THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 51 December 1972 Number 10

On Delta Modulation

By DAVID SLEPIAN

(Manuscript received May 3, 1972)

We show how the steady-state distribution and the mean squared error of a delta modulator with an ideal integrator can be computed exactly when the input signal to the modulator is a stationary Gaussian process with a rational power spectral density. Curves are presented for the mean squared error as a function of the quantizer step size and the sampling interval for several different input spectra. The mathematical development makes use of the Markov properties of the system and involves series ex-expansions in n-dimensional Hermite functions. The key integral equation is generalized to treat the case of a realizable filter in the feedback path, but an analytic method of solving this equation has not been found.

I. INTRODUCTION

Demand for the transmission of digital data grows apace as the computerization of our society continues. This demand, coupled with the many recent striking advances in solid state circuit technology and with new concepts of digital switching, assures an increased role for digital transmission systems in the near future. The existence of such systems in turn gives new importance to digital means of transmitting analog signals. This paper is concerned with one such means—delta modulation.

In its simplest form, as depicted in Fig. 1, the delta modulation transmitter approximates a continuous input signal $X(t)$ by a staircase signal $Z(t)$ that has treads of duration T and risers of height Δ. Every T seconds the staircase either rises one step or falls one step in order to approach $X(t)$ at that instant more closely. At each rise or fall, the delta modulator emits a binary digit that specifies the direction of the step just taken. At the receiver, these transmitted binary digits are then used to reconstruct $Z(t)$, or perhaps a smoothed version of it.

This system was first described in the literature in 1952.[1] Because of its extreme conceptual simplicity, and its relative ease of instrumentation, delta modulation has attracted the attention of theorists and experimentalists alike, and many studies of it and its generalizations have been undertaken in the ensuring years. Many of these have been concerned with calculation or measurement of the mean squared error suffered by signals transmitted by delta modulation and with determination of how this quantity varies with the parameters of the system. Almost without exception, the theoretical studies are based on approximations, the range of validity of which is difficult to determine.

The present paper is also concerned with the mean squared error inherent in delta modulation. Our attention is focused on stationary Gaussian input ensembles $X(t)$ that have rational power density spectra. For this class of inputs we show that the mean squared error can indeed be computed exactly for the simple modulator of Fig. 1.

Since the mathematical analysis entailed tends to become quite involved, we have organized the paper into three main parts. Section II presents definitions, discussion and the results of numerical work. It is free of laborious mathematical derivations and is intended for the

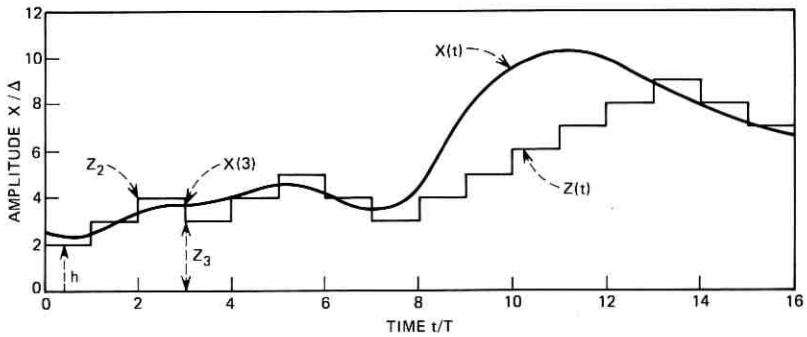

Fig. 1—The waveforms of a simple delta modulator with ideal integrator.

casual reader. In Section III a detailed mathematical treatment leading to a means for computing the mean squared error is given along with some necessary additional theory. Section IV describes a generalization of the present study to systems with realizable filters in the feedback loop.

II. DEFINITIONS, DISCUSSION AND RESULTS

2.1 *Some Definitions and Descriptions*

The modulator described in the Introduction can be defined with mathematical precision as follows. A signal $X(t)$ is given for $t \geqq 0$. Also given are a sampling period $T > 0$, a step size $\Delta > 0$, and an initial value h. Numbers Z_j, $j = 0, 1, 2, \cdots$ are defined recursively by

$$Z_0 = h,$$
$$Z_j = \begin{cases} Z_{j-1} + \Delta, & X(jT) > Z_{j-1} \\ Z_{j-1} - \Delta, & X(jT) \leqq Z_{j-1}, \end{cases} \qquad j = 1, 2, 3, \cdots . \tag{1}$$

The delta modulation approximation signal is then given by

$$Z(t) = Z_j, \qquad jT \leqq t < (j+1)T, \qquad j = 0, 1, 2, \cdots . \tag{2}$$

Notice that $Z(t)$ can only take on values from the set $\mathcal{S} \equiv \{\cdots h - 2\Delta, h - \Delta, h, h + \Delta, h + 2\Delta, \cdots\}$. Indeed, the allowed values of $Z(t)$ are restricted in a periodic way. If t lies in an even interval, i.e., if $2nT \leqq t < (2n+1)T$ for some $n = 0, 1, 2, \cdots$, then $Z(t)$ must take a value from the set

$$\mathcal{S}_e \equiv \{\cdots h - 4\Delta, h - 2\Delta, h, h + 2\Delta, h + 4\Delta, \cdots\}. \tag{3}$$

If t lies in an odd interval, i.e. if $(2n+1)T \leqq t < (2n+2)T$ for some $n = 0, 1, 2, \cdots$, then $Z(t)$ must take a value from the set

$$\mathcal{S}_o \equiv \{\cdots h - 3\Delta, h - \Delta, h + \Delta, h + 3\Delta, \cdots\}. \tag{4}$$

Due to the non-linear nature of (1), it is very difficult to say much about how well $Z(t)$ approximates any given signal $X(t)$, nor is this question of any real importance in a communication setting. What matters is how well $Z(t)$ does on the average in approximating the members of an ensemble of functions that represents an analog information source. Thus we are led to consider the delta modulator described by (1) and (2) when $X(t)$ is a sample function of a stochastic process. Throughout the paper we shall restrict our attention to the case in which $X(t)$ is stationary and satisfies the conditions

$$EX(t) = 0, \qquad EX^2(t) = 1. \tag{5}$$

The latter constraint sets the scale by which signal power is measured.

With $X(t)$ stochastic, $Z(t)$ becomes a dependent stochastic process, and we can speak of the joint distribution of $X(t)$ and $Z(t)$ at any set of times, $0 \leqq t_1 < t_2 \cdots < t_n$. Even when $X(t)$ is a stationary process, $Z(t)$ will not in general be stationary, and the joint distribution of $X(t)$ and $Z(t)$ at times $jT + t_1, jT + t_2, \cdots, jT + t_n$ will depend on the integer j. Real world delta modulators, however, "settle down", and hence one would expect the distribution just referred to to approach a limit as $j \to \infty$. Unfortunately, there are some subtleties to this notion due to the periodic nature of the allowed values of $X(t)$, as already mentioned. Under suitable regularity assumptions, one limiting distribution will be approached as $j \to \infty$ through even values $j = 2m$, $m = 0, 1, 2, \cdots$; another will be obtained as $j \to \infty$ through odd values, $j = 2m + 1$, $m = 0, 1, 2, \cdots$. We call the average of these two limit distributions "the steady-state distribution." It describes the settled down behavior of the delta modulator. The marginal distribution of $X(t)$ computed from this steady-state distribution is, of course, still the original given distribution for $X(t)$.

The conditions under which the statistics of delta modulators approach limiting forms as just described have been investigated by Gersho.[2] His work shows that for the cases treated in this paper, the limits referred to above exist, and that the density of interest here is given by the unique normalized solution of our key equation (22).

We now measure the accuracy of the delta modulator by the mean squared error

$$\epsilon^2 = \epsilon^2(\Delta, T) \equiv E \frac{1}{T} \int_0^T [X(t) - Z(t)]^2 \, dt$$

where $X(t)$ and $Z(t)$ have the steady-state distribution and E denotes expectation. Our main interest is on how ϵ^2 varies with T, Δ, and the statistics of $X(t)$.

Delta modulation is frequently described by passing reference to a block diagram such as is shown in Fig. 2. (The box labelled "filter" is called a "perfect integrator" for the case at hand.) On the surface, this appears to be much more succinct than (1) and (2) and the subsequent limit discussions. Figure 2 describes a recursive situation, however, and so fails to define anything at all unless supplemented with side information that either permits the recursion to be started, or serves otherwise to define a joint distribution for $X(t)$ and $Z(t)$. Analyses of

delta modulation based on Fig. 2 with unstated initial conditions, and no limiting arguments are apt to be approximate.

2.2 *Some Heuristics and History*

Let us consider ϵ^2 as a function of Δ for a fixed sampling period T and for fixed-input ensemble statistics satisfying (5). If Δ is extremely large compared to unity, then for the most part of its history $Z(t)$ will alternate between the level h and one of the two levels $h + \Delta$ or $h - \Delta$. Thus one expects the asymptotic result

$$\epsilon^2(\Delta, T) \sim \tfrac{1}{2}\Delta^2$$

as $\Delta \to \infty$. On the other hand, if Δ is very small compared to unity, $Z(t)$ will rarely wander far from its initial value h and one expects the result

$$\lim_{\Delta \to 0} \epsilon^2(\Delta, T) = E[X(t) - h]^2 = 1 + h^2.$$

(The rate at which $\epsilon^2 \to 1 + h^2$ as $\Delta \to 0$ is a more subtle question that requires detailed analysis.) Thus the curve of $\epsilon^2(\Delta, T)$ vs Δ starts at $\epsilon^2 = 1 + h$ and ultimately rises like $\frac{1}{2}\Delta^2$. How does it behave in between? Does it always dip yielding a best value for Δ, i.e., a positive value for which ϵ^2 is least?

There have been many analyses of delta modulation in the past. The few listed here,[1,3-15] provide entry to the literature. Many of them predict the existence of a best $\Delta > 0$ for any T. Their analysis is based on the notion that the total error is the sum of two kinds of error–quantization error and slope-overload error. The delta modulation signal $Z(t)$ can climb or fall at a maximum average rate of $\Delta/T \equiv \xi$, so that if $| dX/dt |$ exceeds ξ for a length of time much greater than T, a serious tracking error will occur. Such a "region of slope overload" is seen in Fig. 1 for $8 \leqq t/T \leqq 11$. In the region $0 \leqq t/T \leqq 8$ of Fig. 1, $| dX/dt | < \xi$ and the error here is classified as "quantization error".

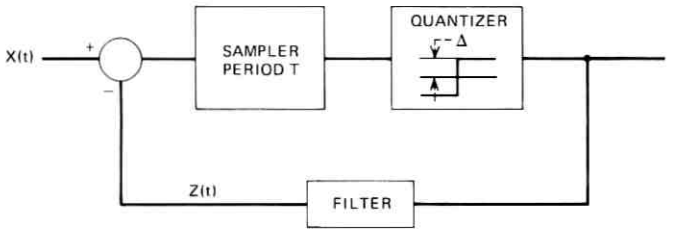

Fig. 2—Block diagram of delta modulator with general feedback filter.

This notion of two sorts of error has been very fruitful, and approximate calculations based on it agree well with experiment when T is small compared to any natural period associated with $X(t)$. This, of course, is the case of interest in practice. The calculations are based on many approximations, however, and it is difficult to determine their precise range of validity without recourse to experiment.

Two notable exceptions to this approach to the mean squared error are the exact treatments by Fine[10] and Aaron and Stanley.[11] The former treats a time discrete model with the input process $X(nT)$, $n = 0, 1, \cdots$, restricted to have independent increments. Aaron and Stanley treat the case in which $X(t)$ is a binary random telegraph signal. Neither of these cases is applicable to the transmission of speech or to continuous amplitude television signals. The work by Aaron and Stanley, however, has much of the flavor of the present study and presents a one-dimensional version of our key integral equation (22). Closely related work is also to be found in the papers of Davisson[13] who presents an integral equation and suggests a solution in a series of Hermite functions.

2.3 *Results of Computations*

The method described later in this paper, in principle, permits exact calculation of ϵ^2 whenever $X(t)$ is a stationary Gaussian process with a rational power spectral density,

$$\Phi(\omega) = K \frac{\prod_1^m (\omega^2 + c_i^2)}{\prod_1^n (\omega^2 + d_i^2)}, \tag{6}$$

where $m < n$ and $\omega = 2\pi f$ is the angular frequency. The complexity of the computation grows rapidly with n and consequently we have done numerical work only for $n = 1$ and $n = 2$. The method involves series that unfortunately converge slowly for small T, so that we have not been able to explore the interesting region of very small T.

Figure 3 shows plots of ϵ^2 vs Δ when the input process $X(t)$ has spectrum

$$\Phi_{RC}(\omega) = \frac{2}{1 + \omega^2}. \tag{7}$$

The corresponding covariance is

$$\rho_{RC}(\tau) \equiv EX(t)X(t + \tau) = e^{-|\tau|}. \tag{8}$$

We refer to this as the RC-noise case.

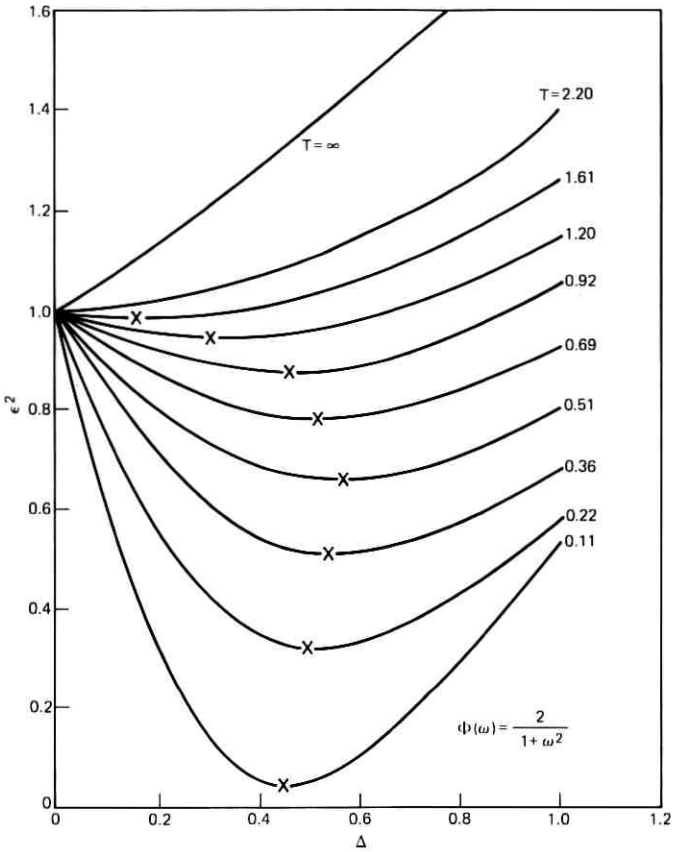

Fig. 3—Curves of ϵ^2 vs Δ for delta modulator with RC Gaussian-noise input.

On the curves of Fig. 3, a cross points out the optimal value of Δ, that is, the value $\Delta_{\min}(T)$ that minimizes ϵ^2. As T goes to zero, $\Delta_{\min}$ decreases (slowly) and the corresponding error decreases rapidly. As T increases, however, $\Delta_{\min}$ reaches a maximum, then starts to decrease once more toward zero. Note that for large sampling times ($T > 2.2$, say) the delta modulator performs poorly indeed. As far as mean squared error is concerned, at these rates one would do better by taking the constant zero as an approximation to the input than by using the delta modulation signal $Z(t)$.

Figures 4 and 5 show the somewhat similar results obtained for the

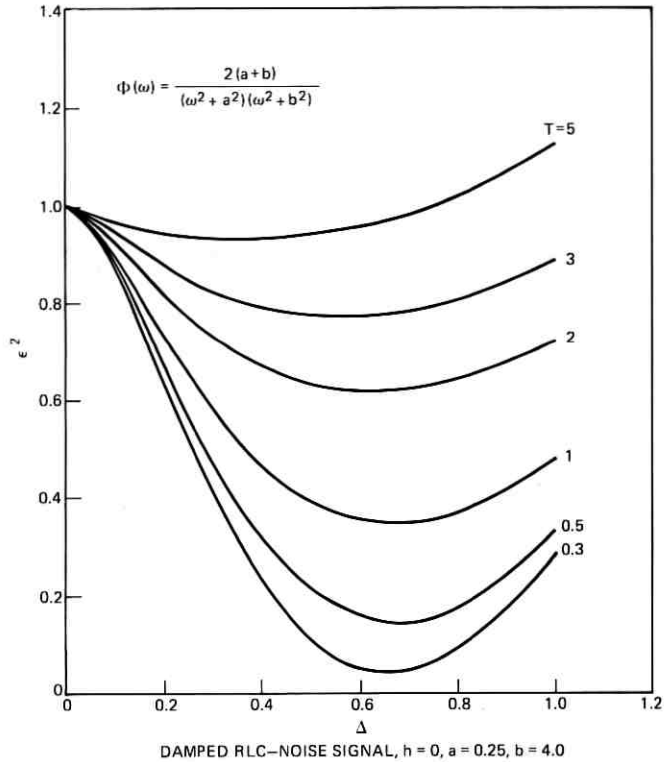

Fig. 4—Curves of ϵ^2 vs Δ for delta modulator with damped *RLC* Gaussian-noise input.

input spectrum

$$\Phi_{DRLC}(\omega) = \frac{2(b + a)}{(\omega^2 + a^2)(\omega^2 + b^2)}, \qquad ab = 1, \tag{9}$$

corresponding to the covariance

$$\rho_{DRLC}(\tau) = \frac{1}{b - a}\left[be^{-a|\tau|} - ae^{-b|\tau|}\right]. \tag{10}$$

When a and b are real, we refer to the input with spectrum (9) as damped *RLC* noise. The spectrum in this case is unimodal with its maximum at the origin.

When

$$a = \alpha + i\beta, \qquad b = \alpha - i\beta, \tag{11}$$

with α and β real, (9) and (10) become

$$\Phi_{RRLC}(\omega) = \frac{4\alpha}{[\omega^2 - (\beta^2 - \alpha^2)]^2 + 4\alpha^2\beta^2}, \qquad \alpha^2 + \beta^2 = 1, \tag{12}$$

$$\rho_{RRLC}(\tau) = \frac{e^{-\alpha|\tau|}}{\beta}[\alpha \sin \beta \mid \tau \mid + \beta \cos \beta\tau]. \tag{13}$$

For this "resonant *RLC* noise" case, the spectrum (12) develops a large narrow peak at $\omega = 1$ as $\alpha \to 0$. Figures 6 through 9 show the curious resonance phenomena that set in as $\alpha \to 0$ and the input signal becomes more and more sinusoidal in nature. For the limiting noise obtained

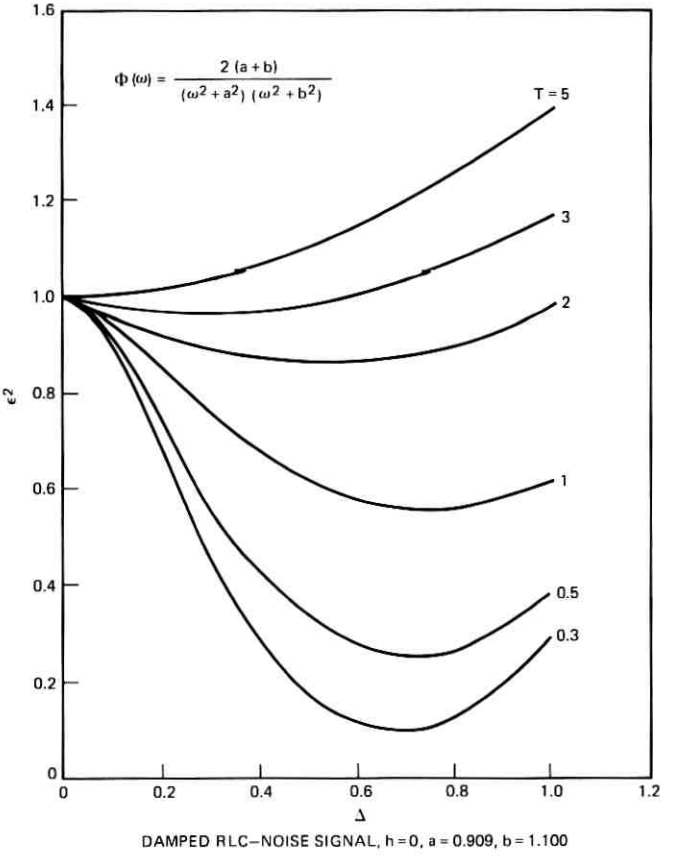

Fig. 5—Curves of ϵ^2 vs Δ for delta modulator with damped *RLC* Gaussian-noise input.

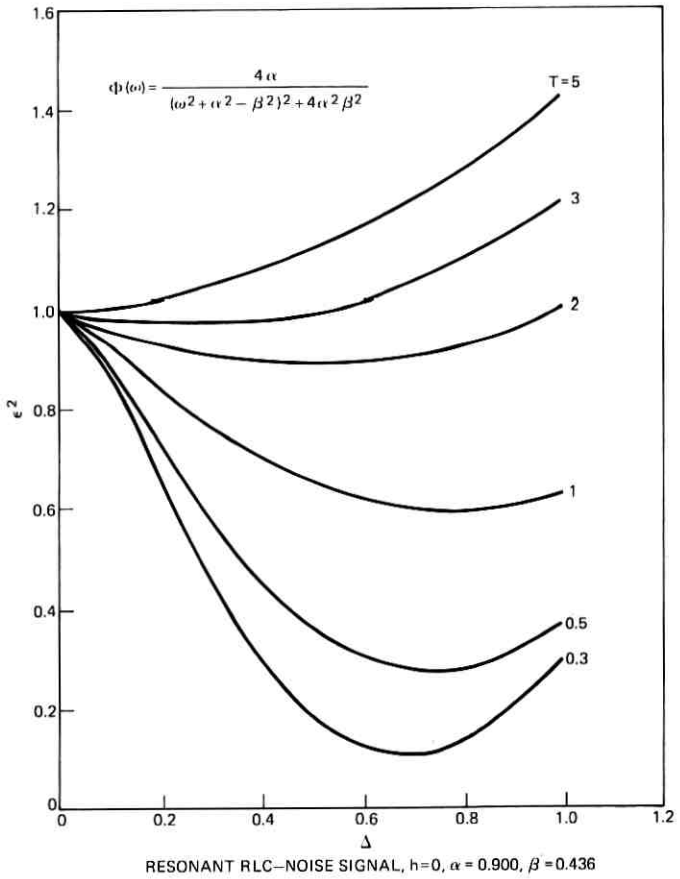

Fig. 6—Curves of ϵ^2 vs Δ for delta modulator with resonant *RLC* Gaussian-noise input.

when $\alpha \to 0$, the single frequency Gaussian ensemble, one can compute ϵ^2 by other methods and anomalous shapes for large T similar to those of Fig. 9 are found.

In the range $\epsilon^2 < 0.4$, the curves of Figs. 3 through 9 agree roughly with values computed by O'Neal[5] and others. These comparisons can, at best, yield approximate agreement, since they are among systems differing in a number of assumptions including the spectral shape of the signal. The approximate methods, based on quantization noise and slope-overload noise will probably continue to be used in practice, as they are much simpler to use than the scheme given here. The present

curves do, however, provide *exact* values for comparison purposes, and this is perhaps the main practical contribution of this paper.

2.4 *Outline of Mathematical Argument*

In this section we outline briefly the mathematical argument of Section III, and point out some of the formulae used to obtain the numerical results of the preceding section.

A stationary Gaussian process $X(t)$ with the rational power spectral density (6) can always be written as the first component of an n-vector Gaussian process

$$\mathbf{X}(t) = \{X_1(t) \equiv X(t), X_2(t), X_3(t), \cdots, X_n(t)\}$$

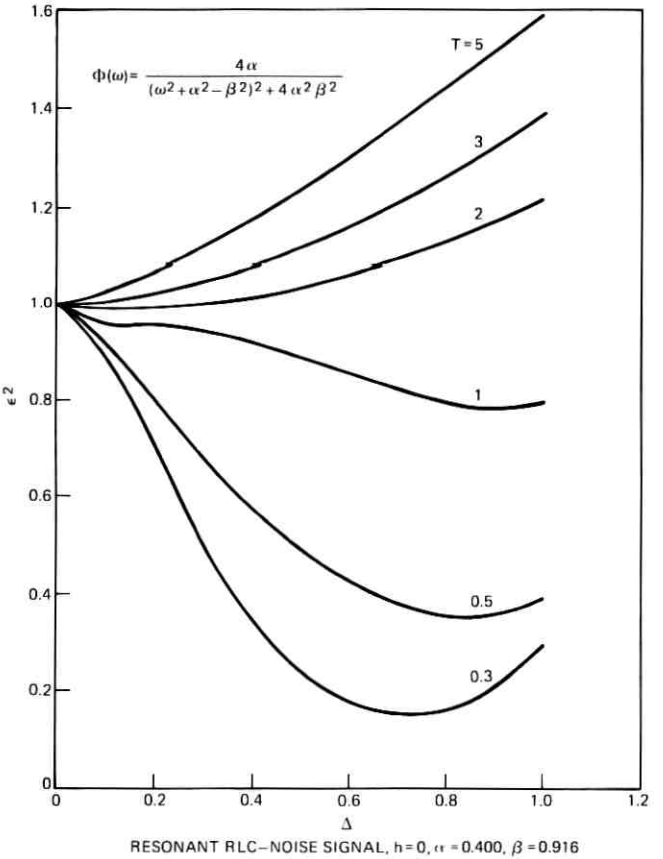

Fig. 7—Curves of ϵ^2 vs Δ for delta modulator with resonant *RLC* Gaussian-noise input.

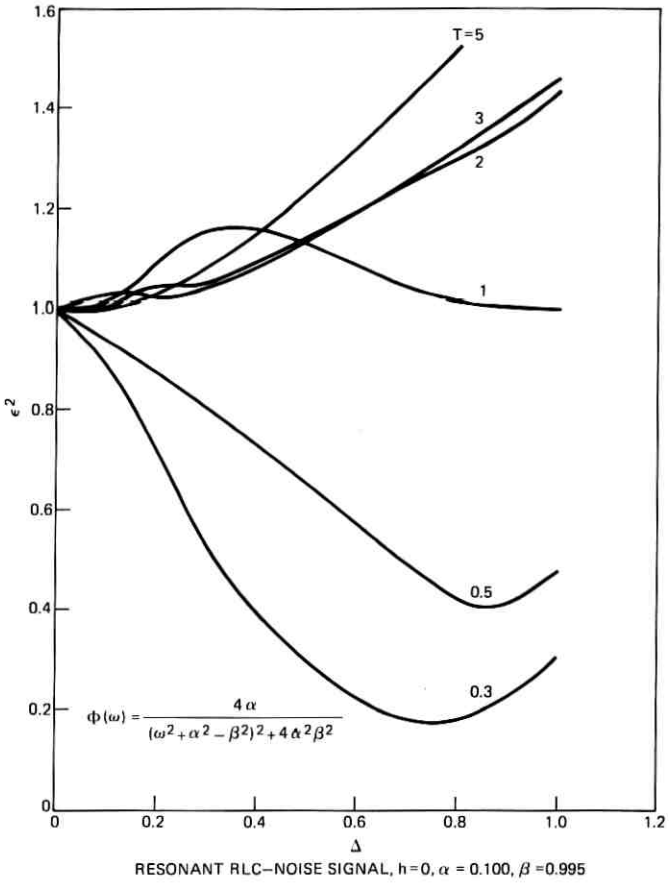

Fig. 8—Curves of ϵ^2 vs Δ for delta modulator with resonant *RLC* Gaussian-noise input–narrow band case.

that is Markovian.[16] It is not difficult to see then that the $n + 1$ quantities $X_1(jT)$, $X_2(jT)$, $\cdots$, $X_n(jT)$, Z_{j-1} for $j = 1, 2, \cdots$, form a time-discrete vector Markov process. The first n components can take any real values, but the last component is restricted to alternate between values in the sets $\mathcal{S}_e$ and $\mathcal{S}_o$ of (3) and (4). The stationary measure, or what we call the steady-state distribution, $m_i(\mathbf{x})$, satisfies the Chapman–Kolmogorov equation (22) with the boundary conditions (23). The notation is explained below (23).

The kernel $p_T(\mathbf{y} \mid \mathbf{x})$ of (22) can be developed in a multiple power series in the cross-correlations β_{ij} defined in (31). This power series resembles Mehler's formula and involves certain functions $\psi_l(\mathbf{x}; \boldsymbol{\alpha})$ of n

variables $x_1, x_2, \cdots, x_n$ that we call n-dimensional Hermite functions. They are defined in (24) and (39). The parameters $\boldsymbol{\alpha}$ here are the correlations (26). The expansion of the kernel is given in (46) in a highly symbolic form. To understand this equation fully, Section 3.2 and the first paragraph of 3.3 must be read.

The expansion (46), in turn, suggests the expansion (47) of the steady-state distribution. We write that symbolic equation in full here:

$$m_i(\mathbf{x}) = \sum_{\nu_{11}=0}^{\infty} \sum_{\nu_{12}=0}^{\infty} \cdots \sum_{\nu_{nn}=0}^{\infty} \sum_{l_1=0}^{\infty} \cdots \sum_{l_n=0}^{\infty} \left[\prod_{j=1}^{n} \prod_{k=1}^{n} \beta_{jk}^{\nu_{jk}} \right] f_{i\nu_{11}\nu_{12}\cdots\nu_{nn}l_1l_2\cdots l_n} \psi_{l_1,l_2,\cdots l_n}(\mathbf{x}, \boldsymbol{\alpha}).$$

Thus $\boldsymbol{\nu}$ is an $n \times n$ matrix of indices, and $\boldsymbol{l}$ is an n-vector of indices.

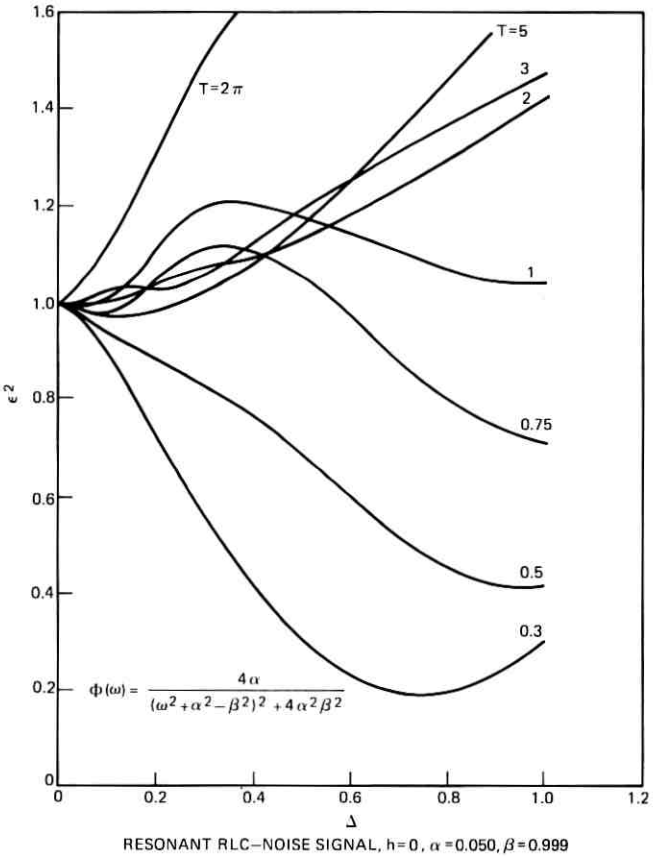

Fig. 9—Curves of ϵ^2 vs Δ for delta modulator with resonant *RLC* Gaussian-noise input–very narrow band case.

Substitution of (47) and (46) in the integral equation (22) yields the recurrence (52) for the expansion coefficients f_{ivl}. This equation involves quantities p_{irs} and q_{irs} defined in (49) and (50). Section 3.4 shows how these quantities can be computed recursively.

The remainder of Section 3.3 is concerned with solving the recurrence (52). The quantities f_{i00} are given explicitly by (64) for each $i = 0, \pm 1, \cdots$. Equation (74), with the definitions (70) and (73), gives f_{0v0}. The remaining f's are given by (75) and (52) when these are applied in proper sequence.

With the expansion coefficients f_{ivl} known, in principle one can write down the joint steady-state distribution of $X(t)$ and $Z(t)$ at any number of times, and from this quantity derive many statistical properties of the delta modulator. Our interest here has centered only on the mean squared error ϵ^2. In Section 3.5, an expression for this quantity in terms of the steady-state distribution $m_i(\mathbf{x})$ is developed. The expansion (47) is then used along with properties of the n-dimensional Hermite functions to obtain a formula, (100), for ϵ^2 involving only the expansion coefficients f_{ivl} and other known quantities. From this formula, the values shown on Figs. 3 through 9 were obtained.

III. MATHEMATICAL TREATMENT

3.1 *The Integral Equation*

Let $X(t)$, the input to the delta modulator, be a continuous stationary stochastic process with mean zero, normalized to have variance unity as shown in (5). We introduce the following notation:

$$a_i \equiv h + i\Delta, \qquad i = 0, \pm 1, \pm 2, \cdots \tag{14}$$

$$X_j \equiv X(jT), \qquad j = 0, 1, 2, \cdots \tag{15}$$

$$p_\tau(y \mid x)\, dy \equiv \Pr\{y \leqq X(t + \tau) < y + dy \mid X(t) = x\} \tag{16}$$

$$m_i^{(j)}(y)\, dy \equiv \Pr\{y \leqq X_j < y + dy, Z_{j-1} = a_i\} \tag{17}$$

$$j = 1, 2, \cdots, \quad i = 0, \pm 1, \pm 2, \cdots .$$

Thus $p_T(y \mid x)$ is the conditional probability density of one sample of the input given the preceding sample, and $m_i^{(j)}(y)\, dy$ is the probability that $Z(t)$ has the value $h + i\Delta$ just before the jth sampling instant *and* that the jth sample of the input, X_j, lies in a small range about the value y.

The event $Z_{j-1} = a_i$ appearing on the right of (17) can occur in two

ways: either Z_{i-2} has the value a_{i-1} and $X_{i-1} > a_{i-1}$; or else $Z_{i-2} = a_{i+1}$ and $X_{i-1} \leqq a_{i+1}$. Thus (17) can be written

$$\begin{aligned} m_i^{(j)}(y)\, dy &= \Pr\{y \leqq X_i < y + dy,\, X_{i-1} > a_{i-1},\, Z_{i-2} = a_{i-1}\} \\ &\quad + \Pr\{y \leqq X_i < y + dy,\, X_{i-1} \leqq a_{i+1},\, Z_{i-2} = a_{i+1}\}. \\ &= dy \int_{a_{i-1}}^{\infty} m_{i-1}^{(j-1)}(x) Q_i(y \mid x,\, i-1)\, dx \\ &\quad + dy \int_{-\infty}^{a_{i+1}} m_{i+1}^{(j-1)}(x) Q_i(y \mid x,\, i+1)\, dx \end{aligned} \tag{18}$$

where

$$Q_i(y \mid x,\, i)\, dy = \Pr\{y \leqq X_i < y + dy \mid X_{i-1} = x,\, Z_{i-2} = a_i\}.$$

Now, *if* $X(t)$ *is Markovian*,

$$Q_i(y \mid x,\, i) = p_T(y \mid x)$$

and (18) becomes

$$\begin{aligned} m_i^{(j)}(y) &= \int_{a_{i-1}}^{\infty} m_{i-1}^{(j-1)}(x) p_T(y \mid x)\, dx \\ &\quad + \int_{-\infty}^{a_{i+1}} m_{i+1}^{(j-1)}(x) p_T(y \mid x)\, dx. \end{aligned} \tag{19}$$

The pair of processes $X(t)$ and $Z(t)$ then form a 2-component vector Markov process. One component, $Z(t)$, takes discrete values; the other, $X(t)$, takes continuous values. Equation (19) is the Chapman–Kolmogorov equation for this vector process.

We have commented in Section II that $m_i^{(2j)}(y)$ and $m_i^{(2j+1)}(y)$ will in general have different limiting forms as $j \to \infty$. By replacing j by $j + 1$ in (19) and adding the result to (19), one finds that

$$\hat{m}_i^{(j)}(y) \equiv \tfrac{1}{2}[m_i^{(j)}(y) + m_i^{(j+1)}(y)]$$

also satisfies (19). Taking the limit as $j \to \infty$, we then have

$$m_i(y) = \int_{a_{i-1}}^{\infty} m_{i-1}(x) p_T(y \mid x)\, dx + \int_{-\infty}^{a_{i+1}} m_{i+1}(x) p_T(y \mid x)\, dx \tag{20}$$

where

$$m_i(y) \equiv \lim_{j \to \infty} \hat{m}_i^{(j)}(y)$$

is the *steady-state* joint distribution for $X(t)$ and $Z(t)$. Equation (20) must be supplemented with the boundary condition

$$\sum_{i=-\infty}^{\infty} m_i(x) = p(x) \tag{21}$$

where $p(x)$ is the probability density for $X(t)$.

The foregoing generalizes readily to the case in which $X(t)$ is not itself Markovian but is one component, say the first, of an n-component stationary-vector Markov process. Denote this process by $\mathbf{X}(t) = \{X_1(t) = X(t), X_2(t), \cdots, X_n(t)\}$. We imagine a delta modulator generating approximations Z_j to $X_1(jT)$ in the manner already described. With an obvious extension of our previous notation, we find

$$m_i(\mathbf{y}) = \int_{-\infty}^{\infty} dx_n \cdots \int_{-\infty}^{\infty} dx_2 \int_{a_{i-1}}^{\infty} dx_1 m_{i-1}(\mathbf{x}) p_T(\mathbf{y} \mid \mathbf{x}) + \int_{-\infty}^{\infty} dx_n \cdots \int_{-\infty}^{\infty} dx_2 \int_{-\infty}^{a_{i+1}} dx_1 m_{i+1}(\mathbf{x}) p_T(\mathbf{y} \mid \mathbf{x}),$$

$$i = 0, \pm 1, \pm 2, \cdots \tag{22}$$

$$\sum_{i=-\infty}^{\infty} m_i(\mathbf{x}) = p(\mathbf{x}). \tag{23}$$

Here, of course, $\mathbf{x}$ and $\mathbf{y}$ are n-vectors, $p_\tau(\mathbf{y} \mid \mathbf{x})$ is the conditional probability density of $\mathbf{X}(t + \tau)$ given $\mathbf{X}(t)$, $p(\mathbf{x})$ is the density of $\mathbf{X}(t)$, and $m_i(\mathbf{y})$ is the steady-state distribution for $\mathbf{X}(t)$ and $Z(t)$, the index i referring to the value $a_i = h + i\Delta$ for $Z(t)$. Equations (22) and (23) are the basic ones on which this paper is built.

In all that follows, we restrict our consideration to inputs $X(t)$ that are Gaussian. It is well-known[16] that if, in this case, $X(t)$ has a rational power density spectrum of form (6), then it can indeed be written as the first component of an n-vector Gaussian process $\mathbf{X}(t)$ that is Markovian. When $m = 0$ in (6), by which we mean that the numerator shown there is a constant independent of ω, the higher order components of $\mathbf{X}(t)$ can be taken as the derivatives of $X(t)$, i.e., $X_{j+1}(t) = d^jX(t)/dt^j$, $j = 1, 2, \cdots, n - 1$. For the more general case $m \geqq 1$, see the article[16] by Helstrom.

