THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLVI MARCH 1967 NUMBER 3

An Adaptive Echo Canceller

By M. M. SONDHI

(Manuscript received November 3, 1966)

A novel method is presented for echo-cancellation in long distance telephone connections. In contrast with conventional echo suppressors, the device described achieves echo-cancellation without interrupting the return path. A replica of the echo is synthesized and subtracted from the return signal. The replica is synthesized by means of a filter which, under the control of a feedback loop, adapts to the transmission characteristic of the echo path and tracks variations of the path that may occur during a conversation.

The adaptive control loop is described by a set of simultaneous, nonlinear, first-order differential equations. It is shown that under ideal conditions, the echo converges to zero. Estimates of the rate of convergence are obtained. Effects of noise are discussed. The results of computer simulations of various alternative configurations of the system are described.

1. INTRODUCTION

In telephone connections that involve both 4-wire and 2-wire links, an echo is generated at the hybrid that connects a 4- to a 2-wire link. The situation at one such hybrid is illustrated schematically in Fig. 1. Here S_1 and S_2 are the two speech signals and E is the echo of S_1 which is returned along with S_2 . In practice, E is an the average about 15 dB lower than S_1 , but in extreme cases may be only 6 dB lower.

This echo has a disturbing influence on the conversation, which appears to increase with increasing round-trip delay.[1] If no steps were taken to reduce this echo, conversation would be seriously impaired

over satellite communication links with round-trip delays of hundreds of milliseconds. The devices used at the present time to combat echo are called echo suppressors. A number of different types of echo suppressors have been designed. They are all, basically, voice-operated switches (albeit ingenious and complex ones) which disconnect the return path or introduce a large attenuation in it whenever a decision mechanism indicates that the level of S_1 is large compared to that of $S_2 + E$. However, since E and S_2 both share the return path, the use of such echo suppressors introduces "chopping" or interruptions of S_2 during periods of double talking.* It has been shown that the

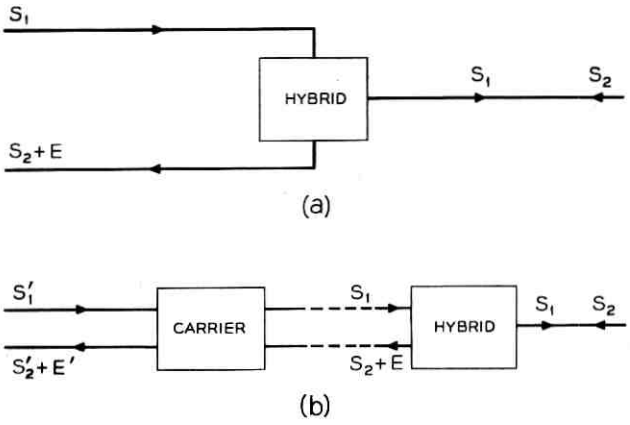


Fig. 1 — Typical situations where an echo canceller could be used.

degrading effect of chopping also increases with increasing round-trip delay.[1] The characteristics, advantages and disadvantages of such echo suppressors have been described in a number of papers.[2,3,4]

It appears that improvements in such echo suppressors are not likely to solve the echo problem satisfactorily. Entirely different approaches are called for. One such approach was an open loop device suggested by J. L. Flanagan and D. W. Hagelberger and implemented by J. de Barbeyrac.[5] In this approach, E is regarded as a linearly filtered version of S_1. The impulse response of this filter is measured by means of a transmitted test pulse, and a transversal filter is synthesized to approximate this impulse response. With S_1 as an input to the transversal filter, the output approximates E, and may, therefore, be sub-

* In this paper, the term "double-talking" will be used for the simultaneous presence, at the echo suppressor, of speech signals of the two speakers.

tracted from the return signal to cancel the echo. In this manner, effective echo cancellation (as opposed to *suppression*) is achieved without interrupting S_2. J. de Barbeyrac demonstrated the feasibility and effectiveness of such a device on four- to two-wire junctions simulated in the laboratory.

