smith, P.G.E.D. Conference of the I.R.E., Washington, 1956.
17. F. M. Smits and R. C. Miller, Phys. Rev., **104**, p. 1242–45, 1956.
18. G. Backenstoss, to be published.
19. C. S. Fuller and J. A. Ditzenberger. J. Appl. Phys., **27**, p. 544–53, 1956.
20. W. T. Read, Jr., B.S.T.J., **35**, p. 1239, 1956.
21. D. J. Rose, Phys. Rev., **105**, p. 413, 1957.
22. C. G. B. Garrett and W. H. Brattain, J. Appl. Phys., **27**, p. 299–306, 1956.

Coincidences in Poisson Patterns

By E. N. GILBERT and H. O. POLLAK

(Manuscript received August 3, 1956)

A number of practical problems, including questions about reliability of Geiger counters and short-circuits in electric cables, reduce to the mathematical problem of coincidences in Poisson patterns. This paper presents the probability of no coincidences as well as asymptotic formulas and simple bounds for that probability under a variety of circumstances. The probability of exactly N coincidences is also found in some cases.

INTRODUCTION

A number of practical problems are questions about what we call "coincidences" in Poisson patterns. In d-dimensional space, a Poisson pattern of density λ is a random array of points such that each infinitesimal volume element, dV, has probability λdV of containing a point, and such that the numbers of points in disjoint regions are independent random variables. Then a volume, V, has probability

$$\frac{(\lambda V)^k}{k!} e^{-\lambda V}$$

of containing exactly k points. A coincidence, in our usage of the word, is defined as follows: We imagine a certain fixed distance δ to be given in advance; two points are then said to be *coincident* if they lie within distance δ of one another.

Examples

The best-known case of a coincidence problem concerns Geiger counters. In the simplest mathematical model, there is a short dead-time δ after each count during which other particles can pass through the counter without registering a count. In our present terminology, a count is missed whenever two particles traverse the counter with coincident times of arrival. The same problem is encountered with telephone call registers.

Another example arises in the manufacture of electric cable. Each wire in a cable is covered with an insulation which contains occasional flaws. When the cable is assembled it will fail a short circuit test if it contains a pair of wires such that a flaw on one wire lies within some distance δ of a flaw on the other wire. In a similar way, coincident flaws in the insulation of the wire from which a coil is wound can lead to failure of the coil.

There are also some problems in the development of certain military systems which lead to the consideration of coincidences in Poisson patterns.

Outline of Work

Our primary aim is to study the probability of no coincidences under various circumstances. In Part I, we examine coincidences of two different Poisson patterns, of densities λ and μ respectively, on a line of length L. Here we do not count two points of the *same* pattern within a distance δ as giving a coincidence. A set of integral equations yields the probability of no coincidences as well as an asymptotic formula and upper and lower bounds.

In Part II, we study the probability, $F_0(L)$, of no coincidences for a single one-dimensional Poisson pattern of density λ. These results may also be interpreted as the distribution function for the minimum distance between pairs of points of a Poisson pattern. Sample formulas are the asymptotic formula (for large L)

$$F_0(L) \approx \frac{\lambda}{(\lambda - a)[1 + \delta(\lambda - a)]} e^{-aL},$$

and the bounds (valid for all L)

$$\left(1 - \frac{a}{\lambda}\right) e^{-aL} e^{-(a-\lambda)\delta} \leqq F_0(L) \leqq e^{-aL} e^{-(a-\lambda)\delta},$$

where $s = -a$ is the largest real root of

$$s + \lambda = \lambda e^{-(s+\lambda)\delta}.$$

The problem of n Poisson patterns, all of the same density λ, is examined in Part III. Coincidences are now counted between points of any two distinct patterns.

The one-dimensional problems of Parts I–III succumb readily to analytic techniques. We can find exact expressions for the probabilities of no coincidences in Parts I–III. Two entirely different methods of deriving

exact results are available and are illustrated in Parts II and III. Unfortunately, the exact formulas, although they are finite sums, contain a number of terms which grows with L. Much of our effort has been directed toward finding good, easily computed bounds and asymptotic formulas.

The probabilities of having exactly N coincidences are also obtainable but they have more complicated formulas. A detailed derivation is given only in Part II.

In Part IV, we consider the probability of no coincidence in higher dimensional problems. The methods of Parts I–III fail in higher dimensions, but we are still able to derive some bounds. An exact formula is derived for the probability of no coincidences within a single two-dimensional Poisson pattern in a rectangle with sides $\leqq 2\delta$. We also give particular attention to coincidences in a three-dimensional cylinder.

Part V contains numerical results.

Reduction of the Examples to the Theory

We now wish to see how answers bearing on the practical problems previously listed may be found from this study.

The literature on Geiger counters (see bibliography in Feller[3]) is concerned with statistics of the number of counts registered in a given long time, t. The basic problem is to test the hypothesis that the particles arrive in a Poisson sequence. To this problem, then, are relevant the formulas for the probability of N coincidences in one pattern given in Part II, and the bounds and asymptotic results there derived.

The problem of coincident flaws in an electric cable is three-dimensional, and we have various approaches leading to the probability of no coincidences which are valid under different circumstances. If the cable contains only two wires (with possibly different flaw densities), then the problem reduces to the one-dimensional case of coincidences between two Poisson patterns treated in Part I. If the diameter of the cable is small with respect to δ, and if the density of flaws is the same on each of the n wires in the cable, we have the situation of n identical patterns treated in Part III. If, in addition, n is very large, we may ignore the fact that coincident flaws on a single wire do not cause short circuits, and think of coincidences within a single pattern (Part II). Without the assumption that the diameter of the cable is small with respect to δ, the problem is no longer reducible to a one-dimensional form. Section 4.4 is especially devoted to thick cable, and to producing a lower bound for the probability of no coincidences in this three-dimensional situation.

The literature on Poisson patterns in a line segment contains the fol-

lowing related papers. C. Domb[1] finds the distribution function for the total length of the set of points lying within distance δ of a pattern point. P. Eggleton and W. O. Kermack[2] and also L. Silberstein[5] consider *aggregates*, which are sets of k pattern points all contained in an interval of length δ. In the special case $k = 2$, aggregates are our coincidences. These authors find the expected number of aggregates but not the probability of N aggregates.

I CONCIDENCES BETWEEN TWO PATTERNS

1.1 Integral Equation

Consider two Poisson patterns of points on the real line, the first with density λ (points per unit length) and the second with density μ. We want the probability $F(L)$ that in the segment from 0 to L there is no coincidence between a point of pattern No. 1 and a point of Pattern No. 2. $F(L)$ will be formulated in terms of the conditional probabilities

$P_1(L)$ = Prob (no coincidence, given Pattern No. 1 has point at L),

$P_2(L)$ = Prob (no coincidence, given Pattern No. 2 has point at L).

If $L \leqq \delta$, $P_1(L)$ and $P_2(L)$ are the probabilities that patterns No. 2 and No. 1 are empty:

$$P_1(L) = e^{-\mu L}, \qquad P_2(L) = e^{-\lambda L}, \quad \text{if} \quad L \leqq \delta. \tag{1-1}$$

If $L > \delta$ and Pattern No. 1 contains a point at L, there are two ways that no coincidences can occur. First, Pattern No. 2 may fail to have any points anywhere in the interval $[0, L]$. The probability of this event is $\exp -\mu L$. The second possibility is illustrated in Fig. 1 (using circles for points of Pattern No. 1 and crosses for points in Pattern No. 2). Pattern No. 2 has points in $(0, L)$; the one closest to L is at $y < L - \delta$. Since the interval (y, L) contains no points of Pattern No. 2, the probability of finding this closest point, y, in an interval, dy, is

$$\exp\,[-\mu(L - y)]\mu\, dy.$$

The interval $(y, y + \delta)$ must be free from points of Pattern No. 1 (prob-

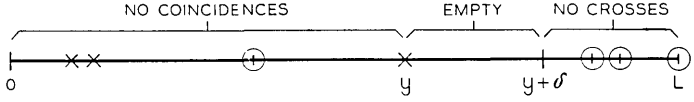

Fig. 1 — Patterns without coincidence.

ability $\exp(-\lambda\delta)$) and the interval $(0, y)$ must contain no coincidences (probability $P_2(y)$). One obtains finally

$$P_1(L) = e^{-\mu L}\left[1 + \mu e^{-\lambda\delta}\int_0^{L-\delta} e^{\mu y}P_2(y)\,dy\right], \tag{1-2}$$

and similarly

$$P_2(L) = e^{-\lambda L}\left[1 + \lambda e^{-\mu\delta}\int_0^{L-\delta} e^{\lambda y}P_1(y)\,dy\right]. \tag{1-3}$$

The solutions $P_1(L)$ and $P_2(L)$ are determined uniquely by (1-1), (1-2) and (1-3). For (1-1) determines them for $0 \leqq L \leqq \delta$ and the integrations indicated in (1-2) and (1-3) will provide the solutions in $0 \leqq L \leqq (n+1)\delta$ when they are known in $0 \leqq L \leqq n\delta$. $P_1(L)$ and $P_2(L)$ are piecewise analytic; the analytic form of the solution changes each time L passes an integer multiple of δ. These analytic expressions soon become complicated and are less useful than the bounds and approximations given later on.