To indicate in full the quantities appearing in (22) and (23) in this Gaussian case, we introduce some further notation. Denote the n-variate Gaussian density with zero means by

$$\psi(\mathbf{x}; \varrho) = \frac{1}{(2\pi)^{n/2} \mid \rho \mid^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \sum_{1}^{n} \rho_{ij}^{-1} x_i x_j\right) \tag{24}$$

where ϱ is a positive definite $n \times n$ matrix, the inverse of which has elements ρ_{ij}^{-1}. The right side of (23) is then given by

$$p(\mathbf{x}) = \psi(\mathbf{x}; \boldsymbol{\alpha}) \tag{25}$$

where $\boldsymbol{\alpha}$ is the covariance matrix of $\mathbf{X}(t)$, i.e.,

$$\alpha_{ij} = EX_i(t)X_j(t), \qquad i, j = 1, 2, \cdots, n. \tag{26}$$

The kernel of (22) is given explicitly by

$$p_T(\mathbf{y} \mid \mathbf{x}) = p_T(\mathbf{x}, \mathbf{y})/p(\mathbf{x}) \tag{27}$$

where

$$p_T(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{z}; \boldsymbol{\varrho}). \tag{28}$$

Here the $2n$-vector $\mathbf{z}$ has components

$$z_i = x_i, \qquad z_{n+i} = y_i, \qquad i = 1, 2, \cdots, n \tag{29}$$

and $\boldsymbol{\varrho}$ has the special partitioned structure

$$\boldsymbol{\varrho} = \begin{bmatrix} \boldsymbol{\alpha} & \boldsymbol{\beta} \\ \tilde{\boldsymbol{\beta}} & \boldsymbol{\alpha} \end{bmatrix} \tag{30}$$

where

$$\beta_{ij} = E[X_i(t)X_j(t + T)], \qquad i, j = 1, 2, \cdots, n, \tag{31}$$

$\boldsymbol{\alpha}$ is given by (26), and the tilde denotes transpose.

We shall show in later sections how explicit series solutions can be found to (22) and (23) in this Gaussian rational spectrum case. But first some further preliminaries are necessary.

3.2 *A Generalized Mehler's Formula*

When $n = 1$, (24) becomes the standard normal density

$$\psi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}. \tag{32}$$

Denote its derivatives by

$$\psi_l(x) = \frac{d^l}{dx^l} \psi(x), \qquad l = 0, 1, 2, \cdots. \tag{33}$$

Now $\boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta} = \beta$ and (28) has the series representation

$$\begin{aligned} p_T(x, y) &= \frac{1}{2\pi\sqrt{1 - \beta^2}} \exp\left(-\frac{x^2 + y^2 - 2\beta xy}{2(1 - \beta^2)}\right) \\ &= \sum_{\nu=0}^{\infty} \frac{\beta^\nu}{\nu!} \psi_\nu(x)\psi_\nu(y), \end{aligned} \tag{34}$$

an expansion known as Mehler's formula.[17] It is the power series for the normalized bivariate Gaussian density in terms of the correlation β between the variables. We need the corresponding multiple power series expansion of the $2n$-variate density (28) in terms of the correlations β_{ij} of (31). The series is derived in Ref. 18: to present it, yet further introduction of notation is necessary.

Boldface lower-case Greek letters, $\boldsymbol{\mu}$, $\boldsymbol{\nu}$, etc., will be used henceforth to denote matrices; boldface lower-case Latin letters, $\mathbf{l}$, $\mathbf{m}$, etc., will denote vectors. If $\boldsymbol{\nu}$ is a matrix with n_1 rows and n_2 columns, we write

$$\mathbf{r}(\boldsymbol{\nu}) = (r_1, r_2, \cdots, r_{n_1})$$
$$r_i = \sum_{j=1}^{n_2} \nu_{ij}, \qquad i = 1, \cdots, n_1 \tag{35}$$

for the vector whose components are the row sums of $\boldsymbol{\nu}$, and we write

$$\mathbf{c}(\boldsymbol{\nu}) = (c_1, c_2, \cdots, c_{n_2})$$
$$c_j = \sum_{i=1}^{n_1} \nu_{ij}, \qquad j = 1, \cdots, n_2 \tag{36}$$

for the vector, the components of which are the column sums of $\boldsymbol{\nu}$. Throughout we adopt the convenient abbreviations

$$\boldsymbol{\mu}^{\boldsymbol{\nu}} \equiv \prod_{i,j} \mu_{ij}^{\nu_{ij}}, \qquad \mathbf{r}^{\mathbf{l}} \equiv \prod_i r_i^{l_i}$$
$$\boldsymbol{\mu}! \equiv \prod_{i,j} \mu_{ij}!, \qquad \mathbf{l}! \equiv \prod_i l_i!$$
$$\sum_{\boldsymbol{\nu}=0}^{\infty} \equiv \sum_{\nu_{11}=0}^{\infty} \sum_{\nu_{12}=0}^{\infty} \cdots \sum_{\nu_{n_1 n_2}=0}^{\infty}, \qquad \sum_{\mathbf{l}=0}^{\infty} \equiv \sum_{l_1=0}^{\infty} \cdots \sum_{l_n=0}^{\infty} \tag{37}$$

where the entries of $\boldsymbol{\mu}$ are μ_{ij}, the components of $\mathbf{l}$ are l_i, etc. We call a matrix of nonnegative integers, such as $\boldsymbol{\nu}$ in the last line of (37), an *index matrix*; a vector of nonnegative integers, such as $\mathbf{l}$, is an *index vector*. The statement $\mathbf{s} \leqq \mathbf{t}$ means that no component of $\mathbf{s}$ is greater than the corresponding component of $\mathbf{t}$; the statement $\mathbf{s} < \mathbf{t}$ means $\mathbf{s} \leqq \mathbf{t}$ and $\mathbf{s} \neq \mathbf{t}$. Inequalities between matrices, e.g., $\boldsymbol{\mu} \leqq \boldsymbol{\nu}$ are to be interpreted in a similar manner. Finally, we write

$$[\mathbf{l}] = l_1 + l_2 + \cdots + l_n \tag{38}$$

for the sum of the components of a vector, and we define

$$\psi_{\mathbf{l}}(\mathbf{x}; \boldsymbol{\varrho}) = \frac{\partial^{[\mathbf{l}]}}{\partial x_1^{l_1} \partial x_2^{l_2} \cdots \partial x_n^{l_n}} \psi(\mathbf{x}; \boldsymbol{\varrho}) \tag{39}$$

where the Gaussian density $\psi(\mathbf{x}; \boldsymbol{\varrho})$ is given by (24).

The desired generalization of Mehler's formula is

$$p_T(\mathbf{x}, \mathbf{y}) = \sum_{\nu=0}^{\infty} \frac{\beta^\nu}{\nu!} \psi_{\mathbf{r}(\nu)}(\mathbf{x}; \alpha)\psi_{\mathbf{c}(\nu)}(\mathbf{y}; \alpha). \tag{40}$$

It is a multiple power series for the density of two identically distributed Gaussian vectors in terms of the cross correlations between the components of the vectors.

The functions $\psi_l(\mathbf{x}; \alpha)$ defined in (39) that occur in (40) are closely related to the Hermite polynomials of several variables studied by Erdélyi[19] and others. We call $\psi_l(\mathbf{x}; \alpha)$ an *n-dimensional Hermite function* of weight $[l]$ [see (38)]. The following facts about them that will be of use to us later are established in Ref. 18.

(*i*) If $l_1, l_2, \cdots, l_r$ are r distinct n-vectors with nonnegative integers as components, then $\psi_{l_1}(\mathbf{x}, \alpha)$, $\psi_{l_2}(\mathbf{x}, \alpha)$, $\cdots$ $\psi_{l_r}(\mathbf{x}, \alpha)$ are linearly independent functions of $\mathbf{x}$. Hermite functions have the generating function

$$\psi(\mathbf{x} + \mathbf{t}; \varrho) = \sum_{l=0}^{\infty} \frac{\mathbf{t}^l}{l!} \psi_l(\mathbf{x}; \varrho), \tag{41}$$

which is just Taylor's theorem in many variables.

(*ii*) There are

$$N(n, p) = \binom{n + p - 1}{p} \tag{42}$$

n-dimensional Hermite functions of weight p. Functions of different weight are orthogonal with respect to the weight function

$$w(\mathbf{x}; \varrho) = \frac{1}{\psi(\mathbf{x}; \varrho)}. \tag{43}$$

(*iii*) The scalar product of any two functions of the same weight p can be expressed in terms of an $N(n, p) \times N(n, p)$ matrix, $\mathfrak{d}_p(\varrho^{-1})$, known as the symmetrized Kronecker pth power of ϱ^{-1}. We have the formula

$$\frac{1}{\sqrt{l!}\ \sqrt{\mathrm{m}!}} \int_{-\infty}^{\infty} dx_1 \cdots \int_{-\infty}^{\infty} dx_n \psi_l(\mathbf{x}; \varrho)\psi_{\mathrm{m}}(\mathbf{x}; \varrho)w(\mathbf{x}; \varrho) = \delta_{[l][\mathrm{m}]}\mathfrak{d}_{[l]}(\varrho^{-1})_{l\mathrm{m}} \tag{44}$$

where δ_{ij} is the usual Kronecker symbol. An explicit formula for the matrix $\mathfrak{d}_p$ is

$$\mathfrak{d}_p(\alpha)_{l\mathrm{m}} = \sqrt{l!}\ \sqrt{\mathrm{m}!} \sum_{\substack{\mathfrak{y} \\ \mathbf{r}(\mathfrak{y}) = l \\ \mathbf{c}(\mathfrak{y}) = \mathrm{m} \\ [l] = [\mathrm{m}] = p}} \frac{\alpha^{\mathfrak{y}}}{\mathfrak{y}!}. \tag{45}$$

As indicated, the sum here is over all index matrices $\mathfrak{y}$ with row-sum vector $\boldsymbol{l}$ and column-sum vector $\mathbf{m}$, these latter being of weight p.

3.3 *Series Solution of the Integral Equation*

We now return to consideration of the system of eqs. (22) and (23) where the density $p(\mathbf{x})$ and $p_T(\mathbf{y} \mid \mathbf{x})$ are defined by (24) through (31). To simplify notation we shall frequently write $\psi_\mathbf{l}(\mathbf{x})$ for $\psi_\mathbf{l}(\mathbf{x}; \boldsymbol{\alpha})$, the unexpressed matrix always being $\boldsymbol{\alpha}$. Unless otherwise explicitly stated, boldface Greek letters will denote $n \times n$ matrices while boldface Latin letters will denote n-vectors.

Equations (40), (27) and (25) show that the kernel of (22) can be written

$$p_T(\mathbf{y} \mid \mathbf{x}) = \sum_{\mathfrak{y}=0}^{\infty} \frac{\boldsymbol{\beta}^{\mathfrak{y}}}{\mathfrak{y}!} \frac{\psi_{\mathbf{r}(\mathfrak{y})}(\mathbf{x})\psi_{\mathbf{c}(\mathfrak{y})}(\mathbf{y})}{\psi(\mathbf{x})}, \tag{46}$$

the conventions (37) being understood here. This suggests a series solution to (22) and (23) in the form

$$m_i(\mathbf{x}) = \sum_{\mathbf{v},\, \boldsymbol{l}=\mathbf{0}}^{\infty} \boldsymbol{\beta}^{\mathbf{v}} f_{i\mathbf{v}\boldsymbol{l}} \psi_{\boldsymbol{l}}(\mathbf{x}). \tag{47}$$

Conditions on the coefficients $f_{i\mathbf{v}\boldsymbol{l}}$ are then obtained by substituting (47) and (46) into (22). There results

$$\sum_{\mathbf{v}\boldsymbol{l}} \boldsymbol{\beta}^{\mathbf{v}} f_{i\mathbf{v}\boldsymbol{l}} \psi_{\boldsymbol{l}}(\mathbf{y}) = \sum_{\mathbf{d}\mathfrak{y}\mathbf{s}} \frac{\boldsymbol{\beta}^{\mathbf{d}+\mathfrak{y}}}{\mathfrak{y}!} \psi_{\mathbf{c}(\mathfrak{y})}(\mathbf{y})[f_{i-1\ \mathbf{d}\ \mathbf{s}} q_{i-1\ \mathbf{s}\ \mathbf{r}(\mathfrak{y})} + f_{i+1\ \mathbf{d}\ \mathbf{s}} p_{i+1\ \mathbf{s}\ \mathbf{r}(\mathfrak{y})}] \tag{48}$$

where

$$p_{i\mathbf{r}\mathbf{s}} = \int_{-\infty}^{\infty} dx_n \cdots \int_{-\infty}^{\infty} dx_2 \int_{-\infty}^{a_i} dx_1 \frac{\psi_\mathbf{r}(\mathbf{x})\psi_\mathbf{s}(\mathbf{x})}{\psi(\mathbf{x})} \tag{49}$$

and

$$q_{i\mathbf{r}\mathbf{s}} = \int_{-\infty}^{\infty} dx_n \cdots \int_{-\infty}^{\infty} dx_2 \int_{a_i}^{\infty} dx_1 \frac{\psi_\mathbf{r}(\mathbf{x})\psi_\mathbf{s}(\mathbf{x})}{\psi(\mathbf{x})}. \tag{50}$$

On setting $\mathfrak{y} = \mathbf{v} - \mathbf{d}$, the right of (48) becomes

$$m_i(\mathbf{y}) = \sum_{\mathbf{v}} \boldsymbol{\beta}^{\mathbf{v}} \sum_{\mathbf{d}=\mathbf{0}}^{\mathbf{v}} \psi_{\mathbf{c}(\mathbf{v}-\mathbf{d})}(\mathbf{y}) \frac{1}{(\mathbf{v}-\mathbf{d})!} \cdot \sum_{\mathbf{s}} [f_{i-1\ \mathbf{d}\ \mathbf{s}} q_{i-1\ \mathbf{s}\ \mathbf{r}(\mathbf{v}-\mathbf{d})} + f_{i+1\ \mathbf{d}\ \mathbf{s}} p_{i+1\ \mathbf{s}\ \mathbf{r}(\mathbf{v}-\mathbf{d})}]. \tag{51}$$

But this form shows that in (47) the $\boldsymbol{l}$ sum could be restricted to run from $\mathbf{0}$ to $\mathbf{c}(\mathbf{v})$. This in turn restricts the $\mathbf{s}$ sum in (51) to run from $\mathbf{0}$

to $\mathbf{c}(\boldsymbol{\sigma})$. Equating coefficients of powers of $\boldsymbol{\beta}$ on the left of (38) and the right of (51) gives

$$\sum_{l=0}^{\mathbf{c}(\boldsymbol{\nu})} f_{i\boldsymbol{\nu}l}\psi_l(\mathbf{y}) = \sum_{\boldsymbol{\mu}=0}^{\boldsymbol{\nu}} \psi_{\mathbf{c}(\boldsymbol{\mu})}(\mathbf{y}) \frac{1}{\boldsymbol{\mu}!} \cdot \sum_{\mathbf{s}=0}^{\mathbf{c}(\boldsymbol{\nu}-\boldsymbol{\mu})} [f_{i-1\,\boldsymbol{\nu}-\boldsymbol{\mu}\,\mathbf{s}} q_{i-1\,\mathbf{s}\,\mathbf{r}(\boldsymbol{\mu})} + f_{i+1\,\boldsymbol{\nu}-\boldsymbol{\mu}\,\mathbf{s}} p_{i+1\,\mathbf{s}\,\mathbf{r}(\boldsymbol{\mu})}]$$

where we have written $\boldsymbol{\sigma} = \boldsymbol{\nu} - \boldsymbol{\mu}$ to reintroduce $\boldsymbol{\mu}$. Using the linear independence of the $\psi_l(\mathbf{y})$, we find finally

$$f_{i\boldsymbol{\nu}l} = \sum_{\boldsymbol{\mu}}{}' \frac{1}{\boldsymbol{\mu}!} \sum_{\mathbf{s}=0}^{\mathbf{c}(\boldsymbol{\nu}-\boldsymbol{\mu})} [f_{i-1\,\boldsymbol{\nu}-\boldsymbol{\mu}\,\mathbf{s}} q_{i-1\,\mathbf{s}\,\mathbf{r}(\boldsymbol{\mu})} + f_{i+1\,\boldsymbol{\nu}-\boldsymbol{\mu}\,\mathbf{s}} p_{i+1\,\mathbf{s}\,\mathbf{r}(\boldsymbol{\mu})}], \quad \mathbf{0} \leqq \boldsymbol{l} \leqq \mathbf{c}(\boldsymbol{\nu}), \quad i = 0, \pm 1, \pm 2, \cdots, \tag{52}$$

which holds for all $\boldsymbol{\nu} \geqq \mathbf{0}$. Here the sum is over all index arrays $\boldsymbol{\mu}$ with $\mathbf{0} \leqq \boldsymbol{\mu} \leqq \boldsymbol{\nu}$ and $\mathbf{c}(\boldsymbol{\mu}) = \boldsymbol{l}$.

In the remaining paragraphs of this section we develop (52) to show how a recurrence scheme can be arrived at that permits successive determination of the $f_{i\boldsymbol{\nu}l}$.

We note that from (52)

$$\begin{aligned} f_{i\boldsymbol{\nu}0} &= \sum_{\mathbf{s}=0}^{\mathbf{c}(\boldsymbol{\nu})} [f_{i-1\,\boldsymbol{\nu}\,\mathbf{s}} q_{i-1\,\mathbf{s}\,0} + f_{i+1\,\boldsymbol{\nu}\,\mathbf{s}} p_{i+1\,\mathbf{s}\,0}] \\ &= f_{i-1\,\boldsymbol{\nu}\,0} q_{i-1\,0\,0} + f_{i+1\,\boldsymbol{\nu}\,0} p_{i+1\,0\,0} \\ &\quad + \sum_{\mathbf{s}\neq 0}^{\mathbf{c}(\boldsymbol{\nu})} [f_{i-1\,\boldsymbol{\nu}\,\mathbf{s}} q_{i-1\,\mathbf{s}\,0} + f_{i+1\,\boldsymbol{\nu}\,\mathbf{s}} p_{i+1\,\mathbf{s}\,0}], \end{aligned} \tag{53}$$

where we assume $\boldsymbol{\nu} \neq \mathbf{0}$.

Again from (52)

$$f_{i-1\,\boldsymbol{\nu}\,\mathbf{s}} = \sum_{\boldsymbol{\mu}}{}' \frac{1}{\boldsymbol{\mu}!} \sum_{\mathbf{t}=0}^{\mathbf{c}(\boldsymbol{\nu}-\boldsymbol{\mu})} [f_{i-2\,\boldsymbol{\nu}-\boldsymbol{\mu}\,\mathbf{t}} q_{i-2\,\mathbf{t}\,\mathbf{r}(\boldsymbol{\mu})} + f_{i\,\boldsymbol{\nu}-\boldsymbol{\mu}\,\mathbf{t}} p_{i\,\mathbf{t}\,\mathbf{r}(\boldsymbol{\mu})}]$$

where the sum on $\boldsymbol{\mu}$ is over all arrays with $\boldsymbol{\mu} \leqq \boldsymbol{\nu}$ and $\mathbf{c}(\boldsymbol{\mu}) = \mathbf{s}$. A similar expression can be written for $f_{i+1\,\boldsymbol{\nu}\,\mathbf{s}}$. Replace the f's appearing in brackets on the right of (53) by these expressions. There results

$$\begin{aligned} f_{i\boldsymbol{\nu}0} &= f_{i-1\,\boldsymbol{\nu}\,0} q_{i-1\,0\,0} + f_{i+1\,\boldsymbol{\nu}\,0} p_{i+1\,0\,0} \\ &\quad + \sum_{\mathbf{s}\neq 0}^{\mathbf{c}(\boldsymbol{\nu})} q_{i-1\,\mathbf{s}\,0} \sum_{\boldsymbol{\mu}}{}' \frac{1}{\boldsymbol{\mu}!} \sum_{\mathbf{t}=0}^{\mathbf{c}(\boldsymbol{\nu}-\boldsymbol{\mu})} h_{i\mathbf{t}\boldsymbol{\mu}\boldsymbol{\nu}} \\ &\quad + \sum_{\mathbf{s}\neq 0}^{\mathbf{c}(\boldsymbol{\nu})} p_{i+1\,\mathbf{s}\,0} \sum_{\boldsymbol{\mu}}{}' \frac{1}{\boldsymbol{\mu}!} \sum_{\mathbf{t}=0}^{\mathbf{c}(\boldsymbol{\nu}-\boldsymbol{\mu})} h_{i+2\,\mathbf{t}\boldsymbol{\mu}\boldsymbol{\nu}} \end{aligned} \tag{54}$$

where we have written

$$h_{i\mathfrak{t}\mathfrak{y}\nu} = f_{i-2\ \nu-u\ \mathfrak{t}} q_{i-2\ \mathfrak{t}\ \mathbf{r}(\mathfrak{y})} + f_{i\ \nu-\mathfrak{y}\ \mathfrak{t}} p_{i\ \mathfrak{t}\ \mathbf{r}(\mathfrak{y})} .$$

Now in the first sum on the right of (54),

$$T \equiv \sum_{\mathbf{s}\neq 0}^{\mathbf{c}(\nu)} q_{i-1\ \mathbf{s}\ 0} \sum_{\mathfrak{y}}{}' \frac{1}{\mathfrak{y}!} \sum_{\mathfrak{t}=0}^{\mathbf{c}(\nu-\mathfrak{y})} h_{i\mathfrak{t}\mathfrak{y}\nu} ,$$

substitute $\mathfrak{d} = \nu - \mathfrak{y}$ to obtain

$$T = \sum_{\mathbf{s}\neq 0}^{\mathbf{c}(\nu)} q_{i-1\ \mathbf{s}\ 0} \sum_{\mathfrak{d}=0}^{\nu'} \frac{1}{(\nu-\mathfrak{d})!} \sum_{\mathfrak{t}=0}^{\mathbf{c}(\mathfrak{d})} h_{i\ \mathfrak{t}\ \nu-\mathfrak{d}\ \nu}$$

where the middle sum is over all arrays $\mathfrak{d}$ with $\mathbf{0} \leqq \mathfrak{d} \leqq \nu$ and $\mathbf{c}(\mathfrak{d}) = \mathbf{c}(\nu) - \mathbf{s}$. Since s varies in the range $\mathbf{0} < \mathbf{s} \leqq \mathbf{c}(\nu)$, however, $\mathfrak{d}$ indeed ultimately takes on all values $< \nu$. Thus

$$T = \sum_{0\leqq\mathfrak{d}<\nu} q_{i-1\ \mathbf{c}(\nu-\mathfrak{d})\ 0} \frac{1}{(\nu-\mathfrak{d})!} \sum_{\mathfrak{t}=0}^{\mathbf{c}(\mathfrak{d})} h_{i\ \mathfrak{t}\ \nu-\mathfrak{d}\ \nu} .$$

Using a similar rearrangement for the last sum in (54) one finds finally

$$f_{i\nu 0} = f_{i-1\ \nu\ 0} q_{i-1\ 0\ 0} + f_{i+1\ \nu\ 0} p_{i+1\ 0\ 0} + \sum_{0\leqq\mathfrak{d}<\nu} \frac{1}{(\nu-\mathfrak{d})!}$$
$$\cdot \sum_{\mathfrak{t}=0}^{\mathbf{c}(\mathfrak{d})} [f_{i-2\ \mathfrak{d}\ \mathfrak{t}} A_{i-2\ \nu-\mathfrak{d}\ \mathfrak{t}} + f_{i\ \mathfrak{d}\ \mathfrak{t}} B_{i\ \nu-\mathfrak{d}\ \mathfrak{t}} + f_{i+2\ \mathfrak{d}\ \mathfrak{t}} C_{i+2\ \nu-\mathfrak{d}\ \mathfrak{t}}] \tag{55}$$

where

$$\begin{aligned} A_{i\mathfrak{y}\mathfrak{t}} &= q_{i+1\ \mathbf{c}(\mathfrak{y})\ 0} q_{i\ \mathbf{r}(\mathfrak{y})\ \mathfrak{t}} \\ B_{i\mathfrak{y}\mathfrak{t}} &= q_{i-1\ \mathbf{c}(\mathfrak{y})\ 0} p_{i\ \mathbf{r}(\mathfrak{y})\ \mathfrak{t}} + p_{i+1\ \mathbf{c}(\mathfrak{y})\ 0} q_{i\ \mathbf{r}(\mathfrak{y})\ \mathfrak{t}} \\ C_{i\mathfrak{y}\mathfrak{t}} &= p_{i-1\ \mathbf{c}(\mathfrak{y})\ 0} p_{i\ \mathbf{r}(\mathfrak{y})\ \mathfrak{t}} . \end{aligned} \tag{56}$$

If $\nu = \mathbf{0}$, the sums in (55) are to be interpreted as zero.

The quantities A, B and C just introduced are not independent. On using (44) and the definitions (49) and (50), we find

$$p_{i\mathbf{r}\mathbf{s}} + q_{i\mathbf{r}\mathbf{s}} = \delta_{[\mathbf{r}],[\mathbf{s}]}\, \mathbf{r}!\, \mathbf{s}!\, \sigma_{[\mathbf{r}]}(\boldsymbol{\alpha}^{-1})_{\mathbf{r}\mathbf{s}} . \tag{57}$$

Now sum (56) to find

$$\begin{aligned} A_{i\mathfrak{y}\mathfrak{t}} + B_{i\mathfrak{y}\mathfrak{t}} + C_{i\mathfrak{y}\mathfrak{t}} &= q_{i\ \mathbf{r}(\mathfrak{y})\ \mathfrak{t}} [p_{i+1\ \mathbf{c}(\mathfrak{y})\ 0} + q_{i+1\ \mathbf{c}(\mathfrak{y})\ 0}] \\ &\quad + p_{i\ \mathbf{r}(\mathfrak{y})\ \mathfrak{t}} [p_{i-1\ \mathbf{c}(\mathfrak{y})\ 0} + q_{i-1\ \mathbf{c}(\mathfrak{y})\ 0}] \\ &= [q_{i\ \mathbf{r}(\mathfrak{y})\ \mathfrak{t}} + p_{i\ \mathbf{r}(\mathfrak{y})\ \mathfrak{t}}]\ \delta_{\mathfrak{y}0} \end{aligned}$$

or finally

$$A_{i\mu t} + B_{i\mu t} + C_{i\mu t} = \delta_{\mu 0}\,\delta_{t0}\,. \tag{58}$$

We note now that the normalization (23) together with (47) yields

$$\sum_i f_{i\nu l} = \delta_{\nu 0}\,\delta_{l0}\,, \tag{59}$$

a normalization requirement of the f's.

Now consider (55) when $\nu = \mathbf{0}$,

$$f_{i00} = f_{i-1\,0\,0}q_{i-1\,0\,0} + f_{i+1\,0\,0}p_{i+1\,0\,0}\,. \tag{60}$$

Since by (57) $p_{i00} + q_{i00} = 1$, (60) can be rewritten as

$$f_{i00}q_{i00} - f_{i+1\,0\,0}p_{i+1\,0\,0} = f_{i-1\,0\,0}q_{i-1\,0\,0} - f_{i00}p_{i00} \tag{61}$$

which is to hold for all i. Now p_{i00} and q_{i00} are bounded for all i, and by (59) the f's are summable. Both sides of (61) are therefore summable, and summing for $i = l, l+1, l+2, \cdots$, we find

$$f_{l-1\,0\,0}q_{l-1\,0\,0} - f_{l00}p_{l00} = 0 \tag{62}$$

which holds for all l. In addition, from (59),

$$\sum_i f_{i00} = 1. \tag{63}$$

These equations are readily solved by setting

$$\begin{aligned} w_0 &= 1 \\ w_j &= \frac{q_{j-1\,0\,0}}{p_{j00}}\,w_{j-1}\,, \qquad j = 1, 2, \cdots \\ w_{j-1} &= \frac{p_{j00}}{q_{j-1\,0\,0}}\,w_j\,, \qquad j = 0, -1, -2, \cdots \\ f_{i00} &= \frac{w_i}{\sum_j w_j}. \end{aligned} \tag{64}$$

With the f_{i00} now determined, we turn our attention to (55) for $\nu \neq \mathbf{0}$. Replace $B_{i\,\nu-d\,t}$ there by $-A_{i\,\nu-d\,t} - C_{i\,\nu-d\,t}$ as is allowed by (58). Multiply $f_{i\nu 0}$ by $1 = p_{i00} + q_{i00}$ and regroup terms to obtain

$$u_i + v_i = u_{i-1} + v_{i-2} \tag{65}$$

where

$$u_i = f_{i\nu 0} q_{i00} - f_{i+1\,\nu\,0} p_{i+1\,0\,0}$$

$$v_i = \sum_{0 \le \delta < \nu} \frac{1}{(\nu - \delta)!} \sum_{t=0}^{c(\delta)} [f_{i\delta t} A_{i\,\nu-\delta\,t} - f_{i+2\,\delta\,t} C_{i+2\,\nu-\delta\,t}]. \tag{66}$$

Equation (65) is to hold for all $i = 0, \pm 1, \pm 2, \cdots$. The quantities $A_{i\mu t}$, $C_{i\mu t}$, p_{irs}, q_{irs} are all bounded in i. The f's are summable by (59) and hence so are the u_i and v_i. Summing (65) for $i = l, l + 1, \cdots$, there results

$$u_l = -(v_l + v_{l-1}).$$

Using (66) this becomes

$$f_{i+1\,\nu\,0} p_{i+1\,0\,0} - f_{i\nu 0} q_{i00} = d_{i\nu} \tag{67}$$

where

$$d_{i\nu} = \sum_{0 \le \delta < \nu} \frac{1}{(\nu - \delta)!} \sum_{t=0}^{c(\delta)} [f_{i-1\,\delta\,t} A_{i-1\,\nu-\delta\,t} + f_{i\delta t} A_{i\,\nu-\delta\,t} - f_{i+1\,\delta\,t} C_{i+1\,\nu-\delta\,t} - f_{i+2\,\delta\,t} C_{i+2\,\nu-\delta\,t}]. \tag{68}$$

Suppose now that the $d_{i\nu}$ are known for $i = 0, \pm 1, \pm 2, \cdots$. If we can solve (67) subject to

$$\sum_i f_{i\nu 0} = 0 \tag{69}$$

as required by (59), our recurrence is complete, for the $d_{i\nu}$ depend only on the $f_{i\mu l}$ with $\mu < \nu$. Equation (52) for $l > 0$ permits computation of the $f_{i\nu l}$, $i = 0, \pm 1, \pm 2, \cdots$ in terms of the $f_{i\mu l}$, $\mu < \nu$. The values (64) start the recurrence off.

Now the solution to (67) subject to (69) is quite straightforward. Introduce the notation $\xi_i = f_{i\nu 0}$, $i = 0, 1, 2, \cdots$, $\eta_i = f_{-i\nu 0}$, $i = 0, 1, 2, \cdots$,

$$V_i^+ = \frac{q_{i-1\,0\,0}}{p_{i00}} \qquad V_i^- = \frac{p_{-(i-1)\,0\,0}}{q_{-i00}}$$

$$D_i^+ = \frac{d_{i-1\,\nu}}{p_{i00}} \qquad D_i^- = -\frac{d_{-i\nu}}{q_{-i00}}. \tag{70}$$

Equation (67) can be written

$$\xi_{i+1} = V_{i+1}^+ \xi_i + D_{i+1}^+, \qquad i = 0, 1, 2, \cdots$$

$$\eta_{i+1} = V_{i+1}^- \eta_i + D_{i+1}^-, \qquad i = 0, 1, 2, \cdots$$

whence

$$\xi_0 = \eta_0$$

$$\xi_i = \xi_0 \prod_{j=1}^{i} V_j^+ + \sum_{j=1}^{i-1} D_j^+ \prod_{j+1}^{i} V_k^+ + D_i^+$$

$$\eta_i = \eta_0 \prod_{j=1}^{i} V_j^- + \sum_{j=1}^{i-1} D_j^- \prod_{j+1}^{i} V_k^- + D_i^- \qquad i = 1, 2, \cdots . \tag{71}$$

Adding these equations we find

$$\xi_0 + \sum_1 \xi_i + \sum_1 \eta_i = \xi_0(1 + V^- + V^+) + \sum_1 L_i^+ D_i^+ + \sum_1 L_i^- D_i^- \tag{72}$$

where

$$\begin{aligned} V^+ &= V_1^+ + V_1^+ V_2^+ + V_1^+ V_2^+ V_3^+ + \cdots \\ &= \sum_{i=1}^{\infty} \prod_{j=1}^{i} V_i^+ \\ V^- &= \sum_{i=1}^{\infty} \prod_{j=1}^{i} V_i^- \\ L_1^+ &= V^+/V_1^+ , \qquad L_1^- = V^-/V_1^- \\ L_i^+ &= (L_{i-1}^+ - 1)/V_i^+ , \\ L_i^- &= (L_{i-1}^- - 1)/V_i^- \qquad j = 2, 3, \cdots . \end{aligned} \tag{73}$$

But the left of (72) is the sum shown in (69) and hence vanishes. We have then

$$\xi_0 = f_{0v0} = -\frac{\sum_1^{\infty} (L_i^+ D_i^+ + L_i^- D_i^-)}{V^- + 1 + V^+} \tag{74}$$

with the quantities on the right given explicitly by (70) and (73). With f_{0v0} known, one can return to (67) in the form

$$f_{i+1\,v\,0} = \frac{1}{p_{i+1\,00}} [f_{iv0} q_{i00} + d_{iv}], \qquad i = 0, 1, 2, \cdots$$

$$f_{iv0} = \frac{1}{q_{i00}} [f_{i+1\,v\,0} p_{i+1\,0\,0} - d_{iv}], \qquad i = -1, -2, \cdots \tag{75}$$

to compute the remaining f's recursively, or one can utilize the explicit solutions (71).

3.4 *Recursion for the p_{irs}*

The formulas just developed for computing the coefficients f_{ivl} involve the quantities p_{irs} and q_{irs}. The latter are given in terms of the former by (57). We turn our attention now to a recursive method of computing the p's.

From the generating function (41) we find

$$\frac{\psi(\mathbf{x}+\boldsymbol{\xi})\psi(\mathbf{x}+\mathbf{n})}{\psi(\mathbf{x})} = \sum \frac{\boldsymbol{\xi}^{\mathbf{r}}\mathbf{n}^{\mathbf{s}}}{\mathbf{r}!\,\mathbf{s}!}\frac{\psi_{\mathbf{r}}(\mathbf{x})\psi_{\mathbf{s}}(\mathbf{x})}{\psi(\mathbf{x})}$$

$$= \exp\left(\sum \alpha_{ij}^{-1}\xi_i\eta_j\right) \cdot \frac{\exp\left[-\tfrac{1}{2}\sum \alpha_{ij}^{-1}(x_i+\xi_i+\eta_i)(x_j+\xi_j+\eta_j)\right]}{(2\pi)^{n/2}\,|\,\boldsymbol{\alpha}\,|^{\frac{1}{2}}} \tag{76}$$

where $\boldsymbol{\xi}$ and $\mathbf{n}$ are n-vectors. Recall now the definition (49) of p_{irs}. Integration of (76) then gives

$$\sum \frac{\boldsymbol{\xi}^{\mathbf{r}}\mathbf{n}^{\mathbf{s}}}{\mathbf{r}!\,\mathbf{s}!}p_{irs} = \exp\left(\sum \alpha_{ij}^{-1}\xi_i\eta_j\right)\int_{-\infty}^{a_i+\xi_1+\eta_1}\frac{e^{-\frac{1}{2}y^2/\alpha_{11}}}{\sqrt{2\pi\alpha_{11}}}\,dy$$

$$= \exp\left(\sum \alpha_{ij}^{-1}\xi_i\eta_j\right)F(a_i+\xi_1+\eta_1). \tag{77}$$

Take the partial derivative of this relation with respect to ξ_j, $j > 1$, to obtain

$$\sum_{rs}\frac{\xi_1^{r_1}\cdots r_j\xi_j^{r_j-1}\cdots\xi_n^{r_n}\eta^s}{r!\,s!}p_{irs} = \sum_{k=1}^{n}\alpha_{jk}^{-1}\eta_k\sum\frac{\boldsymbol{\xi}^{\mathbf{r}}\mathbf{n}^{\mathbf{s}}}{\mathbf{r}!\,\mathbf{s}!}p_{irs}$$

or

$$p_{irs} = \sum_{k=1}^{n}\alpha_{jk}^{-1}s_k p_{ir_1\cdots(r_j-1)\cdots r_n s_1\cdots(s_k-1)\cdots s_n}, \qquad j > 1. \tag{78}$$

A similar formula, obtained by differentiation with respect to η_k,

$$p_{irs} = \sum_{j=1}^{n}\alpha_{jk}^{-1}r_j p_{ir_1\cdots(r_k-1)\cdots r_n s_1\cdots(s_j-1)\cdots s_n}, \qquad k > 1 \tag{79}$$

also holds.

Repeated use of (78) and (79) permits one to express p_{irs} as a linear combination of the quantities $p_{i\hat{r}_1 00\,\cdots\,0\hat{s}_1 00\,\cdots\,0}$ where $0 \leqq \hat{r}_1 \leqq r_1$ and $0 \leqq \hat{s}_1 \leqq s_1$. Let us now define

$$\hat{p}_{r_1 s_1} \equiv p_{ir_1 00\cdots 0 s_1 00\cdots 0}$$

and seek rules to determine these quantities. We have

$$\sum \frac{\xi_1^{r_1}\eta_1^{s_1}}{r_1!\, s_1!}\hat{p}_{r_1 s_1} = \exp(+\alpha_{11}^{-1}\xi_1\eta_1)F(a_i + \xi_1 + \eta_1) \tag{80}$$

with F defined by (77). Without loss of generality, we take $\alpha_{11} = 1$ and note that

$$\hat{p}_{00} = \int_{-\infty}^{a_i} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\, dt. \tag{81}$$

Now differentiate (80) with respect to ξ_1 to obtain

$$\sum \frac{\xi^{r-1}\eta^s}{(r-1)!\, s!}\hat{p}_{rs} = \alpha_{11}^{-1}\eta \sum \frac{\xi^r\eta^s}{r!\, s!}\hat{p}_{rs} + \exp(\alpha_{11}^{-1}\xi\eta)\frac{e^{-\frac{1}{2}(a_i+\xi+\eta)^2}}{\sqrt{2\pi}} \tag{82}$$

where we have dropped some unnecessary subscripts. Let

$$\exp(\alpha_{11}^{-1}\xi\eta)\frac{e^{-\frac{1}{2}(a_i+\xi+\eta)^2}}{\sqrt{2\pi}} = \sum \frac{\xi^j\eta^k}{j!\, k!}M_{jk}\,. \tag{83}$$

Equation (82) then gives

$$\hat{p}_{r+1\ s} = \alpha_{11}^{-1}s\hat{p}_{r\ s-1} + M_{rs}$$

and its symmetric version obtained by interchanging the roles of r and s. These equations yield

$$\begin{aligned}\hat{p}_{rs} &= \frac{M_{r\ s+1} - M_{s\ r+1}}{\alpha_{11}^{-1}(r-s)}, \qquad r \neq s\\ \hat{p}_{rr} &= \alpha_{11}^{-1}r\hat{p}_{r-1\ r-1} + M_{r-1\ r}\,.\end{aligned} \tag{84}$$

To complete the recurrence we must have rules for generating the M's. Differentiating (83) with respect to ξ gives

$$M_{j+1\ k} = -aM_{jk} - (1 - \alpha_{11}^{-1})kM_{j\ k-1} - jM_{j-1\ k} \tag{85}$$

which permits reduction on j, so that M_{jk} can be expressed in terms of $M_{0k'}$ with $0 \leqq k' \leqq k$. But from its definition, $M_{jk} = M_{kj}$ and from (85) we deduce

$$M_{0k} = -aM_{0\ k-1} - (k-1)M_{0\ k-2}\,. \tag{86}$$

Finally, we find

$$\begin{aligned}M_{00} &= \frac{e^{-a_i^2/2}}{\sqrt{2\pi}}\\ M_{01} &= -a_i\frac{e^{-a_i^2/2}}{\sqrt{2\pi}}.\end{aligned} \tag{87}$$

3.5 *The Mean Squared Error*

The mean squared error of the delta modulator running in the steady state is defined by

$$\epsilon^2 = \frac{1}{T}\int_0^T dtE[X(t) - Z(t)]^2 \tag{88}$$

where the expectation is to be taken using the steady-state distribution of $X(t)$ and $Z(t)$. Then

$$\epsilon^2 = \frac{1}{T}\int_0^T dt \sum_i \int_{-\infty}^{\infty} dy[y - a_i]^2 P_i(y, t) \tag{89}$$

where

$$P_i(y, t)dy = \Pr\{y \leqq X(t) < y + dy, Z(t) = a_i\}$$

and so

$$\begin{aligned} P_i(y, t) = &\int_{-\infty}^{\infty} dx_n \cdots \int_{-\infty}^{\infty} dx_2 \int_{a_{i-1}}^{\infty} dx_1 m_{i-1}(\mathbf{x})\hat{p}_t(y \mid \mathbf{x}) \\ &+ \int_{-\infty}^{\infty} dx_n \cdots \int_{-\infty}^{\infty} dx_2 \int_{-\infty}^{a_{i+1}} dx_1 m_{i+1}(\mathbf{x})\hat{p}_t(y \mid \mathbf{x}). \end{aligned} \tag{90}$$

In this last equation $\hat{p}_t(y \mid \mathbf{x})dy$ is the conditional probability that $y \leqq X(t) < y + dy$ given that $\mathbf{X}(0) = \mathbf{x}$. The expressions (89) and (90) can also be found easily from the alternate definition

$$\epsilon^2 = \lim_{i\to\infty} \frac{1}{2T} E \int_{2iT}^{(2i+2)T} dt[X(t) - Z(t)]^2$$

where the expectation is over the actual time varying distribution of $X(t)$ and $Z(t)$, not the steady-state distribution. We proceed now to express ϵ^2 in terms of the f_{ivl}.

