THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Volume 53 September 1974 Number 7

Spurious Parametric Oscillations in IMPATT Diode Circuits

By W. E. SCHROEDER

(Manuscript received December 26, 1973)

This paper describes a quantitative experimental confirmation of a recent theory for large-signal parametric instabilities in IMPATT diode circuits. The relevant parts of the theory are presented in a concise form. A graphical technique is given that is useful for analyzing circuit stability. A sufficient (though not necessary) condition for well-behaved circuits is also studied, and several diagrams are presented that may prove useful to circuit designers.

I. INTRODUCTION

Spurious oscillations are frequently observed in strongly driven IMPATT amplifiers and oscillators. Spurious oscillation is a generic term that means, in the case of an amplifier, that there is a signal frequency in the output not present in the input and, in the case of an oscillator, that there is a signal present in addition to the main desired oscillation. In IMPATT diode circuits, the type of spurious oscillation most frequently observed fits into one of two categories: (*i*) a parametric-pair-type instability in which there are two spurious signals whose frequencies sum to equal that of the strong signal or (*ii*) a subharmonic instability in which one spurious signal occurs at one-half the frequency

of the strong signal. It is evident that the second category is a special case of the first, and may be considered a degenerate parametric oscillation.

In the course of experiments on IMPATT amplifiers and locked oscillators, both degenerate and nondegenerate parametric instabilities have been encountered. Accompanying the onset of spurious oscillation, it has been noted that there is often a saturation of rf output power, a degradation of the amplifier's noise performance, and/or a shift in the center frequency of the locking band. Although the spurious signals themselves are generally undesirable, the adverse effects upon amplifier performance at the design frequency are even more serious. Therefore, it is important to understand the nature of the spurious oscillations and to determine what must be done to the amplifier design to eliminate them.

In a recent paper, M. E. Hines[1] presented a theory for parametric interactions in IMPATT diodes, which is based on the nonlinear inductive behavior of the avalanche process. He has given an expression for the stability of the IMPATT diode-microwave circuit system.

This paper describes experiments conducted to characterize the diode and circuit conditions at the onset of spurious oscillation, for both degenerate and nondegenerate cases, and to compare the results with Hines' theory. Good quantitative agreement was found between our experimental observations and the predictions of a simplified form of Hines' stability expression. Thus, it appears that Hines' theory forms a sound basis for designing circuits to suppress parametric-type spurious oscillations.

In another section of this paper, we discuss a procedure for determining a region of impedance such that, if the circuit impedance (as seen by the diode) lies within this region, the diode-circuit system is stable against parametric oscillations for all drive levels. It appears that this stable impedance region is physically obtainable, and further that it is compatible with the impedance requirements discussed by Brackett[2] with regard to suppressing bias-circuit oscillations that are due to a phenomenon different from the one considered by Hines.

II. REVIEW OF HINES' THEORY

In this section a brief recapitulation is given of the aspects of Hines' theory that are pertinent to the stability of parametric interactions in IMPATT diodes.

Hines uses the Read[3] model of an IMPATT diode to derive an equivalent circuit for parametric interaction of the form given in Fig. 1.

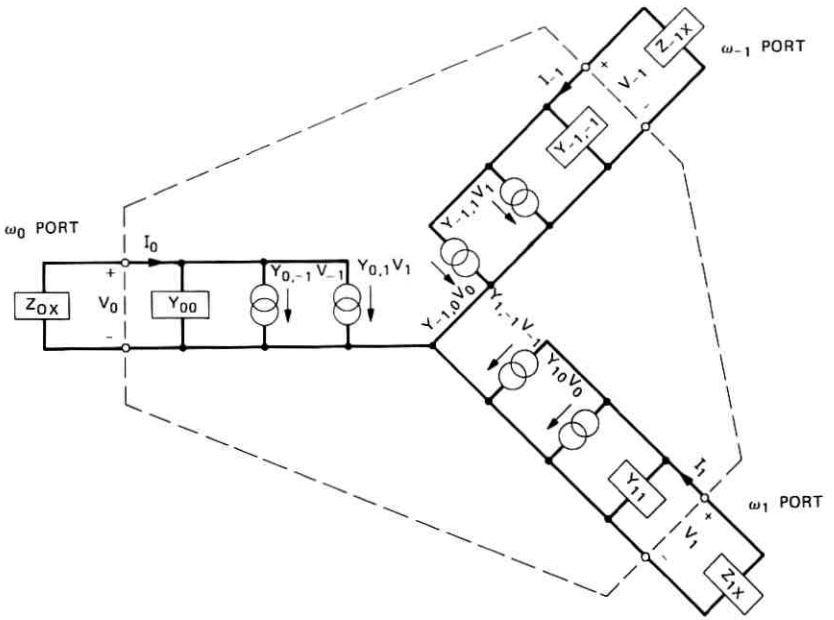


Fig. 1—Equivalent circuit for parametric interaction.

The equivalent circuit is applicable to small-signal parametric interactions when the diode is in some large-signal "pumped" state. The pump frequency ω_p is defined as the fundamental frequency of the large-amplitude pump wave that is present at the diode. The pump wave arises from the impressed signal in the case of an amplifier or from the circuit-controlled oscillation in the case of an oscillator. The parametrically related frequencies ω_m are defined as follows:

$$\omega_0 \triangleq \text{the lowest perturbing frequency (\textit{not} the main oscillation frequency)} \tag{1}$$

and

$$\omega_1 \triangleq \omega_p + \omega_0, \tag{2}$$

$$\omega_{-1} \triangleq \omega_p - \omega_0^*, \tag{3}$$

where the asterisk denotes complex conjugate.

From Fig. 1, it is evident that the current I_0 at ω_0 consists of three components: current due to the self-admittance at ω_0 and current induced by the presence of parametrically related voltages at the upper

and lower sidebands, ω_1 and ω_{-1}, respectively. Contributions to I_0 from the higher-order parametric signals at the frequencies $n \cdot \omega_p \pm \omega_0$ for $n \geqq 2$ have been neglected. The equivalent circuit for the sidebands is similar. The impedance Z_{mx} is the external impedance presented to the diode wafer at the frequency ω_m. The diode package and mount parasitic impedances are included in Z_{mx}. The pump frequency of the diode does not appear directly in the equivalent circuit, but the pump state is introduced through the admittances $Y_{m,n}$.

From this equivalent circuit Hines derived the following characteristic equation, which governs the stability of the diode-circuit system:

$$\begin{aligned} D \triangleq\ & 1 - M_1 M_{-1} S_0 (S^*_{-1} + S_1) - M_2 M_{-2} S_1 S^*_{-1} \\ & + (M_1^2 M_{-2} + M^2_{-1} M_2) S_0 S_1 S^*_{-1} \qquad \text{(Hines eq. 23)} \\ =\ & 0. \end{aligned} \tag{4}$$

The complex quantity M_n is defined as the avalanche particle current at the nth pump harmonic divided by the dc current. Note that here the subscript n signifies the frequency $n \cdot \omega_p$, whereas the subscripts of the S quantities refer to the parametric frequencies ω_m. It is also important to note that

$$M_n = M^*_{-n} \tag{5}$$

and

$$|M_n| \leqq 1. \tag{6}$$

Generally, $|M_n|$ increases monotonically with the pump level and approaches unity as the avalanche current approaches a sharp, narrow pulse.

The complex quantities S_m may be thought of as stability factors. They are defined as:

$$S_m = S(\omega_m) = 1 + \left(\frac{\omega_m^2}{\omega_a^2 - \omega_m^2} \right) \frac{Z_x(\omega_m) + 1/j\omega_m C_T}{Z_x(\omega_m) + Z_d(\omega_m)}, \tag{7}$$

where

ω_a = avalanche resonance frequency of the diode,
C_T = total "cold" capacitance of the diode, i.e., that of the entire depleted region,

and

$$Z_d(\omega) = \frac{\bar{x}_d - \bar{x}_d(\sin \theta_d/\theta_d) - \omega^2/\omega_a^2 + j\bar{x}_d[(1 - \cos \theta_d)/\theta_d]}{j\omega C_T(1 - \omega^2/\omega_a^2)}, \tag{8}$$

where

$\bar{x}_d$ = ratio of drift length to total length of the diode

and

$\theta_d = \omega\tau_d$ = transit angle of the drift region.

The quantity Z_d is the small-signal Read-model admittance previously derived by Gilden and Hines.[4]

The stability of the diode-circuit system may be studied by examining the roots of eq. (4) for complex frequency. It is apparent that the effect of the large-signal pump state upon stability is completely described by the two factors M_1 and M_2, since the stability factors S depend only upon the *small-signal* diode impedance and the circuit impedance. This separation of pump level and circuit impedance effects is a very useful aspect of Hines' approach. It avoids entangling the stability discussion with the details of the large-signal nonlinear diode behavior. If, for a given circuit and diode, the stability factors are such that there are no unstable roots for $0 < |M_1| < 1$ and $0 < |M_2| < 1$, then the system is stable for any pump level.

Under certain conditions, simplified forms of the characteristic equation are valid and useful. For instance, if $|M_2| \ll |M_1|$, which corresponds to intermediate pump levels, then eq. (4) becomes

$$1 - |M_1|^2 S_0(S^*_{-1} + S_1) \cong 0. \tag{9}$$

Alternatively, if $|S_1| \ll |S_0|$ and $|S_{-1}|$ then eq. (4) simplifies to

$$1 - |M_1|^2 S_0 S^*_{-1} \cong 0. \tag{10}$$

The simplified criterion given by eq. (10) is the one suggested by Hines for the degenerate case, but it is applicable to the nondegenerate case as well. We have found good agreement between our experimental observations and eq. (10), as is discussed in the next section.

The approximation which leads to eq. (10) may be justified by examination of the definition of S: For $\omega \ll \omega_a$, the term $\omega^2/(\omega_a^2 - \omega^2)$ is much smaller than 1 and $S(\omega)$ approaches unity. For ω in the vicinity of ω_a, $Z_d(\omega)$ is quite different from $1/j\omega C_T$ and of course $\omega^2/(\omega_a^2 - \omega^2)$ is large: therefore, $S(\omega)$ is not in general negligible. However, there is no pole for $S(\omega = \omega_a)$, since the pole in Z_d cancels that of $\omega^2/(\omega_a^2 - \omega^2)$. Since it is usually desirable to pump the diode where the negative conductance is large, $\omega_p^2 \geqq 2\omega_a^2$ is generally true. Therefore, for $\omega > \omega_p$

$$\frac{\omega^2}{\omega_a^2 - \omega^2} \to -1$$

and also

$$Z_d \to \frac{1}{j\omega C_T},$$

and, therefore, $S(\omega)$ tends ultimately to zero. Therefore, eq. (10) seems to be a reasonable approximation, especially for $\omega_0 \approx \omega_p/2$.

In order to study stability, the characteristic function D is considered a function of complex frequency $p = \sigma + j\omega$. Instability occurs if the zeros of D lie in the right half of the p-plane. In the case of spurious oscillation in IMPATT diodes, it is an experimental fact that for sufficiently small M_1 there is no instability; thus, the zeros of D (and the poles of $S_0S^*_{-1}$) must lie in the left half of the p-plane for small M_1. As M_1 is increased to its maximum value of 1, the roots of D may or may not cross into the right half of the p-plane. Assuming that the roots of D vary continuously with M_1, then any crossing into the right half of the plane implies that there is first an intersection of the root locus with the imaginary axis for some value of M_1. The frequency for which the zero of D lies on the $p = j\omega$ axis may be designated as $\bar{\omega}$ and the corresponding value of M_1 as $\bar{M}_1$. For $M_1 = \bar{M}_1$, the system is said to be marginally stable or at the threshold of spurious oscillation, where the spurious frequency is $\bar{\omega}$.

Thus, in order to determine approximately if a given diode and circuit are unstable for a physically attainable pump condition, it is only necessary to evaluate $D \cong 1 - |M_1|^2 S_0 S^*_{-1}$ for *real* ω (actually, just for $0 < \omega_0 < \omega_p/2$) and determine if there are any zeros for $0 < |M_1| < 1$. If not, then the system is stable for any pump level. If there are zeros, then the zero with the smallest value of $|M_1|$ corresponds to the first spurious oscillation which we would expect to observe upon increasing the pump level monotonically from zero. In this case, we may predict the drive level and the frequency for the onset of spurious oscillation.

It is evident that zeros of D correspond to real values of $S_0S^*_{-1}$. Thus, we may plot $S_0S^*_{-1}$ and note the frequencies where the locus intercepts the real axis. The basic construction for one value of ω_0 is indicated schematically in Fig. 2. There is at least one real axis intercept since, for $\omega_0 = \omega_p/2$, $S_0 = S_{-1}$. This is not necessarily an instability, since $|S(\omega_p/2)|$ may be less than 1. Since the phase condition for instability is always satisfied at the subharmonic, this suggests that the half-pump frequency oscillation may be a more likely form of spurious output, which appears to agree with the experience of many workers.

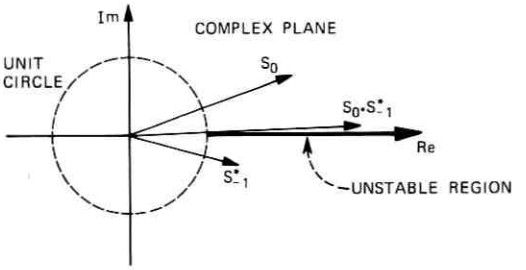


Fig. 2—Schematic diagram of the simplified stability criterion.

It is interesting to note that eq. (10) has the same form as the expression for stability of a single-loop feedback control system, $1 + kG(\omega)H(\omega) = 0$, where $G(\omega)$ is the forward gain and $kH(\omega)$ is the gain of the feedback loop. Thus the modulation index M_1 is analogous to the gain of a feedback loop. This suggests that some concepts of classical control theory may be useful in the present problem.

III. COMPARISON OF OBSERVED SPURIOUS OSCILLATIONS WITH THE STABILITY CRITERION

In this section, an experiment is discussed in which the onset of spurious oscillation was observed and characterized for three diode-circuit cases. Both degenerate and nondegenerate parametric-type oscillations were obtained. The predictions of the simplified stability criterion of eq. (10) are found to be in good agreement with the experiments.

The basic idea of the experiment is to place the diode-circuit system in a state that corresponds to the onset of spurious oscillation. Then that threshold state is characterized by measuring $Z_x(\omega)$, $Z_d(\omega)$, the pump level and the spurious frequencies. The results may be compared with the predictions of the stability theory.

There are several ways in which the diode-circuit system may be placed in the threshold state. One way is to operate the circuit as an amplifier; with the diode bias current and the circuit tuning held constant, the input signal level may be increased until spurious output begins. Another way is to operate the circuit as an oscillator where the diode current is fixed and the tuning is varied in such a way as to increase the oscillation strength until spurious oscillation begins. Still another way is to operate the circuit as a fixed-tuned oscillator and increase the dc bias current, thereby increasing the oscillation strength

until instability occurs. We have experimented with each method and found that the nature of the spurious oscillations appears to be the same in each case. For reasons of experimental convenience, the second method has been adopted, with the results described here.

The experimental setup is shown in the schematic diagram of Fig. 3. The circuit is a circulator-coupled, locked-oscillator-type amplifier designed to operate at 11 GHz with a single 2-watt GaAs IMPATT diode. The oscillation frequency is primarily controlled by the capacitive disc, and the tuning screws are used for impedance trimming. The experimental procedure is to position the screws so that the diode may be biased at full operating current without oscillation of any kind. Next, the tuning screws are adjusted so that oscillation is initiated at about 11 GHz, the pump signal frequency. Then the screws are adjusted to further load the diode so that it delivers progressively more pump power until spurious oscillation begins. At the onset of spurious output, the pump output power and frequency are recorded along with the frequency of the spurious signal or signals, and the circuit impedance is then measured.

Several cases of spurious oscillation were investigated using this procedure. The data characterizing the situation at the onset of instability are summarized in Table I. A more complete discussion of the cases is given below.

Case 1. The diode was biased to $I_{dc} = 300$ mA. In the initial position of the tuning screws, there was no oscillation of any kind. As the screw penetration was increased, the diode began to oscillate near 11 GHz and the output power increased monotonically with screw penetration. There was a slight continuous shift in f_p of about 200 MHz, as the output power was increased up to and somewhat beyond the threshold of spurious output. The output pump power is defined as the power available (at f_p) at the output port of the circulator in Fig. 3.

When the tuning was adjusted past the threshold point, both the subharmonic signal and the pump grew stronger. No additional spurious signals were observed.

At the threshold condition, the output spectrum was carefully searched from 10 MHz to 40 GHz with the following results. No spurious signals were found other than the weak subharmonic signal. In particular, no signal could be detected at the upper sideband frequency, $\omega_1 = 1.5\omega_p$. This provides support for the idea that S_1 may often be negligible, which is the approximation used to reduce eq. (4) to eq. (10). A relatively weak signal was detected at $f = 2f_p$, but not at the third harmonic of the pump. Harmonics of the pump signal are

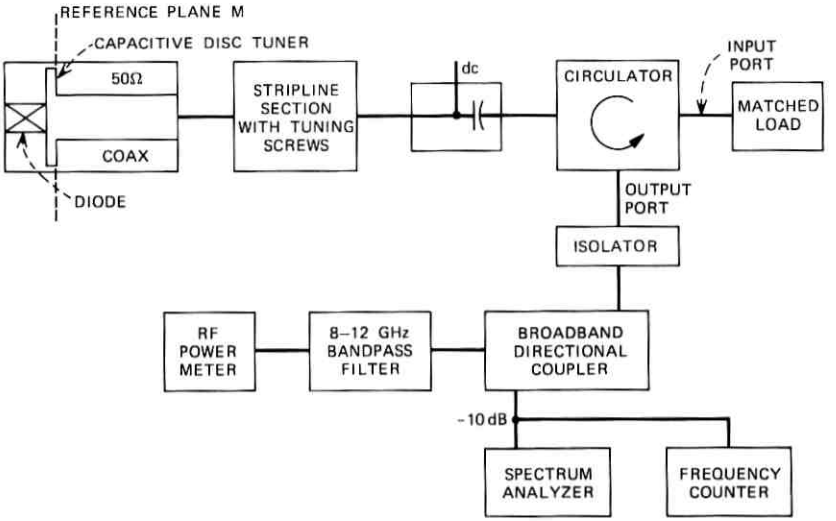


Fig. 3—Schematic diagram of the experimental setup.

not considered to be spurious signals, since in any nonlinear device harmonic components are expected when strongly driven. Note that the second harmonic factor M_2 does not affect the spurious oscillation condition when S_1 is small.

The diode used for this case, diode 15, is a one-sided, uniformly doped, Schottky-barrier, GaAs diode.

Case 2. The circuit used in this case is the same as in case 1; the only difference is in the positions of the tuning screws. The diode used here, diode 2, is similar in type to diode 15. The major difference between

Table I — Summary of experimental data at the threshold of parametric instability

	Case 1	Case 2	Case 3
Diode used	#15	#2	#15
Pump frequency, GHz	11.0988	11.1135	10.94
Pump power, dBm	25.5	28.4	27.0
Spurious type and frequency, GHz	Subharmonic $f_0 = f_{-1}$ = 5.5494	Subharmonic $f_0 = f_{-1}$ = 5.5567	Two-frequency $f_0 = 4.61$; $f_{-1} = 6.33$
I_{dc}, mA	300	300	300
V_{dc}, V	59.7	67.2	59.5

the diodes is that diode 2 has a smaller area (by a factor of 0.73) and, thus, a higher impedance level than diode 15.

The spurious threshold for this case corresponds to nearly 3 dB more output power than for case 1. Otherwise, the two cases are very similar. The pump frequency is essentially the same, the spurious type is subharmonic, and no other spurious signals were observed at the threshold.

Case 3. In this case, diode 15 is used again. The circuit is modified from that shown in Fig. 3 by the inclusion of a shunt line which is $\lambda/4$ long at about 5.5 GHz. The line is connected to the capacitive disc through a small resistor and is terminated with an open circuit at the distant end. The idea is to alter Z_x at the subharmonic without affecting Z_x at the pump frequency.

It is apparent that the desired effect was obtained, in that the diode may now be driven to a higher output level than in case 1 without obtaining spurious tones. Further, when an instability develops, it turns out to be a parametric pair rather than the subharmonic described in case 1. Other than f_0 and f_{-1}, which are given in Table I, no other spurious signals were observed.

It was possible to vary the tuning screws to keep the diode-circuit system at the threshold condition while reducing f_p by about 80 MHz from the value given in the table. As this was done, both f_0 and f_{-1} decreased by approximately 40 MHz each. Thus, apparently neither spurious signal is frequency-dominant in this case.

3.1 Calculation of theoretical spurious threshold

To apply the simplified stability criterion of eq. (10) to predict the spurious threshold for the experimental cases, the diode impedance $Z_d(\omega)$ and the external circuit impedance $Z_x(\omega)$ must be determined. The small-signal impedance of the diodes in question was directly measured for $I_{dc} = 300$ mA by a method that has been described in detail elsewhere.[5] Since Hines' theory is cast in terms of the Read-diode model, what are actually needed are the values of the parameters in the Read-model expression for Z_d [see eq. (8)], which best fit the measured impedance of the real diode. A comparison of the measured small-signal admittance data and the equivalent Read admittance expression is given for both diodes in Figs. 4a and 4b. It is apparent that the agreement is not exact, but it is good enough for our purpose. The corresponding diode parameter values are given in Table II.

The other impedance required is Z_x, the external impedance, which is defined as the impedance seen by the diode wafer. It, therefore,

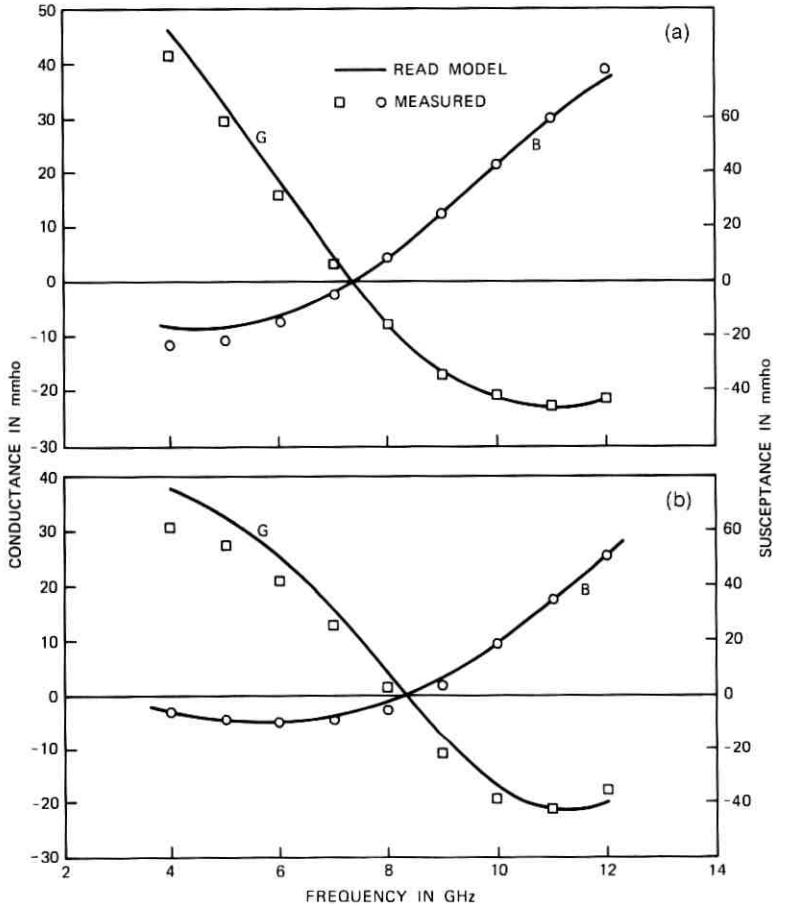


Fig. 4—Comparison of measured small-signal wafer admittance with the Read-model equivalent (I_{dc} = 300 mA). (a) Diode 15. (b) Diode 2.

Table II — Parameter values for the Read-model equivalent of diodes 2 and 15

Parameter	Diode 2	Diode 15
C_T^*, pF	0.86	1.176
f_a, GHz	8.3	7.35
τ_d, *ps*	44.8	41.5
$\tilde{x}_d$, dimensionless	0.83	0.83

* These values of C_T are appropriate for the operating temperature.

must include the parasitic elements contributed by the diode package and mount. The equivalent circuit for this is shown in Fig. 5. The quantity Z_m is the impedance at the reference plane M (see Fig. 3) looking toward the circulator. The element values given in Fig. 5 were determined by making a series of impedance measurements at the plane M looking toward the wafer and replacing the packaged diode with each of several standard impedances. Comparison of the measured element values with the geometry of the circuit reveals that: (*i*) L_s is equal to the inductance associated with the lead wire internal to the package, (*ii*) C_1 is composed partly of the package capacitance and partly of mount capacitance directly in parallel with the conventional C_p, (*iii*) L_M is due to magnetic energy storage within the annular region between the package and the outer conductor, and (*iv*) C_2 is due to the fringing and gap capacitance of the disc. Because of differences in internal construction, $L_s = 0.330\ nH$ for diode 2 and $L_s = 0.278\ nH$ for diode 15. The parameters C_1, C_2, and L_M are the same for each of the three spurious oscillation cases. The impedance Z_M was measured at 100-MHz intervals using a Hewlett-Packard automatic network analyzer. In each case, the tuning screws were in the position that corresponded to the onset of spurious oscillation.

The resulting Z_x for cases 1 and 3 are shown in Smith chart form in Figs. 6a and 6b. In case 2, Z_x is found to be similar but not identical to case 1; the differences are due to different screw positions and different L_s values. An independent corroboration of the measured Z_x may be obtained by observing that, since the diode oscillated at a known frequency in each case, then $-Z_x(\omega_p)$ must be equal to the large-signal diode impedance. Note that $-Z_x(\omega_p) \neq Z_d(\omega_p)$, since the large-signal impedance is somewhat different from Z_d. Such a comparison was made for case 1 using the theoretical large-signal result computed[6] for a very similar diode structure when operated at the same rf efficiency, and the agreement was found to be quite good.

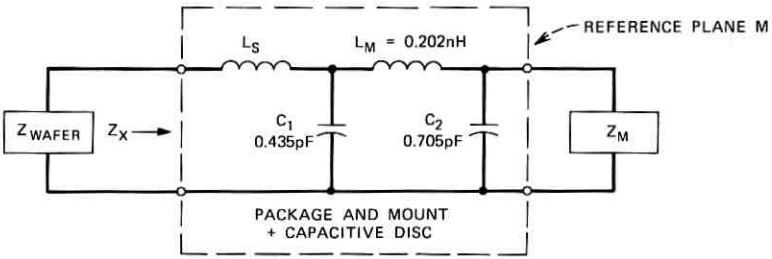
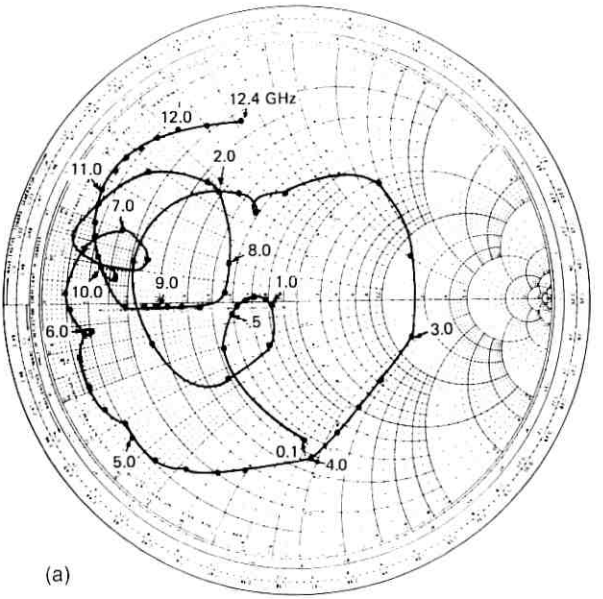
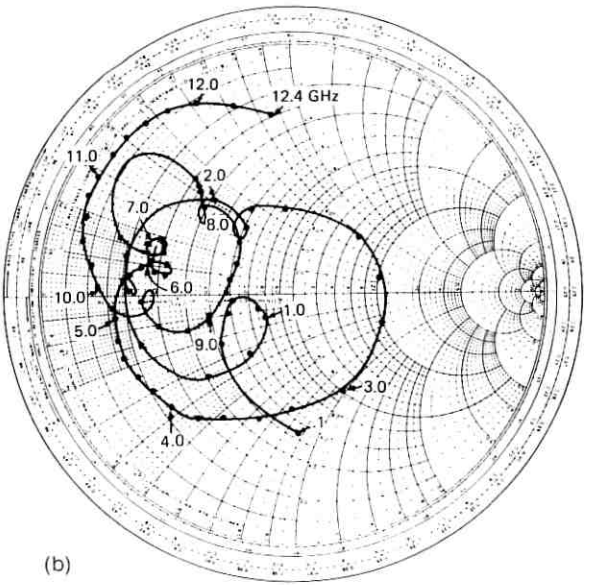


Fig. 5—Equivalent circuit for the diode package and mount.



(a)



(b)

Fig. 6—External impedance normalized to 50 ohms. (a) Case 1. (b) Case 3.

Using the Z_x and Z_d data, the S-factor has been computed for each case. When the pump frequency is established, the quantity $S_0S^*_{-1}$ may then be evaluated. This dimensionless quantity is plotted in Figs. 7a to 7c for each of the three cases.

In case 1, it is evident that there is only one intercept with the real axis. This occurs at $S_0S^*_{-1} = 2.68$ for $f_0 = 5.55$ GHz, i.e., at the subharmonic frequency. Thus, there is only one potential spurious oscilla-

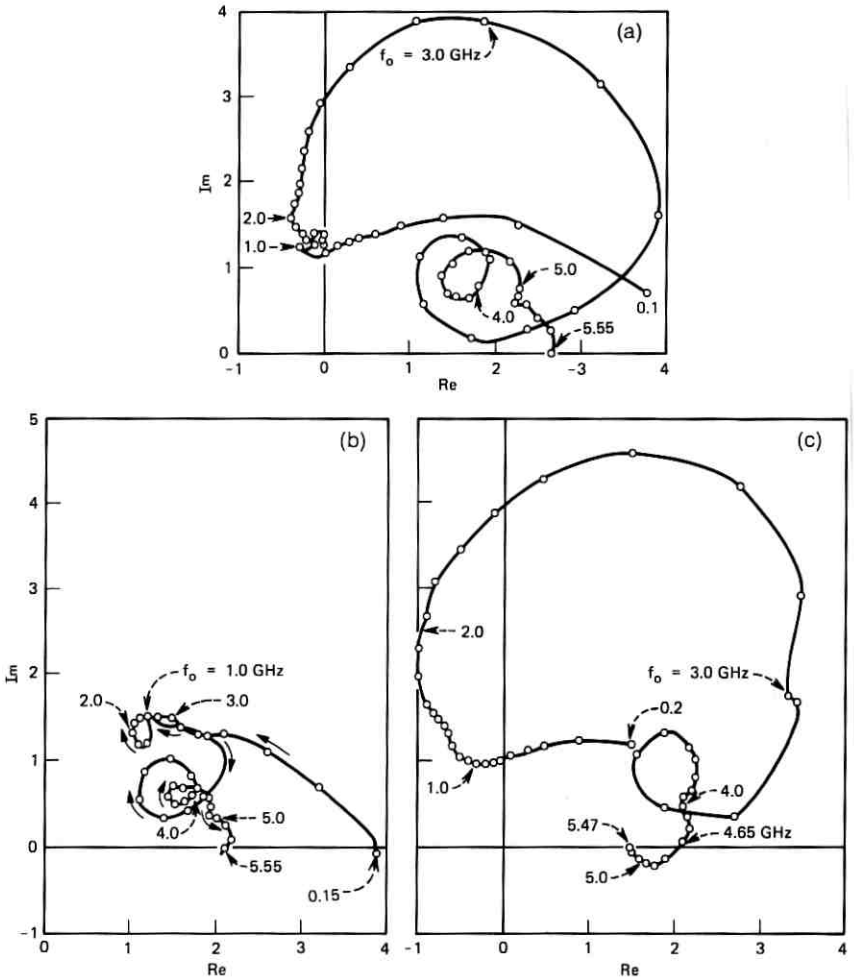


Fig. 7—Stability factor product $S_0S^*_{\pm 1}$ as a function of ω_0. (a) Case 1. (b) Case 2. (c) Case 3.

tion for this case. Since the intercept is larger than one, the subharmonic oscillation is expected to occur for physically attainable pump levels, specifically for $|M_1| \geqq (2.68)^{-\frac{1}{2}} = 0.611$. Experimentally, the spurious oscillation observed was indeed of the subharmonic type, and it began at 25.5-dBm pump power.

In case 2, it is seen that if the behavior of $S_0S^*_{-1}$ at the extreme low frequency limit is neglected for the moment, there is again only one intercept with the real axis. Thus, again the diode is potentially unstable at the subharmonic only, and instability is expected for $|M_1| \geqq 0.678$. Experimentally, subharmonic oscillation was observed to begin for pump power equal to 28.4 dBm. This is significantly higher than the threshold pump power for case 1 and is in agreement with the fact that the predicted threshold M_1 level is higher for case 2.

The general appearance of the $S_0S^*_{-1}$ plot in case 2 is quite different from case 1. This is primarily due to the higher diode impedance level in case 2, since $Z_x(\omega)$ is very similar in both cases. A check of the Z_x data in the vicinity of the subharmonic reveals that they are virtually identical for cases 1 and 2. Thus, the predicted and observed higher degree of stability for diode 2 in the basic amplifier circuit is primarily due to the smaller junction cross-sectional area of diode 2.

Let us now consider the approach of $S_0S^*_{-1}$ toward the real axis for f_0 less than 200 MHz. This behavior indicates a potential instability for $f_0 \cong 150$ MHz and $f_{-1} \cong 10.96$ GHz, which was not experimentally observed. A likely explanation for this is that, for very small ω_0, the upper and lower sidebands are close to ω_p and, therefore, the quantity S_1 is no longer negligible. That is, the more complete stability criterion is probably required for $\omega_0 \ll \omega_{p/2}$. Also, for very small ω_0 the rectification effect which is neglected in this large-signal model becomes important and may affect the stability.

In case 3, there are two intercepts of the real axis:

$$S_0S^*_{-1} = \begin{cases} 2.07 \text{ for } f_0 = 4.65; f_{-1} = 6.29 \text{ GHz} \\ 1.52 \text{ for } f_0 = 5.46 = f_{-1}. \end{cases}$$

Thus, there are two potentially unstable parametric pairs, and it is expected that the nondegenerate pair would be the instability encountered by increasing the pump level until spurious oscillation begins. The associated threshold pump level is $|M_1| = 0.695$. The observed spurious oscillation was a parametric pair at 4.61 and 6.33 GHz, which is considered to be in quite good agreement with the predicted values. The threshold pump power was 27.0 dBm, which is clearly larger than that required to initiate instability for the same

diode in case 1. The latter fact agrees with the predicted relative stability of the two cases.

Thus, there is good agreement between the instability theoretically predicted, using the simplified stability criterion of eq. (10), and that experimentally observed in all three cases.

IV. THEORETICAL STUDY OF CIRCUIT IMPEDANCE SUFFICIENT FOR UNCONDITIONAL STABILITY

As was discussed in the introduction, spurious oscillation is usually an undesirable phenomenon for a number of reasons. Thus, it is of interest to determine what, if anything, can be done to the diode-circuit system to insure that an instability cannot occur. Applying a stability criterion to this synthesis problem is generally more difficult than analyzing a given situation for stability. However, some useful information can be obtained from the simple stability criterion of eq. (10), which was shown in the previous section to predict spurious oscillation accurately for several experimental cases.

To obtain unconditional stability, it is necessary to have an $S_0S^*_{-1}$ function that does not intersect the real axis beyond unity for $0 \leqq \omega_0 \leqq \omega_{p/2}$, as indicated in Fig. 2. It is sufficient for unconditional stability if $S_0S^*_{-1}$ lies within the unit circle. However, the latter condition is not required for stability except in the special case of the subharmonic. Figures 7a to 7c contain numerous examples of frequencies for which $|S_0S^*_{-1}|$ is quite large, yet instability did not occur there, because the phase requirement was not satisfied.

Although the requirement that $S_0S^*_{-1}$ should lie within the unit circle is somewhat restrictive, it has the advantage that one may concentrate on reducing the magnitudes of S_0 and S_{-1} and completely disregard the phase of the stability factors. This simplifies the problem greatly. If the additional requirement is made that S_0 and S_{-1} should each lie within the unit circle, a further simplification results. It is then no longer necessary to consider the pump frequency explicitly since, no matter what the value of ω_p is, $S_0S^*_{-1}$ must lie within the unit disc. Therefore, most of this section is concerned with the conditions that must be satisfied to give $|S(\omega)| \leqq 1$, for $0 < \omega < \omega_p$. Note that it is generally necessary to consider the entire frequency range and not just frequencies below $\omega_{p/2}$ since, no matter how well-controlled S_0 may be, S_{-1} could be such as to yield an unstable product.

The equation for the stability factor, eq. (7), may be solved for Z_x to yield

$$Z_x = -Z_d + \left(\frac{r}{1 + r - S}\right)\left(Z_d - \frac{1}{j\omega C_T}\right), \tag{11}$$

where

$$r \triangleq \left(\frac{\omega_a^2}{\omega^2} - 1\right)^{-1}. \tag{12}$$

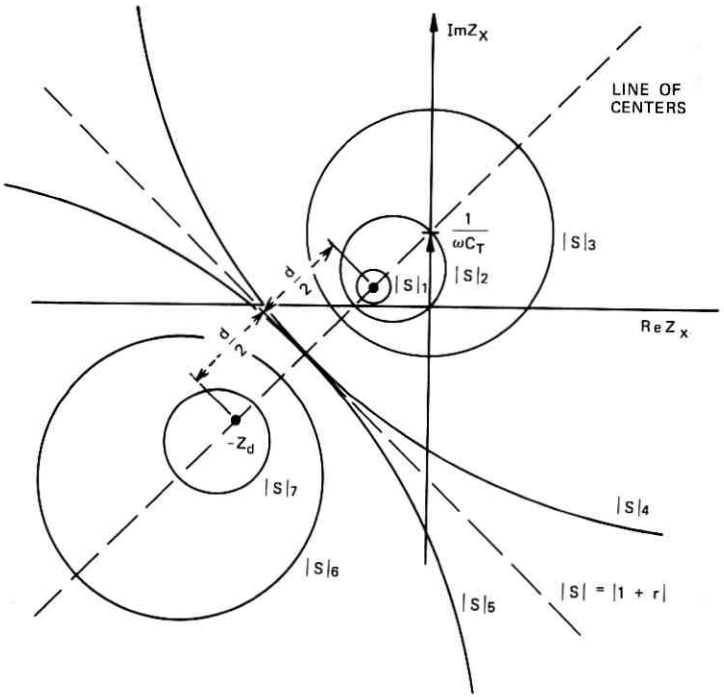
For any set of Z_d, r, ω, and C_T, eq. (11) may be recognized as a bilinear transformation[7] between Z_x and S. Several properties of a bilinear transformation are useful here:

(*i*) A chain of bilinear transformations is also a bilinear transformation.
(*ii*) A bilinear transformation is one-to-one.
(*iii*) A bilinear transformation always maps circles into circles (the straight line is regarded as a circle of infinite radius).

The third property is particularly useful, since it states that circles of constant $|S|$ map into circles in the external impedance plane. It may be shown that the family of constant $|S|$ circles in the Z_x plane are determined by the graphical construction given in Fig. 8.

It is evident from Fig. 8 that there exists a simply connected region in the Z_x plane such that, for any Z_x within this region, $|S| \leqq 1$. Further, the region is bounded by a circle; in some cases, it may be the entire Z_x plane "outside" a circle. The circle in the Z_x plane for $|S| = 1$ has several important properties. It must pass through the point $Z_x = j(1/\omega C_T)$, as may be readily seen from eq. (7). Also, the circle cannot be tangent to the imaginary axis at $Z_x = j(1/\omega C_T)$, but rather must subtend a portion of the positive-real half of the Z_x plane. The latter property can be seen from Fig. 8, where it is evident that the line of centers cannot be horizontal since, in general, Im $(Z_d) \neq 1/j\omega C_T$. Thus, it is apparent that there is always some region of positive-real Z_x such that $|S| \leqq 1$. Therefore, we can contemplate synthesis of a Z_x function sufficient to give stability using only passive elements.

The boundaries of the stable regions of the Z_x plane have been computed for a specific case, and the results are shown in Fig. 9. By "boundary of the stable region" we mean here the locus of Z_x for which $|S| = 1$. The case considered is that of a diode with Read-model parameters; $C_T = 1$ pF, $\bar{x}_d = 0.83$, $\tau_d = 40$ ps, and $f_a = 7.5$ GHz. These parameters are comparable to the experimental diodes of the previous section, i.e., they are representative of a typical X-band GaAs diode. In Fig. 9, the stable regions are shown for a discrete set of frequencies spanning the range from zero through X-band. The boundaries of the stable regions are circular arcs in the Smith chart, since the transformation from rectangular to Smith chart coordinates is another bilinear transformation.



$$|S|_1 < |S|_2 < \cdots < |S|_7$$

$$\beta = \frac{(1 + r)r}{(1 + r)^2 - |S|^2}$$

$$\gamma = Z_d + j\frac{1}{\omega C_T}$$

$$d = \frac{\omega^2}{\omega_a^2}|\gamma|$$

$$Z_{CENTER} = -Z_d + \beta\gamma, \text{ RADIUS OF Z CIRCLE} = \left|\beta\frac{S}{1 + r}\right||\gamma|$$

Fig. 8—Construction of constant $|S|$ circles in the Z_x plane.

A graph such as Fig. 9 may be used to synthesize a Z_x function to give unconditional stability or to check a proposed Z_x in the following manner. If a proposed $Z_x(\omega)$ is such that at $f = 2$ GHz it falls within the 2-GHz circle, at 4 GHz Z_x lies within the 4-GHz circle, and so on for all frequencies $0 < f < f_p$, then the diode for which Fig. 9 is applicable is unconditionally stable in the circuit corresponding to Z_x. That is, it will not develop parametric-type spurious oscillation for any physically attainable pump level.

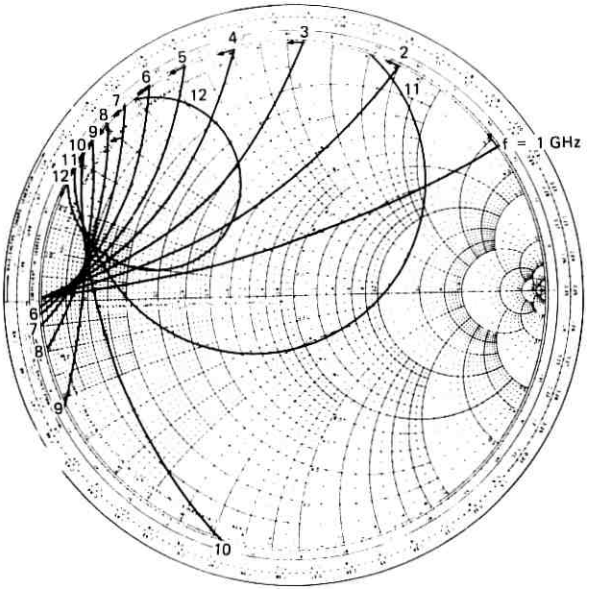


Fig. 9—Stable regions of external impedance for a typical diode. (Arrows indicate stable region. $C_T = 1$ pF, $\bar{x}_d = 0.83$, $\tau_d = 40$ ps, $f_a = 7.5$ GHz, $Z_0 = 50$ ohms.)

The location of the stable regions in Fig. 9 is governed entirely by the four independent parameters of the Read-diode model. Thus, it is of interest to determine the effect of these parameters individually and to construct a more universal diagram. Figure 9 may be interpreted in a more general fashion as follows. Equation (11) may be rewritten as

$$Z_x = \frac{\tau_d}{C_T}\frac{1}{\omega\tau_d}\left[-z + \left(\frac{r}{1+r-S}\right)(z+j)\right]$$
$$= \frac{\tau_d}{C_T}\, g(\bar{x}_d, \omega\tau_d, \omega_a\tau_d, S), \tag{13}$$

where

$$z \triangleq \omega C_T Z_d. \tag{14}$$

In eq. (13), the function g depends only upon the parameters $\bar{x}_d$, $\omega_a\tau_d$, S, and the normalized frequency $\omega\tau_d$. Therefore, it is apparent that Fig. 9 may be applied to *any* diode that has $\bar{x}_d = 0.83$ and $\omega_a\tau_d = 0.6\pi$ radians. Of course, the labelled frequencies must be reinterpreted as normalized frequencies, i.e., $f = 2, 4, \cdots, 12$ GHz corresponds to $\omega\tau_d = 0.16\pi, 0.32\pi, \cdots, 0.96\pi$ radians. Also, the

normalizing impedance needs to be reinterpreted as

$$Z_0 C_T/\tau_d = 1.25 \text{ ohm } F\ S^{-1}. \tag{15}$$

Thus, it is evident that the parameters τ_d and C_T affect only scale factor but not the relative size or location of the stable regions in Fig. 9.

Since a chart such as Fig. 9 is applicable to the family of diodes with fixed $\bar{x}_d$ and $\omega_a\tau_d$, it is worthwhile to consider briefly the usual ranges for these parameters that are obtained in practice. The parameter $\bar{x}_d$ is the fraction of the total depletion layer occupied by the drift region. Although this depends in general upon the semiconductor material and doping profile, for a uniformly doped diode $\bar{x}_d$ is essentially constant over a wide range of doping levels.[6] That is, for a GaAs one-sided, uniformly doped, nonpunch-through IMPATT, it is a reasonable approximation to take $\bar{x}_d = 0.83$ for devices that operate within the approximate frequency range 4 to 20 GHz. The transit time τ_d determines the frequency for optimum large-signal operation. In practice, the diode is usually designed so that $\omega_p\tau_d \cong 0.8\pi$. The avalanche frequency depends upon material parameters and also upon the bias current density. However, in practice, it is usually found that $\omega_p/2 < \omega_a < \omega_p$. Thus, the usual range of $\omega_a\tau_d$ is $0.4\pi < \omega_a\tau_d < 0.8\pi$.

Figures 10a and 10b give the stability charts for $\bar{x}_d = 0.83$ and $\omega_a\tau_d = 0.4\pi$ and 0.8π rad. These normalized charts together with Fig. 9 may be used for studying the stability of the most common type of GaAs IMPATT diodes, namely, one-sided, uniformly doped diodes. For other doping profiles or semiconductor material, it is simple to construct similar charts using eq. (13) or the graphical procedure of Fig. 8. It is evident from Fig. 10b that the region of Z_x which gives $|S| < 1$ is quite small at the higher frequencies. This suggests that diodes with large values of $\omega_a\tau_d$ may be quite difficult to stabilize.

Some additional remarks should be made concerning the choice of a Z_x that is sufficient for unconditional stability. One concerns the possibility of compensation at complementary frequencies. Thus far, the discussion has been devoted to finding $Z_x(\omega)$ so that $|S(\omega)| \leqq 1$ for $0 < \omega < \omega_p$. In some practical situations, there may be a portion of this frequency range where it is undesirable or impossible to keep $|S(\omega)| \leqq 1$, perhaps because of the influence of package parasitics. In such a situation, it is still possible to guarantee stability by introducing compensation at the complementary frequency. For example, if for some $\omega_0 |S_0| > 1$, the S-product still lies within the unit disc if $|S_{-1}| < |S_0|^{-1}$. Thus, it may be useful to construct circles in the Z_x plane corresponding to $|S|$ other than unity, to assist in the design of

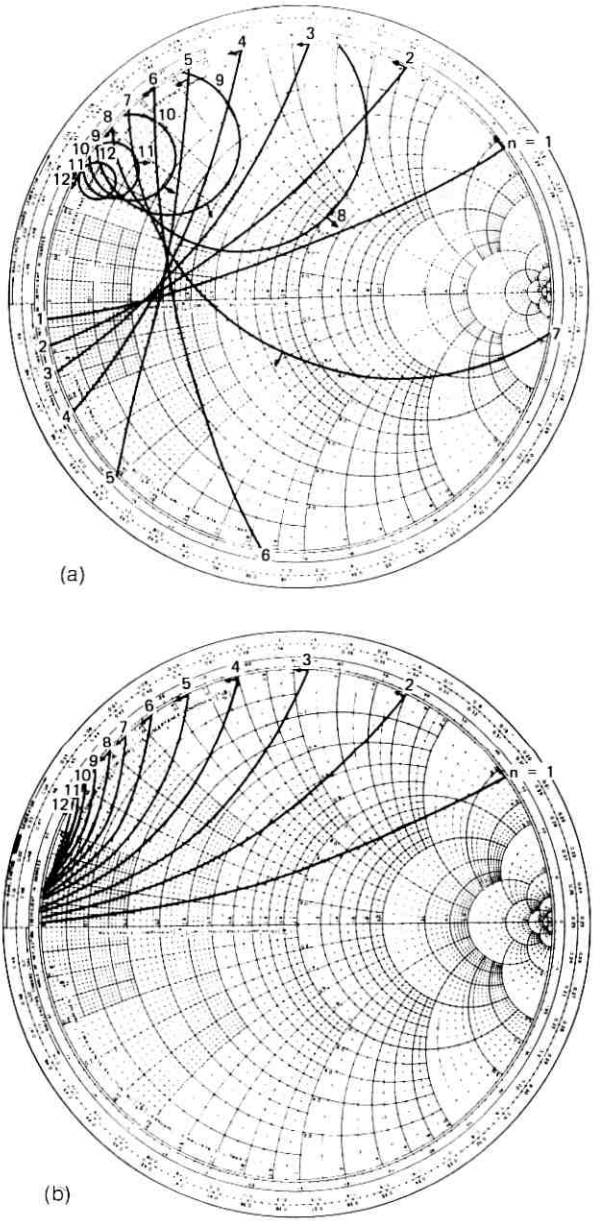


Fig. 10—Normalized stability charts for external impedance. (Arrows indicate stable regions. $Z_0C_T/\tau_d = 1.25$ ohm F S^{-1}.) (a) $\tilde{x}_d = 0.83$, $\omega_a\tau_d = 0.4\pi$. (b) $\tilde{x}_d = 0.83$, $\omega_a\tau_d = 0.8\pi$.

a compensated Z_x. One drawback to any compensation procedure, however, is that the complementary frequencies are related by ω_p. Therefore, a diode-circuit system that is compensated at ω_{p0}, i.e., that has $S_0 S^*_{-1}$ within the unit circle for ω_{p0}, may be unstable for pump frequencies that are sufficiently different from ω_{p0}.

Furthermore, it is not necessary for unconditional stability that $|S_0 S^*_{-1}| \leqq 1$, except at the subharmonic frequency. In applications where this condition is too restrictive of Z_x, it is possible to analyze proposed circuits for stability and study the effect of various circuit parameters. The stability criterion of eq. (10) may be used or, since such an analysis is most conveniently done by computer, the more general expression of eq. (4) may be used with little more cost.

A question arises concerning the compatibility of the impedance requirement resulting from Hines' theory of parametric interaction and the impedance requirements for preventing bias-current oscillation which were recently discussed by Brackett.[2] The origin of the oscillation which Brackett considered is in the dc negative resistance produced in an IMPATT by the large-signal rectification properties of the avalanche process. This is a different physical process from the one Hines considers and, in fact, Hines assumed the induced dc resistance to be zero. Brackett concluded that, to prevent bias-circuit oscillations, the diode should see a high impedance below some cutoff frequency f_c. The value of f_c depends upon the Q of the microwave circuit, but generally f_c is less than about 500 MHz. It may readily be seen from eq. (3) that, for $f < f_c$, the factor $\omega^2/(\omega_a^2 - \omega^2)$ becomes so small that $S \cong 1$ for any finite value of Z_x. Thus, Z_x may be chosen to satisfy Brackett's requirements without appreciably affecting the stability factor.

V. CONCLUSIONS

The aspects of Hines' recent theory of parametric interactions of IMPATT diodes that are most pertinent to the study of stability in strongly driven oscillators and amplifiers have been reviewed. A simple, approximate stability criterion has been developed from Hines' theory. An experiment was performed in which several cases of spurious oscillation were observed, and the diode and circuit impedances were carefully characterized for each case. The predictions of the stability criterion were found to be in good quantitative agreement with the experimental observations in each case.

Using the simple stability criterion, a procedure has been developed for determining a range of circuit impedance sufficient for unconditional stability, i.e., stability under all physically attainable drive

levels. The desired circuit impedance is expressed in terms of the Read-model equivalent of the IMPATT diode in question. A set of normalized stability charts has been given that may be used for designing stable circuits for a class of GaAs IMPATT diodes. The procedure is also given for constructing similar charts for other types of diodes.

The desired circuit behavior appears to be synthesizable using positive-real impedance functions. Finally, it has been shown that the circuit requirements are compatible with the restrictions described by Brackett to prevent bias-circuit oscillations and low-current burnout.

VI. ACKNOWLEDGMENTS

The author wishes to express his gratitude to M. E. Hines for providing a preprint of his paper. The author is also grateful to N. R. Dietrich for the use of his amplifier circuit and for his invaluable cooperation in the measurements. The diode-admittance characterization was performed by R. L. Frank and H. M. Olson. The guidance and suggestions of J. W. Gewartowski throughout this work are gratefully acknowledged.

APPENDIX

This appendix compares the values of modulation index M_1 which were measured at the threshold of spurious oscillation to those predicted by the stability criterion, for the cases discussed in Section III. The measured values of M_1 are calculated as follows:

$$|M_1| \triangleq \frac{|I_o|}{2I_{dc}},$$

where

$$\begin{aligned} I_o &= (1 - \bar{x}_d - j\beta\bar{x}_d)^{-1} I_e, \\ \beta &= (1 - e^{-j\omega_p \tau_d})/\omega_p \tau_d, \\ I_e &= I_T(Y_d - j\omega_p C_T)/Y_d, \\ |I_T| &= (2P/R_x)^{\frac{1}{2}}, \\ Y_d &= (-Z_x)^{-1}, \end{aligned}$$

and

I_o = injected current (*full* wave amplitude),
I_e = induced current,
I_T = terminal current,
Y_d = diode large-signal admittance at the pump frequency,
P = output power at the pump frequency,
R_x = Re (Z_x).

Table III — Modulation index M_1 at threshold of spurious oscillation

	Measured	Theoretical
Case 1	0.49	0.61
Case 2	0.69	0.68
Case 3	0.58	0.69

The magnitude of the modulation index was evaluated using the data given in Section III. The results are compared to the theoretical spurious threshold in Table III. The agreement is considered to be reasonably good, when the approximations of the stability criterion and the diode model are considered.

REFERENCES

1. M. E. Hines, "Large-Signal Noise, Frequency-Conversion and Parametric Instabilities in IMPATT Diode Networks," Proc. IEEE, *60*, No. 12 (December 1972), pp. 1534–1548.
2. C. A. Brackett, "The Elimination of Tuning-Induced Burnout and Bias Circuit Oscillations in IMPATT Oscillators," B.S.T.J., *52*, No. 3 (March 1973), pp. 271–306.
3. W. T. Read, Jr., "A Proposed High-Frequency Negative Resistance Diode," B.S.T.J., *37*, No. 2 (March 1958), pp. 401–446.
4. M. Gilden and M. E. Hines, "Electronic Tuning Effects in the Read Avalanche Diode," IEEE Trans., *ED-13* (January 1966), pp. 169–174.
5. C. N. Dunn and J. E. Dalley, "Computer-Aided Small-Signal Characterization of IMPATT Diodes," IEEE Trans. MTT, *17*, No. 9 (September 1969), pp. 691–695.
6. W. E. Schroeder and G. I. Haddad, "Nonlinear Properties of IMPATT Devices," Proc. IEEE, *61*, No. 2 (February 1973), pp. 153–182.
7. R. V. Churchill, *Complex Variables and Applications*, 2nd ed., New York: McGraw-Hill, 1960, pp. 73–80.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 7, September 1974
Printed in U.S.A.

Theory of Noise in Charge-Transfer Devices

By K. K. THORNBER

(Manuscript received January 8, 1974)

The noise introduced into charge packets transferred through and stored in charge-transfer devices is calculated in a manner that includes all important relaxation, suppression, and correlation effects. First, the noise induced into each packet during each transfer phase from thermal, trapping, emission-current, and leakage-current fluctuations, whose statistics are nonstationary, and from clock-voltage fluctuations, whose statistics are stationary, is determined. Relaxation of the transferring charge to these fluctuations is found to suppress their size. Second, the accumulation (collecting) of the noise as each packet is transferred through the device is calculated neglecting the role of incomplete charge transfer. Attention is drawn to the significant differences between the collecting of storage-process noise, which is unsuppressed, transfer-process noise, whose spectral density is nearly totally suppressed at low frequencies, and modulation noise, which is nearly totally suppressed for digital and analog signals. Third, the role of incomplete charge transfer in suppressing the collecting of the noise is shown for digital signals and indicated for analog signals. We conclude with a numerical calculation of the maximum possible signal-to-noise ratio that can be expected from charge-transfer devices. The presentation is sufficiently general and detailed that, with a minimum of background in formal noise theory, one can use the approach to evaluate noise in many novel, solid-state devices.

I. INTRODUCTION

Charge-transfer devices (CTD's), such as the bucket-brigade[1] (BB) and charge-coupled-device[2] (CCD) shift registers, are currently of great interest. These devices consist of a chain of charge-storage elements along which charge packets are transferred from input to output. Noise accompanying each individual transfer of each charge

packet will be introduced into each packet. It is the purpose of this article to calculate this noise and its cumulative effects on the output signal.

Noise generated in solid-state devices has been treated extensively,[3–6] and this prior work will be of great assistance to us here. However, there are three major, significant differences between much of this prior work and the present treatment.[7,8] The first difference arises from the nature of the transfer process.[9] The usual treatments of noise discuss situations in which the (noiseless) currents, charge densities, conductances, etc. associated with the signal are time-invariant. Under such conditions, the statistics of the noise are also time-invariant and the noise is said to be stationary. Stationary noise is readily treated in the frequency domain using spectral-density functions. Frequency-domain, linear-circuit analysis of equivalent circuits greatly facilitates the usual treatments. (Sometimes, as in mixer theory, periodic rather than time-invariant situations are considered, and frequency-domain analyses are still convenient.)

In CTD's the situation is quite different. During the transfer of the charge from one storage region to the next, the (noiseless) currents, charge densities, conductances, etc. associated with the transfer of the signal packet are rapidly time-varying.[9] As a result, the noise generated during transfer is nonstationary; that is, the statistics of the current fluctuations that give rise to noise vary appreciably with time. Once the transfer is complete, a certain amount of noise has been introduced into the signal independently of the noise acquired in prior or subsequent transfers. The processes of interest, therefore, are nonperiodic as well as nonstationary. Under such conditions, we have found it best to work in the time domain using correlation functions.

The second basic difference between this treatment and other treatments arises from the physical structure and operation of the device. Ordinarily, we can externally control the noiseless portion of the voltages and currents associated with the signal. In CTD's, however, the size of the charge packets, in addition to the characteristics of the clock voltages, controls the charge transfer. Thus, the noise currents that we usually calculate function as additional driving terms in calculating charge transfer and, of course, result in fluctuations in the sizes of the packets. But it is the reaction of the charge transfer to these currents, and not the currents themselves, from which we must calculate charge-packet fluctuations.[10,11] (Normally, we control either the voltage across, or the current through, the device of interest

and calculate fluctuations in the other. Here we control neither directly.)

A third basic difference is that it by no means suffices to calculate only the noise introduced during a single transfer. The collecting[7] of the noise must also be carefully calculated to include correlation effects that lead to a suppression in the cumulative noise added to each packet while being transferred by the device. The number of unexpected effects makes treating noise in CTD's truly interesting.[12]

At first the nonstationary feature of the noise coupled to the reaction of the charge transfer might seem to preclude a reasonably simple, analytic treatment that produces useful results. Indeed, some of our expressions will be a bit complicated. However, owing to certain suppression effects, which will be discussed in detail, the mean-square induced fluctuation in the size of a given charge packet at the end of a single transfer is nearly independent of the size of the packet. Thus, having calculated the magnitude of this fluctuation, it will be possible to obtain a simple, meaningful expression for the spectral density of the accumulated noise at the output which is independent of the signal. This quantity is quite useful for evaluating the effects of noise on the analog performance of a CTD. By contrast, for digital-performance evaluation, only the accumulated mean-square fluctuation in the size of a charge packet at the output is needed. As we shall see, this quantity is arrived at in a straightforward manner using our time-domain analysis without the necessity of working in the frequency domain at all.

In what follows, we shall outline briefly the noise sources whose effect on the output charge packets we shall calculate. Then, following a review of the lumped-charge model of a CTD, which has proven to be so useful in the analysis and calculation of incomplete transfer coefficients, we calculate the statistics of the noise charge introduced in a single transfer in terms of the statistics of the microscopic fluctuations inducing this noise. Thermal, trapping, clock-voltage, emission-current, and leakage-current fluctuations are considered. Three different types of compounding, storage-process, transfer-process, and modulation, are then treated neglecting incomplete-transfer effects. Following this, the general problem of compounding in the presence of incomplete charge transfer is treated and interesting suppression effects are uncovered. The paper ends with calculations of the maximum signal-to-noise ratio expected for state-of-the-art devices. As the techniques employed in calculating nearly all aspects of the noise are

of necessity different from those usually encountered, we have chosen to elaborate our methods in some detail. From a knowledge of the noise at the output, the operational limitations[13] and error rates[14] associated with CTD's can be assessed.

The remainder of this article is long and intricate. This arises not from the nature of the noise treated per se, but rather because of the complexity of the devices considered. Nonetheless, familiarity with the physical operation of CTD's as well as acquaintance with Brownian motion and shot noise are the only prerequisites needed. No use is made of highly developed, sophisticated techniques of noise theory. This is not because such techniques are not considered applicable; rather, it is felt that the methods used here are the simplest available for a rigorous treatment. In addition, it is felt that these methods will be useful in treating noise in other dynamic devices as well. Readers well versed in noise theory may feel that this problem is amenable to existing methods. While the contrary is not claimed here, we believe that such treatments will be more involved than might be expected at first sight.

II. SOURCES OF RANDOM NOISE

We should at the outset stress that we are concerned here only with *random* noise. We are not concerned with signal distortion arising from sampling, incomplete transfer,[15,16] direct clock coupling, nonideal regeneration, or any other deterministic process. This is not because we feel that such problems are unimportant. Rather, our philosophy is that random noise is unavoidable, whereas, in principle, deterministic "noise" can be greatly reduced or compensated for by careful design.[13,14] Thus, it is random noise that plays a major role in setting the operational limits of CTD's, which is a problem of general current interest.

Figure 1 indicates schematically the sources of CTD noise with which we shall be concerned. At the input, if the charge packet is created by photon absorption as in imaging application, or if it is injected by an emission-limited mechanism, full shot noise $[\langle (Q - Q_0)^2 \rangle = eQ_0]$ will accompany the signal Q. If, on the other hand, the packet is injected into the input via a resistor-controlled circuit, as in regeneration and storage applications, no shot noise will be introduced.

The CTD itself simultaneously transfers charge and stores charge.[8] In its charge-transfer capacity, thermal noise from the Brownian motion of the carriers composing the transfer current and trapping noise from the fluctuation in occupancy of interface states[17] arise. Care

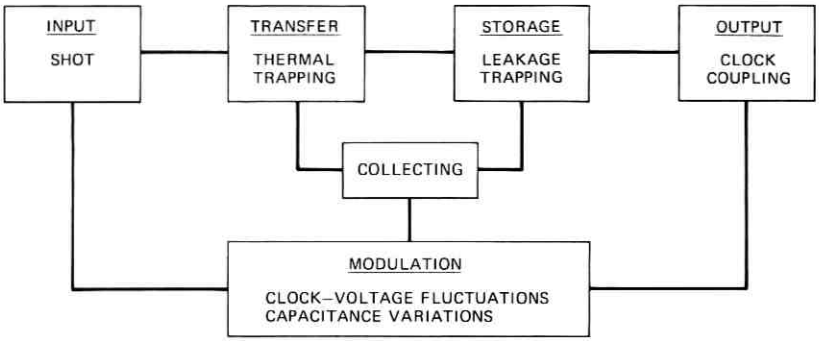


Fig. 1—Sources of noise for a CTD.

must be exercised in calculating this transfer noise because of the tendency of relaxation effects to suppress thermal and trapping fluctuations.[7] Since the transfer of charge from one storage region to the next is controlled by the conductance of the storage regions themselves (or portions thereof), shot noise in the transfer current is totally suppressed for all practical purposes. (Exception: Should such current become barrier-limited, then some shot noise can result. This special case is treated in Section 6.1.) In its capacity of charge storage, noise from leakage-current fluctuations and trap-occupancy fluctuations is introduced. Although intrinsically much smaller than transfer-noise sources, storage noise is unsuppressed, and, hence, it can be important in some cases. At the output, we have the usual problems with detector noise, but we shall not consider these problems at this time.

What is of basic interest is, of course, the cumulative noise in the charge packet by the time it reaches the output of the device. As it turns out, one must be very careful in calculating the collecting of the noise introduced into each packet during each phase of each transfer cycle. For example, if at the end of a transfer phase, a transfer noise of $+\Delta Q$ has been added to the signal, by conservation of charge, a quantity of charge $-\Delta Q$ has been added to the charge left behind. These two contributions to the noise are highly correlated, and this correlation must be taken into account in calculating the noise spectral density at the output.[7] Incomplete transfer of charge distorts the noise as well as the signal, and must also be included in collecting effects. Fluctuations in the clock voltage coupled to fluctuations in the sizes of the storage capacitances of the individual storage cells of the CTD give rise to modulation noise, which collects quite differently

from the noise induced by fluctuations in the clock voltage in the absence of cell-parameter variations along the CTD.

These, then, are the aspects of the noise produced by CTD's that we shall discuss. In calculating these various contributions to CTD noise, we were surprised by the variety of the results: the presence or absence of suppression in individual transfer phases, differences in compounding suppression, stationary fluctuations in charge-packet fluctuations induced by nonstationary noise sources, etc. Athough we by no means treat all aspects of CTD noise, the methods we develop should be helpful in calculating the influence of nearly any source of noise on the output of a CTD.

III. MODEL OF CHARGE-TRANSFER DEVICE

All device noise arises ultimately from fluctuations inherent in the transfer of charge carriers between states characterizing the flow of charge through the device.[18] It follows that to calculate CTD noise, we must first understand how a CTD operates under noiseless conditions.[10,11] As this topic has been the subject matter of a large number of papers,[10,11] a detailed elaboration here is not necessary. We shall, however, briefly review a lumped-charge model[10,11] of charge transfer within a single transfer unit of a CTD, which has proven useful in discussing and calculating incomplete charge transfer in CTD's. This model enables us to express the device current as a function of several characteristic voltages, which in turn control the charge passing through the device. Using this model, we can then calculate fluctuations in the sizes of the transferred charge packets in terms of the current fluctuations which accompany the charge transfer and which can be calculated by standard means.[18]

The lumped-charge models we shall use to calculate CTD noise are shown in Figs. 2a and 2b. Charge Q_s stored on capacitor C_s is transferred to capacitor C_d during the transfer cycle. (The clock voltage which determines the duration of the transfer cycle is V_c.) In subsequent cycles, the charge on C_d is transferred to the right step by step to other storage sites, and charge from the left is brought into C_s—both processes are modeled by repetitions of this model for a single transfer unit. For the present, we shall be concerned only with a single transfer unit, either of type (a), if the charge transfer is characterized by a single-step process, or of type (b), if the transfer is characterized by a two-step process. In the latter process, an intermediate capacitor C_b is inserted between C_s and C_d to enhance their

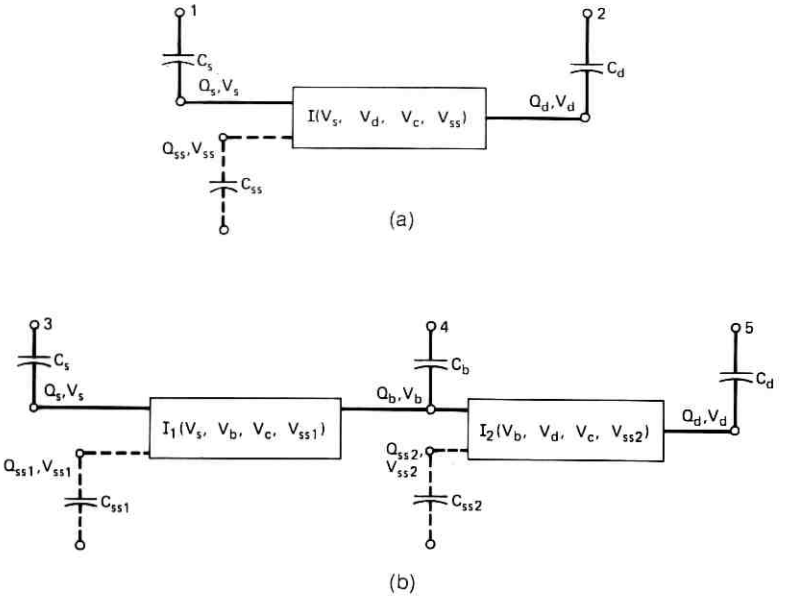


Fig. 2—Lumped-charge model including the effect of interface states for the transfer of charge through a single stage of a CTD. (a) Single-step transfer. (b) Two-step transfer.

mutual isolation so as to reduce the incomplete transfer. As we shall see, this also has the effect of reducing the noise acquired by the transferred packet.

Let us now review the dynamic features of the lumped-charge model of Fig. 2a in some detail. Stored on $C_s = C_s(V_c, V_d, V_c, V_{ss})$ is a time-dependent quantity of stored charge Q_s given by

$$Q_s = \int_{V_{so}}^{V_s} C_s(V_s', V_d, V_c, V_{ss}) dV_s'; \tag{1a}$$

stored on $C_{ss}(V_s, V_d, V_c, V_{ss})$ is a relatively small, time-dependent quantity of interface (trapped) charge Q_{ss} given by

$$Q_{ss} = \int_{V_{sso}}^{V_{ss}} C_{ss}(V_s, V_d, V_c, V_{ss}') dV_{ss}'. \tag{1b}$$

The total charge Q to be transferred through the conductance $I = I(V_s, V_d, V_c, V_{ss})$ is given by

$$Q = Q_s + Q_{ss}. \tag{1c}$$

Stored on C_d is a charge Q_d given by

$$Q_d = Q_o - Q = \int_{V_{do}}^{V_d} C_d(V_s, V_d', V_c, V_{ss}) dV_d', \tag{1d}$$

where Q_o is the total charge to be transferred through this single transfer unit. (Q_o is some constant for each transfer event: it is the quantity of charge initially distributed between Q_s and Q_{ss} to be transferred to C_d.) The time dependence of the decay of Q_s during the charge transfer is governed by the equation

$$\dot{Q}_d = -\dot{Q} = I(V_s, V_d, V_c, V_{ss}) + i_n(t), \tag{2}$$

where $i_n(t)$ is the device noise current induced by microscopic fluctuations within the conductance I. To determine CTD noise we must (i) solve eqs. (1) and (2) for fluctuations in Q_d induced by $i_n(t)$ (and by fluctuations in V_c and in trap occupancy), and (ii) express $i_n(t)$ in terms of the microscopic fluctuations from which it is induced.

For two-step transfer processes, as modeled in Fig. 2b, the dynamics of the charge transfer are more complicated. One must now be concerned with Q_s, Q_b, Q_d, Q_{ss1}, and Q_{ss2} defined by

$$Q_s = \int_{V_{so}}^{V_s} C_s(V_s', V_b, V_c, V_{ss1}) dV_s', \tag{3a}$$

$$Q_b = \int_{V_{bo}}^{V_b} C_b(V_s, V_b', V_d, V_c, V_{ss1}, V_{ss2}) dV_b', \tag{3b}$$

$$Q_d = \int_{V_{do}}^{V_d} C_d(V_b, V_d', V_c, V_{ss2}) dV_d', \tag{3c}$$

$$Q_{ss1} = \int_{V_{ss10}}^{V_{ss1}} C_{ss1}(V_s, V_b, V_c, V_{ss1}') dV_{ss1}', \tag{3d}$$

$$Q_{ss2} = \int_{V_{ss20}}^{V_{ss2}} C_{ss2}(V_b, V_d, V_c, V_{ss2}') dV_{ss2}'. \tag{3e}$$

In addition, we have

$$Q = (Q_s + Q_{ss1}) + (Q_b + Q_{ss2}) \equiv Q_1 + Q_2, \tag{3f}$$

and

$$Q_d = Q_o - Q = Q_o - (Q_1 + Q_2). \tag{3g}$$

The corresponding dynamic equations governing the time dependence

of Q_s, Q_b, and Q_d are

$$\dot{Q}_1 = \dot{Q}_s + \dot{Q}_{ss1} = -I_1(V_s, V_b, V_c, V_{ss1}) - i_{n1}(t), \tag{4a}$$

$$\dot{Q}_2 = \dot{Q}_b + \dot{Q}_{ss2} = I_1(V_s, V_b, V_c, V_{ss1}) + i_{n1}(t) - I_2(V_b, V_d, V_c, V_{ss2}) - i_{n2}(t), \tag{4b}$$

$$\dot{Q}_d = I_2(V_b, V_d, V_c, V_{ss2}) + i_{n2}(t), \tag{4c}$$

where again the noise currents i_{n1} and i_{n2} are the device noise currents induced by microscopic fluctuations within the conductances I_1 and I_2, respectively. To calculate the fluctuations in Q_d, we shall solve eqs. (3) and (4) for the influence of i_{n1}, i_{n2}, and fluctuations in V_c and in trap occupancy Q_{ss1} and Q_{ss2}, on Q_d.

IV. NOISE INTRODUCED IN ONE CHARGE-TRANSFER CYCLE

Let us now inquire into the influence of device noise on the size Q_d of a charge packet transferred through one single-step transfer element of a CTD. (Owing to the additional complexity associated with a two-step transfer element, we shall relegate the noise treatment of this case to Appendix A). It is the purpose of the remainder of this section to write the mean-square fluctuation in Q_d, q_d, in terms of the independently fluctuating voltages and currents characterizing the transfer and storage process. By expressing all dynamic quantities in terms of a minimum number of independent ones at the outset, we can greatly reduce the number of cross-correlations which we must eventually include.