An actual echo path is, however, not perfectly constant. Besides obvious step changes such as the connection or disconnection of extension phones during a conversation, or transfer of calls via key telephones or PBX's, there may be slow changes in gain and other fluctuations of the transfer function of the echo path. Also, economic consideration would force placement of echo cancellers at switching offices high in the hierarchical structure rather than at the hybrids where the echoes are generated. (The number of echo cancellers required in the latter case would be many thousand times the number required in the former.) In such a situation, one or more carrier links intervene between the echo canceller and the hybrid. A large percentage of these (e.g., the N carrier) use compandors which are nonlinear elements with memory. Thus, what are available to the echo canceller are not S_1 and $S_2 + E$ but the modified signals S_1' and $S_2' + E'$ (see Fig. 1(b)). E' is no longer a linearly filtered version of S_1', although a linear filter with an impulse response dependent upon the power level of S_1' could approximately transform S_1' to E'. Thus, for the open loop device to work in practice, it would seem necessary to intermittently adjust the transversal filter during a conversation. The transmission of test pulses required for such adjustments might prove quite intolerable to the customers.

A proposal made by John L. Kelly, Jr. avoids these difficulties. The speech signal itself is used in place of test pulses and a control loop continuously adapts the transversal filter to take care of fluctuations in the echo path.

In this paper, we will describe the system proposed by Kelly. We will describe various modifications of the system which simplify and improve it. Finally, we will report the results of tests of these systems by computer simulation, using artificially created echoes and also using two-track tape recordings made on an actual N-carrier link.

II. KELLY'S PROPOSAL

The adaptive control loop shown in Fig. 2 is the system proposed by Kelly, except for the introduction of the nonlinear function F. This function is chosen to be an odd nondecreasing function with $F(0) = 0$. (Kelly's proposal obtains as a special case when $F(e) = e$.)

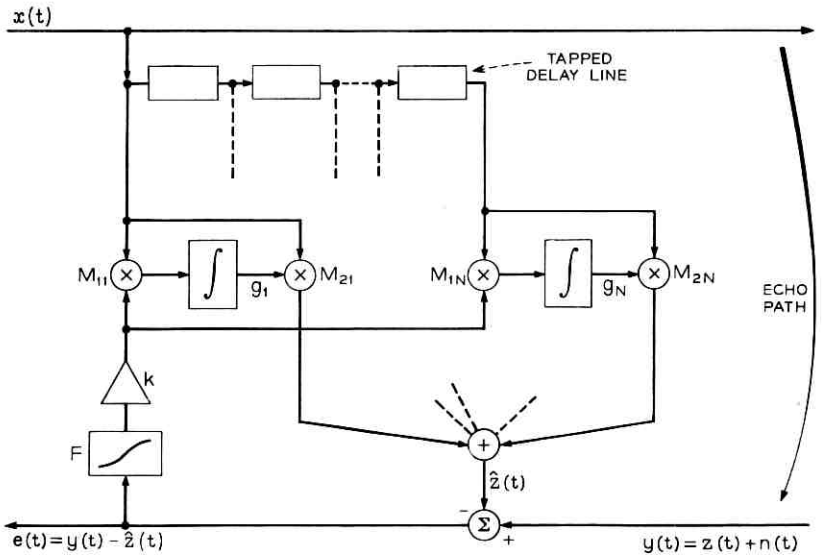


Fig. 2 —Schematic of the echo canceller using a transversal filter.

The signal $x(t)$ is the input speech signal (corresponding to S_1 or S_1' of Fig. 1). The signal $y(t)$ is the return signal (corresponding to $S_2 + E$ or $S_2' + E'$ of Fig. 1) and is given by

$$y(t) = n(t) + z(t),$$

where $z(t)$ is the echo of the input signal and $n(t)$ is a noise which is assumed statistically independent of $z(t)$. The noise may include a second speech signal besides circuit noise.