To compute $F(L)$, consider the last place before L at which either Pattern No. 1 or No. 2 has a point. The probability that this last point lies between x and $x + dx$ and belongs to Pattern No. 1 is $\exp[-(\lambda+\mu)(L-x)]\lambda\,dx$ (Fig. 2). This term multiplied by $P_1(x)$ and integrated from 0 to L gives the probability of no coincidences if the last point is a circle. A similar integral gives the probability if the last point is a cross. Finally there is probability $\exp[-(\lambda+\mu)L]$ that neither pattern has a last point [i.e., $(0, L)$ empty]. Then

$$F(L) = e^{-(\lambda+\mu)L}\left[1 + \int_0^L e^{(\lambda+\mu)x}(\lambda P_1(x) + \mu P_2(x))\,dx\right]. \tag{1-4}$$

1.2 Solution by Laplace Transforms

For $i = 1$ or 2, let

$$p_i(s) = \int_0^\infty P_i(L)e^{-sL}\,dL. \tag{1-5}$$

Replacing $P_1(L)$ in (1-5) by (1-1) for $0 \leqq L \leqq \delta$, by (1-2) for $\delta \leqq L$,

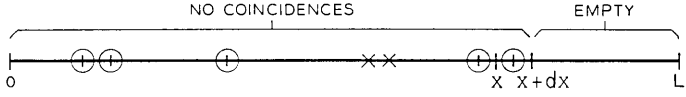

Fig. 2 — Patterns without coincidence.

and interchanging the order of integration of a double integral,

$$(s + \mu)p_1(s) = 1 + \mu e^{-(\lambda+\mu+s)\delta}p_2(s).$$

Similarly,

$$(s + \lambda)p_2(s) = 1 + \lambda e^{-(\lambda+\mu+s)\delta}p_1(s),$$

so that

$$p_1(s) = \frac{s + \lambda + \mu e^{-(\lambda+\mu+s)\delta}}{(s + \lambda)(s + \mu) - \lambda\mu e^{-2(\lambda+\mu+s)\delta}}, \tag{1-6}$$

and

$$p_2(s) = \frac{s + \mu + \lambda e^{-(\lambda+\mu+s)\delta}}{(s + \lambda)(s + \mu) - \lambda\mu e^{-2(\lambda+\mu+s)\delta}}. \tag{1-7}$$

Likewise, using (1-4), the Laplace transform $f(s)$ of $F(L)$ is

$$f(s) = \frac{1 + \lambda p_1(s) + \mu p_2(s)}{\lambda + \mu + s}.$$

As one might expect from the piecewise analytic character of $P_1(L)$ and $P_2(L)$ there is no convenient way of transforming $f(s)$ back to $F(L)$. By evaluating residues of $f(s)$ exp (sL) at the poles of $f(s)$ one might express $F(L)$ as an infinite series of exponential terms. The most slowly damped term in this series can be expected to approximate $F(L)$ when L is large. The poles of $f(s)$ are at the zeros of the denominator $D(s)$ of $p_1(s)$ and $p_2(s)$:

$$D(s) = (s + \lambda)(s + \mu) - \lambda\mu e^{-2(\lambda+\mu+s)\delta}. \tag{1-9}$$

Since $D(x) > 0$ for $x \geqq 0$ and both $D(-\lambda)$ and $D(-\mu)$ are negative, it follows that $D(s)$ has a real zero $s = -a$ with $a < \text{Min } (\lambda, \mu)$.

The zero $s = -a$ of $D(s)$ is the one with the largest real part. For, letting $s = x + iy$, we have in the half plane $x \geqq -a$

$$\begin{aligned} &| (s + \lambda)(s + \mu) | - \lambda\mu | e^{-2(\lambda+\mu+s)\delta} | \\ &\qquad = | s + \lambda | \cdot | s + \mu | - \lambda\mu e^{-2(\lambda+\mu+x)\delta} \\ &\qquad \geqq (x + \lambda)\cdot(x + \mu) - \lambda\mu e^{-2(\lambda+\mu+x)\delta} \geqq 0. \end{aligned}$$

Also, if $y \neq 0$ the $\geqq$ sign in the above proof can be replaced by $>$ and one concludes that all other zeros of $D(s) = 0$ satisfy

$$\text{Re } s < -b$$

for some $b > a$ (note that the left hand side of the preceding inequality does not approach 0 as y approaches $\pm\infty$).

The pole of $f(s)$ at $s = -a$ contributes to $F(L)$ a dominant term

$$F(L) \approx \frac{\lambda^2 + \mu^2 - (\lambda + \mu)a + 2\lambda\mu e^{-(\lambda+\mu-a)\delta}}{(\lambda + \mu - a)[\lambda + \mu - 2a + 2\delta(\lambda - a)(\mu - a)]} e^{-aL}. \quad (1\text{-}10)$$

In (1-10) the error is $0(\exp - bL)$ for large L.

When δ is small, we find $a = 2\lambda\mu\delta + 0(\delta^2)$ and (1-10) becomes

$$F(L) \approx [1 + 0(\delta^2)] \exp - [2\lambda\mu\delta + 0(\delta^2)]L. \quad (1\text{-}11)$$

It is interesting to note that a simple heuristic argument also leads to a formula like (1-11). When δ is small and L is large, one expects that the intervals of length 2δ which contain points of Pattern No. 1 at their centers will comprise a total length near $(\lambda L)(2\delta)$ of the line segment $(0, L)$. The probability that a set of length $2\lambda L\delta$ shall be free of points of Pattern No. 2 is $\exp - 2\lambda\mu\delta L$.

1.3 Bounds

In this section we derive some relatively simple expressions which are good upper and lower bounds on $F(L)$. Both bounds have the same functional form:

$$K(A, B; L) = \frac{\lambda A + \mu B}{\lambda + \mu - a} e^{-aL} + \left(1 - \frac{\lambda A + \mu B}{\lambda + \mu - a}\right) e^{-(\lambda+\mu)L}. \quad (1\text{-}12)$$

In (1-12), a is again the smallest real solution of $D(-a) = 0$. A and B are positive constants which are related by

$$\frac{A}{B} = \frac{\mu}{\mu - a} e^{-(\lambda+\mu-a)\delta} = \frac{\lambda - a}{\lambda} e^{(\lambda+\mu-a)\delta}. \quad (1\text{-}13)$$

$K(A, B; L)$ becomes an upper bound or a lower bound depending on additional restrictions which will be placed on A and B.

To get the lower bound, we restrict A and B by the inequalities

$$A < e^{(a-\mu)\delta}, \qquad B < e^{(a-\lambda)\delta}, \quad (1\text{-}14)$$

and

$$A < \left(1 - \frac{a}{\lambda}\right) e^{\mu\delta}, \qquad B < \left(1 - \frac{a}{\mu}\right) e^{\lambda\delta}. \quad (1\text{-}15)$$

We first prove that (1-13), (1-14), and (1-15) imply

$$P_1(L) > Ae^{-aL}, \qquad P_2(L) > Be^{-aL}. \quad (1\text{-}16)$$

When $0 \leqq L \leqq \delta$, (1-16) holds because of (1-1), (1-14), and the inequalities $a < \lambda$, $a < \mu$. If (1-16) were not true for all L there would be a smallest value, say $L = X > \delta$, at which at least one of the inequalities (1-16) would become an equality. Suppose the inequality (1-16) on $P_1(X)$ fails. Using (1-16) for $L < X$, and (1-2),

$$P_1(X) > e^{-\mu X}\left(1 + B\mu e^{-\lambda\delta}\,\frac{e^{(\mu-a)(X-\delta)} - 1}{\mu - a}\right)$$

$$> Ae^{-aX} + \left(1 - B\,\frac{\mu e^{-\lambda\delta}}{\mu - a}\right)e^{-\mu X} \quad \text{by} \quad (1\text{-}13),$$

$$> Ae^{-aX} \quad \text{by} \quad (1\text{-}15).$$

This contradicts our assumption that (1-16) fails for $P_1(X)$. A similar proof shows (1-16) cannot fail for $P_2(X)$.

Having proved (1-16) we now substitute these bounds into (1-4) and integrate to get $F(L) > K(A, B; L)$.

To make (1-12) into an upper bound it is only necessary to replace (1-14) and (1-15) by

$$A > 1, \qquad B > 1, \tag{1-17}$$

and

$$A > \left(1 - \frac{a}{\lambda}\right)e^{\mu\delta}, \qquad B > \left(1 - \frac{a}{\mu}\right)e^{\lambda\delta}. \tag{1-18}$$

The proof that now $F(L) < K(A, B; L)$ proceeds exactly as before but with all the inequality signs reversed.

Both bounds are dominated by an exponential term $\exp - aL$, as is the asymptotically correct formula (1-10). In typical numerical cases the coefficients multiplying this term in the three formulas agree closely. A numerical case is given in Part V.

1.4 Probability of N Coincidences

The methods of Sections 1.1 and 1.2 can also be used to find the probability $F_N(L)$ that there be exactly N coincidences in the interval $(0, L)$. It might appear most natural to define N to be the number of pairs of points (x, z), x from Pattern No. 1, z from Pattern No. 2, such that

$$|x - z| < \delta. \tag{i}$$

However, we add the additional requirement that x and z be "adjacent" points; i.e.

$$\text{the interval } (x, z) \text{ is empty.} \tag{ii}$$

For example, in Fig. 3, we would count $N = 6$ coincidences even though there are 18 pairs which satisfy (i). In cable problems it appears reasonable to count coincidences as above. If we assume that all flaws are equally bad, then a short circuit is likely to develop only across an adjacent coincidence; our N is the number of places on the cable at which a short circuit can form. Another interpretation is that the cable can be cut into exactly $N + 1$ pieces each of which contain no coincidences.

Let $P_{1,N}(L)$ be the conditional probability of having N coincidences in $(0, L)$ knowing that there is a point of Pattern No. 1 at L. The Laplace transform of $P_{1,N}(L)$ turns out to be the coefficient of t^N in a generating function of the form

$$p_1(t, s) = \frac{\lambda + s + \mu\Omega}{(\lambda + s)(\mu + s) - \lambda\mu\Omega^2},$$

where $\Omega = e^{-(\lambda+\mu+s)\delta}(1 - t) + t$. Interchanging λ and μ one gets the generating function $p_2(t, s)$ for the Laplace transform of the probability $P_{2,N}(L)$ of N coincidences, given a point of Pattern No. 2 at L. Finally the Laplace transform of $F_N(L)$ is the coefficient of t^N in the generating function

$$f(t, s) = \frac{1 - e^{-(\lambda+\mu+s)\delta} + \lambda p_1(t, s) + \mu p_2(t, s)}{\lambda + \mu + s}.$$

Since $f(t, s)$ is a rational function of t, it is easy to find the coefficient of t^N. The poles of this function are again just zeros of $D(s)$. Now, however, the poles are higher order poles. For large L an asymptotic formula for $F_N(L)$ has the form $\exp - aL$ times a polynomial in L with degree depending on N.

For more details about this method we refer the reader to Part II where a similar, but less involved, calculation is carefully done.

II SELF-COINCIDENCES IN ONE POISSON PATTERN

2.1 Integral Equation

In this part we shall consider a single one-dimensional Poisson pattern with density λ and ask for the probability $F_N(L)$ that in the interval $(0, L)$ the pattern have exactly N coincidences. We count coincidences

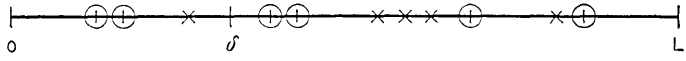

Fig. 3 — Patterns with six coincidences.

as in Section 1.4; a pair (x, z) of pattern points contributes one coincidence to the total number N only if both $|x - z| < \delta$ and the interval between x and z is empty.

Note that $F_0(L)$ is related to the distribution function for the minimum distance between the points of the pattern in $(0, L)$:

$$\text{Prob (min. dist.} \leqq \delta) = 1 - F_0(L),$$

where it must be remembered that $F_0(L)$ is a function of δ.