If we assume that the fluctuations themselves are sufficiently small, then we can linearize eqs. (1) and (2) about their time-dependent, noiseless solutions. The form taken by eq. (2) upon expanding to lowest order in the fluctuations (and subtracting out the noiseless portion) is

$$\dot{q}_d = -\dot{q} = \frac{\partial I}{\partial V_s} v_s + \frac{\partial I}{\partial V_d} v_d + \frac{\partial I}{\partial V_c} v_c + \frac{\partial I}{\partial V_{ss}} v_{ss} + i_n, \tag{5}$$

in which lower-case letters indicate the fluctuation portion of the quantity of interest, e.g., $Q = Q^o + q$, $Q_d = Q_d^o + q_d$, etc. Since v_s and v_d depend upon q, we must also linearize eq. (1), which will yield this dependence. In calculating incomplete charge transfer, it was necessary to pay special attention to the dependence of C_s and C_d on V_s, V_d, V_c, V_{ss}. This was because most terms in the small-signal expansion of Q_s and Q_d contributed to the incomplete transfer of charge. For our

purposes here, however, we can recognize such terms as small; that is, we can drop terms of the order of incomplete transfer times noise compared to terms of the order of unity times noise. As incomplete transfer per transfer is at most 3×10^{-3} in devices of interest, this is certainly warranted. Proceeding in this manner, we obtain

$$q = C_s(V_s^o, V_d^o, V_c^o, V_{ss}^o)v_s - C_s(V_{so}^o, V_d^o, V_c^o, V_{ss}^o)v_{so} + C_{ss}(V_s^o, V_d^o, V_c^o, V_{ss}^o)v_{ss} - C_{ss}(V_s^o, V_d^o, V_c^o, V_{sso}^o)v_{sso} \equiv C_s v_s - C_s^o v_{so} + C_{ss} v_{ss} - C_{ss}^o v_{sso} \quad (6a)$$

and

$$q_o = -q = C_d(V_s^o, V_d^o, V_c^o, V_{ss}^o)v_d - C_d(V_s^o, V_{do}^o, V_c^o, V_{ss}^o)v_{do} \equiv C_d v_d - C_d^o v_{do}. \quad (6b)$$

It is these equations that must be solved for v_s and v_d in terms of q, v_c, and v_{ss}.

At this point a minor subtlety enters. In Fig. 2a, contact 1 is tied to the clock voltage V_c. Since a fluctuation v_c in V_c cannot instantaneously alter the amount of charge stored on C_s, fluctuations v_s and v_{so} must be correlated, and indeed this correlation is contained in eq. (1a). A similar argument also applies to contact 2, v_d and v_{do}, and eq. (1d). (Since the capacitance C_{ss} is not tied directly to the clock, we need not concern ourselves with trapped charge at this point.) It follows from eqs. (1a) and (1d) that

$$v_{so} = \frac{\partial V_{so}}{\partial V_c} v_c = \frac{\partial V_s}{\partial V_c} \frac{C_s}{C_s^o} v_c = (+1) \frac{C_s}{C_s^o} v_c \quad (7a)$$

and that

$$v_{do} = \frac{\partial V_{do}}{\partial V_c} v_c = \frac{\partial V_d}{\partial V_c} \frac{C_d}{C_d^o} v_c = (-1) \frac{C_d}{C_d^o} v_c. \quad (7b)$$

[In (7b), $\partial V_d / \partial V_c = -1$ because it is $-V_c$ that is connected to contact 2.] With respect to the traps, since V_{ss} is the effective level to which traps are occupied during the transfer cycles, V_{sso} is unaffected by the dynamics of the operation of the device and, hence, $v_{sso} = 0$. We may now solve eq. (6a) for v_s, (6b) for v_d, and insert into eq. (5) to obtain

$$\dot{q}_d = -\left(\frac{g_m}{C_s} + \frac{g_r}{C_d}\right) q_d + \left(g_m + g_r + \frac{\partial I}{\partial V_c}\right) v_c - \left(g_m \frac{C_{ss}}{C_s} - \frac{\partial I}{\partial V_{ss}}\right) v_{ss} + i_n, \quad (8a)$$

from which it follows at once that

$$q_d(t) = \int_0^t dt' \exp\left[-\int_{t'}^t dt''/\tau(t'')\right]\left[\left(g_m + \frac{\partial I}{\partial V_c}\right)\Big|_{t'} v_c(t') - \left(g_m \frac{C_{ss}}{C_s} - \frac{\partial I}{\partial V_{ss}}\right)\Big|_{t'} v_{ss}(t') + i_n(t')\right], \quad (8b)$$

where we have taken $g_r \ll g_m$ and defined the time-dependent relaxation time τ by

$$1/\tau \equiv g_m/C_s. \quad (8c)$$

(The capacitances C_s and C_d are of comparable size.) In eq. (8), g_m is the forward conductance $(\partial I/\partial V_s)$ and g_r is the reverse conductance $(-\partial I/\partial V_d)$. Because of the inherent undirectionality, $g_m \gg g_r$. The statistical distribution of $q_d(t)$, which we seek, can be determined from (8b) and the statistical properties of v_c, v_{ss}, and i_n.

Our expression for the noise fluctuation in the transferred charge given in eq. (8) has several interesting features which should be carefully noted. The charge fluctuation that accumulates is not simply the integral of the noise current over the time interval. Rather, it is the integral of the noise current times a damping (relaxation) factor. Physically, this arises because if an excess quantity of charge $-q(=q_d)$ has been transferred, then the subsequent current (which depends most strongly on q_s) is reduced from what it would have been in the absence of the fluctuation. The reduced current causes less charge to flow from C_s to C_d, which in turn tends to partially compensate (null out) the effect of the previous fluctuation. This leads to a suppression of the device noise, which is shown explicitly in eq. (8b). During the initial portion of the charge-transfer cycle, the transfer current is largest, as is the conductance g_m. Hence, although the larger the current, the more noise is present in the charge transfer, because the damping is also largest initially, the effect of this noise on the transferred charge can be expected to be greatly suppressed. On the other hand, near the end of the charge-transfer cycle, while the noise from the transfer of charge and its accompanying noise is much reduced in size, so is the damping. As it turns out, it is the noise produced during the end of the transfer cycle that is most important. This is convenient, because it means that the size of the noise produced by the end of the transfer cycle is independent of the size of the initial charge packet (as long as some charge is transferred).

Of primary interest is the mean-square fluctuation in the transferred charge $\langle q_d(t_f)^2 \rangle$, where t_f is the time at the end of the transfer cycle.

Since the fluctuations v_c, v_{ss}, and i_n are mutually independent, as we shall see when we determine their statistical distribution, it follows that

$$\langle q_d(t_f)^2 \rangle = \int_0^{t_f} dt_1 \int_0^{t_f} dt_2 \exp\left[-\int_{t_1}^{t_f} dt_1'/\tau(t_1') \right] \times \exp\left[-\int_{t_2}^{t_f} dt_2'/\tau(t_2') \right] \left[\left(g_m + \frac{\partial I}{\partial V_c} \right)\Big|_{t_1} \left(g_m + \frac{\partial I}{\partial V_c} \right)\Big|_{t_2} \times \langle v_c(t_1) v_c(t_2) \rangle + \left(\frac{1}{\tau} - \frac{\partial I}{\partial Q_{ss}} \right)\Big|_{t_1} \left(\frac{1}{\tau} - \frac{\partial I}{\partial Q_{ss}} \right)\Big|_{t_2} \times \langle q_{ss}(t_1) q_{ss}(t_2) \rangle + \langle i_n(t_1) i_n(t_2) \rangle \right]. \quad (9)$$

In (9), we have used $C_s v_{ss} = q_{ss}$ and $g_m/C_s = 1/\tau$. Thus, to determine $\langle q_d^2 \rangle$, we must calculate the autocorrelation function of v_c, v_{ss}, and i_n. For stationary noise, such quantities are well known.[3-6] The purpose of the next section will be to calculate these autocorrelation functions for the nonstationary conditions that enter the present problem. The reader should note carefully at this point, moreover, that the factors multiplying these correlation functions involve time-dependent quantities characteristic of the detailed (but noiseless) solutions to the nonlinear device equations [eq. (1)]. Although further simplifications can be made in some cases, it should be evident that, in general, one must understand the noiseless problem in order to do the problem with noise.

Two other important aspects of our results (8b) for $q_d(t)$ and (9) for $\langle q_d(t_s)^2 \rangle$ are these. First, as a result of the suppression factor, $q_d(t_f)$ and, hence, $\langle q_d(t_f)^2 \rangle$ are for all practical purposes independent of the size of the signal. Owing to the suppression, $q_d(t_s)$ depends most strongly on details of the charge transfer for $t \approx t_f$. However, considering the typical size of α, the coefficient of incomplete charge transfer ($\alpha \leqq 10^{-3}$), the details of the charge transfer for $t \approx t_f$ can deviate by only about 10^{-3}, which for our purposes is quite insignificant. Second, $q_d(t_f)$ is a stationary random variable, even though its statistics must be derived from the nonstationary distributions of v_c, v_{ss}, and i_n. These two results greatly simplify our treatment of compounding of the noise in Section V.

(Note that, to avoid undue complication, we have left out leakage current into the storage regions, as well as the noise associated with this current. As this noise source is uncorrelated with the other sources considered above, we can treat it in a separate section. When this component of noise is included, it is no longer the case that $q_d = -q$

[eq. (6b)], and, hence, greater care must be exercised when we sum the noise added with each transfer. This effect also shows up in the distinction between transfer process and storage noise, which is discussed in more detail in Section V.)

V. NOISE INTRODUCED BY MICROSCOPIC FLUCTUATIONS

In the preceding section, we expressed the mean-square fluctuation in the charge transferred during a single transfer cycle in terms of the autocorrelation functions of the various contributions to the device noise. In this section, we shall calculate these correlation functions and then use the results to estimate the noise of each type introduced during a single transfer cycle. Our task is eased considerably because of the extensive effort that has already gone into the study of noise in solid-state devices.[3–6]

Although at first sight the character of the charge transfer from C_s to C_d in a CCD appears to be rather different from that in an IGFET bucket-brigade device, in fact, the two types of charge transfer can be treated in a similar manner. We shall carry over this similarity in treating the device noise associated with $i_n(t)$: we shall make use of the understanding available of noise in IGFET's and then apply these results to the CCD as well. In so doing, we must (and shall) be careful to make note of certain important differences between CCD-mode and BB-mode transfers which can affect the noise calculated.

5.1 Thermal noise

Of the primary sources of noise present in IGFET's—thermal noise at high frequency,[19] generation-recombination ($g - r$) noise at intermediate frequency,[20] and $1/f$ noise at low frequency[21]—by far the most important source of noise associated with $i_n(t)$ is the thermal noise arising from the Brownian motion of the charge carriers in the inverted region of the semiconductor, which forms the conductance I. (This is evident because the spectral densities of $1/f$ and $g - r$ noise have dropped considerably from their peak values by 10^5 Hz, the lower bound on CTD transfer frequencies.) (In CCD's the proximity to the semiconductor-insulator interface of the charge being stored as well as transferred greatly enhances the role of interface trapping on $\langle q_d^2 \rangle$. In fact, for CCD's, it appears that interface noise is the dominant form of device noise in these devices. For reasons that are apparent in the next subsection, we shall treat this contribution as a trapping noise associated with v_{ss} rather than as $1/f$ noise associated with i_n.)

Of the early theoretical work on thermal noise in IGFET's, the treatment by Jordan and Jordan[22] seemed clearest to us on first reading. These authors note that a spontaneous current fluctuation $i_s(x, t)$ at x along the channel of the IGFET ($0 \leqq x \leqq L$) will give rise to a voltage fluctuation $v_s(x, t) = i_s(x, t)dx/\mu\sigma(x, t)$, where μ is the mobility of the carriers and $\sigma(x, t)$ is the mobile charge per unit length of channel at time t. The spontaneous voltage fluctuation $V_s(x, t)$ will in turn induce an x-dependent voltage fluctuation all along the channel, which in its turn induces a fluctuation in the source-to-drain current i_d given by

$$i_d(t) = \frac{\mu}{L}\sigma(x, t)v_s(x, t). \tag{10a}$$

Making use of the above relation between v_s and i_s, we obtain

$$i_d(t) = i_s(x, t)dx/L. \tag{10b}$$

[The same result can be obtained using the impedance field method (IFM).[23] Alternatively, one can develop a current-current method analogous to the current-voltage method employed in the IFM. This is outlined briefly in Appendix B.] The contribution to the noise current $i_n(t)$ due to $i_d(t)$ induced by fluctuations all along the channel is from (10b)

$$i_n(t) = \int_0^L i_s(x, t)dx/L, \tag{11a}$$

so that

$$\langle i_n(t_1)i_n(t_2)\rangle = \int_0^L dx_1 \int_0^L dx_2 \langle i_s(x_1, t_1)i_s(x_2, t_2)\rangle/L^2. \tag{11b}$$

The autocorrelation function of $i_s(x, t)$ for thermal noise can be found from that of the current density $j_s(\mathbf{x}, t)$ obtained from microscopic noise theory [18] and given by

$$\langle j_s(\mathbf{x}_1, t_1)j_s(\mathbf{x}_2, t_2)\rangle = 2kT\mu\rho(\mathbf{x}_1, t_1)\delta(\mathbf{x}_1 - \mathbf{x}_2)\delta(t_1 - t_2), \tag{12}$$

where $\rho(\mathbf{x}_1)$ is the charge density at $\mathbf{x}_1$. Inserting (12) into (11b) and noting that $\rho(\mathbf{x}) = \sigma(x)/A(x)$, $A(x)$ being the cross-sectional area of the channel at x, we obtain

$$\langle i_n(t_1)i_n(t_2)\rangle = 2kT\delta(t_1 - t_2)\int_0^L \mu\sigma(x, t_1)dx/L^2. \tag{13a}$$

The integral in (13) must be found from a knowledge of the noiseless operation of the device at time t_1. In general, the integral is directly

proportional to the forward conductance[22] so that

$$\langle i_n(t_1)i_n(t_2)\rangle = 2kT\delta(t_1 - t_2)g_m(t_1)H_n(t_1), \tag{13b}$$

where $H_n = 1$ for the IGFET operated in the linear region, $H_n = \frac{2}{3}$ for the IGFET operated in saturation, and $H_n = \frac{1}{2}$ for diffusion-limited current.[11] In most cases of interest, the CTD is operated in saturation so that $H_n = \frac{2}{3}$. For very long clock periods, $H_n = \frac{1}{2}$, appropriate near the end of the cycle, can be used. As the 30 percent uncertainty between $\frac{1}{2}$ and $\frac{2}{3}$ is tolerable at present for noise calculations in these devices, we shall not concern ourselves with the additional dependence of H_n on the time t_1. (Setting $H_n = \frac{2}{3}$ and $H_n = \frac{1}{2}$ gives adequate upper and lower bounds on the thermal noise.)

[Although (13b) was derived for the bucket-brigade type of charge transfer, we shall assume that it is valid for CCD-type transfer as well, where, of course, one uses g_m appropriate to the CCD device.[10,11] This assumption clearly breaks down, however, when $\langle q_d^2\rangle^{\frac{1}{2}}$ becomes comparable to or larger than the amount of *free* charge incompletely transferred in a single transfer phase. The reason for this difficulty goes back to the assumption of linearity in eq. (5). This, of course, also applies, *mutatis mutandis*, to all other noise contributions as well. Treating these interesting nonlinear problems is, unfortunately, beyond the scope of this paper.]

Returning now to (9), we can calculate the contribution to $\langle q_d^2\rangle$ arising from thermal noise during a single transfer cycle. If we set $H_n = \frac{2}{3}$, ignore the time dependence of C_s (a 0.1 percent effect if $\alpha = 10^{-3}$), and recall that $\tau^{-1} = g_m/C_s$, inserting (13b) into (9) and integrating over t_1 and t_2, we obtain

$$\langle q_d^2\rangle|_{kT} = \tfrac{2}{3}kTC_s \tag{14a}$$

in the limit that

$$\exp\left[-\int_0^{t'} dt'/\tau(t')\right] \ll 1,$$

as it must if the incomplete transfer α is to satisfy $\alpha \ll 1$. (In deriving the above, we have not assumed that g_m is independent of time.) A corresponding result was first given by Boonstra and Sangster;[24] however, their $\langle q_d^2\rangle$ was four times larger than the right-hand side of (14a). We believe that (14a) is in fact correct.

[If one ignores the nonstationarity of the noise, a result similar to that of Boonstra and Sangster can be obtained in the following manner. For thermal noise, the mean-square current fluctuation is given by

$\langle i_d^2 \rangle = 4kT\frac{2}{3}g_m B$, where g_m is the conductance and B is the bandwidth. If $\tau = C_s/g_m$ is the characteristic relaxation time, then $B = \tau^{-1}$ and $\langle q_d^2 \rangle = \langle i_d^2 \rangle \tau^2 = 4kT\frac{2}{3}C_s$. (The quantities $\langle q_d^2 \rangle$ and $\langle i_d^2 \rangle$ are mean-square fluctuations, not Fourier coefficients. The quantity $\langle i_d^2 \rangle$ is obtained from the spectral density of i_d using the Wiener-Khinchin theorem.) While this result is not too bad, attempting to calculate noise in more complicated, nonstationary situations using stationary results can lead to trouble. For example, for a two-step transfer process, we obtain for the thermal contribution to $\langle q_d^2 \rangle$ the result [from eq. (62)]

$$\langle q_d^2 \rangle|_{kT} = \tfrac{2}{3}kTC_b + \tfrac{2}{3}kTC_s(1 + b^{-1})^{-1}, \tag{14b}$$

where $b = (g_{m2}/C_b)(g_{m1}/C_s)^{-1}$, and is assumed to be independent of time. This result is more of a challenge to obtain from arguments based on stationary noise sources.]

5.2 Interface (trapping) noise

We now focus attention on calculating the statistics of

$$q_{ss}(t) = C_{ss}(t)v_{ss}(t),$$

the fluctuations in occupancy of the interface states during charge transfer. To proceed, we must first write down the dynamic equations governing the trapping. These are simple since we are working in the lumped-charge approximation. If greater accuracy is desired, one can be more microscopic and include the position and energy dependence of the trapping states, as well as their capture cross section and thermal-release time. The procedure is similar but much more involved.

The dynamic equation relating the flow of charge into and out of the traps is

$$\dot{Q}_{ss} = I_t(I, Q_{ss}) - I_r(Q_{ss}) + i_{nss}, \tag{15}$$

where the trapping current I_t is given by

$$I_t = I(\sigma_x/A)(Q_{ss}^t - Q_{ss})/Q_{ss}^t, \tag{16a}$$

and where the release current I_r is given by

$$I_r = Q_{ss}/\tau_r(t). \tag{16b}$$

In (15), i_{nss} is the noise current associated with the filling and emptying of traps; in (16a), I is the transfer current, σ_x/A is the ratio of the effective cross section of all the traps to the area A through which I flows, and the factor $(Q_{ss}^t - Q_{ss})/Q_{ss}^t$ is the fraction of the total number of active sites Q_{ss}^t that are empty; in (16b), τ_r is the thermal-release

time. Writing $Q_{ss} = Q_{ss}^{o} + q_{ss}$, expanding to first order in the noise q_{ss}, and subtracting off the zero-order equation, we obtain for the noise

$$\dot{q}_{ss} = \frac{\partial I_t}{\partial I}\frac{\partial I}{\partial Q_{ss}} q_{ss} + \frac{\partial I_t}{\partial Q_{ss}} q_{ss} - \frac{\partial I_r}{\partial Q_{ss}} q_{ss} + i_{nss} \tag{17a}$$

$$= -\left[-\frac{I_t}{I}\frac{\partial I}{\partial Q_{ss}} + \frac{1}{\tau_t(t)} + \frac{1}{\tau_r(t)} \right] q_{ss} + i_{nss} \tag{17b}$$

$$= -q_{ss}/\tau_{ss}(t) + i_{nss}, \tag{17c}$$

where τ_t is the effective trapping time $(\partial I_t/\partial Q_{ss})$, and where τ_{ss}^{-1} equals the bracketed quantity in (17b). (Ordinarily $I_T \ll I$, so that the first term in τ_{ss}^{-1} can be neglected. Under these circumstances, the noise contribution from different types of traps will be independent, and the traps may be treated independently.)

Proceeding now, we solve (17c) for $q_{ss}(t)$, obtaining

$$q_{ss}(t) = \int_0^t dt' \exp\left[-\int_{t'}^t dt''/\tau_{ss}(t'') \right] i_{nss}(t'). \tag{18}$$

To obtain the correlation function of $q_{ss}(t)$, we must calculate $\langle q_{ss}(t_1) q_{ss}(t_2) \rangle$. The resulting expression involves the correlation function of $i_{nss}(t)$, which, since i_{nss} is an elementary microscopic process, is given by[18]

$$\langle i_{nss}(t_1) i_{nss}(t_2) \rangle = e[I_t^o(t_1) + I_r^o(t_1)]\delta(t_1 - t_2), \tag{19}$$

where e is the charge on an elementary carrier. (The above is not obtained from a spectral density, since for a time-dependent current, a nonstationary process, spectral density is not defined.) For $t_2 > t_1$, it follows that

$$\langle q_{ss}(t_1) q_{ss}(t_2) \rangle = \exp\left[-\int_{t_1}^{t_2} dt/\tau_{ss}(t) \right] \cdot \int_0^{t_1} dt \exp\left[-\int_t^{t_1} dt'/\tau_{ss}(t') \right] e[I_t^o(t) + I_r^o(t)]. \tag{20}$$

If this result is inserted into (9), summation over all trapping states yields the contribution to the transfer noise arising from these states. As is evident from (20), in general, one must know the noiseless solution, especially $\tau_{ss}(t)$, $I_t^o(t)$ and $I_r^o(t)$, in order to actually calculate the noise. A few simplifications, however, permit us to recover Tompsett's result[17] for a single-step, CCD-mode transfer, which is valid in the limit of $I_t \ll I_r$, i.e., near the end of a transfer cycle.

To calculate the contribution to $\langle q_d^2 \rangle$ arising from interface-state trapping, we must, according to (9), know not only $\langle q_{ss}(t_1) q_{ss}(t_2) \rangle$, which we found above (20), but also $(\partial I / \partial Q_{ss})$. This quantity entered into a previous discussion[10] of the role of surface states on incomplete transfer. There it was noted that one could write

$$\partial I / \partial Q_{ss} = -a(t)/\tau(t), \tag{21}$$

where for CCD-mode transfer $a = 0$ and for bucket-brigade-mode transfer a falls from $a = C_s/C_{ch}$ during most of the transfer cycle (C_{ch} being the channel capacity) to $a \approx 0$ toward the end of the cycle, where the IGFET current becomes emission limited. [The vanishing of a for CCD-mode operation is due to the fact that the field felt by a mobile carrier due to another carrier is independent of whether the other carrier is free or trapped. The large value of a for BB-mode transfer arises from the fact that relatively small changes in the threshold voltage, induced by changes in trap occupancy in the channel region, can result in relatively large changes in transfer current. If a is large, it enters (9) as a^2, while, as we shall see, $\langle q_{ss}^2 \rangle$ is proportional to C_{ss} ($\ll C_{ch}$ for the IGFET channel). This implies that the ratio (R) of the surface-state-noise contribution to $\langle q_d^2 \rangle$ for bucket-brigade to that for CCD is C_s/C_{ch}. If, on the other hand, the BB-mode is not turned off until the channel current becomes emission limited, then $a \approx 0$. If we assume that the suppression factors in (9) damp out the noise introduced while $a \gg 1$, then the above ratio (R) becomes C_{ch}/C_s, which means that the contribution to $\langle q_o^2 \rangle$ of q_{ss} can be ignored for BB-mode transfer. We shall assume that this is the primary operational region of interest, pointing out, however, that if the BB-transfer mode is terminated while a is large, one should expect an increase in that portion of $\langle q_d^2 \rangle$ arising from trapping.]

Returning to CCD-mode transfer where we expect interface states to play the largest role, if we assume in eq. (9) that $\langle q_{ss}(t_1) q_{ss}(t_2) \rangle$ varies in time slowly compared to $\tau(t)$, then we can integrate over t_1 and t_2 to obtain

$$\langle q_d(t_f)^2 \rangle \approx \langle q_{ss}(t_f)^2 \rangle. \tag{22}$$

If, in addition, we focus attention on long transfer cycles (10^{-6} second or longer), taking $I_\tau^o \gg I_t^o$ and $\tau_t \gg \tau_\tau$, then from (20) we obtain

$$\langle q_{ss}(t_f)^2 \rangle = e \int_{t_c}^{t_f} dt \cdot \exp\left[-2 \int_t^{t_f} dt'/\tau_\tau(t') \right] Q_{ss}^o(t)/\tau_\tau(t) + \langle q_{ss}(t_c)^2 \rangle. \tag{23}$$

If we make one more assumption that at time $t_c (0 < t_c < t_f)$, $Q_{ss} = Q_{ss}^o = Q_{ss}^i$, and for subsequent $t (t_c < t < t_f)$, Q_{ss}^o decays as if $I_t = 0$, then

$$Q_{ss}^o(t) = Q_{ss}^i \exp\left[- \int_{t_c}^{t} dt'/\tau_r(t') \right], \tag{24a}$$

$$\langle q_{ss}(t_c)^2 \rangle = 0. \tag{24b}$$

Inserting these into (23), we find that

$$\langle q_{ss}(t_f)^2 \rangle = qQ_{ss}^i \exp\left[- \int_{t_c}^{t_f} dt'/\tau_r(t') \right] \left\{ 1 - \exp\left[- \int_{t_c}^{t_f} dt'/\tau_r(t') \right] \right\}, \tag{24c}$$

as found by Tompsett[17] using other reasoning. Should conditions be such that any of these assumptions are unwarranted, then, of course, the result [(20) in (9)] is more complicated.

The result given in (24c) has the following significance. If on the average $\tau_r \gg (t_f - t_c)$ or $\tau_r \ll (t_f - t_c)$, then $\langle q_{ss}^2 \rangle$ for such states is negligible. In the former case, the transfer process occurs too rapidly for the traps to respond; in the latter case, the trap occupancy can follow the transfer current quite closely; in either case, the noise is greatly suppressed as a consequence. Thus, only when $\tau_r \approx t_f - t_c$ can the traps influence the charge transfer.

If we now sum over the distribution of traps, assumed uniform in energy E, and take τ_r^{-1} proportional to $\exp(-E/kT)$, then following Strain[25] we obtain

$$\langle q_{ss}^2 \rangle = e^2 kTN_{ss}A \ln 2, \tag{24d}$$

Tompsett's result,[17] where N_{ss} is the number of interface states per unit energy per unit area, and A is the active trapping area. Like Tompsett,[17] we conclude that the noise introduced into Q_d from trapping has a mean-square value of

$$\langle q_d^2 \rangle = e^2 kTN_{ss}A \ln 2, \tag{25}$$

that is,

$$\langle q_d^2 \rangle = \langle q_{ss}^2 \rangle.$$

5.3 Clock-voltage noise

The influence of fluctuations in clock voltage on the noise added to the transferred charge in a single transfer can be treated very quickly, especially if it is white as we shall assume. If the spectral density of

the noise at zero frequency is $S_v(0)$, then [18]

$$\langle v_c(t_1)v_c(t_2)\rangle = \tfrac{1}{2}S_v(0)\delta(t_1 - t_2). \tag{26}$$

From (9) it follows that the contribution to $\langle q_d^2\rangle$ from clock-voltage fluctuations is

$$\langle q_d^2\rangle|_{\text{clock}} = \tfrac{1}{2}S_v(0)\int_0^{t_f} dt \cdot\exp\left[-2\int_t^{t_f} dt'/\tau(t')\right](g_m + \partial I/\partial V_c)^2|_t. \tag{27a}$$

We recall that $g_m = C_s/\tau$. The direct dependence of I on V_c depends strongly on the type of device, and one must be careful not to include in $\partial I/\partial V_c$ terms already included in $(\partial I/\partial V_s)(\partial V_s/\partial V_c)$. Thus, for a single-step bucket brigade, $\partial I/\partial V_c = +g_m$, while for a single-step CCD, $\partial I/\partial V_c = 0$. If we set $\partial I/\partial V_c = bg_m$, then

$$\langle q_d^2\rangle|_{\text{clock}} = \tfrac{1}{2}S_v(0)C_s^2(1+b)^2\int_0^{t_f} dt \cdot\exp\left[-2\int_t^{t_f} dt'/\tau(t')\right]\tau^{-2}(t). \tag{27b}$$

In general, (27b) must be evaluated from a knowledge of $\tau(t)$ obtained from the noiseless charge-transfer characteristic. If, however, we assume that we can write $\partial I/\partial Q_s = dI/dQ_s$, then

$$dt/\tau(t) = g_m dt/C_s = dt\partial I/\partial V_s/C_s = dt\partial I/\partial Q_s = dtdI/dQ_s = -dI/I.$$

It follows that

$$\langle q_d^2\rangle|_{\text{clock}} = \tfrac{1}{2}S_v(0)C_s^2(1+b)^2\int_{I_f}^{I_o}\frac{dI}{I}\left(\frac{I_f}{I}\right)^2\frac{1}{\tau(I)}, \tag{27c}$$

where $I_f = I(t = t_f)$ and $I_o = I(t = 0)$. In this form it is clear how the integral yields the effective bandwidth B of the white clock-voltage noise. One can replace the τ^{-1} in (27c) by dI/dQ_s, which in turn may be calculated as a function of I using eq. (8) of Ref. 11 ($Q_s = C_sV_s$). For our purposes, we shall be content with

$$\langle q_d^2\rangle|_{\text{clock}} = \tfrac{1}{2}S_v(0)BC_s^2(1+b)^2. \tag{27d}$$

Of course, if the clock-voltage noise is not white, (26) should be replaced with the actual correlation function, which then introduces an additional time-dependent function in the integrand of (27). Relation (27d) will be useful in discussing modulation noise.

VI. OTHER NOISE SOURCES

In the preceding section, we discussed the three primary sources of noise that contribute to the fluctuations in the size of a single charge packet as the result of a single transfer. In this section we examine two other sources of noise which do not readily fit into (9) without undue complication. For simplicity, we shall not include the thermal, trapping, or clock-voltage fluctuations discussed in Section V, as these can be superimposed linearly on the results discussed below.

6.1 Emission-limited-current shot noise

In describing the types of noise expected to be generated in CTD's, we noted that shot noise would *not* play a significant role under ordinary circumstances. This is because, in simplest terms, a single transfer unit of a CTD is like two capacitors connected by a resistor, and, owing to strict charge neutrality in the resistor, shot noise is not present. If, however, the clock period t_0 of an IGFET CTD becomes long ($t_0 > 10^{-6}$ second), the channel current becomes partially emission limited at the source towards the end of the transfer cycle.[11] In this case the shot noise associated with the emitted current will not be totally suppressed. We shall now show how this can be treated. At the same time it will become clear why "ideal" resistors totally suppress shot noise.

In Fig. 3 we represent the barrier region between the diffused source and the channel by a conductance with current I_e and the channel by a conductance with current I. The voltage V_s at the source end of the channel is set V_a as in Ref. 11. In Section V we ignored I_e, assuming its conductance was much greater than that of I. We now consider the more realistic situation in which the conductances of I_e and I are comparable near the end of the transfer cycle. (In such circumstances, the g_m introduced in Section IV and used extensively in Section V, must be replaced by the series conductance of I_e and I in the expressions for thermal, trapping, and clock-voltage noise.)

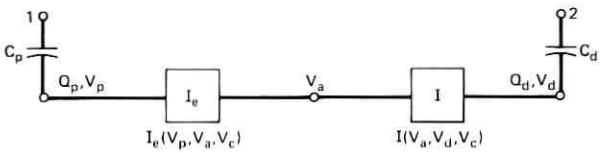


Fig. 3—Lumped-charge model for a CTD including a barrier conductance I_e.

What is happening physically can be outlined as follows. Voltages V_a as well as V_p and V_d can fluctuate in response to fluctuations induced in the conductances. If a small, spontaneous current fluctuation i_s (positive, say) occurs through I_e, then V_a will increase, simultaneously inducing a (positive) fluctuation i in I and a (negative) fluctuation i_e in I_e. The net current fluctuation is $i = i_s + i_e$. If the conductance of I_e is much larger than that of I, then V_a will change very little, $i \approx 0$, and hence $i_e \approx -i_s$. In other words, a current i_e is induced by i_s which nearly cancels i_s; with the larger conductance of I_e, a small v_a induces an i_e sufficient to suppress the i_s fluctuation, while with the smaller conductance of I, v_a induces a much smaller i and I. In the limit of an "ideal" resistor, since it is the bulk which controls the current and not the contacts, $i \to 0$ in the limit of zero contact resistance, and the shot noise is totally suppressed.

Let us now calculate the shot noise introduced during emission-limited operation. We start as before with the equations of the model:

$$I_e(V_p, V_a) = I_e(V_p - V_a), \tag{28a}$$

$$I = I(V_a, V_d, V_c), \tag{28b}$$

$$Q_s = \int_{V_{po}}^{V_p} C_s(V_p', V_a, V_c), \tag{28c}$$

$$-\dot{Q}_s = I_e + i_s = I. \tag{28d}$$

In (28a) we have assumed that I_e is primarily a function of $(V_p - V_a)$, anticipating that for barrier current, I_e is well-described by a diode equation. Since the variations of I with V_d or V_c and of C_s with V_a or V_c are small, we shall ignore these dependences as we did in Section IV. Setting $I_e = I_e^o + i_e$ and $I = I^o + i$, it follows from (28) that

$$i_e = \frac{\partial I_e}{\partial V_p}(v_p - v_a), \tag{29a}$$

$$i = \frac{\partial I}{\partial V_a} v_a, \tag{29b}$$

$$q_s = C_p v_p, \tag{29c}$$

$$-\dot{q}_s = i_e + i_s = i. \tag{29d}$$

Solving for q_s in terms of i_s, we obtain

$$-\dot{q}_s = \frac{\partial I}{\partial V_a} \frac{(\partial I/\partial V_p)(q_s/C_p) + i_s}{\partial I/\partial V_a + \partial I/\partial V_p}, \tag{30a}$$

or letting $g_m = \partial I/\partial V_a$, $g_b = \partial I/\partial V_p$ and solving for q_s,

$$q_d = -q_s = \int_0^t dt' \exp\left[-\int_{t'}^t dt''/\tau_b(t'')\right][i_s(1 + g_b/g_m)^{-1}]|_{t'}, \quad (30b)$$

where $\tau_b = [(g_m^{-1} + g_b^{-1})C_p]$. This expression tells us the fluctuation induced in the transferred charge packet q_d resulting from the shot noise associated with the current emitted into the channel from the source. In the $g_b \to \infty$ limit (zero barrier resistance), $q_d \to 0$. In the $g_b \to 0$ limit (infinite barrier resistance), $q_d \to \int i_s dt$, full, unsuppressed shot noise.

In deriving (30b) our primary goal was to determine $\langle q_d^2 \rangle$. Since $\langle i_s(t_1) i_s(t_2) \rangle = eI^o(t_1)\delta(t_1 - t_2)$,

$$\langle q_d^2 \rangle|_{\text{shot}} = \int_0^{t_f} dt \exp\left[-2\int_t^{t_f} dt'/\tau_b(t')\right] eI^o(t) \cdot [1 + g_b(t)/g_m(t)]^{-2}. \quad (31)$$

To evaluate (31) requires, in general, a knowledge of the noiseless solution to the charge-transfer equations. We can obtain a feel for the result if we recall van der Ziel's expression for the current spectral density of diode shot noise:[6]

$$S_i(f) = 2e(I^o + I_o) = 2kTg_b \frac{I^o + 2I_o}{I^o + I_o}. \quad (32a)$$

(I_o is the diode leakage current that we have neglected.) Although derived for stationary noise, it is straightforward to redo the derivation for nonstationary noise to obtain an autocorrelation function of the form

$$\langle i_s(t_1) i_s(t_2) \rangle = e[I^o(t) + I_o]\delta(t_1 - t_2] = kTg_b(t_1) \frac{I^o(t_1) + 2I_o}{I^o(t_1) + I_o} \delta(t_1 - t_2) \quad (32b)$$

or

$$eI^o(t)\delta(t_1 - t_2) \approx\!\!\approx kTg_b(t_1)\delta(t_1 - t_2), \quad (32c)$$

where in (32c) we have (again) ignored the diode leakage current I_0. Now then, given the barrier, $\langle q_d^2 \rangle$ will be largest if $g_b \ll g_m$. Inserting (32c) into (31) and integrating on t, it follows that

$$\langle q_d^2 \rangle|_{\text{shot}} < \tfrac{1}{2}kTC_p, \quad (33)$$

which is comparable with the thermal noise produced by I ($\frac{2}{3}kTC_p$). If one may assume that towards the end of the cycle $g_b(t) = dg_m(t)$,

then

$$\langle q_d^2 \rangle|_{\text{shot}} = \tfrac{1}{2}kTC_p(1 + d)^{-1}. \tag{34}$$

While (33) and (34) should be sufficient for estimating the contribution to q_d of shot noise, if more accuracy is desired (31) should be used.

6.2 Leakage-current noise

Although leakage currents are small and, hence, fluctuations in them even smaller, it is of interest to briefly analyze leakage-current noise. This is because (*i*) a portion of the contribution to q_d from leakage-current fluctuations is not suppressed and (*ii*) no longer does $q_d = -q_s$. We shall restrict ourselves to the role of leakage current which flows into the source, drain, and channel of an IGFET, bucket-brigade, single-step transfer CTD. Leakage current into the source and sink of a CCD can probably be treated in a similar manner.

If we assume that a leakage current per unit length of the channel $J_{ch}(x, t)$ enters the channel, following Jordan and Jordan,[22] we find that a fraction $(L - x)/L$ flows towards the source and x/L flows towards the drain. We have assumed, of course, that $\int J_{ch}(x, t)dx \ll I$. From the standpoint of noise this means that

$$\dot{q}_s = -q_s/\tau + \int_0^L dx j_{ch}(x, t)(L - x)/L + \int_s dx j_s(x, t), \tag{35a}$$

and

$$\dot{q}_d = \frac{q_s}{\tau} + \int_0^L dx \frac{j(x, t)x}{L} + \int_d dx j_d(x, t), \tag{35b}$$

where $J = J^o + j$. Solving (35a) for q_s and (35b) for q_d, we obtain

$$q_s(t) = \int_0^t dt' \exp\left[-\int_{t'}^t dt''/\tau(t'')\right] \cdot \left[\int_0^L dx \frac{L - x}{L} j_{ch}(x, t') + \int_s dx j_s(x, t')\right] \tag{35c}$$

and

$$q_d(t) = \int_0^t dt' \left[\frac{q_s(t')}{\tau(t')} + \int_0^L dx \frac{j(x, t')x}{L} + \int_d dx j(x, t')\right]. \tag{35d}$$

Comparing (35c) and (35d), we note that while the fluctuation j is suppressed in q_s and, therefore, in the first contribution to q_d, it is *not* suppressed in the second or third contributions to q_d. Thus, even though leakage currents may in themselves be small, since a portion of their noise is *not* suppressed, one may expect to see some contribution

to q_d from this source. As a rough estimate, if after many transfers the size of Q_d is increased by Q_e (leakage charge), one may expect $\langle q_d^2 \rangle$ due to leakage to be about qQ_e. The autocorrelation and cross-correlation functions of q_d and q_s can be calculated at once from (35b, c) if one notes that for full shot noise,

$$\langle j(x_1, t_1) j(x_2, t_2) \rangle = eJ(x_1, t_1)\delta(t_1 - t_2)\delta(x_1 - x_2). \tag{36}$$

Since an accurate calculation of the results again requires some detailed knowledge of the noiseless charge transfer and the role of leakage current, we shall not pursue this topic further.

VII. STORAGE-PROCESS AND TRANSFER-PROCESS NOISE

Up to this point in our discussion, we have been concerned solely with the various contributions to q_d that result from a single charge transfer of a single charge packet. A charge packet reaching the output of a CTD has, however, been transferred typically 10^2 to 10^3 times. The noise in the output packet is an accumulation not only of the noise acquired by the packet of interest during each transfer, but also the noise contained in incompletely transferred portions of preceding packets. In addition, there is correlation between the noise in successive packets at the output. Some of this correlation arises, of course, from the incompletely transferred portions picked up along the line.[7] However, even in the absence of incomplete charge transfer, there is substantial correlation from packet to packet (see Fig. 4). For example, for thermal and interface-state noise, we noted that throughout a single transfer cycle $q_d = -q$. While q_d accompanies the packet of interest, $q(= -q_d)$ is picked up by the next packet.[8] At the output the correlation will affect the spectral density of the total noise accompanying the signal. In this section, we consider this correlation and in the next section, we consider modulation noise, both in the absence of incomplete transfer; in Section IX, we discuss the output noise including incomplete transfer effects. (Finally we remind the reader that all along we have been concerned with *random* noise which is generated in addition to the signal distortion resulting from incomplete transfer. In many cases, physical processes which contribute to random noise also contribute to incomplete transfer. However, it should be kept in mind that while the noise is random and characterized by stochastic processes, the incomplete transfer is deterministic and characterized by a specific transfer function for the entire device.)

We noted in the case of contributions to q_d arising from thermal, trapping, clock-voltage and barrier-current fluctuations that conserva-

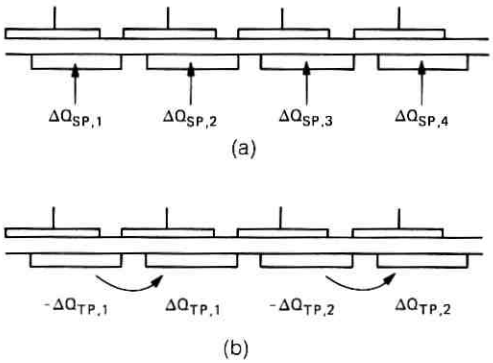


Fig. 4—Schematic illustrating the difference between the origin of storage-process and transfer-process noise.

tion of charge led to the relation $q_d + q = 0$, where q_d is the noise added to the transferred charge packet and q, which is picked up by the following packet, is the noise added to the untransferred charge. Using the notation of Ref. 8 in which $q_{s,\sigma}^{m,\mu}$ is the transfer-noise charge introduced during the σth phase of the sth transfer cycle as charge flows from the $(\mu - 1)$th to the μth phase of the mth CTD cell, the accumulated noise ΔQ_r^N in the charge at the last phase of the Nth cell at the end of the rth transfer cycle is given by

$$\Delta Q_r^N = \sum_{m=1}^{N} \sum_{\mu=1}^{p} (q_{r-(N-m),\mu}^{m,\mu} - q_{r-(N-m)-1,\mu}^{m,\mu}). \tag{37}$$

[In (37), p is the number of charge transfers per cell. If n is the total number of charge transfers in the CTD, then $n = Np$. During the $(r + 1)$th transfer cycle, $Q_r^N (= {}_oQ_r^N + \Delta Q_r^N)$ flows from the device to the output through a low-pass filter. If by $g(t)$ we designate the output current per total charge transferred for a single packet, then the output noise current is given by

$$i(t) = \sum_r \Delta Q_r^N g(t - rt_o), \tag{38}$$

where t_o is the clock period. One can calculate the spectral density of $i(t)$, obtaining

$$S_i(f)|_{\text{TP}} = 4n\langle q^2\rangle_{\text{TP}} f_o |g(f)|^2 (1 - \cos 2\pi f/f_o), \tag{39a}$$

where TP denotes transfer process and f_o is the clock frequency $(f_o = t_o^{-1})$. [Here we have assumed that $\langle q^2\rangle$ is independent of (s, σ, m, μ) or, equivalently, of the signal. This is quite reasonable

since the suppression effect renders the noise dependent only on the final portions of the charge transfer. And owing to the smallness of the coefficient of incomplete transfer, the dependence of the final portion of charge transfer on the signal is negligible for noise purposes.] For $f \ll f_o/2$, S_{iTP} is greatly suppressed below its mean value, $S_i(f_o/4)$ (see Fig. 5). This effect results from the fact that at low frequencies one is averaging over such long times that nearly all the q_d are cancelled by their corresponding $q = -q_d$. For $f \approx f_o/2$ (f is constrained by the relation $|f| < f_o/2$), S_{iTP} is in fact enhanced by the strong, mutual correlation between adjacent packets. The suppression of the transfer process noise at low frequencies is advantageous, since by increasing f_o one can increase the signal-to-noise ratio (S/N).

For the moment, we ignore the transfer-process noise associated with leakage and consider only storage-process (SP) noise, which we define as fluctuations that influence the size of each packet independently during each transfer phase. Since, under such circumstances, there is no correlation between the noise in different packets, the spectral density of the filtered current at the output is independent of frequency (white) and is given by

$$S_i(f)|_{SP} = 2n\langle q^2\rangle_{SP} f_o |g(f)|^2 \tag{39b}$$

[$|g(f)| \approx 1$ for $|f| < f_o/2$]. If $\langle q^2\rangle_{TP} = \langle q^2\rangle_{SP}$, the integral of $S_i(f)$ over $0 < f < f_o/2$ is twice as large for TP as for SP noise. This is because in the case of TP noise each fluctuation contributes to two charge packets, whereas in the SP case, only one packet is affected. We note that $S_i(f)$ for SP noise is *not* suppressed for $f \ll f_o/2$. Thus, although leakage noise is expected to be small, since neither is a portion of it suppressed in forming $\langle q^2\rangle_{SP}$ nor is $S_i(f)$ suppressed by packet-packet correlation, the role of leakage-current noise may in some cases be more important than is usually appreciated.

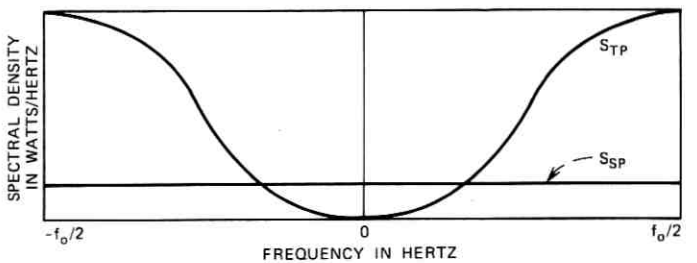


Fig. 5—Noise spectral densities S_{SP} and S_{TP} plotted versus frequency f.

In discussing TP and SP noise above (and in Ref. 8), we have assumed that different, or rather independent, mechanisms contribute to $\langle q^2 \rangle_{SP}$ and $\langle q^2 \rangle_{TP}$, and, for TP noise, that $q_d = -q$. For leakage-current noise, we found that in fact the same fluctuation could contribute to both SP- and TP-like noise, and that $q_d \neq -q = -q_s$. The question is raised: What is the spectral density of such noise?

To calculate the spectral density of the noise in cases where TP- and SP-like effects are correlated, we proceed as follows. Let the noise accumulated on the source in a single transfer phase be $q_s(s, \sigma, m, \mu)$ and the noise on the drain be $q_d(s, \sigma, m, \mu)$, where (s, σ, m, μ) is defined above. Then

$$\Delta Q_r^N = \sum_{m=1}^{N} \sum_{\mu=1}^{p} \{q_d[r - (n - m), \mu, m, \mu] + q_s[r - (N - m) - 1, \mu, m, \mu]\}, \quad (40)$$

where we continue to ignore incomplete transfer effects. It follows that the spectral density of $i(t)$ defined by (38) is given by

$$S_i(f) = 2nf_o |g(f)|^2 [\langle q_d^2 \rangle + \langle q_s^2 \rangle + 2\langle q_d q_s \rangle \cos(2\pi f/f_o)]. \quad (41a)$$

[For pure TP noise, $q_s = -q_d$, and we recover (39a); for pure SP noise, $q_s \ll q_d$ owing to suppression, and we recover (39b).] By rewriting (41a) slightly, it separates into SP and TP portions:

$$S_i(f) = 2nf_o |g(f)|^2 \langle (q_d + q_s)^2 \rangle + 4nf_o |g(f)|^2 \langle -q_s q_d \rangle [1 - \cos(2\pi f/f_o)]. \quad (41b)$$

Thus, we find that even when the sources of noise leading to TP and SP noise are correlated and, hence, more complicated than those treated in Ref. 8, we still obtain an unsuppressed, white-noise contribution and a suppressed contribution. If incomplete transfer effects are included, the spectral densities become much more complicated. We take up this matter again in Section IX.

VIII. MODULATION NOISE

In Section 5.3, we calculated the noise introduced into the signal during a single transfer resulting from fluctuations in the clock voltage. To calculate the effect that this modulation noise has on the output of the CTD, we must notice that each clock-voltage fluctuation is felt simultaneously by each transferring packet throughout the entire device. If the elements of the CTD are physically quite similar, each fluctuation will induce nearly the same noise contribution to the $(q_s, q_d = -q_s)$ pair generated in each CTD element during each cycle.

This means that modulation noise can be expected to be substantially suppressed, which is in fact what we find. In calculating modulation noise, we shall first assume that the transfer parameters of each pair of transfer elements of the CTD are identical, and then we shall include some variation in these parameters.

If by q we represent that portion of q_d introduced by clock-voltage fluctuations, since q is purely TP noise, it follows from (37) that

$$\Delta Q_r^N = \sum_{\mu=1}^{p} \left[-q_{r-N,\mu}^{1,\mu} + q_{r,\mu}^{N,\mu} + \sum_{n=1}^{N-1} (q_{r-(N-n),\mu}^{n,\mu} - q_{r-(N-n),\mu}^{n+1,\mu}) \right] \quad (42a)$$

$$= \sum_{\mu=1}^{p} (q_{r,\mu}^{N,\mu} - q_{r-N,\mu}^{1,\mu}). \quad (42b)$$

The simplification attained in (42b) results from the above observation concerning the similarity of contributions to the noise throughout the CTD, which stated quantitatively is that

$$q_{s,\mu}^{m,\mu} = q_{s,\mu}^{n,\mu} \quad (43)$$

for all cells m, n during a given cycle s. It follows from (42b) that the contribution to the total mean-square fluctuation in the output due to clock-voltage fluctuations is simply

$$\langle (\Delta Q_r^N)^2 \rangle |_{\text{clock}} = 2p \langle q_d^2 \rangle |_{\text{clock}}, \quad (44)$$

where $\langle q_d^2 \rangle$ is given by (27d) and p is the number of phases per cycle. The most interesting feature of (44) is that it is independent of n, the number of transfers each charge packet is subjected to in the CTD. Thus, this portion of modulation noise is not compounded and as such can be expected to be small.

If we introduce the possibility that the physical parameters characterizing each transfer stage of the CTD are slightly different, then (43) will not hold, and, as a consequence, the clock-voltage fluctuations will not essentially cancel. Let us assume that these physical parameters are distributed according to some probability distribution. Then, in place of q in (42a), let us write $(\bar{q} + q')$ where $\bar{q}$ is the charge fluctuation averaged over the distribution of the device parameters, and q' is the deviation from this average. Inserting $(\bar{q} + q')$ into (42a) for q, expressing ΔQ as $(\Delta \bar{Q} + \Delta Q')$, a straightforward calculation leads to

$$\langle \Delta \bar{Q}_r^N \Delta \bar{Q}_s^N \rangle = 2p \langle \bar{q}^2 \rangle [\delta_{r,s} - \tfrac{1}{2}(\delta_{r,s-N} + \delta_{r,s+N})] \quad (45a)$$

and

$$\langle \Delta Q_r'^N \Delta Q_s'^N \rangle = 2n \langle \overline{q'^2} \rangle [\delta_{r,s} - \tfrac{1}{2}(\delta_{r,s-1} + \delta_{r,s+1})], \quad (45b)$$

in which the *bar* denotes averaging over the device parameters. To calculate $\langle \overline{q'^2} \rangle = \langle \overline{(q - \bar{q})^2} \rangle = \langle \overline{q^2} \rangle - \langle \bar{q}^2 \rangle$ for clock-voltage fluctuations, it is perhaps simplest to carry out the average first over the clock-voltage fluctuations obtaining (27), and then average over the device parameters. If we assume that the primary variation from cell to cell is the size of the storage capacitance C_s, then (27d) yields

$$\langle (\Delta Q_r'^N)^2 \rangle = 2n \cdot \tfrac{1}{2} S_V(0) B (1 + b)^2 \langle (C_s - \bar{C}_s)^2 \rangle. \tag{45c}$$

This result expresses the effect on the size of the charge packet of the coupling between fluctuations in clock voltage and deviations in storage capacitance. Unlike $\langle (\Delta \bar{Q}_r^N)^2 \rangle$ given in (44), $\langle (\Delta Q_r'^N)^2 \rangle$ is proportional to n, expressing the fact that during each transfer event, noise is added to the signal. Finally, we note from (45b) that $\Delta Q_r'^N$ has the character of transfer process noises so that its spectral density is suppressed for $f \ll f_o/2$. The spectral density of $\Delta \bar{Q}_r^N$, which can be obtained from (45a), is weird. It oscillates once between 0 and $4p\langle \bar{q}^2 \rangle$ each time f changes by only f_o/N. It is probably best approximated as white of size $2p\langle \bar{q}^2 \rangle$.

IX. INFLUENCE OF INCOMPLETE CHARGE TRANSFER ON COLLECTING

In the two preceding sections we have ignored the influence of incomplete charge transfer on the noise accumulated in charge packets by the time they reach the output. In Section VII, we took $\alpha = 0$ to illustrate as simply as possible the suppression in the spectral density of transfer-process noise at low frequencies. In Section VIII, we took $\alpha = 0$ to treat the multicorrelated charge fluctuations induced by the clock voltage in a straightforward manner. We now include incomplete transfer and find that, even though it alters only a small fraction of the signal and, hence, of the noise on each transfer, typically 10^{-3} to 10^{-4}, its accumulated effects are extremely important in some cases.[7]

The effects of incomplete transfer on a charge packet are not simple, even when linearity is assumed. The key is to write down a general expression for the charge, seen as a function of time at the output, which arises from a given charge packet introduced at an earlier time on an arbitrary cell. Once this is done, one can combine the effects on the output of fluctuation-induced noise charge which is created on each stage during each phase of each cycle. The result is a complex, combinatorial expression which for $n\alpha \geqq 1$ is nontrivial to evaluate numerically. Despite these complications, the results obtained are worth the effort needed to obtain them.[7]

To calculate the effect of incomplete charge transfer on the collecting of noise, we first calculate the portion of the noise q_r^m added to the charge packet in the mth cell during the rth clock cycle which is observed at the output during the sth clock cycle. As defined, q_r^m is related to the $q_{r,\sigma}^{m,\mu}$ of Section VII according to

$$q_r^m = \sum_{\mu=1}^{p} q_{r,\mu}^{m,\mu}. \tag{46}$$

Thus, for purposes of simplicity, we shall ignore for the moment the fact that the transfer of charge within each cell involves p distinct (independent) transfers. The error involved in so doing is of the order of $n\alpha^2$, which for devices of interest is less than 0.01. Since the linear model we shall use to treat incomplete transfer is probably not this accurate, this approximation is justified. However, we must be careful to use (46) when we calculate terms such as $\langle q_s^m q_r^n \rangle$ so that we do not neglect correlation effects.

The dynamic equation that governs the transfer of the charge $q(r, m)$ on the mth cell during the rth cycle is

$$q(r+1, m+1) = q(r, m) - \epsilon q(r, m) + \epsilon q(r, m+1), \tag{47}$$

where ϵ is the coefficient of incomplete transfer per cell.[15] [One can relate ϵ to α, the coefficient of incomplete transfer per transfer,[16] through $\epsilon = p\alpha$, or more accurately, through $(1-\epsilon) = (1-\alpha)^p$.] Using (47), it is straightforward to calculate the charge $q_m(s, N+1)$ observed during the sth cycle at the output, the $(N+1)$th stage of an N-cell register, as a function of charges $q_o(r, m)$, which are added to the packet present in the mth cell during the rth cycle. The result is

$$q_m(s, N+1) = (1-\epsilon)^{N-m+1} \sum_{r=0}^{\infty} \binom{r+N-m}{r} \cdot \epsilon^r q_o[s-(r+N-m+1), m]. \tag{48}$$

[If $\epsilon = 0$, the limit of negligible incomplete transfer $q_m(s, N+1)$ becomes

$$q_m(s, N+1) = q_o[s-(N-m+1), m].$$

Thus, the additional output charge seen during the sth cycle is just the charge added to the mth cell, $N-m+1$ cycles earlier.] Using (48), we calculate the total noise charge ΔQ_s^N observed at the output of an N-cell CTD during the sth cycle by replacing $q_o(r, m)$ by the

noise charge q_r^m and summing overall cells m. The result is

$$\Delta Q_s^N = (1 - \epsilon) \sum_{m=1}^{N} (1 - \epsilon)^{N-m} \sum_{r=0}^{\infty} \binom{r + N - m}{r} \cdot \epsilon^r q_{s-(r+N-m+1)}^m, \quad (49)$$

where

$$q_r^m = q_d(r, m) + q_s(r - 1, m). \quad (50)$$

In (50), $q_d(r, m)$ is the noise added to the charge on drain during the cycle of interest, and $q_s(r - 1, m)$ is the noise added to the charge remaining on the source during the preceding cycle and picked up by the charge packet coming by on the next cycle. [The expression in (49), as well as our treatment of collecting, is valid whether or not the statistics of the individual q_d's and q_s's must be calculated using a nonlinear approach.] From eq. (49) we can calculate nearly all compounding effects of interest, a few of which we now consider in some detail.

For digital purposes, the most important quantity of interest is $\langle(\Delta Q_s^N)^2\rangle$, the mean-square fluctuation in the size of the output charge packet. This quantity can be calculated from (49) keeping in mind that the only nonzero, cross-correlation function that enters is $\langle q_d(r, m) q_s(r, m)\rangle$ for all r and m. The result is

$$\langle(\Delta Q^N)^2\rangle = \langle(q_d + q_s)^2\rangle H_{SP}(p, N) + 2\langle -q_s q_d\rangle H_{TP}(p, N), \quad (51a)$$

where q_d and q_s are noise added to the charge in the drain and source, respectively, during a single transfer cycle, the statistics of which we calculated in Section V and Section VI, H_{SP} is the collecting factor for storage-process noise $q_s = 0$, and H_{TP} is that for transfer-process noise $(-q_s = q_d)$. The analytical expressions for H_{SP} and H_{TP}, ignoring an unimportant factor of $(1 - \epsilon)^2$, are

$$H_{SP}(p, N) = p \sum_{m=1}^{N} (1 - \epsilon)^{2(N-m)} \sum_{r=0}^{\infty} \binom{r + N - m}{r}^2 \epsilon^{2r} \quad (51b)$$

and

$$H_{TP}(p, N) = p \sum_{m=1}^{N} (1 - \epsilon)^{2(N-m)} \cdot \sum_{r=0}^{\infty} \binom{r + N - m}{r}^2 \epsilon^{2r} \left(1 - \epsilon \frac{r + 1 + N - m}{r + 1}\right). \quad (51c)$$

These can be evaluated exactly (Appendix C) as well as approximately (Appendix D). [The latter is necessary because the former, although exact, is difficult to evaluate for the large N and small ϵ of greatest

interest. These calculations are nontrivial. For example, except when $N\epsilon \ll 1$, attempting to calculate (51b) or (51c) using Stirling's formula to approximate the factorials is doomed to failure.] One may note at the outset that for $\epsilon = 0$, $H_{SP} = H_{TP} = pN = n$ the number of transfers, as expected. This result is also approximately valid for $N\epsilon = n\alpha \ll 1$.

The approximate results obtained from evaluating the sums in (51b, c),

$$H_{SP}(p, N) = p\chi_N(\epsilon) = p(N + \tfrac{1}{2})t^{-1}e^{-b}[I_o(b) + I_1(b)] - \tfrac{1}{2} \quad (52a)$$

and

$$H_{TP}(p, N) = p\varphi_N(\epsilon)$$
$$= p\{(1 - \epsilon) + \epsilon^{-1}(x + 1)^{-1}t[\chi_{N-1}(\epsilon) + 1 - \chi_N(\epsilon)]\}, \quad (52b)$$

where

$$t = \frac{(1 - \epsilon)}{(1 + \epsilon)}, \quad (52c)$$

$$b = 2(N + \tfrac{1}{2})\epsilon(1 - \epsilon)^{-2}, \quad (52d)$$

and

$$x = \frac{(1 + \gamma)}{(1 - \gamma)}, \qquad \gamma = \epsilon^2, \quad (52e)$$

are quite interesting. (χ_N and φ_N are given in Appendix D.) For $N\epsilon \ll 1$, where we expect incomplete transfer to play a very minor role in the compounding of noise, indeed we find that H_{SP} and H_{TP} are nearly equal to $pN = n$, the total number of transfers experienced by each packet in the device. This is just what one expects: The cumulative, mean-square noise charge after n independent transfers is just n times the mean-square noise charge following a single transfer. [The factor of two in the TP term of (51a) arises because, as explained in Section VII, for each $+q$ noise contribution there is a $-q$ contribution. Thus for each transfer, two noise terms are produced.] As N increases (ϵ is fixed), however, incomplete transfer plays a more significant role, altering the noise in two important ways. First, of course, the noise is incompletely transferred along with the signal. For $N\epsilon \gg 1$, H_{SP} increases only as $n^{\frac{1}{2}}$, reflecting this attenuation of the noise. Second, and even more important, for TP noise, incomplete transfer enables each $+q$, $-q$ pair created during each transfer of each packet to mix and, hence, null out or suppress the total noise. Thus, for $N\epsilon \gg 1$, H_{TP} approaches constant value, $(2\alpha)^{-1}$, independent of N.

In other words, for $N\epsilon = n\alpha \gg 1$, the collecting saturates and further collecting is totally suppressed (see Fig 6).[7] Of course, the signal is greatly distorted by the incomplete transfer if $N\epsilon \gg 1$. Nonetheless, a CTD has maximum storage capacity[13] for $1 < N\epsilon < 4$, and by using dynamic detection or optimum linear filtering (transversal filtering), one can greatly suppress signal distortion from incomplete transfer.[13] Thus, calculating the compounding factors, H_{SP} and H_{TP}, for $N\epsilon$ other than $N\epsilon \ll 1$ is not an academic exercise.

For analog purposes it is necessary to calculate the autocorrelation function $\langle \Delta Q_s^N \Delta Q_i^N \rangle$ using (49) and (50), from which the current (or voltage) spectral density of the noise can be obtained as in Section VII. We shall pursue this no further than to point out that since $q_s = -q$ must be transferred one more time than q, the effect of incomplete transfer on each q, q_s pair will be slightly different, and this will reduce their mutual correlation at the output. Thus, total suppression at zero frequency is no longer expected.

Incomplete charge transfer will also affect fluctuations at the output caused by modulation noise. In Section VIII we found that modulation noise was so highly correlated that, in the absence of incomplete transfer, the largest portion of modulation noise was not compounded. Introducing incomplete transfer, however, will destroy

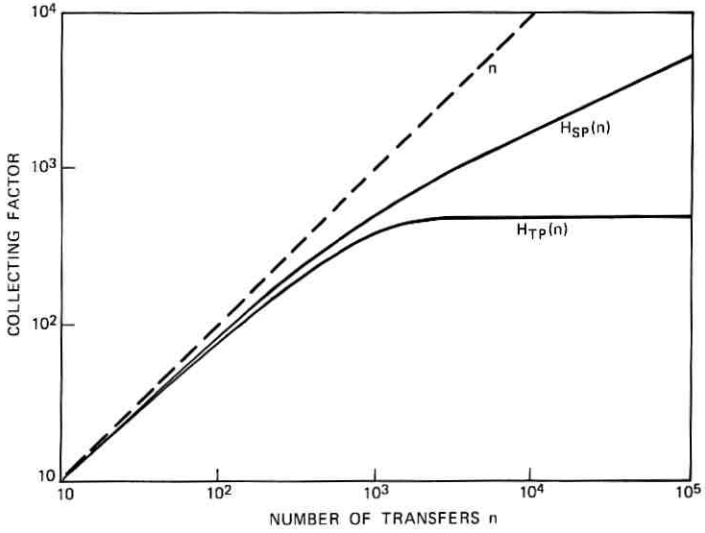


Fig. 6—Suppressed, collecting factors $H_{SP}(n)$ and $H_{TP}(n)$ plotted versus n, the number of transfers, for $\alpha = 10^{-3}$.

the exact correlation which led to this cancellation. Again, if desired, the effect of the collecting of incomplete charge transfer on the modulation noise for digital or analog applications can be calculated using (49) and (50).

Noise present in the input signal can clearly be treated simply as part of the signal with respect to incomplete-transfer effects. A somewhat different situation arises, however, when the signal enters the CTD all along its extent, as is the case in imaging applications. Here full (photon) shot noise results in a signal-dependent contribution to the mean-square, input noise of each packet equal to $eQ_{s,m}^{o}$, where $Q_{s,m}^{o}$ is the average size of the packet injected at site m during cycle s. However, in this case, the input packets undergo different numbers of transfers, $p(N - m)$, between the cell in which they are injected and the CTD output. Thus, the influence of incomplete transfer affects the signal and noise originating at each different cell differently. As with the other compounding effects, the collecting of this shot noise in the presence of incomplete transfer can be calculated if desired using (49) and (50).

X. CALCULATIONS

In the foregoing we have discussed a number of sources of noise in CTD's and their effect on the output signal in the absence and presence of incomplete charge transfer. In most cases, inserting the appropriate physical parameters for the noiseless operation of the device of interest suffices to calculate $\langle q^2 \rangle$, the mean-square fluctuation in size of a given charge packet acquired on a given transfer. Then, using (51), the influence of this fluctuation on the output signal can be determined. Such calculations are, in general, difficult, owing to the necessity of evaluating integrals such as those in (27b). However, realistic approximations can be made as indicated to obtain useful results.

It is of interest, however, to determine the minimum amount of noise expected to be present in CTD's assuming one can minimize clock-voltage fluctuations, surface states, incomplete transfer, etc., and operate each device so as to avoid emission-limited currents, etc. In this ideal situation, one is left only with thermal noise, or with thermal noise plus shot noise on the input signal, and incomplete transfer (intrinsic and modulation in the sense of Refs. 10 and 11). The signal-to-noise ratio (S/N) was calculated for two characteristic coefficients of incomplete transfer ($=10^{-3}$, 10^{-4}) and four characteristic capacitances (C = 1, 0.1, 0.01, 0.001 pF) as a function of the number

of transfers n from input to output:

(*i*) Using eq. (14a) for the thermal noise acquired per transfer.
(*ii*) Using eq. (51a) for the influence of this noise on the output.
(*iii*) Taking the thermal noise to be purely TP, including shot noise at the input when present.
(*iv*) Including incomplete transfer effects on both the signal and the noise.

The results are plotted in Figs. 7 and 8. The ratio Q_o/C designates the maximum signal level (10 volts), and one-half this amount (5 volts) is the minimum signal level. Since the square of the signal charge is proportional to C^2, while the mean-square of the noise charge is proportional to C for both thermal and input shot noise, the S/N decreases proportionately with smaller C (small CTD cells). As the number of charge transfers is increased, the contribution of device noise to the total noise soon dominates that of the input noise. (From S/N one can also calculate the maximum information storage capacity of the CTD as a function of n, α, and C.) In general, other noise sources are present which reduce S/N from the ideal results shown here. While valid in general for BB-mode transfers, in the case of

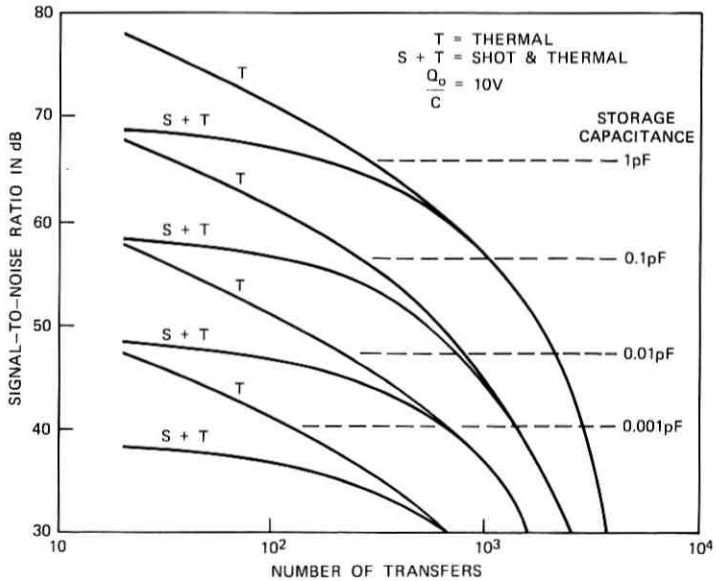


Fig. 7—Signal-to-noise ratio for CTD ($\alpha = 10^{-3}$) with storage capacitance of 1 pF, 0.1 pF, 0.01 pF, 0.001 pF.

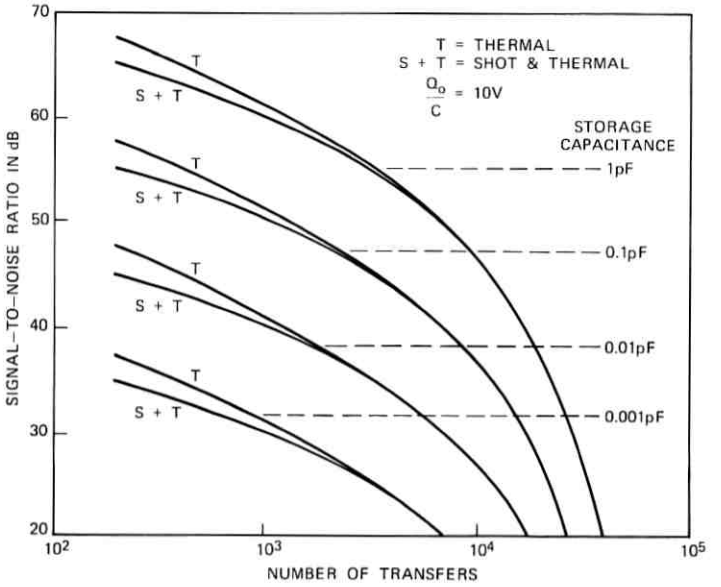


Fig. 8—Signal-to-noise ratio for CTD ($\alpha = 10^{-4}$) with storage capacitance of 1 pF, 0.1 pF, 0.01 pF, 0.001 pF.

CCD-mode transfers, if the rms thermal noise becomes comparable to the free charge incompletely transferred on a single transfer, then the noise predicted by our linear approach will overestimate the true noise.

XI. CONCLUSIONS

In the foregoing we have calculated the influence of several important sources of noise on the output of a CTD in a manner which includes all important relaxation, suppression, and correlation effects. In so doing, we encountered interesting effects, such as nonstationary noise, and we uncovered a number of unexpected results, such as the nearly total suppression of the spectral density of transfer-process noise at low frequencies and the total suppression of the collecting of transfer-process noise after many transfers of a charge packet. Because of the novelty of these and other effects, they were treated in considerable detail in a manner which did not presuppose considerable prior experience with formal treatments of device noise. In noting the results, the proportionate decrease of the S/N with decreasing storage capacitance was illustrated. This unavoidable feature will ultimately limit the practical size of CTD's.

Several of our results, despite the relatively complicated manner in which they were obtained, are remarkably simple, e.g., thermal noise, $\langle q^2 \rangle \propto kTC$, surface-state noise, $\langle q^2 \rangle \propto ekTN_{ss}A$, etc. One wonders whether there exists a general approach which circumvents the necessity of paying such careful attention to the detailed processes that accompany the transfer of charge through a CTD. We feel that there does not exist such an approach. Nonetheless, in Appendix E we outline a very rough method which ignores nearly all details of the charge transfer event. We find the intuitively appealing result that the mean-square noise acquired per transfer is to within a factor of the order of unity equal to full shot noise on the (total and not differential) incompletely transferred portion of the charge. Although providing a rough rule of thumb, since the approach is not totally reliable for calculating incomplete transfer, its accuracy for treating noise is not guaranteed. By contrast, the methods used in the bulk of this paper should be applicable in many types of integrated-circuit, dynamic devices of which CTD's are the first examples.

There are several interesting noise problems that we did not consider here. For example, in discussing modulation noise we indicated the possibility of nonuniformity in the physical parameters of each cell coupling to clock-voltage fluctuations to produce a collecting source of noise. Such cell nonuniformities also, of course, will result in a distribution of α's, the coefficients of incomplete charge transfer. This will, in turn, result in an additional effect on the nature of the compounding of the noise acquired in each transfer. The results are expected to be no less surprising than the effect such a distribution of α's has on the signal. A distribution of α's about their mean α_o leads to *less* signal distortion than if all the α's were α_o.[15,26] (This is actually not too significant for application purposes, since usually all deviations from the desired $\alpha = \alpha_d$ will be to larger α, thereby increasing α_o and enhancing distortion.)

A second problem worthy of attention is how to treat the noise in cases where the clock-voltage waveform does not turn off the flow of transferring charge abruptly.[9] In such cases, the nonlinear terms in the noise fluctuations are not small relative to the linear terms. The noise problem is then nonlinear and much more complicated.

A third problem, straightforward but tedious, is to calculate the S/N's, error rates, and device storage capacities for devices including regenerators, optimum linear filtering, and/or dynamic detection.[13,14] Considering the many new features that have arisen in the present

study, as well as the many types of noise that enter in such different ways, we feel such problems will not prove unrewarding.

XII. ACKNOWLEDGMENTS

It is a pleasure to thank G. E. Smith for initially interesting me in this problem and C. N. Berglund, H. J. Boll, J. R. Brews, J. E. Iwersen, D. Kahng, R. H. Krambeck, W. Shockley, G. E. Smith, R. J. Strain, N. S. Thornber, and M. F. Tompsett for many helpful discussions and several crucial comments.

APPENDIX A

Noise Introduced During a Single, Two-Step, Transfer Cycle

The purpose of this appendix is to calculate the noise fluctuations induced in the transferred charge during a single, two-step, transfer cycle. The procedure followed is the same as that in Section IV of the text for a single-step process. Here we shall merely outline the essential steps leading to the desired result.