The integration on y in (89) can be carried out directly. Using standard formulae for Gaussian variates, one finds

$$\begin{aligned} E[X(t) \mid \mathbf{X}(0) = \mathbf{x}] &= \int_{-\infty}^{\infty} dyy\hat{p}_t(y \mid \mathbf{x}) \\ &= \sum_1^n c_i x_i \end{aligned} \tag{91}$$

where

$$c_i = \sum_l \alpha_{il}^{-1}\beta_{l1}(t), \tag{92}$$

and where we now exhibit explicitly the time dependence of $\beta_{ij}(t) = EX_i(u)X_j(u+t)$. We find further that

$$E[X^2(t) \mid \mathbf{X}(0) = \mathbf{x}] = \int_{-\infty}^{\infty} dy y^2 \hat{p}_t(y \mid \mathbf{x})$$

$$= \alpha_{11} - \sum_{i,j=1}^{n} \alpha_{ij}^{-1}\beta_{i1}(t)\beta_{j1}(t) + \left(\sum_{1}^{n} c_i x_i\right)^2.$$

With these results, (89) and (90) can be rearranged to give

$$\epsilon^2 = \frac{1}{T}\int_0^T dt \sum_i \left[\int_{-\infty}^{\infty} dx_n \cdots \int_{a_{i-1}}^{\infty} dx_1 m_{i-1}(\mathbf{x}) + \int_{-\infty}^{\infty} dx_n \cdots \int_{-\infty}^{a_{i+1}} dx_1 m_{i+1}(\mathbf{x})\right][A + iB + i^2\Delta^2] \tag{93}$$

where

$$A = \alpha_{11} - \sum \alpha_{ij}^{-1}\beta_{i1}(t)\beta_{j1}(t) + \sum c_i c_j x_i x_j - 2h \sum c_j x_j + h^2 \tag{94}$$

and

$$B = -2\Delta \sum c_j x_j + 2\Delta h \tag{95}$$

are independent of the index i of (93). Now

$$\sum_i \int_{-\infty}^{\infty} dx_n \cdots \int_{a_{i-1}}^{\infty} dx_1 m_{i-1}(\mathbf{x})A + \sum_i \int_{-\infty}^{\infty} dx_n \cdots \int_{-\infty}^{a_{i+1}} dx_1 m_{i+1}(\mathbf{x})A = \int_{-\infty}^{\infty} d\mathbf{x}\psi(\mathbf{x})A$$

by (23). But

$$\int_{-\infty}^{\infty} d\mathbf{x} x_j \psi(\mathbf{x}) = 0, \qquad \int_{-\infty}^{\infty} d\mathbf{x} x_i x_j \psi(\mathbf{x}) = \alpha_{ij}, \qquad \int_{-\infty}^{\infty} d\mathbf{x}\psi(\mathbf{x}) = 1,$$

so that one finds finally

$$\int_{-\infty}^{\infty} d\mathbf{x}\psi(\mathbf{x})A = \alpha_{11} - \sum \alpha_{ij}^{-1}\beta_{i1}(t)\beta_{j1}(t) + \sum c_i c_j \alpha_{ij} + h^2$$

$$= \alpha_{11} + h^2 \tag{96}$$

on using (92). The mean squared error can thus be written

$$\epsilon^2 = \alpha_{11} + h^2 + I_1 + I_2 \tag{97}$$

where

$$I_1 = \sum_i \left[\int_{-\infty}^{\infty} dx_n \cdots \int_{a_{i-1}}^{\infty} dx_1 m_{i-1}(\mathbf{x}) + \int_{-\infty}^{\infty} dx_n \cdots \int_{-\infty}^{a_{i+1}} dx_1 m_{i+1}(\mathbf{x}) \right] [2i\Delta h + i^2\Delta^2] \tag{98}$$

and

$$I_2 = \frac{1}{T} \int_0^T dt \sum_i \left[\int_{-\infty}^{\infty} dx_n \cdots \int_{a_{i-1}}^{\infty} dx_1 m_{i-1}(\mathbf{x}) + \int_{-\infty}^{\infty} dx_n \cdots \int_{-\infty}^{a_{i+1}} dx_1 m_{i+1}(\mathbf{x}) \right] [-2i\Delta \sum c_j x_j]. \tag{99}$$

The expressions (98) and (99) can be reduced further by using (47). One finds directly, for example, that

$$\begin{aligned} I_1 &= \sum_i \sum_{\nu l} \mathfrak{F}^{\nu}(T) [f_{i-1\,\nu\,l} q_{i-1\,l\,0} + f_{i+1\,\nu\,l} p_{i+1\,l\,0}][2i\Delta h + i^2\Delta^2] \\ &= \sum_{\nu} \mathfrak{F}^{\nu}(T) \sum_i f_{i\nu 0}[i2\Delta h + i^2\Delta^2]. \end{aligned}$$

Here we have used (49), (50) and (52) with $l = 0$. To reduce I_2, we first note that from (24) one has

$$-\frac{1}{\psi} \frac{\partial \psi}{\partial x_m} = \sum_j \alpha^{-1}_{mk} \mathbf{x}_j \ .$$

From (92) we thus obtain

$$-\sum c_j x_j = \frac{1}{\psi} \sum \beta_{j1}(t) \frac{\partial \psi}{\partial x_j}.$$

Using this result and (47), we find

$$I_2 = \frac{1}{T} \int_0^T dt \sum_i \sum_{\nu l} \mathfrak{F}^{\nu}(T) \cdot \sum_i \beta_{i1}(t) [f_{i-1\,\nu\,l} q_{i-1\,l} \mathbf{e}_j + f_{i+1\,\nu\,l} p_{i+1\,l} \mathbf{e}_j] 2i\Delta$$

where $\mathbf{e}_j$ is the vector having unity for its jth component and zero for all other components. Finally defining

$$Q_{il} = \sum_{j=1}^{n} q_{il\mathbf{e}_j} \frac{1}{T} \int_0^T dt \beta_{i1}(t)$$

$$P_{il} = \sum_{j=1}^{n} p_{il\mathbf{e}_j} \frac{1}{T} \int_0^T dt \beta_{i1}(t)$$

we have our desired result

$$\epsilon^2 = \alpha_{11} + h^2 + \sum_{\nu} \mathfrak{g}^{\nu}(T) \sum_{i} f_{i\,\nu 0}[i2\Delta h + i^2\Delta^2] + \sum_{\nu} \mathfrak{g}^{\nu}(T) \sum_{il} [f_{i-1\,\nu\,l} Q_{i-1\,l} + f_{i+1\,\nu\,l} P_{i+1\,l}] 2i\Delta. \quad (100)$$

IV. GENERALIZATION TO SYSTEMS WITH REALIZABLE FEEDBACK FILTERS

4.1 *Description of the System*

In this section we consider the modulator of Fig. 2 where the feedback filter is a realizable one with a rational transfer function. Again we assume that the input $X(t)$ to the modulator is the first component of a vector Markov process

$$\mathbf{X}(t) = \{X_1(t) = X(t), X_2(t), \cdots, X_n(t)\} \quad (101)$$

with $X(t)$ normalized as in (5). We shall show how an integral equation (131) that generalizes (22) can be written for the steady-state probability distribution of this system.

Let us first describe the system more precisely. The sampler acts at the instants kT, $k = 0, 1, 2, \cdots$, and its output at time jT is $X_1^{(j)} - Z^{(j)}$ where we write

$$X_i^{(j)} = X_i(jT)$$

$$Z^{(j)} = Z(jT-) \equiv \lim_{\epsilon \to 0} Z(jT - \epsilon), \qquad \epsilon > 0$$

$$i = 1, 2, \cdots, n, \quad j = 0, 1, 2, \cdots. \quad (102)$$

This output is acted upon by the quantizer, which at time jT produces an impulse of magnitude U_j that is applied instantaneously to the filter. We suppose a K-level quantizer with representative values $a_1, a_2, \cdots, a_k$ and decision regions $\mathfrak{R}_1, \mathfrak{R}_2, \cdots, \mathfrak{R}_k$. Thus,

$$U_j = a_i \quad \text{if} \quad X_1^{(j)} - Z^{(j)} \in \mathfrak{R}_i$$

$$i = 1, 2, \cdots, K, \qquad j = 0, 1, 2, \cdots. \quad (103)$$

The $\mathfrak{R}$'s are disjoint sets, the union of which exhausts the real line. We suppose the filter described by a real impulse function $h(\tau)$ with

$$h(\tau) = 0, \qquad \tau < 0, \quad (104)$$

and that the filter output is

$$Z(t) = \sum_{k=0}^{j} U_k h(t - kT), \qquad jT \leqq t < (j+1)T$$

$$j = 0, 1, 2, \cdots. \quad (105)$$

Finally, we define

$$Z^{(0)} = Z(0-) = 0 \tag{106}$$

and suppose the system inactive before time $t = 0$. Thus the filter input and output are zero for all $t < 0$. The system starts up at $t = 0$ when U_0, which depends only on $X(0)$, is applied as the first input to the feedback filter.

4.2 *The Markov Nature of the Filter*

Suppose now that the transfer function of the filter is rational,

$$h(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega e^{i\omega\tau} \frac{P(\omega)}{Q(\omega)} \tag{107}$$

where $P(\omega)$ and $Q(\omega)$ are polynomials in ω. Let the degree of Q be m and denote its roots by $i\sigma_j$, $j = 1, 2, \cdots, m$ so that

$$Q(\omega) = d \prod_{j=1}^{m} (\omega - i\sigma_j) \tag{108}$$

where d is independent of ω. For simplicity we shall assume that all m roots are distinct and that none are also roots of $P(\omega)$. Expansion of P/Q in partial fractions shows that

$$\begin{aligned} h(\tau) &= \sum_{1}^{m} \frac{1}{2\pi i} \int_{-\infty}^{\infty} d\omega e^{i\omega\tau} \frac{A_j}{\omega - i\sigma_j} \\ &= \sum_{1}^{m} A_j f(\sigma_j, \tau) \end{aligned} \tag{109}$$

where

$$f(\sigma, \tau) = \begin{cases} e^{-\sigma\tau}, & \tau > 0 \\ 0, & \tau < 0. \end{cases} \tag{110}$$

For convenience, we define $f(\sigma, 0) = 1$. Here we must have

$$\text{Re } (\sigma_j) > 0 \qquad j = 1, 2, \cdots, m \tag{111}$$

to insure (104), while the reality of $h(\tau)$ requires that non-real σ's occur in complex conjugate pairs. The corresponding A's of each such pair in (109) are complex conjugates of each other.

Now it is well known and easy to establish that when such a filter is excited by impulses as in (105), its output at all times can be given by the first component of an m-dimensional state vector

$$\mathbf{Z}(t) = \{Z_1(t) = Z(t), Z_2(t), \cdots, Z_m(t)\}, \tag{112}$$

that satisfies the equation

$$\mathbf{Z}(jT + \xi) = \mathbf{D}(\xi)\mathbf{Z}^{(j)} + U_j\mathbf{h}(\xi), \tag{113}$$

$$0 \leqq \xi < T \qquad j = 0, 1, 2, \cdots .$$

Here the $m \times m$ matrix $\mathbf{D}(\xi)$ and the m-vector $\mathbf{h}(\xi)$ are independent of j. The time-discrete state vector

$$\mathbf{Z}^{(j)} \equiv \mathbf{Z}(jT-) = \lim_{\epsilon \to 0} \mathbf{Z}(jT - \epsilon), \qquad \epsilon > 0 \tag{114}$$

satisfies the recurrence

$$\mathbf{Z}^{(j+1)} = \mathbf{D}\mathbf{Z}^{(j)} + U_j\mathbf{h} \tag{115}$$

where

$$\mathbf{D} = \mathbf{D}(T), \qquad \mathbf{h} = \mathbf{h}(T), \tag{116}$$

which follows from (113) by letting $\xi \to T$.

The validity of (113) can be established in a few lines. Let

$$\mathbf{h}(\tau) = \{h_1(\tau) = h(\tau), h_2(\tau), \cdots , h_m(\tau)\} \tag{117}$$

be the m-vector, the lth component of which is

$$h_l(\tau) \equiv \sum_1^m A_i\sigma_i^{l-1}f(\sigma_i , \tau)$$

$$= \sum_1^m c_{li}f(\sigma_i , \tau) \qquad l = 1, 2, \cdots , m \tag{118}$$

where

$$c_{li} = A_i\sigma_i^{l-1}. \tag{119}$$

For $p = 0, 1, 2, \cdots$, one then has the system of equations

$$h_l(pT) = \sum_{i=1}^m c_{li}e^{-\sigma_i pT}, \qquad l = 1, 2, \cdots , m \tag{120}$$

which can be solved inversely to give

$$e^{-\sigma_i pT} = \sum_{k=1}^m c_{ik}^{-1}h_k(pT), \qquad j = 1, 2, \cdots , m. \tag{121}$$

Here c_{ik}^{-1} is an element from the matrix inverse to $c = (c_{ij})$. The latter is non-singular since its determinant as computed from (119) is

$$|c| = \left(\prod_1^m A_i\right) \prod_{i<k} (\sigma_i - \sigma_k)$$

which is not zero by our assumption of distinct roots for $Q(\omega)$.

Now from (120) and (121) it follows that for $\xi > -pT$

$$\begin{aligned} h_l(pT + \xi) &= \sum_{j=1}^{m} c_{lj} e^{-\sigma_j \xi} e^{-\sigma_j pT} \\ &= \sum_{j=1}^{m} c_{lj} e^{-\sigma_j \xi} \sum_{k=1}^{m} c_{jk}^{-1} h_k(pT) \\ &= \sum_{k=1}^{m} d_{lk}(\xi) h_k(pT) \end{aligned}$$

or, in vector notation, that

$$\mathbf{h}(pT + \xi) = \mathbf{D}(\xi)\mathbf{h}(pT), \qquad pT + \xi > 0 \tag{122}$$

with

$$\mathbf{D}(\xi) = (d_{ij}), \qquad d_{ij} = \sum_{k=1}^{m} c_{ik} e^{-\sigma_k \xi} c_{kj}^{-1}. \tag{123}$$

This is the key to (113). We now define

$$\mathbf{Z}(t) \equiv \sum_{k=0}^{j} U_k \mathbf{h}(t - kT), \qquad jT \leqq t < (j+1)T$$

$$j = 0, 1, 2, \cdots \tag{124}$$

which has (105) for its first component. Then

$$\begin{aligned} \mathbf{Z}(pT + \xi) &= \sum_{k=0}^{p} U_k \mathbf{h}(pT - kT + \xi) \\ &= \sum_{k=0}^{p-1} U_k \mathbf{h}[(p-k)T + \xi] + U_p \mathbf{h}(\xi) \\ &= \mathbf{D}(\xi) \sum_{k=0}^{p-1} U_k \mathbf{h}[(p-1-k)T + T] + U_p \mathbf{h}(\xi) \\ &= \mathbf{D}(\xi)\mathbf{Z}(pT-) + U_p \mathbf{h}(\xi) \end{aligned} \tag{125}$$

by (122). But this is (113). For this equation to hold for $p = 0$, we must define

$$\mathbf{Z}^{(0)} \equiv 0. \tag{126}$$

4.3 *The Integral Equation*

From (115) and (102), it is seen that $\mathbf{Z}^{(j+1)}$ can be defined in terms of the random variables $\mathbf{X}^{(j)}$ and $\mathbf{Z}^{(j)}$. Since $\mathbf{X}^{(j)}$ is assumed Markovian,

it readily follows that the $m + n$ quantities

$$X_1^{(j)}, X_2^{(j)}, \cdots, X_n^{(j)}, \qquad Z_1^{(j)}, Z_2^{(j)}, \cdots, Z_m^{(j)} \tag{127}$$

constitute the components of an $(m + n)$-dimensional time-discrete vector Markov process. Denote by $m^{(j)}(\mathbf{x}, \mathbf{z})$ the joint density of $\mathbf{X}^{(j)}$ and $\mathbf{Z}^{(j)}$,

$$\begin{aligned} m^{(j)}(\mathbf{x}, \mathbf{z}) \prod_{i=1}^{n} dx_i \prod_{j=1}^{m} dz_j \\ = \Pr\{x_1 \leqq X_1^{(j)} \leqq x_1 + dx_1, \cdots, x_n \leqq X_n^{(j)} \leqq x_n + dx_n, \\ z_1 \leqq Z_1^{(j)} \leqq z_1 + dz_1, \cdots, z_m \leqq Z_m^{(j)} \leqq z_m + dz_m\}. \end{aligned}$$

Then

$$m^{(j)}(\mathbf{x}', \mathbf{z}') = \int d\mathbf{x} \int d\mathbf{z} p(\mathbf{x}', \mathbf{z}' \mid \mathbf{x}, \mathbf{z}) m^{(j-1)}(\mathbf{x}, \mathbf{z})$$

where $p(\mathbf{x}', \mathbf{z}' \mid \mathbf{x}, \mathbf{z})$ is the transition density for the process (127) and is independent of j. The steady-state distribution $m(\mathbf{x}, \mathbf{z})$ for the process must then satisfy

$$m(\mathbf{x}', \mathbf{z}') = \int d\mathbf{x} \int d\mathbf{z} p(\mathbf{x}', \mathbf{z}' \mid \mathbf{x}, \mathbf{z}) m(\mathbf{x}, \mathbf{z}). \tag{128}$$

For the case at hand, the transition density takes a very special form. Let

$$\chi_i(\mathbf{x}, \mathbf{z}) = \begin{cases} 1, & (x_1 - z_1) \in \mathcal{R}_i \\ 0, & (x_1 - z_1) \notin \mathcal{R}_i \end{cases} \qquad i = 1, 2, \cdots, K \tag{129}$$

describe the quantizer decision regions. Then from (113) and (103) we find that

$$p(\mathbf{x}', \mathbf{z}' \mid \mathbf{x}, \mathbf{z}) = \sum_{i=1}^{K} \chi_i(\mathbf{x}, \mathbf{z}) p_T(\mathbf{x}' \mid \mathbf{x}) \delta(\mathbf{z}' - D\mathbf{z} - a_i \mathbf{h}) \tag{130}$$

where δ is the usual Dirac symbol and as in (110) $p_\tau(\mathbf{y} \mid \mathbf{x})$ is the probability density of $\mathbf{X}(t + \tau)$ given $\mathbf{X}(t)$. Inserting (130) into (128) and carrying out the $\mathbf{z}$-integration gives the desired integral equation

$$\begin{aligned} |\mathbf{D}| m(\mathbf{x}', D\mathbf{z}) \\ = \sum_{i=1}^{K} \int d\mathbf{x} \chi_i(\mathbf{x}, \mathbf{z} - a_i \mathbf{D}^{-1}\mathbf{h}) p_T(\mathbf{x}' \mid \mathbf{x}) m(\mathbf{x}, \mathbf{z} - a_i \mathbf{D}^{-1}\mathbf{h}), \end{aligned} \tag{131}$$

with $|\mathbf{D}|$ the determinant of $\mathbf{D}$. This equation is to be augmented with the condition

$$\int m(\mathbf{x}, \mathbf{z})\, d\mathbf{z} = p(\mathbf{x})$$

with $p(\mathbf{x})$ as in (111).

We have not seen how to solve (131). The simplest example occurs when $m = n = 1$. We then have RC noise for the signal and an RC filter with impulse response $h(\tau) = e^{-\sigma\tau}$, $\tau > 0$, say, in the feedback path. Taking $a_1 = \Delta$, $a_2 = -\Delta$ and $\mathcal{R}_1$ the positive axis gives for (131)

$$\gamma m(x', \gamma z) = \int_{z-\Delta}^{\infty} dx m(x, z - \Delta) p(x' \mid x) + \int_{-\infty}^{z+\Delta} dx m(x, z + \Delta) p(x' \mid x)$$

where $\gamma \equiv e^{-\sigma T}$.

REFERENCES

1. de Jager, F., "Delta Modulation, A Method of PCM Transmission Using a 1-Unit Code," Philips Res. Rep., *7*, 1952, pp. 442–466.
2. Gersho, A., "Stochastic Stability of Delta Modulation," B.S.T.J., *51*, No 4 (April 1972), pp. 821–842.
3. Van de Weg, H., "Quantization Noise of a Single Integration Delta Modulation System with an N-Digit Code," Philips Res. Rep., *8*, 1953, pp. 367–385.
4. Zetterberg, L. H., "A Comparison Between Delta and Pulse Code Modulation," Ericsson Technics, *11*, No. 1, 1955, pp. 95–154.
5. O'Neal, J. B., Jr., "Delta Modulation Quantizing Noise–Analytic and Computer Simulation Results for Gaussian and Television Input Signals," B.S.T.J., *45*, No. 1 (January 1966), pp. 117–141.
6. Protonotarios, E. N., "Slope Overload Noise in Differential Pulse Code Modulation Systems," B.S.T.J., *46*, No. 9 (November 1967), pp. 2119–2161.
7. Aaron, M. R., Fleischman, J. S., McDonald, R. A., and Protonotarios, E. N., "Response of Delta Modulation to Gaussian Signals," B.S.T.J., *48*, No. 5 (May-June 1969), pp. 1167–1195.
8. Goodman, David J., "Delta Modulation Granular Quantizing Noise," B.S.T.J., *48*, No. 5 (May-June 1969), pp. 1197–1218.
9. Iwersen, J. E., "Calculated Quantization Noise of Single-Integration Delta-Modulation Coders," B.S.T.J., *48*, No. 7 (September 1969), pp. 2359–2389.
10. Fine, Terrence L., "The Response of a Particular Nonlinear System with Feedback to Each of Two Random Processes," IEEE Trans., *IT-14* (March 1968), pp. 255–264.
11. Aaron, M. R., and Stanley, T. P., "Some Statistical Properties of a Delta Modulation System Driven by a Markov Process," unpublished work (May 1968). Presented at 1969 Int. Symp. on Info. Theory, Ellenville, N. Y., January 1969.
12. Abate, J. E., "Linear and Adaptive Delta Modulation," Proc. IEEE, *55* (March 1967), pp. 298–308.
13. Davisson, L. D., "The Theoretical Analysis of Data Compression Systems," Proc. IEEE, *56* (February 1968), pp. 176–186, especially eq. (16), p. 181. Also, "Information Rates for Data Compression," IEEE, 1968 WESCON Technical Papers, Session 8 (August 1968).
14. Protonotarios, E. N., "Application of the Fokker-Planck-Kolmogorov Equation to the Analysis of Differential Pulse Code Modulation Systems," J. Franklin Inst., *289* (January 1970), pp. 31–45.

15. Jayant, N. S., "Adaptive Delta Modulation with a One-bit Memory," B.S.T.J., *49*, No. 3 (March 1970), pp. 321–342.
16. Helstrom, C. W., "Markov Processes and Applications," Balakrishnan, A. V., editor, *Communications Theory*, New York: McGraw-Hill Book Co., 1968, pp. 26–87.
17. Cramer, H., *Mathematical Methods of Statistics*, Princeton: Princeton University Press, 1946, p. 290.
18. Slepian, D., "On the Symmetrized Kronecker Power of a Matrix and Extensions of Mehler's Formula for Hermite Polynomials," SIAM J. Math. Anal., *3*, No. 4 (November 1972), pp. 606–616.
19. Erdélyi, A., *Higher Transcendental Functions*, Vol. II, New York: McGraw-Hill Book Co., 1953, pp. 283–291.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 10, December, 1972
Printed in U.S.A.

Conditions of High Gain in Mixers and Their Relation to the Jump Phenomenon

By C. DRAGONE

(Manuscript received June 21, 1972)

A study of the stability of periodically driven nonlinear networks (mixers), motivated by recent work on low-noise down-conversion with Schottky barrier diodes, is presented. Necessary and sufficient conditions for the unconditional stability of a mixer are derived and discussed. It is shown that potential instability is always associated with the jump phenomenon in the sense that a mixer will (under suitable circumstances) exhibit the phenomenon if, and only if, the above stability conditions are violated. Application of these conditions to frequency multipliers is also discussed.

I. INTRODUCTION

The Schottky barrier diode down-converter is a frequency converter that is capable of noise figures below 1 dB in the microwave range with operation at room temperature.[1] However, this converter is potentially unstable, i.e., is capable of arbitrarily high conversion gain. Evaluation of the noise performance at high gain requires a knowledge of the mechanism of instability, and of the conditions necessary and sufficient for instability. Torrey and Whitmer[2] derived a simple stability condition assuming weak reciprocity and also studied a particular case in detail, but their results are not applicable to the down-converter of Ref. 1. Here we derive general stability conditions, in closed form, and show that instability is intimately related to the jump phenomenon, a type of instability peculiar to periodically driven nonlinear networks. These stability conditions are applicable to any periodically driven nonlinear network (henceforth simply called a mixer) provided it is driven by a source (pump) that generates power at a single frequency ω_1. Because they are very general, these conditions can be used for a variety of purposes; for instance, suitable design criteria for harmonic generators can be determined in order to obviate the jump phenomenon and related instabilities in these devices. A brief discussion of this application is

given in Section 3.2 following a detailed discussion of the stability of the Schottky barrier diode down-converter in Section 3.1.

Mixer stability is reexamined using the method of Torrey and Whitmer[2] in their phenomenological theory of frequency conversion. This method determines the small-signal terminal behavior of a mixer at the input ($\omega_1 \pm p$), image ($\omega_1 \mp p$), and output (p) frequencies without any knowledge or assumptions regarding the internal structure of the mixer. The method requires that the output frequency p be very small, in which case the behavior can be derived directly from the terminal behavior of the mixer at dc and at the pump frequency ω_1. No other assumptions are made.

The small-signal terminal behavior of a mixer can be represented by a nonreciprocal three-terminal-pair network, but no simple stability criterion in closed form is known for such a network; Ku[3] has resorted to graphical and numerical methods. However, because p is assumed small, the three-port assumes special properties that permit study of its stability analytically.

II. THEORY

2.1 *Description*

A mixer that is potentially unstable can exhibit the jump phenomenon and, vice versa, a mixer exhibiting this phenomenon is potentially unstable. To acquaint the reader with our definitions and notation we begin in Section 2.2 with some preliminary considerations, including derivation of a result of Torrey and Whitmer (Ref. 2)[‡] [eq. (11)]. In Section 2.3 conditions necessary and sufficient for avoiding the jump phenomenon are derived. In Sections 2.3.3 and 2.3.4, these conditions are shown to be necessary and sufficient for unconditional stability of the mixer; their sufficiency is shown by proving that if they are fulfilled the mixer has passive behavior at $\omega_0 \pm p$, provided it is terminated in a passive impedance at p. In Section 2.4, it is pointed out that because of such behavior at $\omega_0 \pm p$, a certain type of interconnection of stable nonlinear networks is unconditionally stable.

In Section 2.5, we introduce the concept of a stable nonlinear impedance, and discuss its significance.

2.2 *Preliminary Considerations*

Suppose the mixer is represented (Fig. 1) by a two-terminal-pair network M with two ideal filters F_1 and F_0 permitting currents to flow

[‡] We are indebted to H. E. Rowe for suggesting including derivation of eqs. (11).

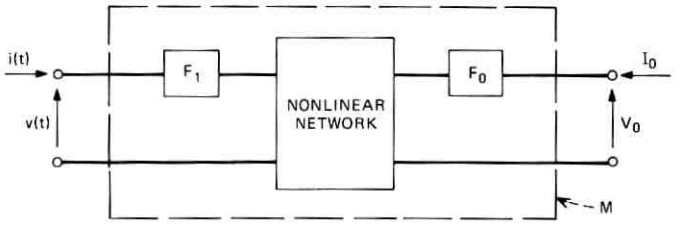

Fig. 1—Periodically driven nonlinear network M.

only in narrow bands centered about ω_1 and dc. This network is assumed to be nonlinear and to contain no sources of energy.

Let the dc and sinusoidal terminal currents be I_o and $i(t)$, the latter with complex amplitude I,

$$i(t) = 2\,Re\,(Ie^{j\omega_1 t}); \tag{1}$$

a periodic steady-state is assumed. Let $v(t)$ and V_o be the terminal voltages arising at ω_1 and dc, and let V denote the complex amplitude of $v(t)$,

$$v(t) = 2\,Re\,(Ve^{j\omega_1 t}). \tag{2}$$

Both the dc voltage and the impedance presented by the network at ω_1 are functions of I_o and $|\,I\,|$. We write

$$\begin{aligned} V_o &= \mathcal{V}_o(I_o\,,\,|\,I\,|) \\ V &= \mathfrak{z}(I_o\,,\,|\,I\,|)\cdot I. \end{aligned} \tag{3}$$

It is convenient to choose the time origin so that $i(t)$ is a cosine wave,‡

$$i(t) = 2I\cos\omega_1 t, \qquad I = |\,I\,|\,, \quad \angle I = 0. \tag{4}$$

If we superimpose small perturbations δI_o and

$$\delta i(t) = 2\,Re\,(\delta I e^{j\omega_1 t}), \tag{5}$$

on I_o and $i(t)$, and note that

$$|\,I + \delta I\,| = I + Re\,\delta I \tag{6}$$

because I is real, then eqs. (3) lead to the variational relationships

$$\begin{aligned} \delta V_o &= \frac{\partial \mathcal{V}_o}{\partial I_o}\,\delta I_o + \frac{\partial \mathcal{V}_o}{\partial\,|\,I\,|}\,Re\,\delta I \\ \delta V &= I\,\frac{\partial \mathfrak{z}}{\partial I_o}\,\delta I_o + I\,\frac{\partial \mathfrak{z}}{\partial\,|\,I\,|}\,Re\,\delta I + \mathfrak{z}\,\delta I, \quad \text{for} \quad I = |\,I\,| \end{aligned} \tag{7}$$

‡ This assumption is not used in the following section.

where δV_o and δV are perturbations on V_o and V.

When the network of Fig. 1 is used as a mixer, a small signal is applied at the input frequency ($\omega_1 + p$, or $\omega_1 - p$), and terminations are provided at the output frequency (p) and at the image frequency ($\omega_1 - p$, or $\omega_1 + p$). The input signal causes small perturbations at frequencies $\omega_1 \pm p$ and p to appear at the terminals of M. We wish to derive from eqs. (7) the relations among the various frequency components of these perturbations. Equations (7) hold without change even if δI and δI_o vary with time, provided the variations are *very slow*. Let

$$\begin{aligned} \delta I &= \delta I(t) = I_\alpha e^{ipt} + I_\gamma e^{-ipt} \\ \delta I_o &= \delta I_o(t) = I_\beta e^{ipt} + I_\beta^* e^{-ipt}. \end{aligned} \tag{8}$$

Then the terminal currents of M become

$$\begin{aligned} i(t) + \delta i(t) &= 2Re(Ie^{i\omega_1 t} + I_\alpha e^{i(\omega_1+p)t} + I_\gamma e^{i(\omega_1-p)t}) \\ I_o + \delta I_o(t) &= I_o + 2Re(I_\beta e^{ipt}). \end{aligned} \tag{9}$$

Substituting eq. (8) in eqs. (7), after replacing I with $|I|$,

$$\begin{aligned} \delta V &= \delta V(t) = V_\alpha e^{ipt} + V_\gamma e^{-ipt} \\ \delta V_o &= \delta V_o(t) = V_\beta e^{ipt} + V_\beta^* e^{-ipt}, \end{aligned} \tag{10}$$

where

$$\begin{bmatrix} V_\alpha \\ V_\beta \\ V_\gamma^* \end{bmatrix} = \begin{bmatrix} \mathfrak{z} + \frac{1}{2}|I| \frac{\partial \mathfrak{z}}{\partial |I|} & |I| \frac{\partial \mathfrak{z}}{\partial I_o} & \frac{1}{2}|I| \frac{\partial \mathfrak{z}}{\partial |I|} \\ \frac{1}{2} \frac{\partial \mathcal{V}_o}{\partial |I|} & \frac{\partial \mathcal{V}_o}{\partial I_o} & \frac{1}{2} \frac{\partial \mathcal{V}_o}{\partial |I|} \\ \frac{1}{2}|I| \frac{\partial \mathfrak{z}^*}{\partial |I|} & |I| \frac{\partial \mathfrak{z}^*}{\partial I_o} & \mathfrak{z}^* + \frac{1}{2}|I| \frac{\partial \mathfrak{z}^*}{\partial |I|} \end{bmatrix} \begin{bmatrix} I_\alpha \\ I_\beta \\ I_\gamma^* \end{bmatrix}$$

$$(\text{for } I = |I|). \tag{11}$$

Thus we have the Torrey and Whitmer result.[2] Note that according to eqs. (2) and (10) the terminal voltages produced by the currents of eqs. (9) can be written

$$\begin{aligned} v(t) + \delta v(t) &= 2Re(Ve^{j\omega_1 t} + V_\alpha e^{j(\omega_1+p)t} + V_\gamma e^{j(\omega_1-p)t}) \\ V_o + \delta V_o(t) &= V_o + 2Re(V_\beta e^{ipt}). \end{aligned} \tag{12}$$

V_α, V_γ and V_β are the complex amplitudes of the terminal voltages at $\omega_1 + p$, $\omega_1 - p$ and p, respectively.

Equation (11) describes the mixer performance subject to three assumptions

(*i*) Quasi-static (p small)
(*ii*) Small-signal ($| \delta I | \ll | I |$, $\delta I_o \ll | I_o |$)
(*iii*) Zero phase for I [eq. (4)] (not restrictive).

The matrix elements in (11), and hence the performance of the mixer, depend exclusively upon the same coefficients $\mathfrak{d}$, $| I |$ ($\partial\mathfrak{d}/\partial | I |$), etc. that characterize the small-signal terminal behavior at ω_1 and dc [see eqs. (7)].

Stability: The network M is unconditionally stable, for a given steady-state condition, if the powers

$$Re(V_\alpha I_\alpha^*), \qquad Re(V_\beta I_\beta^*), \qquad Re(V_\gamma I_\gamma^*) \tag{13}$$

absorbed at $\omega_1 + p$, p and $\omega_1 - p$ cannot *simultaneously* become negative. On the other hand, if

$$Re(V_\alpha I_\alpha^*) < 0, \qquad Re(V_\beta I_\beta^*) < 0, \qquad Re(V_\gamma I_\gamma^*) < 0 \tag{14}$$

simultaneously, then M is potentially unstable. In this case, spurious oscillations at $\omega_1 \pm p$ and p can be produced (without sources at these frequencies) by terminating M with appropriately chosen impedances at $\omega_1 \pm p$ and p. A potentially unstable mixer can have (in principle) unlimited conversion gain.

2.2.1 *The Jump Phenomenon*[‡] *and Stability*

Now suppose a one-terminal-pair network N is constructed by connecting M to a linear circuit consisting of a fixed resistance R_o in series with a constant voltage E_o, as shown in Fig. 2a. For given values of R_o and E_o, the impedance

$$Z = R + jX = \frac{V}{I} \tag{15}$$

presented at ω_1 is now a function only of the magnitude of I. Let us connect in series to this network a linear passive impedance Z_1 as shown in Fig. 2b, and let E denote the complex amplitude of the voltage $e(t)$ arising at the terminals. The behavior at ω_1 is described by the equation $E = I(Z + Z_1)$. Thus, if R_1 and X_1 denote the real and

[‡] This phenomenon is discussed in various texts on nonlinear differential equations (e.g. Ref. 4) for systems governed by Duffing's equation.

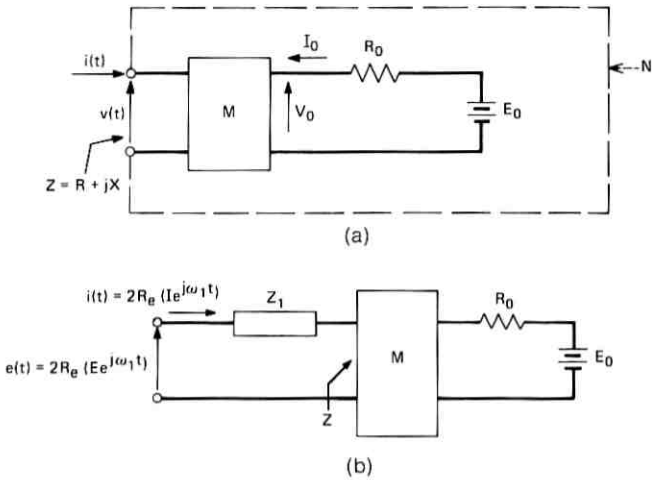

Fig. 2—Networks consisting of (a) M connected to a dc supply, and (b) M connected to a dc supply and a passive impedance Z_1.

imaginary parts of Z_1, we can write the following relation for the magnitude of E

$$|E| = |I| \sqrt{(R_1 + R)^2 + (X_1 + X)^2}, \tag{16}$$

where it is important to keep in mind that R and X are functions of $|I|$. The form of these functions depends, as seen from Fig. 2a, upon the values of R_o and E_o and the behavior of M.