As in Part I, we first define the conditional probabilities $P_N(L) =$ Prob (exactly N coincidences in $(0, L)$, given a point at L). We then have the following equations:

$$\text{If } L \leqq \delta, \qquad P_N(L) = \frac{(\lambda L)^N}{N!} e^{-\lambda L}, \qquad \text{all } N. \tag{2-1}$$

$$\text{If } L \leqq \delta, \qquad F_N(L) = \frac{(\lambda L)^{N+1}}{(N+1)!} e^{-\lambda L}, \qquad N \geqq 1. \tag{2-2}$$

If $L > \delta$, and $N \geqq 1$, the probability of exactly N coincidences in $(0, L)$ equals the probability of N coincidences up to the last point of the pattern in the interval $(0, L)$ — and if there are to be any coincidences, there must be points of the pattern in $(0, L)$. Hence, if $L > \delta$, $N \geqq 1$,

$$F_N(L) = \int_0^L P_N(L - y) e^{-\lambda y} \lambda \, dy. \tag{2-3}$$

If $N = 0$, the same argument applies, but there is also the possibility that there are no points at all of the pattern in $(0, L)$. Hence, if $L > \delta$,

$$F_0(L) = e^{-\lambda L} + \int_0^L P_0(L - y) e^{-\lambda y} \lambda \, dy. \tag{2-4}$$

Now let us consider the case where there is a point of the pattern at L. Then if the last point preceding L is between $L - \delta$ and L, this point and the point at L will create a coincidence; if there is no point within $(L - \delta, L)$, then all coincidences are within $(0, L - \delta)$. Hence, if $L > \delta$, and $N \geqq 1$,

$$P_N(L) = \int_0^\delta P_{N-1}(L - y) \lambda e^{-\lambda y} \, dy + e^{-\lambda \delta} F_N(L - \delta). \tag{2-5}$$

For the case $N = 0$, we cannot allow a point in the interval $(L - \delta, L)$, and hence, if $L > \delta$,

$$P_0(L) = e^{-\lambda \delta} F_0(L - \delta). \tag{2-6}$$

2.2 Laplace Transform of $F_N(L)$

To analyze the system of equations which is given by relations (2-1) through (2-6), we introduce the generating functions

$$f(L, t) = \sum_{N=0}^{\infty} F_N(L)t^N,$$

and

$$p(L, t) = \sum_{N=0}^{\infty} P_N(L)t^N.$$

If $L > \delta$, we obtain from (2-3) and (2-4) the relation

$$e^{\lambda L}f(L, t) = 1 + \int_0^L p(w, t)e^{\lambda w}\lambda \, dw, \tag{2-7}$$

and from (2-5) and (2-6) the relation (again if $L > \delta$)

$$e^{\lambda L}p(L, t) = \lambda t \int_{L-\delta}^{L} p(w, t)e^{\lambda w} \, dw + e^{\lambda(L-\delta)}f(L - \delta, t). \tag{2-8}$$

If we differentiate (2-7) and (2-8) with respect to L, and then apply (2-7) differentiated to simplify the last terms of (2-8) differentiated, we obtain, still only for $L > \delta$,

$$f'(L, t) + \lambda f(L, t) = \lambda p(L, t), \tag{2-9}$$

$$p'(L, t) + \lambda(1 - t)p(L, t) = \lambda e^{-\lambda\delta}(1 - t)p(L - \delta, t). \tag{2-10}$$

It is easy to check from (2-1) and (2-2) that if $L \leqq \delta$, then

$$p(L, t) = e^{-\lambda L(1-t)},$$

and

$$f(L, t) = e^{-\lambda L}\left(\frac{(e^{\lambda L t} - 1)}{t} + 1\right),$$

and hence (2-9) is valid for *all* L, but the left side of (2-10) *vanishes* if $L \leqq \delta$. Hence we may take Laplace transforms of (2-9) and (2-10). If we define

$$A(s, t) = \int_0^{\infty} f(L, t)e^{-Ls} \, dL,$$

and

$$B(s, t) = \int_0^{\infty} p(L, t)e^{-Ls} \, dL,$$

we obtain from (2-9), which we now know to be valid for all L,

$$(\lambda + s)A(s, t) - 1 = \lambda B(s, t), \tag{2-11}$$

and from (2-10), by recalling that the left side vanishes for $L \leqq \delta$,

$$sB(s, t) - 1 + \lambda(1 - t)B(s, t) = \lambda(1 - t)e^{-(s+\lambda)\delta}B(s, t). \tag{2-12}$$

Hence

$$B(s, t) = \frac{1}{s + \lambda(1 - t)[1 - e^{-(s+\lambda)\delta}]}, \tag{2-13}$$

and

$$A(s, t) = \frac{1}{\lambda + s}(1 + \lambda B(s, t)).$$

If we denote the Laplace transforms of $P_N(L)$ and $F_N(L)$ by $p_N(s)$ and $f_N(s)$ respectively, then

$$p_N(s) = \frac{\lambda^N[1 - e^{-(s+\lambda)\delta}]^N}{[s + \lambda - \lambda e^{-(s+\lambda)\delta}]^{N+1}}, \tag{2-14}$$

and

$$\begin{aligned} f_0(s) &= \frac{1}{\lambda + s}(\lambda p_0(s) + 1), \\ f_N(s) &= \frac{\lambda}{\lambda + s} p_N(s) \quad \text{for} \quad N = 1, 2, \cdots. \end{aligned} \tag{2-15}$$

2.3 Exact Formula for $F_0(L)$

It is possible to solve (2-1) through (2-6) in piecewise analytic form by computing recursively from each interval of length δ to the next one. We shall obtain the piecewise analytic form for $F_0(L)$ by a direct derivation essentially due to E. C. Molina.[4]

Suppose k is the number of pattern points which fall into $(0, L)$. Let x_i denote the distance between the $i - 1^{st}$ point and the i^{th} point (x_1 is the distance from 0 to the first point) as shown in Fig. 4. The configura-

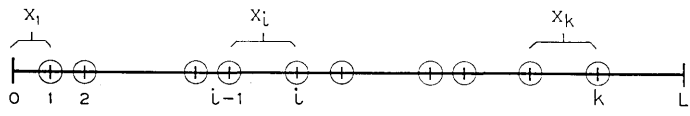

Fig. 4 — Definition of x_i.

tion of points $1, \cdots, k$ on the line is represented by a single point $(x_1, \cdots, x_k)$ in the polyhedron T in k-dimensional space defined by the inequalities

$$T: \quad 0 \leqq x_1, \cdots, 0 \leqq x_k, \qquad x_1 + x_2 + \cdots + x_k \leqq L,$$

and the probability distribution of the point $(x_1, \cdots, x_k)$ in T is uniform. The configurations with no coincidences lie in a smaller polyhedron T' consisting of all points of T for which $\delta \leqq x_2, \cdots, \delta \leqq x_k$. Given k, the conditional probability that there be no coincidences is the ratio of two k-dimensional volumes Vol (T')/Vol (T).

$$\text{Vol } (T') = 0 \quad \text{if} \quad L \leqq (k-1)\,\delta.$$

For larger values of L let $y_1 = x_1$, $y_2 = x_2 - \delta$, $y_3 = x_3 - \delta$, $\cdots$, $y_k = x_k - \delta$. Then T' becomes a polyhedron of the form

$$T'': \quad 0 \leqq y_1, 0 \leqq y_2, \cdots, 0 \leqq y_k, \qquad y_1 + y_2 \cdots + y_k \leqq L - (k-1)\delta.$$

Since the transformation from x's to y's has determinant equal to one, T'' has the same volume as T'. However, T'' is now seen to be similar to T but with sides of length $L - (k-1)\delta$ instead of L. The volume ratio sought must be

$$\left(\frac{L - (k-1)\delta}{L}\right)^k.$$

Since k has the Poisson distribution with mean λL we obtain finally

$$F_0(L) = e^{-\lambda L} \sum_{k=0}^{1+[L/\delta]} \frac{(\lambda L)^k}{k!} \left(1 - \frac{(k-1)\delta}{L}\right)^k.$$

The piecewise-analytic character of $F_0(L)$ is evident; increasing L by an amount δ increases the upper limit on the sum by one and thereby adds a new term to the analytic expression for $F(L)$.

2.4 Asymptotic Formula for $F_N(L)$

Similar exact formulas could be found for all the $F_N(L)$, but they are both complicated and inconvenient for computing if L/δ becomes large. It is thus natural to aim for asymptotic results and for bounds connected with them.

The Laplace transform of $F_N(L)$ is given through (2–14) and (2–15) above. The pole of $f_N(s)$ with largest real part is a pole of order $N + 1$

at a real negative point

$$s = -a > -\lambda.$$

For large L, the asymptotic behavior is given by

$$F_N(L) \approx \frac{\lambda e^{-aL}}{(\lambda - a)[1 + \delta(\lambda - a)]N!}\left[\frac{aL}{1 + \delta(\lambda - a)}\right]^N,$$

where the error term is $0(L^{N-1}\, e^{-aL})$ if $N \geqq 1$. Such a formula, then, is a good approximation for fixed N as L increases; for fixed L, however, it will fail to be good for sufficiently large N.

If $N = 0$, the asymptotic form is

$$F_0(L) \approx \frac{\lambda}{(\lambda - a)[1 + \delta(\lambda - a)]}\, e^{-aL},$$

but the error term now decreases at a more rapid rate, as may be seen by including the contributions of some of the complex poles of $f_0(s)$. To find these poles, set

$$s + \lambda = \lambda e^{-(s+\lambda)\delta}.$$

If

$$s = -\lambda + r \exp(i\theta),$$

one obtains the simultaneous real system

$$2\pi m - \theta = \delta r \sin\theta \qquad (m \text{ integer}),$$

$$\log(r/\lambda) = -\delta r \cos\theta.$$

The first equation defines an infinite family of curves in the s-plane (see Fig. 5). The second equation defines a single curve which intersects the family at poles of $p(s)$.

2.5 Bounds on $F_0(L)$

As in Part I, we may derive bounds on $F_0(L)$ from the integral equation, and obtain

$$\left(1 - \frac{a}{\lambda}\right) e^{-aL} e^{-(a-\lambda)\delta} \leqq F_0(L) \leqq e^{-aL} e^{-(a-\lambda)\delta}.$$

Since $a = \lambda^2\delta + 0(\delta^2)$ for small δ, the bounds are very close if $\lambda\delta$ is not too large.