The N-tap transversal filter synthesizes an estimate of $z(t)$ given by

$$\hat{z}(t) = \sum_{1}^{N} g_k(t)x[t - (k - 1)T_d]$$

where T_d is the delay of each section of the transversal filter. The control loop uses the error $e(t) = y(t) - \hat{z}(t)$ to continuously improve the estimate $\hat{z}(t)$.

Ideally, the system should drive itself to the condition $e(t) = n(t)$ (not necessarily $e(t) = 0$, for as mentioned earlier $n(t)$ may contain a speech signal which must be left as undistorted as possible). Such ideal echo cancellation is possible with this system only if $n(t) \equiv 0$ and if $z(t)$ is exactly representable by passing $x(t)$ through an N-tap transversal filter with constant (or slowly varying) tap gains. In the

next section we will exhibit a proof that under these conditions the system does indeed converge monotonically to this echo-less steady state. In the presence of noise the final state is one in which the tap gains fluctuate around their settings for perfect echo cancellation. The response of the system averaged over the noise ensemble will be shown to converge monotonically to this final state.

The system is a first-order system and is stable for any arbitrary input. As is the case with most control systems, speed of response can be traded for immunity to noise. The constant multiplier K of Fig. 2 allows adjustment of this trade-off. (Unfortunately, in the present case speed of response depends not only upon the feedback factor K but also upon the level and properties of the signal $x(t)$. Thus, K can be adjusted to give a certain speed of response only for some average level of $x(t)$.)

For immunity from noise, K should be made as small as possible; for fast convergence, it should be made as large as possible. We do not have any theory at present to calculate the optimum setting for K. This must be done by computer simulation and/or experiments with hardware implementations of the system.

III. CONVERGENCE

The proof of convergence given in this section is very similar to a proof given by Kelly. The introduction of the nonlinear function F necessitates only minor modifications. However, as we shall show in Section V, a judicious choice of this nonlinearity can considerably simplify and improve the performance and implementation of the system.

To simplify our discussion we introduce the following notation. We denote the output $x[t - (k - 1)T_d]$ of the kth tap of the transversal filter as $x_k(t)$. We will refer to N-tuples as vectors and consider them as column matrices. Thus, $\mathbf{X}(t)$ will be a column matrix with elements $x_1(t), x_2(t), \cdots, x_N(t)$, and $\mathbf{G}(t)$ the column matrix with elements $g_1(t), g_2(t), \cdots, g_N(t)$. The signal $\hat{z}(t)$ of Fig. 2 then becomes

$$\hat{z}(t) = \sum_{k=1}^{N} g_k(t)x_k(t)$$

$$= \mathbf{G}^T\mathbf{X}.$$

Here the superscript T denotes the transpose of a matrix, and for brevity the dependence of $\mathbf{G}$ and $\mathbf{X}$ on t is not explicitly shown. The echo $z(t)$ will likewise be represented as

$$z(t) = \mathbf{H}^T\mathbf{X},$$

where $\mathbf{H}$ has elements $h_1, \cdots, h_N$ which are assumed fixed (or so slowly varying that their time derivatives may be neglected). Thus, we have

$$\begin{aligned} y(t) &= z(t) + n(t) \\ &= \mathbf{H}^T\mathbf{X} + n(t) \\ e(t) &= y(t) - \hat{z}(t) \\ &= \mathbf{R}^T\mathbf{X} + n(t), \end{aligned} \tag{1}$$

where $\mathbf{R} = \mathbf{H} - \mathbf{G}$.