We begin by linearizing the dynamic equations (4a, b, c) to obtain

$$\dot{q}_1 = -\left(\frac{\partial I_1}{\partial V_s} v_s + \frac{\partial I_1}{\partial V_b} v_b + \frac{\partial I_1}{\partial V_c} v_c + \frac{\partial I_1}{\partial V_{ss1}} v_{ss1}\right) - i_{n1}, \quad (53a)$$

$$\dot{q}_2 = -\dot{q}_d - \dot{q}_1, \quad (53b)$$

$$\dot{q}_d = \left(\frac{\partial I_2}{\partial V_b} v_b + \frac{\partial I_2}{\partial V_d} v_d + \frac{\partial I_2}{\partial V_c} v_c + \frac{\partial I_2}{\partial V_{ss2}} v_{ss2}\right) + i_{n2}. \quad (53c)$$

The constraint equations (3a–e), when linearized and when terms of the order of the incomplete transfer are dropped, yield the following relations among the fluctuating quantities:

$$q_s = C_s v_s - C_s^o v_{so}, \quad (54a)$$

$$q_b = C_b v_b - C_b^o v_{bo}, \quad (54b)$$

$$q_d = C_d v_d - C_d^o v_{do}, \quad (54c)$$

$$q_{ss1} = C_{ss1} v_{ss1} - C_{ss1}^o v_{ss1}^o, \quad (54d)$$

$$q_{ss2} = C_{ss2} v_{ss2} - C_{ss2}^o v_{ss2}^o. \quad (54e)$$

In addition, by the same reasoning that led to eqs. (7a, b) in the text,

we find that

$$v_{so} = \frac{\partial V_{so}}{\partial V_c} v_c = \frac{\partial V_s}{\partial V_c} \frac{C_s}{C_s^o} v_c = (+1) \frac{C_s}{C_s^o} v_c, \tag{55a}$$

$$v_{bo} = \frac{\partial V_{bo}}{\partial V_c} v_c' = \frac{\partial V_b}{\partial V_c} \frac{C_b}{C_b^o} v_c' = a \frac{C_b}{C_b^o} v_c', \tag{55b}$$

$$v_{do} = \frac{\partial V_{do}}{\partial V_c} v_c = \frac{\partial V_d}{\partial V_c} \frac{C_d}{C_d^o} v_c = (-1) \frac{C_d}{C_d^o} v_c, \tag{55c}$$

where $a = \partial V_b / \partial V_c$. [For a two-step process, transfer cells whose second step is characterized by a CCD-type of charge transfer, $a = -1$ and $v_c' = v_c$. For such cells whose second step is BB-like, the value of a and v_c' will depend on the details of the device. For example, for the C4D, $a = 0$; for the tetrode BB, $a = 1$ and $v_c' = v_{bl}$ is the voltage on the tetrode bias line. Thus, for purposes of generality we must use the form given in (55b).] Also v_{ss1}^o and v_{ss2}^o can be set equal to zero, as discussed in the text.

The next step is to solve eqs. (54a–e) for v_s, v_b, etc. and insert the results into (53a–c). The result of this substitution is

$$\dot{q}_1 = -\frac{g_{m1}}{C_s} q_1 - \left(g_{m1} + \frac{\partial I_1}{\partial V_c} \right) v_c + \left(g_{m1} \frac{C_{ss1}}{C_s} - \frac{\partial I_1}{\partial V_{ss1}} \right) v_{ss1} - i_{n1}, \tag{56a}$$

$$q_2 = -q_d - q_1, \tag{56b}$$

$$\dot{q}_d = \frac{g_{m2}}{C_b} q_2 + g_{m2} a v_c' + \frac{\partial I_2}{\partial v_c} v_c + \left(g_{m2} \frac{C_{ss2}}{C_b} - \frac{\partial I_2}{\partial V_{ss2}} \right) v_{ss2} + i_{n2}, \tag{56c}$$

in which we have assumed that the forward conductances

$$g_{m1} \equiv \partial I_1 / \partial V_s \quad \text{and} \quad g_{m2} \equiv \partial I_2 / \partial V_b$$

greatly exceed the reverse conductances

$$g_{r1} \equiv -\partial I_1 / \partial V_b \quad \text{and} \quad g_{r2} \equiv -\partial I_2 / \partial V_d.$$

It is now straightforward to solve (56a) for $q_1(t)$, insert into (56b) to obtain $q_2(t)$, and then solve (56c) for $q_d(t)$. The result is

$$q_d(t) = \int_0^t dt' \exp\left[-\int_{t'}^t dt''/\tau_2(t'') \right] \left[\frac{g_{m2}}{C_b} (-q_1) + \left(g_{m2} a + \frac{\partial I_2}{\partial V_c} \right) v_c + g_{m2} a (v_c' - v_c) - \left(g_{m2} \frac{C_{ss2}}{C_b} - \frac{\partial I_2}{\partial V_{ss2}} \right) v_{ss2} + i_{n2} \right]_{t'}, \tag{57a}$$

where

$$-q_1(t') = \int_0^{t'} dt'' \exp\left\{-\int_{t''}^{t'} dt'''/[\tau_1(t''')]\right\}\left[\left(g_{m1} + \frac{\partial I_1}{\partial V_c}\right) v_c - \left(g_{m1}\frac{C_{ss1}}{C_s} - \frac{\partial I_1}{\partial V_{ss1}}\right) v_{ss1} + i_{n1}\right]\Big|_{t''}, \quad (57b)$$

$$1/\tau_1 = g_{m1}/C_s \quad \text{and} \quad 1/\tau_2 = g_{m2}/C_b. \quad (57c)$$

If we define the two suppression factors according to

$$S_1(t, t') = \exp\left(-\int_{t'}^{t} dt''/\tau_1(t'')\right) \quad (58a)$$

and

$$S_2(t, t') = \exp\left(-\int_{t'}^{t} dt''/\tau_2(t'')\right), \quad (58b)$$

let

$$n(t, t') = \int_{t'}^{t} dt'' \exp\left(-\int_{t''}^{t} [1/\tau_2(t''') - 1/\tau_1(t''')]dt'''\right)\Big/\tau_2(t''), \quad (58c)$$

insert (57b) into (57a), and regroup terms, we obtain

$$\begin{aligned} q_d(t) = \int_0^t dt' \Big\{ & S_1(t, t')n(t, t')i_{n1}(t') + S_2(t, t')i_{n2}(t') \\ & - S_1(t, t')n(t, t')\left(\frac{C_{ss1}}{\tau_1} - \frac{\partial I_1}{\partial V_{ss1}}\right)\Big|_{t'} v_{ss1}(t') \\ & - S_2(t, t')\left(\frac{C_{ss2}}{\tau_2} - \frac{\partial I_2}{\partial V_{ss2}}\right)\Big|_{t'} v_{ss2}(t') \\ & + \left[S_1(t, t')n(t, t')\left(g_{m1} + \frac{\partial I_1}{\partial V_c}\right)\Big|_{t'}\right. \\ & \left. + S_2(t, t')\left(g_{m2}a + \frac{\partial I_2}{\partial V_c}\right)\Big|_{t'}\right] v_c(t') \\ & + S_2(t, t')[g_{m2}a(v_c' - v_c)]|_{t'}\Big\}. \end{aligned} \quad (59)$$

Although it is most convenient to keep the $S_1 n$ product as two terms, their combination is of interest:

$$S_1(t, t')n(t, t') = \int_{t'}^{t} dt'' \exp\left(-\int_{t'}^{t''} dt'''/\tau_1(t''') - \int_{t''}^{t} dt'''/\tau_2(t''')\right)\Big/\tau_2(t''). \quad (60)$$

It is clear that to calculate $\langle q_d^2 \rangle$, it is necessary to know $\langle i_{n1}(t_1) i_{n1}(t_2) \rangle$, $\langle i_{n2}(t_1) i_{n2}(t_2) \rangle$, $\langle v_{ss1}(t_1) v_{ss1}(t_2) \rangle$, $\langle v_{ss2}(t_1) v_{ss2}(t_2) \rangle$, $\langle v_c(t_1) v_c(t_2) \rangle$, $\langle v_c'(t_1) v_c'(t_2) \rangle$, and $\langle v_c'(t_1) v_c(t_2) \rangle$. These are discussed in the text.

It is of interest to calculate the contribution to $\langle q_d^2 \rangle$ that results from thermal noise. From (59) and (13b) ($H_n = \frac{2}{3}$), we obtain

$$\langle q_d^2 \rangle |_{kT} = 2kT \frac{2}{3} \int_0^{t_f} dt [S_1^2(t_f, t) n^2(t_f, t) g_{m1}(t) + S_2^2(t_f, t) g_{m2}(t)]. \quad (61)$$

Now ordinarily a two-step device is constructed so that

$$g_{m1}/C_a = \tau_1^{-1} < \tau_2^{-1} = g_{m2}/C_b.$$

Let us assume that the decay current is such that $\tau_1^{-1}(t) b = \tau_2^{-1}(t)$, where b can be taken independent of time ($b > 1$). Then we can perform the integrals in (61), obtaining

$$\langle q_d^2 \rangle |_{kT} = \tfrac{2}{3} kTC_b + \tfrac{2}{3} kTC_a (1 + b^{-1})^{-1} \quad (62)$$

for the thermal contribution to the noise acquired during a single, two-step transfer. The first term is expected from our result (14a) for an individual, single-step transfer. However, the second term includes a suppression factor whose presence one certainly would not have expected *a priori* using arguments assuming stationarity.

APPENDIX B

Outline and Modification of the Impedance-Field Method

As originally presented,[23] the impedance-field method (IFM) consists of dividing the problem of calculating device noise into two simpler problems. First, it was recognized that a given fluctuation in the velocity of a charge carrier at a given location induces a calculable fluctuation in the open-circuit voltage. Second, it was shown how, from a knowledge of the spectral density of the velocity fluctuations of the individual charge carriers, the spectral density of the open-circuit voltage fluctuations could be calculated. The first problem requires only an understanding of the operation of the device of interest under noiseless conditions, while the second can be deduced from the microscopic behavior of charge carriers in a small region of the device.[18] Thus, as long as the microscopic noise is simple, which it nearly always is, the device-noise problem is reduced to integrating the microscopic noise sources over the device, weighted by the influence of a unit fluctuation in each volume element on the output noise.

Let us now outline the IFM in more detail.[23] If $\mathbf{u}_t$ is the time-dependent fluctuation of carrier velocity $\mathbf{v}_t$ from its mean $\mathbf{v}_{ot}$,

$$\mathbf{u}_t \equiv \mathbf{v}_t - \mathbf{v}_{ot},$$

then the dipole current $\delta\dot{\mathbf{P}}_\alpha$ in a small volume element ΔV_α produced by the carriers in this volume is given by

$$\delta\dot{\mathbf{P}}_\alpha = q \sum_j \mathbf{u}_{jt}, \tag{63}$$

where the j-sum is over all carriers in ΔV_α. To calculate the effect of $\delta\dot{\mathbf{P}}_\alpha$ on the open-circuit voltage of the device, we note that if δI_α is injected into the device at $\mathbf{r}_\alpha$ (and if the device is grounded at $\mathbf{r} = \mathbf{0}$), the voltage induced at the contact labeled N, δV_N, is given by

$$\delta V_N = Z_{N\alpha}\delta I_\alpha. \tag{64}$$

If $Z_{N\alpha}$ is evaluated for all α, one maps out the "impedance field." If now one injects δI_α at $\mathbf{r}_\alpha + \delta\mathbf{r}$ and removes δI_α at $\mathbf{r}_\alpha$, linear superposition of small signals implies that

$$\delta V_N = [Z_{N\alpha}(\mathbf{r}_\alpha + \delta\mathbf{r}) - Z_{N\alpha}(\mathbf{r}_\alpha)]\delta I_\alpha \tag{65a}$$

$$= (\nabla Z_{N\mathbf{r}}).\delta\mathbf{r}\delta I_\alpha = \nabla Z_{N\mathbf{r}}.\delta\dot{\mathbf{P}}_\alpha. \tag{65b}$$

The last equality follows because the dipole current $\delta\dot{\mathbf{P}}_\alpha$ equals $\delta I_\alpha\delta\mathbf{r}$ if δI_α is chosen appropriately. Since we can relate $\delta\dot{\mathbf{P}}_\alpha$ to δV_N, from a knowledge of the statistics of the former, we can calculate those of the latter. In particular, from (63) and (65b) we can immediately write down the spectral density $S_{V_N}(f)$ of δV_N in terms of that of $\mathbf{u}_\alpha$:

$$S_{V_N}(f) = \sum_{\alpha,i,j} (\nabla_i Z_{N\alpha})(\nabla_j Z_{N\alpha})^* q^2 n_\alpha S_{\alpha u_i u_j}(f), \tag{66}$$

where i, j each run over x, y, z, and n_α is the number of carriers in ΔV_α. Since in the text we work in the time domain, we work directly with (65).

In calculating CTD noise, it is most convenient to work with short-circuit current fluctuations rather than with open-circuit voltage fluctuations. One may redo the above, calculating the short-circuit current fluctuation δI_N induced by δI_α. The result is that

$$\delta I_N = B_{N\alpha}\delta I_\alpha, \tag{67a}$$

where

$$B_{N\alpha} = Z_{N\alpha}/Z_N, \tag{67b}$$

and where Z_N is the impedance of the device between contact N and

ground. In terms of dipole currents, one obtains

$$\delta I_N = \nabla B_{Nr} \cdot \delta \mathbf{r} \delta I_\alpha = \nabla B_{Nr} \cdot \delta \dot{\mathbf{P}}_\alpha. \tag{68}$$

For an IGFET,

$$\nabla B_{Nr} = L^{-1} \hat{l}, \tag{69}$$

where L is the channel length and $\hat{l}$ is a unit vector along the channel.[22] In complete analogy to (66) one finds that

$$S_{I_N}(f) = \sum_{\alpha,i,j} (\nabla_i B_{N\alpha})(\nabla_j B_{N\alpha})^* q^2 n_\alpha S_{\alpha u_i u_j}(f). \tag{70}$$

In arriving at (66) and (70), we have taken into account the independence between spontaneous fluctuations which occur in separate regions ΔV_α. Thus, the expression for $\langle u_i(t_1) u_j(t_2) \rangle$ for the IFM, which corresponds to (12) in the text, is that for thermal noise

$$\langle u_i(t_1) u_j(t_2) \rangle = \frac{2kT}{q} \mu \delta(t_1 - t_2) \delta_{ij}, \tag{71}$$

which lacks the spatial delta function. The equivalence between the Langevin method used in the text and the IFM outlined there is discussed in some detail in Refs. 27 and 28.

APPENDIX C

*Exact Calculation of H_{SP} and H_{TP}**

In this appendix, we shall evaluate the following sums exactly:

$$\chi_N(\epsilon) = \sum_{m=1}^{N} \sum_{r=0}^{\infty} \left[(1-\epsilon)^{N-m} \epsilon^r \binom{r+N-m}{r} \right]^2$$

$$\varphi_N(\epsilon) = \sum_{m=1}^{N} \sum_{r=0}^{\infty} \left[(1-\epsilon)^{N-m} \epsilon^r \binom{r+N-m}{r} \right]^2 \cdot \left(1 - \epsilon \frac{r+1+N-m}{r+1} \right). \tag{72}$$

As the first step, set $n = N - m$, $\gamma = \epsilon^2$, and perform the sum on r. This yields

$$\chi_N(\epsilon) = \sum_{n=0}^{N-1} \frac{(1-\epsilon)^{2n}}{n!} \frac{d^n}{d\gamma^n} \left(\frac{\gamma^n}{(1-\gamma)^{n+1}} \right) \tag{73a}$$

$$\varphi_N(\epsilon) = \frac{1-\epsilon}{1-\gamma} + \sum_{n=1}^{N-1} \frac{(1-\epsilon)^{2n}}{n!} \frac{d^n}{d\gamma^n} \left(\frac{\gamma^n}{(1-\gamma)^{n+1}} \right) - \epsilon \frac{d^n}{d\gamma^n} \left(\frac{\gamma^{n-1}}{(1-\gamma)^{n+1}} \right). \tag{73b}$$

* Derivation due to N. S. Thornber.

The second step involves taking the n derivatives with respect to γ using the relation[29]

$$\frac{d^n}{dy^n}(fg) = \sum_{s=0}^{n} \binom{n}{s} \frac{d^s f}{dy^s} \frac{d^{n-s} g}{dy^{n-s}}$$

and recognizing that the resulting sum on s is to within certain simple factors equal to a Legendre polynomial,[29] $P_n(x)$, where $x = (1 + \gamma)/(1 - \gamma)$. This yields

$$\chi_N(\epsilon) = (1 - \gamma)^{-1} \sum_{n=0}^{N-1} t^n P_n(x) \tag{74a}$$

$$\varphi_N(\epsilon) = (1 - \gamma)^{-1} \left[(1 - \epsilon) + \sum_{n=1}^{N-1} t^n P_n(x) - \epsilon t^n \frac{x - 1}{n\gamma} P_n'(x) \right], \tag{74b}$$

where $t = (1 - \epsilon)/(1 + \epsilon)$. In (74b), if we note that[30]

$$P_n'(x)/n = [xP_n(x) - P_{n-1}(x)]/(x^2 - 1),$$

then all that remains is to evaluate $\sum_n t^n P_n(x)$.

Before proceeding, we should call attention to a potential source of trouble. Since $0 < \gamma = \epsilon^2 < 1$, it follows that $x > 1$ and $t < 1$. While Legendre polynomials for $|x| > 1$ are well-defined, their properties are not nearly so simple as they are for x in the usual region of interest, $|x| \leqq 1$. Thus, for fixed ϵ (and γ), if n becomes large, evaluating $P_n(x)$ for x only slightly larger than 1 is quite tricky. This provides motivation for the approach adopted in Appendix D.

Returning to the remaining sum in (74), we take the generating function[31] for Legendre polynomials, valid for $|t| < 1$ and $|y| < 1$,

$$\sum_{S=0}^{\infty} t^s P_s(y) = (1 - 2yt + t^2)^{-\frac{1}{2}},$$

multiply both sides by $P_m(y)$, and integrate on y from -1 to $+1$. Using the orthogonality of these polynomials, we obtain

$$t^m = \frac{2m + 1}{2} \int_{-1}^{1} dy P_m(y)(1 - 2yt + t^2)^{-\frac{1}{2}}. \tag{75}$$

Using the summation formula[32] for Legendre's polynomials, it follows that

$$\sum_{n=0}^{N-1} t^n P_n(x) = \int_{-1}^{1} dy (1 - 2yt + t^2)^{-\frac{1}{2}} \frac{N/2}{x - y} \cdot [P_N(x)P_{N-1}(y) - P_N(y)P_{N-1}(x)]. \tag{76}$$

To perform the final integral over y we make use of (75), noting that

$2tx = 1 + t^2$, and hence that

$$(1 - 2yt + t^2)^{-\frac{1}{2}}(x - y) = 2^{\frac{1}{2}}(x - y)^{\frac{1}{2}}t^{\frac{1}{2}}.$$

Thus, since

$$\int_{-1}^{1} dy(x - y)^{-\frac{1}{2}}P_n(y) = -2\frac{d}{dx}\int_{-1}^{1} dy(x - y)^{-\frac{1}{2}}P_n(y)$$

and since from (75)

$$2t^m(2m + 1)^{-1} = \int_{-1}^{1} dy[2t(x - y)]^{-\frac{1}{2}}P_m(y),$$

it follows after some algebra that

$$\sum_{n=0}^{N-1} t^nP_n(x) = \frac{2N}{1 - t^2}[P_N(x)t^N - P_{N-1}(x)t^{N+1}]. \tag{77}$$

Using (77) for the sums present in (55a, b) yields exact expressions for χ_N and φ_N in terms of two or three Legendre polynomials, respectively. The difficulties encountered in evaluating these expressions for $N \gg 1$ and $x \gtrsim 1$ made it clear that another form of the result was needed, one in which N and ϵ enter on an equal footing, preferably as a product. Such a result is derived in Appendix D.

APPENDIX D

Approximate Calculation of H_{SP} and H_{TP}

In this appendix, we evaluate the sum

$$\psi_{N-1}(\epsilon) = \sum_{n=0}^{N-1} t^nP_n(x), \tag{78}$$

where, as in Appendix C, $t = (1 - \epsilon)/(1 + \epsilon)$, $x = (1 + \gamma)/(1 - \gamma)$, and $\gamma = \epsilon^2$. The form we obtain will be a good approximation for $N \gg 1$ and will be very easy to evaluate numerically.

If we define $\psi_{-1} = 0$, then we can write

$$\psi_n - \psi_{n-1} = t^nP_n(x) \tag{79}$$

and evaluate the z-transform ψ_z of ψ_n defined by

$$\psi_z = \sum_{n=0}^{\infty} \psi_n z^{-n}. \tag{80}$$

Thus, from (79) we obtain

$$\psi_z - \frac{1}{z}\psi_z = \sum_{n=0}^{\infty}\left(\frac{t}{z}\right)^nP_n(x) = \left[1 - 2x\frac{t}{z} + \left(\frac{t}{z}\right)^2\right]^{-\frac{1}{2}}. \tag{81}$$

Solving for ψ_z, we obtain, using $2xt = 1 + t^2$,

$$\psi_z = (1 - z^{-1})^{-\frac{1}{2}}(1 - t^2 z^{-1})^{-\frac{1}{2}}. \tag{82}$$

Thus,

$$\psi_n = \frac{1}{2\pi i} \oint \psi_z z^{n-1} dz, \tag{83a}$$

where the closed contour in the complex plane includes $z = 0$. Letting $z = e^s$, (83a) becomes ($\sigma = 0^+$)

$$\psi_n = \frac{1}{2\pi i} \int_{\sigma - i\pi}^{\sigma + i\pi} ds\, e^{ns}(1 - e^{-s})^{-\frac{1}{2}}(1 - t^2 e^{-s})^{-\frac{1}{2}} \tag{83b}$$

an exact expression for $\psi_N(\epsilon)$ if $n = N$.

An exact integration of (83b) would recover (77). However, a very useful approximation for large n can be obtained at once if we note that under such circumstances most of the contribution to the integral comes in the vicinity of $s = 0$. Thus, expanding $\exp(-s) \approx 1 - s$ and then, taking the limits of integration to be from $(\sigma - i\infty)$ to $(\sigma + i\infty)$, yields

$$\psi_n \approx \frac{1}{2\pi i} \int_{\sigma - i\infty}^{\sigma + i\infty} ds\, e^{ns} s^{-\frac{1}{2}} t^{-1}[s + (t^{-2} - 1)]^{-\frac{1}{2}}. \tag{84a}$$

This inverse Laplace transform can be found in the tables.[33] The result is

$$\psi_n \approx t^{-1} n e^{-b}[I_o(b) + I_1(b)], \tag{84b}$$

where

$$b = n(t^{-2} - 1)/2 = 2n\epsilon(1 - \epsilon)^{-2}. \tag{84c}$$

We can match this result, valid for $n \gg 1$, to the $n \ll 1$ limit to obtain an approximate result good for all n. At this point, however, it must be stressed that the major difficulty in approximating ψ_n for large n arises from the necessity to cancel exactly any exponential dependence of ψ_n on n. Clearly, physically compounding can increase with n no more rapidly than linearly. This delicate cancellation is in fact achieved in (84b) since $I_0(b)$ and $I_1(b)$ go as $\exp(b)/(2\pi b)^{\frac{1}{2}}$ for $b \gg 1$. There are other ways of arriving at (84), but this is one of the simplest.

To determine the form of ψ_n for n sufficiently small ($n\epsilon \ll 1$), that is, such that an expansion of ψ_n in powers of $n\epsilon$ rapidly converges in a few terms we note from (71), (73a), and (78) that we have the relation

$$\psi_{N-1} = (1 - \gamma) \sum_{n=0}^{N-1} \sum_{r=0}^{\infty} \left[(1 - \epsilon)^n \epsilon^r \binom{r + n}{r} \right]^2 \tag{85a}$$

$$= N[1 - \epsilon(N - 1) + 0(N^2\epsilon^2)]. \tag{85b}$$

On the other hand, expanding (84b) to similar order yields

$$\psi_n \approx n(1 + 2\epsilon - \epsilon n). \tag{86}$$

Since (84b) and consequently (86) can be expected to be valid only for $n\epsilon \gg 1$, we are not required to set $n = N - 1$ in (84b). [Nonetheless, it is encouraging to note that even for $n\epsilon \ll 1$, if we do set $n = N - 1$ in (86), then (85b) and (86) agree as to the coefficients of N and $N^2\epsilon$, the two largest terms.] Even further improvement can be obtained if we replace ψ_n by $\psi_n + \delta_1$ and n by $N - 1 + \delta_2$, where δ_1 and δ_2 are of the order of unity. This alteration clearly does not alter significantly our result valid for large n ($n\epsilon \gg 1$), and it improves the result for $n\epsilon \ll 1$. After a little algebra, one obtains $\delta_1 = -\frac{1}{2}$ and $\delta_2 = \frac{3}{2}$, which ensure that terms of order unity and $N\epsilon$ match. We conclude that $\chi_N(\epsilon)$ is given quite well by

$$\chi_N(\epsilon) = (N + \tfrac{1}{2})t^{-1}e^{-b}[I_0(b) + I_1(b)] - \tfrac{1}{2}, \tag{87a}$$

where

$$b = 2(N + \tfrac{1}{2})\epsilon(1 - \epsilon)^2 \quad \text{and} \quad t = (1 - \epsilon)/(1 + \epsilon), \tag{87b}$$

in which we have ignored the $(1 - \gamma)^{-1}$ factor in (73a). (Recall that $\gamma = \epsilon^2 \approx 10^{-6}$.) The expression (87a) enables us to calculate the compounding factor for storage-process noise for arbitrary N and ϵ with relative ease.

The compounding factor for transfer-process noise is obtained from $\varphi_N(\epsilon)$, which can be expressed in terms of $\chi_N(\epsilon)$ in the form

$$\varphi_N(\epsilon) = (1 - \epsilon) + [\chi_N(\epsilon) - 1] - \epsilon\gamma^{-1}(x + 1)^{-1}\{x[\chi_N(\epsilon) - 1] - t\chi_{N-1}(\epsilon)\} \tag{88a}$$

$$= (1 - \epsilon) + \epsilon^{-1}(x + 1)^{-1}t[\chi_{N-1}(\epsilon) + 1 - \chi_N(\epsilon)], \tag{88b}$$

where we have used (73) and ignored the prefactor of $(1 - \gamma)^{-1}$. Since $\gamma = \epsilon^2$, the prefactor of the last term of (88) is approximately $(2\epsilon)^{-1}$. Nonetheless, this term is well-behaved even where our approximate result (87) is used in (88) to obtain an approximate $\varphi_N(\epsilon)$.

It is of interest to calculate χ_N and φ_N in the limit of $N\epsilon \gg 1$. The former limit is just N in both cases. In the latter limit

$$\chi_N(\epsilon) \to (N/\pi\epsilon)^{\frac{1}{2}} = p^{-1}(n/\pi\alpha)^{\frac{1}{2}}, \quad N\epsilon \gg 1 \tag{89a}$$

and

$$\varphi_N(\epsilon) \to (2\epsilon)^{-1} = (2p\alpha)^{-1}, \quad N\epsilon \gg 1. \tag{89b}$$

The second limit is most easily obtained by noting that $x(x + 1)^{-1}$

$= (1 + \gamma)/2$, $(x + 1)^{-1} = (1 - \gamma)/2$, $\epsilon\gamma^{-1}[(1 + \gamma) - t(1 - \gamma)]/2 = 1$, and $(x_{N-1} + 1 - x_N) = 1 + 0(N^{-\frac{1}{2}})$, $N\epsilon \gg 1$.

APPENDIX E

Another Approach to CTD Noise

Throughout our treatment we have paid careful attention to the details of the transfer and storage of charge in CTD's in order to calculate the noise induced by the device. We have noted, however, that by making reasonable approximations, we can often reduce rather complicated, exact results to much simpler, approximate results adequate for most purposes. One cannot help but wonder, therefore, if there does not exist an approximate but adequate short cut for calculating CTD noise. One clue is to compare Tompsett's results[17] for incompletely transferred charge Q_i with his results for mean-square noise charge induced by interface states $\langle q^2 \rangle$. If we ignore factors of the order unity we find that

$$\langle q^2 \rangle = eQ_i, \tag{90}$$

where

$$Q_i = ekTN_{ss}. \tag{91}$$

In other words, the noise induced by interface states is just the shot noise associated with the incompletely transferred portion of the interface charge.[34] While one will never recover thermal noise from such arguments, one expects not to do too badly in estimating the influence on the output of those noise sources, such as interface states or emission-limited (barrier-limited) currents, which are closely tied to incomplete charge transfer. Anyway, the basic idea, that of shot noise on incompletely transferred charge, is appealing, however approximate and/or incomplete it may in fact actually be.

We can derive such a result as follows. We make the assumption that we can approximate the current I which flows from one storage region to the next during a single transfer as $I(Q)$, where Q is the charge to be transferred at any given time during the transfer phase. (If this assumption had been made in calculating the coefficient of incomplete charge transfer,[10,11] then certain of the results obtained would have been erroneous. Nonetheless, without this assumption, details of the charge transfer enter, which we wish to avoid.) As in the text we also assume that we can linearize the equations governing the noise. Thus, writing $Q = Q^0 + q$, the equation of motion

$$\dot{Q} = -I(Q) \tag{92}$$

becomes

$$\dot{Q}^0 = -I(Q^0) = -I^0 \tag{93a}$$

for the noiseless transfer, and for the noise

$$\dot{q} = -\frac{dI}{dQ^0}\, q + d_q(t), \tag{93b}$$

where $d_q(t)$ is the naturally arising statistical driving term. Since I depends only on Q, the derivative appearing in (93b) is total and is evaluated using $Q^0(t)$ from (93a). Solving for $q(t)$ is straightforward:

$$q(t) = \int_0^t dt' d_q(t') \exp\left(-\int_{t'}^t (dI/dQ^0)_{t''} dt''\right) \tag{94a}$$

$$= \int_0^t dt' d_q(t') \left(\frac{I^0(t)}{I^0(t')}\right), \tag{94b}$$

in which we have used (93a) in going from (94a) to (94b). The $[I(t)/I(t')]$ factor suppresses the shot-like noise associated with the transfer.

The statistics of q follow from those of d_q using eq. (94b). The statistics of d_q are such that

$$\langle d_q(t_1) d_q(t_2)\rangle = eI[Q^0(t_1)]\delta(t_1 - t_2). \tag{95}$$

(This is not an additional assumption, but rather (95) follows from our initial assumption that I depends only on Q.) It follows that at the end of the transfer cycle

$$\langle q^2\rangle = e\int_{Q_1}^{Q_0} dQ[I^0(Q_1^0)/I^0(Q)]^2, \tag{96}$$

where Q_0 is the initial charge $Q^0(0)$ to be transferred, and Q_1 is the mean charge $Q^0(t_f)$ left behind at the end of the transfer. [We note that in the absence of the expression factor, $\langle q^2\rangle = e(Q_0 - Q_1) \approx eQ_0$, full shot noise.]

To proceed, we must know $I(Q^0)$. If, toward the end of the transfer cycle, $Q \to Q_1 \gg Q_d$, where Q_d is the charge packet size above which the primary force driving the transfer current arises from the packet itself, then $I(Q)$ will be proportional to Q^2, and, using (97), $\langle q^2\rangle \approx eQ_1/3$. If, on the other hand, toward the end of the transfer cycle, $Q \to Q_1 \ll Q_d$, so that the primary force driving the transfer current is diffusion or fringing fields, the $I(Q)$ will be proportional to Q, and, using (96), $\langle q^2\rangle \approx eQ_1$. Taking a more specific example, let us set $CV = Q$ in eq. (8) of Ref. 11, as a realistic approximation. It

follows that

$$\langle q^2 \rangle = e \int_{Q_1}^{Q_0} dQ \frac{Q_1^2 \left(Q_1 + \frac{2kT}{e} C \right)^2}{Q^2 \left(Q + \frac{2kT}{e} C \right)^2}. \tag{97}$$

Equation (97) can be integrated exactly. For our purposes, it suffices to consider two limiting cases: $Q_1 \gg 2kTC/e$, in which case $\langle q^2 \rangle \approx eQ_1/3$, and $Q_1 \ll 2kTC/e$, in which case $\langle q^2 \rangle \approx eQ_1$. Thus, we find in fact that $\langle q^2 \rangle$ can be viewed roughly as the shot noise on the incompletely transferred charge. However, it should be noted that Q_1 is the *total*, and not the much smaller differential (αQ_0), charge incompletely transferred. While the results of this appendix are appealing as a short cut, the reader is strongly advised to keep the basic assumption [$I = I(Q)$] firmly in mind and to use extreme caution in generalizing this approach to other problems.

REFERENCES

1. F. L. J. Sangster and K. Teer, "Bucket-Brigade Electronics—New Possibilities for Delay, Time-Axis Conversion, and Scanning," IEEE J. Solid State Circuits, *SC-4* (June 1, 1969), pp. 131–136.
2. W. S. Boyle and G. E. Smith, "Charge Coupled Semiconductor Devices," B.S.T.J., *49*, No. 4 (April 1970), pp. 587–593.
3. A. van der Ziel, "Noise in Solid State Devices and Lasers," Proc. of the IEEE, *58* (1970), pp. 1178–1206.
4. A. van der Ziel, *Noise*, Englewood Cliffs, N.J.: Prentice-Hall, 1956.
5. A. van der Ziel, *Fluctuation Phenomena in Semiconductors*, New York: Academic Press, 1959.
6. A. van der Ziel, *Noise: Sources, Characterization, Measurement*, Englewood Cliffs, Prentice-Hall, 1970.
7. K. K. Thornber, "Noise Suppression in Charge Transfer Devices," Proc. IEEE, *60* (1972), pp. 1113–1114.
8. K. K. Thornber and M. F. Tompsett, "Spectral Density of Noise Generated in Charge Transfer Devices," IEEE Trans. Electron Devices, *ED-20* (1973), p. 456.
9. K. K. Thornber, "Incomplete Charge Transfer in IGFET Bucket-Brigade Shift Registers," IEEE Trans. Electron Devices, *ED-18* (1971), pp. 941–950.
10. C. N. Berglund and K. K. Thornber, "Incomplete Transfer in Charge-Transfer Devices," IEEE J. Solid-State Circuits, *SC-8* (1973), pp. 108–116. References to earlier discussions of CTD's can be found herein.
11. C. N. Berglund and K. K. Thornber, "A Fundamental Comparison of Incomplete Charge Transfer in Charge Transfer Devices," B.S.T.J., *52*, No. 2 (February 1973), pp. 147–182.
12. J. E. Carnes and W. F. Kosonocky, "Noise Sources in Charge-Coupled Devices," RCA Review, *33* (1972), pp. 327–343.
13. K. K. Thornber, "Operational Limitations of Charge Transfer Devices," B.S.T.J., *52*, No. 9 (November 1973), pp. 1453–1482.
14. K. K. Thornber, "Error Rates of Digital Signals in Charge Transfer Devices," B.S.T.J., *52*, No. 10 (December 1973), pp. 1795–1809.
15. W. B. Joyce and W. J. Bertram, "Linearized Dispersion Relation and Green's Function for Discrete-Charge-Transfer Devices with Incomplete Transfer," B.S.T.J., *50*, No. 6 (July-August 1971), pp. 1741–1759.

16. C. N. Berglund, "Analog Performance Limitations of Charge-Transfer Dynamic Shift Registers, IEEE J. Solid-State Circuits, *SC-6* (1971), pp. 391–394.
17. M. F. Tompsett, "Quantitative Effects of Interface States on the Performance of Charge-Coupled Devices," IEEE Trans. Electron Devices, *ED-20* (1973), pp. 45–55.
18. K. K. Thornber, "Treatment of Microscopic Fluctuations in Noise Theory," B.S.T.J., *53*, No. 6 (July-August 1974), pp. 1041–1078.
19. M. Shoji, "Analysis of High-Frequency Thermal Noise of Enhanced Mode MOS Field-Effect Transistors," IEEE Trans. Electron Devices, *ED-13* (1966), pp. 520–524.
20. S. Y. Wu, "Theory of Generation-Recombination Noise in MOS Transistors," Solid-State Elec., *11* (1968), pp. 25–32.
21. I. Flinn, G. Bew, and F. Berg, "Low Frequency Noise in MOS Field-Effect Transistors," Solid-State Elec., *10* (1967), pp. 833–845.
22. A. G. Jordan and N. A. Jordan, "Theory of Noise in Metal-Oxide-Semiconductor Devices," IEEE Trans. Electron Devices, *ED-12* (1965), pp. 148–156.
23. W. Shockley, J. A. Copeland, and R. P. James, "The Impedance Field Method of Noise Calculation in Active Semiconductor Devices," in *Quantum Theory of Atoms, Molecules, and the Solid State*, edited by P. O. Lowdin, New York: Academic Press, 1966, pp. 537–563.
24. L. Boonstra and F. L. J. Sangster, "Progress in Bucket-Brigade Charge-Transfer Devices," 1972 IEEE Solid-State Conf., Digest of Technical Papers, *15* (1972), pp. 140–141.
25. R. J. Strain, "Properties of an Idealized Traveling-Wave Charge-Coupled Device," IEEE Trans. Electron Devices, *ED-19* (1972), pp. 1119–1130.
26. W. Feller, *An Introduction to Probability Theory and Its Applications*, New York: John Wiley, 2nd Edition (1957), pp. 216 and 263; 3rd Edition (1968), pp. 230 and 282.
27. T. C. McGill, M.-A. Nicolet, and K. K. Thornber, "Equivalence of the Langevin Method and the Impedance-Field Method of Calculating Noise in Devices," Solid-State Elec., *17* (1974), pp. 107–108.
28. K. K. Thornber, T. C. McGill, and M.-A. Nicolet, "Structure of the Langevin and Impedance-Field Methods of Calculating Noise in Devices," Solid-State Elec., *17* (1974), pp. 587–590.
29. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Washington, D. C.: National Bureau of Standards (1965), p. 775, eq. 22.3.2.
30. Ibid., p. 334, eq. 8.5.4.
31. Ibid., p. 783, eq. 22.9.1.
32. Ibid., p. 335, eq. 8.9.1.
33. Ibid., p. 1024, eq. 29.3.51.
34. J. E. Iwerson, private communication.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 7, September 1974
Printed in U.S.A.

A Statistical Analysis of Telephone Noise

By B. W. STUCK and B. KLEINER

(Manuscript received January 28, 1974)

The results of a statistical analysis of telephone noise are presented. The analysis consists of two stages: an exploratory data analysis stage, where the data are characterized through various nonparametric statistics and a model-building stage, where the data are matched to models.

The exploratory data analysis stage involved examination of noise waveforms, power spectra, and covariance estimates. The results show that telephone noise consists of a deterministic component (sinusoids at various frequencies) and a stochastic component. It is assumed that these components add. The data are filtered to remove the deterministic component. Next, central moment estimates are presented, as well as first-order amplitude statistics (histograms and empirical cumulative distributions) for these filtered data. The results indicate that the filtered data appear wide-sense stationary over short periods of time (typically 1 second) and, although close to gaussian, are distinctly nongaussian.

The model-building stage involved fitting the filtered data to two classes of models. The first class of models is based on symmetric stable distributions that arise from the central limit theorem. The second class of models assumes two different physical processes that contribute to the random component of telephone noise: The low-variance process is assumed to be gaussian, while the high-variance component is assumed to be a filtered Poisson process. Both classes of models agree intuitively with the physical processes generating telephone noise and are mathematically tractable. Based largely on graphical tests, both models appear to fit the filtered data reasonably well.

I. INTRODUCTION

Noise on telephone lines has puzzled and plagued people since the invention of the telephone. While it is common knowledge that telephone channel noise is nongaussian, nowhere in the literature is there a clear account of an adequate statistical characterization of telephone

noise. In part, this is due to the fact that only recently have statistical tools been developed that are equal to the task.

This paper attempts to adequately characterize some statistical properties of telephone channel noise by means of various nonparametric statistics and by mathematical models. It is encouraging to note that the results presented here do not contradict those found in earlier works. However, since only a small number of telephone line noise sample functions were examined, the results must be regarded as tentative, awaiting independent checks by other investigators. It is hoped the results presented here will stimulate communication theorists to investigate new methods for optimally processing signals corrupted by the nongaussian noise models presented here. Work along these lines might lead to optimum and practical suboptimum receiver structures for combating telephone noise. This in turn might permit greater insight into how noise limits telephone channel performance with regard to voice communication or data transmission.

The authors have tried to keep the notation and nomenclature consistent with that used in statistics and probability theory. The reader is reminded, for example, that "empirical cumulative distribution function" refers to an estimate of the true "cumulative distribution function" based on observations of "empirical" data. The words "sample" and "empirical" are often used interchangeably, as in "sample mean" and "empirical mean," as compared with the ensemble mean.

1.1 Summary of past work

Broadly speaking, previous investigators characterized telephone channel noise in two different ways, based on different ways of measuring the data and with different problems in mind. First, direct measurements of sample functions of telephone channel noise have been carried out[1–3] and mathematical models for the noise statistics have been constructed. Second, digital signals have been transmitted over telephone lines and the difference between the transmitted and received signals has been analyzed, providing an indirect measurement of telephone channel noise.[4–16] It is extremely difficult to correlate these two types of measurements. This paper is solely concerned with direct measurements of telephone channel noise sample functions.

Both types of measurements indicated the nongaussian nature of the noise. The analog measurements suggested that telephone noise could be considered a mixture of a nongaussian random process with sinusoids at various frequencies.[1–3] The first-order amplitude statistics for the random process appeared to be adequately modeled by a Pareto

distribution.[2] Analysis of errors in digital signals transmitted over telephone lines revealed that errors cluster in time, an indirect measure that the noise cannot be adequately modeled as white and gaussian.[4,13]

Some causes of telephone noise are thermal noise in amplifiers,[1,17] switches sparking on opening or closing contact,[1–3,17–19] lightning,[3,17] electromagnetic crosstalk,[17] fading on microwave links and switching to guard bands,[3,17] echo suppressor turnaround,[17] disturbances because of maintenance,[17] power line harmonics as well as harmonics of all sinusoids used in the telephone switching system,[17] and noise generated at switch interfaces.[17] The main causes of telephone noise are felt to be due to thermal noise, switch arcing, and pickup of unwanted sinusoidal harmonics.[17] The main cause of error clustering in digital signals is felt to be due to the impulsive component of the noise, generated by switch arcing.[17]

1.2 Problem statement

The problem is twofold:

(*i*) To provide an adequate statistical characterization of telephone channel noise by means of various nonparametric statistics.

(*ii*) To allow the data plus knowledge of the physical processes generating telephone noise to lead to a mathematically tractable class of models.

1.3 Outline of the paper

The data from five telephone lines and the processing necessary to convert the data into a form suitable for further analysis are described first. The analysis is broken down into two steps, an exploratory data analysis stage where the data are characterized through various nonparametric statistics and a model-building stage where the data are matched to models.

The exploratory data analysis stage involved examination of noise waveforms, power spectra, and covariance estimates. The results show that the data consisted of a deterministic component (sinusoids at various frequencies) plus a stochastic component, which were assumed to be independent. The data were filtered to remove the sinusoids that were significantly larger than the stochastic component. Histograms and empirical cumulative distribution functions for the filtered data were examined, as well as central moment estimates. The filtered data appeared to be wide-sense stationary over short periods of time, typically 1 second. Based largely on quantile-quantile plots, it was concluded that, although close to gaussian, the filtered data for three out

of the five lines were distinctly nongaussian; the filtered data for the remaining two lines appeared to be gaussian.

The model-building stage involved fitting the filtered data to two classes of models. The first class of models is highly unstructured; it was based on stable distributions, infinitely divisible distributions that arise from the central limit theorem. Based on a series of parameter estimation procedures including robust estimation, maximum likelihood estimation, and quantile-quantile plots and backed up by a likelihood ratio test on the goodness-of-fit, the three nongaussian lines could be adequately modeled by a stable distribution with characteristic index ≈ 1.95 (a gaussian has characteristic 2.0).

The second class of models is much more structured than the first. Two different physical processes were assumed to contribute to the filtered data: the low-variance component was a stationary gaussian process, while the high-variance component was a filtered Poisson process. Parameters for the gaussian component were estimated using trimmed means and trimmed variances. The parameters specifying the filtered Poisson process were much more complicated to estimate. The instants of time at which noise bursts occurred and the intervals between bursts were first examined; based on power spectra as well as covariance estimates, the intervals appeared to come from a renewal process. Histograms and empirical cumulative distribution functions indicated that the time intervals came from a Poisson process; empirical survivor and hazard function plots showed that a Poisson process with constant rate parameter was not an adequate model, however. Because of the small number of bursts observed, it was quite difficult to fit the time intervals to a Poisson process with varying rate parameter, and for expediency a constant Poisson rate parameter was chosen to model noise burst times of occurrence. The amplitudes of the noise bursts were adequately modeled by a log normal and power Rayleigh, or generalized gamma. The durations of actual noise bursts were used to estimate parameters in the noise burst shaping filter. A simple indication is presented of how well the filtered data fit the gaussian-plus-filtered-Poisson-process model. A number of other models and extensions of these models are discussed.

II. EXPLORATORY DATA ANALYSIS

2.1 Description of the data

The data, supplied by J. Fennick, consist of analog tape recordings of noise on five telephone lines. Figure 1 is a diagram of the measure-

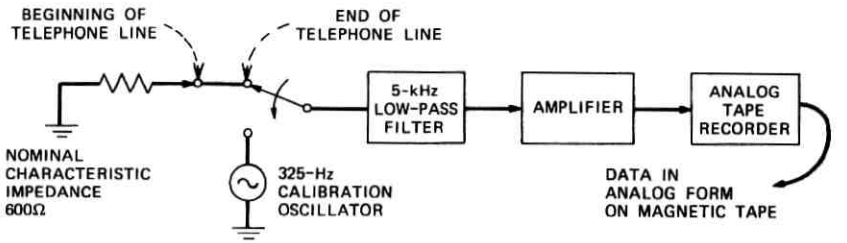


Fig. 1—Telephone noise measurement system.

ment system. Table I describes the origin and destination of each line, as well as the date and time of the recording.

The origin of the telephone line was terminated with the nominal characteristic impedance of the line, 600 ohms. The output of the line was low-pass filtered to remove spurious out-of-band signals, amplified, and recorded on an analog tape at 18.75 cm per second. The system was calibrated before each recording with a 15-second burst of a 325-Hz sinusoid at a predetermined amplitude. The dynamic range of the recording system was approximately 100.[20] No attempt was made to eliminate dc offset and drift. Each recording was approximately 15 minutes long.

Figure 2 is a block diagram of the analog-to-digital tape conversion system. The analog tapes were played back on an analog tape recorder (of a different model than that on which they were recorded) at 18.75 cm per second. The calibration signal set the gain on the playback amplifier so that the calibration signal amplitude was roughly equal to its value at the recording site. There was no attempt to remove wow and flutter in the tape recording.[21] The signal was low-pass filtered (to lessen the chance for spectral aliasing), amplified, sampled 10,000 times a second, passed through an analog-to-digital converter, and subsequently put into digital format on a tape suitable for further processing

Table I — Description of data

Line	Origin	Destination	Approximate Length (km)	Date	Time
1	Monroe, Mich.	Detroit, Mich.	55	7/8/64	2:06 p.m.
2	Saginaw, Mich.	Detroit, Mich.	160	7/8/64	2:49 p.m.
3	St. Louis, Mo.	Ft. Smith, Ark.	690	8/4/64	10:30 a.m.
4	Little Rock, Ark.	Ft. Smith, Ark.	250	8/4/64	1:55 p.m.
5	Newark, N. J.	Binghamton, N. Y.	400	12/12/63	12:52 p.m.

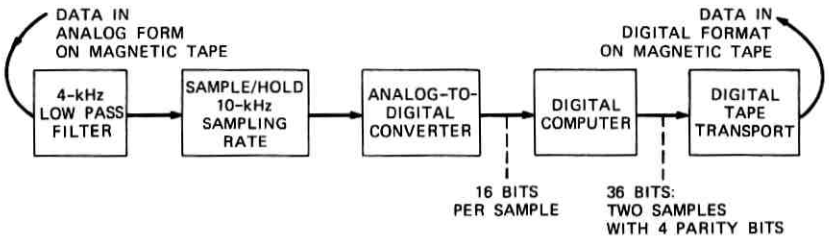


Fig. 2—Analog-to-digital conversion of telephone noise data.

by a Honeywell 6070 digital computer. The levels of the analog-to-digital converter will be the units specifying the amplitudes of the telephone noise data. No measures of the degradation in data on digital tapes resulting from jitter during the sampling process are available; it is assumed to be negligible compared to other sources of measurement error. No bounds are available on the loss of information entailed by examining a continuous-time random process at only discrete instants of time.[22]

Two critical remarks concerning the method of recording the data should be made: first, there are no quantitative measures available on the amount of noise introduced into the data by the measurement system alone. Presumably, any measurement system noise was insignificant compared to the telephone channel noise. Second, the dynamic range of the recording system is probably insufficient to faithfully record telephone noise; a much more satisfactory dynamic range would be 1000 to 10,000. Both issues have been dealt with elsewhere (in a different context); a possible solution would be to convert the data into digital format directly at the recording site.[23] Consideration of these problems is left to future research; the data analysis proceeded with these caveats in mind.

Figure 3 shows a typical telephone channel noise waveform from line 1 after conversion to digital format.

How typical are these data compared with that observed on other telephone lines? A search of the literature as well as private communications from engineers shows that the data discussed here appear to be typical of telephone channel noise. Throughout this investigation, nothing was uncovered that contradicted earlier work; rather, this work tends to clarify and place in perspective that of earlier investigators. Note too that the telephone lines examined here were typically several hundred miles long, presumably passing through a variety of equipment, and hence quite representative of telephone noise. Finally,

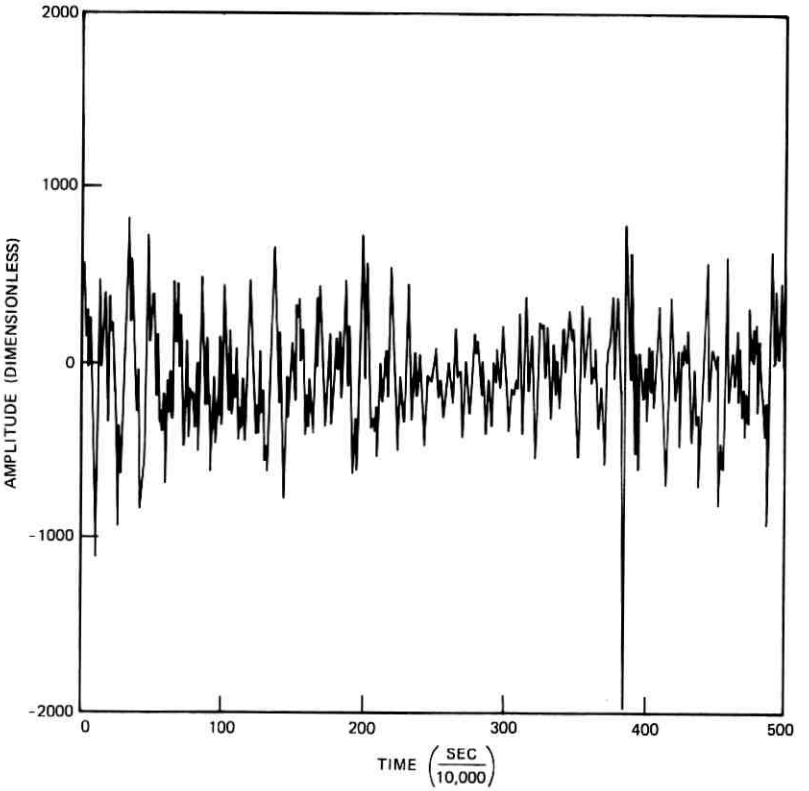


Fig. 3—Typical line 1 telephone noise waveform (unfiltered).

the analog tape recordings were played into loudspeakers, and the authors felt the noise sounded typical.

A much more serious objection to the present analysis is that not enough of the data at hand was examined. If all five 15-minute noise recordings were sampled 10,000 times a second and then put on to tape in digital format, more than 45 million noise data must be analyzed; in particular, 9 million data must be processed and statistically characterized for just one telephone line sample function. In practice, only 10-second segments from the beginning and middle of a recording were examined in detail and compared with each other. No unusual statistical differences were observed between these segments for any telephone line. The main reason for examining so little data was the great cost of analyzing these data statistically and, in particular, of digitally filtering the data to remove sinusoids. It is difficult to predict

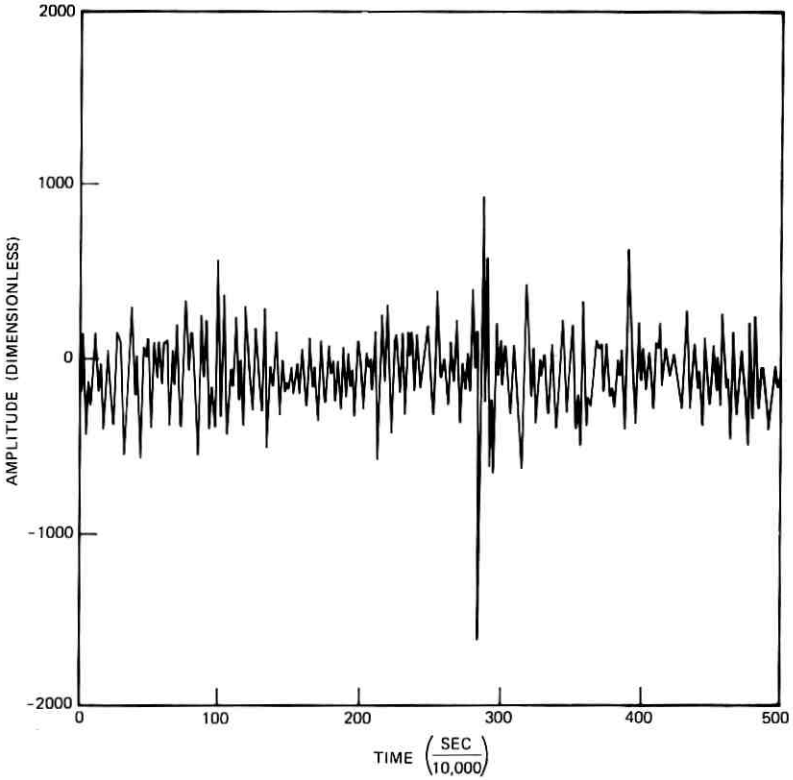


Fig. 4—Typical line 1 telephone noise waveform (filtered).

in advance exactly which sinusoids are present on a particular telephone line; it is easier to filter these out digitally after the measurement without distortion than to accomplish this with analog filters.

2.2 Data modifications and estimation of power density spectra

As mentioned above, the data went through several stages before they were available in digital form. Some further processing is necessary to remove the effects of this pre-processing, as well as to remove unwanted sinusoids. Since the frequency response of each line was unknown, nothing was done to compensate for it.

The first step is to compute estimates of the power spectrum. The data were segmented into blocks (typically of length 1000, corresponding to $\frac{1}{10}$ of a second of noise). Each block was tapered and enlarged to 1024 values by adding zeros, then transformed into the frequency

domain by the fast Fourier transform (FFT);[24] the power spectrum was estimated from the transformed data. The method used is comparable in bias and variance to other spectral estimation procedures, but requires considerably less computer time than other non-FFT-based methods.[25–27] Furthermore, it is possible to calculate cross-spectrum estimates easily, as well as to check for nonstationarity by computing the spectra of successive segments of data. The discrete Fourier transform of N successive data at frequency $w_l = l/N\,\Delta t$ is $\hat{f}(w_l)$,

$$\hat{f}(w_l) = \sum_{K=1}^{N} n_K e^{-i2\pi Kl/N} \qquad l = 0, 1, \cdots, \left[\frac{N}{2}\right],$$

where

n_K = sample of noise waveform at time $K\Delta t$

and

Δt = time interval between samples.

The estimate of the power spectrum density at frequency w_l is $\hat{S}(w_l)$,

$$\hat{S}(w_l) = \frac{\Delta t}{N} \sum_{j=-M}^{M} |\hat{f}(w_{l+j})|^2 g_j,$$

where

$$\sum_{j=-M}^{M} g_j = 1,$$

and the weights $\{g_j\}$ are introduced to smooth the estimate. Unequal weights can be used to lower the bias of the estimate, but increase its variance. The value $\hat{S}(w_l)$ represents the average noise power density in a frequency band centered at w_l. All power spectrum density estimates shown here were computed with g_j equal to $(M - |j|)/M^2$, where $M = 5$. Figure 5 shows the power spectrum for the waveform in Fig. 3 with two sharp peaks probably reflecting sinusoids at 650 and 4300 Hz. Figure 6 shows a succession of 24 power spectrum estimates for line 1 for $\frac{1}{10}$-second segments of filtered data. The first 13 are from successive segments recorded at the beginning, while the final 11 are from successive segments recorded 5 minutes later. These results indicate that line 1 data can be regarded as wide-sense stationary over at least 1-second time intervals. The variance of these spectral estimates is unknown; if the process were gaussian, the distribution of $S(\hat{w}_l)$ can be approximated by a χ^2 distribution with $[2/\sum_{j=-M}^{+M} g_j^2]$ degrees of freedom.[28] For $g_j = (M - |j|)/M^2$ with $M = 5$, this results in approximately 14 degrees of freedom. From the data shown here for line

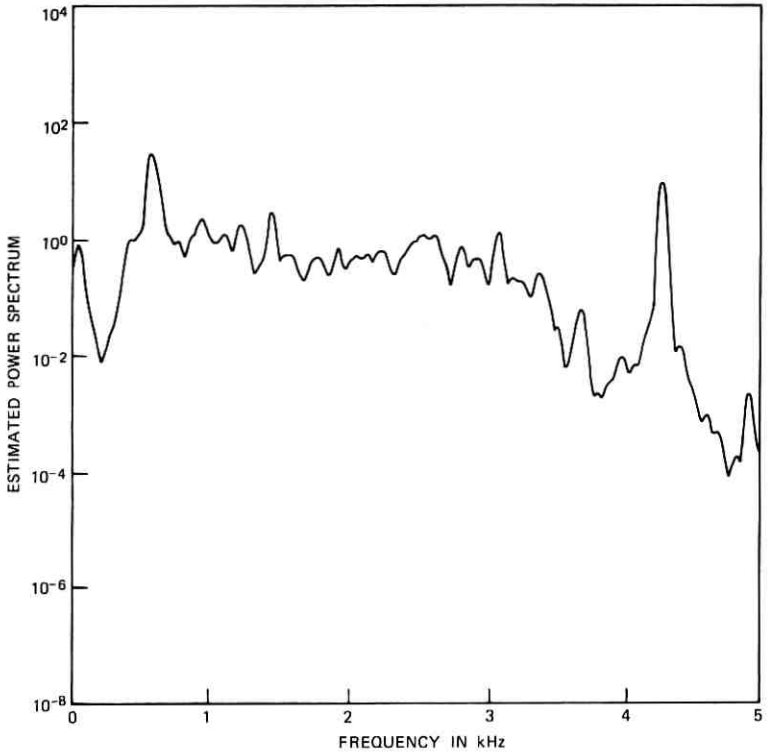


Fig. 5—Typical power spectrum for unfiltered line 1 data ($N = 1000$).

1, as well as data from the other four lines, telephone noise power density spectra appear to have the same shape, but different scales.

Table II summarizes the estimates of frequencies of signals which were quite probably sinusoids, and whose estimated power spectrum density was at least a factor of 10 above the estimated wideband power

Table II — Estimated frequencies of sinusoids that were subsequently filtered out

Line	Estimated Frequency of Sinusoid (Hz)
1	650, 4300
2	650, 4300
3	3900
4	2000, 3600
5	60

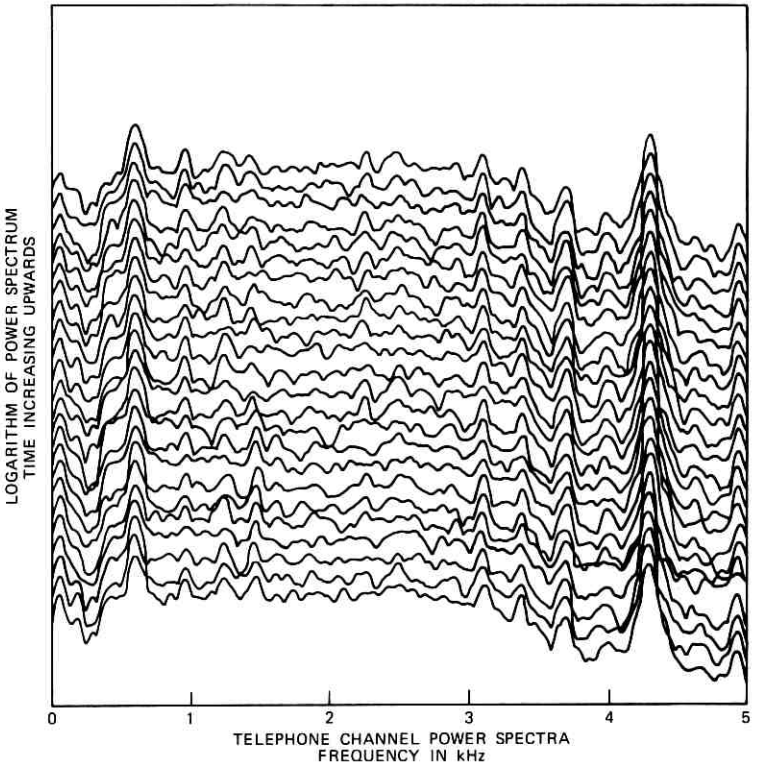


Fig. 6—Line 1 power spectra: bottom 13 from beginning of line 1 data, top 11 from middle of line 1 data ($N = 1000$).

spectrum density. While this criterion is arbitrary, independent experimental evidence to be discussed shortly indicates it is adequate from a statistical point of view.

Since many statistical tests require uncorrelated samples, it is necessary to filter out these sinusoids, as well as to compensate for distortions in the data from the measurement system. This implicitly assumes that telephone noise can be modeled as the sum of a deterministic process, sinusoids at various frequencies, and a purely stochastic process, which will be characterized in greater detail. This was accomplished using low-pass, band-stop, and high-pass linear-phase digital filters designed by computer programs developed by L. Rabiner; the filtering was carried out in the discrete time domain by convolution. Figure 7 shows the power spectrum of a filtered segment of the data shown in Fig. 4.

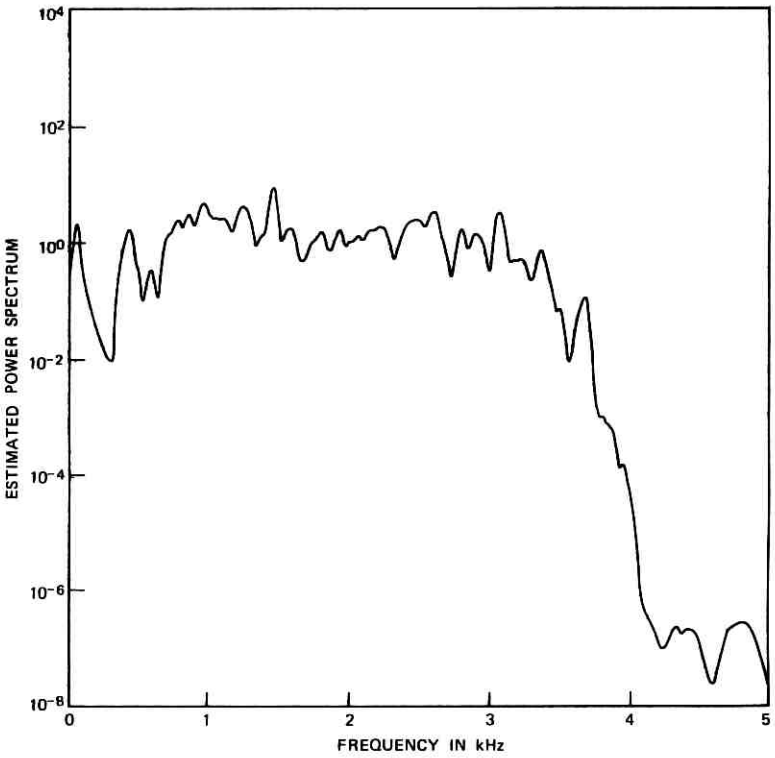


Fig. 7—Typical power spectrum for filtered line 1 data.

Because of the difficulty in finding telephone lines completely free from sinusoidal interference, the question arises as to how much harmonic content can be tolerated in performing various statistical tests. Work carried out elsewhere in a different context has examined this issue from an experimental viewpoint;[29] the principal findings were that the amplitude statistics are apparently not significantly degraded by the linear filtering, if the sinusoid is the same size or smaller than the observed noise levels. This topic can be a subject for future research.

2.3 Covariance estimation

It is assumed in many statistical computations that the data are statistically independent. In practice, the data usually depend to some extent on each other, and it is often quite difficult to quantitatively estimate the effects of this lack of independence. One indication of

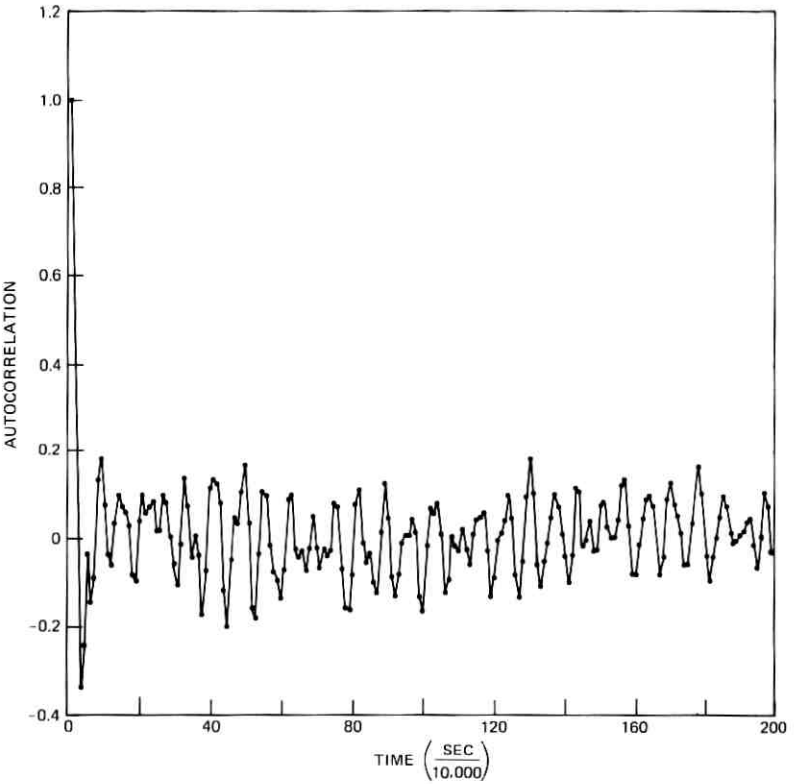


Fig. 8—Typical line 1 autocorrelation function ($N = 1000$).

independence is the estimated autocorrelation function,

$$\hat{R}_{nn}(l\Delta t) = \frac{1}{\bar{R}_{nn}(0)} \cdot \frac{1}{N} \sum_{K=1}^{N-l} n'(K\Delta t)n'(K\Delta t + l\Delta t) \quad l = 1, \cdots, N-1,$$

where

$$n'(K\Delta t) = n(K\Delta t) - \frac{1}{N} \sum_{l=1}^{N} n(l\Delta t)$$

and

$$\bar{R}_{nn}(0) = \frac{1}{N} \sum_{K=1}^{N} n'^2(K\Delta t).$$

A typical autocorrelation of filtered data is plotted in Fig. 8. A sinusoid that was not filtered out is quite evident at approximately 1400 Hz (see also Fig. 7); ignoring this sinusoid,[29] the autocorrelation appears to be approximately zero for $l \geqq 3$.

If the data are wide-sense stationary and ergodic, then the autocorrelation and the power density spectrum are a Fourier transform pair.[30]

Examination of the filtered waveform in Fig. 4 indicates that the samples appear uncorrelated, i.e., they are scattered in a random fashion about a location parameter.

The sample normalized cross covariance for two different segments of data, $\{x(\Delta t), x(2\Delta t), \cdots, x(N\Delta t)\}$ and $\{y(\Delta t), y(2\Delta t), \cdots, y(N\Delta t)\}$, is defined as

$$\hat{R}_{xy}(l\Delta t) = \frac{\frac{1}{N}\sum_{K=1}^{N-l} x'(K\Delta t)y'(K\Delta t + l\Delta t)}{\tilde{R}_{xx}^{\frac{1}{2}}(0)\tilde{R}_{yy}^{\frac{1}{2}}(0)} \quad l = 0, 1, \cdots, N-1,$$

where

$$x'(K\Delta t) = x(K\Delta t) - \frac{1}{N}\sum_{l=1}^{N} x(l\Delta t),$$

$$y'(K\Delta t) = y(K\Delta t) - \frac{1}{N}\sum_{l=1}^{N} y(l\Delta t),$$

$$\tilde{R}_{xx}(0) = \frac{1}{N}\sum_{K=1}^{N} x'^2(K\Delta t),$$

and

$$\tilde{R}_{yy}(0) = \frac{1}{N}\sum_{K=1}^{N} y'^2(K\Delta t),$$

and is shown in Fig. 9 for two typical segments of filtered data. From this as well as other data, the filtered telephone noise data examined appear to be uncorrelated over short time intervals.

Since the data, after filtering, appear approximately uncorrelated, they will now be characterized in greater detail.

2.4 First-order filtered data amplitude statistics

A nonparametric statistical description of first-order noise amplitude statistics provides a great deal of useful information. For example, if the data are independent identically distributed random variables, they can be completely characterized by their empirical cumulative distribution function.[31] The work in this section relies heavily on graphical methods for data analysis, to give more physical insight into the nature of the data.[32]

The empirical or sample cumulative distribution function is defined as

$$\hat{F}(X) \triangleq \frac{\text{number of observations less than or equal to } X}{\text{total number of observations}},$$

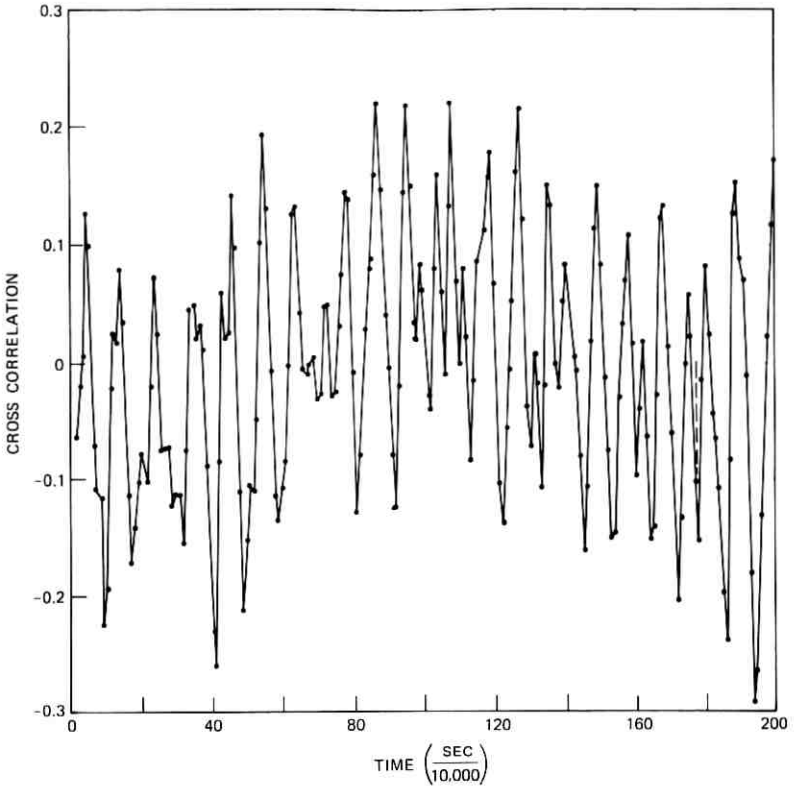


Fig. 9—Typical line 1 crosscorrelation function ($N = 1000$).

which is a function of $\{x_K\}$, the set of observations. The sample histogram is defined as

$$P(X, X + \Delta) \triangleq \text{number of sample values in } [X, X + \Delta],$$

where Δ is the bin width. Figure 10 is a plot of a typical empirical cumulative distribution function, and Fig. 11 shows a typical sample histogram. These two figures imply that the first-order probability density for the data is roughly bell-shaped and symmetric. A simple graphical symmetry check on the empirical cumulative distribution is shown in Fig. 12; x_K is plotted against x_{N-K+1}, where $K = 1, 2, \cdots, [N/2]$, $N = 1000$ is the total number of observations, and $\{x_K\}$ is the set of ordered observations. If the empirical cumulative distribution

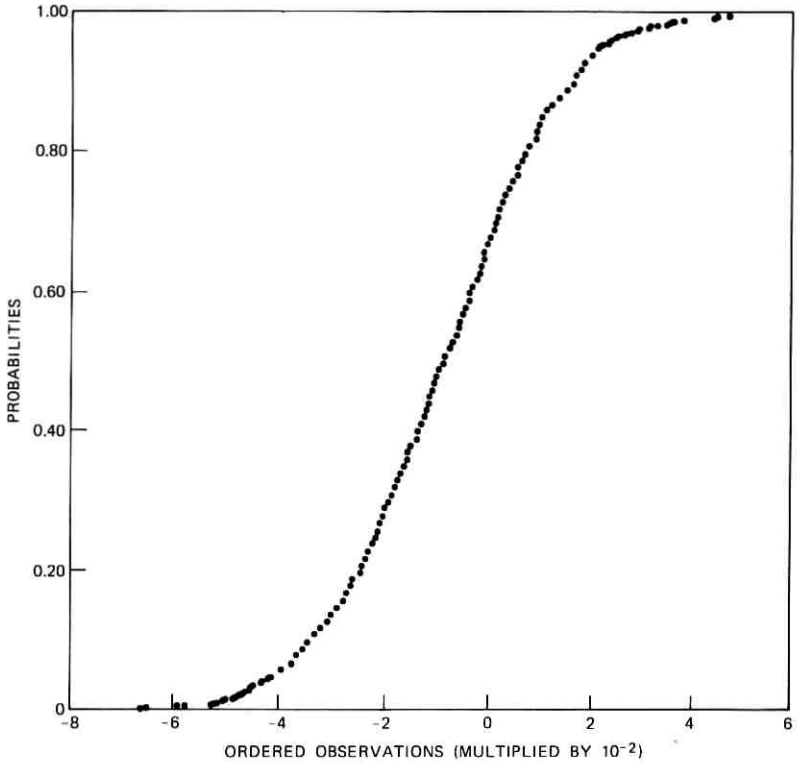


Fig. 10—Typical line 1 empirical cumulative distribution function ($N = 1000$).

is symmetric, these points lie on a straight line with negative unit slope; this is apparently the case.