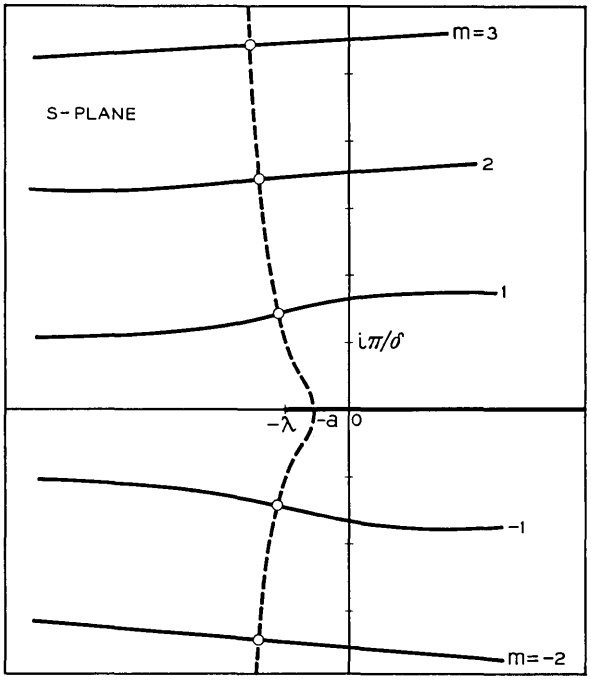

FIG. 5 — Solution of $s + \lambda = \lambda e^{-(s+\lambda)\delta}$

III COINCIDENCES BETWEEN n POISSON PATTERNS

3.1 Integral Equation

In this part we consider n one-dimensional Poisson patterns and ask for the probability, $F(L)$, that in the interval $(0, L)$ no pair of points from different patterns are coincident. Unlike Part I, we now consider only the case in which all n patterns have the same density λ. Let $P(L)$ be the conditional probability, given that Pattern No. 1 has a point at L, that there are no coincidences in $(0, L)$.

If $0 \leqq L \leqq \delta$, $\quad P(L) = \exp - (n - 1)\lambda L.$

If $\delta < L$,

$$P(L) = e^{-(n-1)\lambda L}\left(1 + (n - 1)\lambda e^{-\lambda\delta}\int_0^{L-\delta} e^{(n-1)\lambda y}P(y)\,dy\right)$$

by the same sort of argument used in Part I. Then $F(L)$ will be given by

$$F(L) = e^{-n\lambda L}\left(1 + n\lambda\int_0^L e^{n\lambda x}P(x)\,dx\right).$$

3.2 Bounds and Asymptotic Formula

The Laplace transform of $P(L)$ is

$$p(s) = \{s + (n - 1)\lambda(1 - e^{-(n\lambda+s)\delta})\}^{-1} \tag{3-1}$$

which has one real pole at a negative point $s = -a$, $a < (n - 1)\lambda$. Again it is this pole which contributes the dominant term to both $P(L)$ and $F(L)$ for large L. We find

$$F(L) \approx \frac{n\lambda e^{-aL}}{(1 + [(n - 1)\lambda - a]\delta)(n\lambda - a)}.$$

To bound $P(L)$ by expressions of the form $A \exp(-aL)$ one finds that $A > 1$ will give an upper bound and

$$A < \left(1 - \frac{a}{(n - 1)\lambda}\right) e^{\lambda\delta}$$

will give a lower bound. The corresponding bounds on $F(L)$ are of the form

$$\left(1 - \frac{n\lambda A}{n\lambda - a}\right) e^{-n\lambda L} + \frac{n\lambda A}{n\lambda - a} e^{-aL}.$$

3.3 Exact Solution

As in Part II an exact formula for $F(L)$ may be given as a finite sum. We now derive it from the Laplace transform,

$$f(s) = (s + n\lambda)^{-1} (1 + n\lambda p(s)),$$

of $F(L)$. We may use (3-1) to expand $f(s)$ into the series

$$f(s) = \frac{1}{s + n\lambda} \left\{1 + n\lambda \sum_{k=0}^{\infty} \frac{((n - 1)\lambda e^{-(n\lambda+s)\delta})^k}{(s + (n - 1)\lambda)^{k+1}}\right\}. \tag{3-2}$$

The identity

$$\begin{aligned}(s + n\lambda)^{-1}&(s + (n - 1)\lambda)^{-k-1} \\ &= \frac{1}{\lambda} \sum_{j=0}^{k} (-\lambda)^{-k+j}(s + (n - 1)\lambda)^{-j-1} + (-\lambda)^{-k-1}(s + n\lambda)^{-1}\end{aligned} \tag{3-3}$$

provides a partial fraction expansion for the k^{th} term of the series (3-2). Transforming (3-2) term by term with the help of (3-3) we find

$$\begin{aligned}F(L) = {} & e^{-n\lambda L}[-(n - 1)]^{[L/\delta]+1} \\ & + ne^{-(n-1)\lambda L} \sum_{k=0}^{[L/\delta]} [-(n - 1)e^{-\lambda\delta}]^k \sum_{j=0}^{k} \frac{[-\lambda(L - k\delta)]^j}{j!}.\end{aligned}$$

This is the desired formula for $F(L)$.

IV MULTIDIMENSIONAL PROBLEMS

4.1 Two-Pattern Lower Bound

We now derive some results on the probabilities of no coincidences in some multi-dimensional situations. The simplest one is a lower bound for the case of two Poisson patterns.

Theorem: Consider a d-dimensional region of volume V containing two Poisson patterns with densities λ and μ. Let $S(\delta)$ be the volume of the d-dimensional sphere of radius δ. The probability of no coincidences between the two patterns has the lower bound

$$e^{-\lambda V(1-e^{-\mu S(\delta)})}.$$

Proof

Let the pattern with density λ be called the λ-pattern and the other the μ-pattern. Given any λ-pattern of k points there will be no coincidences provided only that a certain region T contains no points of the μ-pattern. T consists of all points of the volume V which lie in any of the spheres of radius δ centered on the k points of the λ-pattern. Since these spheres may overlap and may extend partly outside the volume V, we have

$$\text{volume of } T \leqq k\, S(\delta),$$

and

$$\begin{aligned}\text{Prob (no coinc., given } k \text{ points)} &= \exp(-\mu \text{ volume of } T)\\ &\geqq \exp(-k\mu\, S(\delta)).\end{aligned}$$

Since the number, k, of points of the λ-pattern has the Poisson distribution with mean λV the (unconditional) probability of no coincidences has the lower bound

$$\sum_{k=0}^{\infty} \frac{(\lambda V)^k}{k!} e^{-\lambda V} e^{-k\mu S(\delta)}.$$

Summing the series one proves the theorem. Interchanging λ and μ in the theorem gives another lower bound. The one stated above is the better of the two if $\lambda < \mu$.

The difference between the lower bound and the true probability comes from two sources: (a) The overlap between the k spheres; this will be a small effect if $\lambda^2 S(2\delta)V$ is small, and (b) the spheres which extend partly outside the volume V; there will be relatively few such spheres if only a small fraction of the volume V lies within distance δ

of its boundary. Hence in some cases the lower bound will be a good approximation to the correct value.

It may also be noted that no real use was made of the spherical shape of the volumes $S(\delta)$. If one wants to consider a point of the μ-pattern to be coincident with a point of the λ-pattern if it lies in some other neighborhood, not of spherical shape, the same lower bound applies but with $S(\delta)$ replaced by the volume of the neighborhood.

4.2 Single-Pattern Lower Bound

A similar derivation in the case of a single Poisson pattern leads to:

Theorem: Let a Poisson pattern of density λ be distributed over a d-dimensional region of volume V. Let $S(\delta)$ be the volume of the d-dimensional sphere of radius δ. Then the probability of no coincidences is at least as large as

$$e^{-\lambda V}\{1 + \lambda S(\delta)\}^{V/S(\delta)}.$$

The theorem will follow from another bound which is slightly more accurate but much more cumbersome.

Lemma

In the above theorem a lower bound is

$$e^{-\lambda V}\left(1 + \lambda V + \sum_{k=2}^{[V/S(\delta)]} \frac{(\lambda V)^k}{k!} \prod_{j=1}^{k-1} [1 - jS(\delta)/V]\right). \tag{4-1}$$

Proof of Lemma

The probability sought is of the form

$$\sum_k e^{-\lambda V} \frac{(\lambda V)^k}{k!} p_k , \tag{4-2}$$

where p_k is the probability that, when exactly k points are distributed at random over V, there are no coincidences. To estimate p_k, imagine the k points to be numbered 1, 2, $\cdots$ k and placed in the region one at a time. If no coincidences have been created among points 1, $\cdots$, j (which is an event of probability p_j) the probability that the addition of point $j + 1$ creates no coincidence is just the probability that this new point lies in none of the j spheres of radius δ centered on points 1, $\cdots$, j. The union of these j spheres intersected with the volume V

is always of volume $\leqq jS(\delta)$. Hence

$$p_{j+1} \geqq p_j[1 - jS(\delta)/V],$$

or

$$p_k \geqq \prod_{j=1}^{k-1} [1 - jS(\delta)/V]. \tag{4-3}$$

When $(k - 1)S(\delta) > V$ the above argument fails because the later terms of the product are negative; in this case we use the trivial bound $p_k \geqq 0$. Combining (4-2) with (4-3) the lemma follows.

Once more the bound may be expected to be almost correct if $\lambda^2 VS(2\delta)$ is small and if most of the region V lies farther than δ away from its boundary. The bound is also correct for non-spherical neighborhoods (see discussion of previous theorem).

When $V/S(\delta)$ is large, the sum (4-1) is unwieldy. If we let H equal $V/S(\delta)$, we may rewrite the typical term in the sum as

$$\frac{(\lambda V)^k}{k!} \prod_{j=1}^{k-1} (1 - j/H) = \frac{(\lambda V/H)^k}{k!} H(H - 1) \cdots (H - k + 1).$$

If H happens to be an integer, this equals

$$\binom{H}{k} (\lambda V/H)^k,$$

so that the complete sum (4-1) equals

$$e^{-\lambda V} \left(1 - \frac{\lambda V}{H}\right)^H. \tag{4-4}$$

We will now prove that if H is not an integer, the sum always *exceeds* (4-4), so that (4-4) is a lower bound in all cases. We wish to prove that

$$1 + \sum_{k=1}^{[H]+1} \frac{x^k}{k!} H(H - 1) \cdots (H - k + 1) \geqq (1 + x)^H \tag{4-5}$$

for any positive H, in which event the theorem follows with

$$x = \frac{\lambda V}{H} \quad \text{and} \quad H = V/S(\delta).$$

The inequality (4-5) will be proved by induction on $[H]$. If $[H] = 0$, then we are required to show that

$$1 + Hx \geqq (1 + x)^H$$

for $0 \leqq H < 1$. This follows immediately from the concavity of $(1 + x)^H$.