The jump phenomenon occurs when $|E|$ is not a strictly monotonic function of $|I|$ (i.e., when $|E|$ has a negative slope for some $|I|$). For instance, suppose that $|E|$ has the behavior of Fig. 3, and let an ideal voltage source with zero internal impedance and variable $|E|$ be connected to N as indicated in Fig. 3. Then, if $|E|$ is gradually increased, starting from $|E| = 0$, $|I|$ will increase smoothly until it

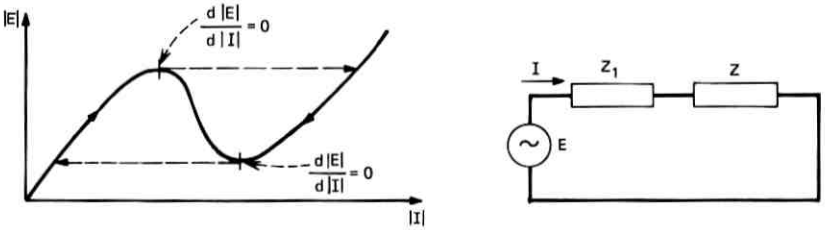

Fig. 3—Jump phenomenon.

reaches a critical value for which

$$\frac{d\,|\,E\,|}{d\,|\,I\,|} = 0.$$

At this point $|\,I\,|$ will suddenly jump to another value‡, as indicated in Fig. 3. If $|\,E\,|$ is then decreased, $|\,I\,|$ will decrease smoothly until another jump occurs, as shown in Fig. 3, for $d\,|\,E\,|/d\,|\,I\,| = 0$.

The following stability criterion, the validity of which will be proven in the following two sections, plays a central role in this paper.

Stability Criterion: Suppose one wants to determine whether or not M is unconditionally stable for a given steady-state condition. Assume in Fig. 2b that R_o and Z_1 are arbitrary, but that E_o and $|\,I\,|$ have been chosen to produce the given steady-state condition in M. It will be shown that M is unconditionally stable if, and only if, the following property is obeyed:

$$\frac{d\,|\,E\,|}{d\,|\,I\,|} > 0 \quad \text{for all} \quad R_o \geqq 0, \qquad R_1 \geqq 0, \quad \text{and} \quad X_1\,. \tag{17}$$

This result implies that, if for a given steady state M is unconditionally stable, then discontinuous jumps from the steady state in question cannot occur, no matter what the values of R_o, R_1 and X_1 may be. If, on the other hand, M is potentially unstable, then the jump phenomenon can be produced by certain choices of R_o, R_1, and X_1.

2.3 *Stability Criteria*

In the first part of this section we will show that requirement (17) demands that the behavior of $R + jX$ as a function of $|\,I\,|$ satisfy the inequality

$$4R\,\frac{d(|\,I\,|R)}{d\,|\,I\,|} > \left(|\,I\,|\,\frac{dX}{d\,|\,I\,|}\right)^2. \tag{18}$$

Since the derivatives of R and X in this inequality depend not only upon the properties of M but also upon the value of R_o, this inequality must be fulfilled for all $R_o \geqq 0$. In Section 2.3.2, we determine the relationship between R_o and the derivatives of R, X, and show that if inequality (18) is fulfilled in the two particular cases

$$R_o = \infty \tag{19}$$

‡ Assuming, of course, that for the steady-state condition corresponding to this new value of $|I|$ the circuit is stable, and that a transient leading to this new condition (from the unstable condition) exists.

and

$$R_o = 0, \tag{20}$$

then it is in general also fulfilled for all positive R_o ("in general" means: except when $\partial \mathcal{V}_o/\partial I_o \leqq 0$, as we shall see). Furthermore, it will be shown (see Section 2.3.2) that in these two cases inequality (18) becomes, respectively,

$$4\mathcal{R}\frac{\partial(|I|\mathcal{R})}{\partial|I|} > |I|^2\left(\frac{\partial\mathfrak{X}}{\partial|I|}\right)^2 \tag{21}$$

and

$$4R\frac{\partial\mathcal{V}_o}{\partial I_o}\left[\frac{\partial\mathcal{V}_o}{\partial I_o}\frac{\partial(|I|\mathcal{R})}{\partial|I|} - \frac{\partial\mathcal{V}_o}{\partial|I|}\frac{\partial(|I|\mathcal{R})}{\partial I_o}\right] > \left(|I|\frac{\partial\mathfrak{X}}{\partial|I|}\frac{\partial\mathcal{V}_o}{\partial I_o} - |I|\frac{\partial\mathfrak{X}}{\partial I_o}\frac{\partial\mathcal{V}_o}{\partial|I|}\right)^2 \tag{22}$$

where the functions $\mathcal{R} = \mathcal{R}(I_o, |I|)$ and $\mathfrak{X} = \mathfrak{X}(I_o, |I|)$ are the real and imaginary parts of $\mathfrak{z} = \mathfrak{z}(I_o, |I|)$. At the end of Section 2.3.2 we will find that for requirement (17) to be fulfilled it is necessary and sufficient that the above two inequalities be fulfilled, and that

$$R > 0, \qquad \frac{\partial\mathcal{V}_o}{\partial I_o} > 0. \tag{23}$$

In Sections 2.3.3 and 2.3.4, these inequalities are shown to be necessary and sufficient for unconditional stability of the mixer.

2.3.1 *Significance of Inequality (18)*

For a given steady-state of M, and given values of R_o and E_o in Fig. 2, we wish to show that the requirement

$$\frac{d|E|}{d|I|} > 0 \quad \text{for all} \quad R_1 \geqq 0, X_1 \tag{24}$$

is fulfilled if, and only if, $R > 0$ and inequality (18) is fulfilled.

First, note that if $R \leqq 0$ then requirement (24) is certainly violated because one can verify, using eq. (16), that $d|E|/d|I| = 0$ for

$$R_1 = -R, \qquad X_1 = -X - |I|\frac{dX}{d|I|}.$$

Thus for fulfillment of requirement (24) it is necessary that $R > 0$.

Next, we show that if $R > 0$, then for fulfillment of (24) it is necessary

and sufficient that inequality (18) be satisfied. We begin by noting that requirement (24) is equivalent to

$$\frac{|E|}{|I|}\frac{d|E|}{d|I|} > 0 \quad \text{for all} \quad R_1 \geqq 0, X_1 . \tag{25}$$

(Note that $|E| \neq 0$ because $R > 0$.) Let us therefore examine the dependence of the quantity

$$\frac{|E|}{|I|}\frac{d|E|}{d|I|} = \frac{1}{2|I|}\frac{d(|E|^2)}{d|I|} \tag{26}$$

upon X_1 and R_1. If one calculates this quantity, using eq. (16), it is found that its minimum value as a function of X_1 occurs for

$$X_1 = -\frac{1}{2}\left[X + \frac{d(|I|X)}{d|I|}\right], \tag{27}$$

while its minimum value as a function of R_1‡ occurs either for

$$R_1 = -\frac{1}{2}\left[R + \frac{d(|I|R)}{d|I|}\right] \tag{28}$$

or for $R_1 = 0$, according to whether the value given by eq. (28) for R_1 is positive or negative. In the former case one finds, using eqs. (16), (26), (27) and (28),

$$\left(\frac{|E|}{|I|}\frac{d|E|}{d|I|}\right)_{\min} = -\frac{1}{4}\left[|I|^2\left(\frac{dR}{d|I|}\right)^2 + |I|^2\left(\frac{dX}{d|I|}\right)^2\right].$$

Thus, requirement (24) is surely violated if the value given by eq. (28) for R_1 is positive. To satisfy (24) it is therefore necessary that (28) be negative. That is, it is necessary that

$$R + \frac{d(|I|R)}{d|I|} > 0, \tag{29}$$

in which case one has

$$\left(\frac{|E|}{|I|}\frac{d|E|}{d|I|}\right)_{\min} = R\frac{d(|I|R)}{d|I|} - \frac{1}{4}|I|^2\left(\frac{dX}{d|I|}\right)^2.$$

This expression is positive if (and only if) inequality (18) is satisfied. Thus, requirement (24) is fulfilled provided inequalities (29), (18) and $R > 0$ are satisfied; since the first of these is implied by the latter two,

‡ Note that we assume $R_1 \geqq 0$.

it is necessary and sufficient that only inequality (18) be satisfied, and $R > 0$.

2.3.2 *Derivation of Inequalities (21), (22) and (23)*

We can write

$$\frac{dX}{d\,|\,I\,|} = \frac{\partial \mathfrak{X}}{\partial I_o}\frac{dI_o}{d\,|\,I\,|} + \frac{\partial \mathfrak{X}}{\partial\,|\,I\,|} \quad (\mathfrak{X} \equiv \mathfrak{X}(I_o\,,\,|\,I\,|)). \tag{30}$$

An analogous expression can be written for $d\,(|\,I\,|\,R)/d\,|\,I\,|$ [substitute $X \rightarrow |\,I\,|\,R$ throughout eq. (30)]. Thus, we can write for inequality (18)

$$4R\left[\frac{\partial(|\,I\,|\,\mathfrak{R})}{\partial\,|\,I\,|} + \frac{\partial(|\,I\,|\,\mathfrak{R})}{\partial I_o}\frac{dI_o}{d\,|\,I\,|}\right] - |\,I\,|^2\left[\frac{\partial \mathfrak{X}}{\partial\,|\,I\,|} + \frac{\partial \mathfrak{X}}{\partial I_o}\frac{dI_o}{d\,|\,I\,|}\right]^2 > 0. \tag{31}$$

From Fig. 2a, $\mathfrak{V}_o(I_o\,,\,|\,I\,|) + R_o I_o = E_o$. Differentiating this relation we obtain

$$\frac{\partial \mathfrak{V}_o}{\partial I_o}\,dI_o + \frac{\partial \mathfrak{V}_o}{\partial\,|\,I\,|}\,d\,|\,I\,| + R_o\,dI_o = 0.$$

Thus,

$$\frac{dI_o}{d\,|\,I\,|} = -\frac{\partial \mathfrak{V}_o}{\partial\,|\,I\,|}\left[\frac{\partial \mathfrak{V}_o}{\partial I_o} + R_o\right]^{-1}. \tag{32}$$

Using this relation we obtain from inequality (18) in the two cases $R_o = \infty$ and $R_o = 0$, inequalities (21) and (22) respectively, as stated at the beginning of this section.

The conditions necessary and sufficient for fulfillment of inequality (31), for all $R_o \geqq 0$, are obtained by noting that this requirement demands that

$$\frac{\partial \mathfrak{V}_o}{\partial I_o} > 0, \tag{33}$$

because the magnitude of $dI_o/d\,|\,I\,|$ becomes infinite‡ (and consequently inequality (31) is violated) for $R_o \cong -(\partial \mathfrak{V}_o/\partial I_o)$. Therefore let $\partial \mathfrak{V}_o/\partial I_o > 0$. Then $dI_o/d\,|\,I\,|$ is (for $R_o \geqq 0$) a continuous function of R_o, and it varies from the value

‡ We assume $(\partial \mathfrak{Z}/\partial I_o)\cdot(\partial \mathfrak{V}_o/\partial|I|) \neq 0$. One can show that the necessary and sufficient conditions for the stability of M are also given by inequalities (21) through (23) in the special case $(\partial \mathfrak{Z}/\partial I_o)\cdot(\partial \mathfrak{V}_o/\partial|I|) = 0$.

$$\beta = -\frac{\partial \mathcal{V}_o}{\partial \mid I \mid} \Big/ \frac{\partial \mathcal{V}_o}{\partial I_o}$$

to 0, as R_o varies from 0 to ∞. If y denotes the left-hand side of inequality (31), and x denotes $dI_o/d \mid I \mid$, we have from inequality (31)

$$\frac{d^2y}{dx^2} = -2 \left| I \frac{\partial \mathfrak{X}}{\partial I_o} \right|^2 < 0.$$

It follows that y cannot have interior minima in the interval $\beta \leqq x \leqq 0$; therefore the lowest value attained by the left-hand side of inequality (31), as $dI_o/d \mid I \mid$ varies from β to 0, must occur at one of those end points. Since those two points correspond to two cases $R_o = 0$ and $R_o = \infty$, one concludes that if inequality (31) is fulfilled in these two cases then it is also fulfilled for all $R_o > 0$. The conditions necessary and sufficient that inequality (31) be fulfilled for all $R_o > 0$ are, therefore, inequalities (21), (22) and (33).

2.3.3 *Necessity of Inequalities (21) through (23) for the Unconditional stability of* M

We have derived inequalities (21) through (23) from the behavior of the network of Fig. 2a at ω_1, by requiring Z to satisfy inequality (18) for all $R_o \geqq 0$. Alternatively, these inequalities could have been derived from the dc behavior of the network of Fig. 4, by requiring that the derivative of V_o (with respect to I_o) be positive for all passive Z_1.[‡] In fact it is shown in the Appendix that this requirement and requirement (17) are equivalent; this implies that if inequalities (21) through (23) are violated, then by properly choosing Z_1, one can make the network of Fig. 4 exhibit a negative differential resistance at dc as illustrated in Fig. 5. Since such a network is potentially unstable, we conclude that inequalities (21) through (23) are necessary conditions

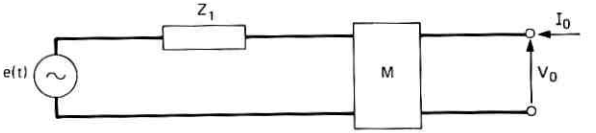

Fig. 4—Network consisting of M driven by a pump with internal impedance Z_1.

‡ This requirement is discussed in Ref. 2. We have chosen to derive our stability conditions from requirement (17), rather than the requirement $dV_o/dI_o > 0$, because one of the purposes of our derivation is to point out the relation between inequalities (21) and (22) and inequality (18). This relation is essential for the proof in the following section. The significance and practical importance of inequality (18) is pointed out in Section 2.5.

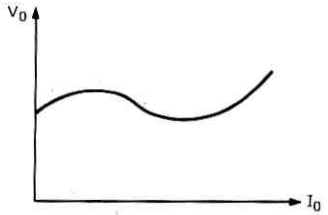

Fig. 5—Example of a dc characteristic with a negative slope.

for the unconditional stability of M. In the following section it is shown that they are also sufficient conditions.

2.3.4 *Sufficiency of Inequalities (21) through (23) for the Unconditional stability of* M

In Fig. 2a, assume that the internal impedance of the linear circuit connected to M is, instead of a frequency-independent resistance R_o, a passive impedance $Z_o(\omega)$ with the arbitrary value Z_β at $\omega = p$. Let a small perturbation $\delta i(t)$ containing the frequencies $\omega_1 \pm p$ be superimposed on the terminal current $i(t)$ of this network, as shown in Fig. 6. According to the definition of Section 2.2, M is unconditionally stable if it is impossible that

$$\text{Re } [I_\alpha V_\alpha^*] < 0, \qquad \text{Re } [I_\gamma V_\gamma^*] < 0, \tag{34}$$

simultaneously. Recall that I_α, I_γ, V_α and V_γ are the Fourier coefficients of $\delta i(t)$ and $\delta v(t)$ of $\omega_1 \pm p$ [see eqs. (9) and (12)]. In this section we show that inequalities (21) through (23) guarantee

$$\text{Re } (I_\alpha V_\alpha^* + I_\gamma V_\gamma^*) > 0. \tag{35}$$

An obvious consequence of this result is that Re $[I_\alpha V_\alpha^*]$ and Re $[I_\gamma V_\gamma^*]$ cannot simultaneously become negative, that is, M is unconditionally stable if inequalities (21) through (23) are fulfilled.

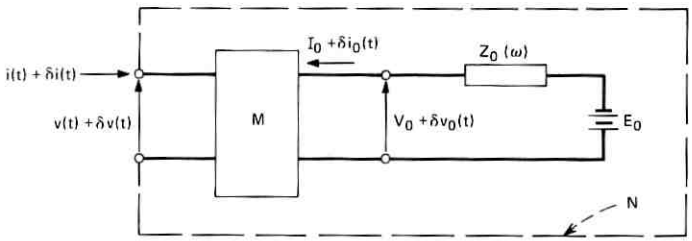

Fig. 6—Network N with perturbations at $\omega_1 \pm p$ and p.

Theorem:[‡] *If inequalities* (21) *through* (23) *are fulfilled and* $Re\ (I_\beta V_\beta^*) < 0$ *then necessarily* $Re\ (I_\alpha V_\alpha^* + I_\gamma V_\gamma^*) > 0$.

Proof:[§] The relations imposed by M among I_α, I_β, I_γ, V_α, V_β and V_γ are given by eq. (11). By using the constraint $V_\beta = -Z_\beta I_\beta$, which is imposed by the linear circuit at $\omega = p$, one may easily eliminate from eq. (11) the variables V_β and I_β, so as to obtain the following relations among I_α, I_γ, V_α and V_γ,

$$\begin{bmatrix} V_\alpha \\ V_\gamma^* \end{bmatrix} = [Z_{\alpha,\gamma}] \begin{bmatrix} I_\alpha \\ I_\gamma^* \end{bmatrix}, \tag{36}$$

where

$$[Z_{\alpha,\gamma}] = \begin{bmatrix} \mathfrak{z} + \frac{1}{2}\left[|I| \frac{\partial \mathfrak{z}}{\partial |I|} + |I| \frac{\partial \mathfrak{z}}{\partial I_o} \xi \right], & \frac{1}{2}\left[|I| \frac{\partial \mathfrak{z}}{\partial |I|} + |I| \frac{\partial \mathfrak{z}}{\partial I_o} \xi \right] \\ \frac{1}{2}\left[|I| \frac{\partial \mathfrak{z}^*}{\partial |I|} + |I| \frac{\partial \mathfrak{z}^*}{\partial I_o} \xi \right], & \mathfrak{z}^* + \frac{1}{2}\left[|I| \frac{\partial \mathfrak{z}^*}{\partial |I|} + |I| \frac{\partial \mathfrak{z}^*}{\partial I_o} \xi \right] \end{bmatrix} \tag{37}$$

$$\xi = -\frac{\partial \mathfrak{V}_o}{\partial |I|} \left[\frac{\partial \mathfrak{V}_o}{\partial I_o} + Z_\beta \right]^{-1}. \tag{38}$$

Condition (35) demands that $[Z_{\alpha,\gamma}] + [Z_{\alpha,\gamma}]^\dagger$ (the superscript $(\ )^\dagger$ denotes the Hermitian conjugate) be positive definite. If we introduce the new quantities

$$Z_u = R_u + jX_u = \frac{\partial(|I| \mathfrak{R})}{\partial |I|} + \frac{\partial(|I| \mathfrak{R})}{\partial I_o} \xi \tag{39}$$

$$Z_v = R_v + jX_v = |I| \frac{\partial \mathfrak{X}}{\partial |I|} + |I| \frac{\partial \mathfrak{X}}{\partial I_o} \xi \tag{40}$$

then from eq. (37) we obtain

$$[Z_{\alpha,\gamma}] + [Z_{\alpha,\gamma}]^\dagger = \begin{bmatrix} R_u + \mathfrak{R} - X_v & R_u - \mathfrak{R} + jR_v \\ R_u - \mathfrak{R} - jR_v & R_u + \mathfrak{R} + X_v \end{bmatrix}. \tag{41}$$

[‡] Note that the condition $Re(I_\alpha V_\alpha^* + I_\gamma V_\gamma^*) > 0$ if $Re(I_\beta V_\beta^*) < 0$ is more restrictive than the condition imposed by the requirement of stability (see Section 2.2.1).

[§] Throughout the proof it is assumed implicitly $Re(I_\beta V_\beta^*) < 0$, since the impedance $Z_\beta = Z_o(p)$ in Fig. 6 is assumed passive.

One can verify that this matrix is positive definite if, and only if, $R > 0$ and

$$4R_u\mathfrak{R} - |Z_v|^2 > 0. \tag{42}$$

Thus, in order to prove the above, theorem one must show that inequality (42) is satisfied if inequalities (21) through (23) are satisfied.

Comparison of (38) and (32) shows that for $Z_\beta = R_o$ the quantity ξ appearing in eqs. (39) and (40) equals $dI_o/d|I|$. Furthermore, if $\xi = dI_o/d|I|$, then inequality (42) reduces to inequality (31), as one can verify using eqs. (39) and (40). Thus, for $Z_\beta = R_o$, inequalities (42) and (31) are equivalent. It follows that if inequalities (21) through (23) are fulfilled then inequality (36) is certainly satisfied for $X_\beta = 0$. We will now show that if inequalities (21) through (23) are fulfilled, inequality (42) is satisfied even if $X_\beta \neq 0$.

It is convenient to introduce the quantity

$$Q = (4R_u\mathfrak{R} - |Z_v|^2)\left[\left(\frac{\partial \mathcal{V}_o}{\partial I_o} + R_\beta\right)^2 + X_\beta^2\right],$$

a product of two factors. This second factor is always positive and the first is the expression appearing in inequality (42); it follows that inequality (42) is equivalent to the condition $Q > 0$. Now let us consider the behavior of Q as a function of X_β. Using eqs. (38) through (40), it can be verified that

$$\frac{\partial Q}{\partial X_\beta} = 2X_\beta\left[4R\frac{\partial(|I|\mathfrak{R})}{\partial|I|} - \left(|I|\frac{\partial \mathfrak{X}}{\partial|I|}\right)^2\right].$$

From this relation we see that if inequality (21) is fulfilled, then the minimum value of Q (as a function of X_β) occurs for $X_\beta = 0$. That is, Q is positive for all X_β provided it is positive for $X_\beta = 0$. Since we already know that inequalities (21) through (23) insure $Q > 0$ for $X_\beta = 0$, we conclude that they also insure $Q > 0$ for all X_β. Thus, if inequalities (21) through (23) are fulfilled, then $\mathfrak{R} > 0$, $Q > 0$ and condition (35) is fulfilled.

2.4 *Lossless Interconnection of n Nonlinear Networks*

In this section certain properties of a lossless interconnection of stable nonlinear networks are discussed; these illustrate the significance of the theorem of the preceding section.

Consider n networks $N_1, N_2 \cdots, N_n$ of the same type as the network of Fig. 6. Let them be connected as shown in Fig. 7, through a $(n+1)$-terminal-pair linear time-invariant lossless network L, resulting in a

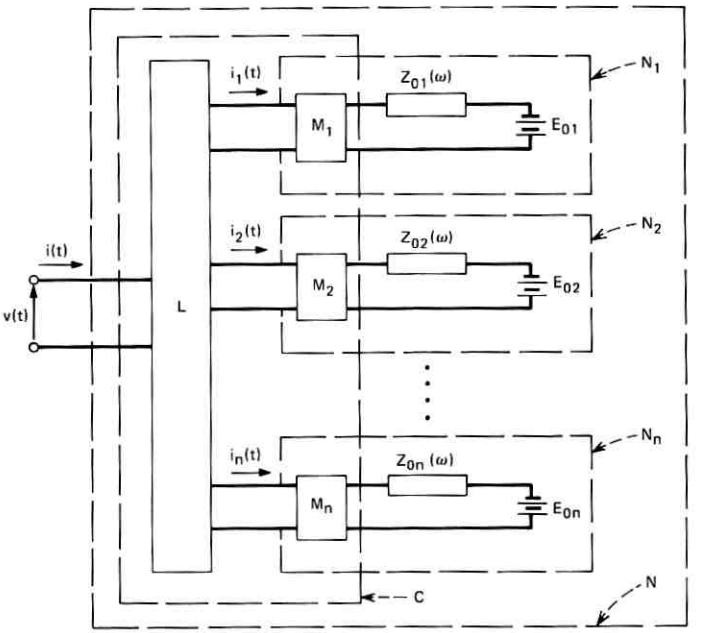

Fig. 7—Lossless interconnection of n networks $N_1, \cdots, N_n$.

one-terminal-pair network N. Let N be driven by the sinusoidal current $i(t)$ of eq. (1) and assume that $i(t)$ produces in N a periodic steady-state with frequency ω_1. Assume that $M_1, M_2, \cdots, M_n$ are unconditionally stable for this steady-state. Then the small-signal terminal behavior of N satisfies condition (35) (i.e., N is passive at $\omega_1 \pm p$), no matter what the values of the passive impedances $Z_{01}(\omega)$, $Z_{02}(\omega)$, $\cdots$, $Z_{0n}(\omega)$ for $\omega = p$ may be. This is a direct consequence of the theorem of the preceding section, which shows that if we superimpose on $i(t)$ the perturbation $\delta i(t)$ of eqs. (9), then the total power absorbed at $\omega_1 \pm p$ by N_r $(r = 1, 2, \cdots, n)$ is necessarily positive. Thus, the power absorbed by N (the sum of the powers absorbed by $N_1, \cdots, N_n$, because L is lossless) is positive.

Note that this result implies that when $M_1, \cdots, M_n$ are unconditionally stable, then the $(n + 1)$-terminal-pair network C consisting of $M_1, \cdots, M_n$ interconnected through L (Fig. 7) is also unconditionally stable. Thus, a lossless interconnection (of the type represented by the network C) of unconditionally stable networks $M_1, \cdots, M_n$ is unconditionally stable.

Another consequence of the above result is that the impedance Z

presented by N at ω_1 must satisfy inequality (18) no matter what the values of the positive dc resistances $R_{01} = Z_{01}(0)$, $R_{02} = Z_{02}(0)$, $\cdots$, $R_{0n} = Z_{0n}(0)$ may be. In fact, in the following section we will see that if inequality (18) were violated for some choice of R_{01}, R_{02}, $\cdots$, R_{0n} then the small-signal terminal behavior of N at $\omega_1 \pm p$ would be potentially unstable, which cannot be, because we have already seen that C is unconditionally stable.

2.5 *Concept of a Stable Nonlinear Impedance*

Definition: The nonlinear impedance Z presented at ω_1 by a one-terminal-pair nonlinear network N which can exchange power only in the vicinity of ω_1 and *does not* contain time-varying sources of energy (such as the network N shown in Fig. 2a or the network N of Fig. 7) is said to be stable if (and only if) it satisfies inequality (18) and $R > 0$.

An important property of a stable, nonlinear impedance has already been pointed out in Section 2.3, where it was shown that such an impedance cannot give rise to the jump phenomenon. We now want to point out another property of this impedance in connection with the small-signal terminal-behavior of N at $\omega_1 \pm p$.

If the frequency p is so small that the value of $Z_o(\omega)$ (Fig. 6) for $\omega = p$ can be assumed equal to its value for $\omega = 0$, $Z_\beta = R_o$, then the small-signal terminal behavior of the network N of Fig. 6 at $\omega_1 \pm p$ is uniquely specified by Z and the derivative of Z with respect to $|I|$. In fact, if in eq. (38) we set $Z_\beta = R_o$, then ξ can be identified as the derivative of I_o with respect to $|I|$ [see eq. (32)], and therefore according to eq. (37) the small-signal terminal behavior at $\omega_1 \pm p$ can be expressed in the form

$$\begin{bmatrix} V_\alpha \\ V^*_\gamma \end{bmatrix} = \begin{bmatrix} Z + \tfrac{1}{2}|I| \dfrac{dZ}{d|I|} & \tfrac{1}{2}|I| \dfrac{dZ}{d|I|} \\ \tfrac{1}{2}|I| \dfrac{dZ^*}{d|I|} & Z^* + \tfrac{1}{2}|I| \dfrac{dZ^*}{d|I|} \end{bmatrix} \begin{bmatrix} I_\alpha \\ I^*_\gamma \end{bmatrix}, \qquad (43)$$

provided $I = |I|$. This equation[‡] is also applicable to the network of Fig. 7, in which case p must be sufficiently small such that $Z_{or}(p) \cong R_{or}$ $(r = 1, \cdots, n)$. Now, one can easily verify using this equation that the conditions necessary and sufficient for passivity are identical to the

[‡] Note that this equation can be obtained by the same method used in Section 2.2 to derive eq. (11). In fact, the matrix of eq. (43) can be formed directly from the matrix of eq. (11) by deleting from this matrix the second row and second column and then replacing $\partial Z/\partial|I|$ with $dZ/d|I|$ throughout the resulting 2×2 matrix.

conditions necessary and sufficient for unconditional stability, and are given by inequality (18) and $R > 0$. Thus, we can say that a stable nonlinear impedance insures passive behavior against small perturbations at frequencies very close to ω_1 (this property is in accord with the theorem of the preceding section).

An interesting consequence of these results is now discussed in connection with the circuit of Fig. 7, which has been redrawn schematically in Fig. 8. Consider the small-signal terminal behavior of this network about some given steady-state condition and assume that the impedances $Z_1, \cdots, Z_n$ presented at ω_1 by the nonlinear networks $N_1, \cdots, N_n$ are stable. Let p be sufficiently small so that eq. (43) is applicable to each nonlinear network N_r (i.e., let $Z_{or}(p) \cong R_{or}$). Then each nonlinear network is passive at $\omega_1 \pm p$ and therefore N is also passive at $\omega_1 \pm p$; this implies that the impedance resulting from a lossless interconnection (of the type shown in Fig. 8) of stable nonlinear impedances is a stable impedance.‡ In particular, if two stable nonlinear impedances are connected in series, or in parallel, the resulting nonlinear impedance is stable.

This last result has an important application in connection with harmonic generators. Often such nonlinear networks are driven by pumps that are *not* linear and that can be represented by an equivalent circuit consisting of a nonlinear impedance Z_1 in series with an ideal voltage $e(t)$. The result in question shows that, even in the case of a harmonic generator driven by such a pump, the jump phenomenon can be prevented by designing the pump and harmonic generator so that both of their impedances (Z_1 and Z) satisfy inequality (18). A particular

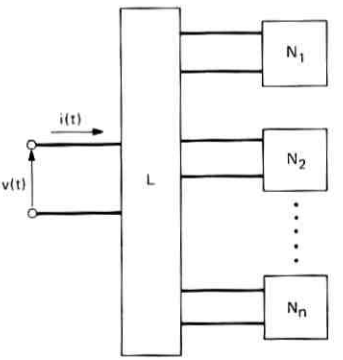

Fig. 8—Network N.

‡ Such an interconnection will therefore satisfy requirement (24).

case discussed in the following section will show that a harmonic generator can actually be designed to satisfy inequality (18) for *all* magnitudes of its input current I.

III. APPLICATIONS

Two applications are now discussed, but first we summarize some results of a previous study[1], concerning stability and noise in a Schottky barrier down-converter. That study motivated the present theory; conversely the present theory was needed in that study. In Section 3.2 results of a study of the jump phenomenon (following certain experimental work on solid-state power sources[5]) are given; other related effects (e.g. starting problems) are also discussed.

3.1 *Schottky Barrier Down-Converter*[1]

Figure 9 shows schematically a network consisting of a Schottky barrier diode connected to two filters F_1 and F_0, which permit currents to flow only in narrowbands centered about $\omega = \omega_1$ and $\omega = 0$ respectively, and which have zero impedance at those frequencies. The diode is represented (to good approximation) by the equivalent circuit of Fig. 9b, consisting of a small resistance R_s and two nonlinear elements, the barrier capacitance $C_b(v_b)$ and the barrier resistance $R_b(v_b)$. $C_b(v_b)$ and the current i_R through $R_b(v_b)$ are assumed to obey the familiar relationships

$$C_b(v_b) = \frac{C_{\min}\sqrt{\phi - V_B}}{\sqrt{\phi - v_b}} \tag{44}$$

$$i_R = i_s\left[\exp\left(\frac{qv_b}{kT}\right) - 1\right], \tag{45}$$

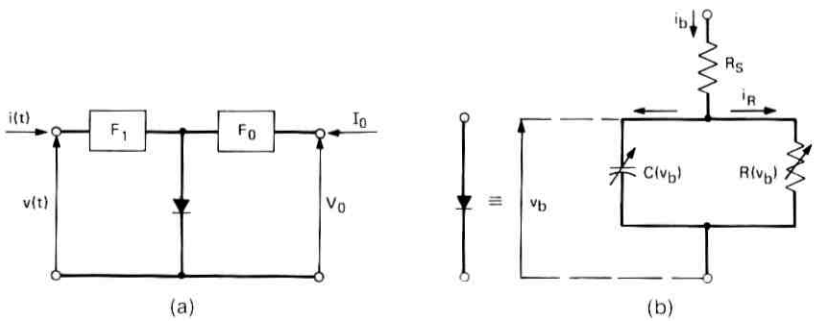

Fig. 9—Down-converter consisting of a Schottky barrier diode and two filters F_0 and F_1.

where V_B is the breakdown voltage of the diode and ϕ the contact potential.

This network is of the same type as the network M considered in the preceding sections. Therefore, its stability can be studied using inequalities (21) and (22) (for this network, inequalities (23) are always fulfilled). Its impedance Z at ω_1 can be written

$$Z = R_s + R_b + jX_b , \tag{46}$$

where $Z_b = R_b + jX_b$ is the impedance presented by the barrier of the diode. Because F_1 and F_0 allow current to flow only in the vicinity of $\omega = \omega_1$ and $\omega = 0$, the current through the diode cannot have components at the harmonics $2\omega_1$, $3\omega_1$, $4\omega_1$, etc. (and at their side frequencies $2\omega_1 \pm p$, $3\omega_1 \pm p$, $4\omega_1 \pm p$, etc.). This condition is an important requirement for low noise down-conversion. Another important requirement is that the diode should be fully pumped. That is, the current I should have the largest magnitude allowed (for a given I_0) by the breakdown voltage V_B of the diode; we assume that this is so. Then, if V_B is sufficiently large, this circuit has the following properties.[1]

First, for unconditional stability, it is sufficient (and of course necessary) that inequality (21) be fulfilled [for this circuit, inequality (22) is always fulfilled if inequality (21) is fulfilled]. Second,

$$|I| \frac{\partial \mathfrak{X}}{\partial |I|} \cong \frac{1}{2\omega_1 C_{\min}} , \tag{47}$$

where, according to eq. (44), $C_{\min}$ is the value of $C_b(v_b)$ for $v_b = V_B$. Third, the power absorbed by the barrier resistance is very small[‡], so that inequality (21) is violated provided R_s is sufficiently small. To find out how small R_s should be for the circuit to be potentially unstable, one can neglect R_b with respect to R_s in eq. (46). Then, inequality (21) requires

$$R_s > \frac{1}{2} \left| I \frac{\partial \mathfrak{X}}{\partial |I|} \right| . \tag{48}$$

From this inequality and eq. (47), we find that high gain is possible provided

$$\frac{4\omega_1}{\omega_c} < 1, \tag{49}$$

where ω_c is the cutoff frequency of the diode, $\omega_c = (R_s C_{\min})^{-1}$. According to this inequality the highest pump frequency for which a given diode

[‡] For this to be true, ω_1 must be large, such that the diode behaves essentially like a variable capacitance for $v_b < 0$ (see Ref. 1).

can be made to exhibit arbitrarily high gain is approximately $\omega_c/4$. This simple relation is valid provided V_B is very large. If V_B is finite, high gain can be obtained only if $\eta\omega_1/\omega_c < 1$, where η is a parameter which is typically less than 6.25 (but always greater than 4; see Ref. 1).

3.2 *Abrupt-Junction Varactor Doubler*[6]

Experimental varactor multipliers[7] exhibit instabilities of the type considered in this paper. In practice it is often found that empirical techniques are necessary to make a varactor multiplier self-starting. Furthermore, the range of frequencies, temperatures and powers over which a varactor multiplier shows stable and efficient operation are often seriously limited by the jump phenomenon. Little is known about the restrictions that should be imposed on multiplier design to prevent such undesirable effects.

In this section, the simplest varactor multiplier, the doubler with abrupt-junction varactor, is considered. Our main result is a stability diagram which gives, for any given varactor characteristics, the input frequencies ω_1 and load resistances R_L for which discontinuous jumps and starting problems may occur. It is shown that these nonlinear effects can always be prevented in an abrupt-junction doubler by properly choosing the output load. In particular, these effects will not occur if the output load is optimized for maximum efficiency. This result, obtained with the help of data from Penfield and Rafuse,[6] shows the importance of optimizing the efficiency of a multiplier in order to reduce starting problems, discontinuous jumps and spurious oscillations.

3.2.1 *Assumptions*[6]

The model of Fig. 9b is assumed for the diode. Since we are interested in converting power from pump frequency ω_1 to its second harmonic $2\omega_1$, the diode is assumed to be terminated by a load impedance $Z_L = R_L + jX_L$ at $2\omega_1$ and to be open-circuited at $3\omega_1$, $4\omega_1$, $5\omega_1$, etc. Furthermore, we assume that it is biased at $\omega = 0$ by a *fixed voltage* V_o.

The doubler can then be represented as in Fig. 10. The three networks F_o, F_1 and F_2 are ideal filters. The impedance of F_r ($r = 0, 1, 2$) is assumed to be zero for $\omega \cong r\omega_1$ and infinite for $|\omega - r\omega_1| > \omega_1/2$ ($\omega > 0$). We are interested exclusively in the behavior of this circuit in the particular case V_o = constant.‡

‡ The behavior for the two cases V_o = constant and I_o = constant is discussed in Ref. 1, for the limiting case $R_L = \infty$ (some of the results of Ref. 1 have been pointed out in Section 2.6). It is shown in Ref. 1 that the condition V_o = constant yields greater stability than for I_o = constant.

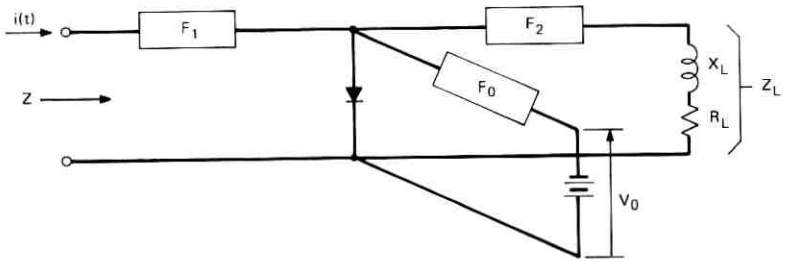

Fig. 10—Doubler.

Let I be the complex amplitude of the diode current at ω_1, and $v_b(t)$ the voltage across the barrier. The doubler will normally be operated about a particular steady state, corresponding to some value I_c of I. Let $v_{bc}(t)$ denote the voltage across the barrier for this particular steady-state, and assume that this steady-state is characterized by the condition

$$X_L = \frac{S_M - S_m}{4\omega_1}, \tag{50}$$

where S_M and S_m denote, respectively, the maximum and minimum value of the elastance of the diode for $v_b = v_{bc}(t)$ (we make this assumption because maximum efficiency for a doubler occurs approximately when this condition is fulfilled[6]). Throughout this section we also assume that the operation of the diode is restricted to the range of voltages for which the barrier capacitance is predominant over the barrier resistance. We therefore neglect the barrier resistance (see Fig. 9b) and represent the diode simply by a resistance R_s in series with a variable elastance $S(v_b) = C^{-1}(v_b)$.

According to Section 2.5, the stability of the doubler at frequencies close to ω_1 depends on the sign of the quantity

$$\eta(|I|) = 4R\frac{d(|I|R)}{d|I|} - \left(|I|\frac{dX}{d|I|}\right)^2 \tag{51}$$

where R and X are the real and imaginary part of impedance Z presented by the diode at ω_1. If, for some value of $|I|$, $\eta(|I|) < 0$, then restrictions must be imposed on the diode terminations at frequencies close to ω_1 in order to prevent spurious oscillations at these frequencies, for that value of $|I|$. Furthermore, restrictions must be imposed on the internal impedance Z_1 of the pump at ω_1, in order to prevent the jump phenomenon for $0 \leqq |I| \leqq |I_c|$. If, on the other hand

$$\eta(|I|) > 0 \quad \text{for} \quad 0 < |I| < |I_c|, \tag{52}$$

then the jump phenomenon and the above spurious oscillations will not be possible for all values of $|I|$ in the interval $0 \leqq |I| \leqq |I_c|$.