The next quantities of interest are central moment estimates, which are defined as follows:[33]

$$\bar{x} = \text{sample mean} = \frac{1}{N}\sum_{j=1}^{N} x_j,$$

$$s^2 = \text{sample variance} = \frac{1}{N}\sum_{j=1}^{N} (x_j - \bar{x})^2,$$

$$\hat{a}_3 = \text{sample skewness} = \frac{1}{N}\sum_{j=1}^{N} (x_j - \bar{x})^3/(s^2)^{\frac{3}{2}},$$

and

$$\hat{a}_4 = \text{sample kurtosis} = \frac{1}{N}\sum_{j=1}^{N} (x_j - \bar{x})^4/(s^2)^2.$$

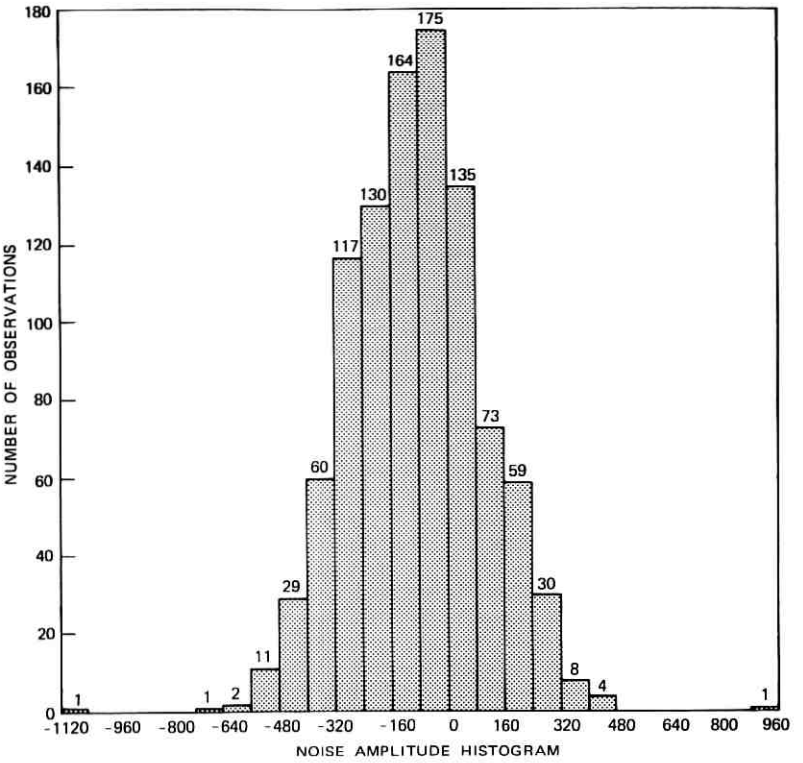


Fig. 11—Typical line 1 histogram ($N = 1000$).

These parameters were estimated for ten segments of 1000 data for each of the five lines. Table III shows these estimates for the segment of each line whose fourth central moment was the median of all ten fourth-central-moment estimates of this line. The 5-percent significance level for 1000 independent identically distributed gaussian random variables with known mean and variance are[33]

$$-0.127 < \hat{a}_3 < 0.127$$
$$2.76 < \hat{a}_4 < 3.26.$$

Figure 13 shows a scatter plot of $\hat{a}_3$ vs. $\hat{a}_4$ for successive segments of 1000 data for each of the five lines. Based on this evidence, it can be conjectured that lines 1, 2, and 4 are nongaussian, while the gaussian hypothesis cannot be rejected for lines 3 and 5. Since quite a large body of literature exists on gaussian random processes and these random processes are well understood, the gaussian hypothesis is not

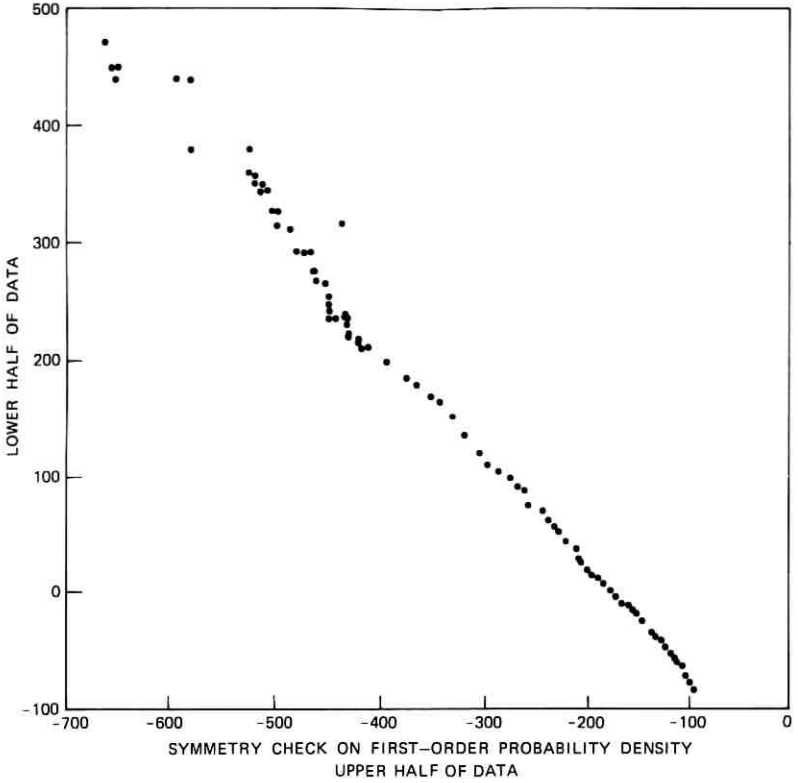


Fig. 12—Typical symmetry check on line 1 empirical cumulative distribution function ($N = 1000$).

lightly discarded: the evidence that the data are nongaussian should be much more convincing than that presented so far.

A very convenient graphical method to check how well data fit a theoretical distribution function is the quantile-quantile, or Q-Q, plot.[32] The qth quantile of a cumulative distribution function $F(x)$ is defined here as the value x for which $F(x) = q$, $0 \leqq q \leqq 1$. A Q-Q plot plots

Table III — Estimated telephone noise central moments

Line	$\bar{x}$	s^2	$\hat{a}_3$	$\hat{a}_4$
1	−87.9	38,600	0.05	3.4
2	−80.1	18,200	0.08	3.5
3	−80.1	44,200	−0.05	3.1
4	−82.3	87,200	0.02	3.3
5	− 1.06	1,990	−0.08	2.9

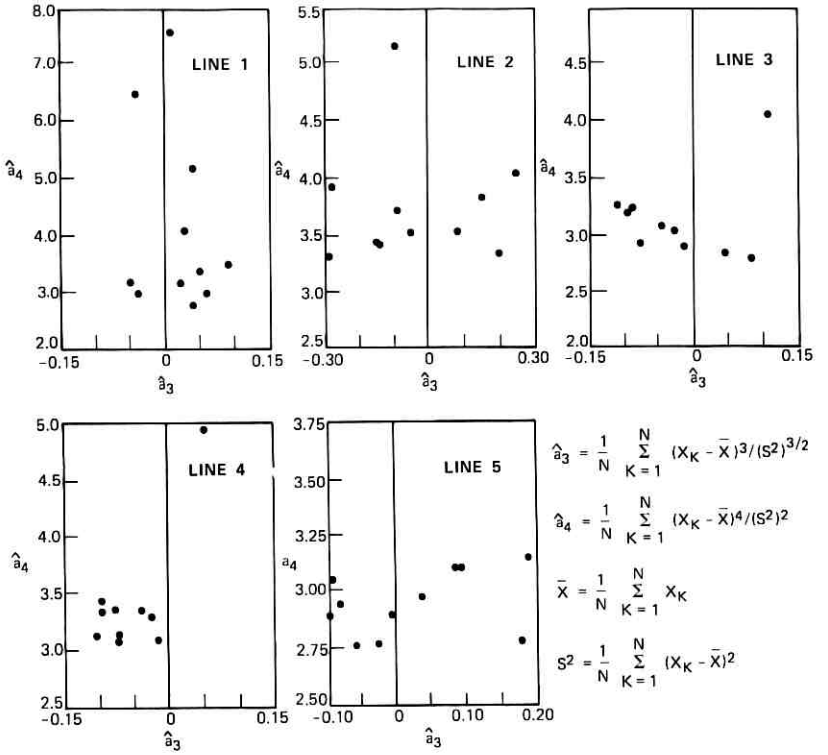


Fig. 13—Scatter plot of estimated third central moment $\hat{a}_3$ vs estimated fourth central moment $\hat{a}_4$ for at least ten successive segments of 1000 data ($N = 1000$) for all five lines.

quantiles of the empirical cumulative distribution function against quantiles of the theoretical distribution. If the empirical and theoretical distribution functions are the same, the plot is a straight line with slope +1 passing through the origin. If the empirical and theoretical distribution functions are the same to within a linear transformation (i.e., to within a scale and location parameter) the plot is still a straight line. A typical quantile-quantile plot for line 1 filtered data against a gaussian distribution is shown in Fig. 14; the sample size was 13,000. The first 100 and last 100 quantiles, as well as every hundredth quantile in the middle, have been plotted, giving the illusion of discontinuity during the transition from the middle to the tail quantiles.[32] Ten observations in each tail are widely scattered. Figure 15 shows the central portion of the quantile-quantile plot with these observations excluded. The tails curve toward horizontal lines,

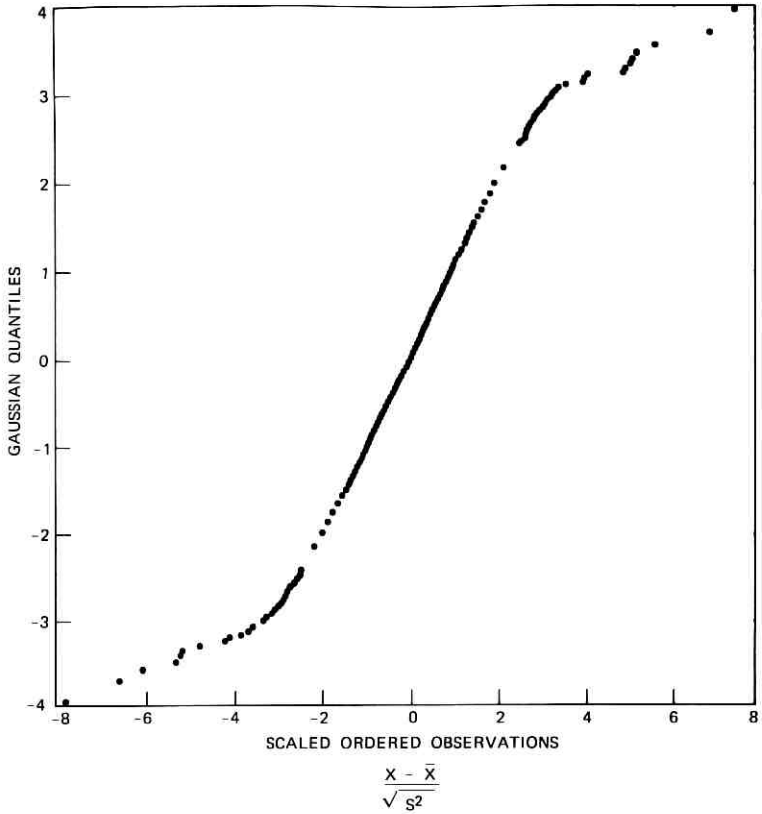


Fig. 14—Q-Q plot for 13,000 line 1 data against gaussian distribution ($\bar{X}$ = sample mean, S^2 = sample variance).

another indication of the long-tailed nongaussian nature of the data. The 10 points on each tail were found to be highly correlated: These very large excursions occurred in clumps of two, three, and five at a time, violating the assumption that the data are independent. For comparison, Fig. 16 shows a quantile-quantile plot for line 5 filtered data against a gaussian distribution; the sample size was 11,000. The straight line is a good indication that these data are gaussian.

III. MODELS

3.1 Central limit theorem

Since noise on telephone lines is presumably due to a large number of independent causes, it is worthwhile to digress and review the central

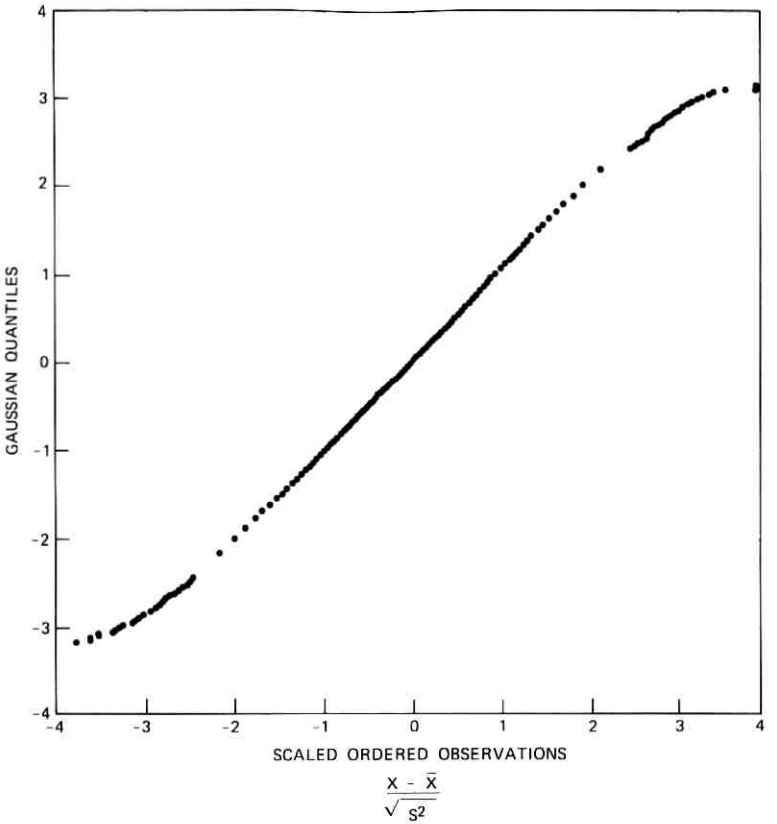


Fig. 15—Center portion of line 1 Q-Q plot for gaussian model (N = 12,980) ($\bar{X}$ = sample mean, S^2 = sample variance).

limit theorem. The material presented here is largely tutorial, following standard references.[34,35] The close association between the central limit theorem and the gaussian distribution is remarkable because of its algebraic closure property: If two independent random variables are both gaussian, their sum is also. It is much less widely known that the gaussian distribution is a special case of a larger family of distributions, which arise from the central limit theorem and exhibit the same closure properties as the gaussian, the stable distributions.

The reason for the importance of the gaussian rather than the entire stable distribution family is that only the gaussian distribution has a finite variance, and infinite variance is felt to be physically inappropriate in virtually any context. However, this is naive in that

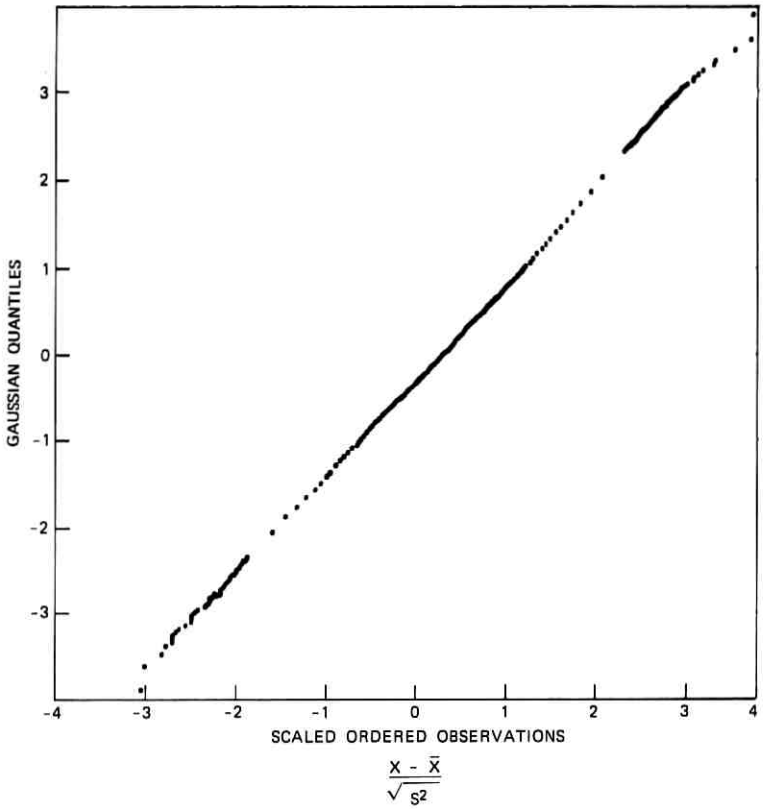


Fig. 16—Q-Q plot for 11,000 line 5 data against gaussian distribution ($\bar{X}$ = sample mean, S^2 = sample variance).

the gaussian distribution is unbounded and unbounded quantities are also felt to be physically inappropriate. The gaussian distribution may describe a particular situation adequately over a certain range; an infinite-variance distribution may model an actual situation over a greater range of a parameter. Both distributions may be physically inappropriate, but the infinite variance may, in this sense, account for the observations better than the gaussian.

Mandelbrot has pioneered in developing and popularizing this notion,[36–38] for example, in connection with economic phenomena[39–42] and in error statistics of digital signals transmitted over telephone lines.[9] Consider, for example, the distribution of changes in stock market prices. Mandelbrot[39] and Fama[43] have shown that, although

the change in stock market prices is bounded, the probability of very large deviations is so great that many statistical techniques that assume an underlying distribution with finite variance are not applicable. Stock market prices may be modeled as a sum of a large number of random variables; similarly, at any instant of time, telephone noise is presumably the sum of a large number of random variables. The sum of a large number of infinite-variance variables is often dominated by one or a few of the summands[44]—a theoretical property of infinite-variance distributions. The key feature common to these models is that the limiting distribution remains the same if an arbitrary but finite number of terms are dropped from the sum. This intuitive notion can be made precise and, subject to a mild restriction on the distribution from which the summands are drawn, leads naturally to the central limit theorem.[45]

Among infinite-variance distributions, the stable distributions play an important role, because only stable distributions can be limiting distributions of suitably normalized sums of independent identically distributed random variables, as well as because stable distributions are closed under convolution. Some pioneering work on the statistical analysis of data from a stable distribution has been carried out already; the analysis described here is a straightforward application of this work.[46–48] Before detailing that work, a summary is presented of some properties of stable distributions.

A distribution function $P(x)$ is called *stable* if, for all $a_1 > 0$, b_1, $a_2 > 0$, b_2, there exist constants $a > 0$, b such that

$$P(a_1x + b_1)*P(a_2x + b_2) = P(ax + b)$$

holds.[49] Every stable distribution has a continuous density; the stable distributions discussed in this work have unimodal densities that are analytic throughout their support.[50] The random variable x is stable if and only if the logarithm of its characteristic function is[51]

$$\ln [E(e^{ixw})] \triangleq \ln [\varphi_x(w)]$$
$$= \begin{cases} -|cw|^{\alpha}[1 - i\beta \cdot \text{sign } (w) \tan (\pi\alpha/2)] + i\delta w & \alpha \neq 1 \\ -|cw|[1 - i\beta 2/\pi \text{ sign } (w) \ln |cw|] + i\delta w & \alpha = 1 \end{cases}$$
$$-1 \leqq \beta \leqq 1 \qquad 0 < \alpha \leqq 2.$$

Thus, every stable law is described by four parameters α, β, c (or $\gamma = c^{\alpha}$), δ, where α is the characteristic index, β is associated with the skewness of the distribution, c is a scale parameter, and δ is a location parameter. If $\beta = 0$, the distribution is symmetric about $x = \delta$. If

$\beta > 0$ and $0 < \alpha < 2$, the distribution is skewed to the right, and the degree of skewness increases as β increases; conversely, if $\beta < 0$, the distribution is skewed to the left and the degree of skewness increases as β decreases. For $\alpha = 2$, β is irrelevant.[52]

If s_n is the suitably normalized sum of n independent identically distributed random variables $x_1, x_2, \cdots, x_n$,

$$s_n = \frac{1}{B_n}(x_1 + x_2 + \cdots + x_n) - A_n$$

where B_n and A_n are normalizing constants, then the distribution of x is said to belong to the domain of attraction of a stable distribution with characteristic index α if the distribution of s_n converges to this stable law as n goes to infinity;[53] this distribution belongs to the domain of partial attraction of a stable distribution if the distribution of s_n converges only for some subsequence.[54] A stable distribution with index α has absolute moments of all orders strictly less than α, i.e., $E[|x|^p] < \infty$ for $0 \leqq p < \alpha$.[55]

Stable distributions and densities can be expressed as a power series.[56] In several cases, this series can be considerably simplified to yield analytic closed-form expressions; these cases are $\alpha = 2$ (gaussian), $\alpha = 1$ and $\beta = 0$ (Cauchy), and $\alpha = \frac{1}{2}$ with $\beta = \pm 1$. Figure 17 depicts several stable density functions.

If $x_1, x_2, \cdots, x_n, \cdots$, are independent random variables drawn from r distributions, each within the domain of attraction of stable laws with indices drawn from the finite set $(\alpha_1, \cdots, \alpha_r)$, then under certain conditions on the number of representatives of each distribution, the suitably normalized sum of these variables converges to a distribution that is the convolution of r stable distributions.[57,58]

The question of rate of approach to the limiting distribution of a sum of suitably normalized, independent, identically distributed random variables is well understood if the limiting distribution is gaussian ($\alpha = 2$).[59–61] If the limiting distribution is in the domain of attraction of a stable distribution, a variety of results are available.[30,35,43,47] The most useful result[62] available at present, from a data analysis point of view (see Ref. 63), loosely states that the difference between the actual distribution of the sum of N suitably normalized random variables and the limiting stable distribution ($0 < \alpha < 2$) is bounded by a linear combination of terms of the order of $N^{-1/\alpha}$ and $N^{-(2-\alpha)/\alpha}$. As an example, consider the case $\alpha = 1.9$: one term is $N^{-1/\alpha} = N^{-0.53}$ while the other term is $N^{-(2-\alpha)/\alpha} = N^{-0.053}$; N must be

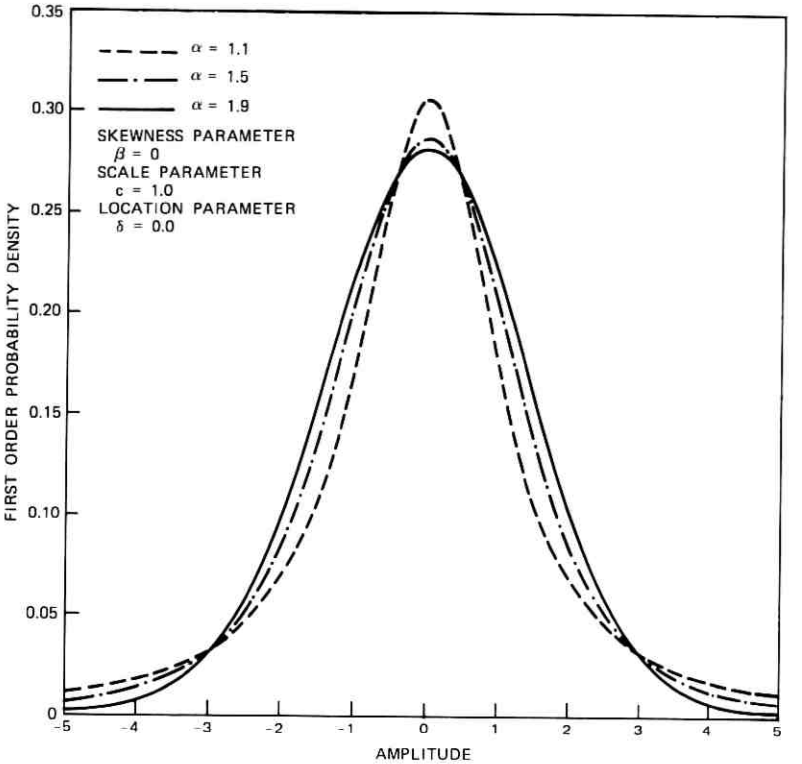


Fig. 17—Some symmetric stable first-order probability density function.

astronomically big to reduce this second term to a value smaller than 0.1, which indicates how slow this rate of convergence to a limiting stable distribution may be.[63] Thus, in many practical situations, caution must be shown in going to the limiting distribution.[64–69] Noise on telephone lines is possibly a case in point.

Section 3.2 discusses how filtered data from the three nongaussian telephone lines are fit to stable distributions. Since these distributions have no second moments, the modifications necessary to properly interpret power spectra and covariance estimates, as well as auto- and crosscorrelation estimates, for these three lines are not clear. This whole area must be subject to further research.[38]

As a final aside, the question of ergodicity, of relating time average statistics to ensemble average statistics, will not be addressed here: The filtered data are assumed to be an ergodic random process.

3.2 Symmetric stable distribution model

In this section, various statistical tests are described for determining if telephone noise on the three lines that appear nongaussian can be adequately modeled by a symmetric stable distribution ($0 < \alpha \leqq 2$, $\beta = 0$).

A series of estimators for symmetric stable distribution ($1 < \alpha < 2$) parameters have recently been developed.[47,48] These estimators are based on statistics easily derived from the empirical distribution function; they have been compared with maximum likelihood estimates and found to offer reasonable agreement when suitable precautions, such as a large sample size for α near 2, are taken.[46] These parameter estimates are

$$\hat{\delta} = \frac{1}{2Np} \sum_{K=(0.5-p)}^{(0.5+p)} \tilde{x}_K \qquad p = 0.125, 0.250, 0.375$$

$$\hat{c} = \frac{1}{1.654} (\tilde{x}_{0.72} - \tilde{x}_{0.28}),$$

where $\tilde{x}_r$ is the value of the rth empirical quantile, $\hat{\delta}$ is a trimmed mean, and $\hat{c}$ measures the spread of the distribution. To estimate the characteristic index α, an auxiliary variable z_q is first computed

$$z_q = \frac{\tilde{x}_q - \tilde{x}_{1-q}}{2\hat{c}} = 0.827 \frac{\tilde{x}_q - \tilde{x}_{1-q}}{\tilde{x}_{0.72} - \tilde{x}_{0.28}}$$

$$q = 0.9995, 0.995, 0.99, 0.985, 0.98, 0.975, 0.97, 0.96, 0.95, 0.94, 0.92,$$

and then $\hat{\alpha}$ is obtained as a function of z_q from tables in Ref. 48. $\hat{\alpha}$ is a measure of how rapidly the distribution approaches its asymptotic values. For line 1, with a sample size of 13,000, it was found that

$$\hat{\delta} = -87.4,$$
$$\hat{c} = 132.0,$$
$$\hat{\alpha}(z_{0.99}) = \hat{\alpha}(z_{0.995}) = 1.95.$$

In addition, this was carried out for a sample size of 1000 thirteen times, a sample size of 2000 six times, a sample size of 3000 four times, a sample size of 4000 three times, and a sample size of 5000 two times. The results are tabulated in Table IV for the $q = 0.98$ fractile. Different choices of q resulted in practically the same estimates.

Larger and larger samples were used because, if the data really come from a stable distribution, then the parameter estimates would presumably converge to their true values with increasing N.

Table IV — Line 1 symmetric stable distribution parameter estimates

	$N = 1000$	$N = 2000$	$N = 3000$	$N = 4000$	$N = 5000$
$\hat{\delta}$	−88.6, −88.8, −86.9, −87.2 −88.2, −88.0, −87.6, −87.6 −88.3, −88.7, −86.6, −87.5 −87.2	−87.0, −88.1 −87.6, −88.5 −87.7, −88.7	−88.1, −87.8 −87.8, −88.0	−87.9, −87.8 −88.1	−87.9, −88.0
$\hat{c}$	132.1, 127.4, 122.9, 124.7 145.2, 134.5, 128.6, 119.8 143.4, 134.5, 130.5, 124.2 137.6	123.7, 141.1 124.0, 137.6 135.1, 128.7	125.8, 135.7 129.1, 135.1	125.5, 132.3 135.9	130.6, 132.1
α	1.92, 1.94, 1.96, 1.90 1.90, 1.89, 2.00 1.90, 1.99, 2.00, 2.00 1.91, 1.98	1.93, 1.94, 1.92 1.92, 2.00, 1.93	1.94, 1.92, 1.93 1.97	1.93, 1.91 1.97	1.94, 1.96

Figure 18 shows a Q-Q plot of 13,000 line 1 data against a symmetric stable distribution with $\alpha = 1.94$, while Fig. 21 shows the same plot with 10 points on either end excluded. These points were excluded because they were possibly atypical observations, and because they were highly correlated to one another. Again, as in the gaussian Q-Q plots, only the first and last 100 empirical quantiles, as well as every one hundredth between have been plotted, giving the false illusion of discontinuity in the observations. The eye is quite sensitive to deviations from a straight line for quantile-quantile plots; in particular, $\alpha = 1.94$, 1.95, 1.96 could easily be distinguished from one another (Figs. 18 to 23). The data appear to be slightly skewed, so a non-symmetric ($|\beta| \ll 1$, $\beta \neq 0$) stable distribution might indeed provide a better fit to the data. As α increases from 1.94 to 1.96, the stable distribution has shorter and shorter tails, and the points in the tails bending towards the vertical for $\alpha = 1.94$ align with the rest of the data for increasing α.

As a check on these estimates, W. DuMouchel has supplied the authors with a computer program that numerically calculates maximum likelihood estimates of parameters of stable distributions, as well as of their covariances.[46] DuMouchel has shown the maximum likelihood parameter estimates are asymptotically normal, so that some statistical techniques developed for data analysis of gaussian samples can be brought to bear.[70] The method used is numerical, nonetheless, so two possible pitfalls must be kept in mind.

(*i*) For ease in numerical calculations, the data were aggregated into bins, thereby losing information.

(*ii*) A Newton-Raphson-type of algorithm was used that approximates the first and second derivatives of the likelihood function with differences.

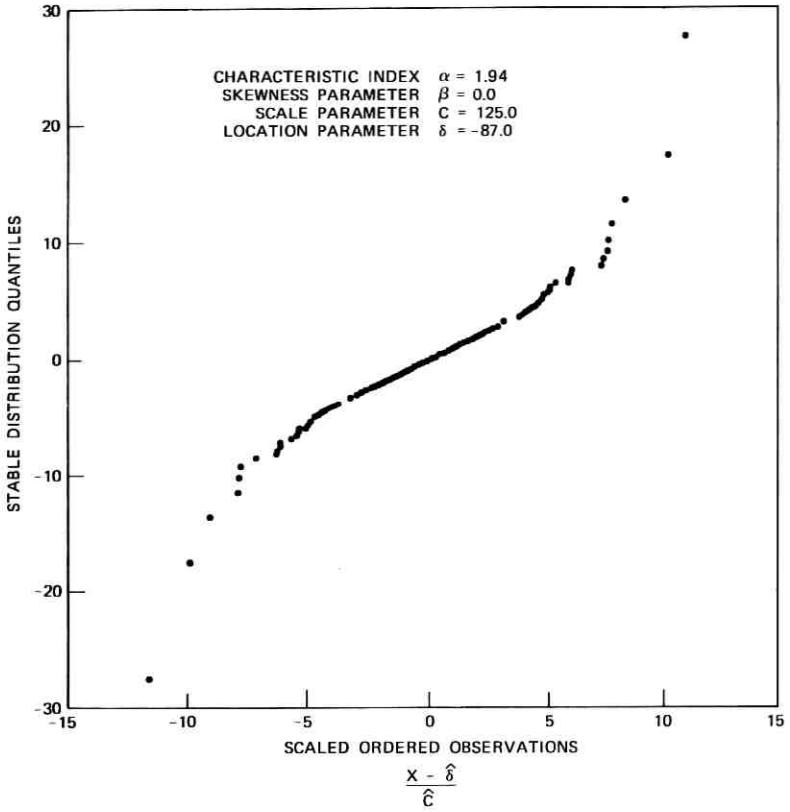


Fig. 18—Line 1 Q-Q plot against a stable distribution, $\alpha = 1.94$ ($N = 13{,}000$).

DuMouchel[71] has observed that the first approximation is the more critical of the two. The second approximation was investigated using a simplex algorithm rather than Newton-Raphson which did not compute discrete approximations to derivatives, with results consistent to those now described.

For line 1 data, with a sample size of 13,000, the numerical maximum likelihood stable distribution parameter estimates were

$$\hat{\alpha} = 1.95$$
$$\hat{\beta} = -0.006$$
$$\hat{c} = 132.7$$
$$\hat{\delta} = -88.8.$$

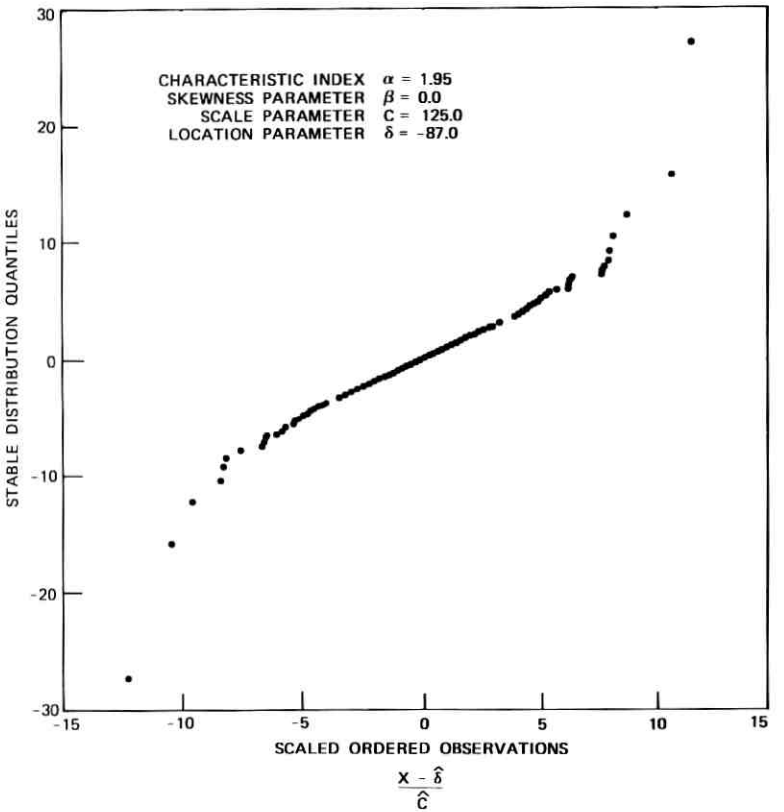


Fig. 19—Line 1 Q-Q plot against a stable distribution, $\alpha = 1.95$ ($N = 13{,}000$).

The numerical approximation to the estimated parameter covariances are shown in Table V.

The large variance of $\hat{\beta}$ compared to the other estimates has been observed by DuMouchel;[46] the cause is unknown. Although Q-Q plots

Table V — Parameter estimate covariances ($\times 10^6$)

	$\hat{\alpha}$	$\hat{c}$	$\hat{\beta}$	$\hat{\delta}$
$\hat{\alpha}$	410	11	−57	−1
$\hat{c}$	11	63	−41	−2
$\hat{\beta}$	−57	−41	13,551	430
$\hat{\delta}$	−1	−2	430	248

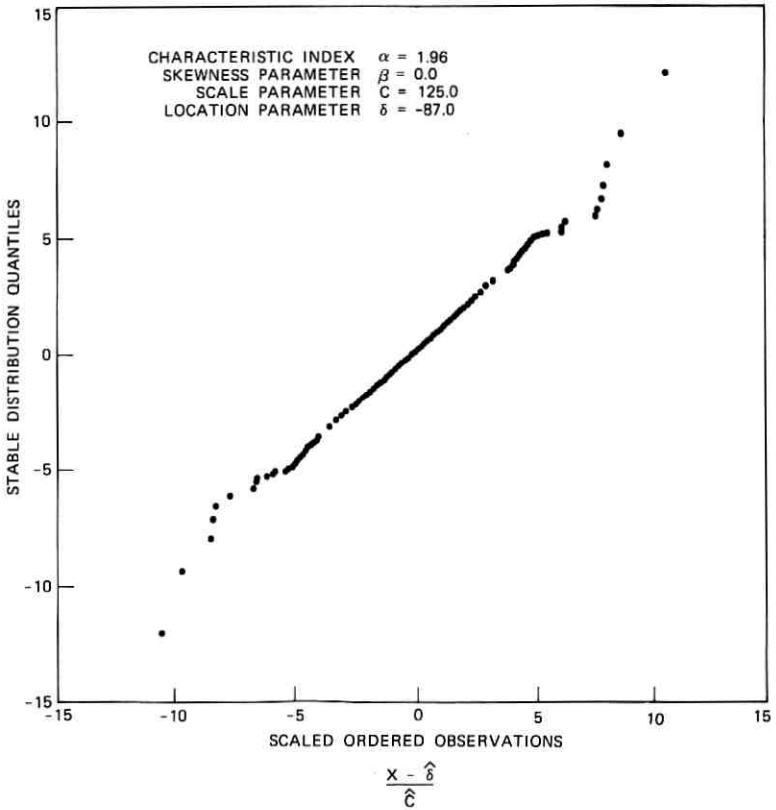


Fig. 20—Line 1 Q-Q plot against a stable distribution, $\alpha = 1.96$ ($N = 13{,}000$).

indicated a slight skewness, i.e., $\beta < 0$ and $|\beta| \ll 1$, the interpretation of the maximum likelihood estimate for β was obscured by this large variance.

As a check on these results, maximum likelihood parameters of stable distributions were estimated for 78,750 filtered data from lines 1 and 2, corresponding to approximately 10 seconds of telephone noise. A Newton-Raphson-type algorithm was used; the parameter estimate covariances were comparable to those just discussed. The results were:

	$\hat{\alpha}$	$\hat{\beta}$	$\hat{c}$	$\hat{\delta}$
Line 1	1.96	−0.0084	218.0	−81.35
Line 2	1.94	−0.0014	93.5	−80.30

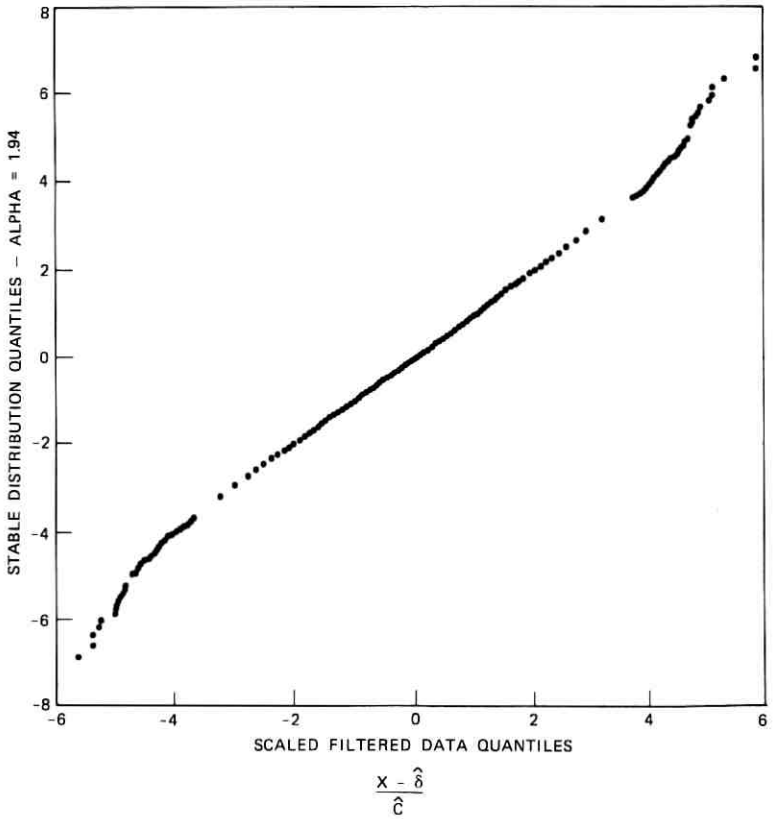


Fig. 21—Center portion of line 1 Q-Q plot against a stable distribution, $\alpha = 1.94$ ($N = 12{,}980$).

Two other pieces of evidence that telephone noise can be fitted by a stable distribution are now presented: the studentized range test and the likelihood ratio test. These have been discussed elsewhere;[47,71] for large amounts of data, caution is necessary to interpret the results of these tests properly. On the other hand, since the data here are apparently close to gaussian, large amounts of data must be examined to make clear the nongaussian nature of the noise. Thus, the results of these tests must be very carefully interpreted, and are included for the sake of completeness.

For line 1 data, testing the gaussian hypothesis at a 0.5-percent significance level via the studentized range test for sample sizes of 1000 led to mixed results: some segments of data fell within these

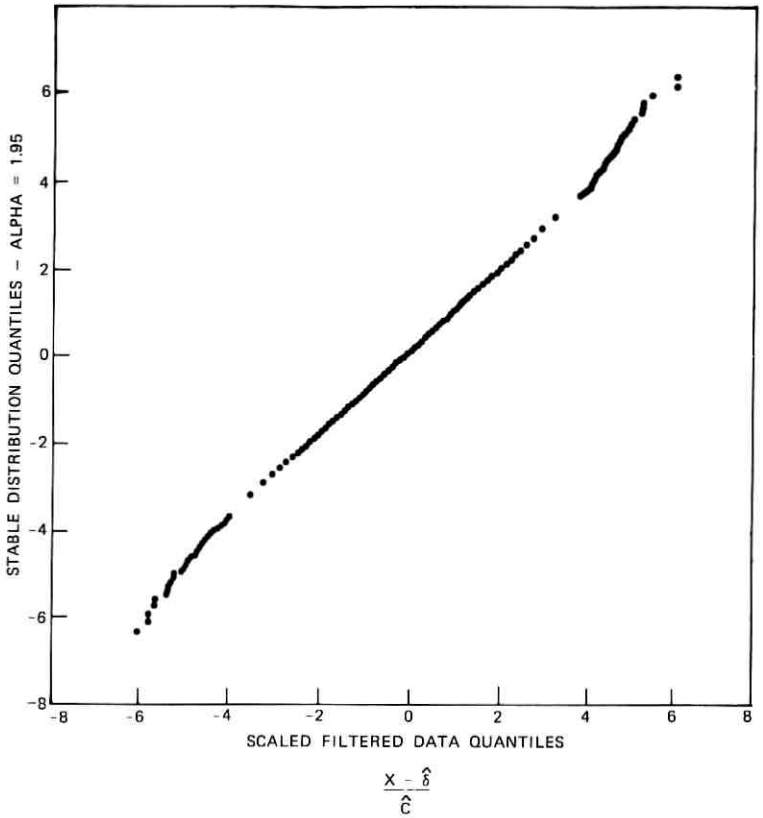


Fig. 22—Center portion of line 1 Q-Q plot against a stable distribution, $\alpha = 1.95$ ($N = 12{,}980$).

limits, others fell outside. However, for a sample size of 10,000 the result of the studentized range test clearly fell outside the confidence intervals.

A likelihood ratio test was used to test 10,000 line 1 data at 1-percent significance levels against five hypotheses. The stable distribution hypothesis was rejected for $\alpha = 2.00, \alpha = 1.98, \alpha = 1.90$, and $\alpha = 1.85$, but could not be rejected for $\alpha = 1.95$.

3.3 Other central-limit-theorem-based models

What other models arise that might adequately account for the data, while having the same central-limit-theorem-based appeal as the stable distributions? First, it is possible the data examined lie in a

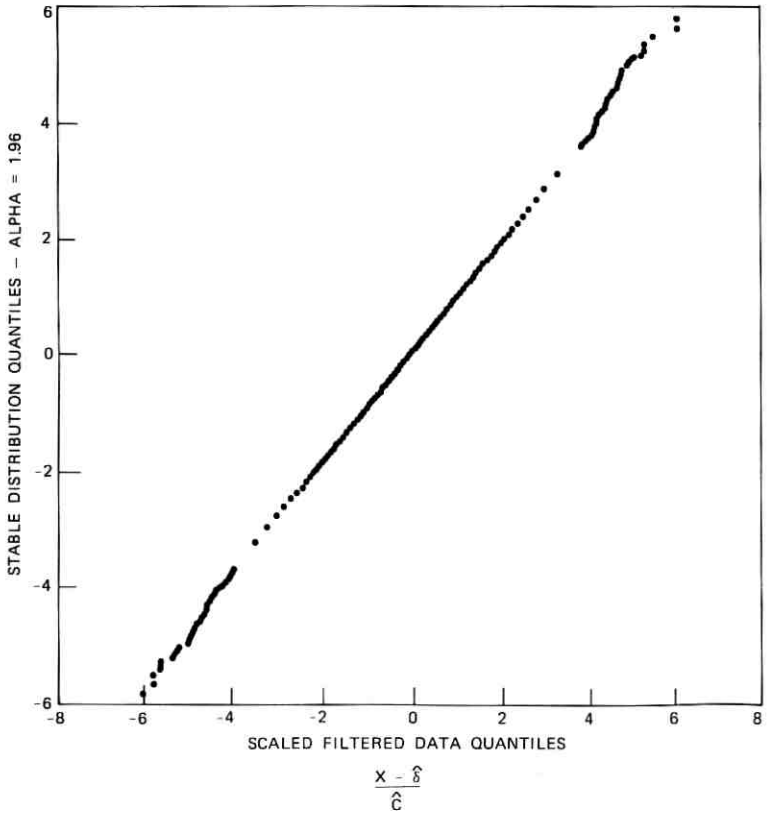


Fig. 23—Center portion of line 1 Q-Q plot against a stable distribution, $\alpha = 1.96$ ($N = 12{,}980$).

domain of partial attraction of a stable distribution (which is wider than the domain of attraction[72]). If enough data were examined, it might be possible that α would approach 2. A second possibility is to model the data as a convolution of r stable distributions, each with its own domain of attraction;[73] presumably, each distribution could be attributed to a separate physical process. A third possibility is that the data are drawn randomly from m gaussian distributions, each with different mean and variance; for example, the data could be drawn from a low-variance gaussian a fraction P of the time, and from a high-variance gaussian a fraction $(1 - P)$ of the time. A fourth possibility is to model the data as a nonstationary gaussian random process, which is a special case of a nonstationary stable random

process, or by a doubly stochastic gaussian random process (where the mean and the variance are themselves random processes), which is a special case of a doubly stochastic stable random process (where all four parameters are themselves random processes). While these nonstationary and doubly stochastic models do not appear to be necessary to adequately model the data discussed here, over longer time intervals, such as days, weeks, months, or years, the simple models might be inadequate while these more complicated models might be more appropriate. Presumably, other classes of models exist.

It is difficult to refute these alternative models offhand. Recall that the original goal was to find a mathematically tractable model for telephone noise; the model discussed here is simple and agrees intuitively with the physics of the noise. The other models are more complicated. To be of practical use, however, they must be so oversimplified that the intuitive agreement with the physics of the noise is lost. It is hoped that the class of models based on stable distributions will lead to more insight into how telephone noise limits voice communication and data transmission and, more important, into new ways for combating this noise.

3.4 Gaussian-plus-filtered-Poisson-process model

A model involving more parameters than the previous one is now investigated. This model assumes that telephone noise is due to a sum of two independent random processes. The low-variance part is assumed to be white and gaussian, while the high-variance process is assumed to be a filtered Poisson process. This type of model was popularized by Snyder,[74] and has been used in optical communication[75,76] and ELF communication[77] to assess theoretically optimum and suboptimum receiver structures. It has intuitive physical appeal: for instance, the low-variance component can be attributed to thermal noise and electromagnetic crosstalk, while the high-variance component can be attributed to switch arcing and thunderstorms. It is convenient to view the filtered Poisson process as the output of a linear dynamical system, whose input is an impulse train; the area of the impulse is

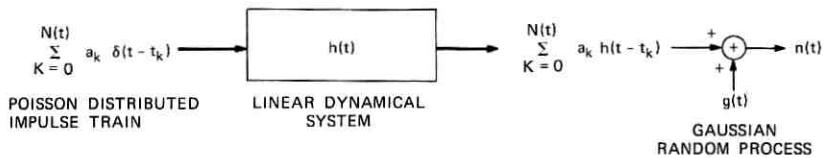


Fig. 24—Block diagram for generating a gaussian-plus-filtered-Poisson process.

assumed to be a random variable, while the instants of time at which the impulses occur follow a Poisson distribution with rate or intensity parameter λ. Figure 24 and the following equation summarize this discussion.

$$n(t) = \begin{cases} g(t) + \sum_{K=1}^{N(t)} a_K h(t - t_K) & N(t) > 0 \\ g(t) & N(t) = 0, \end{cases}$$

where

$n(t)$ = gaussian-plus-filtered-Poisson process,
$g(t)$ = stationary gaussian random process,
a_K = area of Kth impulse,
$h(t)$ = impulse response of linear dynamical system,
t_K = time at which the Kth impulse occurs,
$N(t)$ = number of impulses that occur in $[0, t)$.

To completely describe this model, the following parameters must be estimated

(*i*) The mean and variance of the gaussian random process.
(*ii*) The probability density function for the impulse areas.
(*iii*) The Poisson process rate parameter λ.
(*iv*) The linear system structure.

We recall that the original motivation for this work was to stimulate interest of communication theorists in receiver structures that detect or estimate signals corrupted by nongaussian noise. One advantage of this type model is that parameters can be related to receiver performance limitations as well as to physical causes of noise. This helps in determining how much effort should go into improving the receiver as opposed to reducing the noise (e.g., by designing switches to operate at lower voltages). One disadvantage of this type of model is its great analytical complexity; it may be quite difficult to find analytic performance limitations, and to determine how sensitive these limitations are to model parameters.[75]

If the impulse areas $\{a_K\}$ are assumed to be independent identically distributed random variables that are independent of the times the impulses occur, the characteristic function for the first-order probability density function can be shown to be

$$E[e^{i\omega n(t)}] = \exp\left\{im\omega - \tfrac{1}{2}\sigma^2\omega^2 + \lambda \int_0^t [E_a(e^{i\omega a h(\tau)}) - 1]d\tau\right\},$$

where

ω = frequency,
m = mean of gaussian random process,
σ^2 = variance of gaussian random process,
$E_a(\ )$ = expectation of ( ) with respect to the random variable a, the impulse area.

It is quite difficult to analytically invert $E(e^{i\omega n(t)})$ to find the probability density for $n(t)$. This in turn means maximum likelihood parameter estimates, and Cramér-Rao lower bounds on parameter estimate covariances are difficult to calculate analytically. For this reason, numerical approximations must often be used. To avoid these problems, a suboptimum parameter estimation method was developed: Each parameter of the model was estimated by itself. There is no guarantee that these estimates, when put together, will be close to the true parameter values. The sole reason for doing this was to make the problem tractable. Evidence presented later indicates this method provides an excellent (but perhaps suboptimum) fit to the data.

Although the dynamics of the linear system can be quite complicated, only three simple cases are considered here.

$$(i)\ h(t) = Ae^{-At}u_{-1}(t)$$

$$(ii)\ h(t) = \left(\frac{A^2+\omega^2}{A}\right)e^{-At}\cos\omega t\, u_{-1}(t) \quad u_{-1}(t) = \begin{cases} 1 & t > 0 \\ 0 & t < 0 \end{cases}$$

$$(iii)\ h(t) = \left(\frac{A^2+\omega^2}{\omega}\right)e^{-At}\sin\omega t\, u_{-1}(t),$$

which are perhaps the cases of greatest practical interest.[77]

Assuming the amplitude burst statistics to be independent of the instants of time at which bursts occur, and assuming the gaussian process to be independent of the filtered Poisson process, the mean and variance can be calculated (Table VI) for the steady state noise. $E(a)$ is assumed to be zero in all models presented here. This completes a general discussion of the gaussian-plus-filtered-Poisson-process model; the methods used to estimate the model parameters are now described in detail.

3.5 Gaussian random process parameter estimation

If $E(a) = 0$, then $E[n(t)] = m$, and the sample mean is an unbiased estimate of the true mean of the gaussian process. If the data are

Table VI — Mean and variance for $n(t)$

$h(t)$	$E[n(t)]$	Variance $[n(t)]$
$Ae^{-At}u_{-1}(t)$	$m + \lambda E(a)$	$\sigma^2 + \frac{\lambda A}{2} E(a^2)$
$\frac{A^2+\omega^2}{A} e^{-At} \cos \omega t \, u_{-1}(t)$	$m + \lambda E(a)$	$\sigma^2 + \frac{\lambda}{4}\left(\frac{A^2+\omega^2}{A}\right)^2 \left[\frac{1}{A} + \frac{A}{A^2+\omega^2}\right] E(a^2)$
$\frac{A^2+\omega^2}{\omega} e^{-At} \sin \omega t \, u_{-1}(t)$	$m + \lambda E(a)$	$\sigma^2 + \frac{\lambda}{4}\left(\frac{A^2+\omega^2}{\omega}\right)^2 \left[\frac{1}{A} - \frac{A}{A^2+\omega^2}\right] E(a^2)$

trimmed to exclude a fraction (e.g., 25 percent) of the data with largest absolute deviation from the sample mean, then presumably most values of $n(t)$ that contain large contributions from the filtered Poisson process will be excluded.

The estimates for the mean and variance of the gaussian process were consistent with estimates to be presented later for λ, A, ω, and $E(a^2)$. No bounds are available on the bias or variance of these parameter estimates. The results are summarized in Table VII. The sample variance has been rescaled, based on the assumption that the data were drawn from a truncated gaussian distribution.

3.6 Poisson process parameter estimation

The Poisson process intensity is closely related to the times at which bursts of high-amplitude telephone noise occur. Many definitions of a noise burst are possible. The definition chosen here, although arbitrary, was found to be qualitatively insensitive to the parameters defining a burst. The absolute value of a zero mean waveform is shown

Table VII — Gaussian random process trimmed mean and variance (Total data = 10,000, with a fraction p trimmed from either side)

Line	Truncated Sample Mean			Rescaled Truncated Sample Variance		
	$p = 0.125$	0.250	0.375	$p = 0.125$	0.250	0.375
1	−88.0	−87.6	−87.3	35,414	35,813	35,208
2	−79.5	−78.8	−78.7	17,151	17,058	16,815
3	−77.6	−75.2	−73.6	49,306	48,840	46,554
4	−81.1	−80.9	−81.0	83,076	88,430	90,864
5	− 1.0	− 1.0	− 1.1	1,664	1,672	1,683

in Fig. 25. The duration of the burst is the time difference between the moment that the absolute value of the waveform climbs above an upper threshold, T_{upper}, and the time that the absolute value of the waveform drops below a lower threshold, T_{lower}, provided the waveform stays below the lower threshold for at least a predetermined period of time, called the guard band G_B, which separates one burst from the next. The instant of time a burst occurs, $t_{\max}$, is the first instant at which the absolute value of the burst attains its maximum value, $P_{\max}$.

A large number of statistics can characterize a point process. A number of nonparametric statistics were used to characterize the points in time at which bursts occur, and then, based on this evidence, the burst data were examined in greater detail to see if they could be adequately modeled by a renewal process in general, or a Poisson process in particular (e.g., see Ref. 78).

The two statistics that were first examined were

(*i*) The sample mean time between bursts as a function of T_{upper}, T_{lower}, and G_B.
(*ii*) The empirical cumulative distribution function and the histogram for time intervals between events as a function of T_{upper}, T_{lower}, and G_B.

The effect on these statistics of variations of T_{upper}, T_{lower}, and G_B is now discussed. For line 1 data, for example, T_{lower} was fixed at 600 (roughly three standard deviations from the sample mean), T_{upper} was set at 600, and G_B was varied from 0.1 to 0.9 millisecond, in steps of 0.2 millisecond. T_{upper} was then set at 800, and G_B was varied in an identical manner. Finally, T_{upper} was set at 1000, and G_B was again varied in the same fashion. The number of events observed was found to be insensitive to the choice of guard band as well as to T_{upper}. The guard band was therefore set at 0.5 millisecond, and T_{upper} was set equal to T_{lower} (which also avoids ambiguity in the meaning of threshold).

A typical empirical cumulative distribution function and a histogram for the time intervals between bursts are shown in Figs. 26 and 27, for $T_{\text{upper}} = T_{\text{lower}} = 800$. Typical histograms and empirical cumulative distribution functions for time intervals between bursts for $T_{\text{upper}} = T_{\text{lower}} = 600$ and $T_{\text{upper}} = T_{\text{lower}} = 1000$ had the same shapes as those in Figs. 26 and 27. If the bursts were Poisson-distributed, the distribution function would be completely specified by this information.

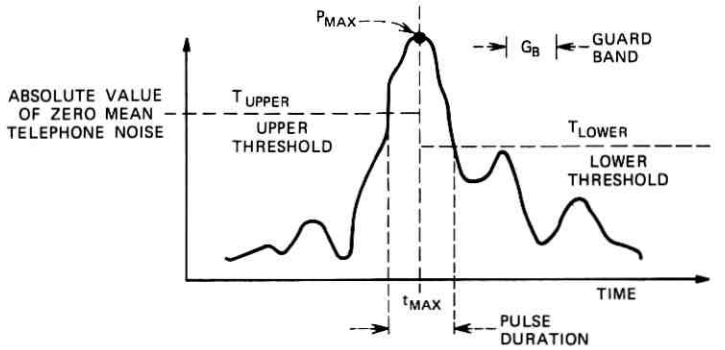


Fig. 25—Definition of burst parameters.

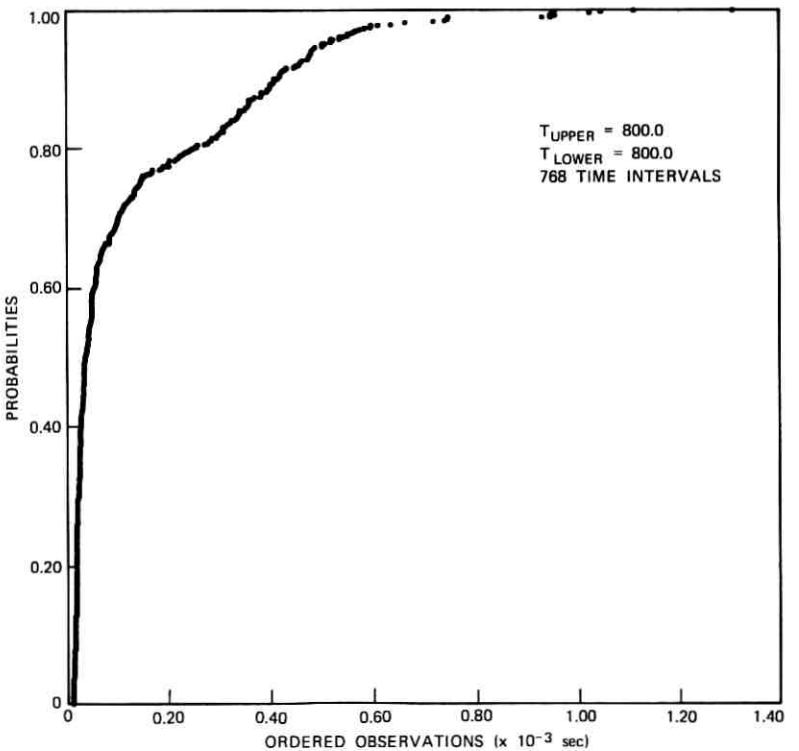


Fig. 26—Empirical cumulative distribution function for line 1 time intervals between bursts ($N = 768$).

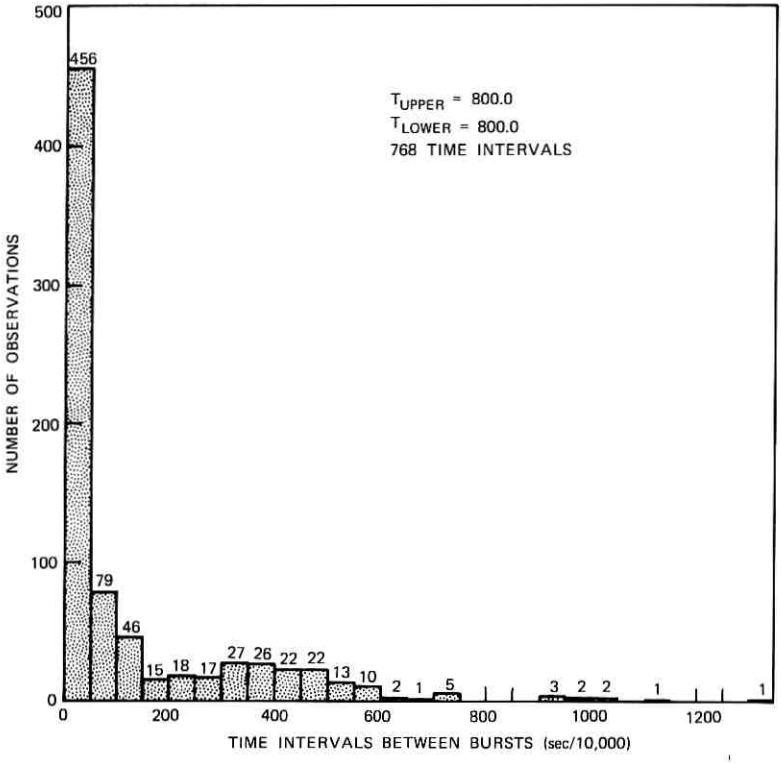


Fig. 27—Histogram for line 1 time intervals between bursts ($N = 768$).

Investigation of the burst statistics of the other two nongaussian lines yielded findings similar to those for line 1.

Next, the second-order statistics of the time intervals were investigated. Figure 28 shows a scatter plot of the $(K + 1)$th interval against the Kth interval. This plot shows that long intervals followed by long intervals are unlikely compared with long intervals followed by short intervals, or short intervals followed by long or short intervals. Note that it is still possible that the intervals are generated by a renewal process with a nonexponential distribution. Note also that approximately half the points plotted fall in the lower left corner square, $0 < \text{time}_K \leqq 100$ and $0 < \text{time}_{K+1} \leqq 100$.

Another set of second-order statistics of interest is

(*i*) The estimated autocorrelation of the time intervals between bursts

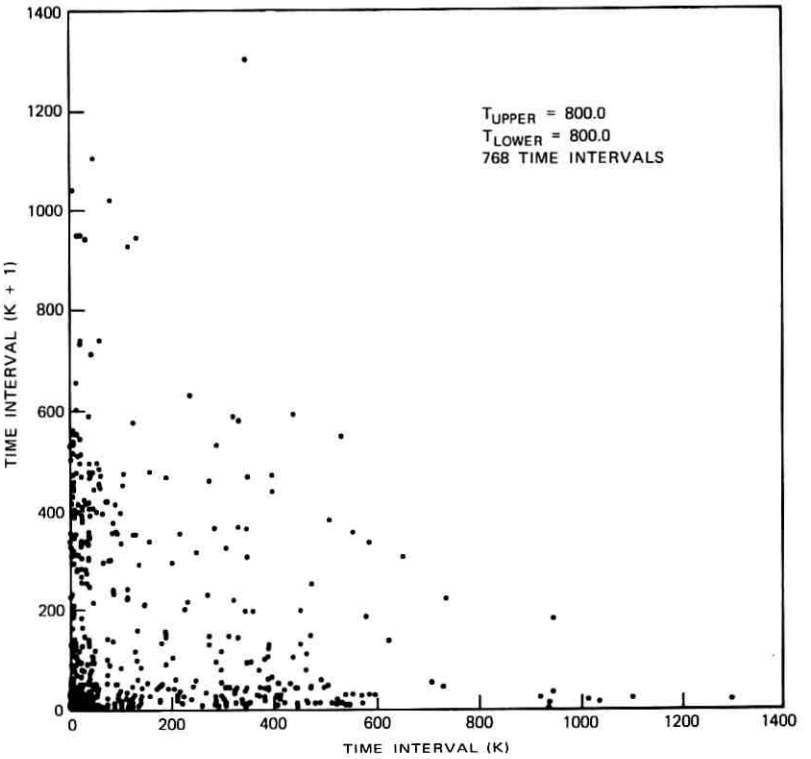


Fig. 28—Scatter plot of time intervals between bursts for line 1 data ($N = 768$).

$$\hat{R}_{tt}(K) = \frac{\frac{1}{N}\sum_{l=1}^{N-K}(t_l - \bar{t})(t_{l+K} - \bar{t})}{\hat{R}_{tt}(0)}, \quad \bar{t} = \frac{1}{N}\sum_{j=1}^{N} t_j,$$

where

$$\hat{R}_{tt}(0) = \frac{1}{N}\sum_{K=1}^{N}(t_K - \bar{t})^2$$

and

t_j = length of jth time interval.

(*ii*) The estimated power spectrum of the time intervals between bursts, which is the Fourier transform of $R_{tt}(K)$ if the process is wide-sense stationary and ergodic.[79] The same issues that were discussed earlier when the noise amplitude power spectrum was estimated are relevant here.

(*iii*) Contingency tables,[80] which correspond to estimates of the second-order joint density of intervals that are a specified number of intervals apart from each other.

(*iv*) The estimate of the conditional expectation of the length of an interval given the length of the interval j intervals earlier

$$E[t_{K+j}|t_K] = \int t_{K+j} dp(t_{K+j}|t_K).$$

The data analysis presented here focuses on the first two of these second-order statistics. Since only a small number of events was observed typically (e.g., 147 for $T_{\text{upper}} = T_{\text{lower}} = 1000$, 768 for $T_{\text{upper}} = T_{\text{lower}} = 800$, for line 1 bursts), statistical fluctuations would have obscured the interpretation of the last two statistics. Figure 29 shows a typical sample autocorrelation function for 1000 intervals,

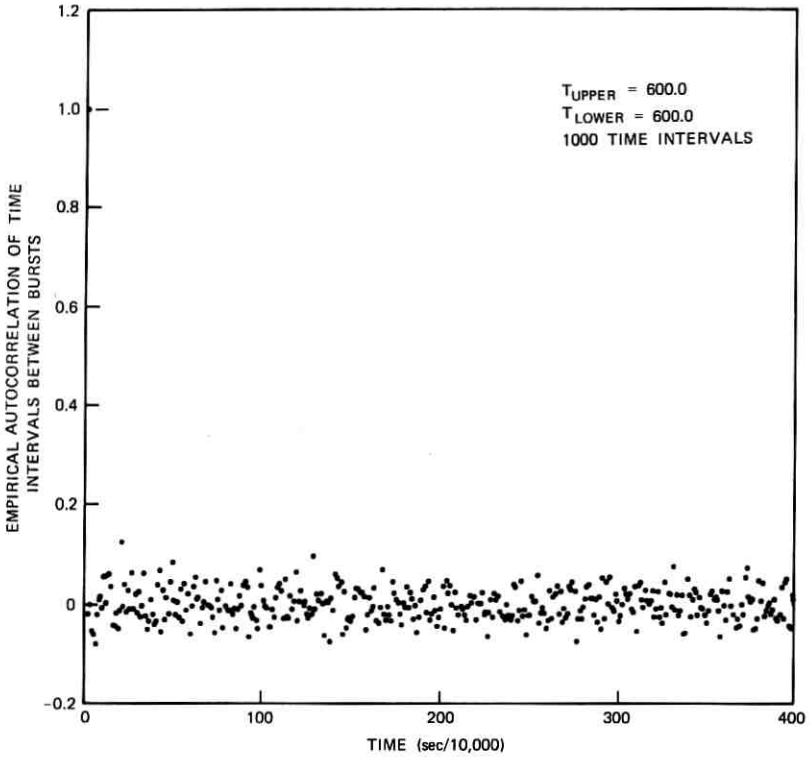


Fig. 29—Autocorrelation function for line 1 time intervals between bursts ($N = 1000$).

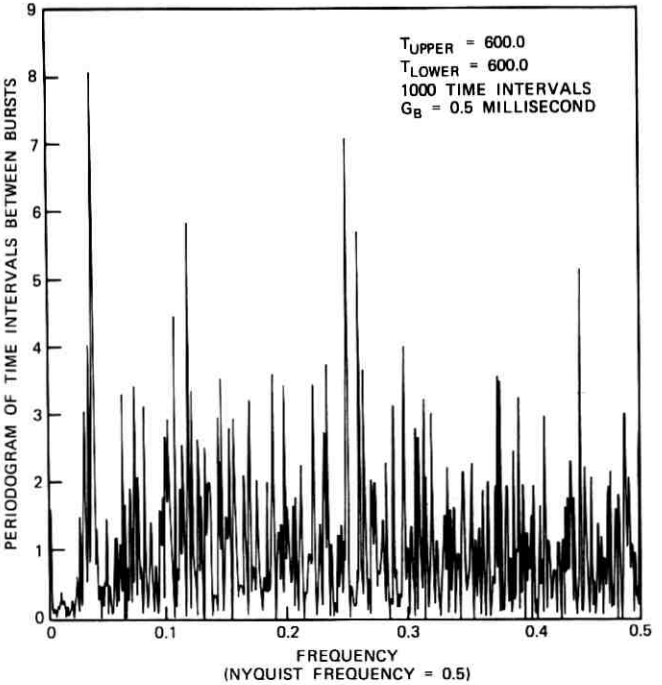


Fig. 30—Periodogram of line 1 time intervals between bursts ($N = 1000$).

while Fig. 30 shows the corresponding periodogram. Based on this evidence, it seems possible the bursts arise from a renewal process. To test this model, the so-called summed empirical periodogram $\tilde{S}(lf_0)$, defined as

$$\tilde{S}(lf_0) = \sum_{K=1}^{l} |\hat{f}(Kf_0)|^2,$$

where $|\hat{f}(Kf_0)|^2$ = periodogram at frequency Kf_0 is plotted in Fig. 31, along with 5 percent significance limits to test the renewal hypothesis, according to which $\tilde{S}$ should be straight. From this and other evidence not presented here, it appears that the bursts analyzed can be reasonably modeled as a renewal process.

Next, it is useful to characterize the renewal process model in greater detail. Such a process can be statistically described by a variety of measures.[81] The two considered here are

(*i*) The survivor function $S(t) \triangleq$ fraction of intervals greater than or equal to t.

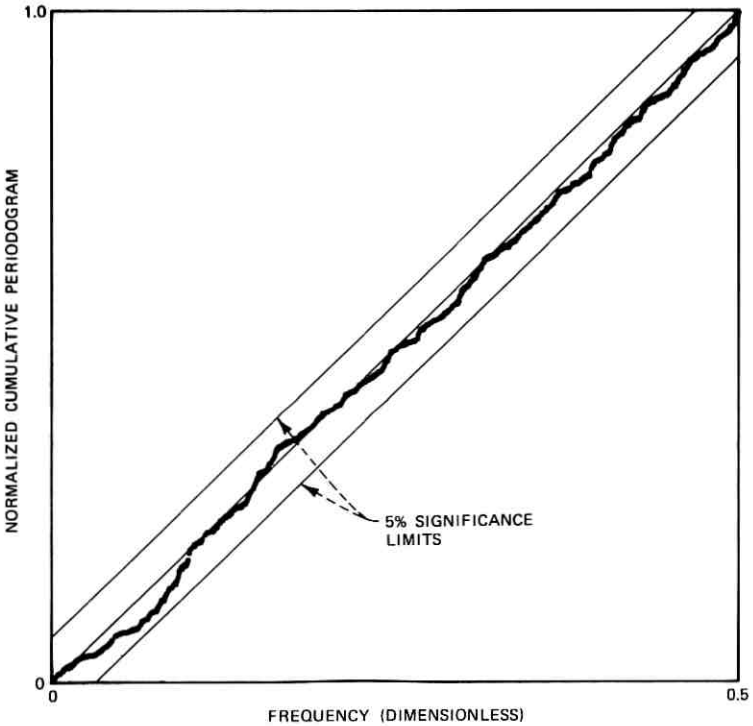


Fig. 31—Normalized cumulative periodogram for line 1 time intervals between bursts, $T_{\text{upper}} = T_{\text{lower}} = 800$ ($N = 768$).

(*ii*) The hazard function $H(t) \triangleq P(t)/S(t)$, where $P(t)$ is the probability density function of the renewal process and $S(t)$ is the survivor function.

Note that $H(t) \cdot \Delta t$ is the probability of an event in an interval of length Δt seconds centered at t, which can be interpreted as the fraction of intervals in the range $(t - \Delta/2, t + \Delta/2)$, given that the last event was t time units ago.

The survivor and hazard functions are related by[82]

$$S(t) = \exp\left[-\int_0^t H(x)dx\right] = \int_t^\infty P(x)dx.$$

In a Poisson process with constant intensity λ, these simplify to

$$\left.\begin{aligned} P(t) &= \lambda e^{-\lambda t} \\ S(t) &= e^{-\lambda t} \\ H(t) &= \lambda \end{aligned}\right\} t > 0.$$

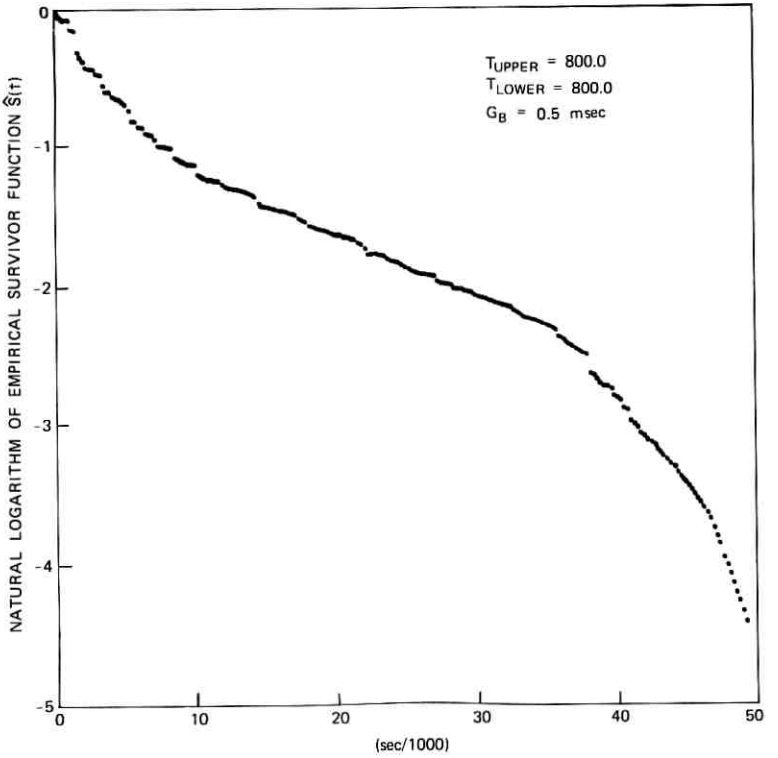


Fig. 32—Natural logarithm of empirical survivor function for line 1 time intervals between bursts ($N = 768$).

In practice, only estimates of the survivor and hazard function are obtained. The empirical survivor function $\hat{S}(t)$ is the fraction of observed intervals greater than or equal to t. The empirical hazard function $\hat{H}(t)$ equals $\hat{P}(t)/\hat{S}(t)$, where $\hat{P}(t)$ is the fraction of observed intervals in the range $[t - \Delta/2,\ t + \Delta/2]$. Figures 32 and 33 show representative empirical survivor and hazard functions for bursts observed on line 1. Statistical fluctuations are quite apparent. For long time intervals, only one or two events fall in any particular bin, giving the appearance of a trend in $\hat{H}(t)$; the survivor function is more stable for long intervals. The log survivor function roughly follows a series of straight-line segments with different slopes, indicating the process can be modeled as a Poisson process whose rate parameter is equal to the absolute value of the slope of the straight line approximations.