Suppose now that (4-5) holds for a value H. If we integrate both sides of (4-5) from 0 to x, we obtain

$$x + \sum_{k=1}^{[H]+1} \frac{x^{k+1}}{(k+1)!} H(H-1) \cdots (H-k+1) \geqq \frac{(1+x)^{H+1} - 1}{H+1},$$

which may be rewritten as

$$1 + \sum_{k=1}^{[H+1]+1} \frac{x^k}{k!} (H+1)(H) \cdots (H-k+2) \geqq (1+x)^{H+1}.$$

This completes the induction, and the proof of the theorem.

4.3 Another Lower Bound (Any Number of Patterns)

Another kind of lower bound can be derived which sometimes will be better than the above bounds when the region V has a large fraction of its volume within δ of the boundary. For example, V might be a three-dimensional circular cylinder (a cable) with a radius which is comparable to δ.

To derive this bound one first finds the expected number, E, of coincidences in V. An upper bound on E will also suffice. Then it is noted that $1 - E$ is a lower bound on the probability of no coincidences. For if Q_N is the probability of finding N coincidences,

$$E = \sum NQ_N \geqq \sum_{N=1}^{\infty} Q_N = 1 - Q_0 . \tag{4-6}$$

4.4 Thick Cable

For example, we now give a lower bound which is of interest in connection with the problem of a cable with many wires.

Theorem: Let a Poisson pattern of points with density λ be placed in a cylinder of length L and radius $R > \delta$. The probability of finding no coincidences in the cylinder is at least as great as

$$1 - \lambda^2\pi^2 L\left(\frac{2R^2\delta^3}{3} - \frac{R\delta^4}{4} + \frac{\delta^5}{15}\right).$$

Proof

Introduce cylindrical coordinates r, φ, Z so that the cylinder is described by

$$r \leqq R, \qquad 0 \leqq Z \leqq L.$$

Consider first any pattern point (r, φ, Z) with Z-coordinate satisfying $\delta \leqq Z \leqq L - \delta$. Let arrows be drawn from this point to all other pattern points (if any) within distance δ. The expected number of arrows drawn from this point will be $\lambda G(r)$ where $G(r)$ is the volume of the intersection of the cylinder with a sphere of radius δ centered at the point. For points near the ends of the cylinder ($Z \leqq \delta$ or $L - \delta \leqq Z$), the expected number of arrows will be less than $\lambda G(r)$. Since the probability of finding a pattern point in a little volume element dV is λdV, we conclude that the expected number of arrows drawn in the entire cylinder will be less than

$$\iiint_{\text{cylinder}} \lambda^2 G(r)\, dV.$$

If the cylinder has N coincidences, there will be $2N$ arrows (each point of a coincident pair appears once at the head of an arrow and once at the tail). Hence the expected number of coincidences is

$$E \leqq \lambda^2 \pi L \int_0^R G(r)\, r\, dr. \tag{4-7}$$

Since an exact formula for $G(r)$ is rather cumbersome, we are content with a simple but close upper bound. If $r \leqq R - \delta$ then clearly $G(r) = 4\pi\delta^3/3$. If $r > R - \delta$ we get an upper bound on $G(r)$ by computing the shaded volume in Fig. 6; the intersection of the sphere with a half-space.

$$G(r) \leqq [2\delta^3 + 3(R - r)\delta^2 - (R - r)^3]\, \pi/3.$$

Substituting these expressions for $G(r)$ in (4-7), integrating, and using (4-6) the theorem follows.

The approximation to $G(r)$ which was made above is bad when R is much less than δ, but in this case good estimates may be obtained from the one-dimensional results of Part II. Note also that if λ is large enough, the bound becomes negative and is therefore useless.

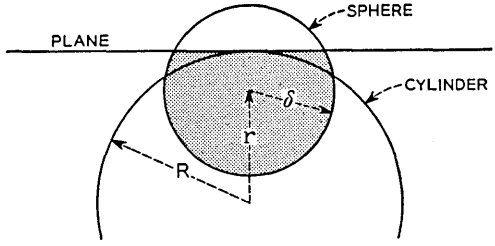

Fig. 6 — A region for estimating $G(r)$.

4.5 Upper Bounds

Good upper bounds appear even harder to get than lower bounds. One procedure is to divide the region V into a number of smaller cells. If each cell has probability, p, of no coincidences and if there are K cells, then p^K is the probability of no coincidence in any cell. If there is no coincidence in V there will be none in any cell; hence p^K is an upper bound on the probability of no coincidence in V.

Of course, p^K is too large because of the possibility of a coincidence between two points in different cells. It follows that p^K will be a close bound only if the cell size is made large; but then p becomes hard to compute.

For example, consider self-coincidences in a single Poisson pattern in a large region of area V in the plane. Cover this area with an array of hexagonal cells of side $\delta/2$ as shown in Fig. 7. The area of each hexagon is $3\sqrt{3}\,\delta^2/8$ so the number of cells used will be about $K = 8V/3\sqrt{3}\,\delta^2$. A cell has no coincidence if it contains at most one pattern point, hence

$$p = (1 + \lambda 3\sqrt{3}\,\delta^2/8) \exp - 3\sqrt{3}\,\lambda\delta^2/8.$$

The upper bound is

$$p^K = e^{-\lambda V}\left(1 + \frac{3\sqrt{3}}{8}\lambda\delta^2\right)^{(8V/3\sqrt{3}\delta^2)}$$

which has an interesting resemblance to the lower bound

$$e^{-\lambda V}(1 + \pi\lambda\delta^2)^{V/\pi\delta^2}.$$

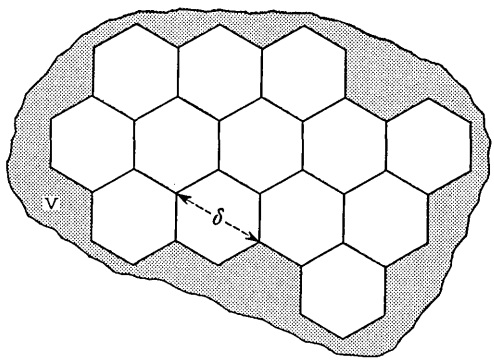

Fig. 7 — Pattern for studying coincidences in a plane region.

4.6 An Exact Calculation

The upper and lower bounds in Section 4.5 are not very close, largely because of the small size of the hexagonal cells. An improved upper bound may be obtained using square cells of side 2δ. We can calculate p for small rectangular cells but only if we redefine our notion of coincidence in terms of square neighborhoods instead of circular neighborhoods. That is, points (x_1, y_1) and (x_2, y_2) are now considered coincident if simultaneously

$$| x_1 - x_2 | \leqq \delta, \quad \text{and} \quad | y_1 - y_2 | \leqq \delta.$$

The result we get is the only exact calculation of a non-trivial multi-dimensional coincidence probability known to us.

Consider the rectangle $0 \leqq x \leqq L$, $0 \leqq y \leqq M$ with L and M both $\leqq 2\delta$. If L is less than δ, two points are coincident if and only if their y-coordinates differ by less than δ. The problem then reduces to a one-dimensional coincidence computation such as we gave in Part II. Therefore, suppose both L and M are greater than δ.

There is probability

$$g_k = \frac{(\lambda LM)^k}{k!} e^{-\lambda LM}$$

that the rectangle contains k points. We therefore subdivide the problem into cases of the form "given k, find the probability that the k points have no coincidences". Only five of these cases have a non-zero answer. To show this, divide the rectangle into four rectangles of sides $L/2$, $M/2$; if $k \geqq 5$ one of these rectangles must contain more than one point, and so a coincidence. The remaining cases $k = 0, 1, 2, 3, 4$ may be further subdivided according to which pairs of x-coordinates are less than δ apart. Let us number the k points $(x_1, y_1), \cdots, (x_k, y_k)$ in such a way that the x-coordinates are in order $x_1 \leqq x_2 \leqq \cdots \leqq x_k$. If, for some i, $x_{i+2} \leqq x_i + \delta$, then the subcase in question contributes zero to the probability of no coincidences because all of $| x_i - x_{i+1} |$, $| x_{i+1} - x_{i+2} |$, $| x_{i+2} - x_i |$ are $\leqq \delta$ and at least one of $| y_i - y_{i+1} |$, $| y_{i+1} - y_{i+2} |$, $| y_{i+2} - y_i |$ is $\leqq \delta$. The only subcases which remain to give a non-zero contribution are the nine listed in Table I. The number in the "subcase" column is k. The next column contains the x-inequalities which define the subcase. The probability that the k ordered x-coordinates satisfy the stated inequalities is listed as prob_x. If the x-inequalities are satisfied there will be no coincidences if and only if $| y_b - y_a | > \delta$ for every inequality $| x_b - x_a | \leqq \delta$ given in the x-inequality column. These y-in-

TABLE I

Subcase	x inequ.	prob$_x$	y inequ.	prob$_y$
0	—	1	—	1
1	—	1	—	1
2(a)	$x_2 - x_1 > \delta$	$(1 - \delta/L)^2$	—	1
2(b)	$x_2 - x_1 \leqq \delta$	$\dfrac{2\delta L - \delta^2}{L^2}$	$\lvert y_2 - y_1 \rvert > \delta$	$(1 - \delta/M)^2$
3(a)	$x_2 - x_1 \leqq \delta$ $x_3 - x_2 > \delta$	$\frac{2}{3}(1 - \delta/L)^3$	$\lvert y_2 - y_1 \rvert > \delta$	$(1 - \delta/M)^2$
3(b)	$x_2 - x_1 > \delta$	$\frac{2}{3}(1 - \delta/L)^3$	$\lvert y_2 - y_3 \rvert > \delta$	$(1 - \delta/M)^2$
3(c)	$x_2 - x_1 \leqq \delta$ $x_3 - x_2 \leqq \delta$ $x_3 - x_1 > \delta$	$\frac{2}{3}(1 - \delta/L)^2 \left(\frac{4\delta}{L} - 1\right)$	$\lvert y_2 - y_3 \rvert > \delta$ $\lvert y_1 - y_2 \rvert > \delta$	$\frac{2}{3}(1 - \delta/M)^3$
4(a)	$x_3 - x_2 \leqq \delta$ $x_3 - x_1 > \delta$ $x_4 - x_2 > \delta$	$\frac{1}{3}(1 - \delta/L)^4$	$\lvert y_2 - y_1 \rvert > \delta$ $\lvert y_3 - y_2 \rvert > \delta$ $\lvert y_4 - y_3 \rvert > \delta$	$\frac{5}{12}(1 - \delta/M)^4$
4(b)	$x_3 - x_2 > \delta$	$\frac{1}{3}(1 - \delta/L)^4$	$\lvert y_2 - y_1 \rvert > \delta$ $\lvert y_4 - y_3 \rvert > \delta$	$(1 - \delta/M)^4$

equalities are listed in the third column and the probabilities that they are satisfied are listed as prob$_y$. The probability of no coincidences is

$$\sum g_k \text{ prob}_x \text{ prob}_y$$

where the sum is over all nine subcases. The sum is

$$\exp(-\lambda LM)\left\{1 + \lambda LM + \frac{\lambda^2}{2}[L^2M^2 - \delta^2(2L - \delta)(2M - \delta)] + \frac{2\lambda^3}{27}(L - \delta)^2(M - \delta)^2(2LM + L\delta + M\delta - 4\delta^2) + \frac{17\lambda^4}{864}(L - \delta)^4(M - \delta)^4\right\}.$$