3.2.2 *Results*

The functional relationship between Z and $|I|$ has been obtained in a straightforward manner using the procedure described in Ref. 6 (pp. 299–335). For given diode characteristics, the form of this relationship depends upon the value of R_L. The effect of this parameter on the stability of Z for $S_m/S_M = 0$ is shown in Fig. 11. The stable region of this diagram gives the values of R_s and R_L for which condition (52) is fulfilled. The boundaries of this region are characterized by the property that the minimum value of $\eta(|I|)$ over the interval $0 \leqq |I| \leqq |I_c|$ is zero. It is interesting to note that there are values of R_L for which condition (52) is fulfilled even if $R_s = 0$. The dashed curve of Fig. 11 is the curve given by Penfield and Rafuse[6] for the load resistances required for maximum efficiency at $|I| = |I_c|$. Note that this curve is inside the stable region, as pointed out earlier in this section.

The unstable regions consist of the points for which $\eta(|I|) < 0$ for some values of $|I|$, $0 \leqq |I| \leqq |I_c|$. These regions can be divided into subregions having different properties, as indicated in Fig. 12. In subregions ② and ③ $\eta(|I_c|) > 0$. In ① and ④ $\eta(|I|) < 0$ even if

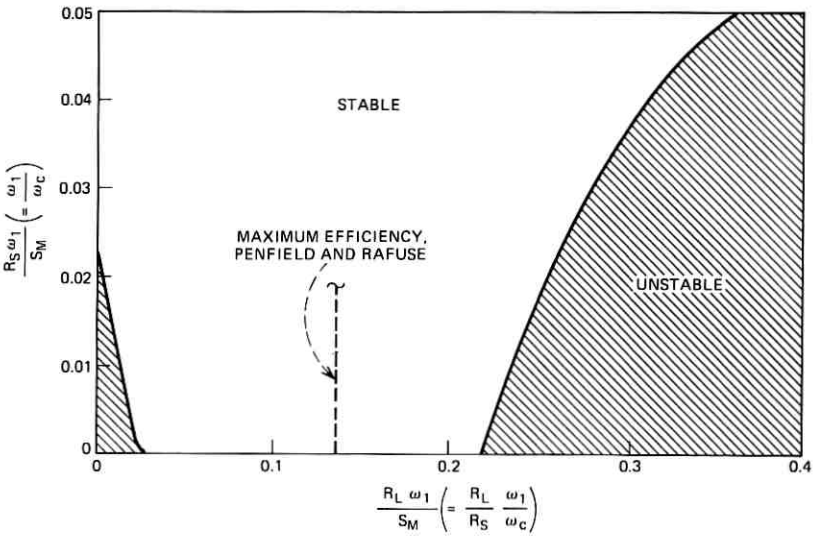

Fig. 11—Stability diagram of the abrupt-junction varactor doubler for $S_m/S_M \cong 0$.

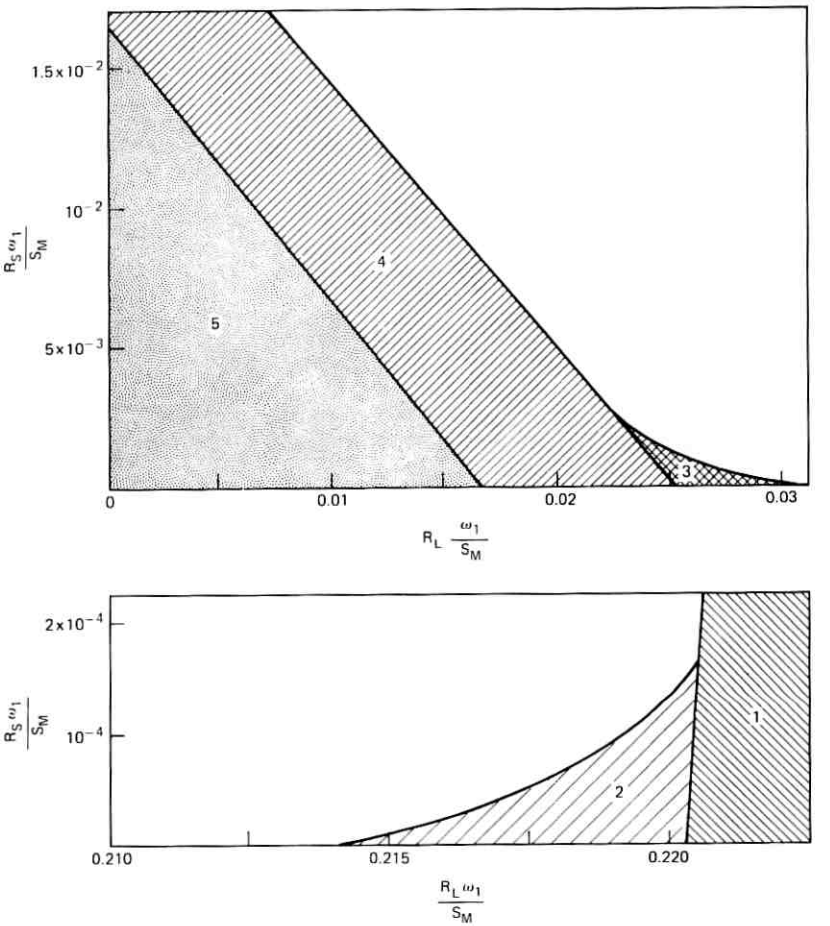

Fig. 12—Details of the diagram of Fig. 11.

$|I| = |I_e|$. For any point in one of these four subregions, the value of I always determines uniquely the voltage $v_b(t)$ across the barrier of the diode. For any point in ⑤, on the other hand, there are values of $|I|$ for which $v_b(t)$ is not uniquely determined by I, as illustrated by the example in Fig. 13, where V_m denotes the minimum value of $v_{bc}(t)$. (Thus, $S_M = C^{-1}(V_m)$). To prevent such undesirable behavior it is necessary (and sufficient) that

$$R_L + R_s > \xi \frac{S_M}{\omega_1}, \tag{53}$$

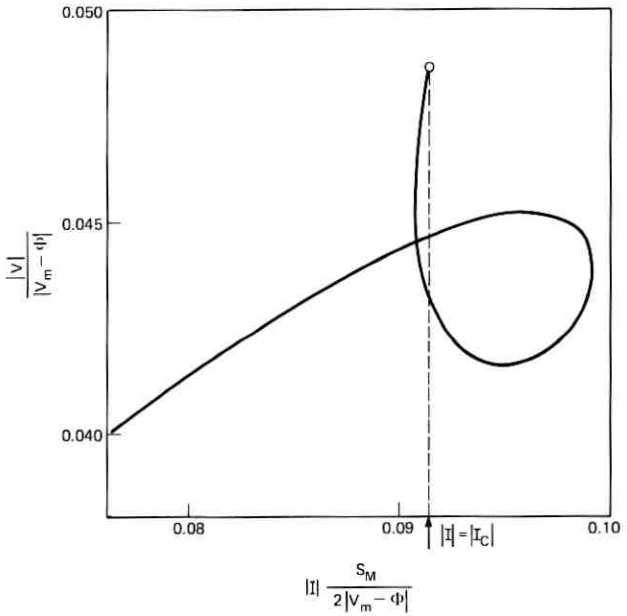

Fig. 13—Example of a $|V| - |I|$ characteristic in subregion ⑤ ($R_L = 1.14225\ 10^{-2}\ S_M/\omega_1$, $R_s = 0$).

where $\xi \cong 0.0166$. Note that in the example of Fig. 13 the desired condition $v_b(t) = v_{bc}(t)$ cannot be obtained by simply increasing $| I |$ very slowly from zero to $| I_c |$.

For points in regions ①, ②, ③ and ④ in Fig. 12, constraints must be placed upon the pump impedances Z_1 in order to avoid discontinuous jumps and starting problems. For instance, consider the case $R_s = 0$ and $R_L = 0.3565\ S_M/\omega_1$. One can see from Figs. 11 and 12 that such a multiplier is potentially unstable, since it corresponds to a point located in subregion ①. The variation of the input voltage with current is shown in Fig. 14. It can be shown that if one connects in series to the input of this multiplier, an inductance (having reactance jX_1) chosen to tune jX for $| I | = | I_c |$, the voltage E across $Z + jX_1$ will exhibit the behavior given in Fig. 15 by the curve corresponding to $x_1 = 0.5$. Figure 15 also shows two examples of the behavior arising for $X_1 < S_M/2\omega_1$ (it can be shown that $S_M/2\omega_1$ is the value of X_1 needed to tune X for $| I | = | I_c |$). All the characteristics of Fig. 15 exhibit a negative differential resistance over part of the range $0 \leqq | I | \leqq | I_c |$. Furthermore, in each case there is a range of voltages for which more than one value of $| I |$ is possible for a given value of $| E |$. In all cases the range

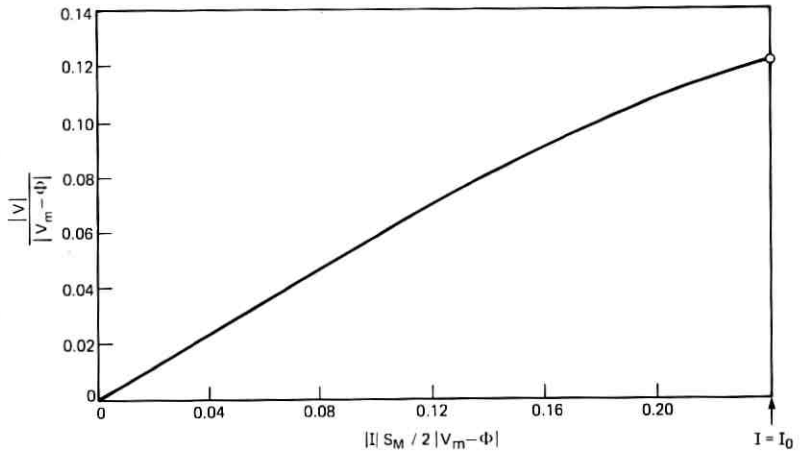

Fig. 14—Example of a $|V| - |I|$ characteristic in subregion ① ($R_L = 0.3565\ S_M/\omega_1$, $R_s = 0$).

in question contains the voltage for which $|I| = |I_c|$. The dotted curves of Fig. 15 show the effect of a small series resistance R_s; they have been calculated for $\omega_1/\omega_c = 5\ 10^{-3}$ ($R_s = 5\ 10^{-3}\ S_M/\omega_1$).

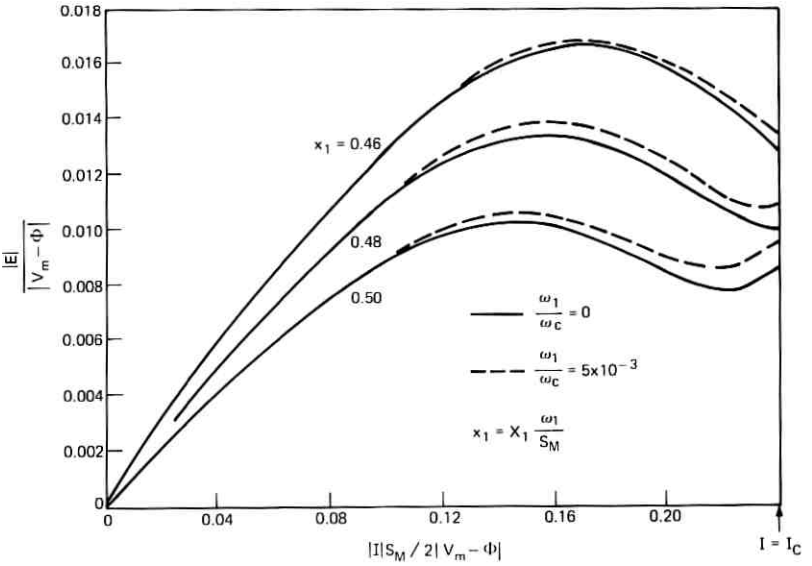

Fig. 15—$|E| - |I|$ characteristic corresponding to the example of Fig. 14 ($R_1 = 0$).

3.2.3 *Differences between this Analysis and that of Refs. 8–10*

For any point in the unstable region of Fig. 11 restrictions must be placed on the diode terminations at $\omega_1 \pm p$ (small p) in order to prevent the appearance (for some value of $| I |$ in the interval $0 \leqq | I | \leqq | I_c |$) of spurious oscillations at $\omega_1 \pm p$. The mechanism responsible for such spurious oscillations differs from that discussed in Refs. 8–10. The spurious oscillations considered in Refs. 8–10 can, in general, be prevented by imposing suitable restrictions on the diode terminations at $2\omega_1 \pm p$. In particular, they cannot occur if p is low enough so that the terminations at $2\omega_1 \pm p$ are essentially equal to Z_L. On the other hand, we have just shown that spurious oscillations are possible even if this condition at $2\omega_1 \pm p$ is fulfilled. This discrepancy between the results of the two analyses arises because the analysis of Refs. 8–10 is not applicable to the circuit of Fig. 10, since in this circuit the diode is short-circuited for $\omega = p$, whereas in Refs. 8–10 the diode was assumed to be open-circuited at $\omega = p$.[‡] Furthermore, in Refs. 8–10 the output load was assumed to be tuned, whereas in our analysis consideration has not been restricted to the particular pump level for which this condition is verified.

3.3 *Concluding Remarks*

The jump phenomenon is a form of instability. Thus it should not be surprising that a mixer capable of producing this nonlinear effect is potentially unstable, and vice versa. The derivation of inequalities (21) through (23), which are the stability conditions necessary and sufficient for the stability of a mixer, has been organized to demonstrate the important relationship between the jump phenomenon and mixer stability. A knowledge of this relation is requisite for an understanding of the mechanism of instability in a mixer; it is useful in experiments whenever one wants to determine whether or not a given mixer is potentially unstable. For that purpose the simplest procedure is to connect the mixer to a pump and a dc bias supply (as shown in Fig. 3) and then determine (in the two cases $R_o = 0$ and $R_o = \infty$) whether, by varying $| E |$ and Z_1, the circuit can be made to exhibit the jump phenomenon. This procedure is straightforward and has been used extensively in experimental work on down-converters.[1]

---

‡ More precisely, in Refs. 8–10 the impedance Z terminating $C(v_b)$ at $\omega = p$ is assumed to be sufficiently large so as to insure negligible charge fluctuations at $\omega = p$ in $C(v_b)$. However, this condition cannot by realized in the limiting case $p \to 0$ because it can be shown that for $p \to 0$ this condition requires that

$$\lim_{p \to 0} pZ_\beta = \infty$$

This requirement is unrealizable.

In Sections 2.6 and 2.7, two applications of practical interest (down-converter and doubler) have been considered; they are quite different in many respects. A fully-pumped down-converter is, in general, potentially unstable if R_s is sufficiently small, whereas a doubler can be unconditionally stable (even when $R_s = 0$), if it is properly designed. Furthermore, in the case of a down-converter, potential instability may be a desirable feature whereas it is highly undesirable in the case of a doubler. Another difference between the two cases is that in the down-converter of Fig. 9, the jump phenomenon always appears to be prevented by proper choice of R_a and Z_1, while in the case of a doubler, the behavior of Fig. 13 may arise when the doubler is improperly designed, in which case the doubler is unusable for all practical purposes. However, in spite of these differences (which arise in part because the two circuits of Figs. 9 and 10 are intended for different purposes), the two cases are related, for in the limit $R_L \to \infty$ the circuit of Fig. 10 becomes that of Fig. 9.

We conclude by summarizing the derivation of inequalities (21) and (22). A nonlinear impedance Z (with $R > 0$) obeying inequality (18) has the following property: if an arbitrary passive impedance Z_1 is connected in series with Z (Fig. 3), and E denotes the voltage $I(Z + Z_1)$ across $Z + Z_1$, then necessarily $d\,|E|/d\,|I| > 0$. In Section 2.5 such an impedance has been termed stable.

We derived inequalities (21) through (23) by connecting the network M to a dc voltage supply, and by requiring that the resulting impedance Z be stable for all nonnegative internal resistances R_a of the dc voltage supply. This procedure is analogous to that used in ordinary linear time-invariant network theory for deriving the stability conditions of a two-terminal-pair network. In fact, the stability conditions of such networks are usually derived by connecting one of its two terminal pairs to an arbitrary passive impedance, and then requiring that the resulting impedance at the other terminal pair be stable (i.e., that its real part be positive).

In Section 2.3 it was shown that if $\partial \mathcal{V}_a/\partial I_a > 0$, then Z has the following property: if inequality (18) is fulfilled in the two particular cases, $R_a = 0$ and $R_a = \infty$, then it is also fulfilled for all positive R_a. Thus it was concluded that if Z is to be stable for all nonnegative R_a, it is necessary and sufficient that $\partial \mathcal{V}_a/\partial I_a > 0$, $R > 0$, and that the two inequalities (21) and (22) [which are inequality (18) for $R_a = \infty$ and $R_a = 0$, respectively] be fulfilled. Then, in Section 2.3, we have proven a theorem showing (as a corollary) that these two inequalities, and the inequalities $\partial \mathcal{V}_a/\partial I_a > 0$ and $R > 0$, are necessary and sufficient conditions for the stability of M.

IV. ACKNOWLEDGMENT

The author is indebted to H. E. Rowe for many interesting ideas and discussions on this subject, and for his helpful criticism and advice.

APPENDIX

Let us superimpose on E_o and $|E|$ in Fig. 2 small perturbations δE_o and $\delta |E|$, and let $\delta |I|$ and δI_o denote the resulting perturbations of $|I|$ and I_o. We can write

$$\begin{bmatrix} \delta |E| \\ \delta E_o \end{bmatrix} = \begin{bmatrix} \dfrac{\partial |E|}{\partial |I|} & \dfrac{\partial |E|}{\partial I_o} \\ \dfrac{\partial E_o}{\partial |I|} & \dfrac{\partial E_o}{\partial I_o} \end{bmatrix} \begin{bmatrix} \delta |I| \\ \delta I_o \end{bmatrix} \tag{54}$$

where

$$\begin{aligned} E &= I[Z_1 + \mathfrak{z}(I_o, |I|)] \\ E_o &= \mathfrak{v}_o(I_o, |I|) + R_o I_o . \end{aligned} \tag{55}$$

Two particular cases $\delta E_o = 0$ and $\delta |E| = 0$ are of interest. In the former case, from eq. (54),

$$\left(\frac{\delta |E|}{\delta |I|}\right)_{\delta E_o = 0} = \frac{J}{\dfrac{\partial E_o}{\partial I_o}}, \tag{56}$$

where J is the determinant (Jacobian) of eqs. (54),

$$J = \frac{\partial |E|}{\partial |I|}\frac{\partial E_o}{\partial I_o} - \frac{\partial |E|}{\partial I_o}\frac{\partial E_o}{\partial |I|}. \tag{57}$$

In the latter case, from eq. (54),

$$\left(\frac{\delta E_o}{\delta I_o}\right)_{\delta |E| = 0} \quad \frac{J}{\dfrac{\partial |E|}{\partial |I|}}. \tag{58}$$

Equation (56) gives the derivative of $|E|$ with respect to $|I|$ when E_o is held constant. We have already considered this derivative in Section 2.3, where it was shown that inequalities (21) through (23) are necessary and sufficient for this derivative to be positive for all $R_o \geqq 0, R_1 \geqq 0$ and X_1. We now show that inequalities (21) through (23) can also be interpreted as the necessary and sufficient conditions for

$$\left(\frac{\delta E_o}{\delta I_o}\right)_{\delta |E| = 0} > 0. \tag{59}$$

We note from eqs. (56) and (58) that

$$\left(\frac{\delta E_o}{\delta I_o}\right)_{\delta |E|=0} = \frac{\left(\frac{\delta |E|}{\delta |I|}\right)_{\delta E_o=0} \frac{\partial |E|}{\partial |I|}}{\frac{\partial E_o}{\partial I_o}} = \frac{\frac{d |E|}{d |I|} \frac{\partial |E|}{\partial |I|}}{\frac{\partial E_o}{\partial I_o}}, \qquad (60)$$

where $d\,|E|/d\,|I|$ is the derivative discussed in Sections 2.2 and 2.3. If inequalities (21) through (23) are fulfilled, then certainly

$$\frac{d |E|}{d |I|} > 0, \qquad \frac{\partial |E|}{\partial |I|} > 0, \qquad \frac{\partial E_o}{\partial I_o} > 0 \qquad (61)$$

(note that requirement (17) implicitly demands $\partial\,|E|/\partial\,|I| > 0$, because for $R_o = \infty$, $d\,|E|/d\,|I|$ reduces to $\partial\,|E|/\partial\,|I|$). Thus, inequalities (21) through (23) are certainly sufficient conditions for requirement (59) to be fulfilled; they are also necessary because if requirement (59) is violated, then, according to eq. (59), at least one of inequalities (61) is violated for some $R_o \geqq 0$, $R_1 \geqq 0$ and X_1, and we already know from Section 2.3 that in such case, inequalities (21) through (23) are violated.

We have just shown that requirements (17) and (59) are equivalent. It can be shown, in an analogous way, that an equivalent requirement is

$$J > 0 \quad \text{for all} \quad R_o \geqq 0, \quad R_1 \geqq 0, \quad X_1 \qquad (62)$$

[other equivalent requirements may be obtained by replacing $>$ with $\neq$ in (17), (59) and (61)].

REFERENCES

1. Dragone, C., "Performance and Stability of Schottky Barrier Mixers," B.S.T.J., this issue, pp. 2169–2196.
2. Torrey, H. C., and Whitmer, C. A., *Crystal Rectifiers, 15*, Rad. Lab. Series, New York: McGraw-Hill, 1948.
3. Ku, W. H., "Stability of Linear Active Nonreciprocal N-Ports," J.F.I., (September 1963), pp. 207–224.
4. Stoker, J. J., *Nonlinear Vibrations in Mechanical and Electrical Systems*, Inc. 1950.
5. Ruthroff, C. L., Osborne, T. L., and Bodtmann, W. F., "Short Hop Radio System Experiment," B.S.T.J., *48*, No. 6 (July–August 1969), pp. 1677–1605.
6. Penfield, P., and Rafuse, R. P., *Varactor Applications*, Cambridge, Massachusetts: MIT Press, 1962.
7. Hines, M. F., Bloidsell, A. A., Collins, F., and Priest, W., "Special Problems in Microwave Harmonic Generator Chain," Digest of Technical Papers, 1962 International Solid-State Circuits Conference, Philadelphia, Pa.
8. Dragone, C., "Phase and Amplitude Modulation in High Efficiency Varactor Frequency Multipliers of Order $N = 2^n$—General Scattering Properties," B.S.T.J., *46*, No. 4 (April 1967), pp. 775–796.
9. Dragone, C., "Phase and Amplitude Modulation in High Efficiency Varactor Frequency Multipliers of Order $N = 2^n$—Stability and Noise," B.S.T.J., *46*, No. 4 (April 1967), pp. 797–834.
10. Dragone, C., and Prabhu, V. K., "Some Considerations of Stability in Lossy Varactor Harmonic Generators," B.S.T.J., *47*, No. 6 (July–August 1968), pp. 887–896.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 10, December, 1972
Printed in U.S.A.

Performance and Stability of Schottky Barrier Mixers

By C. DRAGONE

(Manuscript received June 21, 1972)

We discuss the performance of a Schottky barrier diode as a mixer when the barrier of the diode is open-circuited at the harmonics $2\omega_o$, $3\omega_o$, *etc. of the pump frequency* ω_o . *Such a mixer is shown to be capable of arbitrarily high conversion gain provided*

$$\omega_c \geqq \eta\omega_o ,$$

where ω_c *is the cutoff frequency of the diode and* η *is a parameter that is typically less than* 6.25 *and approaches* 4 *under certain ideal conditions. It is shown that the limit imposed by the series resistance of the diode on the double-sideband noise figure of the mixer is given by*

$$F_m > \left(1 - \eta \frac{\omega_o}{\omega_c}\right)^{-1} .$$

An experiment is described at 1.25 GHz *on a room temperature mixer whose double-sideband noise figure* F_m *as a function of gain has a minimum of about* 0.7 *dB (for gain less than unity) and a maximum of about* 2.3 *dB (for high gain).*

I. INTRODUCTION

In the past five years the performance of microwave mixers has been substantially improved with the advent of high quality Schottky barrier diodes.[1] The noise figures obtained so far are better by a factor of approximately 2 than those obtained previously using point-contact diodes.[1-4] The ultimate microwave noise figure obtainable with these devices is not yet known, but there is reason to believe that, at room temperature, a figure well under 3 dB is possible.

Calculation in a previous article[5] showed that a Schottky barrier diode with suitable characteristics should have a noise figure well under 1 dB provided the barrier of the diode is open-circuited at the harmonics

$2\omega_o$, $3\omega_o$, etc. of the pump frequency ω_o. However, that calculation neglects the barrier capacitance and is therefore valid only at low frequency. The present purpose is to investigate the effect of the barrier capacitance. Three main assumptions are made: (*i*) the barrier of the diode is open-circuited at $2\omega_o$, $3\omega_o$, etc, (*ii*) the output frequency p of the mixer is very low with respect to ω_o, and (*iii*) the diode can be represented by an equivalent circuit discussed in Section II.

A mixer is commonly considered to be a linear transducer having finite maximum gain and a noise temperature ratio close to unity; the maximum gain is usually considered its most important attribute.[1] However, this picture is not valid in general for the mixer under consideration here. It will be shown that this mixer is potentially unstable if the cutoff frequency ω_c of the diode is sufficiently high with respect to ω_o. Thus its gain is unlimited, in the sense that it can be made arbitrarily high by appropriately choosing the terminations at the input, image, and output frequencies (i.e., at $\omega_o \pm p$ and p).

In a previous article[6] the mechanism responsible for instability in a mixer was discussed and necessary and sufficient conditions for unconditional stability were derived. These conditions, given in Section II, are used to determine the relation between mixer stability and mixer parameters. The main result is that a mixer is potentially unstable (i.e., that high gain is possible) if and only if

$$\omega_c \geqq \omega_o \eta, \tag{1}$$

where η is a parameter the value of which depends primarily on the breakdown voltage V_B of the diode. It is shown that $\eta \to 4$ as $V_B \to \infty$ and that typically

$$4 < \eta < 6.25. \tag{2}$$

The value η has important significance in connection with the noise performance of a mixer *at high gain* because the limit imposed by the series resistance of the diode on the ultimate noise performance is given by the inequality

$$F_m > \left(1 - \eta \frac{\omega_o}{\omega_c}\right)^{-1}, \tag{3}$$

where F_m is the double-sideband noise figure.

Following the analysis of Ref. 5, an experiment was undertaken to determine the performance obtainable from a mixer satisfying assumptions (*i*), (*ii*), and (*iii*). We designed such a mixer and measured its

behavior as discussed in Section VI. It was, as expected, potentially unstable. The double-sideband noise figure as a function of gain was found to have a minimum value of 0.7 dB, occurring at a gain less than unity, and a maximum of about 2.3 dB, at high gain.

High gain in a mixer is no new phenomenon; it was demonstrated both theoretically and experimentally more than 20 years ago.[7] Since then, the effect of the barrier capacitance has been treated by several authors.[8-12] However, to the best of our knowledge, the effect of the barrier capacitance in a mixer satisfying assumption (*i*) has never been studied before. The amplifying ability is not a surprising property (Ref. 6), but good noise performance at high gain is perhaps unexpected.

II. PRELIMINARY CONSIDERATIONS

The equivalent circuit of Fig. 1 is assumed for the Schottky barrier diode; it consists of a small series resistance R_s and two nonlinear elements, the barrier capacitance $C(v_b)$ and the barrier resistance $R(v_b)$. The capacitance $C(v_b)$ and the current i_R through $R(v_b)$ are assumed to obey the familiar relations

$$C(v_b) = \frac{C_0\sqrt{\phi}}{\sqrt{\phi} - v_b} \tag{4}$$

and

$$i_R = i_S\left[\exp\left(\frac{qv_b}{kT}\right) - 1\right], \tag{5}$$

where ϕ is the contact potential, i_S the saturation current, q the electronic charge, k the Boltzman constant, and T the absolute temperature; $q/kT \cong 40$ for $T \cong 290°\text{K}$.

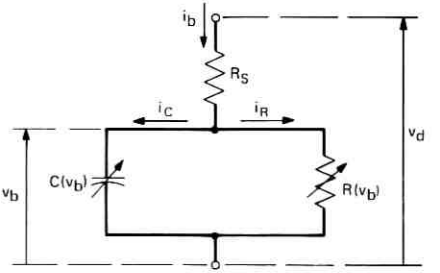

Fig. 1—Schottky barrier diode.

Figure 2 shows a two-terminal-pair network M driven by a sinusoidal current

$$i(t) = 2I \cos \omega_o t \tag{6}$$

and a direct current I_o. This network consists of the diode and two filters F_o and F_{ω_o}. According to assumption (*ii*), we assume for F_o and F_{ω_o} the following characteristics at the harmonics of ω_o (and in their vicinity): F_o is a short circuit at dc and an open circuit at ω_o, $2\omega_o$, $3\omega_o$, etc.; F_{ω_o} is a short circuit at ω_o and an open circuit at dc, $2\omega_o$, $3\omega_o$, etc. From Fig. 2 the terminal current of the diode is

$$i_b(t) = I_o + i(t). \tag{7}$$

The voltage v_b across the barrier is assumed periodic with frequency ω_o,

$$v_b(t) = v_{bo} + 2 \text{ Re } (V_{b1} e^{j\omega_o t} + \cdots), \tag{8}$$

where the dots indicate components at $2\omega_o$, $3\omega_o$, etc. Let Z_b denote the impedance presented by the barrier at ω_o.

$$Z_b = R_b + jX_b = \frac{V_{b1}}{I}. \tag{9}$$

Then the impedance Z presented by the network at ω_o is

$$Z = R + jX = R_S + Z_b. \tag{10}$$

The terminal voltage at dc is

$$V_o = V_{bo} + R_S I_o. \tag{11}$$

When this network (designated here by M; see Fig. 2) is used as a

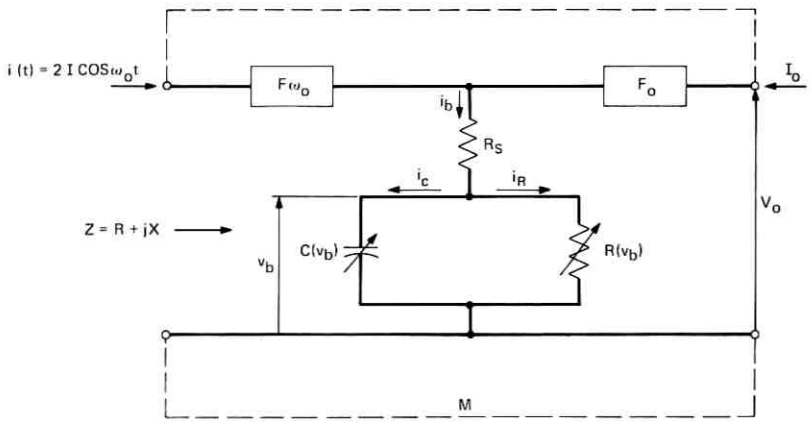

Fig. 2—Network M representing the mixer.

mixer, a small signal is applied at the input frequency $\omega_o + p$ (or $\omega_o - p$) and suitable terminations are provided at the image and output frequencies $\omega_o - p$ (or $\omega_o + p$) and p. Signal power then flows out of the network at $\omega = p$ to the output termination. The conversion gain, which is defined* here as the ratio of the output power to the power available from the input signal generator, has a finite maximum value only if M is unconditionally stable. If M is *not* unconditionally stable (i.e., if M is potentially unstable), its gain can be made arbitrarily high by properly choosing the terminations at $\omega_o \pm p$ and p. To determine the conditions for which M is potentially unstable, we assume $p \ll \omega_o$ [assumption (*ii*)]; this allows us to use the stability criteria of Ref. 6, which are discussed in the following part of this section. Since application of these criteria does not require a knowledge of the conversion properties of M at $\omega_o \pm p$ and p, the analysis will be concerned exclusively with the behavior of M at ω_o and dc.

2.1 *Stability Criteria*[6]

Let a one-terminal-pair network be constructed by connecting M to a dc source as shown in Fig. 3. The nonlinear impedance Z characterizing the terminal behavior at ω_o of this network is a function of the amplitude I of $i(t)$. The form of this function depends upon the characteristics of the dc source. Of particular interest are the two cases arising when the dc source is: (*i*) an ideal current source with infinite internal impedance, or (*ii*) an ideal voltage source with zero internal impedance. It has been shown in Ref. 6 that a necessary and sufficient condition for the network M to be potentially unstable is

$$4R \frac{d(IR)}{dI} \leqq \left(I \frac{dX}{dI}\right)^2 \tag{12}$$

in one of the above two cases.†

The network M imposes a set of nonlinear relations between its terminal voltages and currents at dc and ω_o. Because of these relations, the impedance Z and the dc voltage V_o can be regarded as functions of I_o and I,

$$R = R(I_o, I), \qquad X = X(I_o, I)$$
$$V_o = V_o(I_o, I). \tag{13}$$

* There are several ways of defining the gain of a mixer. A different definition will be used [see eq. (51)] in connection with the experiment described in Section VI, where the input signal generator will contain both $\omega_o + p$ and $\omega_o - p$. However, the particular definition is immaterial to the analysis.

† For the network of Fig. 2 one can show that $R > 0$ and $\partial V_o/\partial I_o > 0$; if these two inequalities are not satisfied, the network would obviously be potentially unstable.

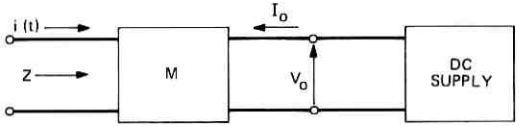

Fig. 3—Mixer connected to a dc bias supply.

These three functions completely describe the terminal behavior of M. If these functions are known, the derivatives appearing in inequality (12) can be evaluated as follows.

Note first that the two cases mentioned above correspond to the two conditions

$$I_o = \text{constant}, \tag{14}$$

$$V_o = \text{constant}. \tag{15}$$

Thus, in the former case, inequality (12) takes the form*

$$4R \frac{\partial(IR)}{\partial I} \leqq \left(I \frac{\partial X}{\partial I}\right)^2, \tag{16}$$

involving only partial derivatives of $IR(I_o, I)$ and $X(I_o, I)$ with respect to I.

In the latter case, differentiating eq. (15) we obtain

$$dI \frac{\partial V_o}{\partial I} + dI_o \frac{\partial V_o}{\partial I_o} = 0.$$

Thus,

$$\frac{dI_o}{dI} = -\frac{\partial V_o}{\partial I}\left(\frac{\partial V_o}{\partial I_o}\right)^{-1}.$$

Using this relation and the rule

$$\frac{d}{dI} = \frac{dI_o}{dI}\frac{\partial}{\partial I_o} + \frac{\partial}{\partial I},$$

we obtain for inequality (12)

$$4R \frac{\partial V_o}{\partial I_o}\left[\frac{\partial V_o}{\partial I_o}\frac{\partial(IR)}{\partial I} - \frac{\partial V_o}{\partial I}\frac{\partial(IR)}{\partial I_o}\right] \leqq \left[I \frac{\partial X}{\partial I}\frac{\partial V_o}{\partial I_o} - I \frac{\partial X}{\partial I_o}\frac{\partial V_o}{\partial I}\right]^2. \tag{17}$$

Inequalities (16) and (17) will be useful in Section IV.

* Throughout this paper $\partial/\partial I_o$ (or $\partial/\partial I$) is used to indicate differentiation with I (or I_o) held constant.

In the following section we consider the behavior of M in a limiting case and find that a simple relation exists between Z and V_o, I. Because of that relation, it is possible to ascertain stability directly from inequality (12), rather than using (16) and (17).

III. ANALYSIS OF A LIMITING CASE

Analysis of the circuit of Fig. 2 is a considerable task, primarily because the current i_c through $C(v_b)$ is not in general simply related to the total current i_b through the barrier. However, if the frequency ω_o is sufficiently high, the current i_R absorbed by $R(v_b)$ is much smaller than i_c for all t, and i_c is approximately equal to the alternating component of i_b.

$$i_c(t) \cong 2I \cos \omega_o t. \tag{18}$$

In the present section we study this case. Our main result is inequality (37).

If q and S denote the charge and elastance of the barrier capacitance respectively, then, since $S = dv_b/dq$ and $i_c = dq/dt$, we can write

$$\frac{dS}{dt} = \frac{dS}{dv_b}\frac{dv_b}{dq}\frac{dq}{dt} = \frac{1}{2}\frac{d[S^2]}{dv_b} i_c . \tag{19}$$

From eq. (4) and the fact that $S = C^{-1}(v_b)$,

$$S^2 = \frac{\phi - v_b}{(C_o\sqrt{\phi})^2}. \tag{20}$$

Taking the derivative and substituting in eq. (19) results in

$$\frac{dS}{dt} = -\frac{1}{2(C_o\sqrt{\phi})^2} i_c . \tag{21}$$

From this equation the alternative component of the elastance $S(t)$ produced by the current $i_c(t)$ of eq. (18) is determined. If S_o denotes the average value of $S(t)$, using eqs. (18) and (21),

$$S(t) = S_o + \frac{1}{(C_o\sqrt{\phi})^2}\left[\frac{jI}{2\omega_o} e^{j\omega_o t} - \frac{jI}{2\omega_o} e^{-j\omega_o t}\right]. \tag{22}$$

Substituting this equation in eq. (20) results in

$$v_b(t) = \phi - (C_0\sqrt{\phi})^2 S_o^2 - \frac{I^2}{(C_0\sqrt{\phi})^2 2\omega_o^2} + \left(-j\frac{S_o}{\omega_o} I e^{j\omega_o t} + j\frac{S_o}{\omega_o} I e^{-j\omega_o t} + \cdots\right), \tag{23}$$

where the dots indicate components at $\pm 2\omega_o$. The amplitudes of $v_b(t)$ at dc and ω_o, obtained from (23), are

$$V_{bo} = \phi - (C_o\sqrt{\phi})^2 S_o^2 - \frac{I}{(C_o\sqrt{\phi})^2 2\omega_o^2} \tag{24}$$

and

$$V = -j\frac{S_o I}{\omega_o}. \tag{25}$$

Thus, the reactance X presented by the barrier capacitance at ω_o can be written

$$X = -\frac{S_o}{\omega_o}. \tag{26}$$

From eq. (24)

$$S_o = \frac{1}{C_o\sqrt{\phi}}\sqrt{\phi - V_{bo} - \frac{I^2}{(C_o\sqrt{\phi})^2 2\omega_o^2}}. \tag{27}$$

Equations (26) and (27) specify the reactance X in terms of the two independent variables I and V_{bo}.

We now make the assumption that the diode is so operated that the power* absorbed by $R(v_b)$ at ω_o is much smaller than the power dissipated in R_S at ω_o. Because of this assumption, which is consistent with eq. (18), the impedance Z presented by the diode at ω_o is simply $R_S + jX$. According to the preceding section the stability of the network of Fig. 2 can be ascertained from the behavior of $R_S + jX$ using inequality (12), which reduces to

$$R_S \leqq \frac{1}{2} I \left| \frac{dX}{dI} \right| \tag{28}$$

because R_S is independent of I. In order for the network of Fig. 2 to be potentially unstable, inequality (28) must be fulfilled under at least one of the two conditions (14) and (15).

Consider first condition (15), in which case V_{bo} can be assumed independent of I ($V_{bo} \cong V_o$ because $R_S I_o \cong 0$). From eqs. (26) and (27) one obtains

$$\frac{dX}{dI} = \frac{I}{2(C_0\sqrt{\phi})^4 \omega_o^3 S_o}. \tag{29}$$

* This power is $\langle 2(Re)V_{bl}e^{j\omega_o t} i_R\rangle_{\text{ave}}$; it can be calculated to a first approximation using eqs. (5), (23), and (27).