Before proceeding to the proof of convergence let us give an interesting heuristic justification for the circuit of Fig. 2. Consider a function $C(e)$ such that $C(e) = C(-e)$ and $d^2C/de^2 \geqq 0$. $C(e)$ is then a monotonically nondecreasing function of the magnitude of e. Let us minimize $C(e)$ by varying the coefficients $g_k(t)$. If we choose to use the steepest descent method, we find the gradient of $C(e)$ with respect to the g_k and make the vector $\mathbf{G}$ change in the direction opposite to this gradient. Now

$$\begin{aligned} \text{grad } C(e) &= \text{grad } C(\mathbf{R}^T\mathbf{X} + n(t)) \\ &= -C'(\mathbf{R}^T\mathbf{X} + n(t))\mathbf{X} \\ &= -F(\mathbf{R}^T\mathbf{X} + n(t))\mathbf{X}, \end{aligned}$$

where $C'(\cdot) = F(\cdot)$ is the derivative of C with respect to its argument. To change $\mathbf{G}$ along the negative of the gradient we may set

$$\frac{d}{dt}\mathbf{G} = KF(\mathbf{R}^T\mathbf{X} + n(t))\mathbf{X}, \tag{2}$$

where K is a positive constant of proportionality.

By inspection, the matrix equation (2) is the equation that governs the dynamic behavior of the system of Fig. 2. Thus, the system is a steepest descent control system in the above sense.

The proof of convergence follows easily from (2) in the case when $n(t) \equiv 0$. Observe that since

$$\mathbf{R} = \mathbf{H} - \mathbf{G}, \qquad \frac{d}{dt}\mathbf{G} = -\frac{d}{dt}\mathbf{R},$$

premultiplying (2) with $-2\mathbf{R}^T$ gives

$$2\mathbf{R}^T \frac{d}{dt}\mathbf{R} = \frac{d}{dt}\mathbf{R}^T\mathbf{R}$$

$$= -2K\mathbf{R}^T\mathbf{X}F(\mathbf{R}^T\mathbf{X}). \tag{3}$$

By its definition F is a monotonic nondecreasing function and also an odd function. Thus, the right-hand side of (3) is always negative, hence $\mathbf{R}^T\mathbf{R}$ is nonincreasing. It is strictly decreasing whenever $\mathbf{R}^T\mathbf{X} \neq 0$, i.e., whenever there is an uncancelled echo. Now $\mathbf{R}^T\mathbf{R} = l_R^2$ is the square of the length of the vector $\mathbf{R} = \mathbf{H} - \mathbf{G}$. Thus, l_R is nonincreasing, and as long as there is an uncancelled echo the length keeps decreasing, i.e., $\mathbf{G}$ keeps approaching $\mathbf{H}$. To show that the echo goes to zero we integrate (3) between 0 and some time τ and obtain

$$l_R^2(0) - l_R^2(\tau) = K\int_0^\tau \mathbf{R}^T\mathbf{X}F(\mathbf{R}^T\mathbf{X})\,dt. \tag{4}$$

As l_R^2 is nonincreasing the left-hand side is bounded and $\leqq l_R^2(0)$. Thus, as $\tau \to \infty$ we note that the integrand on the right must approach zero. However, the integrand is a monotonic nondecreasing function of the magnitude of $\mathbf{R}^T\mathbf{X}$ (which is the uncancelled echo). Thus, the echo power must approach zero.*

If $y(t)$ contains a noise $n(t)$ besides the echo, then proceeding as before we find that

$$\frac{d}{dt}\mathbf{R}^T\mathbf{R} = -K\mathbf{R}^T\mathbf{X}F(\mathbf{R}^T\mathbf{X} + n(t)). \tag{5}$$

From our previous discussion it follows that the right-hand side of (5) is negative, if and only if $\mathbf{R}^T\mathbf{X} + n(t)$ has the same sign as $\mathbf{R}^T\mathbf{X}$. As long as the magnitude of the uncancelled echo is large compared to $n(t)$, this condition is met for a large percentage of time and convergence proceeds essentially monotonically as before. When the level of the uncancelled echo becomes of the same order as or lower than $n(t)$ the convergence clearly cannot be monotonic. There will be intervals when $n(t)$ is greater than R^TX in magnitude and of opposite sign. However, if the feedback gain constant K is small then $\mathbf{G}$ is a slowly varying function of time (typically K would be adjusted so that $\mathbf{R}$ has a "time constant" on the order of 0.5 sec or so). In such a quasi-stationary case it is justified to assume that $\mathbf{R}^T\mathbf{X}$ is independent of $n(t)$ provided

* I. W. Sandberg has pointed out that, strictly speaking, this argument does not prove that $|\mathbf{R}^T\mathbf{X}| \to 0$ for certain pathological cases. Although these cases are of no practical concern, it is interesting that weak conditions suffice to rule these out. For example, it is sufficient that: (*i*) $|\mathbf{X}|$ and $d/dt\,|\mathbf{X}|$ be bounded and (*ii*) the function F be such that $eF(e)$ and $d/de\,(eF(e))$ be bounded for all finite e.