If $L = M = 2\delta$, this reduces to

$$\exp(-4\delta^2\lambda)\left[1 + 4\delta^2\lambda + \frac{7}{2}\delta^4\lambda^2 + \frac{16}{27}\delta^6\lambda^3 + \frac{17}{864}\delta^8\lambda^4\right].$$

A sample of one of the above computations may be instructive. Consider, for example, Case 4(a). We have $0 \leqq x_1 \leqq x_2 \leqq x_3 \leqq x_4 \leqq L$, and require:

$$x_3 - x_2 \leqq \delta,$$

$$x_3 - x_1 > \delta,$$

$$x_4 - x_2 > \delta.$$

The probability of this is

$$(L^4/8)^{-1} \int_{x_3=\delta}^{L} \int_{x_2=x_3-\delta}^{L-\delta} \int_{x_1=0}^{x_3-\delta} \int_{x_4=x_2+\delta}^{L} dx_4\, dx_1\, dx_2\, dx_3$$

$$= \frac{8}{L^4} \int_{x_3=\delta}^{L} \int_{x_2=x_3-\delta}^{L-\delta} (L - x_2 - \delta)(x_3 - \delta)\, dx_3\, dx_2$$

$$= \frac{8}{L^4} \frac{(L-\delta)^4}{24} = \frac{1}{3}\left(1 - \frac{\delta}{L}\right)^4.$$

In the y-direction we require $|y_2 - y_1| > \delta$, $|y_3 - y_2| > \delta$, $|y_4 - y_3| > \delta$, and there are no order restrictions. Assume first that $y_2 < y_3$. Then the probability that y_1 and y_4 satisfy their restrictions is

$$\left(\frac{y_3 - \delta}{M}\right)\left(\frac{M - y_2 - \delta}{M}\right).$$

Hence, the probability for satisfying all the conditions is

$$\int_{\delta}^{M} \int_{0}^{y_3-\delta} \left(\frac{y_3 - \delta}{M}\right)\left(\frac{M - y_2 - \delta}{M}\right) \frac{dy_2}{M} \frac{dy_3}{M} = \frac{5}{24}\left(1 - \frac{\delta}{M}\right)^4.$$

Interchanging y_2 with y_3 and y_1 with y_4 shows that the assumption $y_2 > y_3$ yields the same answer, so that the required probability is

$$\frac{5}{12}\left(1 - \frac{\delta}{M}\right)^4.$$

V NUMERICAL WORK

5.1 Coincidences between Two Patterns

5.1.1 Machine Computation of $F(L)$

To compute the probability of no coincidences in a line of length L directly, it is convenient to transform equations (1–2) through (1–4) into

the following differential difference equations:

$$P_1'(x) + \mu P_1(x) = \begin{cases} 0 \text{ if } x \leqq \delta \\ P_2(x - \delta)\mu e^{-(\lambda+\mu)\delta} & \text{if } x > \delta, \end{cases}$$

$$P_2'(x) + \lambda P_2(x) = \begin{cases} 0 \text{ if } x \leqq \delta \\ P_1(x - \delta)\lambda e^{-(\lambda+\mu)\delta} & \text{if } x > \delta, \end{cases}$$

$$F'(x) + (\lambda + \mu)F(x) = \lambda P_1(x) + \mu P_2(x),$$

$$P_1(0) = P_2(0) = F(0) = 1.$$

These have been solved on a general purpose analog computer with the aid of a lumped-element approximate delay line for a number of cases. We have chosen for illustrative purposes the parameters $\lambda = 5$, $\mu = 10$, $\delta = 0.02$, and $L \leqq 1$. The exact solution, together with various approximations to be described in the sequel, is plotted in Fig. 8, where the exact solution is labelled y_1.

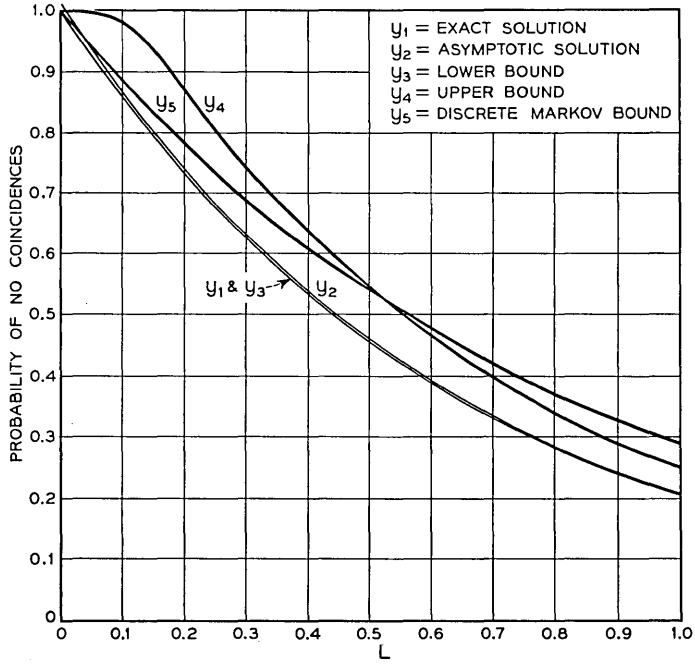

FIG. 8 — Probability of no coincidences between two one-dimensional Poisson patterns with $\lambda = 5$, $\mu = 10$, if $\delta = 0.02$.

5.1.2 The Asymptotic Formula

An approximation to the probability $F(x)$ of no coincidences is given by the asymptotic formula (1-10) which, of course, becomes a better approximation the larger L becomes. If $\lambda = 5$, $\mu = 10$, and $\delta = 0.02$, the smallest value, a, such that

$$(\lambda - a)(\mu - a) = \lambda\mu e^{-2(\lambda+\mu-a)\delta}$$

is $a = 1.548$. The asymptotic formula for $F(L)$ now becomes

$$F(L) \approx 1.013e^{-1.548L},$$

which is found in Fig. 8 as y_2.

5.1.3 Bounds Using the Asymptotic Exponent

Formulas (1-12) through (1-18) give a scheme for computing both upper and lower bounds for $F(L)$ which have the right behavior for large L, and also agree with the solution at $L = 0$. They become

$$F(L) \geqq 1.007e^{-1.548L} - 0.007e^{-15L},$$

and

$$F(L) \leqq 1.195e^{-1.548L} - 0.195e^{-15L},$$

respectively, and are represented by y_3 and y_4 in Figure 8.

5.1.4 An Upper Bound by a Discrete Markov Process

If we mark on the positive x-axis the points $n\delta/2$, $n = 0, 1, 2, \cdots$, we can assign to each interval of length $\delta/2$ thus created a state (ij), i, $j = 0$ or 1, as follows: $i = 0$ if no point of the λ-process is present in the interval, $i = 1$ if one or more points of the λ-process are present, and similarly for j and μ. An interval of length δ, made up of two adjacent intervals of length $\delta/2$, may then be represented by a number between 0 and 15 in binary notation, where 3, 6, 7, 9, and 11–15 represent a coincidence within the interval of length δ. We now define a Markov process as follows: in the interval $0 \leqq t < \delta$, let $p_i^{(0)}$, $i = 0, 1, 2, 4, 5, 8, 10$, be the probabilities of occurrence of the i^{th} state, so that, for example, $p_0^{(0)} = e^{-2\lambda\delta}e^{-2\mu\delta}$, and $p_1^{(0)} = e^{-2\lambda\delta}e^{-\mu\delta}(1 - e^{-\mu\delta})$. These are the states in which there is no coincidence in $(0, \delta)$. In addition, let $q^{(0)}$ represent the probability of all the other states put together; i.e., of a coincidence in $(0, \delta)$. We now define $p_i^{(n)}$, $i = 0, 1, 2, 4, 5, 8, 10$ as the probability of the i^{th} state in the interval $(n\delta/2, (n + 2)\delta/2)$, where we

require in addition that all states in the intervals $(k\delta/2, (k + 2)\delta/2)$, $k < n$, are from the same "no coincidence" index set. We define $q^{(n)}$ as the probability of a state 3, 6, 7, 9, or 11–15, in *some* interval $(k\delta/2, (k + 2)\delta/2)$, $k \leqq n$. There are then transition probabilities from states in the $n - 1^{st}$ to states in the n^{th} interval. For example,

$$p_0^{(n)} = e^{-\lambda\delta}e^{-\mu\delta}(p_0^{(n-1)} + p_4^{(n-1)} + p_8^{(n-1)}),$$

and

$$q^{(n)} = q^{(n-1)} + (1 - e^{-\lambda\delta})(1 - e^{-\mu\delta})(p_0^{(n-1)} + p_4^{(n-1)} + p_8^{(n-1)}) + (1 - e^{-\lambda\delta})(p_1^{(n-1)} + p_5^{(n-1)}) + (1 - e^{-\mu\delta})(p_2^{(n-1)} + p_{10}^{(n-1)}).$$

The quantity $1 - q^{(n)}$ is then an upper bound for the probability of no coincidences (upper because it is possible for a coincidence to occur in the process which is not counted in this subdivision of it). The curve y_5 in Fig. 8 is drawn through points at $L = n\delta/2$ computed in this manner.

To summarize the results, we see that the asymptotic formula and the lower bound are both indistinguishable from the right answer; the upper bounds are fairly far off. The upper bound derived by the Markov process is better than that derived from the integral equation until

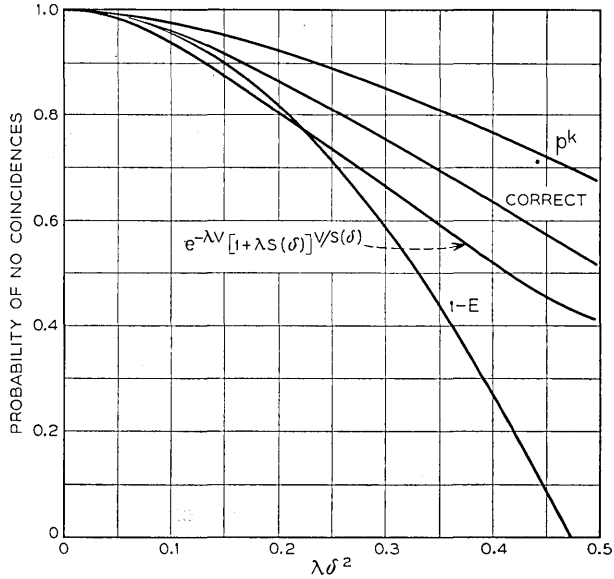

FIG. 9 — Probability of no coincidences in a $2\delta \times 2\delta$ square; neighborhoods are square.

about $L = 0.5$ (25 iterations), when the integral equation upper bound becomes better.