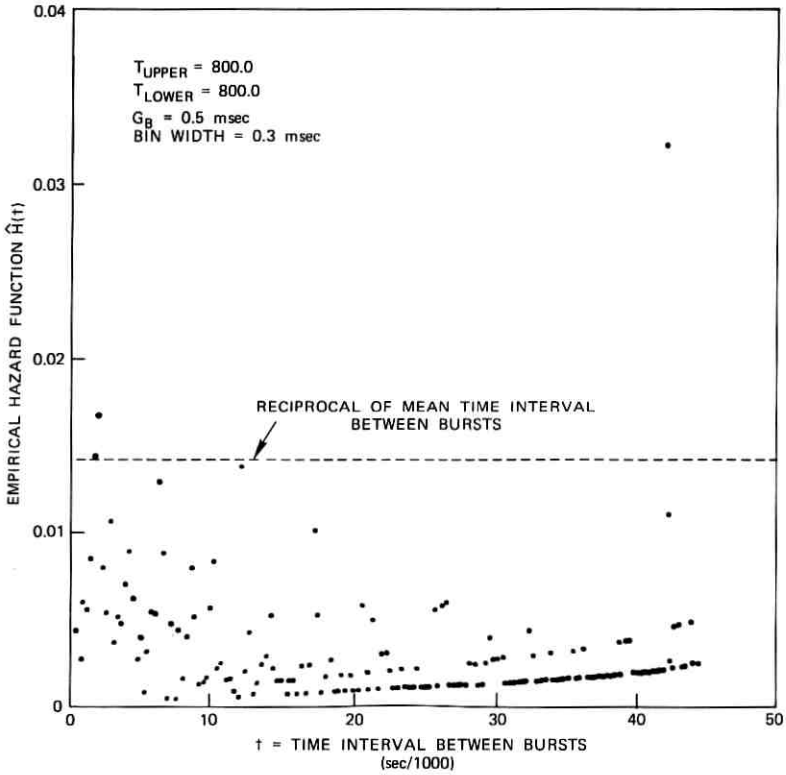


Fig. 33—Empirical hazard function for line 1 time intervals between bursts (N = 768).

In order to adequately model the time intervals between bursts, two models more complicated than a simple Poisson process were investigated. The first model was a pth order autoregressive process, while the second was a doubly stochastic Poisson process, where the Poisson intensity λ was a random variable specified by a two-state Markov process (see Ref. 8). The autoregressive model (for p = 5 and 15) did not adequately account for long time intervals between events. The doubly stochastic Poisson process model did not adequately account for short time intervals followed by long time intervals. Therefore, both these models were dropped in favor of a Poisson process model with constant intensity, even though the log survivor function could not be approximated by a single straight line. Since this is only one of at least six model parameters, it was hoped the overall goodness

of fit of the model would not be seriously degraded; the evidence presented later indicates this might be true.

3.7 Pulse amplitude statistics

It was assumed the burst amplitude and the instant of time at which the burst occurred were independent random variables. Figures 34 and 35 show representative empirical cumulative distribution functions and histograms for line 1 burst amplitude data. Based on these curves, a number of distributions can be fitted to the data; only two will be discussed here, a two-sided log normal and a two-sided power Rayleigh[77] (also known as generalized gamma[83]). A two-sided log normal random variable l equals Rq, where R and q are independent random variables, with q equally likely to be $+1$ or -1 and R defined by a log normal

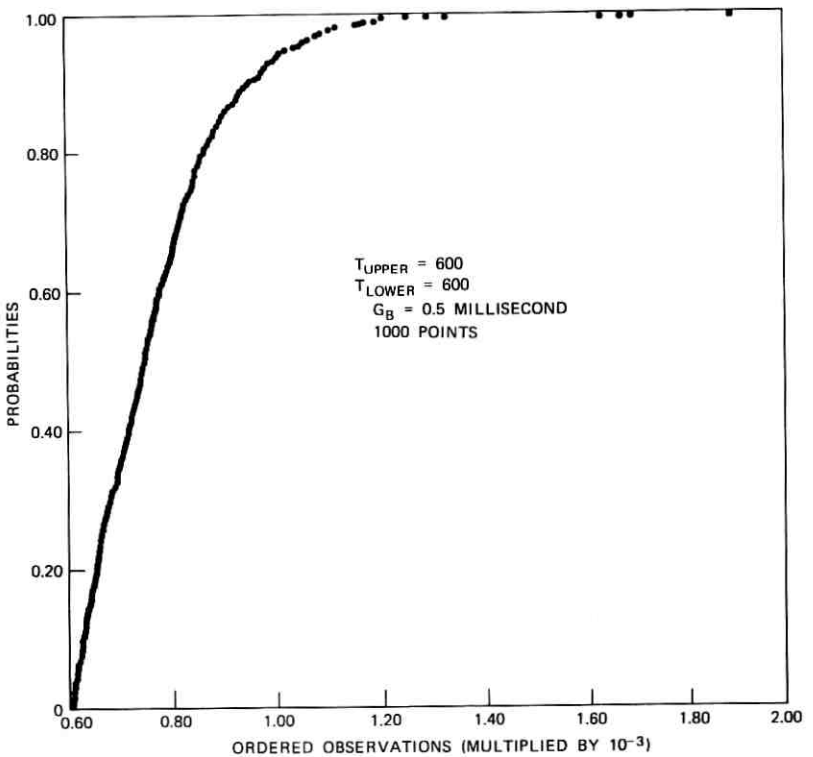


Fig. 34—Empirical cumulative distribution function of maximum burst amplitude (N = 1000).

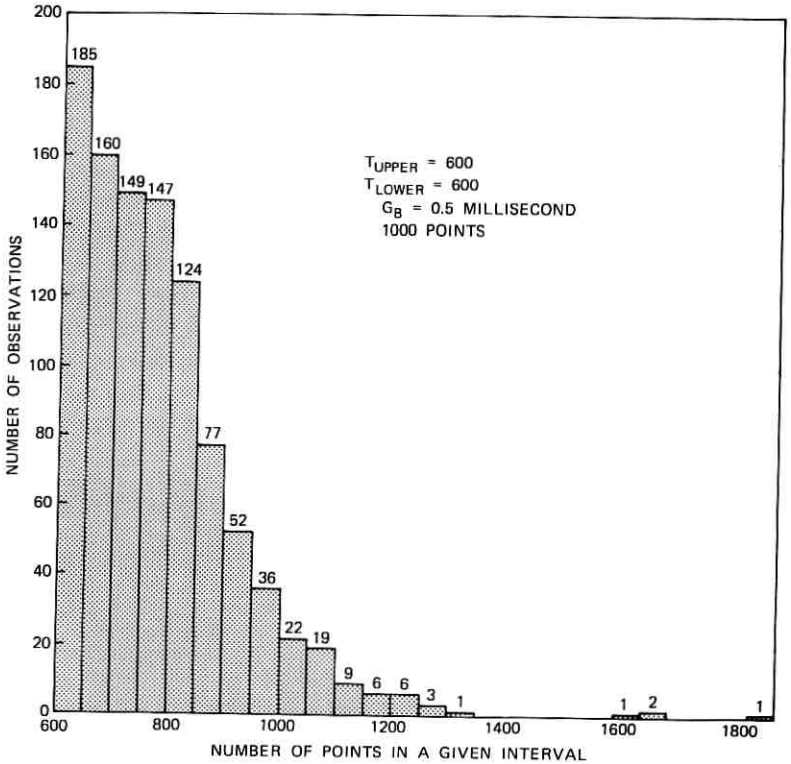


Fig. 35—Histogram of maximum burst amplitudes ($N = 1000$).

density

$$p_R(u) = \frac{1}{\sqrt{2\pi}\sigma u} \exp\left[\frac{-(ln|u| - m)^2}{2\sigma^2}\right] \qquad 0 < u < \infty.$$

Similarly, a two-sided power Rayleigh random variable p equals Rq, where R and q are independent random variables, with q equally likely to be $+1$ or -1, while R is defined by a power Rayleigh density

$$p_R(u) = \left(\frac{K}{R_0}\right)\left(\frac{|u|}{R_0}\right)^{K-1} e^{-(|u|/R_0)^{K-1}} \qquad 0 < K \leqq 2, \quad 0 \leqq |u| < \infty.$$

Each density has zero mean; the log normal variance is $e^{\frac{1}{2}\sigma^2+m}$, while the power Rayleigh variance is $R_0^2\Gamma(1 + 2/K)$.

The parameters of each distribution could be fit to empirical cumulative distribution functions such as that shown in Fig. 34. The range

of parameters for line 1 data, for example, was:

log normal:	mean $5 \leqq m \leqq 6$
	variance $0.01 \leqq \sigma^2 \leqq 0.10$
power Rayleigh:	scale factor $500 \leqq R_0 \leqq 2000$
	exponent $0.50 \leqq K \leqq 0.70$.

The investigation of a much larger amount of data probably would have narrowed the range of these estimates substantially.

3.8 Linear dynamical system parameters

The approach chosen here for estimating linear dynamical system parameters was trial and error. The sample mean burst duration was set equal to the damping constant A^{-1} in both the first- and second-order systems. The second-order system oscillates at frequency ω, which was arbitrarily chosen as $(\frac{1}{2})$ (damping constant)$^{-1}$, to obtain qualitative agreement with actually observed noise bursts (e.g., Fig. 3). For line 1 bursts, for example, $A^{-1} = 0.1$ millisecond for $T_{\text{upper}} = T_{\text{lower}} = 600$, $G_B = 0.5$ millisecond.

3.9 Goodness-of-fit to data of gaussian-plus-filtered-Poisson-process model

Only one test was carried out to provide some heuristic measure of goodness-of-fit of this model to the data. The test was analogous to a quantile-quantile plot. Using typical parameter estimates such as those just described, a computer generated a sample function of a gaussian-plus-filtered-Poisson process. These artificial data were sorted and plotted against actual (sorted) telephone noise from line 1 as shown in Figs. 36 and 37.

The reason for performing just one test is the great difficulty in expressing analytically the distribution function for the gaussian-plus-filtered-Poisson-process model. Hence, it is very difficult to perform quantile-quantile plots of the actual data versus model quantiles, as well as to find maximum likelihood parameter estimates and Cramér-Rao lower bounds on parameter estimate covariances.

3.10 Criticism of the gaussian-plus-filtered-Poisson-process model

Many criticisms of this statistical analysis are possible. First, the question of optimally choosing parameters was never addressed and is still open. Since a large number of parameters must be estimated, a series of presumably suboptimum but easy-to-calculate estimates appeared to be the only feasible course.

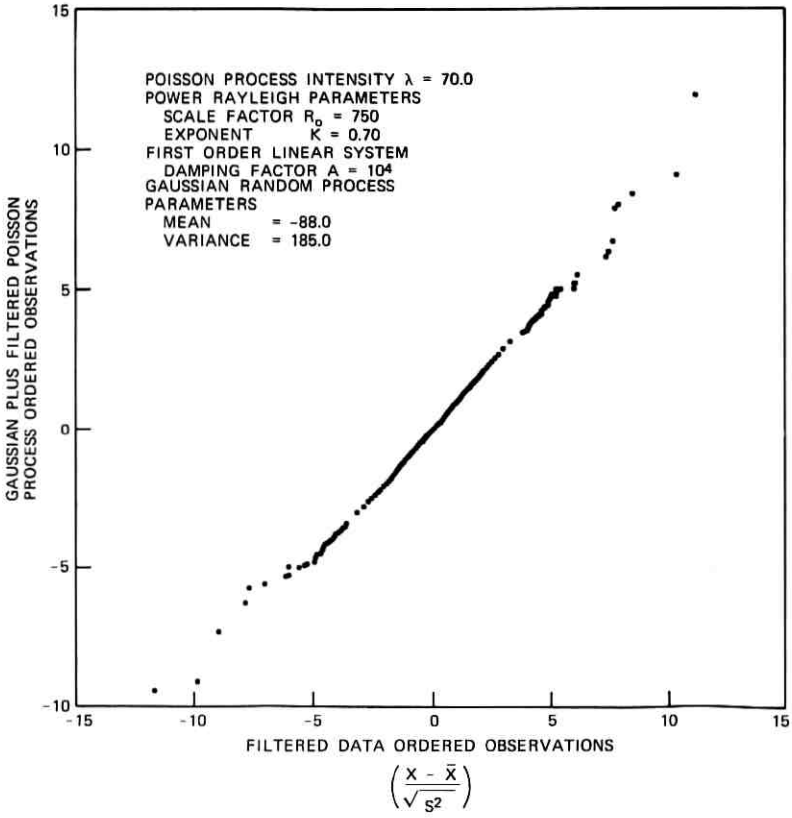


Fig. 36—Line 1 data vs gaussian-plus-filtered-Poisson-process data (N = 13,000) ($\bar{X}$ = sample mean, S^2 = sample variance).

Second, it is not clear how to relate the linear dynamical system parameters in the model to actual telephone system parameters. Where does the filtering occur in reality? Why should it be linear and stationary? Other evidence[19] suggests that the linear dynamical system parameters are not as well defined for other telephone lines as for the data examined in the present work.

Third, the time intervals between bursts are not adequately modeled over the entire observation by a Poisson process. Two other more complicated models were investigated in order to account for this. Many other models can still be investigated.

Fourth, a more general class of models was never investigated that includes the gaussian-plus-filtered-Poisson process as a special case. This model, mentioned briefly earlier, is a mixture of a low-variance gaussian distribution and a high-variance gaussian distribution; during

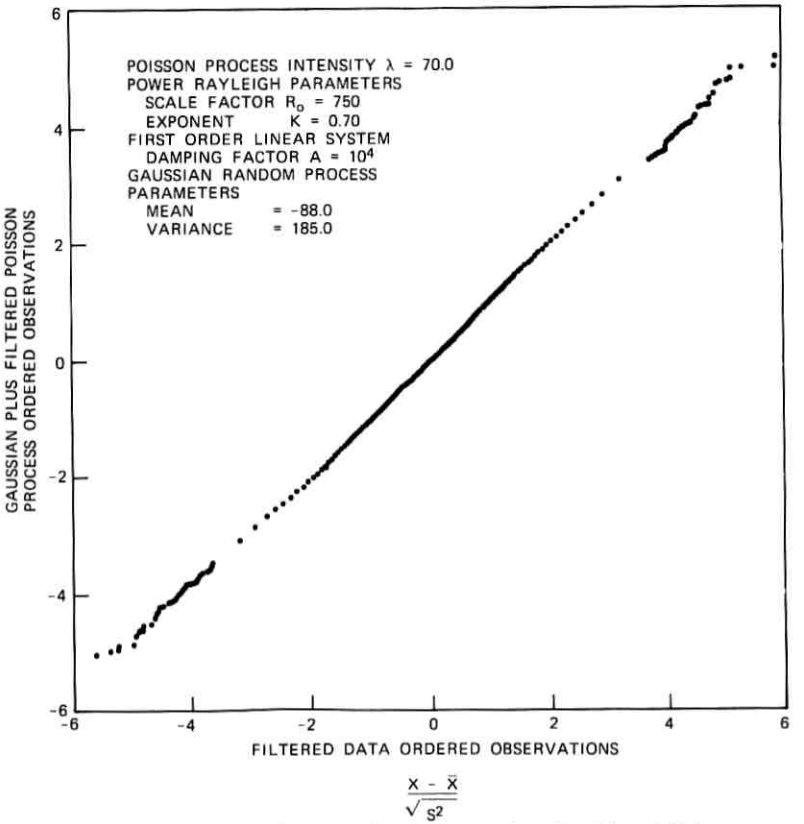


Fig. 37—Center portion of line 1 data vs gaussian-plus-filtered-Poisson-process data (N = 12,980) ($\bar{X}$ = sample mean, S^2 = sample variance).

a fraction P of the time, the low-variance gaussian distribution is chosen to model the data, while during the other $(1 - P)$ fraction of time the high-variance gaussian distribution is chosen. The reasons for not investigating this class of models were that the gaussian-plus-filtered-Poisson-process model comes closer to describing intuitively the physical process of telephone noise generation, and that it has been used by communication theorists in other applications[74–76] more than a mixture of gaussians.

IV. CONCLUSIONS

This study has presented evidence that noise on some lines consists of a deterministic component (e.g., sinusoids at various frequencies) and a purely stochastic component. It was assumed that these com-

ponents add. The data analyzed here suggest that the stochastic component is stationary over short periods of time (typically 1 second) and distinctly nongaussian. Two simple models have been proposed for the nongaussian noise, one based on stable distributions, the other on a mixture of a gaussian process and a nongaussian-filtered Poisson process. Based on the data analyzed here, both models agree intuitively with the physical processes generating telephone noise and appear to fit the data reasonably well.

It is hoped this work will stimulate further research in this area; since only a small number of telephone lines were examined, the models presented here await confirmation by independent investigators. Other models than those discussed here may indeed more adequately account for noise on telephone lines. It is hoped this work will lead to greater insight into how telephone noise limits voice communication and data

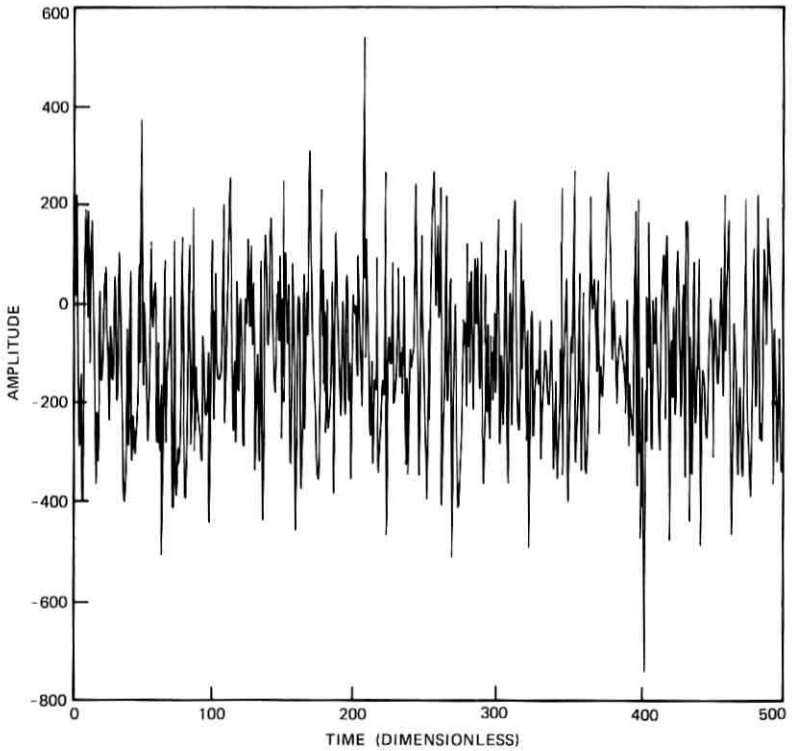


Fig. 38—Computer-generated gaussian-plus-filtered-Poisson-process sample function (same process parameters as in Figs. 36 and 37; $\bar{X}$ = sample mean, S^2 = sample variance).

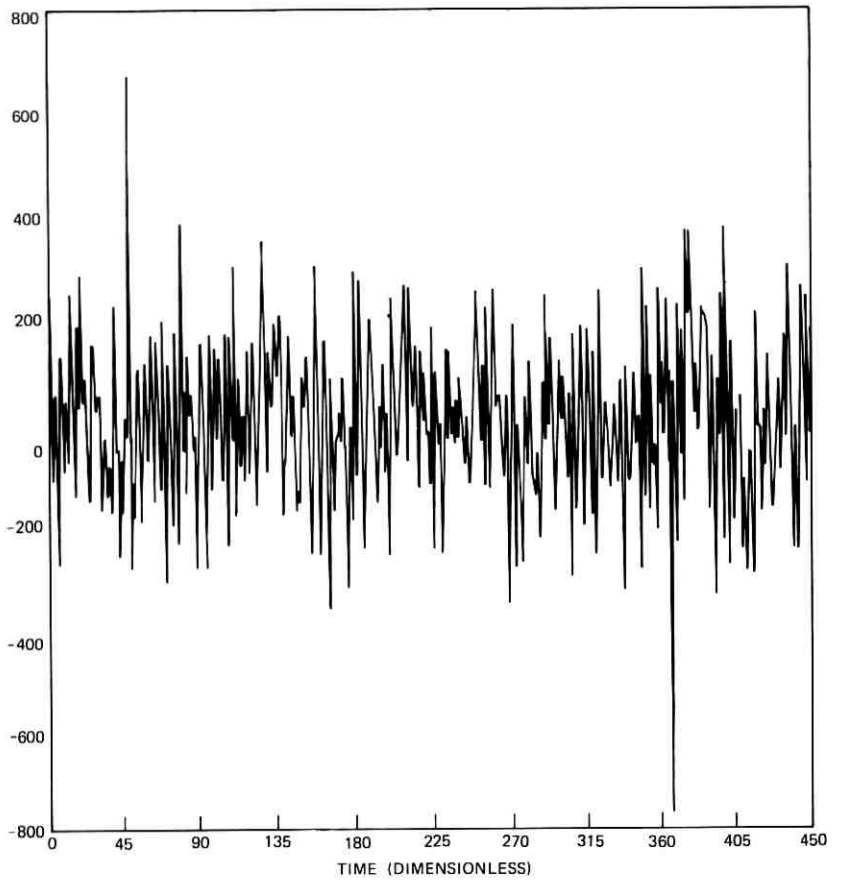
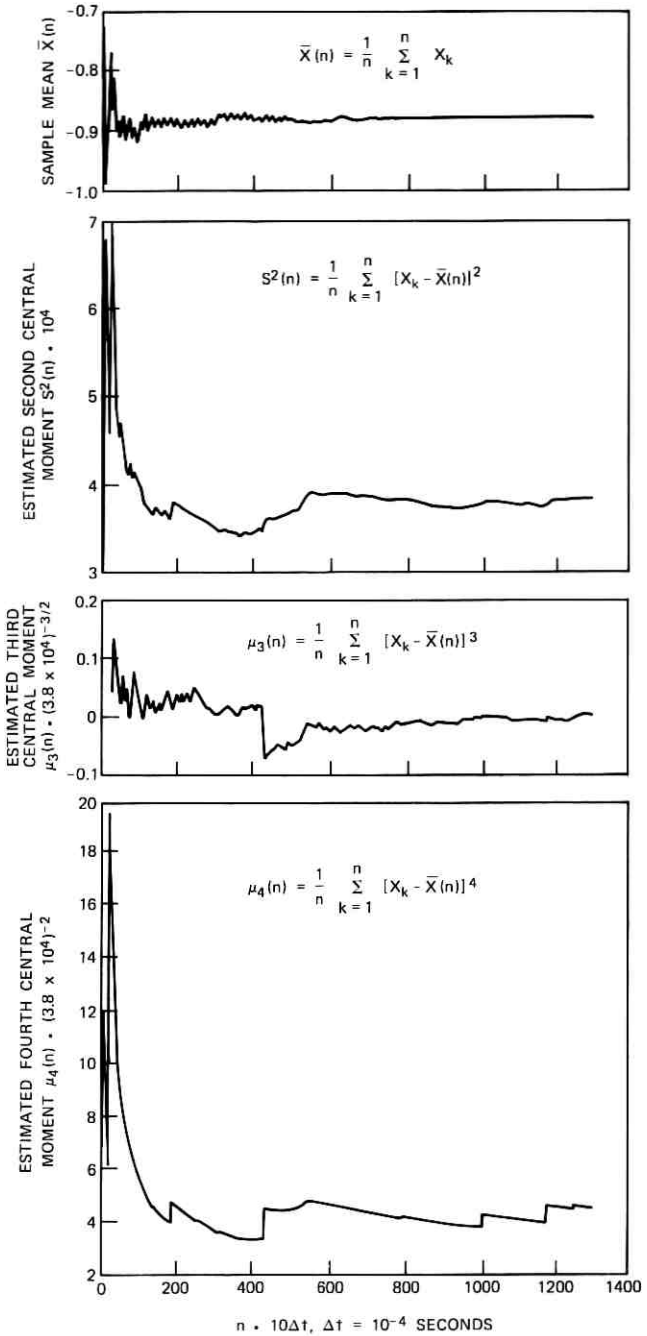


Fig. 39—Computer-generated symmetric stable random process sample function $\alpha = 1.9$ ($c = 100$, $\delta = 0$).

transmission and, more importantly, will lead to new methods for combating this noise.

V. ACKNOWLEDGMENTS

The authors thank J. Fennick for supplying the data, D. Bock for help in the analog-to-digital conversion, L. Rabiner for helpful advice on various aspects of digital filtering, W. DuMouchel for supplying a maximum likelihood stable distribution parameter estimation computer program, M.-J. Cross for accurate calculations of stable distributions used in quantile-quantile plots and elsewhere, J. Chambers for help in calculating and interpreting numerical approximations to


SAMPLE MEAN $\bar{X}(n)$
$\bar{X}(n) = \frac{1}{n}\sum_{k=1}^{n} X_k$
-0.7
-0.8
-0.9
-1.0
ESTIMATED SECOND CENTRAL MOMENT $S^2(n) \cdot 10^4$
$S^2(n) = \frac{1}{n}\sum_{k=1}^{n} [X_k - \bar{X}(n)]^2$
7
6
5
4
3
ESTIMATED THIRD CENTRAL MOMENT $\mu_3(n) \cdot (3.8 \times 10^4)^{-3/2}$
$\mu_3(n) = \frac{1}{n}\sum_{k=1}^{n} [X_k - \bar{X}(n)]^3$
0.2
0.1
0
-0.1
ESTIMATED FOURTH CENTRAL MOMENT $\mu_4(n) \cdot (3.8 \times 10^4)^{-2}$
$\mu_4(n) = \frac{1}{n}\sum_{k=1}^{n} [X_k - \bar{X}(n)]^4$
20
18
16
14
12
10
8
6
4
2
0
200
400
600
800
1000
1200
1400
$n \cdot 10\Delta t$, $\Delta t = 10^{-4}$ SECONDS

maximum likelihood parameter estimates, and J. McKenna and W. M. Boyce for constant encouragement throughout the course of this work. C. Mallows, W. M. Boyce, and D. Slepian had many helpful suggestions for improving this paper. Finally, Miss J. Langoski has done an excellent job in typing and retyping this manuscript.

APPENDIX A

This appendix is included to give the reader some feeling for the two models discussed here. Using typical parameter estimates and a first-order system, a computer generated and plotted a discrete time sample function of a gaussian-plus-filtered-Poisson process (Fig. 38). For comparison, Fig. 39 shows a computer-generated discrete time series, where each point was drawn independently from a symmetric stable distribution (characteristic index $\alpha = 1.9$).

APPENDIX B

A graphic indication that the data can be better modeled by a nongaussian rather than gaussian distribution is now presented. The motivation for this work is found in Mandelbrot.[84]

Estimates for the sample mean, as well as second, third, and fourth central moments can be calculated recursively for larger and larger amounts of data. Figure 40 plots these estimates for 13,000 filtered data (only every tenth estimate is plotted). Note the tendency for the second, third, and fourth central moment estimates to wander rather than stabilize as more and more data are included, as evidenced by the jumps in the estimates. The sample mean, however, does stabilize; note the small ripple in this estimate, which presumably is due to a sinusoid at approximately 600 Hz that was not filtered from the data (see Table II).

These results are qualitatively consistent with results for central moment estimates of computer-generated stable random variables (10,000 independent identically distributed samples, $\alpha = 1.96$, $\beta = 0$). Central moment estimates of computer-generated gaussian random variables (10,000 independent identically distributed samples) did stabilize at the correct values, while exhibiting no apparent jumps such as in Fig. 40. These results are also consistent with the gaussian-plus-filtered-Poisson-process model, with the large jumps in the estimates presumably a result of the filtered Poisson process.

Fig. 40—Sample mean as well as second, third, and fourth central moment estimates for line 1 ($N = 13{,}000$).

REFERENCES

1. J. H. Fennick, "Amplitude Distributions of Telephone Channel Noise and a Model for Impulse Noise," B.S.T.J., *48*, No. 10 (December 1969), pp. 3243–3263.
2. P. Mertz, "Impulse Noise and Error Performance in Data Transmission," Rand Memo RM-4526-PR, DDC AD 614 416, April 1965.
3. H. L. Yudkin, "Some Results in the Measurement of Impulse Noise on Several Telephone Circuits," Proc. Nat. Elec. Conf., Chicago, Ill., *16*, pp. 222–231, 10–12 October 1960.
4. A. Fontaine and R. G. Gallagher, "Error Statistics and Coding for Binary Transmission Over Telephone Circuits," Proc. IRE, *49*, 1961, pp. 1059–1065.
5. E. J. Hofmann, "Error Statistics Utilizing the Code Translation Data-System Over Various Media," M.I.T. Lincoln Laboratory Report 25G-0026, 28 March 1961.
6. E. J. Hofmann, "Error Statistics Utilizing the Milgo Data System Over Various Media," M.I.T. Lincoln Laboratory Report 25G-0025, 23 January 1961.
7. E. J. Hofmann, "Measured Error Distributions on the Bell A1 Facility Over Various Media," Proc. Nat. Elec. Conf., Chicago, Ill., 1960, pp. 37–44.
8. E. N. Gilbert, "Capacity of a Burst-Noise Channel," B.S.T.J., *39*, No. 5 (September 1960), pp. 1253–1265.
9. J. M. Berger and B. Mandelbrot, "A New Model for Error Clustering in Telephone Circuits," IBM J. Research and Dev., *7*, 1963, pp. 224–236.
10. E. O. Elliott, "A Model of the Switched Telephone [Data] Network for Data Communication," B.S.T.J., *44*, No. 1 (January 1965), pp. 89–109.
11. B. D. Frichtman, "A Binary Channel Characterization Using Partitioned Markov Chains," IEEE Trans. Info. Thy., *18*, 1967, pp. 221–227.
12. S. M. Sussman, "Analysis of the Bursts Model for Error Statistics in Telephone Circuits," IEEE Trans. Comm. Sys., *11*, 1963, pp. 213–221.
13. P. A. W. Lewis and D. R. Cox, "A Statistical Analysis of Telephone Circuit Error Data," IEEE Trans. Comm. Tech., *14*, 1966, pp. 382–389.
14. A. A. Alexander, R. M. Gryb, and D. W. Nast, "Capabilities of the Telephone Network for Data Transmission," B.S.T.J., *39*, No. 3 (May 1960), pp. 431–476.
15. H. C. Fleming and R. M. Hutchinson, Jr., "Low-Speed Data Transmission Performance on the Switched Telecommunications Network," B.S.T.J., *50*, No. 4 (April 1971), pp. 1385–1405.
16. M. D. Balkovic, H. W. Klancer, S. W. Klare, and W. G. McGruther, "High-Speed Voiceband Data Transmission Performance on the Switched Telecommunications Network," B.S.T.J., *50*, No. 4 (April 1971), pp. 1349–1384.
17. H. O. Burton and D. D. Sullivan, "Errors and Error Control," Proc. IEEE, *60*, 1972, pp. 1293–1301.
18. A. M. Curtis, "Contact Phenomena in Telephone Switching Circuits," B.S.T.J., *19*, No. 1 (January 1940), pp. 40–62.
19. M. Kurland and D. A. Molony, "Observations on the Effects of Pulse Noise In Digital Data Transmission Systems," IEEE Trans. Comm. Tech., *15*, 1967, pp. 552–556.
20. J. Fennick, private communication.
21. J. E. Evans, "Preliminary Analysis of ELF Noise," M.I.T. Lincoln Laboratory Technical Note 1969-18, DDC AD 691 814, 26 March 1969.
22. A. Ephremides and L. Brandenburg, "On the Reconstruction Error of Sampled Data Estimates," IEEE Trans. on Info. Thy., *IT-19*, 1973, pp. 365–367.
23. J. E. Evans, D. K. Willim, and J. R. Brown, "Description of the Lincoln Laboratory Wideband ELF Noise Recording Systems," M.I.T. Lincoln Laboratory, TN 1972-19, DDC AD 743 006.
24. J. Cooley, P. Lewis, and P. Welch, "The Fast Fourier Transform Algorithm and Its Applications," IBM Research Paper RC-1743, 9 February 1967.
25. J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement Procedures*, New York: Wiley-Interscience, 1971.
26. J. W. Cooley, P. A. W. Lewis, and P. D. Welch, "The Application of the Fast Fourier Transform Algorithm to the Estimation of Spectra and Cross-Spectra," J. Sound Vib., *12*, 1970, pp. 339–352.

27. G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, San Francisco: Holden Day, 1970.
28. R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra*, New York: Dover, 1958.
29. Ref. 21, pp. 6–7.
30. J. L. Doob, *Stochastic Processes*, New York: Wiley, 1953, p. 519.
31. H. Cramér, *Mathematical Methods of Statistics*, Princeton: Princeton University Press, 1946.
32. M. B. Wilk and R. Gnanadesikan, "Probability Plotting Methods for the Analysis of Data," Biometrika, *55*, 1968, pp. 1–17.
33. E. S. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians*, Cambridge: Cambridge University Press, 1966, pp. 207–208.
34. B. V. Gnedenko and A. N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*, rev. ed., Reading, Mass.: Addison-Wesley, 1968, Chs. 3, 4, 6, 7.
35. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II, New York: Wiley, 1966, Chs. 6, 9, 17.
36. B. Mandelbrot, "How Long is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension," Science, *156*, 1967, pp. 636–638.
37. B. Mandelbrot, "Long-Run Linearity, Locally Gaussian Processes, H-Spectra, and Infinite Variances," Int. Econ. Rev., *10*, 1969, pp. 82–111.
38. B. Mandelbrot, "Sporadic Random Functions and Conditional Spectral Analysis: Self Similar Examples and Limits," *Fifth Berkeley Symp. Math. Stat. Prob.*, Vol. III, Berkeley and Los Angeles: University of California Press, 1967, pp. 155–179.
39. B. Mandelbrot, "The Variation of Certain Speculative Prices," J. Business, *36*, 1963, pp. 394–419.
40. B. Mandelbrot, "The Pareto-Levy Law and the Distribution of Income," Int. Econ. Rev., *1*, 1960, pp. 79–106.
41. B. Mandelbrot, "The Stable Paretian Income Distribution When the Apparent Exponent is Near Two," Int. Econ. Rev., *4*, 1963, pp. 111–114.
42. B. Mandelbrot, "The Variation of Some Other Speculative Prices," J. Bus. Univ. Chicago, *40*, 1967, pp. 393–413.
43. E. Fama, "Behavior of Stock Market Prices," J. Bus. U. Chicago, *38*, 1965, pp. 34–105.
44. D. A. Darling, "The Influence of the Maximum Term in the Addition of Independent Random Variables," Trans. Am. Math. Soc., *73*, 1952, pp. 95–107.
45. M. Loève, *Probability Theory*, New York: Van Nostrand Reinhold, 1960, pp. 288–318.
46. W. H. DuMouchel, "Stable Distributions in Statistical Inference," Yale Ph.D. dissertation, 1971.
47. E. F. Fama and R. Roll, "Properties of Symmetric Stable Distributions," J.A.S.A., *63*, 1968, pp. 817–836.
48. E. F. Fama and R. Roll, "Parameter Estimates for Symmetric Stable Distributions," J.A.S.A., *66*, 1971, pp. 331–338.
49. Ref. 34, p. 162.
50. Ref. 34, p. 197.
51. Ref. 34, p. 164.
52. M.-J. Cross, "Tables of Finite Mean Nonsymmetric Stable Distributions as Computed from their Convergent and Asymptotic Series," to appear in J. Stat. Comp. Simulation.
53. Ref. 34, p. 171.
54. Ref. 34, p. 183.
55. Ref. 35, p. 169.
56. Ref. 35, p. 459.
57. J. D. Mason, "Convolution of Stable Laws as Limit Distributions of Partial Sums," Ann. Math. Stat., *41*, 1970, pp. 101–114.
58. V. M. Zolotarev and V. S. Korolyak, "On a Hypothesis Proposed by B. V. Gnedenko," Thy. Prob. Appl., *6*, 1961, pp. 431–435.
59. P. Billingsley, "Limit Theorems for Randomly Selected Partial Sums," Ann. Math. Stat., *33*, 1962, pp. 85–92.

60. A. E. Berry, "The Accuracy of the Gaussian Approximation to the Sum of Independent Variates," Trans. Am. Math. Soc., *49*, 1941, pp. 122–136.
61. R. G. Esseen, "Fourier Analysis of Distribution Functions. A Mathematical Study of the Laplace-Gaussian Laws," Acta. Math., *77*, 1945, pp. 1–125.
62. H. Cramér, "On Asymptotic Expansions for Sums of Independent Random Variables with a Limiting Stable Distribution," Sankhyā, Series A, 1963, pp. 13–24.
63. W. DuMouchel, "Stable Distributions in Statistical Inference: 1. Symmetric Stable Distributions Compared to Other Symmetric Long-Tailed Distributions," J.A.S.A., *68*, 1973, pp. 469–477.
64. R. M. Dudley, "Speeds of Metric Probability Convergence," Z. Wahrschein., *22*, 1972, pp. 323–332.
65. C. C. Heyde, "On Large Deviation Probabilities in the Case of Attraction to a Non-Normal Stable Law," Sankhyā, Series A, *30*, 1968, pp. 253–258.
66. V. Boonyasombut and J. M. Shapiro, "The Accuracy of Infinitely Divisible Approximations to Sums of Random Variables With Application to Stable Laws," Ann. Math. Stat., *41*, 1970, pp. 237–250.
67. T. A. Hern, "Error Estimates for Weak Convergence to Certain Infinitely Divisible Laws," Ann. Math. Stat., *43*, 1972, pp. 1592–1602.
68. M. Lipschultz, "On Strong Bounds for Sums of Independent Random Variables Which Tend to a Stable Distribution," Trans. Am. Math. Soc., *81*, 1956, pp. 135–154.
69. W. Doeblin, "Sur l'Ensemble de Puissances d'une Loi de Probabilité," Studia Math., *9*, 1940, pp. 71–96.
70. W. H. DuMouchel, "On the Asymptotic Normality of the Maximum Likelihood Estimate when Sampling from a Stable Distribution," Ann. Stat., *1*, 1973, pp. 948–957.
71. Ref. 46, p. 87.
72. Ref. 34, p. 189.
73. H. G. Tucker, "Convolutions of Distributions Attracted to Stable Laws," Ann. Math. Stat., *39*, 1968, pp. 1381–1390.
74. D. L. Snyder, "Optimal Detection of Known Signals in a Non-Gaussian Noise Resembling VLF Atmospheric Noise," 1968 Wescon Convention Record, Part 4.
75. E. V. Hoversten, R. O. Harger, and S. J. Halme, "Communication Theory for the Turbulent Atmosphere," Proc. IEEE, *58*, 1970, pp. 1626–1650.
76. S. D. Personick, "Statistics of a General Class of Avalanche Detectors with Applications to Optical Communication," B.S.T.J., *50*, No. 10 (December 1971), pp. 3075–3095.
77. J. W. Modestino, "A Model for ELF Noise," M.I.T. Lincoln Laboratory Tech. Rpt. 493, DDC AD 737 368, 16 December 1971.
78. D. R. Cox and P. A. W. Lewis, *The Statistical Analysis of Series of Events*, London: Methuen, 1966.
79. Ref. 78, p. 71.
80. Ref. 78, p. 176.
81. Ref. 78, pp. 134–171.
82. Ref. 78, p. 135.
83. N. L. Johnson and S. Kotz, *Continuous Univariate Distribution I*, Boston: Houghton Mifflin, 1970, p. 197.
84. Ref. 39, p. 396.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 7, September 1974
Printed in U.S.A.

Low-Noise, Integrated, Millimeter-Wave Receiver

By B. GLANCE and W. W. SNELL, JR.

(Manuscript received December 7, 1973)

A low-noise, hybrid-integrated, millimeter-wave receiver that consists of a local oscillator and a downconverter on a silica substrate is described in this paper. The source for the local oscillator is a Gunn diode, and the mixer element is a beam-leaded Schottky barrier diode. A novel filter circuit is used to combine the local oscillator and the signal with low insertion loss in the signal path. The single-sideband noise figure of the receiver at 30 GHz is 5.5 dB, including 0.8-dB contribution of the if amplifier, and the rms FM frequency variation resulting from the local oscillator is 168 Hz/$\sqrt{KHz}$.

I. INTRODUCTION

Large numbers of solid-state, low-noise receivers that can be built into compact low-cost radio transmission systems are needed for future terrestrial and satellite communication systems. This paper describes a low-noise microstrip receiver with a single-sideband noise figure of 5.5 dB at 30 GHz, including 0.8 dB resulting from the if amplifier. The receiver consists of a Gunn diode, a Schottky barrier device, and a microstrip conductor pattern on a fused-quartz substrate. The circuit is fabricated using thin-film photolithographic processing techniques. It can be readily scaled to frequencies up to 100 GHz.

II. RECEIVER DESCRIPTION

The microstrip circuit pattern consists of a 2-μm-thick evaporated chromium-gold film on a silica substrate, as shown in Fig. 1. It includes an input rf filter, a pump filter, a low-pass if filter, and a Gunn oscillator with a biasing circuit. The pump and rf signal filters are made from a single rectangular resonator that supports two orthogonally polarized microstrip modes.[1] The resonance for the pump

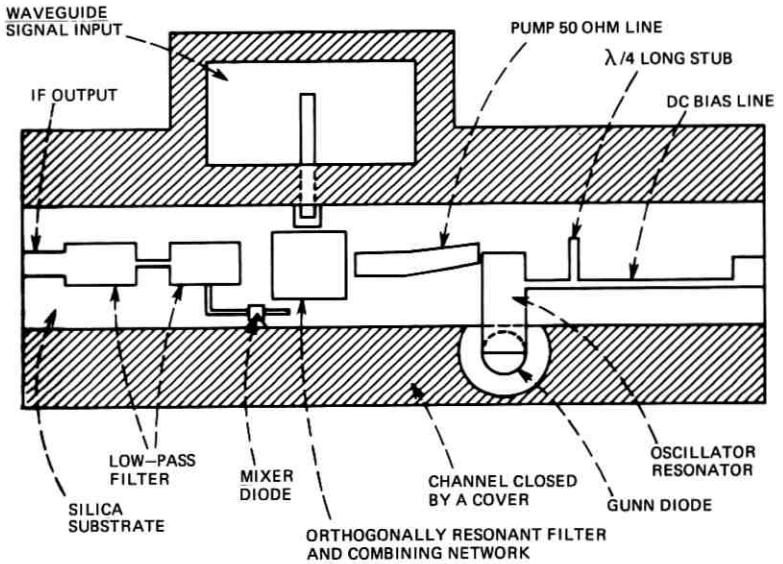


Fig. 1—Schematic drawing of the microstrip receiver circuit in a rectangular channel with a length of 25.4 mm, a width of 4.0 mm, and a depth of 2.5 mm. The silica substrate thickness is 0.34 mm.

signal is directed along the channel, with the resonance for the input signal in the perpendicular direction. Each orthogonal resonance has the response of a single-pole filter. Isolation between the signal and pump ports is >20 dB, which is equal to or better than the isolation normally achieved in a balanced downconverter design. A single beam-leaded Schottky barrier diode (see Table I) is used for downconverting the rf signal to the if frequency. The diode is shunt-mounted to a

Table I — Gunn diode and mixer diode parameters

Diodes	Parameters	
Gunn diode Microwave Associates Model No. MA49173	Frequency Maximum output power dc bias voltage dc bias current	28–30 GHz 50 mW 5 volts 800 mA
Mixer diode Hewlett-Packard Model SO82-2716	Junction resistance Series resistance Junction capacitance Package inductance Package capacitance	120 ohms 10 ohms 0.07 pF 0.1 nH 0.07 pF

high-impedance microstrip line. The length of this line was chosen so that the reactance seen by the diode at the image frequency is optimized to obtain a low noise figure.

A Gunn diode connected to a microstrip resonator provides the pump power at 28.4 GHz. The diode is inserted into the side wall of the rectangular channel that supports the substrate (Fig. 1). The microstrip resonator is coupled capacitively to a 50-ohm line that feeds the pump signal into the orthogonally resonant filter. Direct-current bias to the Gunn diode is provided through a high-impedance line connected to the microstrip resonator at an rf minimum. Radio-frequency leakage is further minimized in the bias line by means of a $\lambda/4$-long microstrip stub. The metallized substrate is mounted in a closed rectangular channel to eliminate radiation losses and parasitic coupling to external circuits. Waveguide modes inside the channel are below cut-off for the frequency range of the receiver. The Gunn diode (see Table I), encapsulated in a ceramic pill package, gives up to 50 mW at 30 GHz. The oscillator frequency can be tuned from 27 to 29 GHz. The tuning mechanism consists of a small piece of dielectric material attached to a nylon screw that is inserted in the channel above the microstrip resonator. The adjustable dielectric loads the fringe field of the resonator and provides a means for tuning the local oscillator frequency. Figure 2 shows the 30-GHz microstrip receiver and Table I lists the Gunn diode and the mixer diode parameters.

III. NOISE FIGURE MEASUREMENT

The overall mixer noise figure of a receiver is[2]

$$F_o = L_c(N_r - 1 + F_{i-f}), \tag{1}$$

where L_c is the downconverter conversion loss, N_r is the mixer diode noise ratio,[2] and F_{i-f} is the noise figure of the if preamplifier.

The parameter N_r can be obtained by measuring the system noise figure at constant diode current, with two if preamplifiers with different noise figures $(F_{i-f})_1$ and $(F_{i-f})_2$. If $(F_o)_1$ and $(F_o)_2$ are the two measured system noise figures, the parameter N_r is given by

$$N_r = \frac{(F_o)_2(F_{i-f})_1 - (F_s)_1(F_{i-f})_2}{(F_o)_1 - (F_o)_2} + 1. \tag{2}$$

The system noise figure was measured by using a wideband preamplifier with a noise figure of 4.5 dB yielding a total noise figure of 9.2 dB. Measurements were also made with a low-noise parametric amplifier with a noise figure of 0.8 dB, giving in this case a system noise figure of

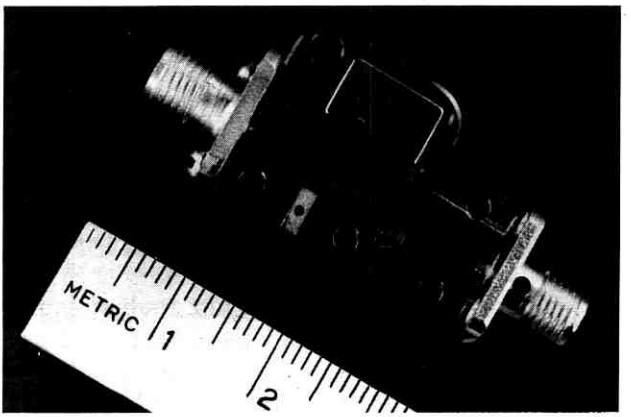


Fig. 2—The receiver. A probe that extends into an adjacent RG-96/U waveguide is used to couple to the signal.

5.5 dB. Substituting these values into eq. (2) gives a figure of merit equal to 1, as expected from a good mixer diode.[2] Equation (1) then becomes

$$F_o = L_c \times F_{i-f}, \tag{3}$$

yielding a downconverter noise figure equal to the mixer conversion loss. Thus,

$$L_c = 4.7 \text{ dB}. \tag{4}$$

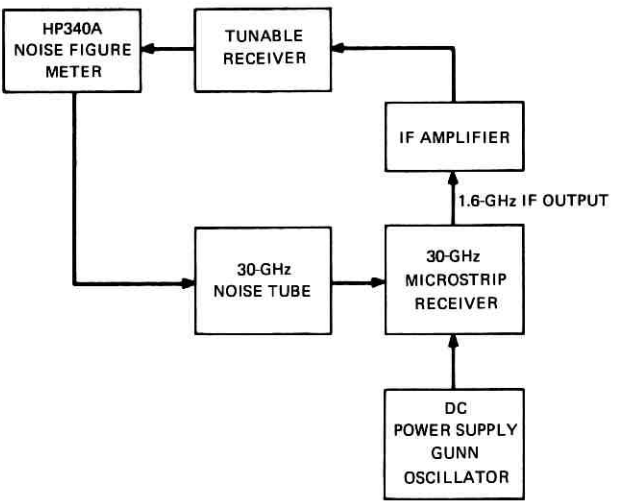


Fig. 3—Noise figure measurement setup.

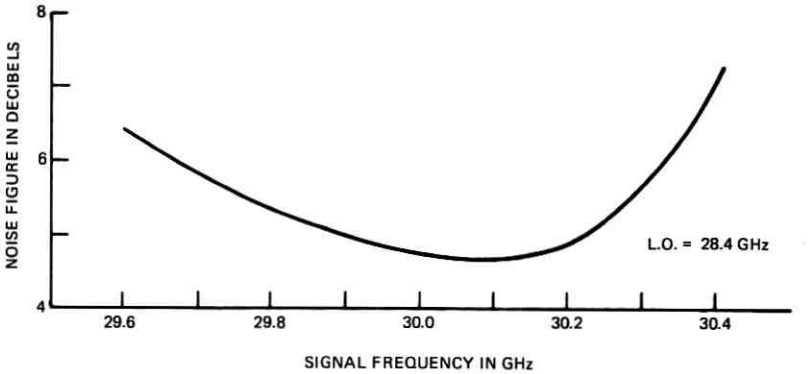


Fig. 4—Measured noise figure vs frequency of the rf receiver, not including the if amplifier contribution.

Figure 3 is a schematic diagram of the noise measuring setup. The signal frequency is 30 GHz, the pump frequency is 28.4 GHz, the if frequency is 1.6 GHz, and the bandwidth is 460 MHz wide at the −1 dB points. Figure 4 shows the single-sideband noise figure of the downconverter, not including the if contribution.

IV. FM NOISE MEASUREMENT

FM noise resulting from the local oscillator is an important parameter in downconverters designed for receiving angle modulated signals, since the FM noise of the local oscillator is directly added to the phase information contained in the downconverted if signal.

Measurement of the FM noise was made by beating the pump frequency with a 30-GHz low-noise signal. Figure 5 shows the power spectral density measured at 1.64 GHz with a 100-Hz wide filter and a scanning time of 2 seconds per division. The Lorentzian shape of the observed power spectrum is characteristic of an oscillator with white gaussian noise. Theoretical analysis for this type of oscillator shows that the spectral density around the carrier is mainly due to FM noise and varies as[3–5]

$$S_{FM}(f) = \frac{\omega_o^2 kT}{2Q^2} \times \frac{1}{(\omega_o^2 kT/4Q^2P_o)^2 + (\omega - \omega_o)^2}, \tag{5}$$

where

ω_o = oscillator frequency
P_o = output power

and

Q = oscillator external loaded Q.

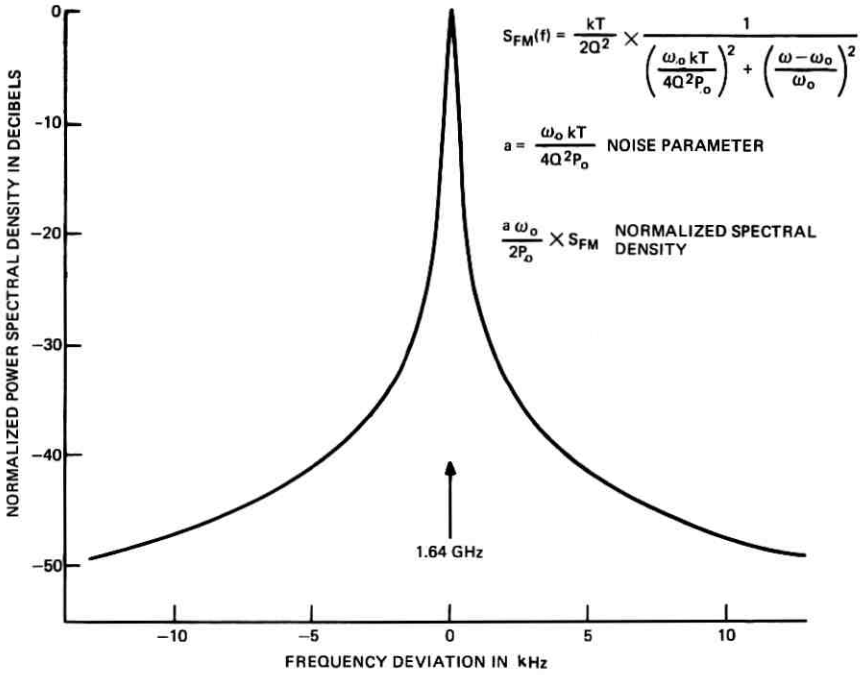


Fig. 5—Measured power spectral density of the local oscillator signal downconverted at the if frequency.

Equation (5) normalized to unity varies as

$$y = \frac{a^2}{a^2 + x^2}, \tag{6}$$

where a is the noise parameter given by

$$a = \frac{\omega_o kT}{4Q^2P_o} \tag{7}$$

and

$$x = \frac{\omega - \omega_o}{\omega_o} \tag{8}$$

$$y = S(f)_{FM} \times \frac{a\omega_o}{2P_o}. \tag{9}$$

The noise parameter a is obtained from the measured power spectrum (shown in Fig. 5) by using eq. (6). Figure 6 shows the values determined for the parameter a for various frequency deviations. The results

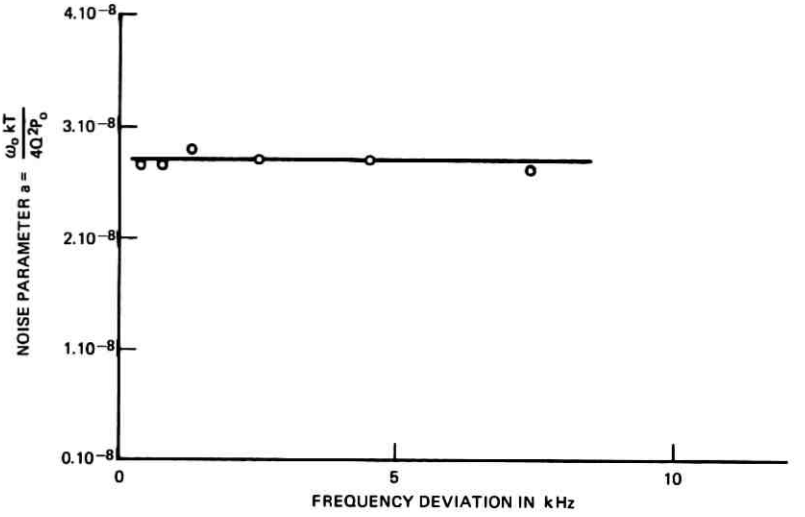


Fig. 6—Calculated values of the noise parameter $a = \omega_o kT/4Q^2P_o$ vs frequency deviation.

are distributed, within the error margin of our measurements, around a constant value of

$$\langle a \rangle = 2.7 \times 10^{-8}. \tag{10}$$

This is a good experimental verification that the noise in the signal is in fact mainly FM in character as predicted by eq. 5.

The rms FM frequency deviation per $\sqrt{\text{Hz}}$ is given by[4]

$$\Delta\omega_{\text{rms}} = \frac{\omega_o}{Q}\sqrt{\frac{kT}{P_o}}, \tag{11}$$

which can be written from eq. (7) as

$$\Delta\omega_{\text{rms}} = 2\omega_o\sqrt{\frac{a}{\omega_o}}. \tag{12}$$

Substituting eq. (10) into eq. (12), the rms FM frequency deviation becomes

$$\Delta f_{\text{rms}} = 168\,\frac{\text{Hz}}{\sqrt{\text{KHz}}}. \tag{13}$$

The receiver described in this paper was designed for use in a PCM system in which the phase of the carrier will be modulated between four levels with a sampling period $T = 7.3$ ns, which will occupy a

bandwidth $B = 200$ MHz. The rms phase deviation resulting from the local oscillator, which occurs in the time interval T, is

$$\Delta\phi_{\text{rms}} = 2\pi\Delta f_{\text{rms}} T\sqrt{B}. \tag{14}$$

The rms phase fluctuation calculated from eq. (13) equals

$$\Delta\phi_{\text{rms}} = 0.2°. \tag{15}$$

Equation (15) shows that the noise due to the local oscillator is an insignificant contribution to the modulated signal.

V. CONCLUSION

A low-noise hybrid integrated receiver that is comparable to the best conventional waveguide circuits[6] has been built and tested at 30 GHz. A unique orthogonal resonator input circuit allows the use of a single mixer diode resulting in a performance comparable to conventional balanced downconverter receivers. The receiver circuit, made on a single silica substrate, can be mass-produced with high reliability. It can be readily scaled to higher frequencies and is especially suited for incorporation into millimeter-wave communication systems.

VI. ACKNOWLEDGMENTS

The authors wish to thank M. V. Schneider for his helpful comments and many discussions. The expert assistance of E. Zack and K. Schwarz in building advanced circuits is also gratefully acknowledged.

REFERENCES

1. W. W. Snell, "Microstrip Downconverter from 30 GHz to 1.5 GHz," unpublished work.
2. H. A. Watson, ed., *Microwave Semiconductor Devices and Their Circuit Applications*, New York: McGraw-Hill, 1969, pp. 387, 389.
3. W. A. Edson, "Noise in Oscillators," Proc. IRE, *48* (August 1960), pp. 1454–1466.
4. K. Kurokawa, "Noise in Synchronized Oscillators," IEEE Trans. on Microwave Theory Tech., *MTT-16*, No. 4 (April 1968), pp. 234–240.
5. J. Josenhans, "Noise Spectra of Read Diode and Gunn Oscillators," Proc. IEEE, *54*, No. 10 (October 1966), pp. 1478–1479.
6. M. Viggh, "Low Noise Receivers for Millimeter Wave Signals," Proc. 1971 European Microwave Conference, Vol. 1, pages A11/1:1, A11/1:4.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 7, September 1974
Printed in U.S.A.

Rain Attenuation on Short Radio Paths: Theory, Experiment, and Design

By W. F. BODTMANN and C. L. RUTHROFF

(Manuscript received January 9, 1974)

The outage time on short radio paths at frequencies above 10 GHz can be estimated from distributions of point rain rates derived from U. S. Weather Service rain gauge charts. In this paper, a previous theory is extended to include the effects of path length and frequency. Experimental corroboration is presented for three locations on the east coast of the United States, and the design of short radio paths is illustrated by examples. One-minute rain rate distributions for 20 locations are also included.

I. INTRODUCTION

Many characteristics of radio systems above 10 GHz are due to the existence of large attenuation by rain. These characteristics have been discussed at length in a previous publication.[1] The economical design of a system operating at these frequencies requires a knowledge of the incidence of rain attenuation at the location of the system; to be more specific, it requires a procedure for predicting the number of minutes per year that a radio path will be out of service because of excessive rain attenuation. This paper describes such a design procedure and includes data required to design systems in several sections of the country; design examples are also given.

The starting point for this design procedure is the theory of attenuation by uniform rainfall[2] and a relationship, derived in Ref. 3, between the attenuation distribution measured on a path and a point rain rate distribution measured near the path. The key result of Ref. 3 is a connection between the theory of attenuation of uniform rain and that of the variable rainfall experienced in practical situations. This connection is, of course, vital to any practical measurement because uniform rain does not usually occur. A conclusion of this work was that the measuring interval used in rain rate measurements has a large

effect on the accuracy of the predicted attenuation distribution. A definition of rain rate suited to the prediction of attenuation distributions was provided, and it was shown that, if the rain rate distribution were measured in the proper way, the attenuation distribution could indeed be estimated from it.

The importance and usefulness of the theoretical result lies in the existence of point rain rate data from a large number of rain gauges all over the country.[4] Copies of charts from many of these recording gauges can be obtained from the U. S. Weather Service.* However, these charts do not record rain rates directly, and initially it was far from obvious that appropriate rain rate data could be obtained from them at all. Thus, the transformation from an interesting theory to practical engineering application came when a method was devised for measuring accurate 1-minute rain rate distributions from these rain gauge charts.

This paper begins with a brief review of the salient results in Ref. 3, and extends that theory to include the effects of frequency and path length. Since a theory of design is only as good as its results, predictions of attenuation distributions computed from rain rate distributions are compared with attenuation distributions measured on several paths, and in all cases the agreement is excellent. The paper concludes with examples of the design of short radio paths in several sections of the country.

II. REVIEW OF THEORY[3]

A theory of attenuation of uniform rainfall has been formulated by Ryde and Ryde and others, and a good account is given by Medhurst.[5] The attenuation resulting from uniform rain has been computed by Ryde and Ryde,[2] Medhurst,[5] and Setzer;[6] Setzer's results, for several frequencies of interest, are shown in Fig. 1.

Any relationship between attenuation and rain rate must take into account the fact that attenuation depends upon the density of rain in the path and is therefore a volume function, whereas rain rate is a surface vector, the magnitude of which is the rate of water flowing through a surface area.

For uniform rainfall, the relationship between rain rate and rain density on the path is simple and not at all dependent upon the time interval used in the rain rate measurement. However, uniform rainfall

* U. S. Department of Commerce, National Oceanic and Atmospheric Administration, Environmental Data Service, National Climatic Center, Federal Building, Asheville, N. C. 28801.

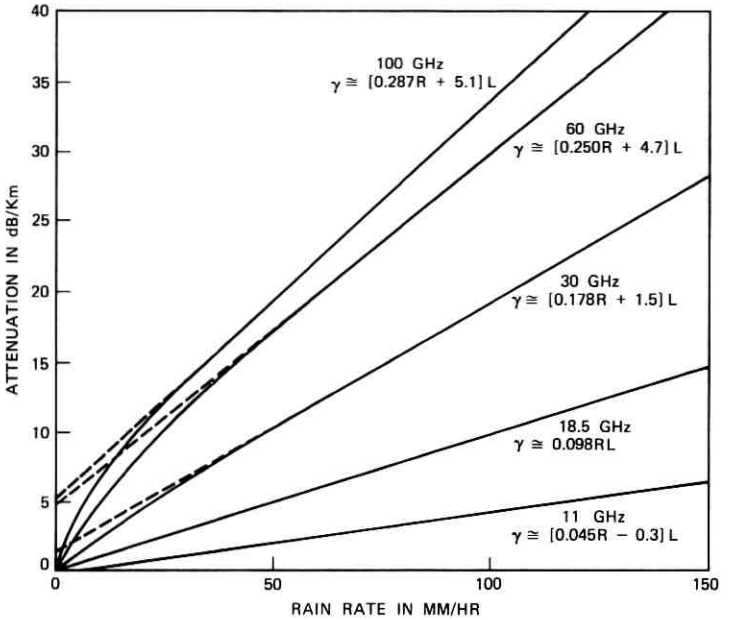


Fig. 1—Computed attenuation as a function of rain rate and frequency.

is not encountered in nature; both the density of rain on the path and the rain rate are functions of position and time. The divergence theorem can be used to relate the rain density, which is a volume function, to the rain rate, which is a surface vector. The volume chosen for the application of the divergence theorem to a radio path is shown in Fig. 2. It is the first Fresnel ellipsoid, a prolate ellipsoid with major axis L and minor axis $\sqrt{\lambda L}$. The path length is L and the wavelength of the transmission is λ. Typically, the ratio of these axes on short paths is on the order of 500. This volume is chosen because substantially all the power that reaches the receiver passes through it.

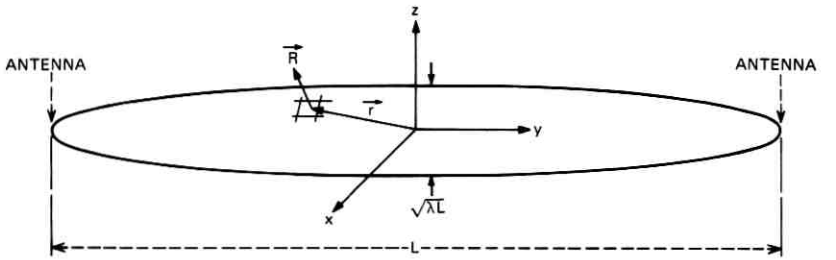


Fig. 2—An ellipsoidal surface enclosing the radio path.

If the rain rate is known at every point on the surface at all times, the volume of rain in the path can be computed.[3] However, rain rate must be measured in a specified time interval—called integration time—and if the interval is too long, important fine structure in the rain rate—and the resulting fades—could be missed.

The path geometry can be used to determine an interval short enough so that no deep fades will be missed. The time required for a deep fade to develop is related to the time required for the rain to fill a large fraction of the volume of the ellipsoid in Fig. 2. For a step function of rain with velocity V, the minimum time for the volume to fill is $\sqrt{\lambda L}/V$. It follows that a deep fade will not occur in a time $T_o \ll \sqrt{\lambda L}/V$. Thus, a rain gauge with an integration time T_o should be sufficiently fast for our purpose. Actually, of course, step functions of rain do not occur, and an integration time T larger than T_o will suffice. For a practical rain gauge design, T_o serves only as a guide and as assurance that a suitable rain gauge is at least possible.

Since rain rate cannot be measured at all points on the surface of the path, it is necessary to consider what can be done with a single rain gauge in the vicinity of the path.

Fortunately, the attenuation as a function of time is not required; for radio path design, knowledge of the fraction of time that the path attenuation exceeds the fading margin is sufficient. To determine this information, we need to relate the point rain rate distribution to the path attenuation distribution. The method of relating these distributions is described in Ref. 3, and only pertinent results will be given here. If we measure the point rain rate distribution for a small enough integration time and are then able to obtain distributions for multiples of this interval, an estimate of the path attenuation distribution can be computed from the point rain rate distribution obtained for a suitable integration time T. The appropriate integration time T is a function of path length and frequency and is discussed in Section III. A rain gauge that measures rain rate as prescribed by the theory is described in Section IV.[7]

III. RAIN RATE INTEGRATION TIME AS A FUNCTION OF PATH LENGTH AND FREQUENCY

The theory has so far produced a rain rate integration time T_o sufficiently short that no significant rates will be missed. It is, of course, expected to be unnecessarily short because it was derived for severe and unrealistic conditions. For a given path and wavelength, there is an integration time T for the point rain rate measurement for which

the point rain rate distribution best approximates the path average rain rate distribution. Thus, the attenuation distribution is related to frequency and path length through the rain rate integration time T. The rate at which the path average rain rate changes depends upon the dimensions of the path. We assume, therefore, that the integration time in seconds is proportional to the distance through the path of Fig. 2, averaged over all trajectories through the center.

$$T = \frac{1}{V_c} \frac{\sqrt{\lambda L}}{\pi} \ln 32 \frac{L}{\lambda}, \tag{1}$$

where

L is the path length in meters,

λ is the wavelength in meters, and

V_c is a constant with the dimension of velocity.

The constant V_c is determined experimentally from measured attenuation and point rain rate distributions. From data to be discussed later, $V_c \approx 0.95$ meter per second.

IV. THE FAST INTEGRATING RAIN GAUGE: RAIN RATE DISTRIBUTIONS AS A FUNCTION OF INTEGRATION TIME

A rain gauge designed to measure rain rates in accordance with the requirements of the theory is shown schematically in Fig. 3.[7] Rain is collected in a funnel with a 16-inch diameter. The water is directed by the small rotating funnel to one of four glass tubes. The tube receives the rain collected in a 1.5-second interval, and then the small funnel rotates to the next tube. After the water has settled in the first tube, its volume is measured by measuring the capacitance between two metallic plates attached to the outer surface of the tube. After the measurement, the tube is emptied and is ready to be used again. In this manner, the rain rate is measured in successive intervals of 1.5 seconds. Since all intervals are measured, it is simple to obtain rain rates for intervals that are multiples of 1.5 seconds.

The effect of rain rate integration time is shown in Fig. 4, in which the data were taken from the fast rain gauge. As expected, the highest rain rates are recorded for the shortest integrating time. From Fig. 4, it is clear that the rain rate distributions depend upon the integration time.

Figure 5 shows rain rate distributions for two years at Holmdel, N. J., measured on this rain gauge with integration time as a parameter. As expected, the distributions are functions of the integration time T.

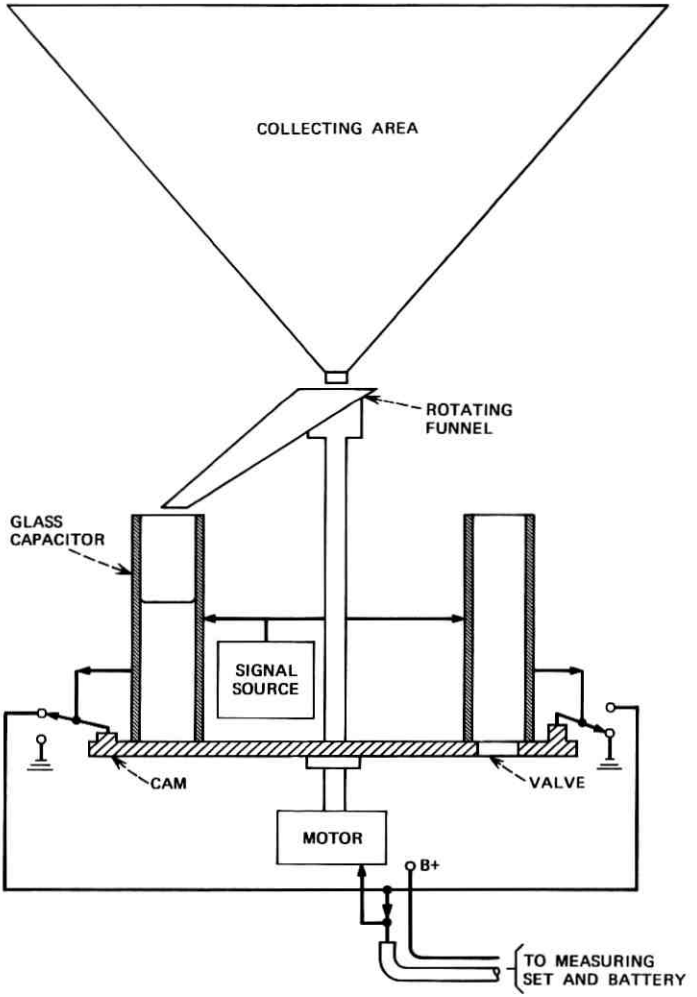


Fig. 3—The fast integrating rain gauge.

At present, only two of these fast rain gauges exist, so it has not been possible to obtain such distributions in many locations.

What is possible and practical is to measure distributions for an integration time of 1 minute from U. S. Weather Service rain gauge charts. We are therefore interested in converting 1-minute rain rate distributions to distributions at other measuring intervals. To facilitate this conversion, Fig. 6 has been prepared from Fig. 5. For many paths

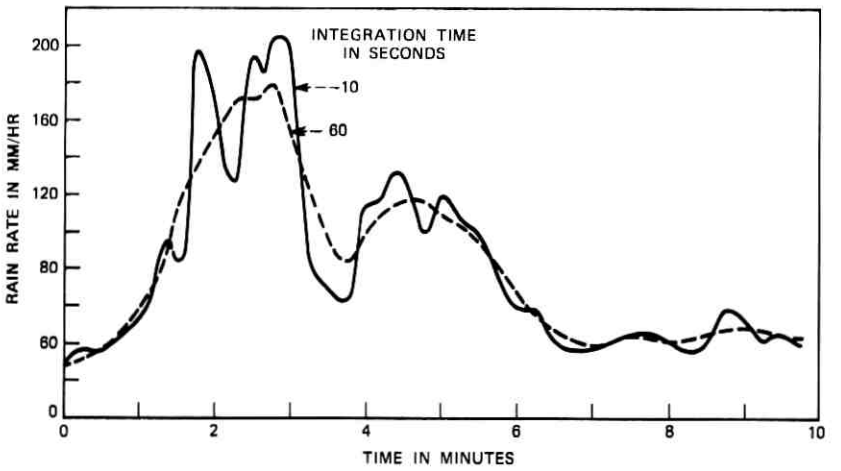


Fig. 4—Integrating rain gauge: storm of April 3, 1972.

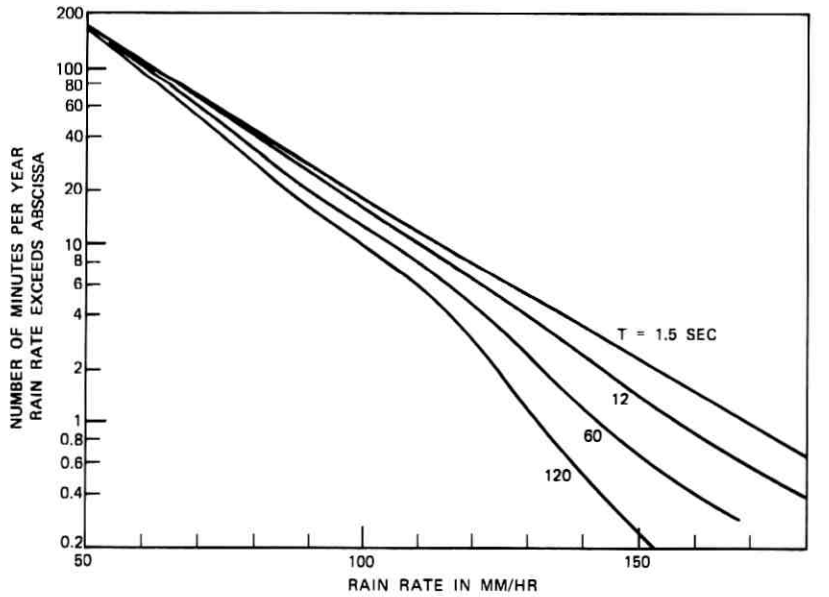


Fig. 5—Distribution of rain rates measured with fast-integrating rain gauge at Holmdel, N. J. for two years (1971 and 1972).

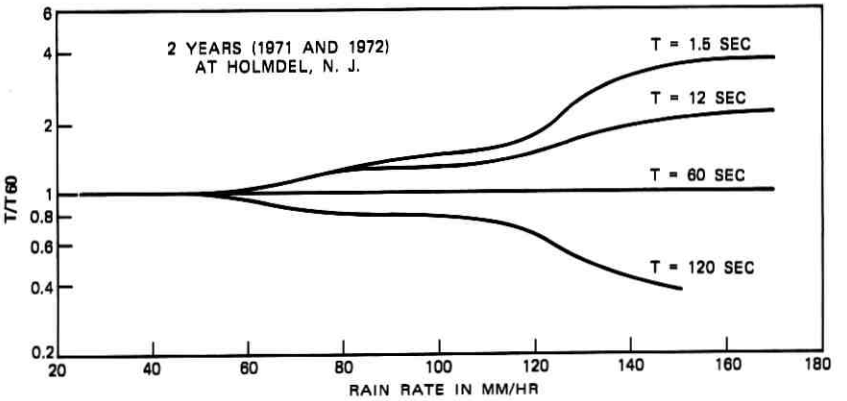


Fig. 6—$T/T60$ = (minutes per year rain rate exceeds abscissa for integration time T)/(minutes per year rain rate exceeds abscissa for 1-minute integration time).

of interest, the proper integration times fall between 12 and 120 seconds with the majority in the vicinity of 60 seconds. Thus, Fig. 6 shows that the distributions are not very sensitive to measuring interval out to rates of 120 millimeters per hour, which may correspond to a deep fade. Therefore, for many paths, 1-minute distributions will be adequate for design. For other paths, adjustments can be made from Fig. 6.