Let S_M and S_m denote the maximum and minimum value of $S(t)$. From eq. (22) one can verify that I and S_o are related to S_M and S_m as follows:

$$S_o = \frac{S_M + S_m}{2} \tag{30}$$

$$I = \omega_o(C_o\sqrt{\phi})^2 \frac{S_M - S_m}{2}. \tag{31}$$

These relations and eq. (29) yield

$$I\frac{dX}{dI} = \frac{1}{4}\frac{S_M}{\omega_o}\frac{(1 - S_m/S_M)^2}{(1 + S_m/S_M)} \tag{32}$$

which is valid provided V_{bo} is independent of I.

Now consider condition (14) where V_{bo} is a function of I. In general, no simple relation exists between X and I; however, in Appendix A it is shown that if for a given value of the ratio S_m/S_M the inequality

$$\left|\frac{qv_m}{kT}\right| \gg 1 \tag{33}$$

obtains [v_m is the minimum value of $v_b(t)$], then condition (14) becomes equivalent to

$$S_m = \text{constant} \tag{34}$$

which leads to a simple relation between X and I. In fact, using eqs. (26), (30), and (31), X can be expressed in the form

$$X = -\frac{1}{\omega_o}\left[S_m + \frac{I}{\omega_o(C_o\sqrt{\phi})^2}\right], \tag{35}$$

and X is linearly related to I. Further, using eqs. (31) and (35),

$$I\frac{dX}{dI} = -\frac{S_M - S_m}{2\omega_o}. \tag{36}$$

We now compare the two cases I_o = constant and V_o = constant, under the assumption that in the former case condition (33) is satisfied. From eqs. (32) and (36), for given S_M, S_m, and ω_o, eq. (36) gives a larger magnitude for IdX/dI than eq. (32). Since eq. (36) corresponds to the condition I_o = constant, we conclude that the network of Fig. 2 is potentially unstable only if inequality (28) is fulfilled in that case. From inequality (28) and eq. (36),

$$R_S \leqq \frac{S_M - S_m}{4\omega_o}, \tag{37}$$

which gives the values of R_S , S_M , S_m , and ω_o for which instability (and therefore high gain) is possible. If ω_c denotes the cutoff frequency of the diode, so that $\omega_c = S_M/R_S$, and if $S_m \ll S_M$, then inequality (37) reduces to

$$\omega_c \geqq 4\omega_o \tag{38}$$

which is eq. (1) for $\eta = 4$.

An understanding of the practical validity of inequality (37) is obtained by examining the restrictions in the above analysis. It has been assumed that the voltage across the barrier can be determined to a first approximation by neglecting the barrier resistance and also that the power absorbed at ω_o by the barrier resistance is negligible compared with that dissipated in R_S . These assumptions are certainly satisfied if operation of the diode is restricted to a range of voltages v_b for which the barrier capacitance is predominant over the barrier resistance. However, this restriction is impractical because it would result in a mixer with very poor performance; for optimum performance the diode should be fully pumped; that is, $v_b(t)$ should vary over the entire usable range of forward and reverse voltages. In the following section it is shown that inequality (37) is valid, approximately, even if the diode is fully pumped (in which case the restriction in question is not satisfied), *provided* the breakdown voltage V_B *is sufficiently large.* Thus, we can say that this requirement on V_B [which is in accord with the fact that inequality (37) has been derived under requirement (33)] is the main restriction on inequality (37).

IV. GENERAL CASE

According to eqs. (4) and (5), the voltage and current at the barrier are related through the nonlinear differential equation

$$\frac{C_o\sqrt{\phi}}{\sqrt{\phi - v_b}}\frac{dv_b}{dt} + i_s e^{(q/kT)v_b} - i_s - i_b = 0, \tag{39}$$

with i_b given by eqs. (6) and (7). This equation cannot in general be solved exactly, but an approximate solution can be obtained fairly simply to any degree of accuracy by the Euler method, as shown in Appendix B. Using that method, Z_b , V_{bo} , and their partial derivatives with respect to I_o and I [these derivatives are needed to test the stability of M using inequalities (16) and (17)] have been calculated for various diode characteristics and terminal currents I_o and I. Table I shows

TABLE I—FOUR EXAMPLES OF THE BEHAVIOR OF M

$\frac{i_s e^{(q\phi/kT)}}{C_o\omega_o\sqrt{\phi}} = 31.25$	$\frac{I_o + i_s}{C_o\omega_o\sqrt{\phi}} = 0.1$			
$\lvert I\rvert/C_o\omega_o\sqrt{\phi}$	0.75	1.00	1.25	1.75
$V_{bo} - \phi$	−0.722	−1.116	−1.603	−2.857
$V_m - \phi$	−1.648	−2.660	−3.919	−7.180
$V_M - \phi$	−0.093	−0.089	−0.086	−0.082
$R_b \cdot \omega_o C_o\sqrt{\phi}$	0.110	0.101	0.096	0.090
$X_b \cdot \omega_o C_o\sqrt{\phi}$	−0.725	−0.904	−1.080	−1.432
$\partial V_{bo}/\partial I_o \cdot \omega_o C_o\sqrt{\phi}$	2.016	2.577	3.133	4.243
$\partial V_{bo}/\partial\lvert I\rvert \cdot \omega_o C_o\sqrt{\phi}$	−1.389	−1.762	−2.134	−2.878
$\partial(\lvert I\rvert R_b)/\partial I_o \cdot \omega_o C_o\sqrt{\phi}$	0.576	0.774	0.966	1.345
$\partial(\lvert I\rvert R_b)/\partial\lvert I\rvert \cdot \omega_o C_o\sqrt{\phi}$	0.0781	0.0763	0.0757	0.075
$\lvert I\rvert\partial X_b/\partial I_o \cdot \omega_o C_o\sqrt{\phi}$	1.403	1.784	2.166	2.937
$\lvert I\rvert\partial X_b/\partial\lvert I\rvert \cdot \omega_o C_o\sqrt{\phi}$	−0.542	−0.710	−0.881	−1.228

four examples where

$$\frac{i_S e^{(q\phi/kT)}}{C_o\omega_o\sqrt{\phi}} = 31.25, \qquad \frac{I_o + i_S}{C_o\omega_o\sqrt{\phi}} = 0.1 \cdot \tag{40}$$

The four examples correspond to various values of the quantity $I/C_o\omega_o\sqrt{\phi}$, which represents the terminal current at ω_o. The values v_m and v_M in Table I are the minimum and maximum value of $v_b(t)$. Note that according to eq. (20) S_M and S_m are related to v_M and v_m through the relations

$$S_m = \frac{\sqrt{\phi - v_M}}{C_o\sqrt{\phi}}, \qquad S_M = \frac{\sqrt{\phi - v_m}}{C_o\sqrt{\phi}}. \tag{41}$$

One can verify that, in all the four cases of Table I, inequality (16) is violated for $R_S = 0$. Thus, in each case the circuit of Fig. 2 is potentially unstable provided the series resistance R_S is sufficiently small.

The values of R_S associated with potential instability can be derived as follows. In all the cases considered it has been found that if inequality (17) is fulfilled, then inequality (16) is also fulfilled; in other words, if Z satisfies inequality (12) under the condition V_o = constant, then it also satisfies inequality (12) under the condition I_o = constant. This property has already been found to be true in the limiting case discussed in the preceding section. It follows that for the network to be potentially unstable it is necessary (and, of course, sufficient) that $Z = R_S + R_b + jX_b$ satisfy inequality (16). That is, it is necessary

that the expression

$$4(R_s + R_b)\left[\left(\frac{\partial(IR_b)}{\partial I}\right) + R_s\right] - I^2\left(\frac{\partial X_b}{\partial I}\right)^2 \tag{42}$$

be nonpositive. This requirement is equivalent to

$$R_s \leqq R_{sc} , \tag{43}$$

where

$$R_{sc} = \frac{1}{2}\left[\sqrt{I\left(\frac{\partial R_b}{\partial I}\right)^2 + I\left(\frac{\partial X_b}{\partial I}\right)^2} - \frac{\partial(IR_b)}{\partial I} - R_b\right] . \tag{44}$$

In fact, one can verify that expression (42) vanishes for $R_s = R_{sc}$ and is negative for $R_s \geqq R_{sc}$. According to inequality (43), R_{sc} is the largest series resistance for which the network of Fig. 2 is potentially unstable.

Figure 4 shows several curves of R_{sc} versus $| v_m - \phi |$ calculated [using Eq. (44)] for different values of $I_o/C_o\omega_o\sqrt{\phi}$ and for $i_s[\exp(q\phi/kT)]/\omega_o C_o\sqrt{\phi} = 31.25$. The four points indicated on the curve relative to $I_o/C_o\omega_o\sqrt{\phi} = 0.1$ correspond to the four cases of Table I, as one can

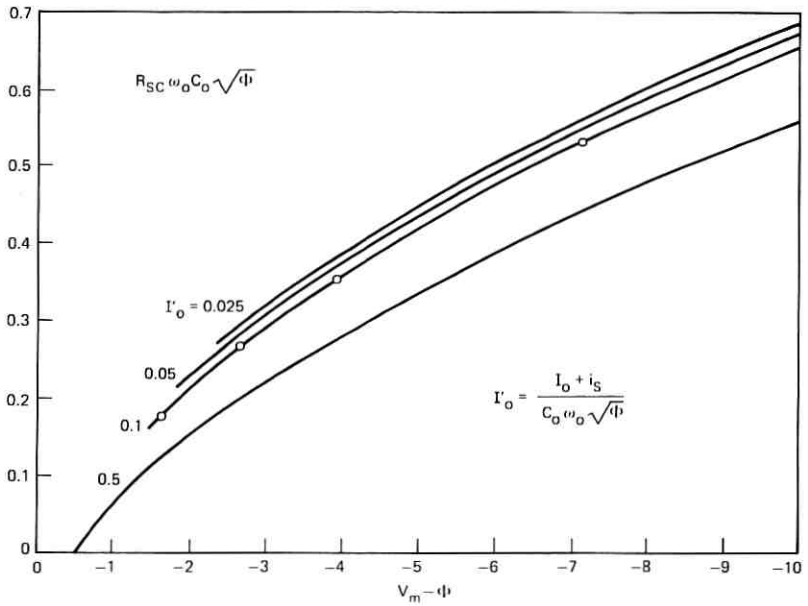

Fig. 4—Behavior of R_{sc} as a function of v_m for different values of I_o.

verify using eq. (44). We see from Fig. 4 that for given diode characteristics and for a given ω_o, the behavior of R_{SC} as a function of I_o and $| v_m - \phi |$ has the following characteristics. For a given I_o, R_{SC} increases monotonically with $| v_m - \phi |$. Since v_m cannot exceed the breakdown voltage, V_B, the largest value of R_{SC} for a given I_o occurs when $v_m = V_B$. For a given v_m, R_{SC} increases with decreasing I_o and approaches a finite limit for $I_o \to 0$. Figure 4 shows that, if $| V_m - \phi | > 2$ volts, R_{SC} is little affected by the value of I_o for $I_o/\omega_o C_o \sqrt{\phi} < 0.1$, approximately.

Now from the discussion of the limiting case of Section III, we know that, if conditions (18) and (33) are satisfied, then according to inequality (37)

$$\frac{R_{SC}\omega_o}{S_M - S_m} = \frac{1}{4} . \tag{45}$$

Condition (18) can be assumed to be fulfilled when $I_o/\omega_o C_o \sqrt{\phi}$ is sufficiently small, in which case the quantity $R_{SC}\omega_o/(S_M - S_m)$ is expected to approach 1/4 for large $| v_m - \phi |$. It is interesting to compare this asymptotic behavior of $R_{SC}\omega_o/(S_M - S_m)$ with the behavior corresponding to the curves of Fig. 4. Figure 5 shows $R_{SC}\omega_o/(S_M - S_m)$ plotted as a function of $| v_m - \phi |$ for the four cases corresponding to

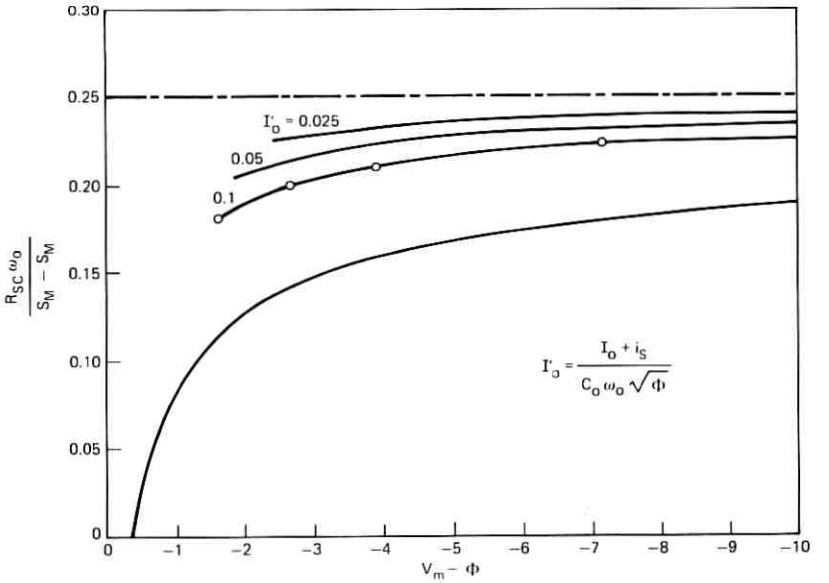

Fig. 5—Behavior of $R_{SC}\omega_o/(S_M - S_m)$.

the curves of Fig. 4. One sees that $R_{SC}\omega_o/(S_M - S_m)$ never exceeds 1/4, and that if $I_o/\omega_o C_o\sqrt{\phi}$ is sufficiently small, $R_{SC}\omega_o/(S_M - S_m)$ approaches 1/4 for large $|v_m - \phi|$, as expected. The four points indicated in Fig. 5 correspond to the cases of Table I.

So far, the quantity $i_S[\exp(q\phi/kT)]/\omega_o C_o\sqrt{\phi}$ has been assumed to be 31.25; Fig. 6 shows how $R_{SC}/\omega_o C_o\sqrt{\phi}$ is affected if $i_S[\exp(q\phi/kT)]/\omega_o C_o\sqrt{\phi}$ is changed from 31.25 to 1.08. The two curves of Fig. 6 have been calculated for $I_o/\omega_o C_o\sqrt{\phi} = 0.1$. It is evident that $R_{SC}/\omega_o C_o\sqrt{\phi}$ does not depend critically on the value of $i_S/\omega_o C_o\sqrt{\phi}$. The four points indicated in Fig. 6 correspond to the four cases of Table I.

4.1 *Region of Instability for Typical Diode Characteristics*

Typically, the breakdown voltage is sufficiently large so that one can assume $\phi - v_m > 2$ volts. For such values of $\phi - v_m$ one can verify from Fig. 4 that $R_{SC}\omega_o C_o\sqrt{\phi}/\sqrt{\phi - v_m} > 0.16$ if $I_o/\omega_o C_o\sqrt{\phi} < 0.5$. Thus, one can assume for typical diodes

$$0.16 < \frac{R_{SC}\omega_o C_o\sqrt{\phi}}{\sqrt{\phi - v_m}} < 0.25. \tag{46}$$

In the introduction, the range of pump frequencies associated with potentially unstable behavior was expressed in terms of the parameter η. A comparison of inequalities (1) and (43) shows that this parameter is

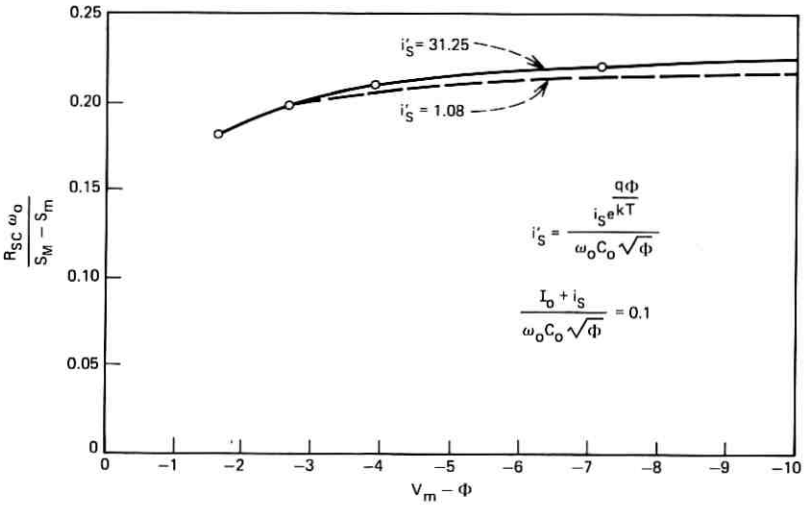

Fig. 6—Effect of i_S.

related to R_{SC} as follows:

$$\eta = \frac{\omega_c}{\omega_o}\frac{R_S}{R_{SC}} = \frac{S_M}{\omega_o R_{SC}} = \frac{\sqrt{\phi - v_m}}{R_{SC}\omega_o C_o \sqrt{\phi}}. \tag{47}$$

Thus, according to inequality (46), η typically lies between 4 and 6.25, as was stated in the introduction.

V. EFFECT OF R_S ON NOISE PERFORMANCE AT HIGH GAIN

The limit imposed by R_S on the optimum noise performance obtainable at high gain is now derived, when the diode is used as a down-converter and is terminated with equal impedances at $\omega_o \pm p$. The effect of R_S on noise can be separated into two parts: one is the effect at $\omega_o \pm p$, and the other at the output frequency p. We will see that if the diode is operated at high gain, the effect at $\omega_o \pm p$ is minimized when the diode is terminated with a very high impedance at $\omega = p$. Under this condition the effect of R_S at $\omega = p$ vanishes and can therefore be ignored.

Consider the effect of R_S at $\omega_o \pm p$. If R_a denotes the real part of the equal terminations at $\omega_o + p$ and $\omega_o - p$, the real part of the total impedance terminating the barrier at $\omega_o \pm p$ is $R_{at} = R_a + R_S$. Now inequality (43) implies that high gain is possible only if $R_{at} \leqq R_{SC}$, that is, only if

$$R_a \leqq R_{SC} - R_S . \tag{48}$$

Furthermore, when the resistance R_{at} equals R_{SC} (i.e., when $R_a = R_{SC} - R_S$), one can show that high gain requires a very high output impedance in all cases of Section IV. (This is a direct consequence of the fact that, when the resistance seen by the barrier at ω_o equals R_{SC}, instability may arise only if I_o = constant; that is, only if the diode is biased by a dc supply with infinite internal impedance.)

Now the ratio of the thermal noise power available (at $\omega_o \pm p$) at the barrier to that available at the terminals of the diode is R_{at}/R_a. This represents the impairment caused by the presence of R_S at $\omega_o \pm p$ on the noise performance of the diode as a down-converter. According to inequality (48), this impairment is minimized when R_a equals $R_{SC} - R_S$, in which case a very high termination is required at $\omega = p$. Thus, since the effect of R_S at $\omega = p$ vanishes, and at $\omega = \omega_o \pm p$ is given by the ratio $R_{SC}/(R_{SC} - R_S)$, the noise figure can be written

$$F_m = \frac{R_{SC}}{R_{SC} - R_S} F, \tag{49}$$

where F is the noise figure obtainable at high gain in the ideal case $R_S = 0$, when the termination at $\omega = p$ is a very high impedance. Using eq. (47) this relation can be rewritten

$$F_m = \left(1 - \eta \frac{\omega_o}{\omega_c}\right)^{-1} F \tag{50}$$

from which one obtains inequality (3).

VI. EXPERIMENTAL RESULTS*

Experimental data have been obtained at a pump frequency of 1.25 GHz using a GaAs Schottky barrier diode,† having $R_S \cong 4$ ohms, $C_o \cong 1.2$ pF, and $V_B \cong 10$ volts. This diode was mounted in a circuit designed to produce at the barrier very high terminating impedances at $2\omega_o$, $3\omega_o$, and $4\omega_o$. The structure is shown in Fig. 7 with the cover plate removed; the diode is inserted between two resonators both of which operate in the TEM mode. One resonator is connected to a 50-ohm coaxial line and consists of a main line with two series-resonant circuits connected in shunt. The purpose of the two series-resonant circuits is to short out transmission at $2\omega_o$ and $4\omega_o$ between the 50-ohm coaxial line and the diode. Each circuit consists of a line element connected to the main line at one end, with a lumped capacitance between the other end of each line element and ground.

The main line has the diode chuck soldered at one end; the other end is open-circuited. The electrical distance between the open-circuited end and the connection point at which the coaxial line is attached is a quarter of a wavelength at $3\omega_o$. The connection point of the coaxial line is therefore also short-circuited at $3\omega_o$. The distance from this connection point to the diode and the dimensions of the other resonator were chosen so as to open-circuit the barrier of the diode at $2\omega_o$, $3\omega_o$, and $4\omega_o$, using the following procedure. Initial dimensions for the two resonators were obtained experimentally by adjusting one resonator at the time (the other resonator and the diode being removed). The L-shaped resonator was adjusted so as to obtain two resonances at $2\omega_o$ and $4\omega_o$. The other resonator was adjusted (with the 50-ohm line terminated in 50 ohms) for a resonance at $3\omega_o$. Then an empty package identical to that of the diode used in this experiment was mounted between the two resonators as shown in Fig. 7, and the resonant frequencies of this circuit were measured by coupling the circuit

* The experiment described in this section was carried out by S. Michael of Bell Laboratories.

† Supplied by J. C. Irvin of Bell Laboratories.

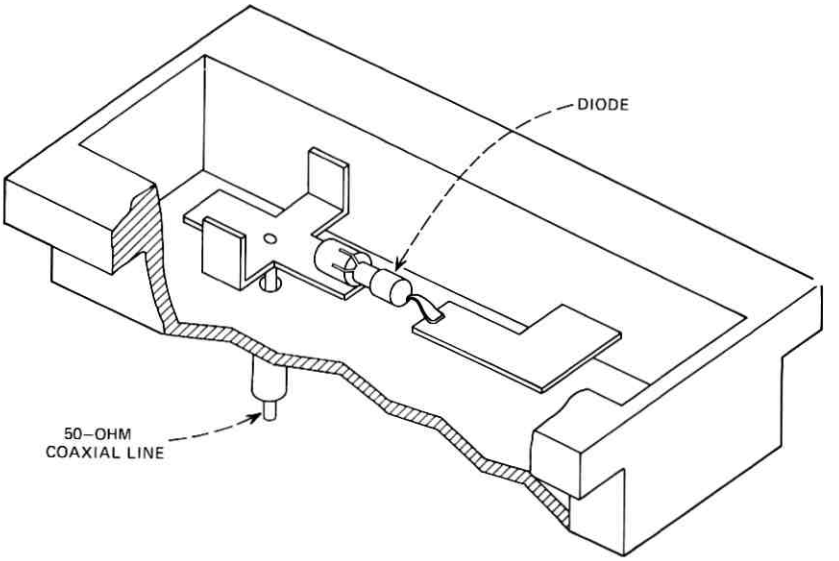

Fig. 7—Circuit used to open-circuit the barrier of the diode at $2\omega_o$, $3\omega_o$ and $4\omega_o$.

to a load and a generator by means of two loosely coupled capacitive probes. The resonant frequencies were found to be appreciably different from $2\omega_o$, $3\omega_o$, and $4\omega_o$ as expected, because of the coupling between the two resonators resulting from the case capacitance of the package. The dimensions of the two resonators were then adjusted to give the desired resonances at $2\omega_o$, $3\omega_o$, and $4\omega_o$ and the empty package was finally replaced with the actual diode.

In order to separate the pump frequency ω_o from dc, the 50-ohm coaxial line of the circuit of Fig. 7 was connected to one of the three arms of a monitor tee consisting of a main line shunted by an auxiliary line. The two lines are provided with a capacitor and an inductor connected in series to their central conductors to block signals in the vicinity of dc and ω_o, respectively. The other two arms of the monitor tee are connected to an output matching network and a tuner, as shown schematically in Fig. 8.

Figure 9 shows a block diagram of the apparatus used to measure the noise characteristics of the mixer of Fig. 8. The noise source is the AIL type 70 Hot-Cold Body Standard Noise Generator [consisting of two terminations, one immersed in liquid nitrogen (77.3°K) and the other mounted in a temperature-controlled oven (373.2°K)]. The pump is connected to a narrowband filter. This filter consists of four identical

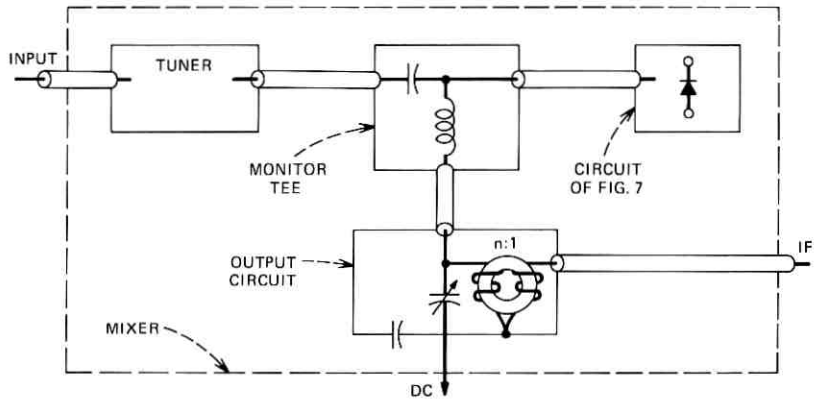

Fig. 8—Mixer.

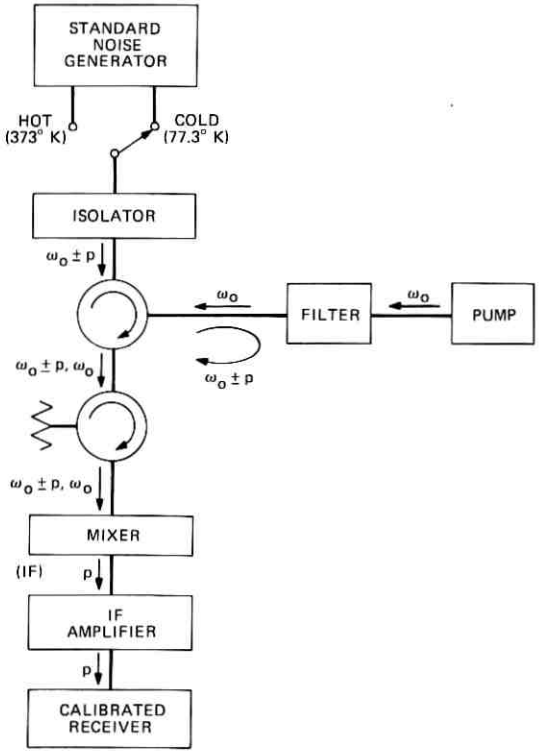

Fig. 9—Block diagram of apparatus used to test the mixer of Fig. 8; ω_o = 1.25 GHz; p = 2 MHz.

cavity resonators tuned at ω_o and separated by transmission lines $\lambda/4$ long at ω_o ; the attenuation at ω_o is approximately 10 dB and, at $\omega_o \pm p$ for $p \geqq 2$ MHz, is greater than 55 dB; its fractional bandwidth is approximately 0.07 percent. The noise source is buffered by an isolator providing more than 22 dB isolation at $\omega_o \pm p$. The noise components originating from the source at $\omega_o \pm p$ ($p \geqq 2$ MHz) enter the first circulator and are directed into the pump filter which reflects them back to the circulator where they are directed into a second circulator. The second circulator, which provides 22 dB of isolation at $\omega_o \pm p$, is followed by the mixer, the circuit of which is shown in Figs. 8 and 9. The noise power entering the mixer at $\omega_o \pm p$ is converted to the output (IF) frequency $p \cong 2$ MHz. The converted power is then amplified and finally measured with a narrowband receiver. The IF amplifier system consists of a calibrated variable attenuator followed by an amplifier, the noise figure of which was optimized at 0.19 dB for $p = 2$ MHz. The purpose of the variable attenuator is to vary the noise figure F_i of the IF amplifier.

The double-sideband noise figure of the mixer-amplifier combination is given by the familiar relation

$$F_r = F_m + \frac{F_i - 1}{G_m}, \tag{51}$$

where F_m is the double-sideband noise figure of the mixer* and G_m is its gain. Note that G_m is not the conversion gain from $\omega_o + p$ to p, or from $\omega_o - p$ to p, but is the *sum* of the two. That is, if $G_{\alpha\beta}$ and $G_{\gamma\beta}$ denote these two gains, $G_m = G_{\alpha\beta} + G_{\gamma\beta}$ (thus $G_m \cong 2G_{\alpha\beta}$ because $G_{\alpha\beta} \cong G_{\gamma\beta}$). Measurement of F_r consists essentially of determining two quantities: (*i*) the ratio $Y = P_2/P_1$, where P_2 and P_1 are the power outputs of the IF amplifier corresponding to the two temperatures $T_2 = 373.2°\text{K}$ and $T_1 = 77.3°\text{K}$ and (*ii*) the insertion loss ℓ of the circuit connected between the source and the mixer; this loss is less than 1 dB and can be measured very accurately. The noise figure is given by the well-known formula $F_r = [1 + (T_2 - T_1 Y)/290(Y - 1)]\ell$. The accuracy of measurement of F_r is limited primarily by the accuracy of the two temperatures T_2 and T_1 ; the estimated error for F_r is less than 0.1 dB.

According to eq. (51), F_m and G_m can be determined indirectly by measuring the effects of F_r of varying the IF noise figure F_i . Figure 10 shows a plot of F_r versus F_i which was obtained after adjusting the

* F_m is the ratio of the total noise power output of the mixer to that portion of this power originating from the terminations of the mixer at $\omega_o + p$ and $\omega_o - p$, assuming these terminations are at 290°K.

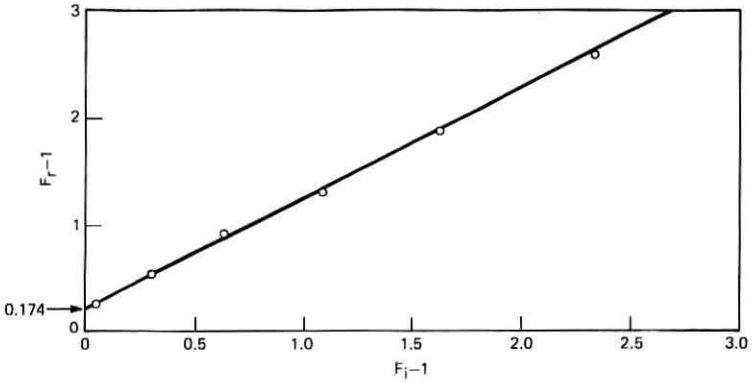

Fig. 10—Behavior of F_r versus F_i; mixer optimized for $F_i = 0.19$ dB.

mixer for minimum F_r under the particular condition $F_i = 0.19$ dB. According to eq. (51) the slope of this curve is the mixer conversion loss $L_m = 1/G_m$ and the point corresponding to $F_i = 1$ is F_m. This curve tells us that for a gain of 0.948 the lowest noise figure obtainable from the mixer is 1.174.

Figure 11 shows a plot of the minimum noise figure of the mixer versus its conversion loss. From this plot we can derive the lowest F_r obtainable for a given F_i as follows. From eq. (51), after replacing $1/G_m$ with L_m and differentiating, we get

$$\frac{dF_r}{dL_m} = \frac{dF_m}{dL_m} + (F_i - 1) \cdot \tag{52}$$

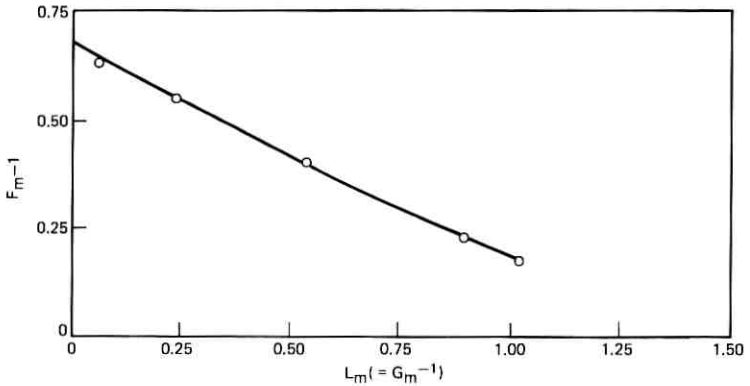

Fig. 11—Behavior of F_m versus $L_m = G_m^{-1}$.

Because the curve of Fig. 11 has $d^2F_m/d^2L_m > 0$, it follows that a point on the curve minimizes F_r if it satisfies $dF_r/dL_m = 0$ which, from eq. (52), results in

$$F_i = 1 - \frac{dF_m}{dL_m} . \tag{53}$$

Using this equation, we obtain from Fig. 11 the curve of Fig. 12, which shows the relation between F_i and the optimum value of L_m. One sees that it is desirable to have $L_m < 1$ (i.e., gain greater than unity) when

$$F_i > 1.37 \ (\sim 1.4 \text{ dB}) . \tag{54}$$

VII. CONCLUSIONS

It has been shown that a Schottky barrier diode is capable of arbitrarily high conversion gain as a mixer provided $\omega_o < \omega_c/\eta$, where η is a parameter typically less than 6.25, but always greater than 4. If $\omega_o \ll \omega_c/\eta$, then according to inequality (3), the ultimate double-sideband noise figure at high gain should be very close to unity. A fortiori, the ultimate noise figure at low gain should be excellent, if $\omega_o \ll \omega_c/\eta$. These conclusions are corroborated by the experimental results. Although the experimental result obtained is very good (F_m, 0.7 to 2.3 dB), it does not achieve the theoretical limit, nor was it expected to; practical limitations such as input circuit losses, estimated to exceed 0.4 dB for $G_m \gg 1$, confine the experimental noise figure for $G_m \gg 1$ to values appreciably higher than the theoretical limit.

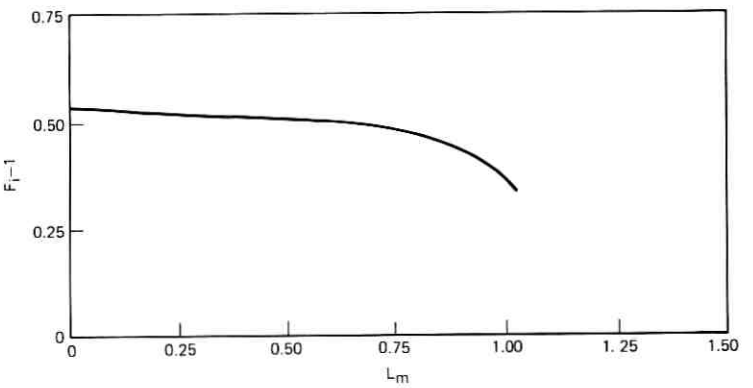

Fig. 12—Relation between F_i and the optimum value of L_m.

VIII. ACKNOWLEDGMENTS

The author is indebted to D. C. Hogg for his help and guidance in the preparation of this paper. Thanks are also expressed to C. L. Ruthroff for a number of useful discussions during the course of the work.

APPENDIX A

In this appendix, we analyze the condition I_o = constant for the case of $|(qv_m/kT)| \gg 1$ of Section III. The average current through the barrier resistance is

$$I_o = \langle i_s e^{(q/kT)v_b(t)} \rangle_{\text{ave}} - i_s . \tag{55}$$

Thus, if a perturbation δI is applied to the diode current at ω_o, while I_o is held constant, from eq. (55) the resulting perturbation of $v_b(t)$ must satisfy the condition

$$\langle e^{+(q/kT)v_b(t)} \, \delta v_b(t) \rangle_{\text{ave}} = 0. \tag{56}$$

According to eq. (20) this condition can be rewritten in the form

$$\langle e^{-BS^2(t)} S(t) \, \delta S(t) \rangle_{\text{ave}} = 0, \tag{57}$$

where

$$B = \frac{q}{kT} (C_o \sqrt{\phi})^2. \tag{58}$$

Note that $S(t)$ is completely specified by its maximum and minimum value S_M and S_m. We will presently show that if for a given value of the ratio

$$r = \frac{S_m}{S_M} \tag{59}$$

we let $S_M \to \infty$ then the time function $\exp [-BS^2(t)]$ over the interval $-T/2 \leqq t \leqq T/2$ approaches an impulse located at $t = 0$. Thus, if A denotes the area of this impulse, we will show that for a given value of r we can write:

$$\exp [-BS^2(t)] \to A u_o(t) \qquad (|t| \leqq T_o/2) \tag{60}$$

for $S_M \to \infty$, where $u_o(t)$ denotes the unit impulse. From this result and the fact that $S(0) = S_m$ we find that for $S_M \to \infty$ Eq. (56) becomes

$$S(0) \, \delta S(0) = S_m \, \delta S_m = 0. \tag{61}$$

We can thus say that for $S_M \to \infty$, condition (14) is equivalent to condition (34). We now show that relation (60) is true for $S_M \to \infty$.

Let ϵ be an arbitrarily small positive quantity. We have to show that for a given value of r

$$\lim_{S_M \to \infty} \frac{\int_{\epsilon}^{T/2} \exp\left[-BS^2(t)\right] dt}{\int_{0}^{T/2} \exp\left[-BS^2(t)\right] dt} = 0 \cdot \tag{62}$$

First, note that $S(t)$ can be written

$$S(t) = S_M\left[\frac{1+r}{2} - \frac{1-r}{2}\cos\omega_o t\right] \cdot \tag{63}$$

Using this expression one can verify that $\exp\left[-BS^2(t)\right] < \exp\left[-BS^2(\epsilon)\right]$ for $\epsilon < t \leqq T/2$. Thus,

$$\int_{\epsilon}^{T/2} \exp\left[-BS^2(t)\right] dt < \frac{T}{2} \exp\left[-BS^2(\epsilon)\right] . \tag{64}$$

Furthermore, it can be shown that if $S_M B$ is sufficiently large then

$$\int_{0}^{T/2} \exp\left[-BS^2(t)\right] dt > \frac{T}{2} \frac{e^{-r^2 S_M^2 B}}{\sqrt{2\pi r(1-r)BS_M^2}} \cdot \tag{65}$$

In fact, since $\cos\omega_o t > 1 - (\omega_o t)^2/2$, from eq. (63)

$$S^2(t) < S_M^2\left[r + \frac{(1-r)}{4}(\omega_o t)^2\right]^2$$

$$= S_M^2\left[r^2 + \frac{(1-r)r}{2}(\omega_o t)^2 + \frac{(1-r)^2}{16}(\omega_o t)^4\right] \cdot$$

This inequality allows one to write

$$\int_{0}^{T/2} \exp\left[-BS^2(t)\right] dt$$

$$> \frac{T}{2\pi} \frac{e^{-r^2 S_M^2 B}}{\sqrt{r(1-r)BS_M^2/2}} \int_{0}^{\pi\sqrt{r(1-r)BS_M^2/2}} \exp\left[-u^2 - \gamma u^4\right] du,$$

where $\gamma = (4BS_M^2 r^2)^{-1}$. If BS_M is sufficiently large, so that $\gamma \ll 1$, then from this inequality we obtain (65).