5.2 A Single Pattern in a Square

To test our higher-dimensional bounds, we consider again coincidences in a single Poisson pattern in a square of side 2δ. The exact probability of no coincidences was given in Part IV assuming square neighborhoods. The lower bound (Sec. 4.2)

$$e^{-\lambda V}(1 + \lambda S(\delta))^{V/S(\delta)}$$

applies using $V = (2\delta)^2$ and $S(\delta) = (2\delta)^2$ for square neighborhoods. To use the lower bound $1 - E$ we note that the exact expected number of coincidences is

$$E = \frac{1}{2}\lambda^2 \int_0^{2\delta} \int_0^{2\delta} A(x, y)\, dx\, dy$$

where $A(x, y)$ is the area of the intersection of the given square with the square neighborhood centered at (x, y). The lower bound is $1 - E = 1 - 9\lambda^2\delta^4/2$. The upper bound p^K can be used if the square is cut into $K = 4$ squares of side δ, each with a probability $p = (1 + \lambda\delta^2) \exp - \lambda\delta^2$ of no coincidence.

These bounds, together with the exact probability, are plotted as functions of $\lambda\delta^2$ in Fig. 9. When $\lambda\delta^2$ is small, the $1 - E$ bound is correct to terms of order $0(\lambda^3\delta^6)$. This might have been predicted from (4-6) since it seems reasonable that $Q_2, Q_3, \cdots$ should be of higher order in λ than Q_1 when λ is small. Ultimately the first lower bound becomes a better estimate. It must be recognized that this other lower bound is being tested under very severe conditions. Since every point of the square has a neighborhood which intersects the boundary, the errors from source (b) of Part V are considerable.

The authors wish to thank D. W. Hagelbarger and H. T. O'Neil for their assistance in the course of the calculations reported in this section, and Miss D. T. Angell for preparing some of the figures.

REFERENCES

1. C. Domb, The Problem of Random Intervals on a Line, Proc. Cambridge Phil. Soc., **43**, pp. 329–341, 1947.
2. P. Eggleton and W. O. Kermack, A Problem in the Random Distribution of Particles, Proc. Royal Soc. Edinburgh, Sec A, **62**, pp. 103–115, 1944.
3. W. Feller, On Probability Problems in the Theory of Counters, Studies and Essays presented to R. Courant, Interscience, pp. 105–115, 1948.
4. E. C. Molina, The Theory of Probability and Some Applications to Engineering Problems, Trans. A.I.E.E., **44**, pp. 294–299, 1925.
5. L. Silberstein, The Probable Number of Aggregates in Random Distributions of Points, Phil. Mag., Series 7, **36**, pp. 319–336, 1945.

Bell System Technical Papers Not Published in this Journal

ANDERSON, O. L., see Andreatch, P.

ANDREATCH, P., and ANDERSON, O. L.[1]

Teflon as a Pressure Medium, Rev. Sci. Instr., **28,** p. 288, April, 1957.

BOYET, H.,[1] and SEIDEL, H.[1]

Analysis of Nonreciprocal Effects in an N-Wire Ferrite-Loaded Transmission Line, Proc. I.R.E., **45,** pp. 491–495, April, 1957.

BOYET, H., see Seidel, H.

BOZORTH, R. M.[1]

Review of Magnetic Annealing, Proc. of 1956 A.I.E.E. Conf. on Magnetism and Magnetic Materials, **T-91,** pp. 69–75, April, 1957.

BURNS, F. P.[1]

Piezoresistive Semiconductor Microphone, J. Acous. Soc. Am., **29,** pp. 248–253, Feb., 1957.

CARRUTHERS, J. A., see Geballe, T. H.

CIOFFI, P. P.[1]

Rectilinearity of Electron-Beam Focusing Fields from Transverse Component Determinations, Commun. and Electronics, **29,** pp. 15–19, March, 1957.

COOK, R. K.[1]

Absorption of Sound by Patches of Absorbent Material, J. Acous. Soc. Am., **29,** pp. 324–329, March, 1957.

[1] Bell Telephone Laboratories, Inc.

COOK, R. K.[1]

Variation of Elastic Constants and Static Strains with Hydrostatic Pressure: A Method for Calculation from Ultrasonic Measurements, J. Acous. Soc. Am., **29,** pp. 445–449, April, 1957.

DEWALD, J. F.[1]

The Kinetics of Formation of Anode Films on Single Crystal Indium Antimonide, J. Electrochem. Soc., **104,** pp. 244–251, April, 1957.

DOWLING, R. C.[4]

Lightning Protection on the Stevens Point-Wisconsin Rapids Intercity Telephone Cable, Commun. and Electronics, **28,** pp. 697–701, Jan., 1957.

FEHER, G.[1]

Electron Structure of F Centers in KCl by the Electron Spin Double-Resistance Technique, Phys. Rev., Letter to the Editor, **105,** pp. 1122–1123, Feb. 1, 1957.

FOSTER, F. G.[1]

The Unconventional Application of the Metallograph, Focus, **18,** pp. 16–20, April, 1957.

GARN, P. D.[1]

An Automatic Recording Balance, Anal. Chem., **29,** pp. 839–841, May, 1957.

GAST, R. W.[5]

Field Experience with the A2A Video System, Commun. and Electronics, **28,** pp. 710–716, Jan., 1957.

GEBALLE, T. H.,[1] CARRUTHERS, J. A.,[6] ROSENBERG, H. M.,[6] and ZIMAN, J. M.[7]

The Thermal Conductivity of Germanium and Silicon Between 2 and 300°K, Proc. Royal Soc., **A238,** pp. 502–514, Jan. 29, 1957.

[1] Bell Telephone Laboratories, Inc.
[4] Wisconsin Telephone Company, Madison.
[5] New York Telephone Company, New York.
[6] Oxford University, England.
[7] Cambridge University, England.

GELLER, S.[1]

Crystallographic Studies of Perovskite-Like Compounds.— IV. Rare Earth Scandates, Vanadites, Galliates, Orthochromites, Acta Crys., **10,** pp. 243–248, April 10, 1957.

GELLER, S.[1]

Crystallographic Studies of Perovskite-Like Compounds.— V. Relative Ionic Sizes, Acta Crys., **10,** pp. 248–251, April 10, 1957.

GELLER, S.,[1] and GILLEO, M. A.[1]

Structure and Ferrimagnetism of Ythium and Rare-Earth-Iron Garnets, Acta Crys., **10,** p. 239, March 10, 1957.

GILLEO, M. A.[1]

Crystallographic Studies of Perovskite-Like Compounds.— III. $La(M_x, Mn_{1-x})0$ with M = Co, Fe and Cr, Acta Crys., **10,** pp. 161–166, March, 1957.

GILLEO, M. A., see Geller, S.

GREEN, E. I.[1]

Nature's Pulses, I.R.E. Student Quarterly, **3,** pp. 3–5, Feb., 1957.

GROSSMAN, A. J.[1]

Synthesis of Tschebycheff Parameter Symmetrical Filters, Proc. I.R.E., **45,** pp. 454–473, April, 1957.

HAGSTRUM, H. D.[1]

Thermionic Constants and Sorption Properties of Hafnium, J. Appl. Phys., **28,** pp. 323–328, March, 1957.

HAWORTH, F. E.[1]

Breakdown Fields of Activated Electrical Contacts, J. Appl. Phys., Letter to the Editor, **28,** p. 381, March, 1957.

HEIDENREICH, R. D.,[1] and NESBITT, E. A.[1]

Stacking Disorders in Nickel Base Magnetic Alloys, Phys. Rev., Letter to the Editor, **105,** pp. 1678–1679, March 1, 1957.

[1] Bell Telephone Laboratories, Inc.

HOLDEN, A. N., see Wood, Elizabeth A.

HUNTLEY, H. R.[2]

Where We Are and Where We Are Going in Telephone Transmission, Commun. and Electronics, **29,** pp. 54–63, March, 1957.

JOEL, A. E.[1]

Electronics in Telephone Switching Systems, Commun. and Electronic, **28,** pp. 701–709, Jan., 1957.

KAPPEL, F. R.[2]

Three-Dimensional Engineers, Elec. Engg., **76,** pp. 267–270, April 1957.

KARLIN, J. E., see Pierce, J. R.

KARP, A.[1]

Backward-Wave Oscillator Experiments at 100 to 200 Kilomegacycles, Proc. I.R.E., **45,** pp. 496–503, April, 1957.

KELLY, M. J.[1]

The Work and Environment of the Physicist Yesterday, Today, and Tomorrow, Phys. Today, **10,** pp. 26–31, April, 1957.

LAW, J. T.,[1] and MEIGS, P. S.[1]

Rates of Oxidation of Germanium, J. Electrochem. Soc., **104,** pp. 154–159, March, 1957.

LAW, J. T.,[1] and MEIGS, P. S.[1]

The High Temperature Oxidation of Germanium, Semiconductor Surface Physics (book), pp. 383–400, 1957, Univ. of Penna. Press, Philadelphia.

LIEHR, A. D.[1]

Structure of $Co(CO)_4H$ and $Fe(CO)_4H_2$, Zeitschrift Für Naturforschung, **12b,** pp. 95–96, Feb., 1957.

[1] Bell Telephone Laboratories, Inc.
[2] American Telephone and Telegraph Company.

LIEHR, A. D.[1]

Structure of π-Cyclopentadienyl Metal Hydrides, Naturwissenschaften, **44,** p. 61, Feb. 1, 1957.

LUNDBERG, C. V., see Vacca, G. N.

LUNGBERG, J. L.,[1] and NELSON, L. S.[1]

The High Intensity Flash Irradiation of Polymers, Nature, Letter to the Editor, **179,** pp. 367–368, Feb. 16, 1957.

MARRISON, W. A.[1]

A Wind-Operated Electric Power Supply, Elec. Engg., **76,** pp. 418–421, May, 1957.

MCCALL, D. W.[1]

Nuclear Magnetic Resonance in Guanidinium Aluminum Sulfate Hexahydrate, J. Chem. Phys., Letter to the Editor, **26,** pp. 706–707 March, 1957.