Note that the data shown in Fig. 6 are for a single location and cover a two-year period. It is not known to what extent these conversions apply to other locations, longer paths, and other periods of time.

V. THE EFFECTS OF NONSPHERICAL RAINDROPS

The theory was derived for spherical raindrops but, because raindrops are not quite spherical, the attenuation for vertical and horizontal polarizations are different. Morrison and Chu have computed the differential attenuation, the difference in attenuation between horizontally and vertically polarized waves, as a function of frequency for a specific raindrop model, and some of their results are shown in Fig. 7.[8] The attenuation of a vertically polarized signal is less than that of a horizontally polarized signal, and we take the average of these to correspond to the attenuation caused by spherical raindrops.

The dashed lines in Fig. 7 are linear approximations to the data and may be used in the design of radio paths described in Section IX. As measured distributions of differential attenuation become available, it may be necessary to modify the results of Fig. 7.

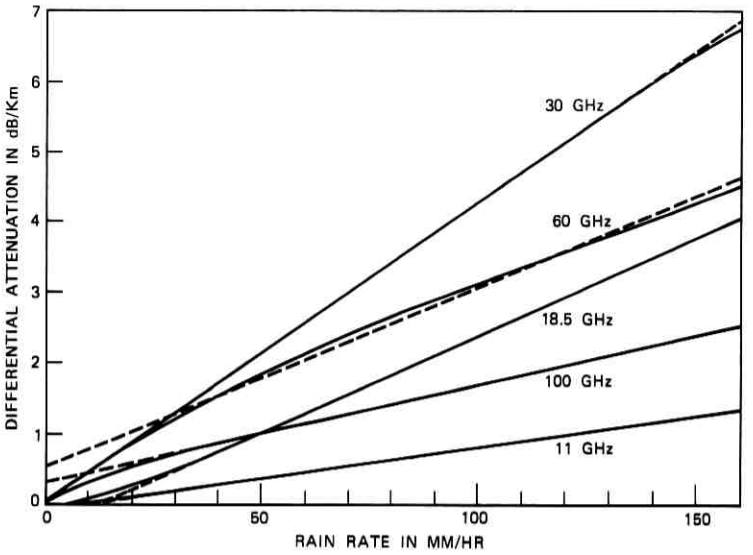


Fig. 7—Differential attenuation as a function of rain rate and frequency.

VI. COMPARISON OF THEORY AND EXPERIMENT USING DATA FROM THE FAST-INTEGRATING RAIN GAUGE

In this section and in Section VIII, comparisons are made between measured attenuation distributions and distributions predicted by the theory from measured rain rate distributions. Since the object of these comparisons is to test the efficiency of the procedure, it is important to use attenuation and rain rate distributions that are coincident in time and as near the same location as possible. In addition, distributions for time periods of less than one year were deemed inadequate.

The problem of estimating future attenuation distributions from existing rain rate distributions is different in that the time periods cannot be coincident. For this purpose, rain rate distributions covering a five-year interval are used in Section IX to illustrate the design of radio paths.

Two of the fast-integrating rain gauges are in service, one at Holmdel, N. J., and one at North Andover, Mass. At Holmdel, A. F. Dietrich and O. E. DeLange measured the attenuation on a 1.03-kilometer path at 60 GHz for the year 1971, and an attenuation distribution was obtained from these data. For the same year, estimated attenuation distributions for several rain gauge integration times were computed

from the fast-rain-gauge data. The distribution computed for an integration time of 12 seconds was the best fit to the measured attenuation distribution; the two distributions are shown in Fig. 8. Using a value of $T = 12$ seconds, the constant V_c in (1) was computed. Thus, based on this fit, we have an expression for rain rate integration time in seconds as a function of frequency and path length:

$$T \approx 1.05 \frac{\sqrt{\lambda L}}{\pi} \ln 32 \frac{L}{\lambda}. \tag{2}$$

At the Bell Laboratories facility in North Andover, Mass., G. H. Lentz measured attenuation on a 4.3-kilometer path at 18.5 GHz and rain rates with a fast-integrating rain gauge for three years.[9] The appropriate rain rate integration time for this path is, from (2), about 45 seconds. By interpolation of Fig. 6, the 45-second and 60-second rain rate distributions are nearly identical, so the attenuation computed from a 1-minute distribution is accurate enough and is shown with the measured distribution of attenuation in Fig. 9.

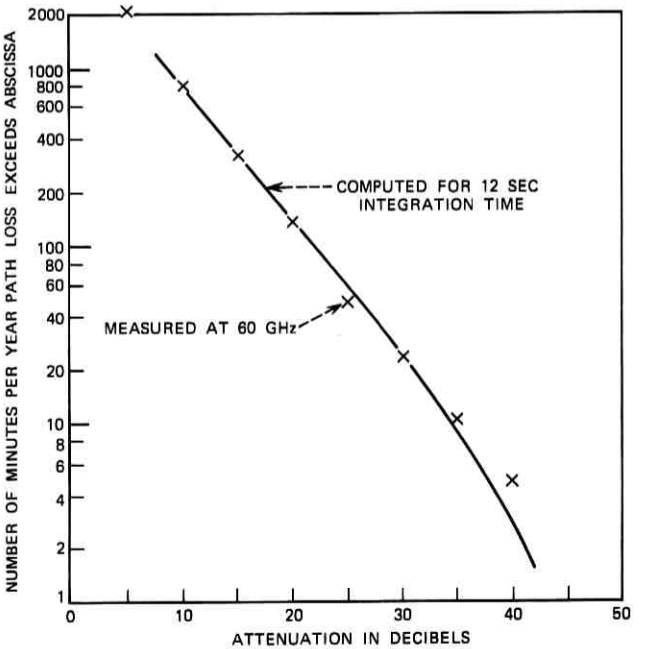


Fig. 8—Computed and measured attenuation distributions at 60 GHz for 1971.

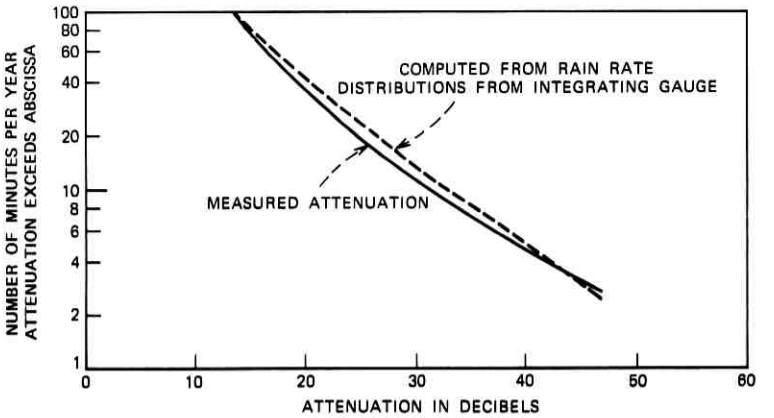


Fig. 9—Measured and computed attenuation distributions at North Andover, Mass., for three years (1971 to 1973), vertical polarization.

This is certainly satisfactory agreement between theory and experiment. The problem remains, however, of how to obtain suitable rain rate distributions for other locations.

VII. MEASUREMENT OF RAIN RATE DISTRIBUTIONS FROM U. S. WEATHER SERVICE DATA

The U. S. Weather Service operates rain gauges at many locations; for example, there are nearly 300 first-class weather stations, all of which operate rain gauges.[4] There is, then, an immense amount of rain gauge data available which, if rain rates can be derived from it, can be used to produce accurate estimates of attenuation distributions anywhere in the country.

A weighing gauge measures the depth of water accumulated as a function of time, and a reproduction of part of such a record is shown in Fig. 10a. This chart does not show the rain rate, which is the derivative of the curve on the chart. However, a method has been devised for taking data at 1-minute intervals and using the data to compute the rain rate and its distribution.

Computing derivatives in this manner is notoriously inaccurate, and considerable processing is necessary to get accurate results, especially at high rain rates. The processing used in this work is described elsewhere.[10] The rain rate computed from the chart in Fig. 10a is shown in Fig. 10b.

One-minute rain rate distributions for a 5-year period are shown in Fig. 12 for the locations indicated on the map of Fig. 11. The use of

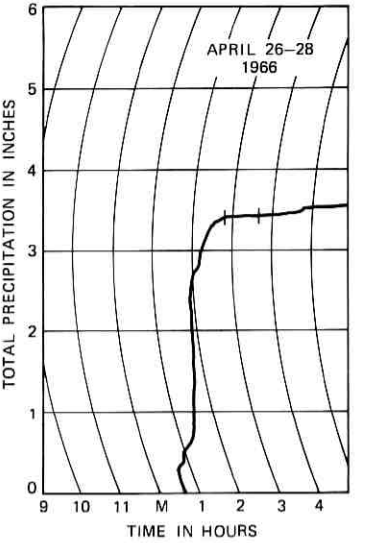


Fig. 10a—National Weather Service chart, Dallas, Texas.

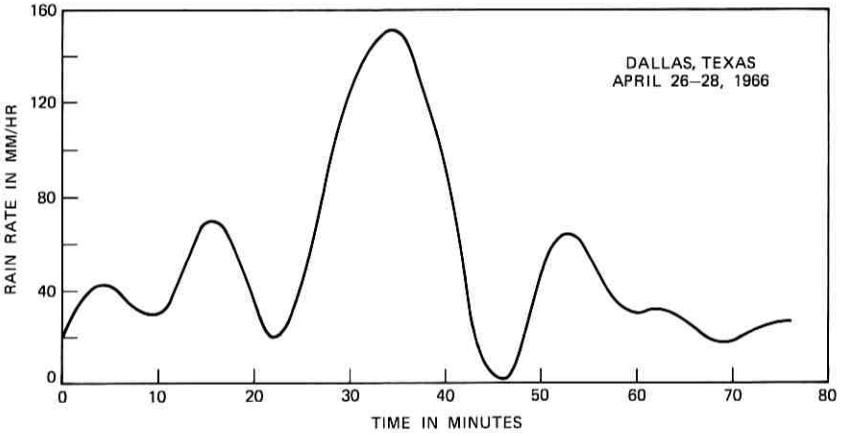


Fig. 10b—Rain rates vs time computed from weighing gauge data.

these distributions for the design of radio paths is demonstrated in Section IX.

VIII. COMPARISON OF THEORY AND EXPERIMENT USING U. S. WEATHER SERVICE DATA

Hudson, N. H., is about ten miles from the radio path at North Andover, Mass. The estimated attenuation distribution for the 4.3-kilometer path at 18.5 GHz was computed from the rain rate distribu-

Fig. 11—Locations for which 1-minute rain rate distributions are shown in Fig. 12 (Key: d3 refers to Fig. 12d, curve 3).

tion measured at Hudson, N. H., as previously described. This result is shown with the measured distribution in Fig. 13.

Figure 14 compares a 3-year attenuation distribution, measured by R. A. Semplak on a 6.4-kilometer path at 18.5 GHz, to the estimated distribution computed from the Newark, N. J. rain rate distribution.[11] The distance from Newark to Crawford Hill, N. J., is about 30 miles.

Similarly, Fig. 15 compares the estimated attenuation distribution computed from the Atlanta, Ga., rain rate distribution with an attenuation distribution reported by S. H. Lin on a 5.15-kilometer path at 17.75 GHz near Palmetto, Ga., a distance of about 25 miles.[12]

It is evident from these comparisons that attenuation distributions can be estimated from rain rate distributions obtained by processing rain gauge charts. Design of radio paths using these distributions are now discussed.

IX. THE DESIGN OF SHORT RADIO PATHS

Determination of the path length for a specified outage time requires a knowledge of the wavelength and fading margin. Rain attenuation depends upon the wavelength, as already discussed in Section II. The fading margin requires a little more discussion. During periods of normal free-space propagation, the power received at the output port of the receiving antenna is given by

$$P_R = P_T \frac{A_T A_R}{\lambda^2 L^2}, \tag{3}$$

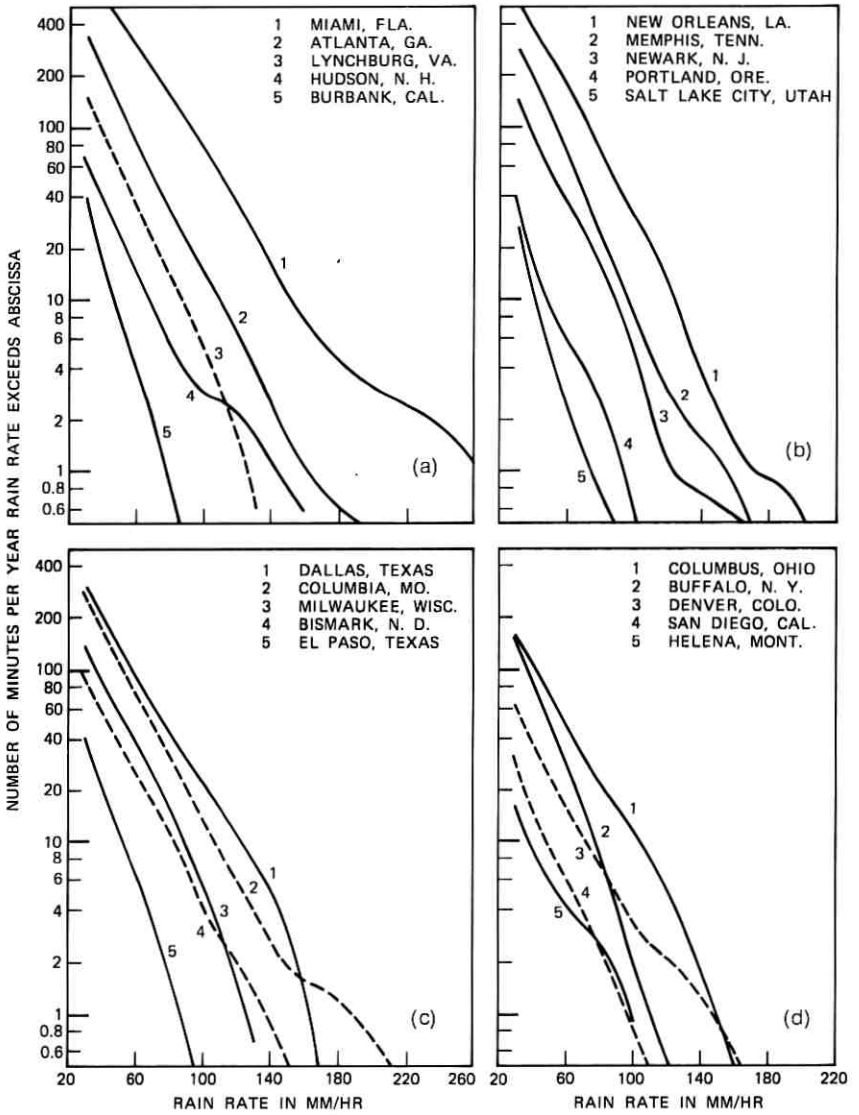


Fig. 12—Rain rate distributions average of five years (1966 to 1970).

where

P_T is the power at the input port of the transmitting antenna,

P_R is the power received at the output port of the receiving antenna,

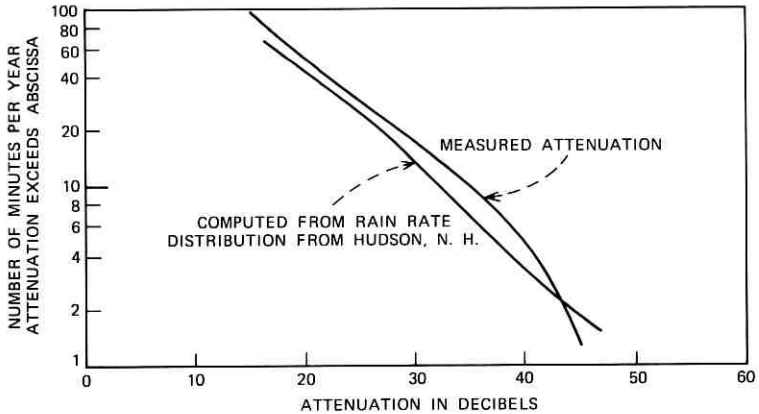


Fig. 13—Measured and computed attenuation distributions at North Andover, Mass., for year 1971, vertical polarization.

A_T, A_R are the effective areas of the transmitting and receiving antennas,

λ is the wavelength of the transmitted signal,

and

L is the distance between the transmitter and receiver.

The relationship between the effective antenna area and the gain, G,

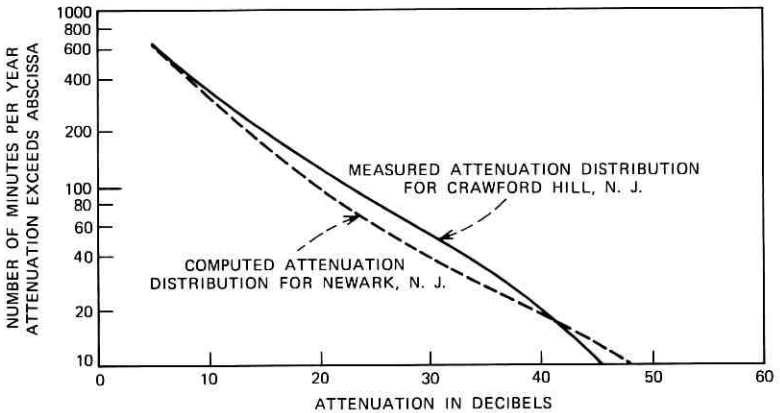


Fig. 14—Comparison of measured attenuation distributions at Crawford Hill, N. J., with computed attenuation distributions from rain rate data from Newark, N. J. for three years (1967 to 1969), vertical polarization.

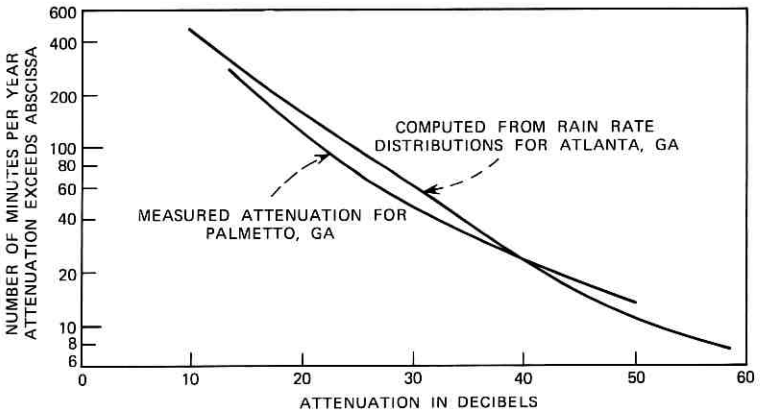


Fig. 15—Measured and computed attenuation distributions at Palmetto, Ga., for year 1971.

of the antenna is

$$G = \frac{4\pi A}{\lambda^2}.$$

In any system there is a minimum power, $P_{\min}$, that is required to meet the transmission objectives. If the received power falls below this value, the system is said to suffer an outage. The duration of the outage is the time interval during which the received power is less than $P_{\min}$, which depends upon the noise factor and bandwidth of the receiver and on the type of modulation and detection. With this convention, the fading margin in decibels is defined as

$$\alpha \equiv 10 \log P_R/P_{\min}. \tag{4}$$

Substituting for P_R from (3) reveals that the fading margin is a function of path length.

$$\alpha(L) = 10 \log \left(\frac{P_T}{P_{\min}} \times \frac{A_T A_R}{\lambda^2 L^2} \right). \tag{5}$$

It is convenient to write the fading margin for path length L in terms of the fading margin for a 1-kilometer spacing between antennas in decibels. Thus, if $\alpha(1)$ is the fading margin for a 1-kilometer path, the fading margin for an L-kilometer path in decibels is

$$\alpha(L) = \alpha(1) - 20 \log L. \tag{6}$$

From Fig. 1 we recall that the attenuation resulting from uniform rain as a function of rain rate can be described by linear functions over

the range of rain rates of interest. Thus, we have, from Fig. 1,

$$\gamma(L) = [aR + b]L, \tag{7a}$$

where

R is rain rate in millimeters per hour,
L is path length in kilometers,

and

a, b depend on wavelength.

The differential attenuation of Fig. 7 can be written

$$\delta\gamma(L) = (\delta a \times R + \delta b)L.$$

Using this expression, the attenuation of vertically and horizontally polarized signals can be written in the same form as (7a).

$$\gamma_V(L) = [(a - \delta a)R + (b - \delta b)]L, \tag{7b}$$

$$\gamma_H(L) = [(a + \delta a)R + (b + \delta b)]L. \tag{7c}$$

The parameters for several frequencies of interest are given in Table I.

When the rain attenuation exceeds the fading margin, the path is out of service. The rain rate for which the attenuation equals the fading margin can be found by setting $\alpha(L) = \gamma(L)$ in (6) and (7) to obtain

$$\alpha(1) = [aR + b]L + 20 \log L \tag{8a}$$

$$\alpha_V(1) = [(a - \delta a)R + (b - \delta b)]L + 20 \log L \tag{8b}$$

$$\alpha_H(1) = [(a + \delta a)R + (b + \delta b)]L + 20 \log L. \tag{8c}$$

These are nonlinear equations, and it is useful to present solutions in graph form, as in Fig. 16, for a frequency of 18.5 GHz.

Table I — Coefficients for attenuation as a function of rain rate

Frequency in GHz	a	b	δa	δb
11	0.045	−0.3	0.0046	−0.06
16	0.077	−0.08		
18.5	0.098	0	0.014	−0.2
30	0.178	1.5	0.0216	0
60	0.250	4.7	0.0129	0.25
100	0.287	5.1	0.0069	0.15
150	0.292	4.9		
300	0.275	4.45		

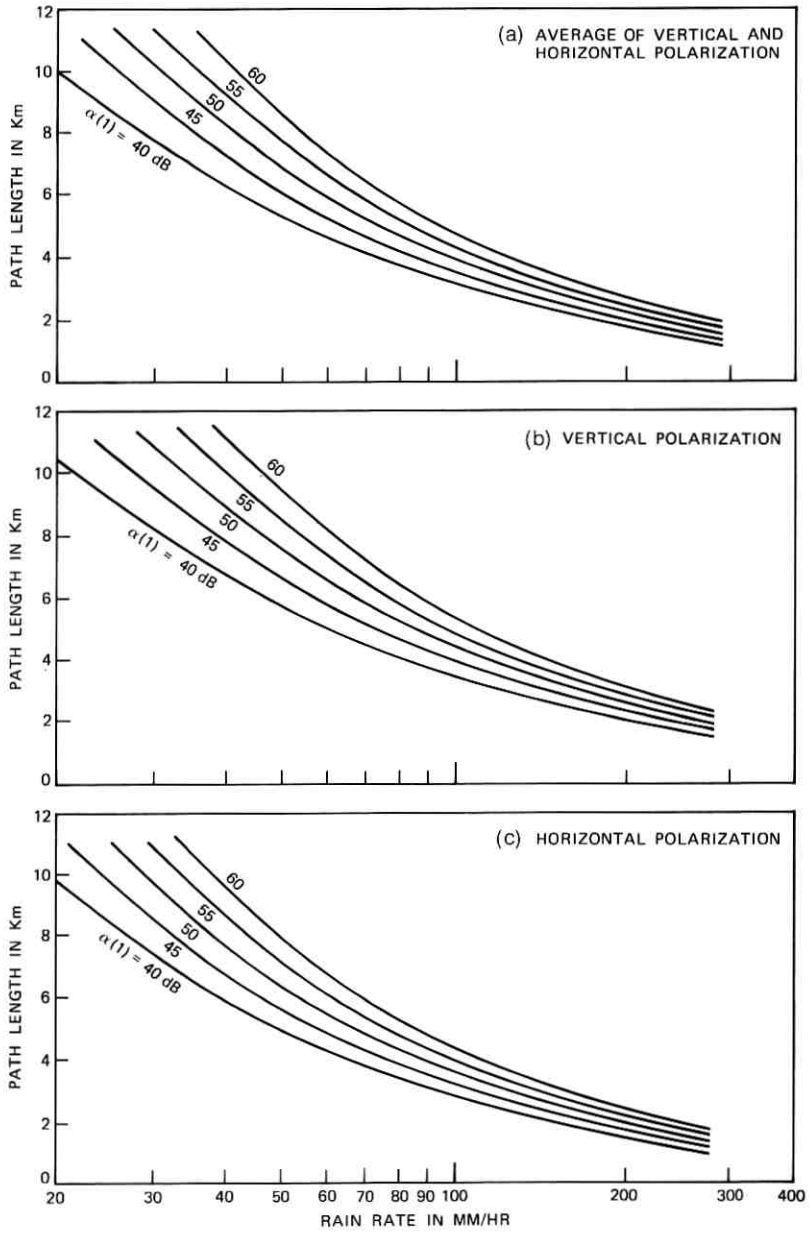


Fig. 16—Path length as a function of rain rate at 18.5 GHz.

We are now in a position to illustrate the design of short radio paths.

Example 1. Determine the outage time for a path of length $L = 6$ kilometers for a system characterized by a 1-kilometer fading margin, $\alpha(1) = 50$ dB, and an operating frequency of 18.5 GHz with vertical polarization. From (2), $T \approx 54$ seconds, so 1-minute rain rate distributions will suffice.

Enter Fig. 16b at $L = 6$ kilometers and $\alpha(1) = 50$ dB. The resulting rain rate is about 66 millimeters per hour. If the system is to be located in New Jersey, enter the rain rate distribution of Fig. 12b(3) at 66 millimeters per hour and note that the average outage time is about 28 minutes per year.

Example 2. A hypothetical 18.5-GHz system is to be built in New Hampshire between terminals separated by 40 kilometers. The service for which the system is intended requires that the average outage time not exceed 0.02 percent per year, which is 105 minutes. The system is characterized by a 1-kilometer fading margin of $\alpha(1) = 50$ dB.

We use the rain rate distribution for Hudson, N. H. Since there must be an integral number of hops in the system, we compute the total outage time for various numbers of hops, and the smallest number of hops resulting in an outage time of 105 minutes or less is the proper answer. We proceed as follows. For five hops, the path length would be 8 kilometers per hop. The polarization has not been specified, so Fig. 16a is used. From Fig. 16a, a rain rate of 41 millimeters per hour is obtained. Using this rain rate and the rain rate distribution for Hudson, N. H., an outage time of 35 minutes for an 8-kilometer path is found. There are five hops, so the total outage time is 175 minutes. It is clear that 8-kilometer paths are too long. Repeating the process for larger numbers of hops, it is instructive to plot the results, as in Fig. 17. From the figure we see that seven hops with average path length of 5.7 kilometers will be required. For paths of this length, $T \approx 52$ seconds, so again the 1-minute distribution will suffice.

It would be rare if the system could be laid out with equal path lengths. When the repeater sites are selected and the path lengths known, the expected system outage can be computed by summing the estimated outages for the individual paths as illustrated in the first example. The system outages computed in this manner will be conservative, since any joint path outages cannot be distinguished and, in effect, are counted as separate outages. To the extent that over-

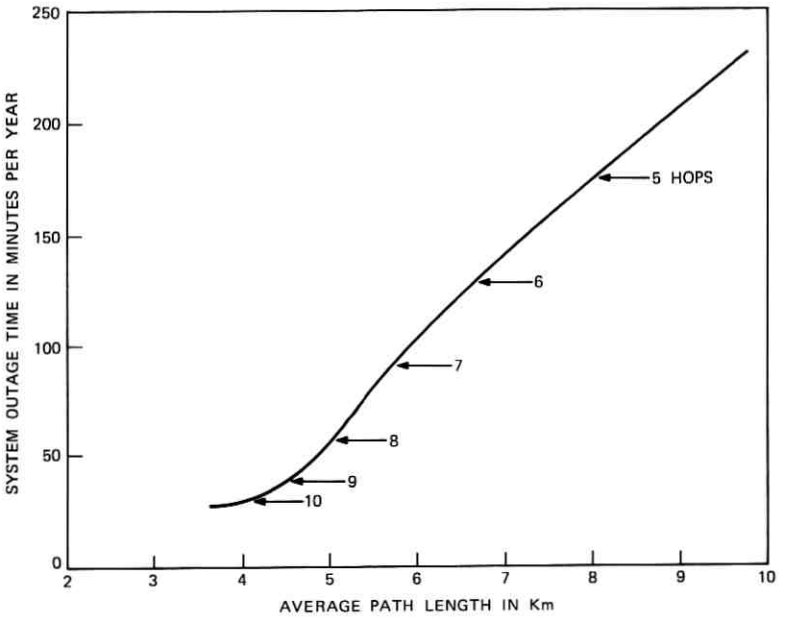


Fig. 17—System outage time, average path length, and number of hops for a total system length = 40 km and $\alpha(1) = 50$ dB.

lapping outages occur, the actual system outage time will be somewhat less than predicted.

X. CONCLUSION

A procedure has been described for designing short radio paths to meet transmission outage objectives when fading is caused by rain attenuation. The method is simple and examples are given. Data required include the operating frequency, the fading margin, the polarization to be used, and a suitable rain rate distribution. For many paths and frequencies of interest, accurate attenuation distributions can be estimated from 1-minute rain rate distributions. Examples of such distributions for a 5-year period, which were obtained by processing rain gauge data supplied by the U. S. Weather Service, have been presented. The design procedure is applicable to any section of the country for which suitable rain gauge data are available. This includes many of the several hundred stations operated by the U. S. Weather Service. Application of this method to longer paths is possible; work to accomplish this is presently under way.

XI. ACKNOWLEDGMENTS

We are indebted to F. M. Garland of the National Climatic Center for selecting the rain gauge charts used in this work. And we are grateful to Mrs. D. Vitello for writing and operating computer programs and Mrs. B. G. Griffin for assistance in the data processing.

REFERENCES

1. C. L. Ruthroff, T. L. Osborne, and W. F. Bodtmann, "Short Hop Radio System Experiment," B.S.T.J., *48*, No. 6 (July-August 1969), pp. 1577–1604.
2. J. W. Ryde, "Echo Intensity and Attenuation due to Clouds, Rain, Hail, Sand, and Dust Storms at Centimeter Wavelengths," Rep. 7831, General Electric Company Research Laboratories, Wembley, England, October 1941. J. W. Ryde and D. Ryde, "Attenuation of Centimeter Waves by Rain, Hail, and Clouds," Rep. 8516, General Electric Company Research Laboratories, Wembley, England, August 1944. J. W. Ryde and D. Ryde, "Attenuation of Centimeter and Millimeter Waves by Rain, Hail, Fog, and Clouds," Rep. 8670, General Electric Company Research Laboratories, Wembley, England, May 1945.
3. C. L. Ruthroff, "Rain Attenuation and Radio Path Design," B.S.T.J., *49*, No. 1 (January 1970), pp. 121–135.
4. A. H. Jennings, "Maximum Recorded United States Point Rainfall for 5 Minutes to 24 Hours at 296 First Order Stations," Technical Paper No. 2, U. S. Department of Commerce, Washington, D. C., Revised 1963.
5. R. G. Medhurst, "Rainfall Attenuation of Centimeter Waves: Comparison of Theory and Measurement," IEEE Trans. Antennas and Propagation, *Ap-13*, No. 4 (July 1965), pp. 550–564.
6. D. E. Setzer, "Computed Transmission Through Rain at Microwave and Visible Frequencies," B.S.T.J., *49*, No. 8 (October 1970), pp. 1873–1892.
7. W. F. Bodtmann, unpublished work, 1974.
8. J. A. Morrison and T. S. Chu, "Perturbation Calculations of Rain-Induced Differential Attenuation and Differential Phase Shift at Microwave Frequencies," B.S.T.J., *52*, No. 10 (December 1973), pp. 1907–1913.
9. G. H. Lentz, unpublished work, 1974.
10. C. L. Ruthroff, "Computing Derivatives from Equally-Spaced Data," unpublished work.
11. R. A. Semplak, "The Influence of Heavy Rainfall on Attenuation at 18.5 GHz and 30.9 GHz," IEEE Trans. Ant. and Prop., *AP-18*, No. 4 (July 1970), pp. 507–511.
12. S. H. Lin, unpublished work, 1972.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 7, September 1974
Printed in U.S.A.

An Improved Antenna for Microwave Radio Systems Consisting of Two Cylindrical Reflectors and a Corrugated Horn

By C. DRAGONE

(Manuscript received January 31, 1974)

A broadband antenna is described with an elliptical beam suitable for efficient illumination of the United States from a satellite in synchronous orbit. The antenna is also suitable for use in terrestrial radio systems above 10 GHz. It consists of a corrugated feed and two parabolic cylinders that efficiently transform the circularly symmetric beam radiated by the feed into an elliptical beam. Depolarization of the incident beam by the two cylinders is very small and essentially independent of the angle of incidence, which can therefore be chosen as large as required to avoid aperture blockage.

The performance is described of an antenna with a 5.8° × 2.9° elliptical beam at 18.5 GHz. For any input polarization, the cross-polarized component of the far field remains over the entire beam at least 33.5 dB below the main component on axis. This cross-polarized component is due in part to imperfections in the corrugated feed and in part to some aperture blockage by the feed and depolarization by the cylinders.

A first-order analysis of the frequency dependence shows that the beamwidths vary little with frequency for an antenna using a properly designed feed and cylinders of sufficiently large apertures. As the frequency is increased from 18.5 to 29 GHz, the measured horizontal and vertical beamwidth variations are +2.7 and −14 percent, respectively.

I. INTRODUCTION

The suitability of corrugated feeds for use in reflector antennas of revolution is well known.[1,2] When this feed is properly designed, the resulting far field of the antenna is a circularly symmetric beam reproducing, in all directions, the input polarization of the feed. It is shown here that this feed is also suitable for use with orthogonal

cylinders, with excellent performance. Although the far field in this case has a cross-polarized component resulting from depolarization by the cylinders, this component is very small, negligible for most applications. An important advantage of cylindrical reflectors over reflectors of revolution* is that this cross-polarized component is essentially independent of the angle of incidence, for incidence in a plane orthogonal to the axis of the first cylinder. Because of this property, aperture blockage can be eliminated by properly orienting the feed without sacrificing the polarization properties of the antenna.

We describe the performance of an antenna consisting of two orthogonal parabolic cylinders[4] and a corrugated feed (Fig. 1). The two cylinders transform the circularly symmetric beam radiated by the feed into an elliptical beam. Thus, this antenna is particularly suitable for applications requiring different beamwidths in the two principal planes. One such application arises when a satellite in synchronous orbit is required to efficiently illuminate a region of approximately elliptical shape, such as the United States. Another application[5] arises in connection with terrestrial microwave radio systems above 10 GHz, where an important limitation arises in the use of antennas with very narrow beams, because of the finite stability of the towers on which the antennas are to be mounted. In that case, the choice of beamwidth in a vertical plane may be governed by the maximum movement of the tower in heavy wind, and it may therefore be desirable[5] to choose different values for the beamwidths in the two principal planes.

In Fig. 1a the two parabolic cylinders are so arranged† that, if a point source is placed at F in front of the first cylinder, a spherical wave is transformed by the two cylinders into a plane wave. This implies that the focal point F lies on the focal line of the first cylinder. The spherical wave is therefore transformed by this reflector into a cylindrical wave originating from a virtual line source behind the reflector (see Fig. 1a). The second reflector, which is disposed so that its focal line coincides with the line source, then transforms the cylindrical wave into a plane wave. All this means that the rays from F that are intercepted by the first cylinder become, after the two reflections, parallel rays, i.e., rays focused at infinity. Because of this property, it can be shown that if the two cylinders are of sufficiently

* The polarization properties of reflectors of revolution for oblique incidence are discussed in Ref. 3.

† Details of the geometry of the two cylinders and their transformations are given in Ref. 4.

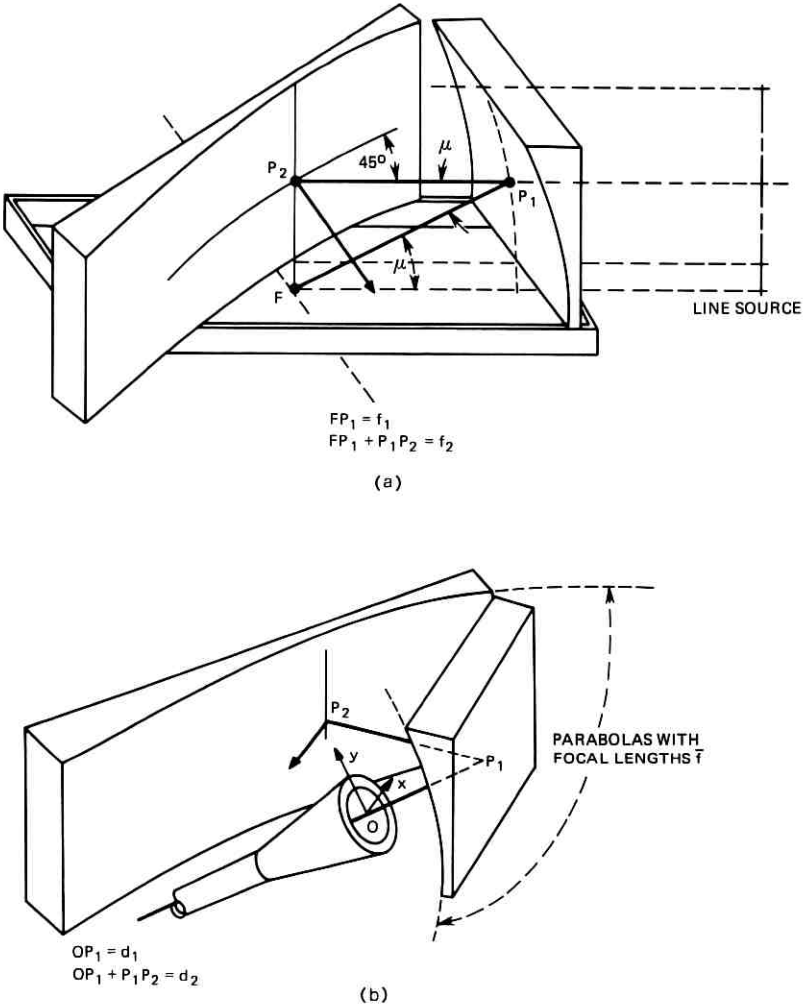


Fig. 1—Double cylinder antenna (a) with feed removed and (b) with feed shown. $FP_1 = f_1$, $FP_1 + P_1P_2 = f_2$, $FQ_1 = f$; $OP_1 = d_1$, $OP_1 + P_1P_2 = d_2$.

large aperture there is a simple relation[6] between the far field and the field over the focal plane Σ_0, which is the plane orthogonal to the feed axis at F. More precisely, consider a plane Σ_1 orthogonal to the beam at a great distance from the antenna. To a first order, Σ_0 and Σ_1 are conjugate planes and therefore their field intensities can be determined

accurately using the imaging laws of geometrical optics.* Since the magnifications of Σ_1 in the horizontal and vertical directions are proportional to the focal lengths f_1 and f_2, respectively, it follows that the circularly symmetric field radiated on Σ_0 by the feed is imaged into an elliptical beam whose beamwidths in the two principal planes have the ratio

$$r = \frac{f_1}{f_2}. \tag{1}$$

An important consequence of the relation of far field to focal plane field is that, since the field distribution over the aperture of a properly designed feed varies little with frequency and since the feed aperture is normally placed close to the focal plane, the beam of the antenna is essentially frequency independent. This can be an important property in many applications. We now derive the antenna characteristics under the assumption that the wave transformation by the two cylinders is efficient, i.e., that each cylinder intercepts essentially all the energy incident on it.

II. ANALYSIS

Throughout this section we assume that the feed radiates a narrow beam. This implies that the radius a of the circular feed aperture is much larger than the wavelength λ,

$$ka \gg 1,$$

where $k = 2\pi/\lambda$.

Suppose the wave incident on the first cylinder is a spherical wave originating from the focal point F. According to geometrical optics, this wave is transformed by the two cylinders into a plane wave having the following characteristics.[4] If, for the incident wave,

$$E_x = 0, \tag{2}$$

then the resulting plane wave is vertically polarized; if, instead,

$$H_x = 0, \tag{3}$$

then the plane wave is horizontally polarized. In the first part of this section, we assume the former condition.

Since the feed radiates a narrow beam, consideration can be restricted to the field in the paraxial region of the principal ray, which

---

* This result is derived in Ref. 6 for systems having rotational symmetry and is extended in Appendix A to the asymmetric system of Fig. 2.

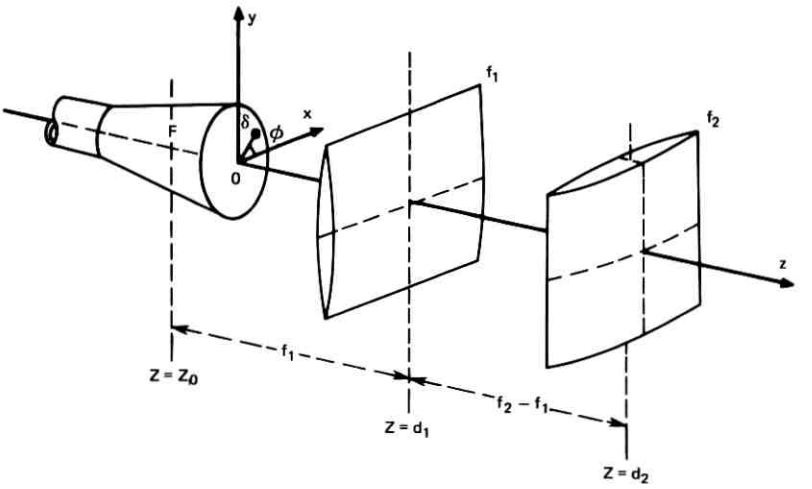


Fig. 2—System consisting of a feed and two ideal cylindrical lenses.

is the ray that corresponds to the feed axis. The system of Fig. 1b can then be replaced by the fictitious system of Fig. 2, consisting of two cylindrical lenses of focal lengths, f_1, f_2. The lenses in Fig. 2 are assumed ideal; they simply multiply the incident field by the phase factors

$$\exp\left(-j\frac{k}{2f_1}x^2\right) \quad \text{and} \quad \exp\left(-j\frac{k}{2f_2}y^2\right). \tag{4}$$

The distances, d_1, d_2, of the two lenses from the feed aperture are the distance in Fig. 1b between the two reflectors and the feed aperture, measured along the principal ray. We assume that the separation $d_2 - d_1$ between the two lenses is equal to $f_2 - f_1$,

$$d_2 - d_1 = f_2 - f_1, \tag{5}$$

so that the focal lines of the two lenses coincide. Then a spherical wave originating from the point $z = z_0$, where

$$z_0 = d_1 - f_1 = d_2 - f_2, \tag{6}$$

is transformed by the two lenses into a plane wave. The rectangular system of x, y, z coordinates in Fig. 2 has been chosen so that the feed aperture lies in the $z = 0$ plane centered at $x = y = z = 0$.

The following correspondence exists between the two systems of Figs. 1b and 2. Let s be a parameter measuring distance from the center of the feed aperture along the principal ray of Fig. 1b. Then

the field distribution over a plane normal to the principal ray at some point $s = s'$ is given by the field in Fig. 2 over the corresponding plane $z = s'$. This correspondence is, of course, valid only in the neighborhood of the principal ray.

2.1 Field for $z \to \infty$

Using the Fresnel diffraction formula,* a simple expression for the far field is derived in Appendix A, relating the far field to the focal plane field. Consider a plane $z = z_1$ at a great distance from the antenna. Because of condition (5), the planes $z = z_0$ and $z = z_1$ are conjugate planes. If x_0, y_0, z_0 and x_1, y_1, z_1 are corresponding points over these two planes, it follows from eq. (32) of Appendix A that

$$|E_y(x_1, y_1, z_1)| \to \frac{\sqrt{f_1 f_2}}{z_1} \left| E_y\left(\frac{x_1}{M_x}, \frac{y_1}{M_y}, z_0\right)\right|, \tag{7}$$

where M_x and M_y are the magnifications in the xy-directions,

$$M_x = \frac{x_1}{x_0} = -\frac{z_1}{f_1}, \qquad M_y = \frac{y_1}{y_0} = -\frac{z_1}{f_2}. \tag{8}$$

Thus, the far field is given by the field over the focal plane $z = z_0$. Since in practice this plane is not too far from the aperture plane of the feed, its field distribution can be determined accurately with little difficulty, using the Fresnel diffraction formula or the procedure of Appendix B.

2.2 Feed characteristics

The corrugated feed is a conical horn with circular symmetry. Its aperture is illuminated by the fundamental mode of the horn, which is a spherical hybrid mode[9] generated from a TE_{11}-mode of a smooth waveguide by a transducer† connected to the input of the horn. Let (ρ, ϕ) be polar coordinates defined by $x = \rho \cos \phi$, $y = \rho \sin \phi$ (see Fig. 2) and assume the input TE_{11}-mode is vertically polarized. Then we can show that the vertical component of the field over the aperture

---

* Fresnel's formula is applicable provided the wave equation $(k^2 + \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2)E_y = 0$ can be approximated by the parabolic wave equation (Ref. 7)

$$[k^2 + \tfrac{1}{2}(\partial^2/\partial x^2 + \partial^2/\partial y^2)]E_y = -jk\partial E_y/\partial z.$$

This approximation is justified in our case (we assume $ka \gg 1$; $f_1, f_2 \gg a$), since the field radiated by the feed is made up of plane waves whose directions of propagation are mostly confined to a small angular region about the z-axis (Ref. 8).

† Details of the feed and the transducer, which are of standard design, are not given here.

is given accurately by the function

$$\psi = \psi(\rho) = J_0(u) \exp\left(jk \frac{\rho^2}{2R}\right), \tag{9}$$

where J_0 is the Bessel function of order zero and

$$u = \frac{\rho}{a} u_{a0},$$

u_{a0} being the first zero of $J_0(u)$,

$$J_0(u_{a0}) = 0, \qquad u_{a0} = 2.4048.$$

The factor $\exp[jk(\rho^2/2R)]$ in eq. (9) arises because the hybrid mode illuminating the feed aperture has a spherical phase front with radius of curvature R approximately equal to the length of the horn from vertex to aperture. Equation (9) is true provided $R \gg a$, a condition approximately satisfied in the experiment ($R \cong 4.17\ a$). Note that $\psi = 0$ for $\rho = a$. This is due to the corrugated wall that imposes to a good approximation the boundary condition

$$E_\phi = H_\phi = 0, \tag{10}$$

where E_ϕ, H_ϕ are the components of E, H in the ϕ-direction. Appendix C shows that a consequence of this condition is that the field over the aperture contains, in addition to the component $E_y = \psi(\rho)$, a small component given accurately by

$$E_x \cong \frac{1}{2k^2} \cdot \frac{\partial^2 E_y}{\partial x \partial y}. \tag{11}$$

As a consequence, the far field contains a small (undesirable) horizontal component* that can be determined accurately by replacing the system of Fig. 1 by that of Fig. 2. The amplitude of this component is therefore given by a formula analogous to eq. (7) [simply replace E_y with E_x in eq. (7)].

2.3 First-order polarization properties of the far field

The location of the feed is normally chosen so that its phase center, the center of curvature P_c of the phase fronts of the far field, coincides with the focal point $z = z_0$ of the two reflectors. We assume this condition. Consider first the ideal case $R = \infty$, in which P_c is at the center

* Actually, E_x also causes a small vertical component. This component is, however, much smaller (for large ka) than the vertical component resulting from E_y and can therefore be neglected.

of the feed aperture and therefore the aperture is placed in the focal plane. Then, over this plane, E_y is given by $\psi(\rho)$ and eq. (7) gives for the far field

$$|E_y| \to \frac{\sqrt{f_1 f_2}}{z} |\psi(v)|, \tag{12}$$

where

$$v = \sqrt{\xi^2 + \eta^2}, \qquad \xi = \frac{x f_1}{z}, \qquad \eta = \frac{y f_2}{z}. \tag{13}$$

A relation identical to eq. (7) can be written for E_x. Therefore, since, for $z = 0$, E_x is given by eq. (11),

$$|E_x| \to \frac{\sqrt{f_1 f_2}}{z} \frac{1}{2k^2} \left| \frac{\partial^2 \psi}{\partial \xi \partial \eta} \right| = \frac{\sqrt{f_1 f_2}}{2} \frac{1}{2k^2} \left| \frac{\xi \eta}{v^2} \left(\psi''(v) - \frac{\psi'(v)}{v} \right) \right|. \tag{14}$$

We note from this relation that $|E_x|$ attains its maximum value for $\xi = \eta = v_0/\sqrt{2}$, where v_0 is the value of v for which $|\psi''(v) - \psi'(v)/v|$ is maximum. In the particular case of eq. (9), with $1/R = 0$,

$$|\psi| = \left| J_0 \left(\frac{u_{a0}}{a} v \right) \right|, \tag{15}$$

$$|\psi''(v) - \psi'(v)/v| = \left(\frac{u_{a0}}{a} \right)^2 \left| J_2 \left(\frac{u_{a0}}{a} v \right) \right|, \tag{16}$$

for $v < a$. We can verify that the maximum value of J_2 in the interval $v \leqq a$ occurs for $v = a$ and is $J_2(u_{a0}) = 0.431$. Therefore, if C denotes the ratio between the peak of $|E_x|$ and the peak of $|E_y|$ (the peak of $|E_y|$ occurs on axis), using eqs. (12), (14), (15), and (16), we find

$$C = \frac{|E_x|_{\max}}{|E_y|_{\max}} = \frac{1}{4} \left(\frac{u_{a0}}{ka} \right)^2 J_2(u_{a0}) \cong \frac{0.6231}{(ka)^2}. \tag{17}$$

Next, consider the case $1/R \neq 0$ but, instead of considering the distribution of eq. (9), assume that E_y over the aperture plane $z = 0$ varies as

$$\exp\left(-\frac{\rho^2}{w^2} \right) \exp\left(jk \frac{\rho^2}{2R} \right). \tag{18}$$

Appendix B points out that, for

$$w \cong 0.6437a, \tag{19}$$

this distribution represents fairly accurately that of eq. (9).

The phase center P_c (and, therefore, the focal plane) is now behind the aperture of the feed, located at[6,10]

$$z_0 = -\frac{R}{1 + (\lambda R/\pi w^2)}, \tag{20}$$

and we can verify that, over the plane $z = z_0$, E_y varies as

$$\psi_0 = \exp\left(-\frac{\rho^2}{w_0^2}\right),$$

where[6,10]

$$w_0^2 = \frac{w^2}{1 + (w^2\pi/\lambda R)^2}. \tag{21}$$

Since the maximum value of $|\psi_0'' - \psi_0'/\rho|$ occurs for $\rho = w_0$ and is

$$|\psi_0'' - \psi_0'/\rho|_{\max} = \frac{4}{ew_0^2}, \tag{22}$$

the ratio C between $|E_x|_{\max}$ and $|E_y|_{\max}$ is now given by

$$C = \frac{1}{e(kw_0)^2}. \tag{23}$$

To allow a direct comparison of eqs. (23) and (17), let us assume $1/R = 0$ and w given by eq. (19). From eqs. (19) and (23) we get

$$C = \frac{0.889}{(ka)^2}, \tag{24}$$

which gives values somewhat larger than eq. (17).

In the case of the experiment $a \cong 2.35\ \lambda$, $R = 9.78\ \lambda$. For these values of a, R, from eqs. (19), (21), and (23), we obtain, for C^2, -44 dB. This, of course, is only a rather crude estimate of the actual depolarization by the two cylinders (the mean square error in representing the actual distribution (9) by means of (18) is almost 2 percent; see Appendix B). For the present purpose, however, this estimate is quite adequate, since in the experiment other effects such as imperfections in the feed and some aperture blockage by the feed are found to be predominant.

III. EXPERIMENT

Two mirrors and a feed were constructed and assembled as shown in Fig. 3. Their radiation characteristics were measured at 18.5 GHz. The cylindrical surfaces of the two mirrors were milled to a tolerance

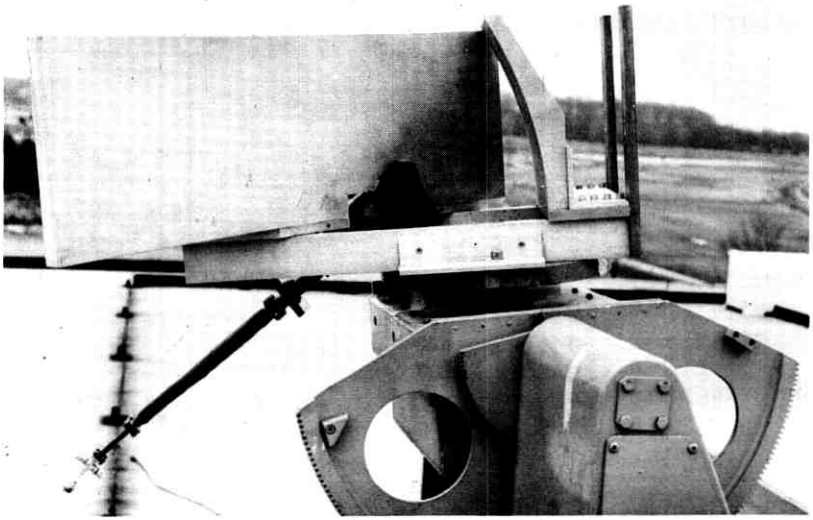


Fig. 3—Double cylinder antenna with corrugated feed.

of about 1 mil. The parabolas generating the two cylinders have the same focal length $\bar{f} = 12.919\ \lambda$.

The feed, a corrugated horn with

$$ka = 14.76, \qquad R = 4.17\ a, \tag{25}$$

satisfies the boundary condition (10) at 18.5 GHz, and therefore its radiation pattern is essentially circularly symmetric.[1,2] A measured pattern is shown in Fig. 4. Figure 4 also shows the pattern calculated for the gaussian distribution (18), with w given by eq. (19). The two patterns are somewhat different, as expected, since the distribution (18) represents only the fundamental term ($m = 0$) of eq. (35). A much closer agreement with the measured pattern could be obtained from eq. (35) by considering also the term relative to $m = 2$ (it turns out that the term $m = 1$ is negligible), but for the present purpose the accuracy of Fig. 4 is satisfactory.

If we consider only the fundamental term (18), the location of the phase center P_c and the beamwidth of the feed can be calculated straightforwardly by using eqs. (20) and (47); the 3-dB beamwidths of the antenna in the principal planes are given by eqs. (45) and (46) of Appendix B.

The distance z_0 of the feed aperture from the focal point F was chosen using eq. (20), in which case the phase center P_c of (18) coin-

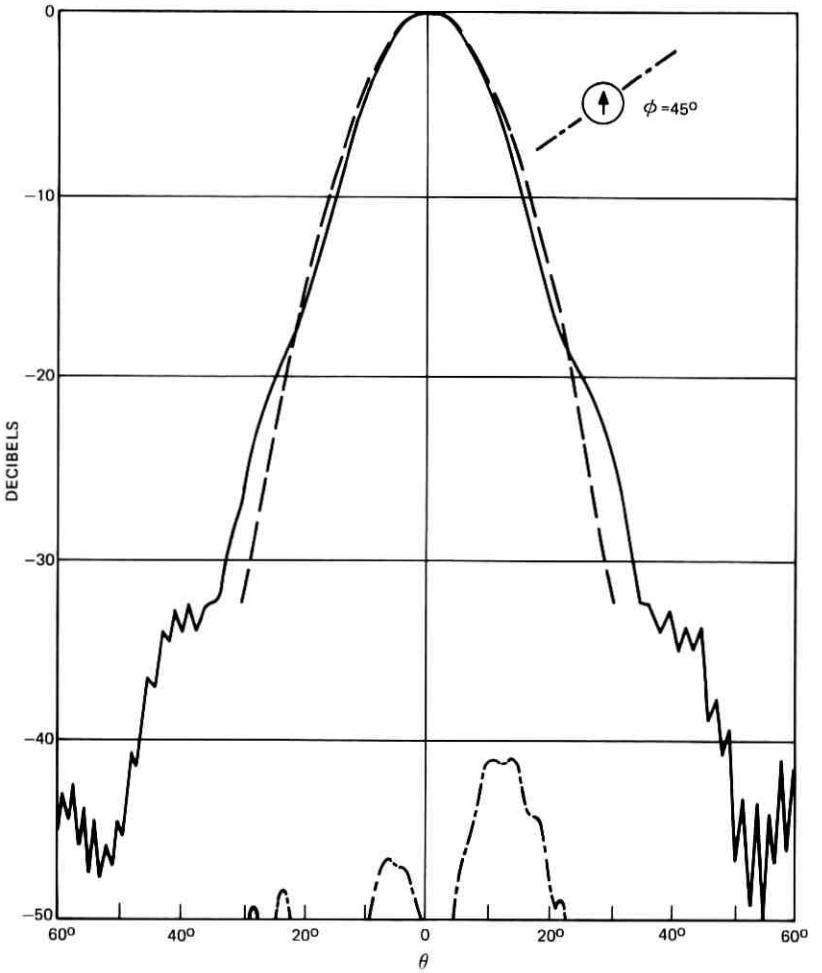


Fig. 4—Radiation patterns of vertically polarized corrugated feed measured at 18.5 GHz in the $\phi = 45°$ plane for the main component (——) and the cross-polarized component (–·–·–) of the far field. The curve (– – – –) is the pattern calculated for the gaussian distribution of eqs. (31) and (32).

cides with F. We can show, using eqs. (46) and (47), that under this condition the beamwidths of the antenna are stationary with respect to small displacements of the feed. The orientation of the feed with respect to the first cylinder was chosen as follows.

When the angle of incidence μ in Fig. 1 is 0, which is the condition assumed in Ref. 4, the axis of the beam reflected by the first cylinder

coincides with the axis of the feed. Because of the relatively large size of the feed aperture, this condition ($\mu = 0$) is undesirable, since a relatively large fraction of the energy reflected by the first cylinder is then intercepted by the feed aperture. Thus, in the experiment, a relatively large value (36°) was chosen for μ. For this value, the energy blocked by the feed is small. Using the distribution of eq. (18), we

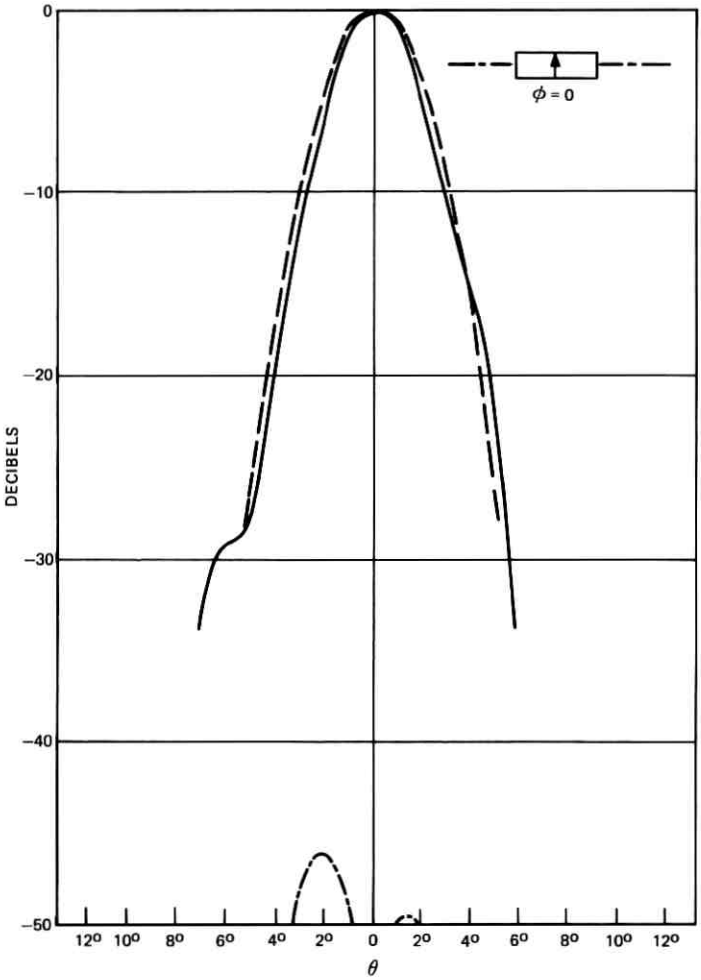


Fig. 5—Radiation patterns of vertically polarized double-cylinder antenna measured at 18.5 GHz in the horizontal plane $\phi = 0$ for the main component (——) and the cross-polarized component (–·–·–) of the far field. Curve (– – – –) is the pattern calculated from eq. (18) for the gaussian distribution of eqs. (31) and (32).

can show that the field incident on the upper edge of the feed is about −10 dB with respect to the field on the axis of the incident beam.

The first cylinder in Fig. 3 (14.91 λ by 14.32 λ) intercepts most radiation from the feed; the illumination of its four edges with respect to the illumination at the center is less than −16.5 dB on the upper edge and less than −24.5 dB on the other three edges.

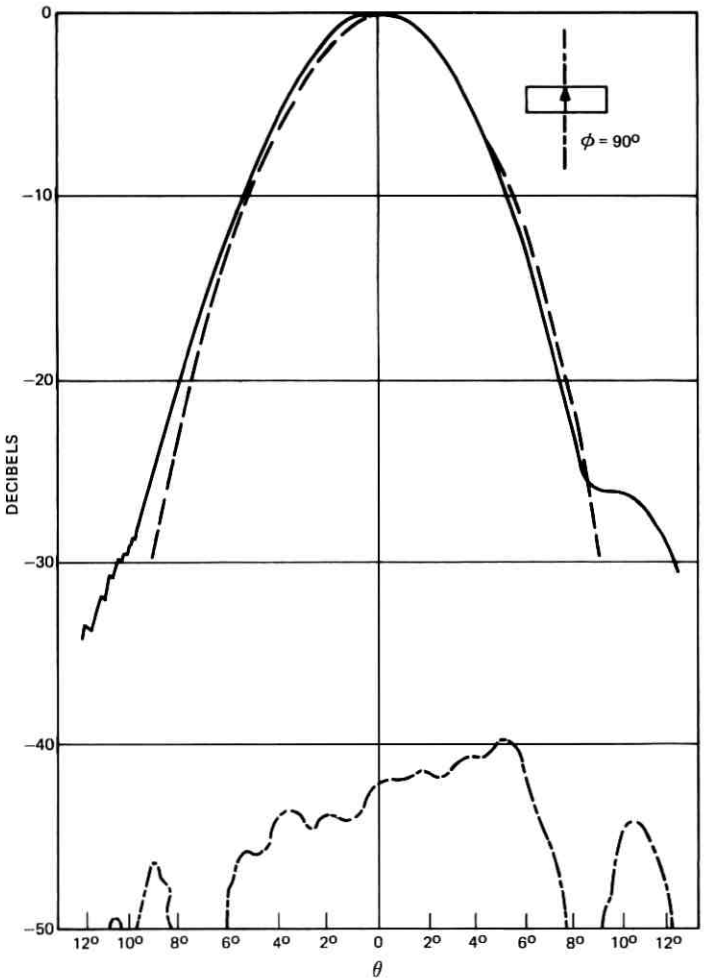


Fig. 6—Radiation patterns of vertically polarized double-cylinder antenna measured at 18.5 GHz in the vertical plane $\phi = 90°$ for the main component (——) and the cross-polarized component (–·–·–) of the far field. Curve (– – – –) is the pattern calculated from eq. (18) for the gaussian distribution of eqs. (31) and (32).

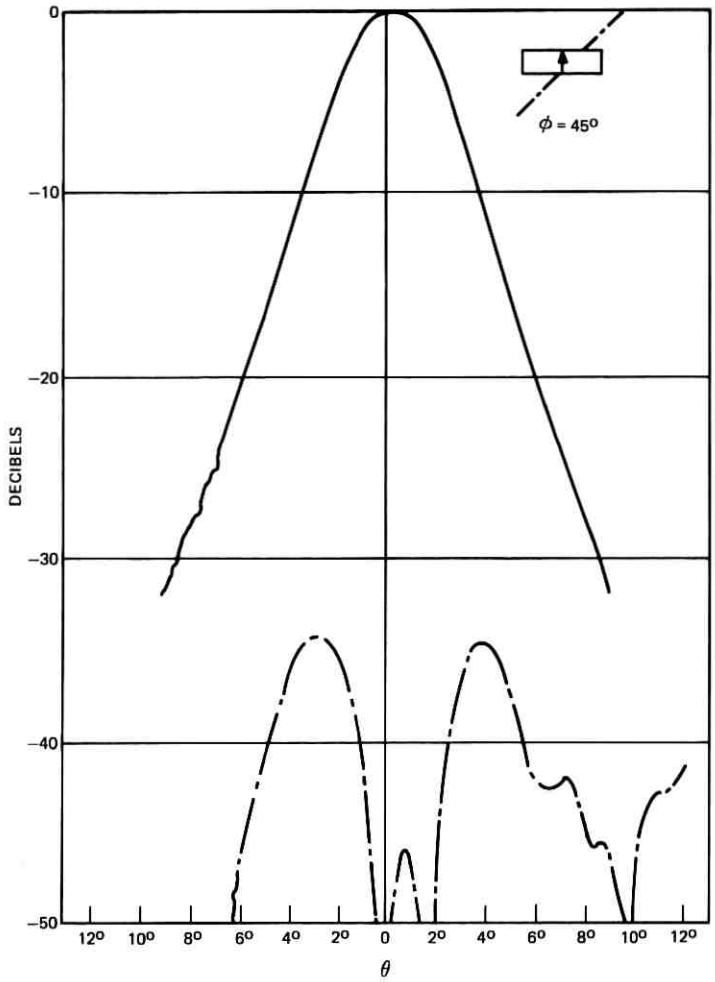


Fig. 7—Radiation patterns of vertically polarized double-cylinder antenna measured at 18.5 GHz in the $\phi = 45°$ plane for the main component (——) and the cross-polarized component (–·–·–) of the far field.

The second reflector (44.77 λ by 14.32 λ) is sufficiently large that it intercepts essentially all the energy reflected by the first cylinder, except for the energy blocked by the feed. The distances f_1, f_2 measured along the principal ray between F and the two reflectors are* 14.282 λ

* We can show (see Ref. 4) that

$$f_1 = \frac{2\hat{f}}{1 + \cos\mu}$$
$$f_2 = 2\hat{f}.$$

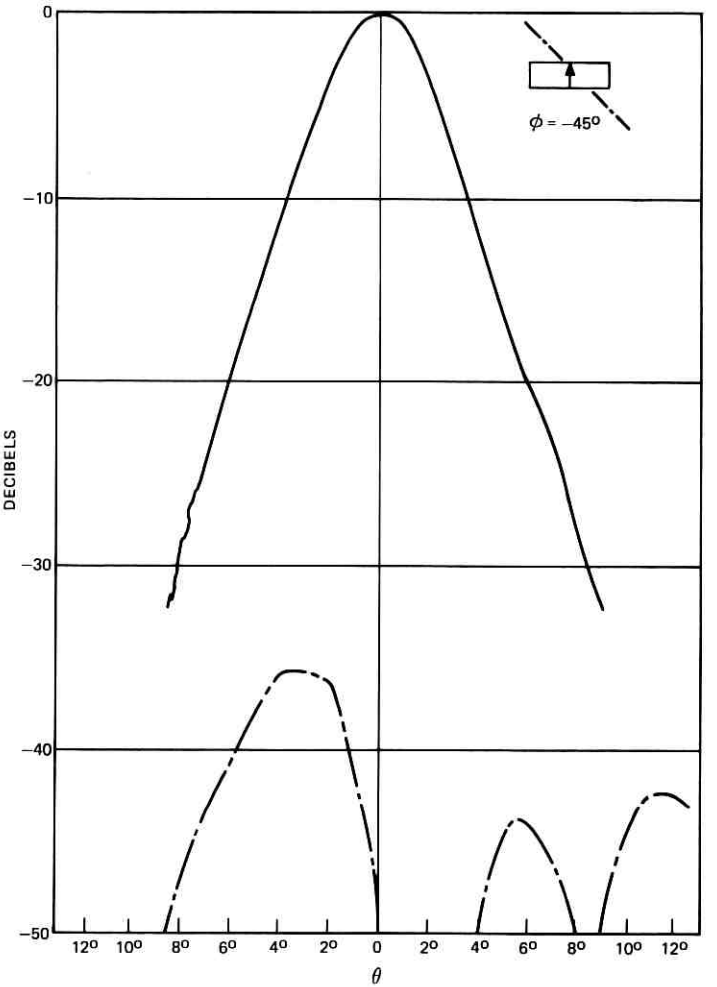


Fig. 8—Radiation patterns of vertically polarized double-cylinder antenna measured at 18.5 GHz in the $\phi = -45°$ plane for the main component (——) and the cross-polarized component (-·-·-) of the far field.

and 25.837 λ. Since $d_1 = f_1 + z_0$ and $d_2 = f_2 + z_0$, using eqs. (46) and (47), we obtain for the beamwidths in the two principal planes

$$2\theta_1 = 5.75°, \qquad 2\theta_2 = 3.18°.$$

The measured values (Figs. 5 and 6) are $2\theta_1 = 5.84°$ and $2\theta_2 = 2.87°$.

Figures 5 to 8 show the measured patterns in the principal planes $\phi = 0$, $\phi = 90°$ and in the planes $\phi = 45°$, $\phi = -45°$. In these figures, θ is the angle from the beam axis. Also shown in Figs. 5 and 6

are patterns calculated for the distribution (18). In all these cases, the feed is vertically polarized; the main patterns of Figs. 5 to 8 (the solid curves) give, therefore, the magnitude of the vertical component E_y, in dB with respect to the field on axis.

The patterns for the horizontal component E_x are given in Figs. 5 to 8 by the dashed curves. In the worst case, the plane $\phi = 45°$, the

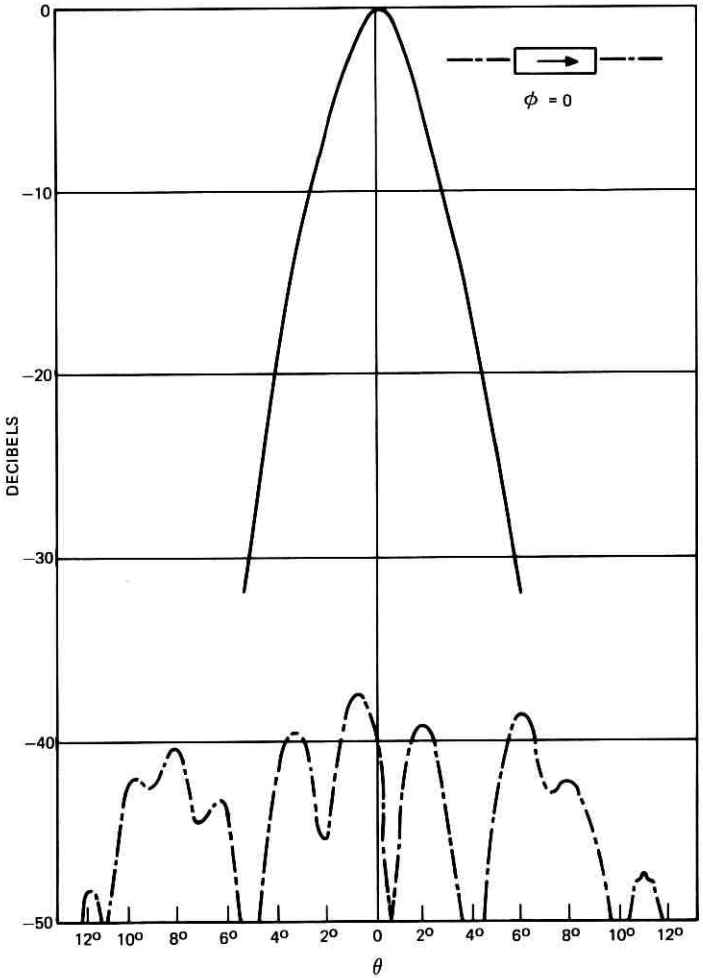


Fig. 9—Radiation patterns of horizontally polarized double-cylinder antenna measured at 18.5 GHz in the horizontal plane $\phi = 0$ for the main component (——) and the cross-polarized component (–·–·–) of the far field.

ratio C between the largest value of $|E_x|$ and the peak of $|E_y|$ is -33.5 dB, which is approximately 10.5 dB larger than the value given by eq. (23). This larger value of C is due in part to blockage by the feed and in part to imperfections in the feed that were found to cause a cross-polarized component in the feed radiation patterns with a peak of approximately -41 dB (see Fig. 4). The effect of blockage by the

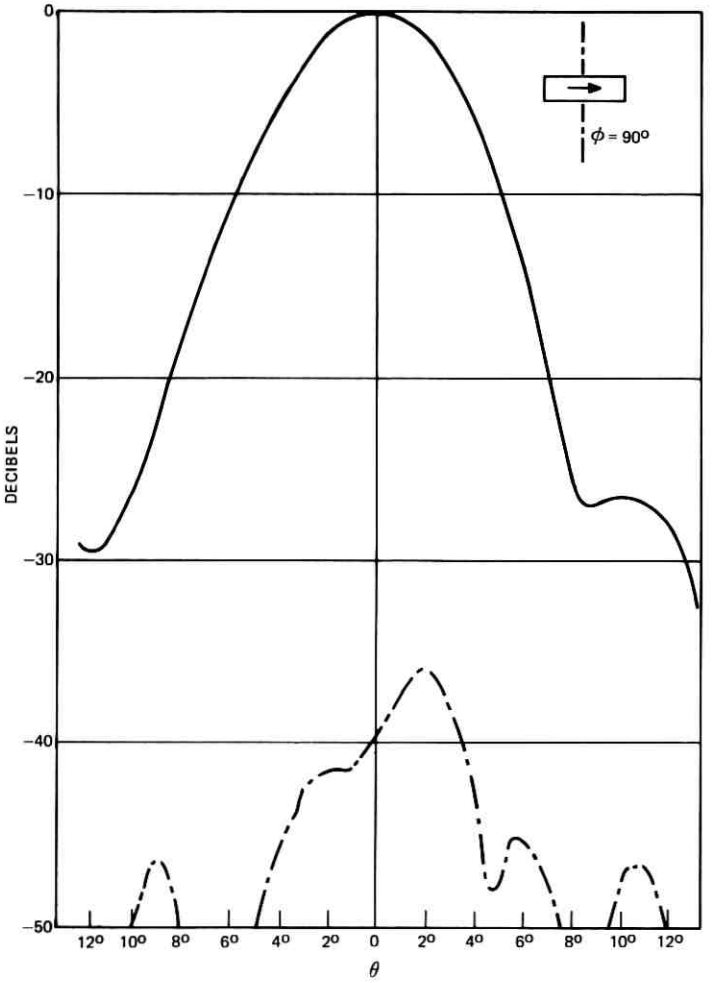


Fig. 10—Radiation patterns of horizontally polarized double-cylinder antenna measured at 18.5 GHz in the vertical plane $\phi = 90°$ for the main component (——) and the cross-polarized component (–·–·–) of the far field.

feed was evaluated by measuring the variation of C with μ. Decreasing μ from 36° to 31° caused an increase in C of approximately 3 dB.