From inequalities (64) and (65) we obtain

$$\frac{\int_{\epsilon}^{T/2} \exp\left[-BS^2(t)\right] dt}{\int_{0}^{T/2} \exp\left[-BS^2(t)\right] dt} < \sqrt{2\pi BS_M^2(1-r)r}\, e^{-B\left[S^2(\epsilon) - r^2 S_M^2\right]} . \tag{66}$$

According to Eq. (63) we can write

$$S^2(\epsilon) - r^2 S_M^2 = S_M^2 \mathcal{E}_r(\epsilon), \tag{67}$$

where $\mathcal{E}_r(\epsilon)$ is a positive quantity which depends upon r and ϵ, but is independent of S_M. From eq. (67) and inequality (66) we therefore conclude that eq. (62) is true. Note that, since ϵ is a small quantity, from eqs. (63) and (67) we can write

$$\mathcal{E}_r(\epsilon) \cong \frac{(1-r)r\epsilon^2}{2}.$$

Thus,

$$\frac{\int_\epsilon^{T/2} \exp\left[-BS^2(t)\right] dt}{\int_0^{T/2} \exp\left[-BS^2(t)\right] dt} \ll 1$$

provided

$$BS_M^2 \gg \frac{2}{(1-r)r\epsilon^2}.$$

APPENDIX B

Figure 13 shows a network consisting of a variable elastance, a variable resistance, and a constant resistance R_s'. Assume that the current i_R' through the variable resistance is related to the voltage v_b' as follows:

$$i_R' = i_S' e^{qv_b'/kT} \tag{68}$$

and that the variation of the elastance S' is characterized by the relation

$$S' = \sqrt{-v_b'}\,. \tag{69}$$

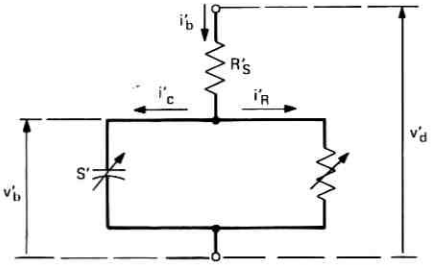

Fig. 13—Network representing a Schottky barrier diode.

Let the terminal current i_b' of this network be a periodic function of τ with period 2π, so that

$$i_b' = i_b'(\tau) = i_b'(\tau + 2\pi) . \tag{70}$$

Then, i_b' and v_b' are related through the differential equation

$$\frac{1}{\sqrt{-v_b'}} \frac{dv_b'}{d\tau} + i_s' e^{(q/kT)v_b'} - i_b' = 0. \tag{71}$$

Furthermore, one sees from Fig. 13 that the terminal voltage v_d' can be written

$$v_d' = i_b' R_s' + v_b' . \tag{72}$$

A Schottky barrier diode can always be represented by such a network. In fact, if we assume

$$\begin{aligned} \tau &= \omega_o t \\ v_b' &= v_b - \phi, \qquad v_d' = v_d - \phi + i_s R_s \\ i_s' &= \frac{i_s e^{(q\phi/kT)}}{C_o \omega_o \sqrt{\phi}} \\ i_b' &= \frac{i_b}{C_o \omega_o \sqrt{\phi}} + \frac{i_s}{C_o \omega_o \sqrt{\phi}} \\ R_s' &= R_s \omega_o C_o \sqrt{\phi} \end{aligned} \tag{73}$$

and substitute these relations in eqs. (71) and (72), we obtain eq. (39) and the relation $v_d = i_b R_s + v_b$. Thus, the two networks of Figs. 11 and 13 are equivalent.

Analysis

Let us now assume for i_b' the form

$$i_b' = i_b'(\tau) = \begin{cases} C, & \text{for} \quad \tau \leqq 0 \\ I_o' + 2I' \cos \tau, & \text{for} \quad \tau > 0 \end{cases} \tag{74}$$

where C is a constant. Since then $i_b' = i_R' = C$ for $\tau \leqq 0$, from eq. (68):

$$v_b' = v_b'(\tau) = \frac{kT}{q} (\ln C - \ln i_s') \quad \text{for} \quad \tau \leqq 0. \tag{75}$$

Since for $\tau \geqq 0$, $i_b'(\tau)$ is periodic, it is reasonable to expect that for τ sufficiently large, $v_b'(\tau)$ will also be periodic. Let us suppose there is a

positive integer N such that one can write to a good approximation

$$v_b'(2\pi N + 2\pi) \cong v_b'(2\pi N), \tag{76}$$

so that $v_b'(\tau)$ can be assumed to be periodic for $\tau \geqq 2\pi N$,

$$v_b'(\tau + 2\pi) \cong v_b'(\tau), \quad \text{for} \quad \tau \geqq 2\pi N. \tag{77}$$

Then we can write

$$v_b'(\tau) \cong v_{bo}' + 2(\mathrm{Re})(v_{b1}' e^{i\tau} + \cdots) \quad \text{for} \quad \tau \geqq 2\pi N, \tag{78}$$

where the dots indicate the components of order 2, 3, etc., and

$$V_{bo}' = \frac{1}{2\pi} \int_{2\pi N}^{2\pi N + 2\pi} v_b'(\tau)\, d\tau \tag{79}$$

and

$$V_{b1}' = \frac{1}{2\pi} \int_{2\pi N}^{2\pi N + 2\pi} v_b'(\tau) e^{-i\tau}\, dt\,. \tag{80}$$

Thus, if we can determine the behavior of $v_b'(\tau)$ over the interval $2\pi N \leqq \tau \leqq 2\pi(N + 1)$, where N is the smallest positive integer for which condition (76) can be assumed to be fulfilled, then the coefficients V_{bo}' and V_{b1}' can be readily determined by these two relations.

For given values of C, I_o' and I', an approximate solution to eq. (71) can be obtained using the Euler method. This method requires that the continuous variable τ be replaced by the discrete variable $n\tau_o$ ($n = 0$, ± 1, ± 2, etc.). Since in eq. (74) for $\tau \geqq 0$ $i_b'(\tau)$ has period 2π, it is convenient to assume for the step size τ_o an exact submultiple of 2π,

$$\tau_o = \frac{2\pi}{P} \tag{81}$$

where P is a positive integer. In the Euler method of solution, eq. (71) is replaced by the difference equation

$$v_{n+1}' = v_n' + D_n \tau_o\,, \tag{82}$$

where v_n' is the approximation to v_b' for $\tau = n\tau_o$, and D_n is the approximation to the derivative $dv_b'/d\tau$ for $\tau = n\tau_o$. From eq. (71)

$$D_n = \sqrt{-v_n'}\,(i_n' - i_s' e^{(q/kT)v_n'}) \tag{83}$$

where

$$i_n' = \begin{cases} I_o' + 2I' \cos(n\tau_o), & \text{for} \quad n > 0 \\ C, & \text{for} \quad n = 0 \end{cases}. \tag{84}$$

Equations (82) to (84) show that the values of v_o , v_1 , v_2 , etc. can be calculated sequentially, starting with

$$D_o = 0, \qquad v_o' = \frac{kT}{q} \ln \left(\frac{i_o'}{i_s'} \right), \tag{85}$$

then calculating v_1' and D_1 , and so on. Let N be the smallest integer for which the condition

$$v_{P(N+1)}' = v_{PN}' \tag{86}$$

is satisfied. Then, according to eqs. (79) and (80) the desired approximations to V_o' and V' are

$$\begin{aligned} V_o' &= \frac{1}{P} \sum_{K=0}^{P-1} v_{K+PN}' \\ V_o' &= \frac{1}{P} \sum_{K=0}^{P-1} v_{K+PN}' e^{-i\tau_{K+PN}} . \end{aligned} \tag{87}$$

Table I* and Figs. 4 to 6 have been calculated choosing as initial condition

$$i_0' = C = I_0' + 2I'. \tag{88}$$

The value of N depends on the value of I_0' . One can show that $N \to \infty$ for $I_0' \to 0$. Thus, the above method is not suitable if I_0' is too small. However, for the values of I_0' that are of practical interest, N is typically a small integer; for instance, $N \leqq 4$ in all the cases in Table I.

REFERENCES

1. Watson, H. A., *Microwave Semiconductor Devices and Their Circuit Applications*, New York: McGraw-Hill, 1969.
2. Abele, T. A., Alberts, A. J., Ren, C. L., and Tuchen, G. A., "Schottky Barrier Receiver Modulator," B.S.T.J., *47*, No. 7 (September 1968), pp. 1257–1287.
3. Elder, H. E., et al., "Active Solid-State Devices," B.S.T.J., *47*, No. 7 (September 1968), pp. 1323–1375.
4. Osborne, T. L., Kibler, L. U., Snell, W. W., "Low Noise Receiving Down-Converter," B.S.T.J., *48*, No. 6 (July–August 1969), pp. 1651–1663.
5. Dragone, C., "Amplitude and Phase Modulation in Resistive Diode Mixers," B.S.T.J., *48*, No. 6 (July–August 1969), pp. 1967–1997.
6. Dragone, C., "Conditions of High Gain in Mixers and their Relation to the Jump Phenomenon," B.S.T.J., this issue, pp. 2139–2167.
7. Torrey, H. C., and Whitmer, C. A., *Crystal Rectifiers*, *15*, Rad. Lab. Series, New York: McGraw-Hill, 1948.

* Note that according to eqs. (73) the various expressions of Table I have the following significance: $(I_o + i_s)/C_o\omega_o\sqrt{\phi} = I_o'$; $I/C_o\omega_o\sqrt{\phi} = I'$; $V_{bo} - \phi = V_{bo}'$; $V_m - \phi$ and $V_M - \phi$ are the minimum and maximum value of $v_b(\tau)$; $Z_b\omega_oC_o\sqrt{\phi} = Z_b' = V_{b1}'/I'$; etc.

8. Edwards, C. F., "Frequency Conversion by Means of a Nonlinear Admittance," B.S.T.J., *35*, No. 6 (November 1956,) pp. . 1403–1416.
9. MacPherson, A. C., "An Analysis of the Diode Mixer Consisting of Nonlinear Capacitance and Conductance and Ohmic Spreading Resistance," *IRE Trans. MTT, MTT-5* (January 1957), pp. 43–51.
10. Engelbrecht, R. S., "Parametric Energy Conversion by Nonlinear Admittances," Proc. IRE, No. 50 (March 1962), pp. 312–320.
11. Becker, L., and Ernst, R. L., "Nonlinear-Admittance Mixers," RCA Rev., *25* (December 1964), pp. 662–691.
12. Liechti, C. A., "Down-Converters Using Schottky-Barrier Diodes," IEEE Trans. Electron Devices, *ED-17*, No. 11 (November 1970), pp. 975–983.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 10, December, 1972
Printed in U.S.A.

The Accuracy of Call-Congestion Measurements for Loss Systems with Renewal Input

By A. KUCZURA and S. R. NEAL

(Manuscript received June 23, 1972)

The concept of a generalized renewal process is used to derive an asymptotic approximation for the variance of the observed proportion of unsuccessful attempts on a trunk group during a given time-interval. Calls are assumed to arrive according to a general renewal process, and those which are blocked leave the system and do not return (loss system).

As an application of our result we examine the special case of an overflow input–an important example from telephone networks with alternate routing. Comparison of our results with values obtained from simulation indicates that the approximation is quite accurate for telephone traffic-engineering purposes.

I. INTRODUCTION

In a communication network, the proportion of unsuccessful attempts on a trunk group during a specified interval of time is called the measured call-congestion, and is used to estimate the single-hour blocking probability for many of the trunk groups in the Bell System network. In order to determine how many measurements should be taken to properly assess system performance, one needs to know the statistical accuracy of the estimated blocking probability.

In the context of telephone traffic-engineering, the measured call-congestion is an unbiased estimate of the blocking probability, and hence we use its variance as an indicator of the precision of the measurements. For loss systems with exponentially-distributed service times, the variance has previously been studied under the assumption that calls originate according to a Poisson process.[1] However, attempts on a trunk group are well approximated by a Poisson process only for those groups which do not serve overflow traffic from subtending groups, so

that earlier results do not cover intermediate high-usage and final groups.

Assuming that the arrival process is of the renewal type, we derive an asymptotic approximation for the variance of the measured call-congestion for loss systems with exponentially-distributed service times. Since a single stream of overflow traffic is a renewal process, our results provide an estimate of the accuracy of call-congestion measurements made on intermediate high-usage and final trunk-groups.*

In Section II we describe the mathematical model used to solve our problem. The asymptotic approximation is derived in Section III. We also consider the "number of calls carried" as an unbiased estimate of the carried load and derive an asymptotic estimate of its variance. Section IV contains numerical results and Section V consists of a summary and our conclusions.

II. MATHEMATICAL MODEL

We consider a system of c servers serving customers, the arrival epochs of which constitute a renewal process. We assume that the interarrival times are independent and identically distributed according to the distribution function F, and that the service times are also independent and identically distributed according to an exponential distribution with unit mean. If all servers are occupied when a customer arrives, he leaves and has no further effect on the system. If an idle server is available when a customer arrives, service begins immediately.

Let $(0, t]$ denote a time interval of length t which commences at a point chosen at random on the time axis. Let $N(t)$ be the number of arrivals and $O(t)$ the corresponding number of blocked attempts in $(0, t]$. The ratio $O(t)/N(t)$ is the measured call-congestion. In Section III we show that the variance of $O(t)/N(t)$ can be approximated in terms of the covariance between $O(t)$ and $N(t)$ and of the individual first two moments of $O(t)$ and $N(t)$. We now describe the mathematical model used to obtain the required moments.

2.1 *A Multi-Dimensional Renewal Process*

Let t_n, $n = 0, 1, 2, \cdots$, be the instant of time at which the nth overflow occurs, $t_0 < 0 < t_1 < \cdots$, and set $X_n = t_n - t_{n-1}$. The interoverflow times X_n, $n = 1, 2, \cdots$, form a sequence of independent (because holding times are exponential) and identically distributed

* For engineering purposes, the total overflow traffic offered to such groups is adequately described by a single overflow process.[2]

random variables. Let K_n, $n = 1, 2, \cdots$, be the number of arrivals occurring in $(t_{n-1}, t_n]$ and define the row vector

$$\chi_n = (1, K_n), \qquad n = 1, 2, \cdots .$$

Since K_n, $n = 1, 2, \cdots$, is also a sequence of independent and identically distributed random variables, the components of the vector sequence χ_n, $n = 1, 2, \cdots$, are independent and identically distributed. Now, set

$$\eta(t) = \sum \chi_n ,$$

where the sum is taken over all n such that $0 < t_n \leqq t$. With these definitions it follows that for large t,

$$\eta(t) \approx (O(t), N(t))$$

and that (χ_n, X_n), $n = 1, 2, \cdots$ is a multi-dimensional renewal process.[3] Consequently, the results communicated by J. M. Hammersley[3] in the discussion of W. L. Smith's paper apply directly to our model. In particular, we shall use his eqs. (25) and (26) to compute the moments of $\eta(t)$.

We could also have used Smith's results on cumulative processes to obtain, in an indirect way, the covariance between $O(t)$ and $N(t)$. However, the concept of a cumulative process is not as naturally suited to our problem as is Hammersley's generalization of a renewal process.

Let $\mu_n(c) = E[X_1^n]$ be the nth moment of the interoverflow times from a group of c servers and

$$\nu_n = \int_0^\infty \xi^n \, dF(\xi).$$

For brevity we shall denote the arrival intensity ν_1^{-1} by λ. Equation (25) of Ref. 3 states that

$$E[O(t)] = \frac{t}{\mu_1(c)} \tag{1}$$

and

$$E[N(t)] = \frac{t}{\mu_1(c)} E[K_1].$$

But since $E[N(t)]$ is also equal to λt, we have (as is clear intuitively)

$$E[K_1] = \lambda \mu_1(c). \tag{2}$$

Now, using eq. (26) of Ref. 3 we have for large t

$$\text{Cov}\,[O(t), N(t)] \sim \frac{t}{\mu_1^2(c)}\{\lambda\mu_2(c) - E[K_1X_1]\}. \tag{3}$$

In this expression, as in others below, we omit terms which behave as $o(t)$ for large t. The results of Section IV indicate their contribution to be negligible.

Since the overflow epochs constitute a renewal process, and since we have a renewal input, we can also use eq. (26) of Ref. 3 to obtain

$$\text{Var}\,[O(t)] \sim \frac{t}{\mu_1^3(c)}[\mu_2(c) - \mu_1^2(c)] \tag{4}$$

and, with a change in definitions,

$$\text{Var}\,[N(t)] \sim \frac{t}{\nu_1^3}[\nu_2 - \nu_1^2] \tag{5}$$

for large values of t. Since ν_1 and ν_2 can be computed directly from F, we only need $\mu_1(c)$, $\mu_2(c)$ and $E[K_1X_1]$ in order to evaluate the asymptotic expressions (3), (4) and (5).

2.2 *The Joint Distribution of K_1 and X_1*

Let

$$g_c(k, t)\,dt = dP\{K_1 = k, X_1 \leqq t\},$$

where the differential on the right-hand side is to be taken with respect to the variable t. Using the same arguments as those presented in Ref. 4, pages 388–389, we obtain

$$g_c(k, t) = e^{-t}g_{c-1}(k, t) + \sum_{m=1}^{k-1}\int_0^t (1 - e^{-u})g_c(k - m, t - u)g_{c-1}(m, u)\,du. \tag{6}$$

If we define

$$\gamma_c(w, s) = \int_0^\infty \sum_{k=1}^\infty w^k g_c(k, t)e^{-st}\,dt, \tag{7}$$

then it follows from (6) that

$$\gamma_c(w, s) = \frac{\gamma_{c-1}(w, s + 1)}{1 - \gamma_{c-1}(w, s) + \gamma_{c-1}(w, s + 1)}. \tag{8}$$

This relation is identical to relation (7) of Ref. 4 derived for the case of

Poisson input, and (contrary to the author's comment) is also valid for any arrival process of the renewal type.

Relation (8) is of the same form as the recurrence relation for the Laplace–Stieltjes transform of the interoverflow distribution for loss systems with renewal input (see Ref. 5, page 37). Consequently, we can follow the outline of the analysis in Ref. 5 to obtain γ_c.

First, Riordan's results imply that $\gamma_c(w, s)$ can be written in the following form:

$$\gamma_c(w, s) = \frac{D_c(w, s)}{D_{c+1}(w, s)}, \tag{9}$$

where $D_0(w, s) = 1$, and, as can be seen by setting $c = 0$ in (7),

$$D_1(w, s) = \frac{1}{w\alpha(s)}, \tag{10}$$

where

$$\alpha(s) = \int_0^\infty e^{-s\xi}\, dF(\xi).$$

Furthermore, for $r \geqq 1$,

$$D_{r+1}(w, s) = D_r(w, s) + \left[\frac{1}{w\alpha(s)} - 1\right] D_r(w, s + 1). \tag{11}$$

Following Riordan, we define

$$\lambda_j = \lambda_j(w, s) = 1 - \frac{1}{w\alpha(s + j)}. \tag{12}$$

Now using (10) and (11) and mathematical induction (as noted by Riordan) one can show that

$$D_r(w, s) = 1 + \sum_{j=1}^{r} (-1)^j \binom{r}{j} \lambda_0\lambda_1 \cdots \lambda_{j-1}. \tag{13}$$

Finally, since

$$E[K_1^m X_1^n] = (-1)^n \frac{\partial^{m+n}}{\partial w^m\, \partial s^n} \gamma_c(w, s) \Big|_{\substack{w=1 \\ s=0}}$$

for $m = 0, 1$ and $n \geqq 0$, we can compute the required moments directly from (8).

First, differentiation of (8) with respect to s yields the following results:

$$\mu_1(c) = \nu_1 D_c$$

and

$$\mu_2(c) = \frac{\nu_2}{\nu_1}\mu_1(c) + 2\mu_1(c)\sum_{k=1}^{c}\mu_1(k) - 2\nu_1 D_c^{(01)}, \tag{14}$$

where

$$D_c = D_c(1, 1) = 1 + \sum_{j=1}^{c}(-1)^j\binom{c}{j}\Lambda_0\Lambda_1 \cdots \Lambda_{j-1} \tag{15}$$

is the reciprocal of the generalized Erlang B blocking,

$$\Lambda_k = 1 - \frac{1}{\alpha(k+1)},$$

and

$$D_c^{(01)} = \frac{\partial}{\partial s} D_c(w, s)\Big|_{\substack{w=1\\s=1}}.$$

Performing the last operation, we have

$$D_c^{(01)} = \sum_{j=1}^{c}(-1)^j\binom{c}{j}\Lambda_0\Lambda_1 \cdots \Lambda_{j-1}\left[\frac{\Omega_0'}{\Lambda_0} + \frac{\Omega_1'}{\Lambda_1} + \cdots + \frac{\Omega_{j-1}'}{\Lambda_{j-1}}\right] \tag{16}$$

where Ω_k' is the derivative of $\lambda_k(1, s)$ evaluated at $s = 1$, i.e.,

$$\Omega_k' = \frac{\alpha'(k+1)}{\alpha^2(k+1)}.$$

Similarly, the joint expectation of K_1 and X_1 is given by

$$E[K_1X_1] = \mu_1(c) + \frac{2}{\nu_1}\mu_1(c)\sum_{k=1}^{c}\mu_1(k) + \nu_1 D_c^{(10)} - D_c^{(01)}, \tag{17}$$

where

$$D_c^{(10)} = \frac{\partial}{\partial w} D_c(w, s)\Big|_{\substack{w=1\\s=1}}$$

is given by

$$D_c^{(10)} = \sum_{j=1}^{c}(-1)^j\binom{c}{j}\Lambda_0\Lambda_1 \cdots \Lambda_{j-1}\left[\frac{1}{\Lambda_0} + \frac{1}{\Lambda_1} + \cdots + \frac{1}{\Lambda_{j-1}} - j\right]$$

and $D_c^{(01)}$ is given by (16).

Before concluding this section, we show that the covariance function

(3) reduces to the expression given by Descloux[1] for the case of Poisson input. Substituting (14) and (17) into (3), we have

$$\text{Cov } [O(t), N(t)] \sim \frac{t}{\mu_1(c)} \left[\frac{\nu_2 - \nu_1^2}{\nu_1^2} - \frac{D_c^{(01)} + \nu_1 D_c^{(10)}}{\mu_1(c)} \right]. \tag{18}$$

Since $\mu_1^{-1}(c) = \lambda B(c, \lambda)$, where $B(c, \lambda) = D_c^{-1}$ is the generalized Erlang B blocking, we can write

$$\text{Cov } [O(t), N(t)] \sim \lambda t B(c, \lambda) \left\{ \frac{\nu_2 - \nu_1^2}{\nu_1^2} - B(c, \lambda)[D_c^{(10)} + \lambda D_c^{(01)}] \right\}.$$

When the input is Poisson, we have the following simplifications:

$$\frac{\nu_2 - \nu_1^2}{\nu_1^2} = 1,$$

$$\Lambda_k = -\frac{k+1}{\lambda},$$

$$\Omega_k' = -\frac{1}{\lambda},$$

$$D_c^{(10)} + \lambda D_c^{(01)} = -\sum_{i=1}^{c} j \binom{c}{j} j! \, \lambda^{-i}$$

$$= \lambda \frac{\partial}{\partial \lambda} E_{1,c}^{-1}(\lambda)$$

where $E_{1,c}(\lambda)$ is the Erlang B blocking probability

$$E_{1,c}(\lambda) = \frac{\lambda^c}{c! \sum_{i=0}^{c} \frac{\lambda^i}{j!}}.$$

With these simplifications the covariance function becomes

$$\text{Cov } [O(t), N(t)] = \lambda t \frac{\partial}{\partial \lambda} [\lambda E_{1,c}(\lambda)].$$

Substituting for the derivative (see Ref. 6, page 1)

$$\frac{\partial}{\partial \lambda} E_{1,c}(\lambda) = \frac{1}{\lambda} [c - \lambda + \lambda E_{1,c}(\lambda)] E_{1,c}(\lambda),$$

we obtain Descloux's result

$$\text{Cov } [O(t), N(t)] \sim \lambda t E_{1,c}(\lambda)[1 + c - \lambda + \lambda E_{1,c}(\lambda)].$$

III. CALL CONGESTION AND CARRIED LOAD VARIANCES

In Section II we derived asymptotic approximations for the variances of $N(t)$ and $O(t)$ and the covariance between the two. These expressions can now be combined to obtain an asymptotic approximation for the variance of $O(t)/N(t)$. Moreover, without additional effort we can also approximate the variance of

$$\mathfrak{L}(t) = \frac{N(t) - O(t)}{t}, \tag{19}$$

which is the number of calls carried per mean holding time, i.e., an estimate of carried load.

3.1 *Variance of Call Congestion*

From the theory of standard errors the variance of the call congestion is given approximately by[7]

$$\text{Var}\left[\frac{O(t)}{N(t)}\right] \approx \left\{\frac{\text{Var}\,[O(t)]}{E^2[O(t)]} + \frac{\text{Var}\,[N(t)]}{E^2[N(t)]} - \frac{2\,\text{Cov}\,[O(t), N(t)]}{E[O(t)]E[N(t)]}\right\} \frac{E^2[O(t)]}{E^2[N(t)]}. \tag{20}$$

The derivation of this expression is based on squaring the first-order terms of a Taylor series expansion of $O(t)/N(t)$ about the means of $O(t)$ and $N(t)$; the higher central moments of $O(t)$ and $N(t)$ are omitted. The accuracy of the approximation is discussed with the numerical results in Section IV. Using eqs. (1), (2), (4), (5), and (18) to substitute for the various quantities on the right-hand side of (20), we obtain

$$\text{Var}\left[\frac{O(t)}{N(t)}\right] \approx \frac{1}{t}\left\{\frac{\mu_2(c) - \mu_1^2(c)}{\mu_1(c)} - \frac{\nu_2 - \nu_1^2}{\nu_1} + \frac{2\nu_1}{\mu_1(c)}\left[D_c^{(01)} + \nu_1 D_c^{(10)}\right]\right\}\left(\frac{\nu_1}{\mu_1}\right)^2. \tag{21}$$

3.2 *Variance of Carried Load*

The expectation of (19) is

$$E[\mathfrak{L}(t)] = \lambda\left[1 - \frac{1}{\lambda\mu_1(c)}\right].$$

Since $1/[\lambda\mu_1(c)]$ is the blocking probability, we have

$$E[\mathfrak{L}(t)] = \lambda[1 - B(c, \lambda)]$$

which shows that (19) is an unbiased estimate of the carried load. The variance of this estimate is given by

$$\text{Var}\,[\mathfrak{L}(t)] = \frac{1}{t^2}\{\text{Var}\,[N(t)] + \text{Var}\,[O(t)] - 2\,\text{Cov}\,[O(t), N(t)]\}. \quad (22)$$

Substituting for the various terms on the right-hand side of (22), we obtain for large t

$$\text{Var}\,[\mathfrak{L}(t)] \sim \frac{1}{t}\left\{\frac{\mu_2(c) - \mu_1^2(c)}{\mu_1^3} + \frac{\nu_2 - \nu_1^2}{\nu_1^3}\left[1 - \frac{2\nu_1}{\mu_1(c)}\right] + \frac{2}{\mu_1^2(c)}[D_c^{(01)} + \nu_1 D_c^{(10)}]\right\}. \quad (23)$$

IV. NUMERICAL RESULTS

We are primarily interested in the accuracy of the measured call-congestion when the input is overflow traffic. Of course, one expects that the variance of the measured call-congestion will increase as the variance-to-mean ratio (peakedness) z of the offered traffic increases. We must also test the accuracy of the various analytical approximations. In particular, we must determine whether the standard measurement period of one hour (about 20 mean holding-times) is long enough for the asymptotic expressions (3) through (5) and the approximation (20) to be accurate enough for engineering purposes.

We computed the estimate (21) of call-congestion variance for trunk groups of 6, 10, 20, 30 and 40 trunks serving overflow traffic with various values of peakedness ranging from one to ten and a measurement interval of $t = 20$ mean holding-times. In each case, the offered load was varied over a range from 0.05 erlangs/trunk to 2 erlangs/trunk. The interarrival-time distribution of the originating traffic was obtained by using the Interrupted Poisson process[8] with a three-moment match.

To check the accuracy of the various approximations, estimates of the variance were obtained by simulation for each of the cases mentioned above. The results for the five different trunk groups were of the same general form as for the ten-trunk case shown in Figs. 1 and 2.

Figure 1 is a graph of the standard deviation of the measured call-congestion $\sigma_B = \{\text{Var}\,[O(t)/N(t)]\}^{\frac{1}{2}}$ vs the offered load $\alpha = \lambda = \nu_1^{-1}$ for several values of z and $c = 10$ trunks. Notice that for a fixed value

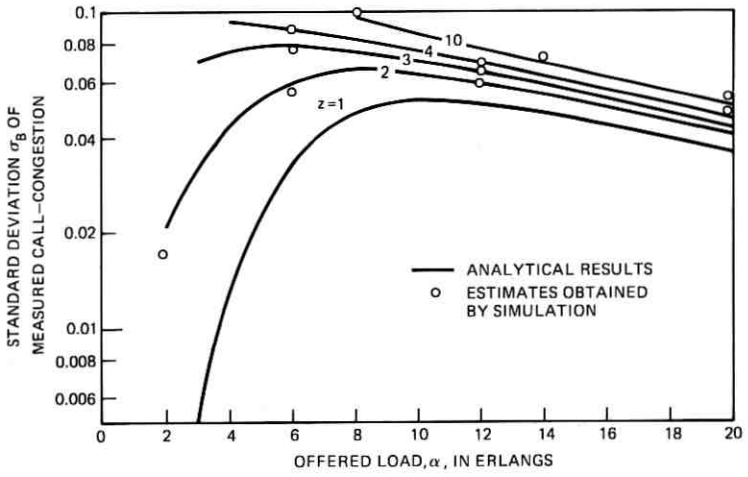

Fig. 1—Standard deviation of measured call-congestion vs offered load for $c = 10$ trunks and overflow input having peakedness z.

of α, σ_B increases as z increases (as was expected). The curves are terminated in the region of $z = \alpha$. In general, we found that our analytic results were in good agreement with the simulation for $\alpha > z$. This verifies the accuracy of our approximations when $\alpha > z$. However, a notable disparity occurred in several cases when $\alpha < z$. The latter inequality rarely arises in telephone traffic; but for other applications, where $\alpha < z$ might occur (e.g., data transmission), further work is required.

Figure 2 illustrates the behavior of the coefficient of variation σ_B/B of the measured call-congestion. For a large range of offered loads, the coefficient of variation decreases as z increases, indicating that the blocking probability increases faster as a function of z than does the standard deviation. Although the variance of the measured call-congestion decreases as α decreases for small α, the coefficient of variation (relative error) increases hyperbolically. Consequently, the relative accuracy of the measured call-congestion decreases as the blocking probability decreases, i.e., as the number $O(t)$ of observed overflows decreases.

Figure 3 displays the coefficient of variation as a function of $\alpha B = E[O(t)]/t$ for several trunk-group sizes. In each case we used both $z = 1$ and $z = 10$. The slope of the curves on the log paper is approximately $-1/2$. Hence, the coefficient of variation of the measured call-congestion is approximately inversely proportional to the square root of the number of blocked calls observed.

V. SUMMARY

For loss systems with renewal input and exponential holding times, we derived asymptotic approximations for the respective means and variances of $N(t)$, the number of arrivals, and $O(t)$, the number of overflows in the measurement period $(0, t]$. We also obtained an asymptotic estimate for the covariance between $O(t)$ and $N(t)$. Using these results, we obtained an estimate of the variance of the measured call-congestion $O(t)/N(t)$, as well as the variance of $\mathfrak{L}(t) = [N(t) - O(t)]/t$ which is an unbiased estimate of carried load, provided the mean holding time is known.

Our analytical approximation for the variance of $O(t)/N(t)$ was checked by simulation for systems serving overflow traffic. In those cases which were tested, the simulation results were in excellent agreement with the analytical results for the range of system parameters (roughly $z < \alpha$) which normally arise in telephone-engineering applica-

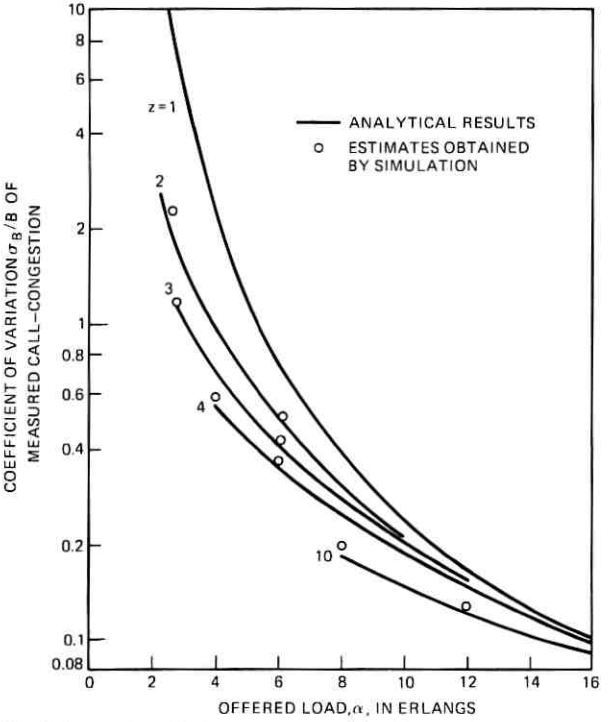

Fig. 2—Coefficient of variation of measured call-congestion vs offered load for $c = 10$ trunks and overflow input having peakedness z.

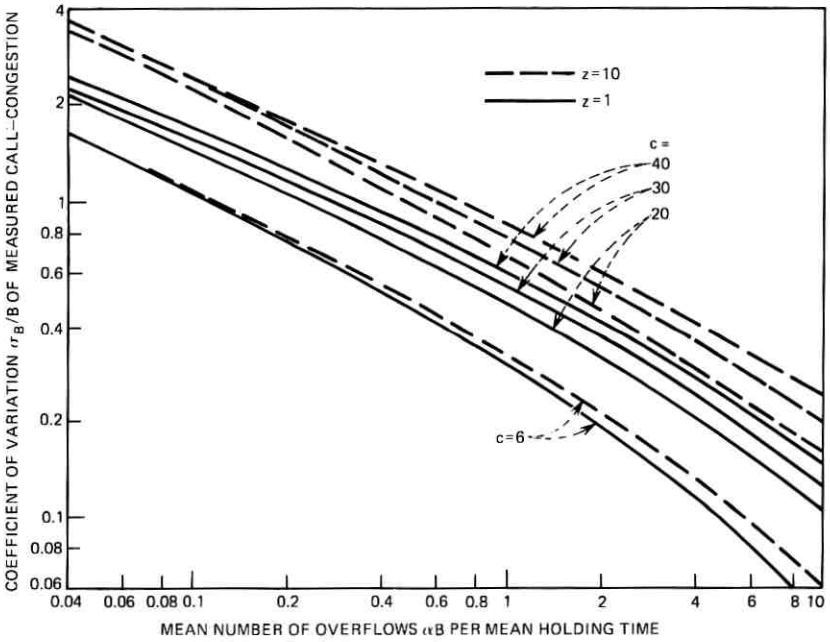

Fig. 3—Coefficient of variation of measured call-congestion vs overflow rate for c = 6, 20, 30, and 40 trunks, and input traffic with peakedness z = 1, 10.

tions. We also found empirically that the coefficient of variation of $O(t)/N(t)$ is approximately inversely proportional to the square root of the mean number of overflows in $(0, t]$.

REFERENCES

1. Descloux, A., "On the Accuracy of Loss Estimates," B.S.T.J., *44*, No. 6 (July–August 1965), pp. 1139–1164.
2. Wilkinson, R. I., "Theories of Toll Traffic Engineering in the U.S.A.," B.S.T.J., *35*, No. 2 (March 1956), pp. 421–514.
3. Hammersley, J. M., Discussion of paper by W. L. Smith, "Renewal Theory and Its Ramifications," J. Roy. Stat. Soc., Series B, *20*, No. 2, 1958, pp. 289–291.
4. Descloux, A., "On overflow Processes of Trunk Groups with Poisson Inputs and Exponential Service Times," B.S.T.J., *42*, No. 2 (March 1963), pp. 383–397.
5. Riordan, J., *Stochastic Service Systems*, New York: Wiley, 1962, pp. 36–40.
6. Nishimura, T., "The Derivatives of Erlang's B Formula," Elec. Commun. Lab. Tech. J., Tokyo, Extra Issue No. 22, 1966.
7. Kendall, M. G., and Stuart, A., *The Advanced Theory of Statistics*, *1*, New York: Hafner, 1958, pp. 231–232.
8. Kuczura, A., "The Interrupted Poisson Process as an Overflow Process," to appear in March 1973 issue of B.S.T.J.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 10, December, 1972
Printed in U.S.A.

A Cost Optimization Model for Seismic Design of Structures

By S. C. LIU and F. NEGHABAT

(Manuscript received July 17, 1972)

Considering the earthquake susceptibility of structures located in seismic regions, the question arises as to what level of protective measures should be provided in order to achieve a certain degree of reliability against possible damage. To address this question, engineering risk and optimal design of structures located in a seismic area are studied. The basic concept is to obtain a tradeoff between the cost of providing a protective measure and the expected cost of earthquake damage.

A simple mathematical approach is presented to determine the optimal earthquake intensity which the structure is designed to withstand. The objective is to minimize the total construction cost of the structure plus the expected cost of earthquake damage throughout the entire service life of the structure. For the case of deterministic structural resistance, and for structural response processes having Poisson (independent) crossings, an objective function is derived in terms of the building and earthquake variables. The optimal design intensity can then be determined by minimizing the objective function with respect to the intensity variable. The resulting equations are relatively simple and can be easily handled for numerical studies and sensitivity analysis. Generalizations of the results for nondeterministic structural resistance and for structural response processes different from those having Poisson crossings are also indicated.

As an illustration of the proposed approach, a hypothetical building with realistic seismicity and structural parameters is analyzed for its optimal design earthquake intensity. The construction, damage and total costs are obtained in terms of the intensity variable. The implications and sensitivity of the results are also discussed.

I. INTRODUCTION

Structures constructed in seismic regions are required to function properly in a forcing environment characterized by random earthquake

occurrences and intensities. The seismic environment including the expected earthquake-magnitude levels and the corresponding frequency of occurrence for different seismic-risk zones was described previously.[1] The study was based on a statistical analysis of nationwide seismic data and may be used as a guide for the development of seismic design requirements on a global basis. Under localized situations, however, the seismic requirement for structures that are expected to adequately withstand the earthquake environments should be based on cost-reliability studies. During an earthquake of given intensity, there exists a probability that the response of the structure is greater than its resistance capability and, therefore, a probability of damage to the structure. The cost associated with this probable damage may be referred to as the "earthquake risk cost". Increasing the design intensity of the structure reduces the probability of damage, but at an increased cost of construction. Therefore, an optimal design earthquake intensity can be determined by achieving an appropriate balance between the construction cost and the earthquake risk cost.

This paper presents a new analytical approach to the determination of the economically optimal earthquake intensity or other design variables for structures. The construction and earthquake risk costs are expressed in terms of design intensity and other parameters reflecting the earthquake and structural characteristics. Minimization of the total expected cost of the structure yields the optimum structural design intensity in terms of such parameters as estimated cost of earthquake damage, unit construction cost, expected earthquake duration, and statistics obtained from seismological data for the particular site.

II. ANALYSIS OF DESIGN INTENSITY MODEL

2.1 *Objective Function*

Consider a certain seismic region in which a structure is located. Let $K_c(i_o)$ and $K_d(i_o)$ represent the construction and the earthquake risk or damage costs of the structure respectively, both being functions of the design intensity i_o measured from I to XII on the Modified Mercalli scale. The function $K_c(i_o)$ may be regarded as a monotonically increasing function of i_o , while the function $K_d(i_o)$, as would be expected, is a monotonically decreasing function of i_o .