$n(t)$ is a wideband signal. Then it is not hard to show that the average of $F(\mathbf{R}^T\mathbf{X} + n(t))$ over the noise ensemble has the same sign as the average of $\mathbf{R}^T\mathbf{X}$ provided only that noise has a symmetric distribution and thus, the system still converges in this average sense.

IV. RATE OF CONVERGENCE

While convergence can be proved by such relatively simple arguments, estimating the convergence rate is an extremely difficult problem. The convergence rate depends upon the properties of $\mathbf{X}$, upon the choice of F, and upon K and we do not have a solution to the problem in the general case. However, we will now derive an estimate of the mean convergence rate for the noiseless case under the assumption that (3) can be averaged over the $\mathbf{X}$ ensemble (which is assumed stationary) and the vector $\mathbf{R}$ assumed independent of $\mathbf{X}$ on the right-hand side. This assumption is justified if K is small, hence $\mathbf{R}$ slowly varying. Under this assumption, the expectation can be calculated for a variety of different functions F and random processes $\mathbf{X}$. We will give the result when $\mathbf{X}$ is a zero mean Gaussian process with (i) $F(x) = x$, and (ii) $F(x) = \text{sgn } (x)$ (here sgn $(x) = 1$ for $x \geqq 0$ and -1 for $x < 0$). In case (i)

$$\frac{d}{dt}\langle \mathbf{R}^T\mathbf{R}\rangle_{av} = -2K\sigma^2\langle \mathbf{R}^T\Phi\mathbf{R}\rangle_{av}, \tag{6}$$

where σ is the standard deviation of $x_i(t)$ (assumed identical for all the $x_i(t)$) and Φ is the normalized NxN correlation matrix of the $x_i(t)$. The angular bracket denotes ensemble averaging. In case (ii)

$$\frac{d}{dt}\langle \mathbf{R}^T\mathbf{R}\rangle_{av} = -2K\sigma\sqrt{\frac{2}{\pi}\langle \mathbf{R}^T\Phi\mathbf{R}\rangle_{av}}. \tag{7}$$

These equations follow easily since any linear combination of Gaussian variables is another Gaussian variable. Equations (6) and (7) give upper and lower bounds to the convergence rates for the two cases, when we observe that

$$\lambda_{min}\mathbf{R}^T\mathbf{R} \leqq \mathbf{R}^T\Phi\mathbf{R} \leqq \lambda_{max}\mathbf{R}^T\mathbf{R},$$

where λ_{min} and λ_{max} are the smallest and largest eigenvalues of Φ. If $x(t)$ is white noise, then Φ is an identity matrix with $\lambda_{min} = \lambda_{max} = 1$. Thus, for case (i)

$$\sqrt{\mathbf{R}^T\mathbf{R}}\,|_{t=0}\exp(-K\sigma^2\lambda_{max}t) \leqq \sqrt{\mathbf{R}^T\mathbf{R}} \leqq \sqrt{\mathbf{R}^T\mathbf{R}}\,|_{t=0}\exp(-K\sigma^2\lambda_{min}t) \tag{8}$$

and in case (*ii*)

$$\sqrt{\mathbf{R}^T\mathbf{R}\,|_{t=0}} - \sqrt{\frac{2}{\pi}}\,K\sigma\lambda_{\max}^{\frac{1}{2}}t \leqq \sqrt{\mathbf{R}^T\mathbf{R}} \leqq \sqrt{\mathbf{R}^T\mathbf{R}\,|_{t=0}} - \sqrt{\frac{2}{\pi}}\,K\sigma\lambda_{\min}^{\frac{1}{2}}t. \quad (9)$$