The radiation patterns for a horizontally polarized feed are shown in Figs. 9 to 12. The cross-polarized component now has a peak of −35 dB in the worst case, the $\phi = 45°$ plane, which is about 1.5 dB lower than the value measured for vertical polarization.

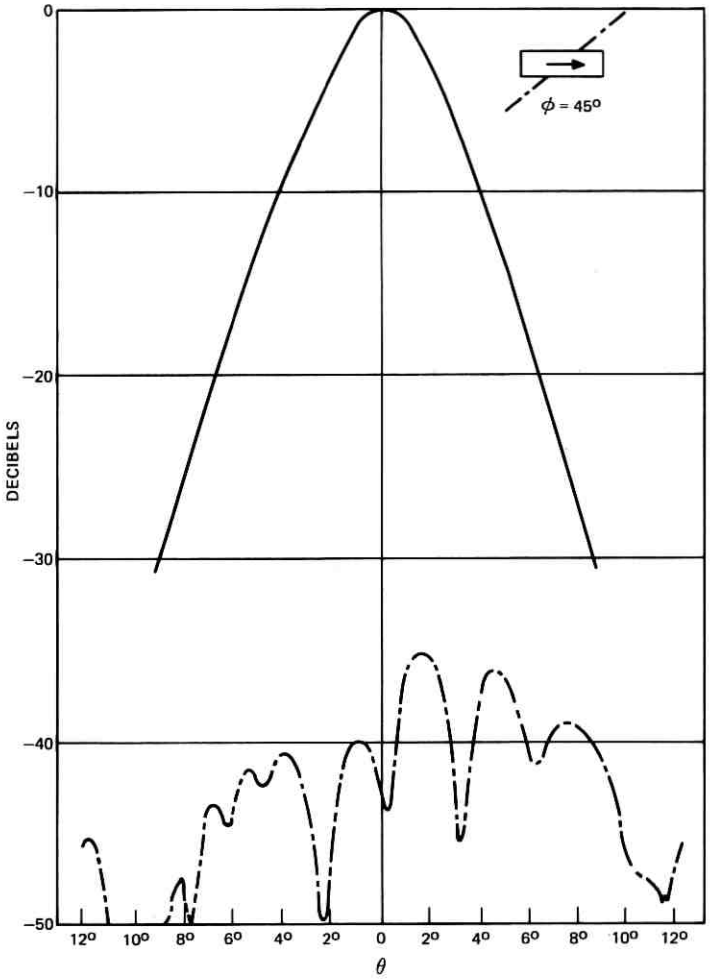


Fig. 11—Radiation patterns of horizontally polarized double-cylinder antenna measured at 18.5 GHz in the $\phi = 45°$ plane for the main component (——) and the cross-polarized component (–·–·–) of the far field.

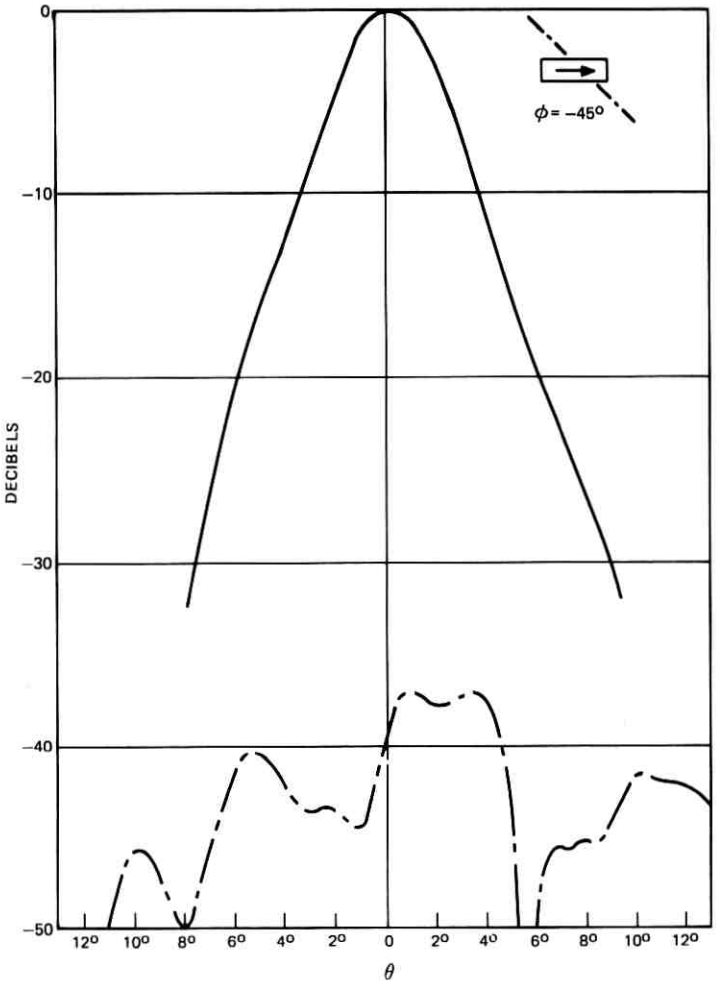


Fig. 12—Radiation patterns of horizontally polarized double-cylinder antenna measured at 18.5 GHz in the $\phi = -45°$ plane for the main component (——) and the cross-polarized component (–·–·–) of the far field.

In all the above measurements, the frequency was 18.5 GHz. Measurements at 29 GHz were also made to determine the frequency dependence of the two beamwidths $2\theta_1$, $2\theta_2$. At 29 GHz the measured values of $2\theta_1$ and $2\theta_2$ with the feed vertically polarized were $2\theta_1 = 6°$ and $2\theta_2 = 2.47°$. Thus, $2\theta_1$ increased by only 2.7 percent, with respect to the value measured at 18.5 GHz, while $2\theta_2$ decreased by 14 percent.

This difference in the variation of $2\theta_1$ and $2\theta_2$ is due to the lack of feed pattern symmetry as the frequency was increased. At 29 GHz in the H- and E-planes, $2\theta_{f1}$ was greater than $2\theta_{f2}$ by approximately 10 percent.

IV. CONCLUSIONS

Cylindrical reflectors are well adapted to efficiently generate an elliptical beam from the circularly symmetric beam radiated by a corrugated feed. Depolarization by the cylinders is negligible ($10 \log_{10} C^2 < -40$ dB, for $a > 1.25\ \lambda$) for typical feed diameters, and it is essentially independent of the angle of incidence μ which can therefore be chosen as large as needed to minimize aperture blockage.

In the experiment, blockage by the feed,* although small, was large enough to cause some deterioration in C^2. The measured value of C^2 in the worst case was -33.5 dB, approximately 10.5 dB higher than the value given by eq. (23). This higher value was due in part to certain imperfections in the feed.

A first-order analysis of the antenna was given in Section II. It was pointed out that, if the feed aperture is located close to the focal point F, then the beamwidths θ_1 and θ_2 vary little with frequency, assuming that each reflector intercepts all the energy incident on it. Simple approximate expressions [eqs. (45) and (46)] were given for θ_1 and θ_2. The measured values, $2\theta_1 = 5.8°$ and $2\theta_2 = 2.9°$ agree well with the values given by those equations.

V. ACKNOWLEDGMENTS

Thanks are due to Mrs. D. Vitello for programming the calculations of $\bar{u}$ and η_{max} of eqs. (38) and (39). The measurements were carried out by W. E. Legg.

APPENDIX A

This appendix derives the field transformation through the two lenses of Fig. 2. Consider first the one-dimensional case of Fig. 13, where it is assumed that E_y is a function of only x, z,

$$E_y = E_1(x, z)e^{jkz}.$$

Let the problem be to find E_y over the image plane $z = z_1$, with E_y

---

*The antenna considered here has an unusually large (≈ 2.75) ratio of focal distance to feed diameter (this is mainly due to the relatively large value required of θ_1). If a much narrower beam is desired, aperture blockage can be entirely eliminated without difficulty.

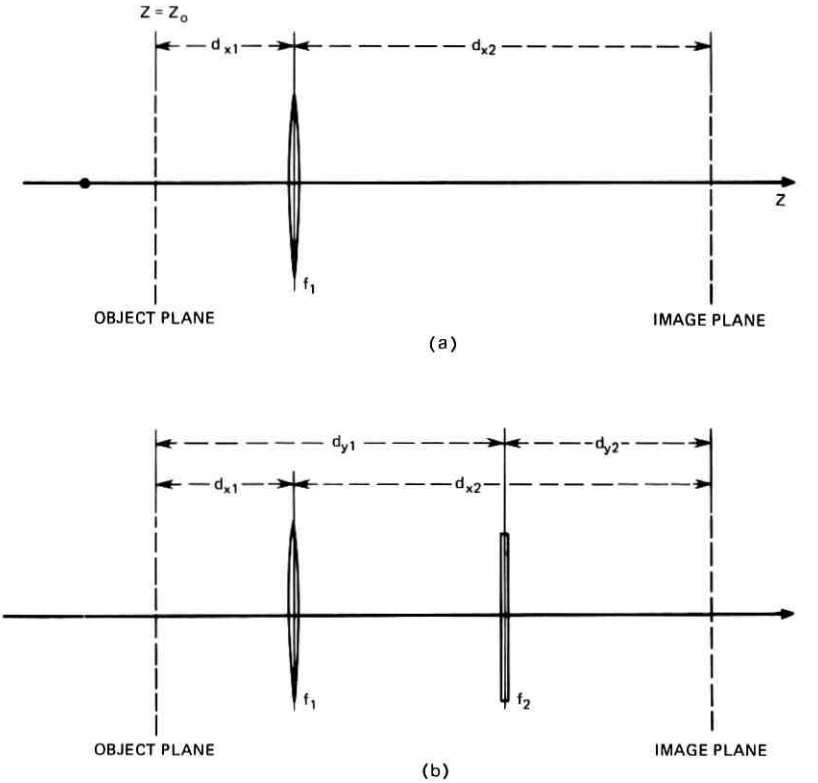


Fig. 13—Imaging (a) of one-dimensional field distribution by a cylindrical lens and (b) of two-dimensional field distribution by two orthogonal cylindrical lenses.

given [by some function $\psi_1(x)$] over the object plane $z = z_0$,

$$E_1(x, z) = \psi_1(x) \qquad \text{for} \quad z = z_0.$$

The distances d_{x1} and d_{x2} between the cylindrical lens and object and image planes are related by

$$\frac{1}{d_{x1}} + \frac{1}{d_{x2}} = \frac{1}{f_1}, \tag{26}$$

where f_1 is the focal length of the lens. The field immediately to the left of the lens can be expressed in terms of the field on the object plane $z = z_0$ using the one-dimensional form of Fresnel's formula which, for $z < z_0 + d_{x1}$, gives

$$E_1(x, z) = \sqrt{\frac{1}{+j\lambda(z - z_0)}} \int_{-\infty}^{\infty} E_1(x_0, z_0) \exp\left[-jk \frac{(x - x_0)^2}{2(z - z_0)} \right] dx_0. \tag{27}$$

By multiplying the field immediately to the left of the lens by the factor

$$\exp\left(-jk\frac{x^2}{2f_1}\right),$$

we obtain the field immediately to the right, which can then be used (with the help of Fresnel's formula) to determine the field over the image plane $z = z_1$. The resulting expression gives the result

$$E_1(x_1, z_1) = \frac{1}{+j}\sqrt{\frac{d_{x1}}{d_{x2}}}\,E_1\left(-\frac{d_{x1}}{d_{x2}}x_1, z_0\right)\exp\left(jk\frac{x_1^2}{2f_1}\frac{d_{x1}}{d_{x2}}\right), \quad (28)$$

as we may verify by using the procedure of Ref. 6.

Next, consider the two-dimensional case of Fig. 13b where f_1, f_2 now satisfy eq. (26) and

$$\frac{1}{d_{y1}} + \frac{1}{d_{y2}} = \frac{1}{f_2}, \quad (29)$$

and assume that over the plane $z = z_0$, E_y is of the type

$$E_y(x, y, z) = \psi_1(x)\psi_2(y)e^{jkz_0}, \qquad \text{for} \quad z = z_0. \quad (30)$$

Then, using the two-dimensional form of Fresnel's formula we find that for any $z > z_0$ the distributions of E_y in the x- and y-directions can be treated separately, and each distribution can be determined using the one-dimensional form of Fresnel's formula. More precisely, E_y can be written in the form

$$E_y(x, y, z) = E_1(x, z)E_2(y, z)e^{jkz}, \quad (31)$$

and the relation between $E_1(x, z)$ and $E_1(x, 0)$ [or between $E_2(y, z)$ and $E_2(y, 0)$] can be found considering the one-dimensional problem of Fig. 13a (replacing, in the case of E_2, d_{x1}, d_{x2}, f_1 in Fig. 13a with d_{y1}, d_{y2}, f_2). Thus, using eq. (28), we find

$$E_y(x_1, y_1, z_1)e^{-jkz_1} = -\sqrt{\frac{d_{x1}d_{y1}}{d_{x2}d_{y2}}}\,E_y\left(-\frac{d_{x1}}{d_{x2}}x_1, -\frac{d_{y1}}{d_{y2}}y_1, z_0\right)$$
$$\cdot\exp\left(jk\frac{x_1^2}{2f_1}\frac{d_{x1}}{d_{x2}}\right)\exp\left(jk\frac{y_1^2}{2f_2}\frac{d_{y1}}{d_{y2}}\right). \quad (32)$$

Note that, because of eqs. (26) and (29), the object and image planes ($z = z_0$ and $z = z_1$, respectively) are conjugate planes. All rays emanating from a point (x_0, y_0) in the object plane intersect each other at the conjugate point (x_1, y_1) in the image plane where x_1 and y_1 are

related to x_0 and y_0 by the magnifications

$$M_x = \frac{x_1}{x_0} = -\frac{d_{x2}}{d_{x1}} \quad \text{and} \quad M_y = \frac{y_1}{y_0} = -\frac{d_{y2}}{d_{y1}}.$$

Equation (32), derived for the particular case of eq. (30), is valid also for arbitrary $E_y(x, y, z_0)$, as may be seen by writing $E_y(x, y, z_0)$ in the form

$$E_y(x, y, z_0) = \sum A_{mn}\psi_{1m}(x)\psi_{2n}(y)e^{jkz_0}, \tag{33}$$

and then applying eq. (32) to each term of this expression. Equation (32) can also be derived using eq. (15a) of Ref. 7. If one conjugate plane, $z = z_1$, is in the far field, eq. (32) gives eq. (7).

APPENDIX B

In this appendix the field radiated by the corrugated horn is represented in terms of Laguerre-gaussian modes of propagation.[10] Let the truncated Bessel function

$$\psi = \begin{cases} J_0(u), & |u| \leqq u_{a0} \\ 0, & |u| > u_{a0} \end{cases} \tag{34}$$

be developed into Laguerre-gaussian functions

$$\psi = \sum_{m=0}^{\infty} B_m L_m\left(\frac{u^2}{\bar{u}^2}\right) \exp\left(-\frac{u^2}{2\bar{u}^2}\right), \tag{35}$$

where L_m are Laguerre polynomials,

$$L_m(x) = \frac{e^x}{n!}\frac{d^n}{dx^n}(e^{-x}x^n), \tag{36}$$

and $\bar{u}$ is a parameter whose optimum choice is, for the present purpose, that which maximizes the ratio

$$\eta = \frac{B_0^2}{\sum_0^{\infty} B_k^2}.$$

We can show that η is maximum for

$$\bar{u} = 1.09549, \tag{37}$$

and that

$$\eta_{\max} = 0.9811. \tag{38}$$

We therefore assume this particular value of $\bar{u}$.

We now derive the field E_y for $z > 0$ subject to the boundary condition

$$E_y(x, y, 0) = \psi \exp\left[j \frac{k}{2R} (x^2 + y^2)\right] \tag{39}$$

for $z = 0$. We assume that Fresnel's formula is applicable, which implies that the wave equation $(\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2 + k^2)E_y = 0$ can be approximated by the parabolic wave equation

$$[k^2 + \tfrac{1}{2}(\partial^2/\partial x^2 + \partial^2/\partial y^2)]E_y = -jk\partial E_y/\partial z.$$

We can verify that the solution of the parabolic wave equation that satisfies the above boundary condition is given by[10]

$$E_y(x, y, z) = e^{jkz} \sum_0^\infty B_m C_m L_m\left(\frac{k(x^2 + y^2)}{A}\right) \exp\left[j \frac{k}{2q} (x^2 + y^2)\right], \tag{40}$$

where C_m, A, and q are functions of z,

$$q = q(z) = q(0) + z \tag{41}$$

$$\frac{1}{A} = \mathrm{IM}\left(\frac{1}{q}\right) \tag{42}$$

$$C_m = \sqrt{\frac{A_0}{A}} \left[\frac{\dfrac{z}{q(0)^*} + 1}{\dfrac{z}{q(0)} + 1}\right]^{m+\frac{1}{2}}, \tag{43}$$

and $A_0 = A$ for $z = 0$. The complex beam parameter $q(0)$ appearing in these formulas is determined by the feed characteristics,

$$\frac{1}{q(0)} = j \frac{1}{(ka)^2} \frac{ku_{a0}^2}{\bar{u}^2} + \frac{1}{R}. \tag{44}$$

We can verify by means of eqs. (35) and (40) to (44) that E_y satisfies condition (39) for $z = 0$.

In the experiment, the focal plane is behind the aperture of the feed. In this case, to derive the far field using eq. (7) we have to consider not the field radiated by the feed in the half space $z > 0$, but its virtual extension into the half space $z < 0$. We can show, however, that eq. (40) is valid also in this case (i.e., for $z = z_0 < 0$).

According to eq. (38), 98.11 percent of the total energy radiated by the feed is caused by the fundamental gaussian mode (18), which is the term $m = 0$ of eq. (35). Therefore, the horn is an efficient gaussian

beam launcher, and a rough estimate of the radiation characteristics of the feed and the antenna can be obtained considering only this $m = 0$ term. In this case using eqs. (7) and (8) we obtain, for the 3-dB beamwidths $2\theta_1$ and $2\theta_2$ of the antenna in the two principal planes,[10,11]

$$2\theta_1 = \sqrt{2 \ln 2}\,\frac{w_0}{f_1} \tag{45}$$

$$2\theta_2 = \sqrt{2 \ln 2}\,\frac{w_0}{f_2}, \tag{46}$$

with w_0 given by eq. (21). The beamwidth $2\theta_f$ of the feed pattern is[10]

$$2\theta_f \cong \frac{\lambda}{\pi w_0}\sqrt{2 \ln 2}. \tag{47}$$

The distance z_0 of the phase center from the feed aperture is[10] given by eq. (20).

The spot size* of the gaussian beam at the horn aperture is[10]

$$w = \sqrt{\frac{\lambda A_0}{\pi}} = \sqrt{2}\left(\frac{\bar{u}a}{u_{a0}}\right) \cong 0.6437\ a. \tag{48}$$

APPENDIX C

Derivation of Eq. (11)†

It can be shown[1,2] that the far field of the feed of Section 2.2 has, because of boundary condition (10), the following property: If the far field is transformed into a plane wave using an optical system of revolution centered around the feed axis, then the resulting plane wave is vertically polarized if the input of the feed is vertically polarized. We now show that this property implies condition (11).

Consider an optical system of revolution centered around the z-axis, and let (ρ, ϕ) be polar coordinates defined by $x = \rho \cos \phi$, $y = \rho \sin \phi$. Consider an input ray entering the optical system with direction parallel to the axis of symmetry and with polarization characterized by unit vector e_0. Let the meridional plane defined by this ray be the plane $\phi = \phi_0$ (see Fig. 14). Because of the symmetry of the system, this ray will be transformed into an output ray leaving the system in the same meridional plane $\phi = \phi_0$. Therefore, if (k_x, k_y, k_z) is the wave vector characterizing the direction of the output ray and θ_1 is the angle between (k_x, k_y, k_z) and the z-axis (see Fig. 14),

* Spot size is defined (Ref. 10) as the radius at which the field amplitude drops to $1/e$ of its value on axis.

† See Chapter VI of Ref. 12 for a justification of the approach used in this derivation.

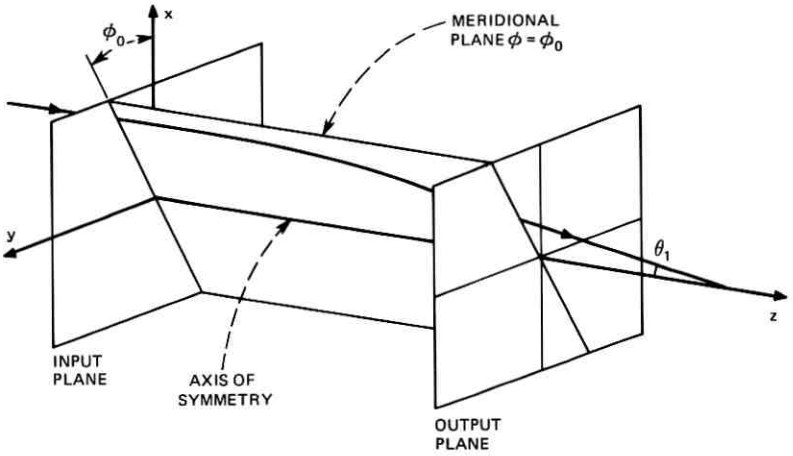


Fig. 14—Transformation by an optical system of revolution of an input ray incident parallel to the axis of symmetry.

then

$$\begin{aligned} k_x &= -k \sin \theta_1 \cos \phi_0 \\ k_y &= -k \sin \theta_1 \sin \phi_0 \\ k_z &= k \cos \theta_1. \end{aligned} \tag{49}$$

Now if e_1 is a unit vector characterizing the polarization of the output ray, there is a simple relation between e_1 and e_0 that can be obtained as follows. The ray in question, as it passes through the system, describes a plane curve since it remains in the plane $\phi = \phi_0$. It immediately follows that if e_0 lies in the plane $\phi = \phi_0$ then e_1 must lie in this plane also, while if e_0 is orthogonal to it, then e_1 is also. In other words,

$$\begin{aligned} e_1 &= i_\phi && \text{if} \quad e_0 = i_\phi \\ e_1 &= \cos \theta_1 i_\rho + \sin \theta_1 i_z && \text{if} \quad e_0 = i_\rho, \end{aligned}$$

where i_ϕ and i_ρ are unit vectors in the ϕ, ρ-directions. From these two relations it follows that, if the input ray is polarized in the y-direction,

$$e_0 = \sin \phi_0 i_\rho + \cos \phi_0 i_\phi,$$

then

$$e_1 = \sin \phi_0 (\cos \theta_1 i_\rho + \sin \theta_1 i_z) + \cos \phi_0 i_\phi. \tag{50}$$

Now the field along the output ray can be written in the form

$$E = e_1 A e^{j\psi},$$

where A is approximately constant and

$$\psi = k_x x + k_y y + k_z z. \tag{51}$$

We now determine the x, y-components of E using eq. (51) and the relations $i_\rho \cdot i_x = \cos\phi_0$, $i_\phi \cdot i_x = -\sin\phi_0$, $i_\rho \cdot i_y = \sin\phi_0$, $i_\phi \cdot i_y = \cos\phi_0$. The result is

$$E_x = \sin\phi_0 \cos\phi_0(\cos\theta_1 - 1)Ae^{j\psi}$$
$$E_y = [1 + \sin^2\phi_0(\cos\theta_1 - 1)]Ae^{j\psi}.$$

From these relations and eqs. (50) and (52), we can verify with little difficulty that for small θ_1

$$E_x \cong -\sin\phi_0 \cos\phi_0 \frac{\sin^2\theta_1}{2} Ae^{j\psi} \cong \frac{1}{2k^2} \frac{\partial^2 E_y}{\partial x \partial y}. \tag{52}$$

Now we assume that the angle θ_1, which is a function of the distance ρ_0 of the input ray from the z-axis, remains small for all ρ_0. Then, according to eq. (53), a plane wave polarized in the y-direction is transformed by the optical system into a wave satisfying eq. (11).

Equation (53) was derived by ray optics that apply in the far field of the feed. However, if a differential relation such as (52) holds in the far field, it must hold everywhere (e.g., in the aperture plane of the feed horn).

REFERENCES

1. H. C. Minett and B. MacA. Thomas, "A Method of Synthesizing Radiation Patterns with Axial Symmetry," IEEE Trans., *AP-14*, 1966, pp. 645–656.
2. V. H. Rumsey, "Horn Antennas with Uniform Power Patterns Around their Axes," IEEE Trans., *AP-14*, 1966, pp. 656–658.
3. T. S. Chu and R. H. Turrin, "Depolarization Properties of Off-Set Reflector Antennas," IEEE Trans. on Antennas and Propagation, *AP-21*, No. 3 (May 1973), pp. 339–345.
4. R. C. Spencer, F. S. Holt, H. M. Johanson, and J. Sampson, "Double Parabolic Cylinder Pencil-Beam Antenna," IRE Trans. on Antennas and Propagation, January 1955, pp. 4–8.
5. L. C. Tillotson, "Use of Frequencies above 10 GHz for Common Carrier Applications," B.S.T.J., *48*, No. 6 (July-August 1969), pp. 1563–1576.
6. H. Kogelnik, "Imaging of Optical Modes-Resonators with Internal Lenses," B.S.T.J., *44*, No. 3 (March 1965), pp. 455–494.
7. J. A. Arnaud, "Nonorthogonal Optical Waveguides and Resonators," B.S.T.J., *49*, No. 9 (November 1970), pp. 2311–2347.
8. G. A. Deschamps, "Gaussian Beam as a Bundle of Complex Rays," Electron. Lett., *7*, No. 23 (1971), pp. 684–685.
9. P. J. B. Clarricoats and P. K. Saha, "Propagation and Radiation Behaviour of Corrugated Feeds," Proc. IEE, *118*, No. 9 (September 1971), pp. 1167–1176.
10. H. Kogelnik and T. Li, "Laser Beams and Resonators," Appl. Opt., *5*, No. 10 (October 1966), pp. 1550–1567.
11. G. Goubau, "Optical Relations for Coherent Wave Beams," *Electromagnetic Theory and Antennas*, E. C. Jordan, ed., New York: Pergamon Press, 1963.
12. R. K. Luneburg, *Mathematical Theory of Optics*, Berkeley: University of California Press, 1966.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 7, September 1974
Printed in U.S.A.

Transverse Coupling in Fiber Optics Part III: Bending Losses

By J. A. ARNAUD

(Manuscript received February 20, 1974)

A general method is proposed to evaluate the radiation loss of bent open waveguides. This method consists in evaluating the coupling between the waveguide mode and the whispering-gallery modes that can be associated with the surrounding medium. The expression obtained for a reactive surface coincides with a previous result by Miller and Talanov.[1] We investigate in detail the radiation loss of the fundamental (HE_{11}) mode of a dielectric rod coupled to a slab. This arrangement, described in Part II of this article series,[2] provides a useful mode-selection technique. The radiation loss is given by a simple closed-form expression. We find that for a wavelength of 1 μm and a rod radius of 5 μm in physical contact with the slab, the bending loss is less than 1 dB/km if the radius of curvature, in the plane of the slab, exceeds 22 mm.

I. INTRODUCTION

Open waveguides support modes whose phase velocity is smaller than the velocity of plane waves in the surrounding medium. Thus, no radiation takes place under normal conditions. If the fiber is bent, however, the phase velocity increases in proportion to the distance from the curvature center. At some radius, it exceeds the velocity of plane waves in the medium and a radiation loss is suffered. This effect is of great practical importance in fiber communication because it sets a limit on how sharp bends can be made without resulting in untolerable loss. For most single-mode optical-glass fibers, a radius of curvature of the order of 1 cm can be tolerated. For gas lenses and weakly guiding millimeter-wave systems, the minimum radius is sometimes as large as 100 m. The relative insensitivity of glass fibers to bends results from the rather large change in refractive index in the cross section that they provide. It constitutes their main advantage compared to other guiding systems.

Different methods have been used to evaluate the radiation losses of curved waveguides. For simple geometries, it is possible to solve the boundary value problem using a cylindrical coordinate system, the loss being given by the imaginary part of the propagation constant. This method was used by Miller and Talanov[1] for a reactive surface, and by Marcatili[3] for dielectric slabs with various imbedding materials and for waveguides with rectangular cross section. Another method consists in evaluating the power radiated at the radius where the phase velocity of the guided mode becomes equal to the velocity in the surrounding medium. This approach was followed by Shevchenko[4] and Marcuse.[5] Ray pictures have also been used to describe the propagation in curved dielectric fibers.[6]

In most practical situations, the fiber curvature is not a constant but varies along the fiber axis (z). This results in an additional radiation loss suffered at the transition between straight and curved sections of the fiber. This transition effect has been discussed in detail by Shevchenko[7] for the case of a reactive surface. Because the radiation from the bend itself has the form $\exp(-\rho/\rho_o)$, where ρ denotes the radius of curvature and ρ_o a constant, and the radiations at the junctions have the form $1/\rho^2$, the latter becomes significant for large ρ, that is, for very low radiation losses. It should be noted, however, that the mechanical strength of a fiber usually prevents an abrupt curve. The transition radiation, therefore, is usually unimportant for single-mode communication fibers (we assume that the fiber is otherwise perfect). It is more important that, when a highly multimoded fiber is bent in some random manner, the ray slopes associated with the guided waves tend to increase in proportion to the square root of the length of the fiber. Eventually, the rays cease to be totally reflected. This is easily understood if we observe that randomly bent fibers are analogous to mechanical oscillators (e.g., harmonic oscillators for the case of square-law fibers) driven by random forces $f(t)$ proportional to the curvature of the fiber $C(z)$.[8–10] The equivalent mechanical oscillator gains energy as time goes on; that is, the amplitude and momentum increase until the limit is reached. Note that, even in nominally single-mode fibers, higher-order modes can be excited and can propagate over a certain length (see Part II of this article for numerical values[2]) and cause significant pulse spreading. This problem will not be discussed further here. In what follows, we assume that the fiber curvature is a constant. This is the case if the fiber is wound on a cylindrical drum. Our calculation, therefore,

constitutes a first step in the evaluation of the losses of randomly bent fibers.

To evaluate the radiation loss of curved fibers, we shall use an approach akin to the one used in Part II of this article series.[2] For curved geometries, the guided mode is coupled to the whispering-gallery type of radiation modes rather than to sinusoidal standing waves. We first assume that the surrounding medium, which we will call a substrate, is bounded at some radius ρ_b and has finite dissipation losses. The loss suffered by the waveguide mode as a result of its coupling to the substrate is evaluated. Radius ρ_b is subsequently allowed to reach infinity and the dissipation loss to vanish.

These whispering-gallery modes are characterized by a circular caustic with radius ρ_s. The behavior of the field is oscillatory outside the caustic and exponential inside the caustic. At the caustic, the phase velocity is just equal to the phase velocity of plane waves. It is not difficult, therefore, to find for what values of ρ_s synchronism with the waveguide mode is achieved. Although ρ_s exceeds only slightly the rod curvature radius ρ, the distance $y_o \equiv \rho_s - \rho$ is the most critical parameter that influences the loss. Once the synchronism conditions have been obtained, we evaluate the coupling coefficients and the mode-number densities using the method described in Part II. We shall use for the waveguide field an expression applicable to scalar fields (described in Part I of this series[2]), which is slightly simpler than the expression applicable to the Maxwell field.

Formulas applicable to arbitrary open waveguides are obtained in Section II. We investigate in Sections III and IV the case of round fibers whose dimensions are large compared with the wavelength, both in free space and coupled to slabs providing mode selection. The latter configuration is closely related to the single-material fiber proposed by Kaiser, Marcatili, and Miller,[11] and to slab-coupled guides.[12] Simple closed-form formulas are obtained for the bending losses in all cases.

II. GENERAL FORMULA

Let us consider an opened waveguide with propagation constant h and radius of curvature ρ as shown in Fig. 1. The general expression for the radiation loss given in Part II is

$$\mathfrak{L} = \pi N C^2, \tag{1}$$

where N denotes the mode number density in the substrate (or surrounding medium), that is, the number of substrate modes whose

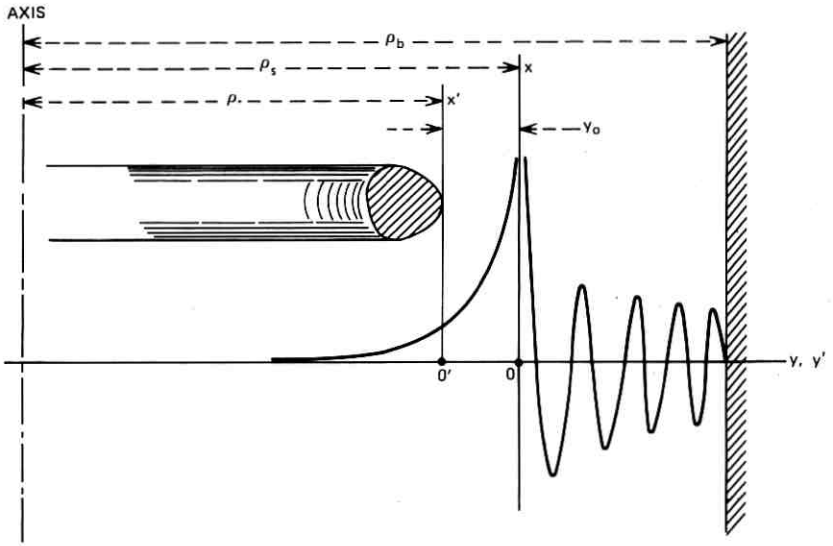


Fig. 1—Curved fiber. The radiation loss is obtained by evaluating the coupling between the straight fiber mode and the whispering-gallery mode shown with the caustic (or turning point) at ρ_s.

propagation constant lies between h and $h + dh$ at the waveguide radius ρ, divided by dh. The term C^2 is the coupling coefficient between the waveguide mode (fields E, H) and a synchronous substrate mode (fields E_s, H_s). For the Maxwell field (see Part I) we have

$$C = \frac{1}{2}\int_{-\infty}^{+\infty} (E_x H_{sz} + E_z H_{sx} - E_{sx} H_z - E_{sz} H_x) dx, \qquad (2a)$$

if the powers are normalized to unity; that is, if

$$\int \mathbf{E} \times \mathbf{H} \cdot d\mathbf{S} = \int \mathbf{E}_s \times \mathbf{H}_s \cdot d\mathbf{S} = 1. \qquad (2b)$$

The integral in eq. (2a) can be evaluated, for instance, along the x' axis shown in Fig. 1.

When the transverse variations of permittivity are small, the scalar parabolic (Fock) approximation is usually applicable. In that case (see Part I),

$$C = \frac{1}{2}\int_{-\infty}^{+\infty} (E \partial E_s/\partial y - E_s \partial E/\partial y) dx, \qquad (3a)$$

with

$$\int kE^2 dS = \int k_s E_s^2 dS = 1. \tag{3b}$$

The terms k and k_s denote propagation constants in the media.* In eq. (1) the product NC^2 is to be evaluated in the limit where the boundaries imposed on the substrate extend to infinity. In that limit $N \to \infty$, and $C^2 \to 0$, but NC^2 remains finite.

The waveguide mode, with propagation constant h, is synchronous to a whispering-gallery mode with propagation constant h_s in the y,z plane, whose caustic radius ρ_s is given by

$$h\rho = h_s \rho_s. \tag{4a}$$

When the waveguide field is uniform along the x axis, h_s is simply equal to k, the bulk propagation constant, and, in free space, $h_s = k = \omega/c$.

More generally, if k_x denotes the x component of the wave vector, then

$$h_s = (k^2 - k_x^2)^{\frac{1}{2}}. \tag{5}$$

The distance y_o between the waveguide and the caustic is from eq. (4a):

$$y_o \approx \rho(1 - h_s/h). \tag{4b}$$

The field of whispering-gallery modes is easily obtained from ray-optics considerations, as illustrated in Fig. 2, in the paraxial JWKB approximation. For a detailed discussion the reader should see Ref. 13. We obtain

$$E_s(y) = \begin{cases} y^{-\frac{1}{4}} \sin\left[(2^{\frac{1}{2}}/3)h_s\rho_s^{-\frac{1}{2}}y^{\frac{3}{2}} + \pi/4\right], & y > 0 \\ \frac{1}{2}(-y)^{-\frac{1}{4}} \exp\left[-(2^{\frac{1}{2}}/3)h_s\rho_s^{-\frac{1}{2}}(-y)^{\frac{3}{2}}\right], & y < 0. \end{cases} \tag{6a}$$

These expressions are asymptotic forms of

$$E_s(y) = \pi^{\frac{1}{2}}(2h_s^2/\rho_s)^{1/2}\mathrm{Ai}[(2h_s^2/\rho_s)^{\frac{1}{3}}y], \tag{6b}$$

where (Ai) denotes an Airy function of the first kind, for large $|y|$. Alternatively, these expressions can be obtained from asymptotic forms of Bessel's functions with large arguments and orders [see, for example, the first eq. (42) in Ref. 3]. With sufficient accuracy we have

$$\int E_s^2(y)dy \approx \frac{1}{2}\int_0^{\rho_b - \rho_s} y^{-\frac{1}{2}}dy = (\rho_b - \rho)^{\frac{1}{2}}, \tag{7}$$

* Within the Fock approximation, the difference between k and k_s can be ignored in evaluating the powers in eq. (3b). However, for improved accuracy, we retain the distinction when n is not close to unity (e.g., $n = 1.41$).

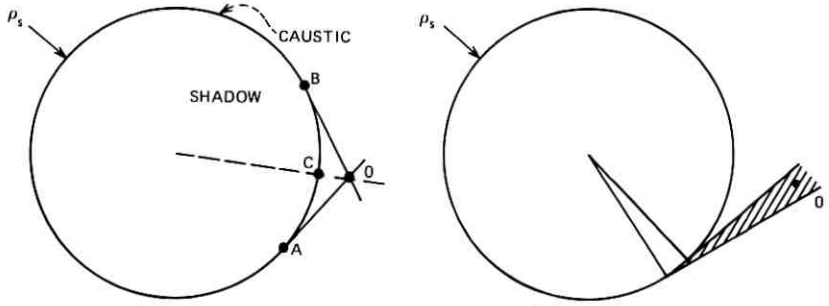


Fig. 2—Illustration of how the field of whispering-gallery modes given in eq. (6) is obtained in the JWKB approximation. (a) The argument of the sine function is h_s times the difference in length between the ray AO and the caustic arc AC. (b) The amplitude term is obtained from a power conservation argument applied to the ray pencil.

because only the oscillatory part of $E_s(y)$ contributes significantly to the total power, $\langle \cos^2 x \rangle \approx \frac{1}{2}$, and $\rho_b - \rho_s \approx \rho_b - \rho$ in the limit $\rho_b \to \infty$.

The reciprocal of the mode-number density N is the change in propagation constant at the waveguide radius ρ corresponding to a change in the argument of the cosine function in eq. (6) equal to π. We find by differentiation of this argument

$$N = (2^{\frac{1}{2}}/\pi)\rho^{\frac{1}{2}}(\rho_b - \rho)^{\frac{1}{2}}. \tag{8}$$

Thus, if we characterize the y dependence of the radiation field by a function

$$S(y) \equiv \pi N E_s^2(y) \Big/ \left[\int k_s E_s^2(y) dy \right], \tag{9a}$$

we obtain, with the help of eqs. (6), (7), and (8),

$$S(y) = (2^{\frac{1}{2}}/4)\rho^{\frac{1}{2}}(-y)^{-\frac{1}{2}}k^{-1} \exp\left[-(2^{\frac{1}{2}}/3)h\rho^{-\frac{1}{2}}(-y)^{\frac{3}{2}}\right], \quad y < 0. \tag{9b}$$

In eq. (9b) we have replaced ρ_s by ρ and h_s by h. Therefore, the value of $S(y)$ at $y = -y_o$, where y_o is given in eq. (4b), is

$$S(y_o) = (2^{\frac{1}{2}}/4)(1 - h_s/h)^{-\frac{1}{2}}k^{-1} \exp\left[-(2^{\frac{1}{2}}/3)h\rho(1 - h_s/h)^{\frac{3}{2}}\right]. \tag{10a}$$

This expression is the asymptotic form of

$$S(y_o) = 2^{\frac{1}{2}}\pi h^{-1}(2h^2\rho^2)^{1/6}\mathrm{Ai}^2\left[(2h^2\rho^2)^{\frac{1}{3}}(1 - h_s/h)\right]. \tag{10b}$$

Because we are limiting ourselves to small delays, $1 - h_s/h \ll 1$, we have

$$p \equiv (h^2 - h_s^2)^{\frac{1}{2}} \approx 2^{\frac{1}{2}}h(1 - h_s/h)^{\frac{1}{2}}. \tag{11}$$

Note, for later use, that at the waveguide radius ρ, $E_s^{-1}(\partial E_s/\partial y)$ is equal to p. Using eq. (11), eq. (10) can be written in the simpler form

$$S(y_o) = (\tfrac{1}{2}p) \exp [-\rho/(\tfrac{3}{2}p^{-3}h^2)]. \tag{12}$$

Let us now consider a waveguide structure, such as a reactive surface or a slab, uniform along the x axis, and fields that are independent of x. Because the field of the waveguide has an exp $(-py)$ dependence on y, the general expression for the radiation loss, eqs. (1) and (3), is

$$\mathfrak{L} = p^2 \left[E^2 \Big/ \int_{-\infty}^{+\infty} kE^2 dy \right] S(y_o), \tag{13}$$

$S(y_o)$ being given in eq. (12).

For a reactive surface with normalized susceptance p (see Fig. 3), we have

$$E^2 \Big/ \int_{-\infty}^{+\infty} kE^2 dy = 2p/k. \tag{14}$$

Therefore, the radiation loss (in nepers/unit length) is, substituting eqs. (12) and (14) in eq. (13),

$$\mathfrak{L} = (p^2/k) \exp (-\rho/\rho_o), \tag{15a}$$

where

$$\rho_o \equiv \tfrac{3}{2}k^2/p^3. \tag{15b}$$

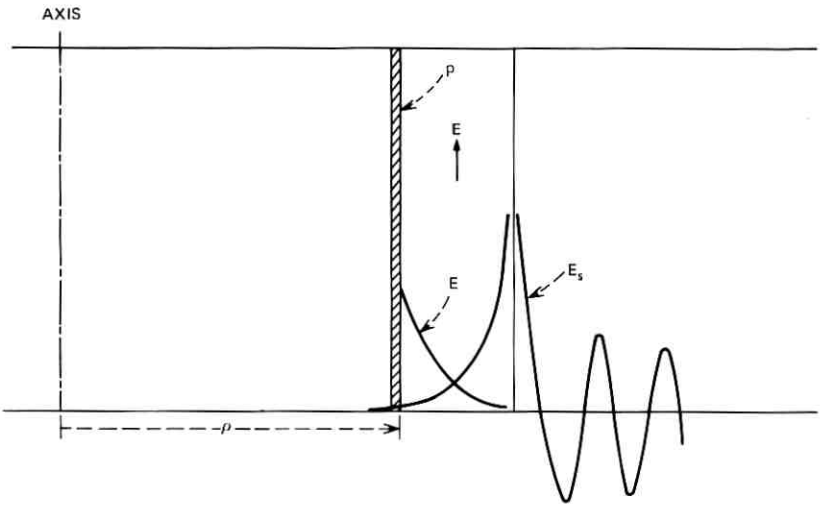


Fig. 3—Radiation loss of a curved reactive surface with normalized susceptance p.

This expression coincides with a previous result by Miller and Talanov.[1*]

For arbitrary waveguides with field $E(x)$, we need consider radiation modes for all values of k_x and not just $k_x = 0$ as before. A simplification comes from the fact that, as k_x increases, the caustic moves away from the waveguide and the coupling becomes negligible for some small value of k_x.

Let us introduce boundaries at $x = \pm L/2$ and assume symmetry for simplicity. The bending loss associated with some k_x is

$$\mathfrak{L}(k_x) = S(y_o, k_x)R(k_x), \tag{16}$$

where $R(k_x)$ is essentially the spatial spectrum of the waveguide field

$$R(k_x) = \left[\int_{-\infty}^{+\infty} \tfrac{1}{2}(p_o E - \partial E/\partial y) \cdot \cos(k_x x)dx\right]^2 \Big/ \left(\frac{L}{2}\int knE^2 dxdy\right), \tag{17}$$

where

$$p_o \equiv (h^2 - k^2)^{\frac{1}{2}} \approx 2^{\frac{1}{2}}h(1 - k/h)^{\frac{1}{2}}. \tag{18}$$

The quantity $S(y_o, k_x)$ given in eq. (10) is a function of k_x through the dependence of h_s on k_x, and is given, according to eq. (5), by

$$h_s = (k^2 - k_x^2)^{\frac{1}{2}} \approx k - \tfrac{1}{2}k_x^2/k. \tag{19}$$

To first order in k_x we have

$$(2^{\frac{1}{2}}/3)h\rho(1 - h_s/h)^{\frac{3}{2}} \approx \rho/\rho_o + (\rho p_o/k^2)k_x^2. \tag{20}$$

This approximation supposes $k_x \ll p_o$ or $\rho \gg \rho_o$, where ρ_o is the critical radius given in eq. (15b). These assumptions are acceptable for long fibers, where only very small bending losses can be tolerated. The dependence of h_s on k_x is important only in the exponential term. Summing $\mathfrak{L}(k_x)$ in eq. (16), with $S(y_o, k_x)$ given by eqs. (10) and (20), over k_x we obtain

$$\begin{aligned}\mathfrak{L} &= (L/2\pi)\int_0^\infty \mathfrak{L}(k_x)dk_x \\ &= (1/4\pi p_o)\exp(-\rho/\rho_o) \times \int_0^\infty LR(k_x)\exp(-\rho p_o k^{-2}k_x^2)dk_x. \end{aligned}\tag{21}$$

$\rho_o \equiv \tfrac{3}{2}k^2/p_o^3.$

*The results in Ref. 1 are applicable to anisotropic bent surfaces and arbitrary k_x. For our case, the expression for the loss $\mathfrak{L}$ reads [as in Ref. 1, eq. (13.18)]

$$\rho\mathfrak{L} = \gamma\alpha^2(1 + \alpha^2)^{-\frac{1}{4}}[\alpha + (1 + \alpha^2)^{\frac{1}{2}}]\uparrow[-2(1 + \alpha^3)^{\frac{1}{2}}\gamma]\exp(2\alpha\gamma),$$

where $\uparrow$ means "to the power," $\gamma \equiv kp$ and $\alpha \equiv p/k$. This expression goes to eq. (15) when $\alpha \to 0$, as one can (but not too easily) verify.

This is a general expression for the bending loss of open waveguides. Note that some of the assumptions could be relaxed with little additional complication. The integration over k_z, for instance, could be performed without using the approximation in eq. (20), and we could use the Maxwell field instead of the scalar field.

III. BENDING LOSS OF A DIELECTRIC ROD

We consider in this section the fundamental HE_{11} mode of a dielectric rod with refractive index n and radius a, with $a \gg \lambda$ (see Fig. 4). The asymptotic form of the field, given in Part II, is

$$E(r) = J_0(gr), \quad r < a, \tag{21a}$$

with

$$\begin{aligned} J_0(ga) &= 0, \\ ga &\equiv (k^2n^2 - h^2)^{\frac{1}{2}}a = 2.4\cdots \\ h &\approx kn - 2.87/kna^2. \end{aligned} \tag{21b}$$

Therefore,

$$\int E^2 2\pi r dr \approx 2\pi \int_0^a J_0^2(gr) r dr = 0.83a^2. \tag{22}$$

By specifying the continuity of the first derivative of the field, we obtain the field near the boundary (which is small but not strictly zero):

$$E(x, y) \approx (1.25/p_o a) \exp(-p_o x'^2/2a) \exp(-p_o y'). \tag{23}$$

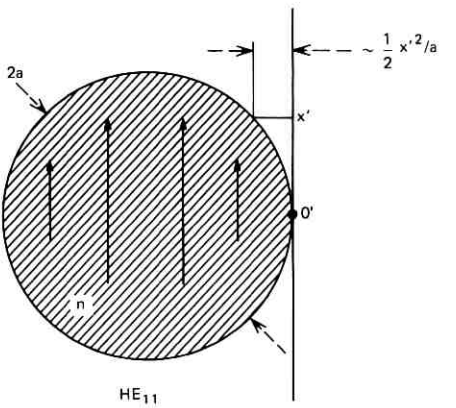


Fig. 4—Field of the fundamental HE_{11} mode of a dielectric rod. The important quantity is the field averaged over the x' axis. Only the close neighborhood of point 0 contributes significantly to the radiation loss.

The parameter $R(k_x)$ defined in eq. (17) is, in the present case,

$$LR(k_x) = (3.75/kna^4)\left[\int_{-\infty}^{+\infty} \exp(-p_o x^2/2a)\cos(k_x x)dx\right]^2. \quad (24)$$

Using the equality

$$\left[\int_{-\infty}^{+\infty} \exp(-\alpha x^2)\cos(k_x x)dx\right]^2 = (\pi/\alpha)\exp(-k_x^2/2\alpha),$$

we obtain

$$LR(k_x) = (7.5\pi/p_o kna^3)\exp(-ak_x^2/p_o). \quad (25)$$

The total loss is obtained by substituting this result in eq. (21) and integrating

$$\mathfrak{L} = (1.6/p_0^2 kna^3)(\rho p_o k^{-2} + a/p_o)^{-\frac{1}{2}}\exp(-\rho/\rho_o). \quad (26a)$$

The term ρ_o is the same as in eq. (15b). The exponential term coincides with the exponential term in Ref. 3 and is applicable to fibers with rectangular cross section. For $n = 1.01$, $\lambda = 1$ μm, and $a = 10$ μm, we obtain

$$\mathfrak{L} \approx 18 \times 10^6(\rho + 480)^{-\frac{1}{2}}\exp(-\rho/80)\ \text{dB/km}, \quad (26b)$$

where ρ is expressed in micrometers.

The loss in dB/km is plotted in Fig. 5 as a function of ρ for $\lambda = 1$ μm, $n = 1.01$, and $a = 10$ μm. The bending loss is considerably

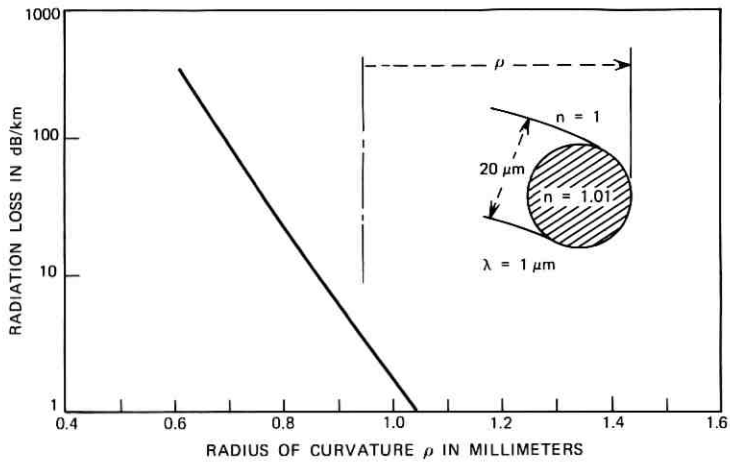


Fig. 5—Bending loss in dB/km for a round fiber with radius $a = 10$ μm and $n = 1.01$ in free space, as a function of the radius of curvature ρ.

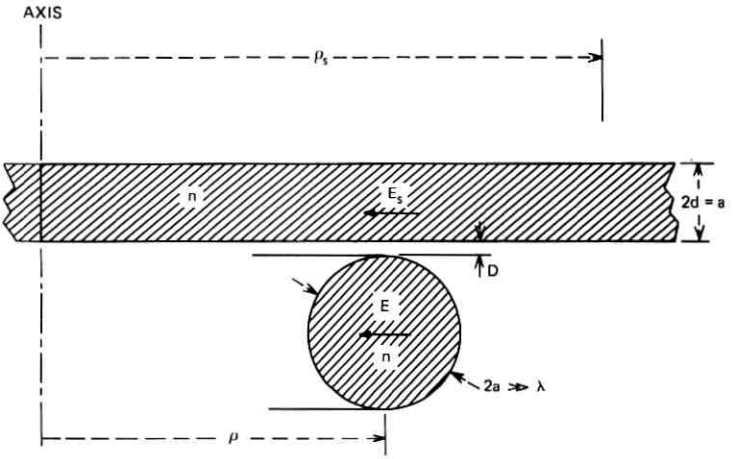


Fig. 6—Slab-loaded rod. For selection of a single mode, the slab thickness must be about half the rod diameter if the refractive indices are equal.

smaller for an oversized dielectric rod than for a reactive surface having the same curvature and the same propagation constant.

IV. RADIATION OF A BENT ROD LOADED BY A SLAB

We now investigate the radiation loss of a dielectric rod coupled to a slab having the same refractive index n shown in Figs. 6 and 7.

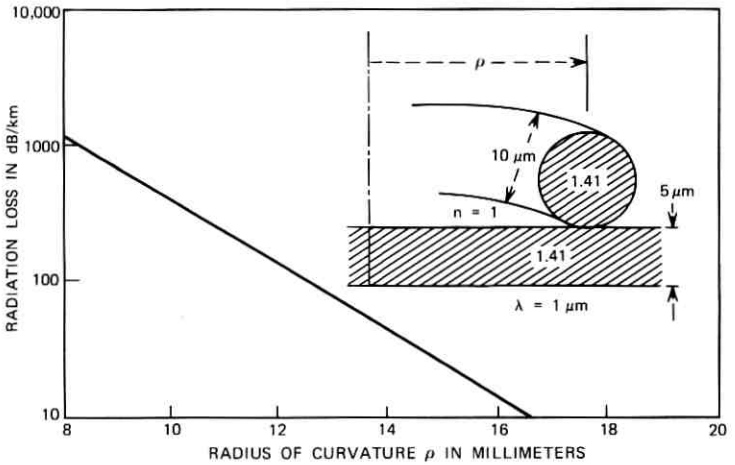


Fig. 7—Bending loss in dB/km as a function of the radius of curvature ρ for a slab-loaded rod with $n = 1.41$, $\lambda = 1$ μm, and $a = 5$ μm.

The mode selection mechanism provided by this arrangement was discussed in Part II. The thickness $2d$ of the slab is chosen equal to half the rod diameter $2a$ so that, in the absence of curvature, only the fundamental (HE_{11}) mode of the rod propagates without radiation loss.

The propagation constant h of the rod is given in eq. (21b). For the fundamental mode (H_1) of the slab, we have $k_x \approx \pi/2d = \pi/a$. Thus, from eq. (5), with k changed to kn, we obtain

$$h_s = kn - 4.9/kna^2. \tag{27}$$

Substituting these results in eq. (4b), we obtain the rod-caustic spacing y_o:

$$y_o = 2\rho/k^2n^2a^2. \tag{28}$$

It can be shown that the higher-order modes of the slab H_2, H_3, $\cdots$, corresponding to $k_x = 2\pi/a$, $3\pi/a$, $\cdots$, respectively, do not couple significantly to the rod and can be ignored.

The normalized field at the surface of the slab (as we have seen in Part II) is, with $d = a/2$,

$$E_s^2 \Big/ \int_{-\infty}^{+\infty} E_s^2(x)dx \approx 2\pi^2/p^2a^3. \tag{29}$$

Substituting the value of y_o given in eq. (28) in eq. (9b) and multiplying by the expression in eq. (29) to take into account the distribution of the radiation field in the x direction, we obtain

$$S(y_o) = (\pi^2/2)/p^2a^2. \tag{30}$$

Here it is legitimate to assume that only the average waveguide field is important. Thus, we can set $k_x = 0$ in eq. (25). Dividing by 2, because the field E_s is a constant along x instead of a cosine function, setting $L = 1$, and taking into account the spacing D between the rod and the slab, we obtain

$$R = (3.75\pi/pkna^3) \exp(-2pD). \tag{31}$$

Finally, we obtain the bending loss

$$\mathfrak{L} = SR = 58(p^3kna^5)^{-1} \exp(-5.3\rho/k^2n^2a^3) \exp(-2pD). \tag{32}$$

The term $\exp(-2pD)$ in eq. (32) is the same as in eq. (76) in Part II, and is applicable to the spurious H_{01} mode of the straight fiber. Thus, moving the fiber away from the slab reduces the mode discrimination and the bending loss of the fundamental mode in the

same proportion. A trade-off is necessary, therefore, that depends on the application, e.g., integrated optics or long-distance communication. The prefactor in eq. (32) also has the same form as for the discrimination loss of the straight fiber, with a very fast a^{-5} dependence on the fiber radius a. The numerical factor, however, is almost 20 times smaller. This, of course, is a welcome result. The first exponential term in eq. (32) has the form

$$\exp(-\rho/\rho_o), \tag{33}$$

where the critical radius is now

$$\rho_o = k^2n^2a^3/5.3 = 15a^3/\lambda^2, \quad n = 1.41. \tag{34}$$

For example, if $\lambda = 1$ μm and $a = 5$ μm, the critical radius is $\rho_0 = 1.9$ mm. If $n = 1.01$ and $a = 5$ μm, the critical radius is $\rho_0 = 0.95$ mm.

The complete expression in eq. (32) is explicitly, for $n = 1.41$,

$$\mathfrak{L} = 2.3 \times 10^8 \lambda_{\mu m}^{-1}(a/\lambda)^{-5} \exp(-12.5D/\lambda) \cdot \exp[-\rho/(15a^3/\lambda^2)] \text{ dB/km}. \tag{35}$$

This expression shows that the radiation loss does not exceed 1 dB/km if $\rho > 24$ mm, when $\lambda = 1$ μm and $a = 5$ μm. If $n = 1.01$ or, almost equivalently, if the space between the fiber and the slab is filled up with a material whose refractive index is 99 percent of the rod and slab index, we have

$$\mathfrak{L} = 12 \times 10^{10} \lambda_{\mu m}^{-1}(a/\lambda)^{-5} \exp(-1.75D/\lambda) \cdot \exp[-\rho/(7.5a^3/\lambda^2)] \text{ dB/km}. \tag{36}$$

For instance, for $\lambda = 1$ μm, $a = 5$ μm, and $D = 0$, a 1-dB/km bending loss is obtained for $\rho = 9$ mm.

The radiation losses of the higher-order modes (e.g., H_{01}) of the rod can be obtained along the same lines. Because h_{spurious} is smaller than h_s, the caustic radius ρ_s is, for spurious modes, smaller than ρ. Thus, the spurious modes see the oscillatory part of the whispering-gallery modes. The radiation losses of these spurious modes, already quite high, are not significantly affected by bending, except near the "sound barrier" $h \approx h_s$. Near this point, the JWKB approximation that we used to describe the radiation field is not applicable, and one needs the description in terms of Airy's functions. For the fundamental mode (HE_{11}) and first spurious mode (H_{01}), only the H_1 mode of the slab need be considered. For high-order modes, however, a sum needs to be performed over the various slab modes. We obtain from

eq. (10b)

$$hS = 2^{\frac{1}{2}}\pi \sum_{m=1}^{\infty} m^2(2h^2\rho^2)^{1/6} \text{Ai}^2\{(2h^2\rho^2)^{\frac{1}{3}}[1 - kn/h + \tfrac{1}{2}m^2(\pi/2hd)^2]\}. \quad (37)$$

The first factor m^2 in the sum expresses the fact that the slab field at the boundary increases in proportion to the slab mode number m. The term $\frac{1}{2}m^2(\pi/2hd)^2$ results from the change in the synchronism condition with the slab mode number m. This function is plotted in Fig. 8 as a function of: $1 - (kn/h)$ for $\lambda = 1$ μm, $n = 1.41$, a slab thickness $2d$ of 10 μm, and for a radius of curvature $\rho = 16$ mm. The propagation constants of HE_{11} and H_{01} modes for a rod with a radius of 10 μm are shown on the same figure. We observe that the radiation loss of spurious modes fluctuates by a factor as large as 100 (in dB/km) for small changes of parameters. Of course, if the radius of curvature is not a constant, these fluctuations tend to smooth out. Fig. 9 gives in more detail the region $h \approx kn$ for the same wavelength, slab thickness, and refractive index as in Fig. 8, and for $\rho = 16$ mm and 48 mm. The propagation constant of the rod modes are shown by arrows for a rod radius of 11.5 μm. This radius is to be preferred to the radius of 10 μm originally chosen because it provides a better discrimination ratio, almost 10^4 in dB/km, for $\rho = 48$ mm. The next 16 higher-order modes (each arrow, except the first, corresponds to a group of four modes) have somewhat less radiation loss than the H_{01} because they are

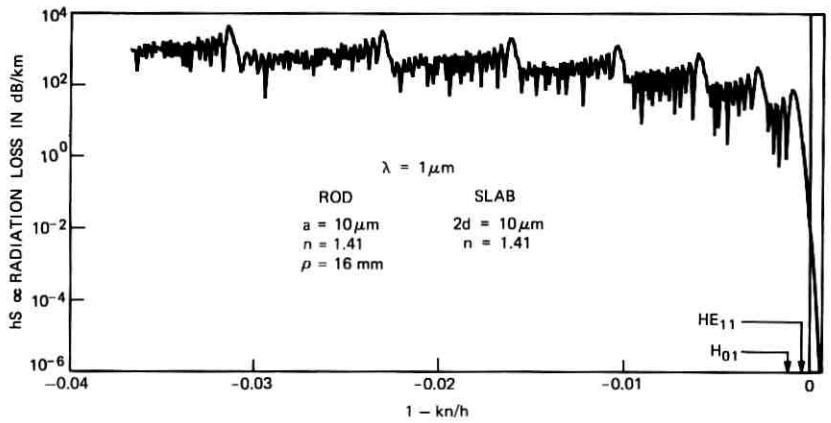


Fig. 8—Radiation loss (proportional to hS) resulting from the coupling of the waveguide to the various slab modes (H_1, H_2 ···), as a function of h (the waveguide propagation constant), for $2d = 10$ μm, $\lambda = 1$ μm, $n = 1.41$, and $\rho = 16$ mm.

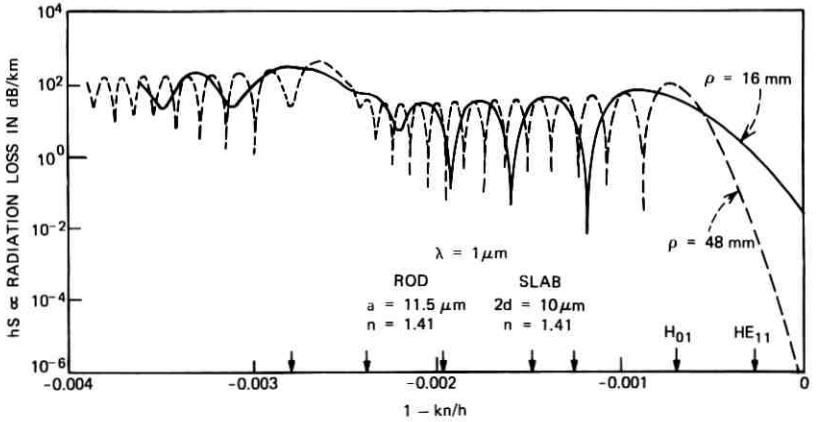


Fig. 9—Detail of Fig. 8 for $h \approx kn$, for $\rho = 16$ mm and 48 mm. The arrows denote the rod modes (HE_{11}, H_{01}, $\cdots$) for a rod radius of 11.5 μm.

beyond the "sound barrier" (we are ignoring the dips that would be smoothed out anyway in a real system). It is desirable, therefore, that a second slab, with smaller thickness, be coupled to the rod to damp strongly the modes just following the H_{01} group, and to improve further the discrimination against the other modes.

The bending loss that we have calculated is applicable to the fiber bent in the plane of the slab. When the fiber is bent in the perpendicular plane and with the same radius, the radiation loss is practically negligible. Slab-loaded fibers should be manufactured, therefore, in the shape of ribbons so that the mechanical rigidity of the fiber favors bending in the plane perpendicular rather than parallel to the slab. Physical considerations also suggest that a slight curvature of the slab, which is unimportant when the fiber is straight, may help to reduce the radiation loss.

V. ACKNOWLEDGMENT

The author expresses his thanks to E. A. J. Marcatili for many stimulating discussions and useful comments.

REFERENCES

1. M. A. Miller and V. I. Talanov, "Electromagnetic Surface Waves Guided by a Boundary with Small Curvature," Zh. T.F., *26*, No. 12 (1956), p. 2755.
2. J. Arnaud, "Transverse Coupling in Fiber Optics—Part I: Coupling Between Trapped Modes," B.S.T.J., *53*, No. 2 (February 1974), pp. 217–224; "Part II: Coupling to Mode Sinks," No. 4 (April 1974), pp. 675–696.
3. E. A. J. Marcatili, "Bends in Optical Dielectric Guides," B.S.T.J., *48*, No. 7 (September 1969), pp. 2103–2132.

4. V. V. Shevchenko, "Surface Wave Radiation Losses in Bent Waveguides," Izv. Vuz, Radiofiz, *14*, No. 5 (1971), p. 768.
5. D. Marcuse, *Light Transmission Optics*, New York: Van Nostrand Reinhold, 1972, p. 398.
6. A. W. Snyder and D. J. Mitchell, "Bending Losses of Multimode Optical Fibers," Elec. Letters, *10*, No. 1 (January 10, 1974).
7. V. V. Shevchenko, *Continuous Transitions in Open Waveguides; Introduction to the Theory*, Boulder, Colorado: The Golem Press, 1971, p. 65.
8. D. W. Berreman, "Growth of Oscillations of a Ray about the Irregularly Wavy Axis of a Lens Light Guide," B.S.T.J., *44*, No. 9 (November 1965), pp. 2117–2132.
9. J. A. Arnaud, "Graded Index Fibers Versus Cladded Fibers: a Mechanical Analogy," February 1972, unpublished work.
10. J. A. Arnaud, "Hamiltonian Theory of Beam Mode Propagation," in E. Wolf, ed., *Progress in Optics, XI*, Amsterdam: North Holland, 1973.
11. P. Kaiser, E. A. J. Marcatili, and S. E. Miller, "A New Optical Fiber," B.S.T.J., *52*, No. 2 (February 1973), pp. 265–269.
12. E. A. J. Marcatili, "Slab-Coupled Waveguides," B.S.T.J., *53*, No. 4 (April 1974), pp. 645–674.
13. J. B. Keller and S. I. Rubinow, "Asymptotic Solution of Eigenvalue Problems," Ann. Phys., *9* (1960), pp. 24–75.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 7, September 1974
Printed in U.S.A.

Optimum Refractive-Index Difference for Graded-Index Fibers Resulting From Concentration-Fluctuation Scattering

By F. W. OSTERMAYER, Jr. and D. A. PINNOW

(Manuscript received October 26, 1973)

Since the use of graded-index optical-fiber waveguides can greatly reduce pulse dispersion, fibers with greater acceptance angles can be used to increase the power coupled from an incoherent source. The increase in acceptance angle is obtained by increasing the total index variation between the core and cladding, which is achieved by increasing the doping of the fiber. This paper shows that the increased loss of the fiber resulting from concentration-fluctuation scattering by the dopant eventually decreases the light at the output end of the fiber. Therefore, there is an optimum magnitude for the total index variation based on loss considerations alone. Analytic expressions are developed using a simplified theory of concentration-fluctuation scattering that assumes the dopant molecules are randomly distributed. These expressions are applied to doped-silica-based fibers, assuming typical values for the source intensity, fiber loss, and receiver sensitivity.

I. INTRODUCTION

As is well known, multimode optical-fiber waveguides with graded refractive-index profiles can have appreciably lower mode dispersion than fibers with a step index profile for the same modal content.[1–3] One consequence of this fact is that, for a given bit rate and fiber diameter, a graded-index fiber can be used that propagates many more modes. This is important when the source is incoherent, such as a light-emitting diode, since the amount of power that can be coupled into the fiber from such a source is proportional to the number of modes it can propagate.

At what distance a signal of a given bit rate must be regenerated depends on how much dispersion has taken place and on the strength of the signal. At first, it seems that the distance would be maximized by increasing the number of modes until the maximum distance resulting from dispersion is equal to that resulting from signal strength. In this paper, we show that this may not be the case, and that a maximum distance can arise from signal strength considerations alone.

Briefly, the argument is as follows. The refractive index of the fiber is graded by a graded concentration of a dopant in a base glass. To increase the number of modes propagating in a fiber of a given diameter, the magnitude of the index variation must be increased and, therefore, so must the dopant concentration. Associated with the dopant is an additional loss resulting from scattering by fluctuations in the composition of the glass.[4] Therefore, we have a situation in which, as the number of modes is increased, the loss coefficient is also unavoidably increased. The competition between the two effects produces an optimum number of modes and, hence, a maximum distance between repeaters.

In this paper we derive the details of this effect and we calculate its magnitude for typical source, fiber, and receiver characteristics, using a simplified theory of concentration-fluctuation scattering[4] that assumes a random distribution of dopant molecules and neglecting any effect of changes in the physical properties and fictive temperature of the glass on density fluctuations.

II. THEORY

We consider a fiber with a cylindrically symmetric refractive-index profile

$$n = \begin{cases} n(r) & r < a \\ n_0 & r \geqq a, \end{cases}$$

where n_0 is a constant. For light guidance, $n(r) > n_0$.

For definiteness, we assume that the dopant increases the refractive index of the base glass, since this is generally the case for fibers based on fused silica.* It has been shown[4] that, assuming the dopant molecules are randomly distributed, a glass doped to produce an index n from a host glass of index n_0 will have an additional loss resulting from scattering by concentration fluctuations given by

$$\alpha_c = A \frac{n - n_0}{n_0}, \tag{1}$$

* Although it has been shown (Ref. 5) that boron oxide can be used to lower the refractive index of fused silica, this is the only dopant that does so.

where α_c is the excess loss coefficient and A is dependent on the refractive properties of the dopant and the host glass and on the inverse fourth power of the wavelength. Strictly speaking, the assumption of a random distribution of dopant molecules limits the validity of eq. (1) to dilute dopant concentrations. There is, however, experimental evidence that, in the Na_2O—CaO—SiO_2 and TiO_2—SiO_2 glass systems, it is valid up to the concentrations of interest[4] and, although its validity for other glass systems of interest has not been established, it does represent a reasonable starting point for estimating the effect of concentration fluctuation scattering. If we neglect any changes in the density fluctuation scattering because of changes in the physical properties and fictive temperature with doping, α_c is the total increase in scattering loss.

If $I(r)$ is the power flow per unit area in the fiber, then from eq. (1) the excess loss coefficient of the fiber will be[6]

$$\bar{\alpha}_c = \frac{\dfrac{A}{n_0}\displaystyle\int_0^\infty [n(r) - n_0]I(r)2\pi r dr}{\displaystyle\int_0^\infty I(r)2\pi r dr}. \tag{2}$$

Although the power flow in a particular mode of a cylindrically symmetric fiber generally has an azimuthal dependence, there is always another degenerate mode that can be expected to carry the same power and produce a total power independent of azimuth. $I(r)$ will be a function of the relative excitation of the different modes as well as of the index profile $n(r)$. For example, Gloge and Marcatili[3] have calculated $I(r)$ for the one-parameter class of index profiles,

$$n(r) = \begin{cases} n_0\left[\dfrac{1 - 2\Delta(r/a)^\gamma}{1 - 2\Delta}\right]^{\frac{1}{2}} & r < a \\ n_0 & r \geqq a, \end{cases} \tag{3}$$

where the parameter $\gamma \geqq 1$. This class includes the parabolic profile $\gamma = 2$ and the step profile $\gamma = \infty$. Assuming that: (*i*) the fiber can propagate many modes, (*ii*) $\Delta \ll 1$; (*iii*) $n(r)$ is relatively constant over distances of a wavelength; and (*iv*) all modes carry the same power, Gloge and Marcatili[3] have shown that $I(r)$ for the above class of index profiles is

$$I(r) = \begin{cases} I_0[1 - (r/a)^\gamma] & r < a \\ 0 & r \geqq a. \end{cases} \tag{4}$$

Multimode fibers of interest will undoubtedly satisfy assumptions (*i*) to (*iii*). Assumption (*iv*) is less realistic but, as the actual distribution

of power among the modes of a given fiber will depend on the relative loss of the modes and the coupling between them, it provides a convenient first approximation.

Since we are restricted to $\Delta \ll 1$ by assumption (*ii*), it is reasonable to make the approximation to the profiles of eq. (3)

$$n(r) \approx n_0 \left[\frac{1 - \Delta(r/a)^\gamma}{1 - \Delta} \right] \quad r < a. \tag{5}$$

Using eqs. (4) and (5) in eq. (2), we obtain

$$\bar{\alpha}_c = \frac{A\gamma}{\gamma + 1} \frac{\Delta}{1 - \Delta} \approx \frac{A\gamma}{\gamma + 1} \Delta. \tag{6}$$

We are interested in considering only those index profiles that have low pulse dispersion. From among the class of profiles of eq. (3), Gloge and Marcatili[3] found that the lowest dispersion occurs for a value of γ very close to 2. Therefore, we will consider specifically the parabolic profile for which

$$\bar{\alpha}_c = \tfrac{2}{3} A \Delta \tag{7}$$

from eq. (6).

III. APPLICATION TO REPEATER SPACING

We now consider the consequences of the Δ-dependent loss of eq. (7) on the distance one can transmit before the signal becomes so weak that it must be regenerated. The loss of the fiber is

$$10 \log_{10} \frac{P_i}{P_0} = (\alpha_0 + \bar{\alpha}_c) L, \tag{8}$$

where P_i is the power at the input, P_0 is the power at a distance L kilometers from the input, $\bar{\alpha}_c$ is the fiber loss coefficient in dB per kilometer because of concentration-fluctuation scattering, and α_0 is the fiber loss coefficient in dB per kilometer resulting from all other scattering and absorption processes.