MEIGS, P. S., see Law, J. T.

MENDIZZA, A.,[1] SAMPLE, C. H.,[8] and TEEL, R. B.[9]

A Comparison of the Corrosion Behavior and Protective Value of Electrodeposited Zinc and Cadmium on Steel, Symposium on Properties, Tests, and Performances of Electrodeposited Metallic Coatings, **A.S.T.M. Special Tech. Publication 197,** pp. 49–64, 1957.

MILLER, S. L.[1]

The Ionization Rates for Holes and Electrons in Si, Phys. Rev., **105,** pp. 1246–1249, Feb. 15, 1957.

NELSON, L. S., see Lundberg, J. L.

NESBITT, E. A., see Heidenreich, R. D.

[1] Bell Telephone Laboratories, Inc.
[8] International Nickel Company, New York City.
[9] International Nickel Company, Wrightsville Beach, North Carolina.

PALMQUIST, T. F.[10]

Multi-Unit Neutralizing Transformers, Commun. and Electronics, **28,** pp. 717–721, Jan., 1957.

PIERCE, J. R.,[1] and KARLIN, J. E.[1]

Information Rate of a Human Channel, Proc. I.R.E., Letter to the Editor, **45,** p. 368, March, 1957.

REA, W. T.[1]

The Communication Engineer's Needs in Information Theory, Commun. and Electronics, **28,** pp. 805–808, Jan., 1957.

ROSENBERG, H. M., see GEBALLE, T. H.

SAMPLE, C. H., see MENDIZZA, A.

SEIDEL, H.,[1] and BOYET, H.[1]

Form of Polder Tensor for Single Crystal Ferrite with Small Cubic Symmetry Anisotropy Energy, J. Appl. Phys., **28,** pp. 452–454, April, 1957.

SEIDEL, H., see Boyet, H.

SLICHTER, W. P.[1]

Nuclear Magnetic Resonance in Some Fluorine Derivatives of Polyethylene, J. Poly. Sci., **24,** pp. 173–188, April, 1957.

SMITH, K. D., see Veloric, H. S.

SUHL, H.[1]

Proposal for a Ferromagnetic Amplifier in the Microwave Range, Phys. Rev., Letter to the Editor, **106,** pp. 384–385, April 15, 1957.

SWANEKAMP, F. W., see Van Uitert, L. G.

TEEL, R. B., see Mendizza, A.

[1] Bell Telephone Laboratories, Inc.
[10] Bell Telephone Company of Canada, Montreal, Quebec.

TREUTING, R. G.[1]

Torsional Strain and the Screw Dislocation in Whisker Cyrstals, Acta Met., Letter to the Editor, **5,** pp. 173–175, March, 1957.

VACCA, G. N.,[1] and LUNDBERG, C. V.[1]

Aging of Neoprene in a Weatherometer, Wire and Wire Products, **32,** pp. 418–457, April, 1957.

VAN UITERT, L. G.[1]

Magnesium-Copper — Manganese-Aluminum Ferrites for Microwave Applications, J. App. Phys., **28,** pp. 320–322, March, 1957.

VAN UITERT, L. G.[1]

Magnetic Induction and Coercive Force Data on Members of the Series $BaAl_xFe_{12-x}O_{19}$ and Related Oxides, J. Appl. Phys., **28,** pp. 317–319, March, 1957.

VAN UITERT, L. G.,[1] and SWANEKAMP, F. W.[1]

Permanent Magnet Oxides Containing Divalent Metal Ions, J. Appl. Phys., **28,** pp. 482–485, April, 1957.

VELORIC, H. S.,[1] and SMITH, K. D.[1]

Silicon Diffused Junction Avalanche Diodes, J. Electrochem. Soc., **104,** pp. 222–227, April, 1957.

WALKER, L. R.[1]

Orthogonality Relays for Gyrotropic Wave Guides, J. Appl. Phys., Letter to the Editor, **28,** p. 377, March, 1957.

WEBER, L. A.[1]

Influence of Noise on Telephone Signaling Circuit Performance, Commun. and Electronics, **28,** pp. 636–643, Jan., 1957.

WEIBEL, E. S.[1]

An Electronic Analogue Multiplier, Trans. I.R.E., **PGEC, EC-6,** pp. 30–34, March, 1957.

[1] Bell Telephone Laboratories, Inc.

WEISS, J. A.[1]

An Interference Effect Associated with Faraday Rotation, and Its Application to Microwave Switching, Proc. Conf. on Magnetism and Magnetic Materials, pp. 580–585, Feb., 1957.

WERTHEIM, G. K.[1]

Energy Levels in Electron-Bombarded Silicon, Phys. Rev., **105,** pp. 1730–1735, March 15, 1957.

WILLIS, F. H.[1]

Some Results with Frequency Diversity in a Microwave Radio System, Commun. and Electronics, **29,** pp. 63–67, March, 1957.

WOOD, ELIZABETH A.,[1] and HOLDEN, A. N.[1]

Monoclinic Glycine Sulfate: Crystallographic Data, Acta Crys., **10,** pp. 145–146, Feb., 1957.

YOUNKER, E. L.[1]

A Transistor Driven Magnetic Core Memory, Trans. I.R.E., **PGEC, EC-6,** pp. 14–20, March, 1957.

ZIMAN, J. M., see Geballe, T. H.

[1] Bell Telephone Laboratories, Inc.

Recent Monographs of Bell System Technical Papers Not Published in This Journal*

BASHKOW, T. R., and DESOER, C. A.

A Network Proof of a Theorem on Hurwitz Polynomials, Monograph 2614.

BEACH, A. L., see Thurmond, C. D.

BIGGS, B. S., see Lundberg, C. V.

CUTLER, C. C.

Instability in Hollow and Strip Electron Beams, Monograph 2711.

DESOER, C. A., see Bashkow, T. R.

ELIAS, P., FEINSTEIN, A., and SHANNON, C. E.

A Note on the Maximum Flow Through a Network, Monograph 2768.

FEINSTEIN, A., see Elias, P.

GULDNER, W. G., see Thurmond, C. D.

HARING, H. E., see Taylor, R. L.

INGRAM, S. B.

Role of Evening Engineering Education in the Training of Technicians, Monograph 2771.

KRAMER, H. P., and MATHEWS, M. V.

A Linear Coding for Transmitting a Set of Correlated Signals, Monograph 2757.

* Copies of these monographs may be obtained on request to the Publication Department, Bell Telephone Laboratories, Inc., 463 West Street, New York 14, N. Y. The numbers of the monographs should be given in all requests.

Contributors to This Issue

VACLAV E. BENES, A.B., Harvard College, 1950; M.A. Ph.D., Princeton University, 1953. Instructor in logic and philosophy of science, Princeton University, 1952–53; Bell Telephone Laboratories, 1953–. Since joining the Laboratories, Mr. Benes has been engaged in mathematical systems research, involving stochastic processes describing the passage of traffic through a switching system. He is the author of a number of papers on mathematical logic and analytic philosophy. Member of the Mind Association, the Association for Symbolic Logic, the Institute of Mathematical Statistics, American Mathematical Society, and Phi Beta Kappa.

E. N. GILBERT, B.S., Queens College, 1942; Ph.D., Massachusetts Institute of Technology, 1948; Mr. Gilbert's early employment was with the M.I.T. Radiation Laboratory. He joined Bell Telephone Laboratories in 1948. His work has been on studies of the information theory and on the switching theory. He now is part of the communication fundamentals group. Mr. Gilbert is a member of the American Mathematical Society.

H. O. POLLAK, B.A., Yale University, 1947; M.A., Harvard University, 1948; Ph.D., Harvard University, 1951; Bell Telephone Laboratories, 1951–. Mr. Pollak has been engaged in mathematical research and military systems analysis. He is a member of Phi Beta Kappa, Sigma Xi, American Mathematical Society and Mathematical Association of America.

M. B. PRINCE, A.B., Temple University, 1947; Ph.D., Massachusetts Institute of Technology, 1951; Bell Telephone Laboratories, 1951–1956; Hoffman Semiconductor Division of Hoffman Electronics Corporation, 1956–. Between 1949–51 he was a research assistant at the Research Laboratories of Electronics at M.I.T. where he was concerned with cryogenic research. At Bell Telephone Laboratories, Mr. Prince was concerned with the physical properties of semiconductors and semiconductor devices and was associated with the development of silicon devices, including the Bell Solar Battery and the silicon power rectifier.

Mr. Prince is a member of the I.R.E., the American Physical Society, the Association for Applied Solar Energy, the Electrochemical Society and Sigma Xi.

W. W. RIGROD, B.S. in E.E., Cooper Union Institute of Technology, 1934; M.S. in Engineering, Cornell University, 1941; D.E.E., Polytechnic Institute of Brooklyn, 1950; State Electrotechnical Institute, Moscow, U.S.S.R., 1935–39; Westinghouse Electric Corporation, 1941–51; Bell Telephone Laboratories, 1951–. His work has been related principally to the study and development of electron tubes, both the gaseous-discharge and the high-vacuum types. He is a member of the American Physical Society, I.R.E. and Sigma Xi.

STEPHEN O. RICE, B.S., Oregon State College, 1929; California Institute of Technology, Graduate Studies, 1929–30 and 1934–35; Bell Telephone Laboratories, 1930–. In his first years at the Laboratories, Mr. Rice was concerned with the non-linear circuit theory, with special emphasis on methods of computing modulation products. Since 1935 he has served as a consultant on mathematical problems and in investigations of the telephone transmission theory, including noise theory, and applications of electromagnetic theory. Fellow of the I.R.E.

ERLING D. SUNDE, E.E., Technische Hochschule, Darmstadt, Germany, 1926; Brooklyn Edison Company, 1927; American Telephone and Telegraph Company, 1927–1934; Bell Telephone Laboratories, 1934–. Mr. Sunde's work has been centered on theoretical and experimental studies of inductive interference from railway and power systems, lightning protection of the telephone plant, and fundamental transmission studies in connection with the use of pulse modulation systems. Author of *Earth Conduction Effects in Transmission Systems*, a Bell Laboratories Series book. Member of the A.I.E.E., the American Mathematical Society, and the American Association for the Advancement of Science.

HAROLD S. VELORIC, B.A., University of Pennsylvania, 1951; M.A., 1952, Ph.D., 1954, University of Delaware; Bell Telephone Laboratories, 1954–. Between 1951–4 he was a research fellow concerned with the synthesis and analysis of boron and silicon compounds. Since joining the Laboratories, Mr. Veloric has been concerned with the properties and development of solid state devices. He has been associated with the development of several classes of silicon diodes, including power rectifiers, voltage-reference and computer diodes. Dr. Veloric is a member of the American Chemical Society and the Electrochemical Society.