The optimum design intensity i_o^*, may be obtained as a trade-off between these two functions by minimizing the total cost

$$K(i_o) = K_c(i_o) + K_d(i_o) \tag{1}$$

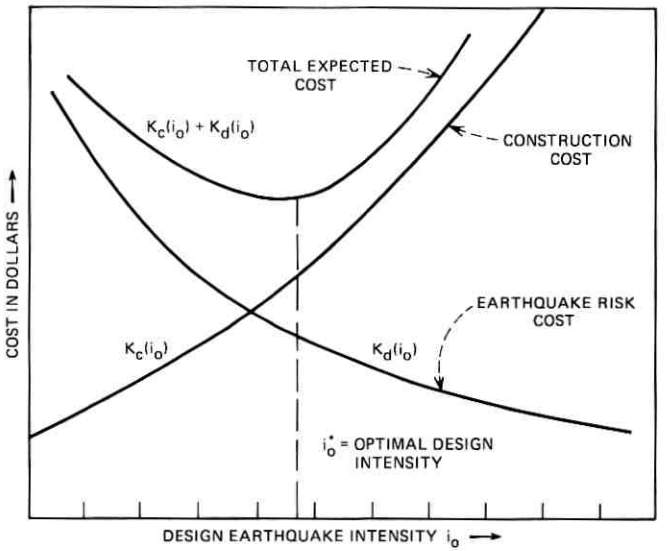

Fig. 1—Sketch showing cost of structure against design intensity of earthquake.

as shown in Fig. 1. The function $K(i_o)$ in eq. (1) is the objective function and i_o is the decision variable.

The construction cost, $K_c(i_o)$, can be written as

$$K_c(i_o) = A_f(f + C) \tag{2}$$

where A_f is the floor area of building, f is the building cost per unit floor area, and C is the cost of earthquake protection per unit floor area. The earthquake protection cost obviously increases with the protection level, such as i_o in this case. Therefore, $C = C(i_o)$ is a monotonically increasing function of i_o . It should be noted that the determination of the function $C(i_o)$ depends on the type of structure, method of design, and is greatly influenced by the designer's personal judgment and experience. A reasonable first approximation can be made assuming $C(i_o)$ is linear with a coefficient c. Under this assumption eq. (2) can be written as

$$K_c(i_o) = A_f(f + ci_o) \tag{3}$$

where $A_f f$ equals a fixed initial cost of building, and c is the earthquake resistance cost of the building per unit floor area per unit intensity.

To determine the function $K_d(i_o)$, analytical procedures can be effectively employed utilizing some basic knowledge about the structure and the random forcing environment. Let $N(t)$ be a random variable

representing the total number of earthquakes to occur in time t. Furthermore, for $k = 1, 2, \cdots, N(t)$, let p_k be the probability that the structure fails given, the kth earthquake occurs, and let e_k , an identically distributed random variable, be the associated loss. Therefore, the total cost of earthquake damage, Z, is

$$Z = \sum_{k=1}^{N(t)} e_k p_k \ . \tag{4}$$

Taking expectation (denoted by overbar) of both sides of eq. (4), one obtains

$$\bar{Z} = \overline{N(t)}\bar{e}\bar{p} \tag{5}$$

in which $\overline{N(t)}$ equals the expected number of earthquakes (of all intensities) in time t, $\bar{e}$ equals the expected value of the random earthquake loss, and $\bar{p}$ equals the mean failure probability of structure given an earthquake occurs. The present worth of the expected value of Z is the earthquake risk cost $K_d(i_o)$:

$$K_d(i_o) = \bar{Z}g(t) = \bar{N}(t)\bar{e}\bar{p}g(t) \tag{6}$$

where $g(t)$ is a discount factor.

The quantities on the right-hand side of eq. (6) will be discussed next in terms of the related design parameters which fall into two categories: the earthquake parameters and the building parameters. The earthquake parameters include the regional seismicity constants, the earthquake magnitude (m), intensity (i), duration (l_o), amplitude (a) of the ground motions, and the statistics of these quantities. The building parameters include the mass (ρ), stiffness (k), natural frequency (ω_o), damping (ξ_o), height (h), the resistance or strength (x) of the structure, etc.

2.2 *The Average Number of Earthquakes, $\bar{N}(t)$*

The quantity $\bar{N}(t)$ depends on and can be estimated from the regional seismicity. Earthquakes can be considered to be a series of events randomly distributed on a real line (representing time), and the sequence of original times $\{t_n\}$ forms a point process.[2] It is further assumed that the joint statistics of the respective number of shocks in any set of intervals are invariant under a translation of these intervals; this implies that $\{t_n\}$ is a stationary point process. The stationary point process generalizes certain aspects of renewal processes; in particular, the interval lengths $\tau_k = t_k - t_{k-1}$ between successive events need not be independently or identically distributed.

The simplest stationary point process is the Poisson process. Intuitively, the process $\{t_n\}$ can be approximated as a Poisson process if it represents rare events. More rigorously, it requires that τ_κ be independently and identically distributed and follow a negative exponential function. The main deficiency of the simple Poisson model is its inability to describe the aftershocks which are often triggered by a large main shock. However, for most practical engineering purposes, the simple Poisson model for earthquakes appears to be adequate. In practice, an engineer is concerned with the earthquake risk of structures located in some specific geographic areas. The risk depends heavily on the statistics of large earthquakes in these areas, and the omission of small earthquakes or aftershock processes is relatively unimportant in terms of earthquake risk.

If $\{N(t);\ t \geqq 0\}$ is assumed to be Poisson with a constant rate, α, then

$$\text{prob}\,[N(t) = n] = \frac{(\alpha t)^n}{n!} e^{-\alpha t} \tag{7}$$

and

$$\bar{N}(t) = \alpha t. \tag{8}$$

The parameter α per unit time t can be determined from regional seismicity data.[1]

2.3 *The Mean Failure Probability of Structure, $\bar{p}$*

The quantity $\bar{p}$ depends on both earthquake and building parameters. The earthquake intensity "i" to which the structure is designed will be the only decision variable considered in this formulation and all other parameters are assumed known. Let $Y(t) = \{\max |\, y(t)\,|\ ;\ t \in [0, t_o]\}$, where, t_o equals the duration of the structural vibration which is assumed approximately equal to the duration of the earthquake ground motion, and $y(t)$ equals the response parameter (displacement, velocity, acceleration, stress, etc.) of the structure. For an earthquake with intensity i, failure could occur when the resulting structural response $y(t)$ equals or exceeds the actual resistance, x, of the structure. The corresponding failure probability, $p(i)$, can be expressed by

$$p(i) = \text{prob}\,[Y \geqq x \mid \text{earthquake with intensity } i \text{ has occurred}]. \tag{9}$$

The quantity $p(i)$ is a function of the random variable i representing the earthquake intensity whose probability density function $f_I(i)$ can be found in terms of the regional seismicity and earthquake source

geometry. Since i, in the Modified Mercalli intensity scale, takes only discrete integer values from one to twelve, the mean failure probability of structure, $\bar{p}$ is given by

$$\bar{p} = \sum_{i=i_o}^{12} p(i) f_I(i), \tag{10}$$

in which i_o equals the design earthquake intensity. From eqs. (9) and (10) it is clear that earthquake parameters enter the formulation of the problem through $p(i)$ and $f_I(i)$, while building parameters enter the formulation through $p(i)$ only.

The density function $f_I(i)$ can be derived from an expression obtained by Cornell[3] for the distribution function of earthquake intensity i:

$$F_I(i) = 1 - \frac{1}{l} \Gamma J \exp\left(-\frac{\beta}{c_2} i\right). \tag{11}$$

The governing assumptions for eq. (11) are:

(*i*) The earthquake magnitude m is a random variable with independent and negative-exponential distribution function

$$F_M(m) = 1 - e^{-\beta m}, \quad m \geqq m_o \tag{12}$$

where m_o is some magnitude small enough, say 4, that events of lesser magnitude may be ignored by engineers, and β is a constant the inverse and inverse square of which represent the mean and variance of earthquake magnitude $m > 0$ respectively;

(*ii*) The intensity attenuation law is given by

$$i = c_1 + c_2 m - c_3 \ln r \tag{13}$$

in which c_1 , c_2 and c_3 are regional seismicity constants,† and r, the focal distance in miles, is the random variable representing the distance from the structural site to the location of an earthquake source on the fault line.

(*iii*) The earthquake has a line source (fault line) of length l (in miles) with uniform distribution.

The parameters Γ and J are given by:

$$\Gamma = \exp\left[\beta\left(\frac{c_1}{c_2} + m_o\right)\right] \tag{14}$$

$$J = 2\int_d^{r_o} \frac{dr}{r^\gamma \sqrt{r^2 - d^2}}, \qquad \gamma = \beta \frac{c_3}{c_2} - 1 \tag{15}$$

† For example, c_i, $i = 1, 2, 3$ are semi-empirical constants on the order of 8, 1.5, and 2.5, respectively, for firm ground in Southern California.[4]

in which $d = \min r$ and $r_o = \max r$.

It follows from eq. (11) that

$$f_I(i) = \frac{dF_I(i)}{di} = \frac{\beta \Gamma J}{lc_2} \exp\left(-\frac{\beta}{c_2} i\right). \tag{16}$$

2.4 *Determination of the Failure Probability of Structure, $p(i)$*

To find the structure's failure probability $p(i)$, as defined by eq. (9), it is necessary to specify the failure mechanisms of the structure. It is also necessary to establish stochastic models for the response parameters $y(t)$ and the corresponding resistance x of the structure. For a linear and deterministic structure which is assumed to experience no plastic deformations and the properties of which are governed by given constants, the response model can be obtained given a stochastic model for the input earthquake ground motion. For the case of random strength, x, it becomes necessary to determine the distribution function $F_X(x)$ based on statistical and laboratory tests on individual building components as related to the overall structural resistance, e.g., Kennedy.[5] Similar tests were used by Freudenthal and Wang to establish a representative distribution of the ultimate strength of aircraft structures.[6]

In this study, consideration will be limited to the first excursion failure only, and the input and response of the structure are both treated as random processes. The probability of the duration of the response amplitude excursion and other failure mechanisms such as wearing and fatigue are not considered. In the first excursion failure, a structure is said to have failed if the response parameter, $y(t)$, exceeds a prescribed resistance or strength level, x, during the vibration cycles caused by the earthquake. Let the duration of the structural vibration be approximated by t_o, then, for any $t \in [0, t_o]$

$$\begin{aligned} p(i) &= \text{prob}\,[\,|\,y(t)\,| \geqq x; 0 \leqq t \leqq t_o] \\ &= 1 - W_i(t_o) \end{aligned} \tag{17}$$

in which $W_i(t_o) = \text{prob}\,[\,|\,y(t)\,| < x; 0 \leqq t \leqq t_o]$ = the reliability of the system. Two different situations in the structure's resistance characteristics will be considered below.

2.4.1 *Deterministic Resistance Variable, x*

A random process model for the response of structures subjected to a stationary earthquake excitation can be established as follows. A simple structure can generally be treated as a lightly damped linear

oscillator and its response, $y(t)$, is related by a second-order differential equation to the excitation, e.g., $a_i(t)$, the ground acceleration of an earthquake with intensity i. A multistory structure can be treated similarly in generalized coordinates considering normal mode vibrations. The structural response $y(t)$ in our case is a Gaussian process which approaches stationarity after a few cycles of initial transient motions. Let $p_1(t)dt$ denote the probability that $y(t)$ exceeds the threshold $y = x$ during the interval $[t, t + dt]$ for the first time since the initial time $t = 0$. The probability density function $p_1(t)$, referred to as the first-crossing density, is related to the reliability function by $-(dW/dt) = p_1(t)$. While establishing the precise behavior of $p_1(t)$ for small t poses some difficulty, for most practical purposes some approximations can be made for large mean failure time t. The simplest approximation to the first-crossing density is to assume that the up-crossings of the threshold occur rarely in the stationary response, so that they can be considered as statistically independent events. If so, the instants at which $| y(t) |$ cross the level x from below would constitute a Poisson process with a constant rate $2\nu_x$, where ν_x is the level crossing rate of $y(t)$ at the level $y = x$. In this situation it can be easily shown[7] that

$$\begin{aligned} p(i) &= 1 - \exp(-2\nu_x t_o) \\ &= 1 - \exp[-2\nu_o t_o \exp(-x^2/2\sigma_y^2)] \end{aligned} \tag{18}$$

in which $\nu_o = \sigma_{\dot{y}}/(\pi\sigma_y)$ = zero crossing rate of $y(t)$; $\nu_x = \nu_y|_{y=x} = \nu_o \exp(-x^2/2\sigma_y^2)$; where σ_y = standard deviation of $y(t)$, $\sigma_{\dot{y}}$ = the standard deviation of $\dot{y}(t) = dy(t)/dt$. All these quantities are dependent on earthquake parameters (therefore, on intensity i) and building parameters. These dependences will be derived later in this study. More specifically, it will be shown that σ_y and $\sigma_{\dot{y}}$ are directly proportional to i and that ν_o is a constant.

Expressions for $p(i)$ under other assumptions on the response process are presented in Appendix A.

2.4.2 *Random Resistance Variable, x*

The resistance variable x is a random variable with probability density function $f_X(x)$. In this situation, the failure probability is given by

$$p(i) = \int_0^\infty \text{prob}[\max_t | y(t) | \geqq x; 0 \leqq t \leqq t_o] f_X(x)\, dx. \tag{19}$$

Let $Y(t_o) = (\max_t | y(t) |, 0 \leqq t \leqq t_o)$; and $N_y(t)$ = the number of

peaks of $|y(t)|$ in the time t; then $\{N_y(t); 0 \leqq t \leqq t_o\}$ is a random process. Assuming it is a stationary Poisson process of intensity λ_y, it follows from eq. (18) that $\lambda_y = 2\nu_x$. For situations as described in Appendix A, $\lambda_y = 2\nu_o ln[1 - \exp(-x^2/2\sigma_y^2)]$ from assumption (i); and $\lambda_y = 2\nu_x\sqrt{2\xi_o}x/\sigma_y$ from assumption (ii).

The first excursion probability can be expressed in terms of λ_y as:

$$\text{prob}\,[Y(t_o) \leqq x] = \exp[-\lambda_y t_o(1 - F_x(x))] \tag{20}$$

in which $F_X(x) = \int_{-\infty}^{x} f_X(x')\,dx'$. From eqs. (19) and (20) the expression for $p(i)$ becomes

$$p(i) = \int_0^\infty \{1 - \exp[-\lambda_y t_o(1 - F_X(x))]\} f_X(x)\,dx. \tag{21}$$

A closed form solution of eq. (21) is possible when $f_X(t)$ has a simple expression such as a uniform, Gaussian, or Rayleigh density function. In general, eq. (21) can be conveniently solved by numerical integration.

2.5 *Ground Motion Statistics*

The statistics characterizing the random ground motion shall now be brought into the formulation. Let $a_i(t)$ be the ground accelerations of earthquakes of intensity i and assume $\{a_i(t); 0 \leqq t \leqq t_o\}$ be a stationary process with a power spectral density function $G_{ai}(\omega)$, where ω is the frequency variable. Such a stochastic ground motion model has been proposed and used extensively, e.g., by Liu[8] and by Jennings et al.[9] Further, assume that the process $\{a_i(t)\}$ is a filtered white noise with a constant power spectrum density G per unit intensity, and that the ground filter is a linear, single-mode oscillator with constant frequency and damping characteristic values ω_g and ξ_g, respectively. The following is a derivation of eq. (22) showing the direct relationship between $G_{ai}(\omega)$ and intensity i.

$$G_{ai}(\omega) = \frac{i^2 G}{(\omega^2 - \omega_g^2)^2 + 4\xi_g^2\omega^2\omega_g^2}. \tag{22}$$

The earthquake intensity value is a measure of the damage potential which is represented by the corresponding response spectrum $S_y = \max_t |\dot{y}(t)|$ for a structure with natural frequency and damping parameters ω_o and ξ_o. Therefore, $S_y = S_y(i, \omega_o, \xi_o, t_o)$ is clearly an increasing function of i. The precise functional relationship between S_y and i is not yet known, but can be obtained from data fittings of calculated response spectral values of past earthquakes with known i.

A simple linear approximation may be made for our analysis by assuming

$$S_{\ddot{y}}(i, \omega_o, \xi_o, t_o) = ki \tag{23}$$

where k is a constant of proportionality. From the ground acceleration model defined above, it can be shown[10] that

$$\bar{S}_{\ddot{y}} = \omega_o K \bar{\sigma}_{\dot{y}} \tag{24}$$

where

$$K = (2 \ln \nu_o t_o)^{\frac{1}{2}} + 0.577(2 \ln \nu_o t_o)^{-\frac{1}{2}}. \tag{25}$$

Since K is independent of i, as will be shown later in this section, it is obvious that

$$G_{a_i}(\omega) = i^2 G_o(\omega) \tag{26}$$

satisfies eqs. (23) and (24); and according to the ground motion model as defined earlier, $G_o(\omega)$ is given[10] by

$$G_o(\omega) = \frac{G}{(\omega^2 - \omega_g^2)^2 + 4\xi_g^2 \omega^2 \omega_g^2}. \tag{27}$$

Finally, eq. (22) follows directly from eqs. (26) and (27).

It may be noted from eqs. (23), (24), and the relation $\sigma_{\dot{y}}^2 = \int_{-\infty}^{\infty} G_{a_i}(\omega) \mid H(\omega) \mid^2 d\omega$, that the power spectrum density $G_{a_i}(\omega)$ of the earthquake process is proportional to i^2, which agrees with eq. (26). Housner and Jennings[11] have used the relationship $G_{a_i}(\omega) = \text{const.}\ S_{\dot{y}}^2$, which also leads to our assertion of eq. (26).

A difficulty exists in determining the value of the constant spectral density G corresponding to a unit Modified Mercalli intensity level. Because the intensity cannot be precisely related to the earthquake waveform parameters such as the amplitude of acceleration, velocity, displacement, response spectrum intensity, etc., some normalization procedures based on these parameters must be used to determine G. For example, a constant power spectral density level for the input white noise to the ground filter is determined by matching the corresponding expected velocity spectra of the filter's response to Housner's average velocity spectra.[12]

Using the well-known relation $\sigma^2 = \int_{-\infty}^{\infty} G(\omega)\, d\omega$, (i.e., the variance of a random process is equal to the integral of its power spectral density over the entire real line representing frequency), it follows from eq. (24) that the variances of $a_i(t)$ and $\dot{a}_i(t)$ are respectively $\sigma_a^2 = i^2 G\pi/(2\xi_g \omega_g^3)$ and $\sigma_{\dot{a}}^2 = \omega_g^2 \sigma_a^2$. Also, from the relation $G_{\dot{y}}(\omega) = \mid H(\omega) \mid^2 G_{a_i}(\omega)$ in which $H(\omega) = (\omega^2 - \omega_o^2 - 2j\xi_o \omega_o \omega)^{-1}$ = the transfer function of the

simple structure for displacement output $y(t)$ and input $a_i(t)$, where j represents the complex unit, it can be shown that[1]

$$\sigma_y = \left(\frac{\pi GB}{A_0A_1B - A_0^2A_3^2}\right)^{\frac{1}{2}} i = \theta_y i \tag{28}$$

$$\sigma_{\dot{y}} = \left(\frac{\pi GA_3}{A_1B - A_0A_3^2}\right)^{\frac{1}{2}} i = \theta_{\dot{y}} i \tag{29}$$

in which $A_0 = \omega_g^2\omega_o^2$, $A_1 = 2\omega_o\omega_g(\xi_o\omega_g + \xi_g\omega_o)$, $A_2 = \omega_o^2 + \omega_g^2 + 4\xi_o\xi_g\omega_o\omega_g$, $A_3 = 2(\xi_o\omega_o + \xi_g\omega_g)$ and $B = A_2A_3 - A_1$.

From eqs. (28) and (29), the zero crossing rate of $y(t)$ is

$$\nu_o = \sigma_{\dot{y}}/\pi\sigma_y = \frac{1}{\pi}\left(\frac{A_0A_3}{B}\right)^{\frac{1}{2}}. \tag{30}$$

To show that K is independent of i in eq. (25), it is sufficient to show that ν_o is likewise independent of i. This is obvious from eq. (30).

This completes the discussion on the determination of the failure probability of structure $p(i)$. It is shown that eqs. (18), (21), and eqs. (37) through (39) in Appendix A define $p(i)$ for various failure mechanisms. Furthermore, substituting eqs. (27) through (30) in the appropriate terms in $p(i)$, indicates that $p(i)$ is a function of intensity i. Finally, substituting eq. (16) and various expressions for $p(i)$ into eq. (10) determines $\bar{p}$, which is a function of i_o.

2.6 *The Expected Random Earthquake Loss, $\bar{e}$*

The earthquake loss depends on earthquake and building parameters, and the extent to which human lives are in danger. The quantity $\bar{e}$ can be determined from statistical data of actual earthquake damage. Unfortunately, empirical values of $\bar{e}$ for different classes of constructions are not presently available. It is logical to assume that $\bar{e}$ is directly proportional to the damage potential of earthquakes, therefore, either of the following relationships, or their combinations, may be appropriate:

$$\bar{e} = C_1 \max_t \overline{a_i(t)} \tag{31a}$$

$$\bar{e} = C_2\overline{t_o} \tag{31b}$$

$$\bar{e} = C_3\{\bar{S}_y \text{ , or } \bar{S}_{\dot{y}} \text{ , or } \bar{S}_{\ddot{y}}\}, \tag{31c}$$

in which C_1, C_2, C_3 are the constants of proportionality and $\bar{S}$ represents the expected response spectrum associated with the subscript response parameter. Equation (31c) in which $\bar{e}$ is expressed in terms of the expected velocity spectrum $\bar{S}_{\dot{y}}$ appears superior to others because the effects of the amplitude, duration as well as the frequency char-

acteristics of the earthquake accelerogram, are all considered. Thus it will be used in the following analysis. Let $C_3 = \epsilon A_f f$, where ϵ is the percent loss in building cost per unit response velocity spectrum and ϵf is the expected earthquake loss per unit floor area per unit response spectrum. It should be noted that no upper bounds for ϵ can be established in situations where human lives are involved. In these situations, the value of ϵ would increase for large-occupancy structures such as hospitals, schools and office buildings, etc., and decrease for small-occupancy structures such as warehouses, machine rooms, unmanned equipment buildings, etc. The determination of ϵ, with due considerations to loss of human lives, needs actual earthquake life-loss statistics and a mathematical model which converts life-loss into dollars. These matters will need further studies and more data collection. Clearly, expressing $\bar{e}$ in terms of initial building investment is a convenient way of incorporating all possible losses in an earthquake environment. A sensitivity analysis for the parameter ϵ should provide some insight to the overall cost structure. From eqs. (24) and (31c)

$$\begin{aligned} \bar{e} &= C_3 \omega_o K \bar{\sigma}_y \\ &= \epsilon A_f f \omega_o K \theta_y \bar{i} \end{aligned} \tag{32}$$

where

$$\bar{i} = \sum_{i=1}^{12} i f_I(i).$$

It can be noted from the above that $\bar{e}$ is independent of the design intensity i_o and the expected service life of a building. This is because according to its definition, $\bar{e}$ is the expected loss associated with a "single" random event. The quantity $\bar{e}$ should not be confused with the total expected loss of building (see Appendix B for its derivation) throughout the entire service life, which should be expected to increase with service life but decrease with design intensity.

2.7 *Earthquake Risk Cost*

The expression for the earthquake loss function $K_d(i_o)$, for the case of deterministic structural resistance and independent crossings of response process can now be established. Substituting eqs. (8), (32), and (10) [$f_I(i)$, $p(i)$ and σ_y given by eqs. (16), (18) and (28)] into eq. (6) leads to:

$$\begin{aligned} K_d(i_o) = \alpha t \epsilon A_f f \omega_o K \theta_y g(t) \left(\frac{\beta \Gamma J}{l c_2}\right)^2 \sum_{i=1}^{12} i e^{-\beta i / c_2} \sum_{i=i_o}^{12} e^{-\beta i / c_2} \\ \cdot [1 - \exp(-2 t_o \nu_o e^{-x^2/2\theta_y^2 i^2})], \qquad i_o = 1, 2, \cdots, 12 \end{aligned} \tag{33}$$

in which K and ν_o are given by eq. (25) and eq. (30) respectively. To show eq. (33) is a monotonically decreasing function of i_o, let

$$\alpha t_0 \epsilon A_f f \omega_o K \theta_{ll} g(t) \left(\frac{\beta \Gamma J}{l c_2}\right)^2 = \Omega_1 ,$$

$$\sum_{i=1}^{12} i \exp\left(-\frac{\beta i}{c_2}\right) = \Omega_2 , \qquad \frac{\beta}{c_2} = \delta, \qquad 2t_o\nu_o = \mu, \qquad \frac{x^2}{2\theta_{ll}^2} = \zeta,$$

and

$$\frac{K_d(i_o)}{\Omega_1 \Omega_2} = Z_o(i_o).$$

Equation (33) is rewritten as

$$\left.\begin{aligned} Z_o(i_o) &= \sum_{k=i_o}^{12} z_k = T_{12} - T_{i_o} , \\ z_k &= e^{-\delta k}[1 - \exp(-\mu e^{-\zeta/k^2})], \\ z_o &= 0, \quad k, i_o = 1, 2, \cdots, 12 \end{aligned}\right\} \tag{34}$$

where $T_{i_o} = \sum_{k=1}^{i_o} z_k$ is a function of i_o and $T_{12} = \sum_{k=1}^{12} z_k$ is a constant. Notice in eq. (34) that for δ and $\zeta \geqq 0$, both $\exp(-\delta k)$ and $\exp(-\zeta/k^2)$ are bounded between zero and unity; furthermore, for $\mu \geqq 0$, the quantities $\exp[-\mu \exp(-\zeta/k^2)]$ and $1 - \exp[-\mu \exp(-\zeta/k^2)]$ are also bounded between zero and one. It is apparent that since $0 \leqq z_k \leqq 1$ and $T_i = \sum_{k=1}^{i} z_k \geqq 0$, therefore Z_o is a monotonically decreasing function of i_o, as is expected.

Substituting eqs. (3) and (34) into eq. (1), the final expression for the objective function is obtained:

$$\begin{aligned} K(i_o) &= K_c(i_o) + K_d(i_o) \\ &= A_f(f + ci_o) + \Omega_1 \Omega_2 Z_o(i_o) \end{aligned} \tag{35}$$

in which the first term in the right-hand side increases with i_o and the second term decreases with i_o. The optimum intensity i_o^* is determined by setting equal to zero the first derivarive of eq. (35) with respect to design intensity i_o, i.e.,

$$\frac{\partial Z_o(i_o)}{\partial i_o} = -\frac{A_f c}{\Omega_1 \Omega_2}. \tag{36}$$

Alternatively, i_o^* can be obtained from eq. (35) by direct, numerical evaluation of the function $K(i_o)$ for all i_o.

As an illustration of the presented approach, a hypothetical building design problem is numerically analyzed for its optimal design earth-

quake intensity. Figure 2 is a plot of i_o versus the normalized total cost $K(i_o)/A_f f$ computed using the following data: The cost data, ϵ = 1 percent, $g(t)$ = 5 percent; the building data, x = deflection = 0.9×10^{-2} ft; the earthquake data, l = 50 miles, α = 3 earthquakes per year, $\beta = 2$, $t_o = 25$ s, $t = 40$ yrs, $c_1 = 8.16$, $c_2 = 1.45$, $c_3 = 2.46$, $m_o = 5$, $\omega_g = 31.4$ rad/s, $\xi_g = 0.5$, $\omega_o = 3.14$ rad/s, $\xi_o = 0.5$, and d = 40 mi, and $G = 1.6\ g^2s^{-3}$. Although these numerical values are used for illustration purposes, nevertheless, the earthquake parameters reflect realistic data based on seismicity in Southern California. Four different seismic protective ratios c/f = 0.01 to 0.04 are considered. The results indicate that for this specific design problem, the optimal design intensity for the building is VII.

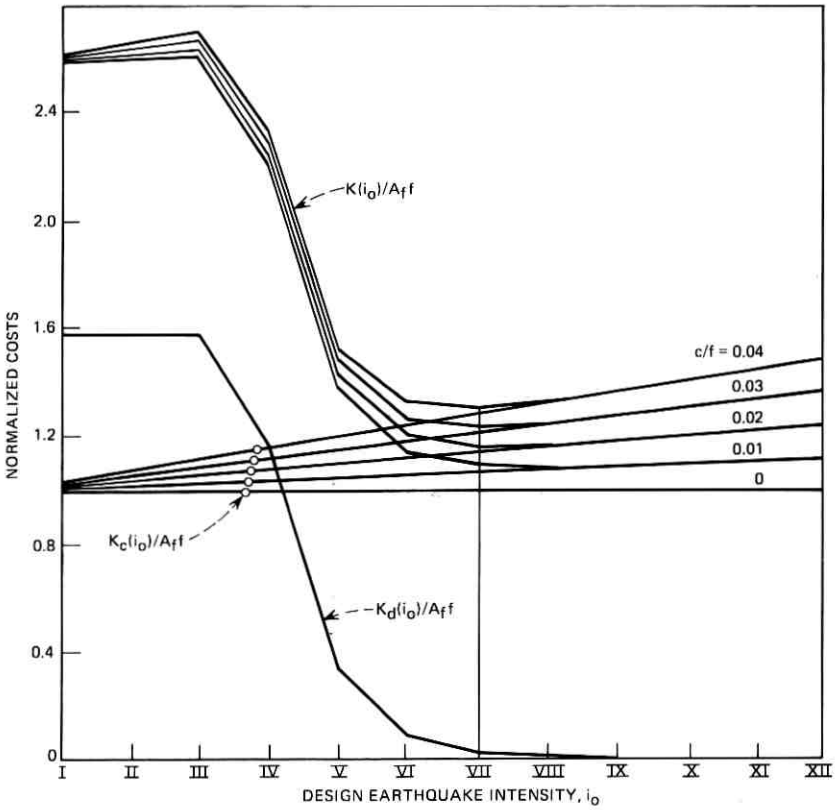

Fig. 2—Optimum design earthquake intensity analysis for a sample building.

A preliminary sensitivity analysis on this numerical example indicates that the convexity of the total cost function becomes more apparent as the building resistance parameter x increases (in this case building deflection measured in feet). For the problem under consideration, the value of x above which the design intensity could be established is found to be 0.7×10^{-2} feet. For smaller values of x, the cost ratio $K(i_o)/A_f f$ becomes insensitive to design intensity i_o and ratio c/f as the total cost curve becomes flat for $i_o > \text{V}$.

III. CONCLUSION

A simple mathematical approach is presented to determine the optimal design intensity of earthquakes for structures. The objective function to be optimized is taken as the total construction cost of the structure plus the expected cost of earthquake damage throughout the entire service life of the structure. For the case of deterministic structural resistance and probabilistic structural response with Poisson (independent) crossings, the objective function is derived in terms of the building and earthquake variables. The optimal design intensity can then be determined by minimizing the objective function with respect to the intensity variable. Other optimum design variables can also be obtained by simply regarding them as the decision variables in the objective function and by performing optimization analysis. The resulting equations are relatively simple and can be easily handled for numerical studies and sensitivity analysis. Generalizations of the results for nondeterministic structural resistance and for response processes different from those having Poisson crossings are also indicated.

IV. ACKNOWLEDGMENTS

The authors wish to express their appreciation to A. H. Carter and J. McDonald for the many valuable comments and suggestions during their careful review of the manuscript.

APPENDIX A

Expressions for $p(i)$ under Different Assumptions

(*i*) *Independent Peaks*—The dispersion in the number of peaks of the narrowband response $y(t)$ is neglected and the magnitudes of these peaks are assumed to be statistically independent variables,

and each having the probability distribution $P[|\ y\ |_{\text{peak}} < x] = 1 - (\nu_x/\nu_o)$, then[7]

$$p(i) = 1 - \exp\,[-2\nu_o t_o \ln\,(1 - e^{-x^2/2\sigma_y^2})]. \tag{37}$$

(*ii*) *Independent Envelope Crossings*—The crossings of the envelope of $y(t)$ are assumed to be independent and in this situation[7]

$$p(i) = 1 - \exp\,[-2(2\xi_o)^{\frac{1}{2}}(x/\sigma_y)t_o \exp\,(-x^2/2\sigma_y^2)] \quad \text{for} \quad \xi_o \ll 1 \tag{38}$$

in which ξ_o equals the damping ratio of the structure.

(*iii*) *Two-State Markov-Process Assumption*[13]—The successive intervals that the envelope of $y(t)$ spends above and below the level x are assumed to be random variables with exponential distributions. In this case

$$p(i) = 1 - \exp\,[-(1 - \nu_x/\nu_o) \exp\,(-\delta_x t_o)] \tag{39}$$

where

$$\delta_x = n_x(1 - \nu_x/\nu_o)^{-1}$$

and $n_x = \int_0^\infty p(r, \dot{r})\dot{r}\ d\dot{r}\ |_{r=x}$ is the envelope crossing rate of $y(t)$.

APPENDIX B

Expressions for Failure Probability, $p(t)$, and Total Expected Loss of Building, D

Let $h(t)$ be the expected loss in case of failure and $p(t)$ be the failure probability density function for the building, then

$$D = \int_0^{t_D} h(t)p(t)\ dt. \tag{40}$$

where t_D is the service life of the building. From the logic leading to eq. (32), and assuming a discount factor of cost $g(t) = 1 - (t/t_D)$, the function $h(t)$ can be written as

$$h(t) = K_1(i_o)(\eta + g(t)) \tag{41}$$

where η is percent of construction cost, representing the earthquake loss. The failure density $p(t) = dP(t)/dt$, where $P(t)$ is the failure probability of the building and is given by

$$P(t) = 1 - \sum_{i=i_o}^{12} \sum_{n=0}^{\infty} \text{prob}\,[N_i(t) = n](1 - p(i))^n \tag{42}$$

where $p(i)$ is given by eqs. (18) and (21), and $N_i(t)$ equals the total (random) number of earthquakes of intensity i in t years. According to eq. (7)

$$\text{Prob}\,[N_i(t) = n] = \frac{1}{n!} \exp\,(-\Delta F_i \alpha t)(\Delta F_i \alpha t)^n \tag{43}$$

where ΔF_i is the probability that given an earthquake occurs, this earthquake has an intensity equal to i. For a linear source of earthquakes it follows from eq. (11) that

$$\Delta F_i = \frac{1}{l}\,\Gamma J \exp\left(-\frac{\beta i}{c_2}\right)\left[1 - \exp\left(-\frac{\beta}{c_2}\right)\right]. \tag{44}$$

Equations (40) through (44) completely define the total loss expectation, D, and from above, it is obvious that D increases with t_D and decreases with i_o.

REFERENCES

1. Liu, S. C., and Fagel, L. W., "Earthquake Environment for Physical Design: A Statistical Approach," B.S.T.J., *51*, No. 9 (November 1972), pp. 1957–1982.
2. Cox, D. R., and Lewis, P. A. W., *The Statistical Analysis of Series of Events*, London: Mathen, 1966.
3. Cornell, C. A., "Engineering Seismic Risk Analysis," Bull. Seismological Society of America, *58*, No. 6 (October 1968), pp. 1583–1606.
4. Esteva, L., and Rosenblueth, E., "Spectra of Earthquakes at Moderate and Large Distances," Soc. Mex. de Ing. Sismica, Mexico II (1964), pp. 1–18.
5. Kennedy, R. P., "A Statistical Analysis of the Shear Strength of Reinforced Concrete Beams," Technical Report No. 78, Civil Eng. Department, Stanford University, Stanford, California, April 1967.
6. Freudenthal, A. M., and Wang, P. Y., "Ultimate Strength Analysis of Aircraft Structures," J. Aircraft, *7*, No. 3 (May-June 1970), pp. 205–210.
7. Crandall, S. H., "First-Crossing Probabilities of the Linear Oscillator," Journal of Sound and Vibrations, *12*, No. 3 (1970), pp. 285–299.
8. Liu, S. C., "Earthquake Response Statistics of Nonlinear Systems," Jr. Eng. Mechanics Div., Proceedings of the Am. Soc. Civil Eng., *95*, EM2 (April 1969), pp. 397–419.
9. Jennings, P. C., Housner, G. W., and Tsai, G. W., "Simulated Earthquake Motions," Technical Report, Earthquake Engineering Research Laboratory, California Institute of Technology, Pasadena, California, April 1968.
10. Shinozuka, M., "Application of Stochastic Theory to Earthquake Engineering," Proceedings of International Conference on Structural Safety and Reliability, Smithsonian Institute, Washington, D. C., 1969.
11. Housner, G. W., and Jennings, P. C., "Generation of Artificial Earthquakes," Jr. Eng. Mechanics, Am. Soc. Civil Engineers, *90*, No. EM1 (1964), pp. 113–150.
12. Penzien, J., "Applications of Random Vibration Theory in Earthquake Engineering," Bull. Int. Institute of Seismology and Earthquake Eng., Tokyo, *2* (1965), pp. 47–70.
13. Vanmarcke, E. H., "On Measures of Reliability in Narrow-Band Random Vibration," Research Report No. R69-20, Civil Eng. Department, MIT, 1969.

Contributors to This Issue

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.

Anatol Kuczura, B.S. (Engineering Physics), 1961, University of Illinois; M.S. (Mathematics), 1963, University of Michigan; M.S.E.E., 1966, New York University; Ph.D. (Mathematics), 1971, Polytechnic Institute of Brooklyn; Bell Laboratories, 1963—. From 1963 to 1966, Mr. Kuczura worked in military systems engineering. Since 1966, he has been engaged in research on the application of probability theory and stochastic processes to the analysis of telephone traffic and queuing systems. Member, ORSA, SIAM, American Mathematical Society, Mathematical Association of America, AAAS, Chi Gamma Iota, Pi Mu Epsilon.

S. C. Liu, B.S. in C.E., 1960, National Taiwan University; M.S., 1964, and Ph.D., 1967, University of California at Berkeley; Bell Laboratories, 1967—. Mr. Liu has done research in structural dynamics, random vibrations, and earthquake engineering. Recently he has been concerned with structural optimization problems. Member, American Society of Civil Engineers, The Seismological Society of America.

Scotty R. Neal, B.A. (Mathematics), 1961, M.A. (Mathematics), 1963, and Ph.D. (Mathematics), 1965, University of California, Riverside; Research Mathematician, Naval Weapons Center, China Lake, California, 1964–1967; Bell Laboratories, 1967—. Since coming to Bell Laboratories, Mr. Neal has been primarily concerned with the analysis of various aspects of telephone traffic systems. He has also worked on applications of optimal linear estimation theory and certain aspects of communication theory. Member, American Mathematical Society.

FARROKH NEGHABAT, B.E.S., 1964, Brigham Young University; M.C.E., 1966, and Ph.D., 1970, University of Delaware; Corning Glass Works, 1966–1967; Bell Laboratories, 1970—. Mr. Neghabat is engaged in research and development of methodologies for central office planning and equipment layout optimization studies. Member, Operations Research Society of America, American Society of Civil Engineers, National Society of Professional Engineers, Sigma Xi, The International Society for Technology Assessment.

DAVID SLEPIAN, University of Michigan, 1941–1943; M.A., 1947, and Ph.D., 1949, Harvard University; Bell Laboratories, 1950—. Mr. Slepian has been engaged in mathematical research in communication theory and noise theory, as well as in a variety of aspects of applied mathematics. During the academic year 1958–59, he was a Visiting Mackay Professor of Electrical Engineering at the University of California at Berkeley and during the Spring semesters of 1967 and 1970 he was a Visiting Professor of Electrical Engineering at the University of Hawaii. He now is Professor of Electrical Engineering at the University of Hawaii and shares his time between that institute and Bell Laboratories. He was Editor of the Proceedings of the IEEE during 1969 and 1970. Fellow, IEEE, Institute of Mathematical Statistics. Member, AAAS, SIAM.