For incoherent excitation of the fiber, P_i will be a function of Δ. An incoherent source is characterized by a brightness B that is the power per unit solid angle per unit area. In general, B is a function of both position and direction. Since the area of the fiber and the range of directions over which the fiber accepts light are small, we assume B constant. The solid acceptance angle at the fiber end face is[3]

$$\Omega(r) = \pi[n^2(r) - n_0^2] = \frac{2\pi n_0^2 \Delta}{1 - 2\Delta} \left[1 - \left(\frac{r}{a} \right)^\gamma \right]$$

and, therefore, the power coupled to the fiber neglecting any reflection loss will be

$$P_i = \int_0^a B\Omega(r)2\pi r dr = 2\pi^2 a^2 n_0^2 B \left(\frac{\gamma}{\gamma+2}\right)\frac{\Delta}{1-2\Delta} \approx C\Delta. \quad (9)$$

The maximum distance of transmission will be that value of L for which P_0 is equal to the minimum detectable signal as determined by the receiver sensitivity. Combining eqs. (7), (8), and (9), we obtain for the maximum transmission distance

$$L = \frac{10\log_{10}(C/P_{0\,\min}) + 10\log_{10}\Delta}{\alpha_0 + \frac{2}{3}A\Delta}. \quad (10)$$

This function has a maximum value as a function of Δ for any given choice of C, $P_{0\,\min}$, α_0, and A, and, therefore, from loss considerations alone there is an optimum Δ.

In Fig. 1, L is plotted as a function of Δ for a possible dopant of fused silica, TiO_2,[7] and for three values of $C/P_{0\,\min}$ (a ratio that is the total allowable loss normalized by the fractional index variation Δ) that cover the range expected for a practical system. They correspond to $P_{0\,\min}$ between 10^{-5} and 10^{-6} milliwatts[8] and C between 1 and 10 milliwatts (P_i between 10 and 100 microwatts for a fiber with $\Delta = 0.01$).* The values used for the other parameters in eq. (10) are $\alpha_0 = 4$ dB per kilometer, corresponding to the loss of the best Corning fibers in the vicinity of 0.85 micrometer,[10] and from the calculations of Ref. 4 a value of A of 170 dB per kilometer, corresponding to TiO_2 at a wavelength of 0.90 micrometer. The maxima are quite broad; this (together with the result from Fig. 2 that the optimum Δ is not a strong function of receiver sensitivity) means that a particular choice of Δ near the optimum will give repeater spacings near the maximum over a considerable range of receiver sensitivities.

Another interesting point from Figs. 1 and 2 is the relatively low value of the optimum Δ, which lies in the range of 0.0037 to 0.0064 for the expected source, receiver, and fiber parameters.

IV. CONCLUSIONS

The purpose of this paper has been to show that, when concentration-fluctuation scattering is taken into account, pulse dispersion does not necessarily determine the total index variation that will maximize

* Burrus and Dawson (Ref. 9) have shown that it is possible to make diodes with $C = 10$ milliwatts and long life. To cover the full range of possibilities, however, we also include the conservative value $C = 1$ milliwatt.

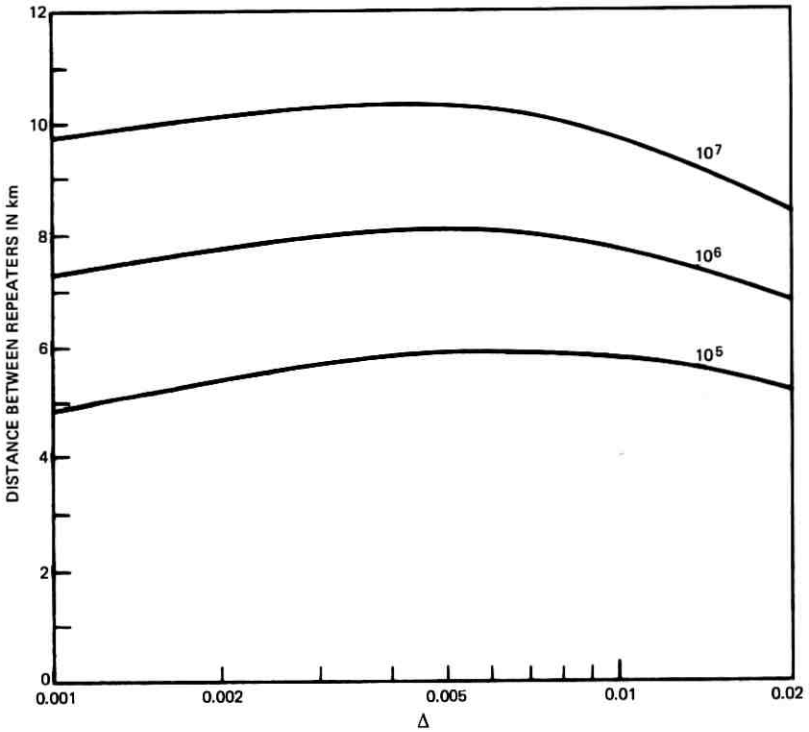


Fig. 1—Plot of the maximum distance between repeaters resulting from signal strength as a function of the total relative index variation Δ of a TiO_2-doped fused-SiO_2 parabolic profile graded-index optical waveguide. The numbers labeling the curves are the values of $C/P_{0\,\min}$, the ratio of the power available from the source normalized by Δ [which, from eq. (9), is a function of the brightness of the source and the diameter and index profile of the fiber] to the minimum power the receiver can detect.

the distance a signal can be transmitted over a graded-index optical-fiber waveguide. We have specifically considered the parabolic index profile, since it is close to the optimum for minimizing pulse dispersion. A theory of concentration-fluctuation scattering has been used that assumes a random distribution of dopant molecules, and changes in density fluctuation scattering have been neglected. With these assumptions, we find for TiO_2 doping and typical source, receiver, and fiber parameters that the value of Δ maximizing the distance between repeaters from loss considerations is in the range of 0.0037 to 0.0064. The corresponding maximum distance is in the range of 10.3 to 5.9 kilometers.

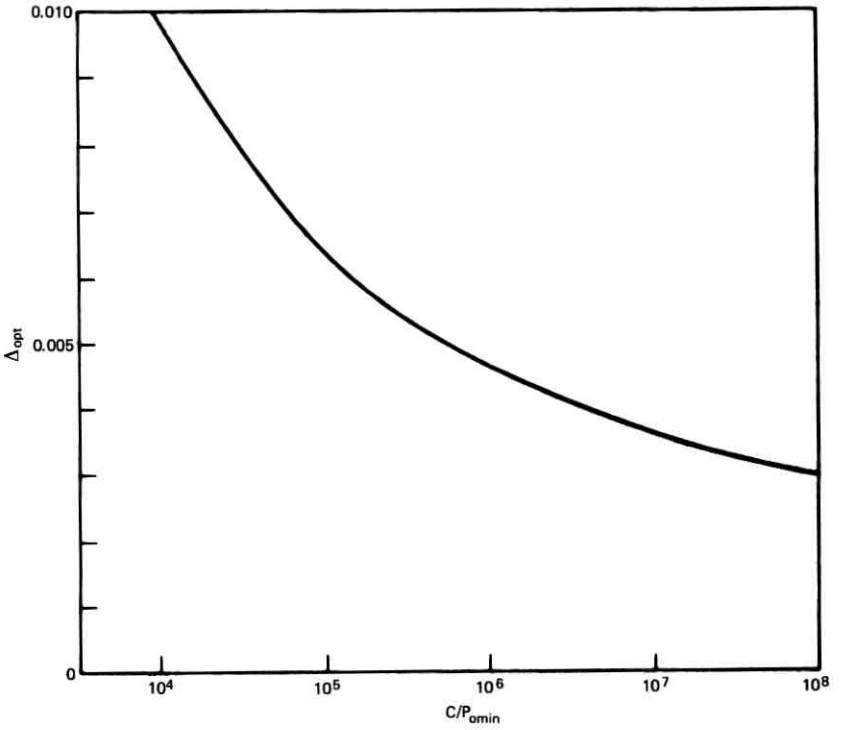


Fig. 2—Plot of Δ_{opt}, the value of the total relative index variation of a parabolic profile graded-index optical waveguide that maximizes the distance between repeaters because of signal strength, as a function of $C/P_{0\,min}$ for TiO_2-doped fused SiO_2.

Since the pulse dispersion τ for a parabolic index profile is approximately $(Ln_0/c)(\Delta^2/2)$,[3] we find, for $\Delta = 0.0047$ and $L = 8.1$ kilometers (the optimum values for $C/P_{0\,min} = 10^6$), that $\tau = 0.44 \times 10^{-9}$ seconds corresponding to a bit rate of 2.3 gigabits per second. Hence, for a system having $C/P_{0\,min} = 10^6$ and a lower bit rate, the upper limit on Δ and the maximum repeater spacing would be determined by loss considerations alone.

REFERENCES

1. S. E. Miller, "Light Propagation in Generalized Lens-Like Media," B.S.T.J., *44*, No. 9 (November 1965), pp. 2017–2064.
2. S. Kawakami and J. Nishizawa, IEEE Trans. Microwave Theory and Tech., *MTT-16*, 1968, p. 814.
3. D. Gloge and E. A. J. Marcatili, "Multimode Theory of Graded-Core Fibers," B.S.T.J., *52*, No. 9 (November 1973), pp. 1563–1578.

4. F. W. Ostermayer and D. A. Pinnow, "Concentration Fluctuation Scattering Applied to Optical Fiber Waveguides," presented at the Annual Meeting of the American Ceramic Society, Cincinnati, Ohio, 1973, unpublished.
5. S. H. Wemple, D. A. Pinnow, T. C. Rich, R. E. Jaeger, and L. G. Van Uitert, "The Binary SiO_2—B_2O_3 Glass System: Refractive Index Behavior and Energy Gap Considerations," J. Appl. Phys., *44*, 1973, pp. 5432–5437.
6. A. W. Snyder, Proc. IEEE, *60*, 1972, p. 757.
7. R. D. Maurer, Proc. IEEE, *61*, 1973, p. 452.
8. J. E. Goell, "An Optical Repeater with High-Impedance Input Amplifier," B.S.T.J., *53*, No. 4 (April 1974), pp. 629–643.
9. C. A. Burrus and R. W. Dawson, Appl. Phys. Lett., *17*, 1970, p. 97.
10. P. C. Schultz, Annual Meeting of the American Ceramic Society, Cincinnati, Ohio, 1973, unpublished.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 53, No. 7, September 1974
Printed in U.S.A.

Application of Response-Bound Method in Shock and Vibration Analysis of Telephone Structures

By S. C. LIU

(Manuscript received December 10, 1973)

Telephone structures must remain in operation during the shock and vibration caused by blasts, earthquakes, or other dynamic forces. Traditionally, various numerical methods, including the finite element approach, have been used to analyze such problems. These methods, although effective, generally require excessive and costly computations. In contrast, the Fourier transform method used in conjunction with a fast Fourier transform algorithm is much more economical. In addition, shock and vibration problems involving frequency-dependent parameters can be effectively treated by the Fourier method. However, to make the Fourier method more effective and widely applicable, various tractable input-transfer-output relations are needed. This paper derives a set of simple equality and inequality relations that allow various response parameters to be conveniently estimated based on partial knowledge of the input and the structures. In particular, two lower bounds and eight upper bounds of the maximum response of linear structures are presented. Simple structures subjected to impulse loads resembling blasts and a random transient load resembling earthquakes are studied. Practical applications of the method to telephone structures are demonstrated by the analysis of a battery stand and a community dial office system in an earthquake area.

I. INTRODUCTION

The Fourier transform method has long been used to characterize linear systems and to identify the frequency content of waveforms. Sneddon[1] demonstrates classical applications of the Fourier transform method to engineering and physics problems. Amba-Rao[2] gives recent applications of this method to elasticity, White[3] illustrates system response calculations, Le Bail[4] demonstrates boundary-value problems

in physics, and Liu and Fagel[5,6] analyze earthquake soil-structure interactions. In the last three applications cited, the fast Fourier transform (FFT) algorithms were used in conjunction with the Fourier method to obtain numerical solutions. The development of FFT algorithms in the past few years has rapidly expanded the application of the Fourier method to include digital analysis of linear-system dynamics.

This work was motivated by the idea that some simple and effective techniques based on the Fourier transform method and response-bound relations of linear systems can be developed for the shock and vibration analysis of telephone structures.

This paper discusses the Fourier transform method applied to the vibration problems of linear structures and presents some simple relationships that make it possible to quickly evaluate and estimate various response parameters. In a typical engineering vibration problem, the primary concerns are the time and frequency aspects of the input and output (response) and the transfer characteristics of the system or structure involved. The parameters or functions of interest are generally the waveforms of input and output, their Fourier transforms, the peak values, the frequency characteristics of these waveforms, and the effects of linear filtering by structures. Of further concern are the time and frequency distributions of energy or power of these waveforms and, particularly, the analytical and empirical relations among all the parameters and functions mentioned above. In view of the advance of computational techniques, some well-known and lesser-known input-transfer-output relations useful to shock and vibration analysis should be systematically presented. The purpose of this paper is to make these relations available, to discuss their implications, and to demonstrate their applications.

II. BACKGROUND AND THEORETICAL BASIS

Applied to the engineering vibration problem, the basic concept of the Fourier method of analysis is very simple. We assume that the structure under consideration is linear, causal, and stable, and its property is completely defined by the associated transfer function pairs $h(t) \leftrightarrow H(i\omega)$, where $H(i\omega)$ is the frequency response function, $h(t)$ is the impulse response function, t is the time variable, ω is the frequency variable, i is the complex unit, and the symbol $\leftrightarrow$ indicates a Fourier transform (FT) pair. When the structure is excited by an arbitrary (deterministic or random) input function, $x(t) \leftrightarrow X(i\omega)$, the response is given as $y(t) \leftrightarrow Y(i\omega)$. If $X(i\omega)$ and $H(i\omega)$ are known, the response can be obtained by using the simple relationship, $y(t)$

$= x(t)*h(t)$, where $*$ indicates convolution, or by using $Y(i\omega) = X(i\omega)H(i\omega)$ and taking the inverse FT of $Y(i\omega)$. Other response quantities of interest can also be obtained from the basic FT pairs $x(t) \leftrightarrow X(i\omega)$, $h(t) \leftrightarrow H(i\omega)$, and $y(t) \leftrightarrow Y(i\omega)$ once their relations and the pertinent response parameter are found. This approach is efficient because no numerical integrations are required (as in many traditional methods of analysis), and it yields all needed solutions in both the time and frequency domains with simple discrete FT or FFT computer routines now generally available. However, the Fourier method of analysis is particularly attractive and powerful because of the availability of some analytical input-transfer-output relations as described in this paper. These relations, in the form of simple equality or inequality, can solve many practical problems. Some important preliminary background material, particularly the definition and properties of an arbitrary shock function, $f(t)$, will be briefly given before the main results.

2.1 Shock function and related quantities

A time function $f(t)$, $t \in [a, b]$ for finite $a > b \geqq 0$ is referred to as a shock function. Without losing generality, it is assumed that $f(t)$ begins at $t = 0$ and has finite duration T, i.e., $a = 0$ and $b = T$. The shock function $f(t)$ could be the input $x(t)$, the output $y(t)$, or the impulse response function of the system $h(t)$, and it could be either deterministic or random. Associated with a shock function $f(t)$, the following quantities of interest are defined:

Fourier Transform:

$$F(i\omega) = \int_0^T f(t) \exp(-i\omega t)dt. \tag{1a}$$

Running FT or Spectrum:

$$F(t, i\omega) = \int_0^t f(\tau) \exp(-i\omega\tau)d\tau. \tag{1b}$$

Energy:

$$E_f(t) = \int_0^t f^2(\tau)d\tau = \frac{1}{2\pi}\int_{-\infty}^{\infty} |F(t, i\omega)|^2 d\omega. \tag{1c}$$

Total Energy:

$$E_f = \int_0^T f^2(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |F(i\omega)|^2 d\omega. \tag{1d}$$

Instantaneous Power Spectrum:

$$\rho_f(t, i\omega) = \frac{\partial}{\partial t} |F(t, i\omega)|^2. \tag{1e}$$

Power:

$$P_f(t) = \frac{dE_f(t)}{dt} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \rho_f(t, i\omega) d\omega. \tag{1f}$$

Frequency Moment:

$$M_{f,n} = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\omega|^n |F(i\omega)| d\omega \quad n = 0, 1, 2, \cdots. \tag{1g}$$

Time Moment:

$$m_{f,n} = \int_0^T |t|^n |f(t)| dt \quad n = 0, 1, 2, \cdots. \tag{1h}$$

Frequency Variation:

$$S_{f,n} = \frac{1}{2\pi} \int_{-\infty}^{\infty} |F^{(n)}(i\omega)| d\omega, \quad n = 0, 1, 2, \cdots. \tag{1i}$$

Time Variation:

$$s_{f,n} = \int_0^T |f^{(n)}(t)| dt, \quad n = 0, 1, 2, \cdots. \tag{1j}$$

Correlation Function:

$$R_f(\tau) = f(t)*f(t) = \int_0^T f(t)f(t - \tau) dt. \tag{1k}$$

Note that some of these quantities, e.g., $M_{f,n}$ and $m_{f,n}$, may not exist. Some simple bounds on $f(t)$, $F(\omega)$, and their derivatives, which will be used to derive response relations, can be readily obtained from the above equations.

2.2 Some bounds of shock function and related quantities

2.2.1 Derivative bounds

The following inequalities relating the derivatives of a shock function or of its FT with various signal parameters can be established.

$$|f^{(n)}(t)| \leqq M_n \qquad n = 0, 1, 2, \cdots. \tag{2a}$$

$$|f^{(n)}(t)| \leqq \Omega^n \left[\frac{E_f \Omega}{(2n + 1)\pi} \right]^{\frac{1}{2}} \quad n = 0, 1, 2, \cdots. \tag{2b}$$

$$|F^{(n)}(i\omega)| \leqq m_n \qquad n = 0, 1, 2, \cdots. \tag{2c}$$

$$|F^{(n)}(i\omega)| \leqq T^n \left(\frac{E_f T}{2n + 1} \right)^{\frac{1}{2}} \quad n = 0, 1, 2, \cdots. \tag{2d}$$

These relations can be proved as follows. The nth derivatives $f^{(n)}(t) = d^n f(t)/dt^n$ and $F^n(i\omega) = d^n F(i\omega)/d\omega^n$ $(n = 0, 1, 2, \cdots)$, can be written as

$$f^{(n)}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (i\omega)^n F(i\omega) \exp(i\omega t) d\omega \tag{3a}$$

and

$$F^{(n)}(i\omega) = \int_0^T (-it)^n f(t) \exp(-i\omega t) dt. \tag{3b}$$

Equations (2a) follows from (3a) and (1g), and (2c) follows from (3b) and (1h). Note that (2a) and (2c) are analogous bounds in terms of frequency and time moments, respectively. Equations (2b) and (2d) follow from (3a) and (3b), respectively, and Schwartz inequality. In (2b), Ω is a constant and is defined as $F(i\omega) \cong 0$ for $|\omega| \geqq \Omega$; i.e., in the practical sense, the shock function $f(t)$ is assumed to be essentially band-limited. Although a shock function cannot be strictly band-limited, in practice most functions (except ideal impulses) possess a frequency value beyond which the Fourier spectra are negligibly small. Furthermore, because the sample rate can never be infinitely large, digitization of the time function always imposes a practical band-limit. Although the details are beyond the scope of this paper, various methods and criteria are available for the construction of equivalent band-limited signals. Equations (2b) and (2d) are also analogous bounds in terms of energy.

2.2.2 Amplitude bounds

$$|f(t)| \leqq \left(\frac{E_f \Omega}{\pi}\right)^{\frac{1}{2}}, \quad \text{for all } t, \tag{4a}$$

$$|F(i\omega)| \leqq (E_f T)^{\frac{1}{2}}, \quad \text{for all } \omega, \tag{4b}$$

$$|f(t)| \leqq \frac{S_{f,n}}{t^n}, \quad \text{for all } t, \quad n = 0, 1, 2, \cdots, \tag{4c}$$

$$|F(i\omega)| \leqq \frac{S_{f,n}}{|\omega|^n}, \quad \text{for all } \omega, \quad n = 0, 1, 2, \cdots. \tag{4d}$$

When $n = 0$, (4a) and (4b) are special cases of (2b) and (2d), respectively. From the relation $(-it)^n f(t) \leftrightarrow F^{(n)}(i\omega)$, the following relation holds:

$$|f(t)| \leqq |(it)^{-n}| \frac{1}{2\pi} \int_{-\infty}^{\infty} |F^{(n)}(i\omega)| d\omega. \tag{5}$$

Relation (4c) is obtained from (1i) and (5). Note that in (4c) no

band-limitedness requirement is needed. Similarly, the relation $f^{(n)}(t) \leftrightarrow (i\omega)^n F(i\omega)$, eq. (1j), and

$$|F(i\omega)| \leqq |(i\omega)^{-n}| \int_{-\infty}^{\infty} |f^{(n)}(t)|\, dt \tag{6}$$

lead to (4d). When (4c) [or (4d)] is used, better bounds are provided for small t (or ω) when n is small, and for large t (or ω) when n is large.

2.2.3 Increment bounds

Let $\delta_f = f(t + \Delta t) - f(t)$ and $\delta_F = F[i(\omega + \Delta\omega)] - F(i\omega)$, then

$$|\delta_f| \leqq M_1 \Delta t, \tag{7a}$$

$$|\delta_f| \leqq \left[\frac{2E_f\Omega}{\pi}\left(1 - \frac{\sin \Omega\Delta t}{\Omega\Delta t}\right)\right]^{\frac{1}{2}} \leqq \Omega\Delta t\left(\frac{E_f\Omega}{3\pi}\right), \tag{7b}$$

$$|\delta_F| \leqq m_1 \Delta\omega, \tag{7c}$$

and

$$|\delta_F| \leqq \left[2E_f T\left(1 - \frac{\sin T\Delta\omega}{T\Delta\omega}\right)\right]^{\frac{1}{2}} \leqq T\Delta\omega\left(\frac{E_f T}{3}\right)^{\frac{1}{2}}. \tag{7d}$$

Relations (7a) and (7c) are special cases (when $n = 1$) of (2a) and (2c), respectively. Equations (7b) and (7d) follow from (1a) and the corresponding inverse FT relation, respectively, and from Schwartz inequality. Note that the band-limitedness requirement is imposed to derive (7b). Equations (7a) and (7c) are analogous increment bounds in terms of moments; (7b) and (7d) are analogous bounds in terms of energy. Kak[7] proves (2d), (4b), and (7d), while Papoulis[8,9] proves (2b), (4a), and (7b).

The equality conditions for the above relations can be easily identified. For example, the equality holds for (2d) when $f(t) = C(it)^n$ for even n and a constant C, for (4b) when $f(t) = C \exp(i\omega_o t)$, when $\omega_o = n\pi/T$, and $n = 0, 1, 2, \cdots$, and for (7d) when $f(t) = 2(1 - \sin \Delta\omega)^{\frac{1}{2}}$. Analogous conditions apply to their analogous counterrelations and can be established straightforwardly.

2.2.4 Energy bounds

Let $f_m = \max |f(t)|$, occurring at $t = t_f$, and $F_m = \max |F(i\omega)|$, occurring at $\omega = \omega_F$. Then

$$F_m^2 T^{-1} \leqq E_f \leqq f_m^2 T \quad \text{for all } \omega. \tag{8}$$

The upper bound is obvious from the definition of E_f given in (1d); the lower bound follows from applying Schwartz inequality to the

relation

$$F(i\omega) = \int_0^T f(t) \exp(-i\omega t)dt,$$

which results in

$$|F(i\omega)|^2 \leqq \int_0^T |f(t)|^2 dt \int_0^T |\exp(-i\omega t)|^2 dt \leqq E_f T \quad \text{for all } \omega. \quad (9)$$

III. RELATIONS FOR VARIOUS RESPONSE PARAMETERS

The basic properties of a general shock function and its FT provide the basis for establishing important equality and inequality relations for various response parameters of linear systems. Let $x(t) \leftrightarrow X(i\omega)$, $h(t) \leftrightarrow H(i\omega)$, and $y(t) \leftrightarrow Y(i\omega)$ represent, respectively, the FT pair of the input, the transfer function, and the response; $x(t)$, $h(t)$, and $y(t)$ are shock functions; therefore, the basic definitions and properties are applicable. The functions $x(t)$, $h(t)$, and $y(t)$ are assumed to possess time duration T_x, T_h, and T_y and are band-limited within ω_X, ω_H, and ω_Y. It is obvious that $T_y = T_x + T_y$, and $\Omega_Y = \min(\Omega_X, \Omega_H)$. Let $x_m = \max|x(t)|$, $h_m = \max|h(t)|$, $y_m = \max|y(t)|$, $X_m = \max|X(i\omega)|$, $H_m = \max|H(i\omega)|$, and $Y_m = \max|Y(i\omega)|$, and t_x, t_h, t_y, ω_X, ω_H, and ω_Y be the corresponding times and frequencies these maxima occur. Note that t_x, t_h, and t_y or ω_X, ω_H, and ω_Y need not be equal. In addition to the well-known basic input-transfer-output relationships $y(t) = x(t)*h(t)$ and $Y(i\omega) = X(i\omega)H(i\omega)$, the important relations that follow can be established.

3.1 Total energy bounds

The total energy of $x(t)$, $h(t)$, or $y(t)$ is bounded by the peak amplitude of the associated time function from above and by the peak amplitude of the corresponding FT from below.

$$X_m^2 T_x^{-1} \leqq E_x \leqq x_m^2 T_x \quad (10a)$$

$$H_m^2 T_h^{-1} \leqq E_h \leqq h_m^2 T_h \quad (10b)$$

$$Y_m^2 T_y^{-1} \leqq E_y \leqq y_m^2 T_y. \quad (10c)$$

These results follow directly from the use of (8).

3.2 Kinetic, potential, and dissipated energies

For a simple oscillator with unit mass, a natural frequency ω_o, and a viscous damping λ, the following relationships exist for the kinetic energy $[T(t) = \dot{y}^2/2 \leftrightarrow T(\omega)]$, the potential energy $[V(t) = \omega_o^2 y^2/2 \leftrightarrow V(\omega)]$, and the dissipated energy $[D(t) = \lambda\omega_o^2 y\dot{y} \leftrightarrow D(\omega)]$,

of the system

$$T(\omega) = \frac{1}{4\pi}[\dot{Y}(i\omega)*\dot{Y}(i\omega)] \tag{11a}$$

$$V(\omega) = \frac{\omega_o^2}{4\pi}[Y(i\omega)*Y(i\omega)] \tag{11b}$$

$$D(\omega) = \frac{\lambda\omega_o^2}{2\pi}[\dot{Y}(i\omega)*Y(i\omega)], \tag{11c}$$

where $\dot{y}(t) \leftrightarrow \dot{Y}(i\omega) = (i\omega)Y(i\omega)$. The proof of these equations is obvious. Using the relationship $Y(i\omega) = H(i\omega)X(i\omega)$ with a simple FT routine, (11a) through (11c) allow various energy functions of time to derive directly from $X(i\omega)$ and $H(i\omega)$.

3.3 Maximum response bounds

Various upper and lower bounds of the maximum temporal response $y_m = \max |y(t)|$ can also be established. Let l_i denote lower bounds and u_i denote upper bounds of $y_m (i = 1, 2, \cdots)$, then the following relation holds:

$$l_2 \leqq l_1 \leqq y_m \leqq \begin{cases} u_1 \leqq \begin{cases} u_5 \leqq u_8 \\ u_6 \\ u_7 \end{cases} \\ u_2 \\ u_3 \\ u_4 \end{cases}, \tag{12}$$

where

$$l_1 = \max_t [E_y(t)/t]^{\frac{1}{2}}, \qquad l_2 = Y_m/T_y, \qquad u_1 = (E_y\Omega_Y/\pi)^{\frac{1}{2}},$$

$$u_2 = (E_xE_h)^{\frac{1}{2}}, \qquad u_3 = S_{y,n}/(t_y)^n, \qquad u_4 = \left| \begin{matrix} \Omega_Y \\ V \\ -\Omega_Y \end{matrix} \quad Y(i\omega) \right| / |2\pi t_y|,$$

$$u_5 = \Omega_Y Y_m/\pi, \qquad u_6 = (\Omega_Y E_x/\pi)^{\frac{1}{2}} H_m, \qquad u_7 = (\Omega_Y E_h/\pi)^{\frac{1}{2}} X_m,$$

and

$$u_8 = \Omega_Y X_m H_m/\pi.$$

The symbol V in the expression of u_4 indicates the total variation.

The lower bounds shall be proved first. Considering the running FT of $y(t)$ according to (1b) and applying Schwartz inequality, the following can be written

$$\begin{aligned} |Y(t, i\omega)|^2 &\leqq \int_0^t |y(\tau)|^2 d\tau \int_0^t |\exp(-i\omega\tau)|^2 d\tau \\ &\leqq E_y(t)t \\ &\leqq [\max_{0 \leqq \tau \leqq t} |y(\tau)|]^2 t^2. \end{aligned} \tag{13}$$

Therefore, for all ω and t,

$$\max_{0 \leq \tau \leq t} |y(\tau)| \geq \left(\frac{E_y(t)}{t} \right)^{\frac{1}{2}} \geq \frac{|Y(t, i\omega)|}{t} \tag{14a}$$

and

$$y_m \geq \max_t \left[\frac{E_y(t)}{t} \right]^{\frac{1}{2}} \geq \frac{Y_m}{T_y}, \tag{14b}$$

which gives the lower bounds l_1 and l_2. Note that $(E_y/T_y)^{\frac{1}{2}}$ and $[E_y(t_y)/t_y]^{\frac{1}{2}}$ are two special cases of l_1, and that l_1 occurs at large (or small) t if the energy function $E_y(t)$ increases faster (or slower) than t. For a given structure, Y_m is fixed; therefore, the bound l_2 decreases with T_y. Note also that, in (14a), the quantity $\max_{t,\omega} [|Y(t, i\omega)|/t]$ should be the exact expression for l_2. However, the current expression Y_m/T_y is more convenient to use than the exact expression that requires the computation of the running spectrum over the entire $t - \omega$ plane. Clearly, better lower bounds for the maximum response can be established on the knowledge of the total energy of the response than on the knowledge of the peak Fourier spectrum of the response.

It is interesting to compare the lower bounds in (12) with the bound Liu[10] establishes for a specific case. For a simple linear oscillator with natural frequency ω_o and viscous damping λ, it has been shown that the maximum displacement response

$$y_m \geq \frac{1}{p} |X(i\omega)|_{\omega=p} = l_3, \tag{15}$$

where $p = (1 - \lambda^2)^{\frac{1}{2}}\omega_o$. In this situation, the corresponding $H(i\omega) = (-\omega^2 + 2i\omega\omega_o + \omega_o^2)^{-1}$ and $H_m = (2\lambda\omega_o p)^{-1}$ occurring at $\omega = \omega_H = (1 - 2\lambda^2)^{\frac{1}{2}}\omega_o$. Therefore, the bound l_2 gives $y_m \geq |X(i\omega)|_{\omega=\omega_H}/(2p\omega_o T)$, and in (15) l_3 gives $y_m \geq |X(i\omega)|_{\omega=p}/(p)$. Clearly, if damping is small, i.e., if $\omega_H \approx p$, then $l_2 < l_3$ when $2\lambda\omega_o T > 1$, $l_2 > l_3$ when $2\lambda\omega_o T < 1$, and $l_2 = l_3$ when $2\lambda\omega_o T = 1$.

The upper bounds, u_1 through u_8, will now be derived. The bound u_1 follows from the use of (4a) and the assumption $t_y \leq T_y$; u_2 is obtained from applying Schwartz inequality to the relation $y(t) = x(t)*h(t)$; u_3 is obtained directly from (4c); u_4 is a special case of u_3 when $n = 1$ (Giardina[11] proves this differently); u_5 and u_8 follow from u_1 and the relation

$$E_y = \frac{1}{2\pi} \int_{-\Omega_Y}^{\Omega_Y} |Y(i\omega)|^2 d\omega \leq \frac{\Omega_Y}{\pi} Y_m^2 \leq \frac{\Omega_Y}{\pi} X_m^2 H_m^2; \tag{16}$$

and u_6 and u_7 follow from u_1 and the relation

$$E_y = \frac{1}{2\pi} \int_{-\Omega_Y}^{\Omega_Y} |X(i\omega)H(i\omega)|^2 d\omega \leq H_m^2 E_x \quad \text{or} \quad X_m^2 E_h. \tag{17}$$

Generally, better bounds can be obtained when more is known about the input, the transfer characteristics, and the output. For example, the poorest bound in (12) is obviously u_8 because the band-limit Ω_Y and the peak values of spectra $|X(i\omega)|$ and $|H(i\omega)|$ are all that are known. Closer bounds can be established if Ω_Y and Y_m are known, or, equivalently, E_x and H_m, or E_h and X_m. Still closer bounds can be obtained if the energy functions E_y or E_x and E_h are known. Knowledge of E_y, in most cases, depends on the knowledge of $y(t)$ or $Y(i\omega)$; the exact value of y_m can be readily obtained from this knowledge. Therefore, bounds depending on E_y explicitly, such as l_1 and u_1, may not help in practical applications. Note that bounds u_1, u_2, and u_6 grow with T_y; therefore, they do not provide good results for large T_y, i.e., when long excitations or small damping sustain the response for some time. The bound u_2 depends only on the knowledge of the input and the transfer function; therefore, it is convenient to use. Furthermore, it is independent of the output information, thus, it cannot be considered one of the better bounds, especially for very small or very large time t. Bounds u_6, u_7, and u_8 also depend mainly on the input and the transfer function, and the only output knowledge required is the band-limit frequency Ω_Y. Bounds u_3 and u_4 are generally better, yet they require the prior knowledge of t_y, the time when the maximum response occurs. This is usually difficult to know without the waveform or FT of the response. Additional calculations are also needed for the following two bounds: for u_3, the calculation of $S_{y,n}$ and the areas under the spectra of $Y(i\omega)$ and its derivatives, and for u_4, the absolute value of the total variation of the complex quantity $Y(i\omega)$ for the entire frequency axis. For u_3, small n provides better bounds for small t, and large n for large t. For u_4, it provides better bounds for large t than for small t. In general, u_1, u_3, and u_4 are better upper bounds, yet they are less convenient to use than others.

Some interesting features about these response bounds are noteworthy. Drenick[12] derives the least favorable excitation for structures based on u_2. An uncertainty condition $T_y\Omega_Y \geqq \pi$, familiar in band-limited and time-limited signal analysis can be arrived at from bounds l_2 and u_5. The bound u_4 can be used to estimate the settling time t_ϵ for the response $y(t)$ to remain with a specified value ϵ[10]. For example, if $Y_1(i\omega)$ is smoother, and thus has less variation than $Y_2(i\omega)$, then $y_1(t)$ takes longer than $y_2(t)$ to settle to the level ϵ. This implies that, for two comparable response Fourier spectra, the smoother one corresponds to the response of a system with higher damping and the rugged one corresponds to the response with less damping.

3.4 Mean square response

If $x(t)$ and hence $y(t)$ are random, the mean square response $\sigma_y^2 = E[|y(t)|^2]$ (where E denotes expectation) is of interest. Let $\sigma_x^2 = E[|x(t)|^2]$, and $S_x(i\omega)$ and $S_y(i\omega)$ be the power spectral density functions, and $R_x(\tau)$ and $R_y(\tau)$ be the autocorrelation functions of $x(t)$ and $y(t)$, respectively, then the following relations regarding σ_y can be found:

$$\sigma_y^2 = R_y(0) = \int_0^\infty R_h(\tau) R_x(\tau) d\tau \tag{18a}$$

$$\sigma_y^2 = \frac{E[E_y]}{T_y} = \frac{E[|X(i\omega)|^2]}{2\pi T_y} E_h \tag{18b}$$

$$\sigma_y^2 \leqq \frac{\sigma_x^2}{2\pi} H_m^2. \tag{18c}$$

In (18a), $R_y(\tau) = R_h(t) * R_x(t)$, $R_h(\tau) = \int_0^{T_h} h(t) h(t - \tau) dt$ = the system autocorrelation function, as defined in (1k). Because the proofs of these equations are straightforward, they are omitted. The mean square response can be provided either by (18a) upon a simple time integration of the product of the autocorrelation functions of input and impulse response, or by (18b) upon the direct knowledge of the mean output energy or the expected square input Fourier spectrum. Equation (18c) provides a simple upper bound based on mean square value of input and the peak amplitude of the frequency response of the system.

3.5 Relations regarding input and output time histories

The following useful equations relate the time histories of input and output:

$$\begin{Bmatrix} x(t) \\ y(t) \end{Bmatrix} = \exp(i\omega t) \begin{Bmatrix} \partial X(t, i\omega)/\partial t \\ \partial Y(t, i\omega)/\partial t \end{Bmatrix}$$

$$= \exp(-i\omega t) \begin{Bmatrix} \partial X^c(t, i\omega)/\partial t \\ \partial Y^c(t, i\omega)/\partial t \end{Bmatrix} \tag{19}$$

$$x(t) y(t) = \frac{1}{2\pi} \int_0^\infty \mathrm{Re}\,[H(i\omega)] \rho_x(t, i\omega) d\omega \tag{20}$$

$$y(t) = \sum_{j=0}^{n} \frac{m_j x^{(j)}(t)}{j!} \tag{21}$$

$$x(t) = \frac{1}{m_0} \sum_{j=0}^{n} \frac{n_j y^{(i)}(t)}{j!}, \tag{22}$$

where

$$m_j = \int_{-\infty}^{\infty} t^j h(t)dt, \qquad n_0 = (m_1/m_0), \qquad n_1 = (m_1^2/m_0 - m_2/2),$$

$$n_2 = [m_1^3/m_0 - (1 + m_0)m_1m_2/2m_0 - m_3/6m_0], \cdots, \quad \text{etc.},$$

and the superscript c in (19) indicates the complex conjugate.

Equation (19) expresses the waveform of input or output in terms of the corresponding running spectrum and can be used to recover it; (20) gives the instantaneous value of the multiple of $x(t)$ and $y(t)$ as a simple frequency integral in terms of the transfer function and instantaneous power spectrum density of the input; (21) [or (22)] expresses the output (or the input) waveform in terms of the moments of the impulse response function of the system and the derivatives of the input (or the output). Liu[13] derives (19) and (20). Papoulis[14] uses Taylor's expansion to derive (21) and (22). Note that (21) and (22) are useful only if the two series converge rapidly; i.e., if the transfer function $H(i\omega)$ only takes significant values and can be approximated by a polynomial of low order within the region of $|\omega| \leqq \Omega_Y$.

IV. SHOCK AND VIBRATION ANALYSIS OF SIMPLE STRUCTURAL SYSTEMS

The relations presented above can be applied in the shock and vibration analysis of structures. Since many structures can be treated as systems with a single mode of oscillation and many others can be solved by superposition of various modal responses in the generalized coordinates, the application can be well illustrated by the analysis of simple structures. We therefore consider a set of single-degree-of-freedom systems with natural frequency in the range of 0.05 to 10 Hz (most commercial and industrial structures have natural frequency in this range), and having a damping ratio of $\lambda = 2$ percent and 10 percent, respectively. These structures are subjected to a random transient load resembling a strong-motion earthquake accelerogram (Fig. 1) and four different types of idealized impulses resembling air or ground shocks (Fig. 2). The various bounds of the relative displacement and acceleration responses of the structures are calculated according to (12) and compared with the actual maximum response y_m. The results in Fig. 3 indicate that, for earthquake type of excitations, the lower bound l_1 and the upper bound u_3 (with $n = 0$) are the better bounds; u_4 and u_1 also appear as reasonably good upper bounds although, for the acceleration response, they show wider fluctuations than u_3. Bounds l_2 and u_5 fluctuate with Y_m and vary with the natural

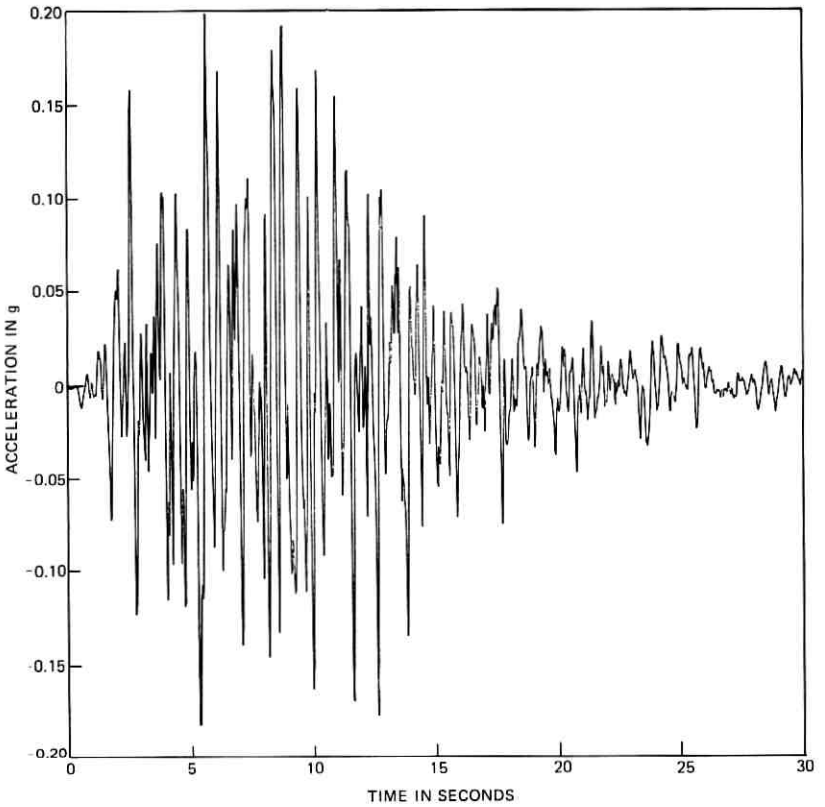


Fig. 1—A random transient waveform.

frequency of the structure, ω_o; therefore, both have the same degree of closeness to the real response. Bounds u_6, u_8, and l_2 provide rough estimates of the response; this is expected because little information is needed about the input, the output, or the structure. Bounds u_6 and u_8 are constant for the acceleration response (see Figs. 3a and 3b) because, for a fixed input, E_x and X_m are both constant, and the simple structures considered, $H_m = (2\lambda)^{-1}(1 - \lambda^2)^{-\frac{1}{2}}$, i.e., the maximum spectrum of the transfer function associated with the relative acceleration response is a constant for a given damping, and is independent of ω_o. Bounds u_6 and u_8, for the displacement response, decrease with ω_o (see Figs. 3c and 3d) because the associated $H_m = (2\lambda p)^{-1}$. Bounds u_2 and u_7, for the acceleration response, grow with $E_h^{\frac{1}{2}}$, which increases with ω_o. For the displacement response, u_2 appears also to be a reasonably good bound.

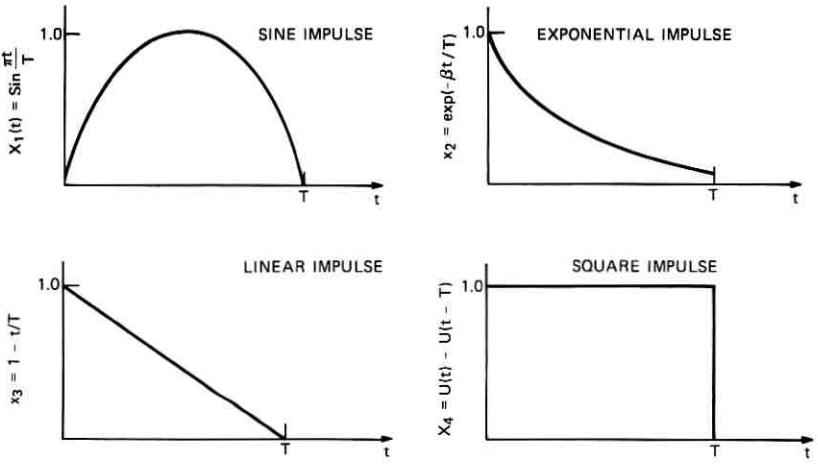


Fig. 2—Impulse loads.

The results for various types of impulse loads are generally similar and can be represented by Fig. 4. For impulsive types of excitations, we can make the following general observations: (*i*) l_1 seems to be a very good bound for most cases ($l_1 \approx y_m$ when $\omega \to 0$ or when ω is small), (*ii*) the best upper bound is u_3 ($n = 0$), which is very close to y_m, particularly for the displacement response, (*iii*) u_6 and u_8, because they are constant for the acceleration response and decrease with ω_o for displacement response, do not appear to be good upper bounds, (*iv*) for the acceleration response, u_2 and u_7 increase with ω_o; the rate of increase is reduced in high-damping cases, (*v*) for the displacement response, $u_5 \to u_8$ and $u_7 \to u_1$, when $\omega \to 0$, and (*vi*) next to u_3, u_2 and u_4 seem adequate as upper bounds.

V. APPLICATIONS TO TELEPHONE STRUCTURES IN AN EARTHQUAKE AREA

The above analysis can be applied to shock and vibration problems of telephone structures. Examples are the dynamic response analysis and aseismic design of various telephone building and equipment systems located in an earthquake area. Some of such applications will be described below.

5.1 Aseismic design of battery stands

It has been recognized[15] that, in large areas of the country susceptible to earthquakes, battery stands on floors of multistory telephone build-

ings require special physical design consideration to protect the safety of operating personnel and the reliability of telephone plant. Figure 5 shows a typical soft-site stand modified for service in earthquake areas. It consists of bracing to support the stand laterally from an overhead ironwork structure and to support the cells on the shelves. Clearance is provided around the entire cell to allow it to rattle around freely in response to ground motion. The failure of the battery stand generally occurs from excessive acceleration, which could lead to sliding and rattling of cells and their banging against one another and against the side racks, and from excessive displacement of the stand that could result in overstress of the column members of the stand. Therefore, the displacement and acceleration responses of the stand are of crucial concern to its aseismic adequacy. These response parameters can be estimated very conveniently by the response-bound relations derived in Section III. We assume that all that is known about the physical property of the battery stand is that it has a damping ratio about 2 percent and a natural frequency in the range of 3 to 5 Hz.[16] It is reasonable to assume that the floor motion corresponding to a large earthquake resembles the waveform shown in Fig. 1. The peak acceleration amplitude would be in the range of 0.75 g because of the soil and building amplification of the ground motion. From the analysis in Section IV, the response bounds of the battery stands shown in Fig. 6 can be readily established. From the acceleration lower bound l_1 in Fig. 6 that provides an underestimate of the actual responses, it is immediately clear that the natural frequency of the battery stand in the range of 1.5 to 3 would not be desirable. Because the high acceleration level in this range would cause lateral forces exceeding the friction force between the base of the cells and the supporting shelves, the battery would then slide and rattle violently in response to ground motion. The displacement bounds u_3 and l_1 indicate a continuously decreasing trend with the natural frequency of the stand. The upper bound u_3, which provides a conservative estimate, indicates that, if the battery stand is designed to have a natural frequency higher than 5 Hz, the actual maximum displacement response would be less than the acceptable limit, e.g., about 3 inches. We can therefore conclude from such a simple analysis that, from consideration of both the acceleration and displacement responses, it is desirable to increase the frequency of the battery stand up to about 5 Hz or above. The increase of the frequency or, equivalently, the stiffness of the stand can be generally achieved by adding bracings to the wall, columns, or to superstructures such as the auxiliary frames.

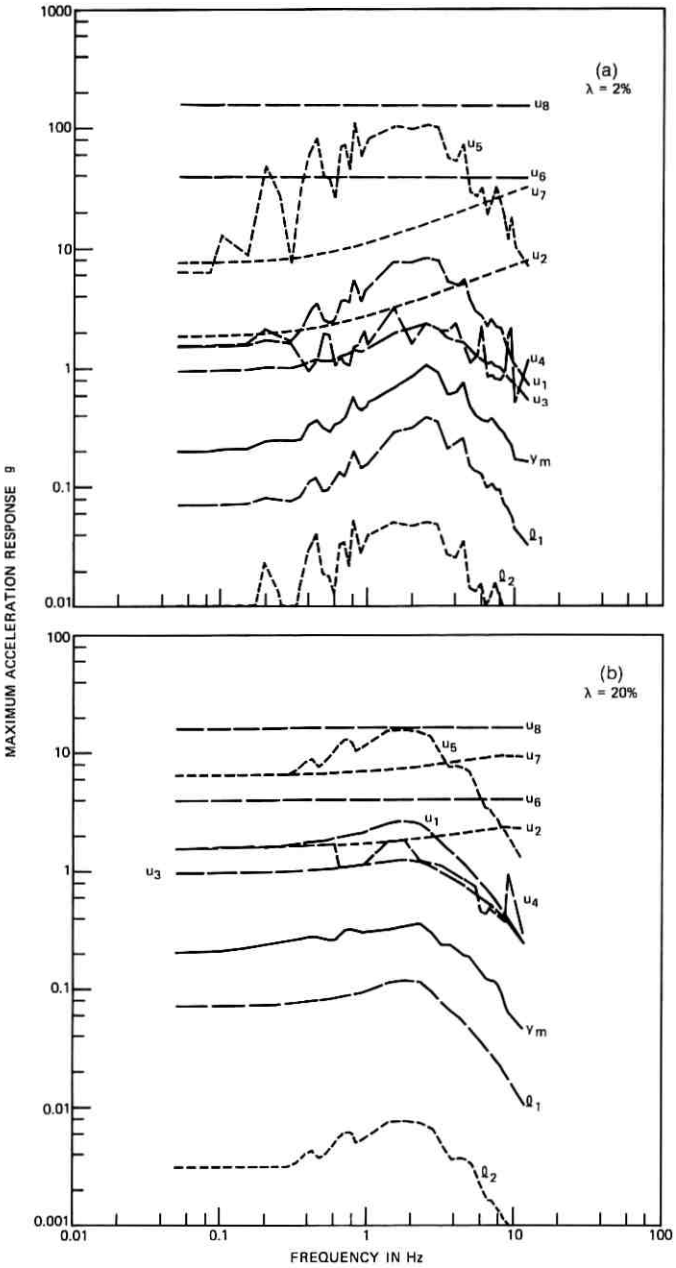


Fig. 3—Response bounds of single-degree-of-freedom structures subjected to random load.

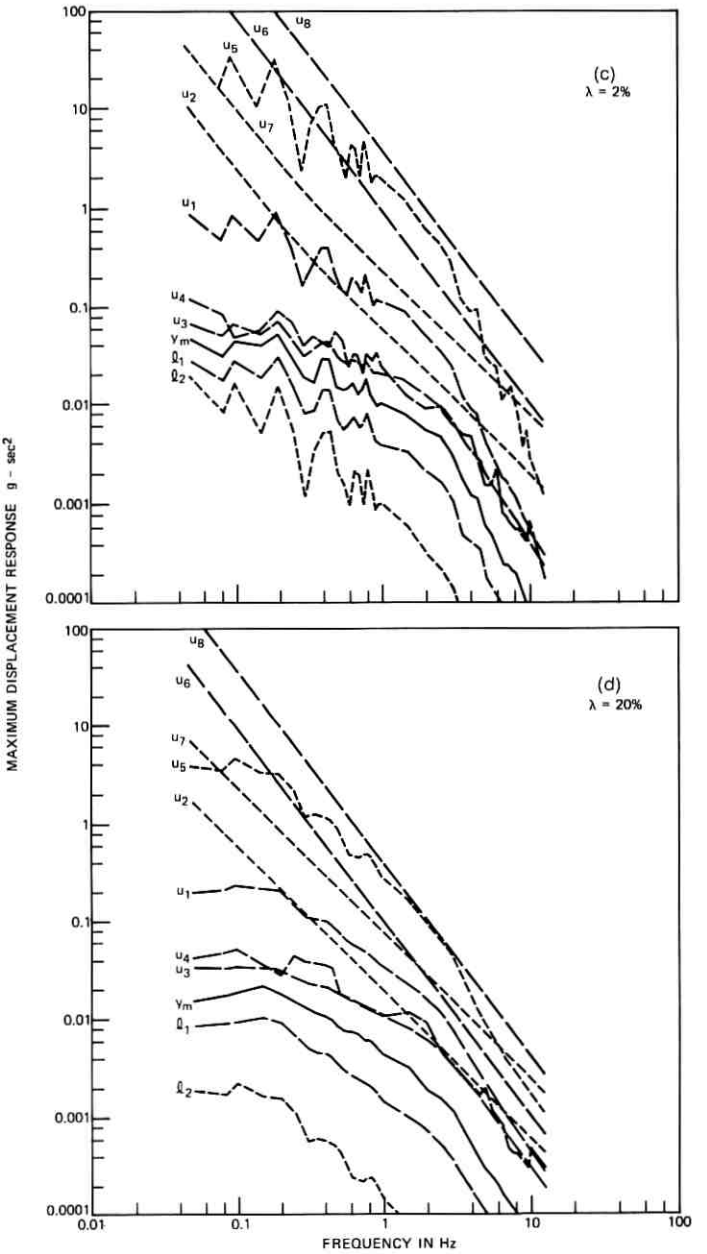


Fig. 3 (continued).

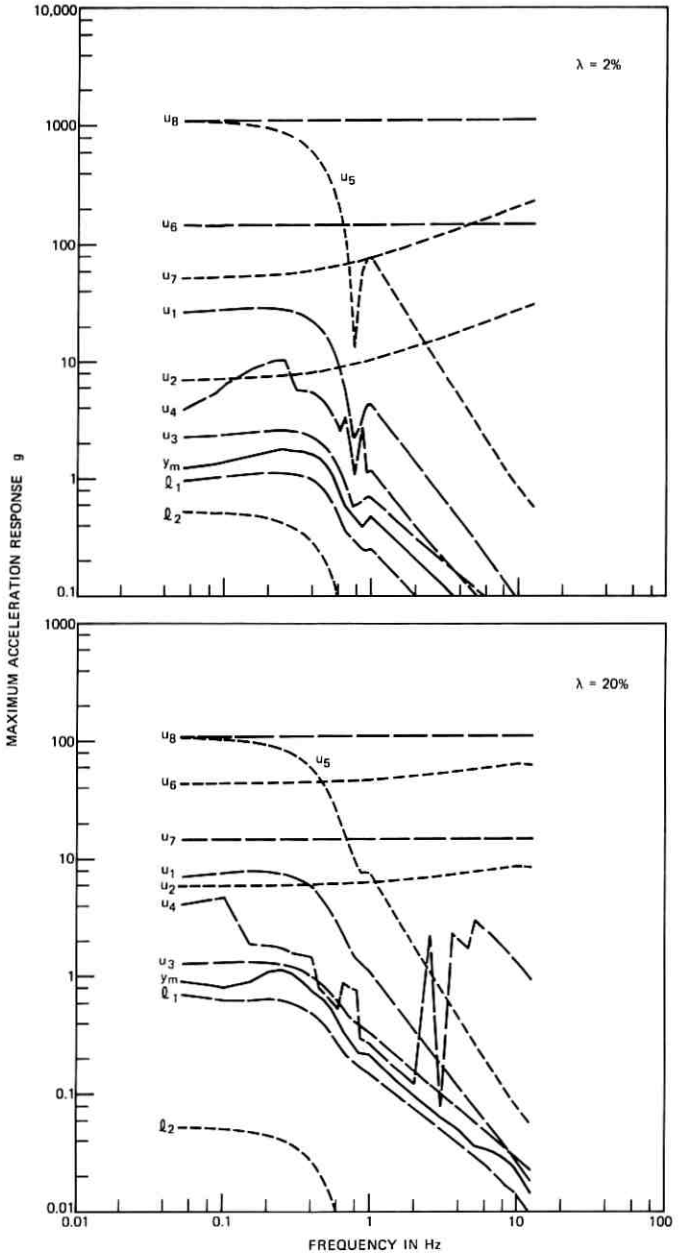


Fig. 4—Response bounds of single-degree-of-freedom structures subjected to sine impulse.

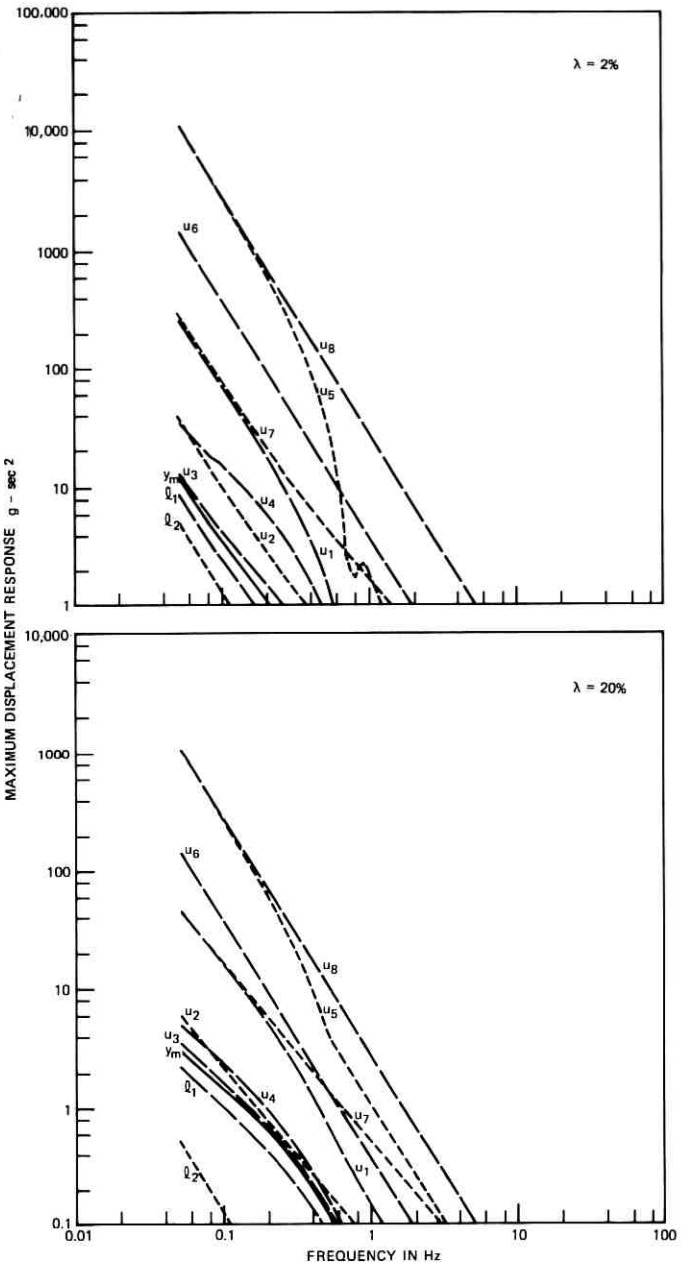


Fig. 4 (continued).

Fig. 5—Soft-site battery stand modified for service in earthquake area.

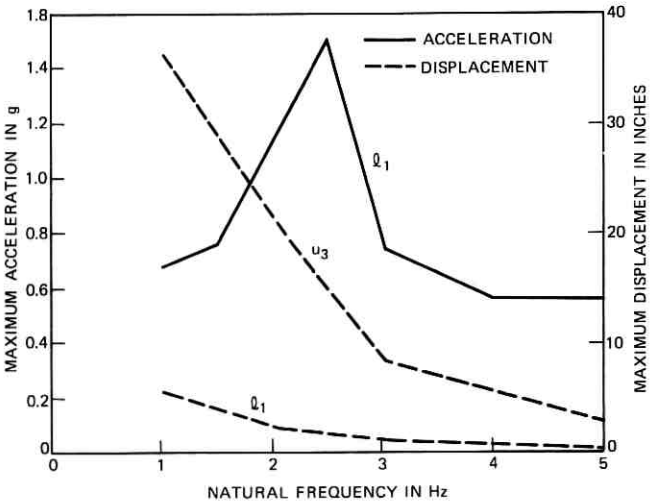


Fig. 6—Response bounds of battery stands.

5.2 Earthquake adequacy of community dial offices

A community dial office (CDO) is a small telephone office housing mostly step-by-step equipment servicing a local community. The CDO buildings are generally one-story concrete block constructions, with a few steel-frame constructions as exceptions.

Tests and analyses relating to typical CDO facilities[17,18] have indicated seismic weaknesses in the equipment connective system, particularly at the junction of the superstructure bracing system to the building. Figure 7 shows a schematic cross-sectional representation of a typical CDO equipment support structure. Equipment frame lineups are generally tied together at the 9-foot level by cross-aisle cable racks and angle braces (indicated as lineup braces). On the common distribution frame (CDF) side, the frames are attached to a wood batten on the building wall by horizontally oriented cable racks and angle braces (indicated as wall braces).

While there are several possible failure modes of CDO equipment, the most likely and the most critical one is the failure of the lateral

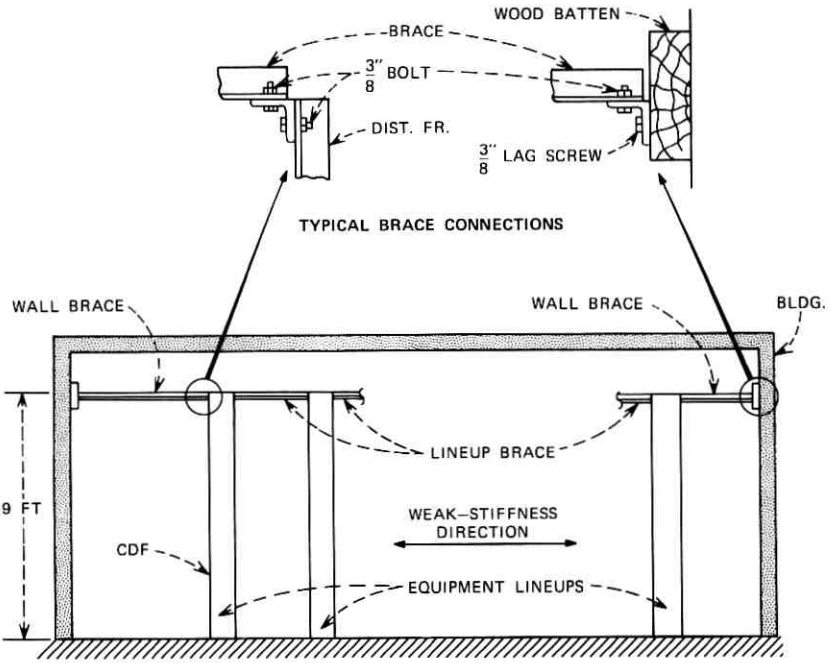


Fig. 7—Typical CDO support system.

support system during severe ground shaking. The wood batten shown in Fig. 7 is attached to the building wall by either lag screws into wood studs, lag into lead expansion shields inserted into concrete blocks, or toggle bolts into concrete blocks, depending on the building construction. This batten could pull away from the wall and result in the collapse of the equipment. To determine whether such a failure might occur, it is necessary to compare the axial force of the wall bracing development during the earthquake with the allowable force. The wall force is given as $F = EA\delta/l$, where E is the modulus of elasticity, A and L are the cross-sectional area and length of the wall brace, respectively, and δ is the relative displacement between the two ends of the wall brace. From testing or analysis we assume the transfer function $H_\delta(\omega)$ for the relative displacement response δ of a hypothetical CDO system as given in Fig. 8. We wish to find the earthquake-resistant capacity of this system, e.g., the highest ground acceleration this system can safely withstand.

A typical wall brace is made of steel with $E = 30 \times 10^6$ pounds per square inch, $A = 0.53$ square inch, and $L = 5.7$ inches. By testing we found that the maximum axial force the wall brace can take without pulling out of lag screws or toggle bolts is 900 pounds, which corresponds to a maximum lateral relative displacement of $\delta = 0.003$ inch between the two ends of wall brace. Based on the transfer function, $H_\delta(\omega)$, and the results described in Section IV, we found that a conservative estimate (u_3) of the maximum displacement corresponding

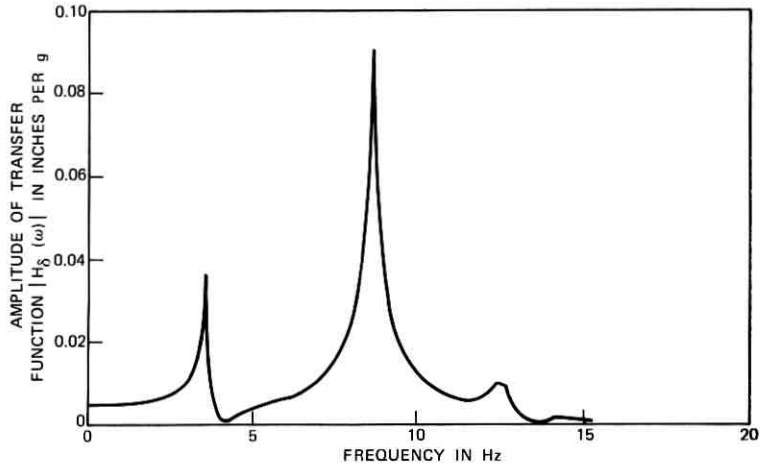


Fig. 8—Transfer function of CDO system.

to a 0.2 g earthquake loading is approximately 0.002 inch. Assuming the system remains linearly independent of the input amplitude, we can conclude that the hypothetical CDO system under consideration is capable of resisting earthquake ground acceleration up to about 0.3 g, a motion equivalent in magnitude to the El Centro 1940 earthquake. Additional braces would be needed if higher acceleration level is expected at the site.

VI. CONCLUSION

Various equality and inequality input-transfer-output relations are derived to facilitate the application of the Fourier transform method in the shock and vibration analysis of linear structural systems. In particular, a set of lower and upper bounds of the maximum response is obtained. Such bounds allow various maximum response parameters to be easily estimated with little computation effort. The accuracy of the estimate depends on the degree of availability of information about the input, the transfer function, and the response of the system. In general, l_1 and u_3 appear to be reasonably good lower and upper estimates of the maximum response, respectively. Estimates of response bounds of simple systems are examined when subjected to a random transient excitation and impulse loadings. For systems whose transform functions roll off quickly at high frequencies, or for inputs that have fast-decaying Fourier spectra, the relations presented apply reasonably well. Care should be taken with short-duration impulses that have wide spectra. Telephone structures subjected to earthquake excitations, e.g., a conventional battery stand and a CDO system, are considered to demonstrate the practical applications.

VII. ACKNOWLEDGMENTS

The author wishes to thank F. X. Prendergast for careful review of the manuscript, N. J. DeCapua for providing information used in Section V, and M. Dougherty for performing most of the numerical analysis of this paper.

REFERENCES

1. I. N. Sneddon, *Fourier Transforms*, New York: McGraw-Hill, 1951.
2. C. L. Amba-Rao, "Fourier Transform Methods in Elasticity Problems and an Application," J. Franklin Inst., *287*, No. 3 (March 1969), pp. 241–269.
3. P. H. White, "Application of the Fast Fourier Transform to Linear, Distribution Systems Response Calculations," J. Acoust. Soc. Am., Pt 2, *46*, No. 1 (1969), pp. 273–274.
4. R. C. Le Bail, "Use of Fast Fourier Transforms for Solving Partial Differential Equations in Physics," J. Comput. Physics, *9*, 1972, pp. 440–465.

5. S. C. Liu and L. W. Fagel, "Earthquake Interaction by Fast Fourier Transform," J. EM Div., ASCE, *97*, No. EM4 (August 1971), pp. 1223–1237.
6. L. W. Fagel and S. C. Liu, "Earthquake Interaction for Multistory Buildings," J. EM Div., ASCE, *98*, No. EM4 (August 1972).
7. S. C. Kak, "Causality and Limits on Frequency Functions," Int. J. Electronics, *32*, No. 1 (1971), pp. 41–47.
8. A. Papoulis, "Limits on Bandlimited Signals," Proc. IEEE, *55*, No. 10 (October 1967), pp. 1677–1686.
9. A. Papoulis, "Maximum Response with Input Energy Constraints and the Matched Filter Principle," IEEE Trans. on Circuit Theory, *CT-17*, No. 2 (May 1970), pp. 175–182.
10. S. C. Liu, "An Approach to Time-Varying Spectral Analysis," J. EM Div., ASCE, *98*, No. EM1, Proc. paper 8733 (February 1973), pp. 245–253.
11. C. R. Giardina, "An Upper Bound on the Settling Time," IEEE Trans. on Circuit Theory, *CT-19*, No. 9 (July 1972), pp. 372–373.
12. R. F. Drenick, "Model-Free Design of Aseismic Structures," J. EM Div., ASCE, *96*, No. EM4, Proc. paper 7496 (August 1970), pp. 483–493.
13. S. C. Liu, "Time-Varying Spectra and Linear Transformation," B.S.T.J., *50*, No. 7 (September 1971), pp. 2365–2374.
14. A. Papoulis, "Approximations of Point Spread for Deconvolution," J. Opt. Soc. Am., *62*, No. 1 (January 1972), pp. 77–80.
15. H. J. Luer, "Incorporating the New Battery into the Telephone Plant," B.S.T.J., *49*, No. 7 (September 1970), pp. 1447–1470.
16. N. J. DeCapua and M. G. Hetman, "Earthquake Analysis of Battery Stands," unpublished work.
17. R. M. Riley, private communication to W. H. Turner, 1974.
18. N. J. DeCapua and F. X. Prendergast, unpublished work.

Contributors to This Issue

Jacques A. Arnaud, Dipl. Ing., 1953, Ecole Supérieure d'Electricité, Paris, France; Docteur Ing., 1963, University of Paris; Docteur es Science, 1972, University of Paris; Assistant at E.S.E., 1953–1955; C.S.F., Centre de Recherche de Corbeville, Orsay, France, 1955–1966; Warnecke Elec. Tubes, Des Plaines, Illinois, 1966–1967; Bell Laboratories, 1967—. At C.S.F., Mr. Arnaud was engaged in research on high-power traveling-wave tubes and supervised a group working on noise generators. He is a supervisor, currently studying microwave quasi-optical devices and the theory of optical wave propagation. Senior Member, IEEE; Member, Optical Society of America.

William F. Bodtmann, Monmouth College, 1957–61; Bell Laboratories, 1941—. Mr. Bodtmann has been engaged in research on long- and short-haul microwave radio systems, frequency feedback receivers, and FM multiplex systems. He is presently engaged in work associated with communication systems operating at millimeter wavelengths.

Corrado Dragone, Laurea in E.E., 1961, Padua University (Italy); Libera Docenza, 1968, Ministero della Pubblica Istruzione (Italy); Bell Laboratories, 1961—. Mr Dragone has been engaged in experimental and theoretical work on microwave antennas and solid-state power sources. He is currently concerned with problems involving electromagnetic wave propagation and microwave antennas.

Bernard Glance, Dipl. Ing., 1958, Ecole Speciale de Mecanique et Electricité, Dipl. Ing., 1960, Ecole Supérieure d'Electricité, Paris (France); C.S.F., Research Center of Corbeville, Orsay, France, 1960–1966; Dipl. Docteur Ing., Sorbonne, Paris, 1964; S.F.D. Laboratories (subsidiary of Varian Assoc.), Union, N. J. 1966–1968; Bell Laboratories, 1968—. At C.S.F., Mr. Glance was engaged in research on microwave tubes. At S.F.D. Laboratories, he worked on high-power microwave amplifiers. Since joining Bell Laboratories, he has been working on microwave solid-state integrated circuits.

Beat Kleiner, Diploma (Mathematics), 1968; Dr. Sc. Math., 1971, Swiss Federal Institute of Technology, Zurich, Switzerland; University of California, Berkeley, 1971–1972; Bell Laboratories, 1972—. Mr.

Kleiner worked initially on statistical analysis of brainwaves; after joining Bell Laboratories he was mainly engaged in time series analysis. He is currently involved in a large statistical investigation of air pollution data.

S. C. Liu, B.S.C.E., 1960, National Taiwan University; M.S., 1964, and Ph.D., 1967, University of California at Berkeley; Bell Laboratories, 1967—. Mr. Liu has been working in the areas of mechanical vibrations, random processes, structural dynamics, and earthquake engineering. Member, American Society of Civil Engineers.

F. W. Ostermayer, Jr., S.B. and S.M. (Electrical Engineering), 1959, Massachusetts Institute of Technology; Ph.D. (Physics), 1968, Lehigh University; Bell Laboratories, 1959–1965, 1968—. Mr. Ostermayer initially worked on microwave masers and ferrite devices. Since rejoining Bell Laboratories in 1968, he has worked on infrared-to-visible conversion phosphors, electroluminescent pumping of YAG:Nd lasers, and, presently, on optical fiber waveguides. Member, American Physical Society, Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

Douglas A. Pinnow, B. Eng. Phys., 1961, Cornell University; M.S., 1965, and Ph.D., 1967, the Catholic University of America; Naval Reactors Division of the U. S. Atomic Energy Commission, 1961–1965; Bell Laboratories, 1967—. Mr. Pinnow has been engaged in research and exploratory development in ultrasonics, nonlinear optics, and optical information processing and display. Since 1970 he has been involved principally in the evaluation and development of fiber optical waveguides and waveguide materials. Member, Optical Society of America, IEEE, American Ceramic Society, Tau Beta Pi.

Clyde L. Ruthroff, B.S.E.E., 1950, and M.A., 1952, University of Nebraska; Bell Laboratories, 1952—. Mr. Ruthroff has published contributions on the subjects of FM distortion theory, broadband transformers, FM limiters, threshold extension by feedback, microwave radio systems, rain attenuation, multiple-path propagation, linear phase modulators, injection-locked FM receivers, and coherent detection. He is interested in the extension of radio communication into the millimeter and optical wavelengths. Fellow, IEEE; member, Sigma Xi and the American Association for the Advancement of Science.

William E. Schroeder, B.S. (E.E.), 1965, M.S. (E.E.), 1967, Ph.D. (E.E.), 1972, The University of Michigan; Bell Laboratories, 1972—. Mr. Schroeder has worked on microwave solid-state devices and is currently studying circuit problems associated with such devices. Member, Eta Kappa Nu, Tau Beta Pi, Sigma Xi, IEEE.

William W. Snell, Jr., Williamsport Technical Institute, 1951; Bell Laboratories, 1955—. In his first years with Bell Laboratories, Mr. Snell was concerned with the design of waveguide components for use in the 4-, 6-, and 11-GHz common carrier band. He later participated in the Shotput and Project Echo satellite communications experiments, during which he designed several components of the Holmdel Space Communication Receiver. He is presently concerned with the design of hybrid integrated circuit frequency-converting devices for use at frequencies above 10 GHz.

Barton W. Stuck, B.S.E.E., M.S.E.E., 1969, Massachusetts Institute of Technology; Sc.D., 1972, Massachusetts Institute of Technology; Bell Laboratories, 1972—. Mr. Stuck is currently working on problems in mathematical physics, and detection and estimation of signals corrupted by nongaussian noise. Member, SIAM, MAA, IEEE.

K. K. Thornber, B.S., 1963, M.S. (E.E.), 1964, Ph.D. (E.E.), 1966, California Institute of Technology; Research Associate, Stanford Electronics Laboratories, 1966–68; Research Assistant, Physics Department, University of Bristol, 1968–69; Bell Laboratories, 1969—. Mr. Thornber is a member of the Unipolar Integrated Circuit Laboratory. Member, Sigma Xi, Tau Beta Pi.