These upper and lower bounds are close to each other only if Φ is nearly diagonal (i.e., when $x(t)$ is broadband). More elaborate methods can be used to estimate the convergence rate (e.g., perturbation methods). However, although these methods would be extremely interesting from a theoretical point of view, they are not likely to yield much more insight into the convergence process. This is especially true in view of the fact that no satisfactory statistical description of a speech signal is available at present. For Gaussian noise (8) and (9) have been checked by computer simulation.

V. CHOICE OF NONLINEARITY

In the formulation of the echo suppressor problem discussed in Section III, the choice of the nonlinearity F depends upon the choice of the function C. This choice has a profound influence upon the behavior of the resulting system. One could set up the problem of determining the optimum F which would provide, according to some reasonable criterion, the fastest convergence and the maximum immunity from interfering noise. We do not know the solution to such a general optimization problem. In any case, since convergence rate depends upon the statistics of the signal and noise, any such optimization would be practically impossible for signals as difficult to characterize as speech signals. We have, however, found that a number of improvements over the linear case result upon making F an infinite clipper. The use of an infinite clipper has also been suggested independently by B. F. Logan.

One of the main drawbacks of making $F(e) = e$ (i.e., Kelly's proposal) is the dependence of the time constant of the control system on the signal level. Equation (6), although approximate, nevertheless indicates that the time constant (at least for a wide band signal) is proportional to the signal power. A 20-dB change in signal level thus changes the time constant by a factor of 100. If, however, an infinite clipper is used the same change in signal level changes the time constant by a factor of only 10, which is a considerable improvement.

Another important advantage of using an infinite clipper is the

considerable simplification of the circuitry. The multipliers $M_{11} \cdots M_{1N}$ can all be replaced by switch modulators, which are far cheaper and simpler than broadband multipliers.

There is another advantage in using an infinite clipper which may be described as follows. Suppose the system is near equilibrium and the echo has been reduced to a very low value. If now there is a sudden burst of noise (say a spurt of double-talking) then in the linear system the rate at which the system moves away from equilibrium is proportional to the level of the noise, whereas it is more or less independent of noise level if an infinite clipper is used. (In Section VI we will give a detailed example of this effect.) We will argue in Section VII that during intervals of double-talking the control loop be opened and the vector **G** frozen at its last value. However, any decision mechanism that would make this possible would require a finite time to make the decision. It is therefore important that the system should not depart from equilibrium too rapidly upon the introduction of a large noise in the return signal.

VI. OTHER MODIFICATIONS

It may be noted that although we started out by taking the components of the vector $\mathbf{X}(t)$ to be delayed versions of the input $x(t)$, this fact was nowhere of any importance in the proof of convergence. All that is required is that the vector $\mathbf{X}(t)$ be derived from $x(t)$ in such a way that all transformations of $x(t)$, that may possibly be produced by the round trip transmission path, should be representable as $\mathbf{H}^T\mathbf{X}$ with a suitable choice of H. This immediately gives us the possibility of generalizing the circuit of Fig. 2 to that of Fig. 3. In Fig. 3 the $w_i(t)$ are a set of impulse responses such that linear combinations of them are good approximations to most practical echo path impulse responses.

Now there is an infinite variety of sequences of functions that are complete on the semi-infinite interval. One such set is the set of Laguerre functions which is of particular interest because it can be synthesized as a simple tapped RC ladder network. The impulse response of the nth Laguerre network is given by

$$l_n(t) = e^{\alpha t} \frac{d^n}{dt^n} \left(\frac{t^n}{n!} e^{-2\alpha t} \right) \tag{10}$$

with the corresponding transfer function

$$L_n(s) = \frac{\alpha}{s + \alpha} \left(\frac{\alpha - s}{\alpha + s} \right)^n. \tag{11}$$

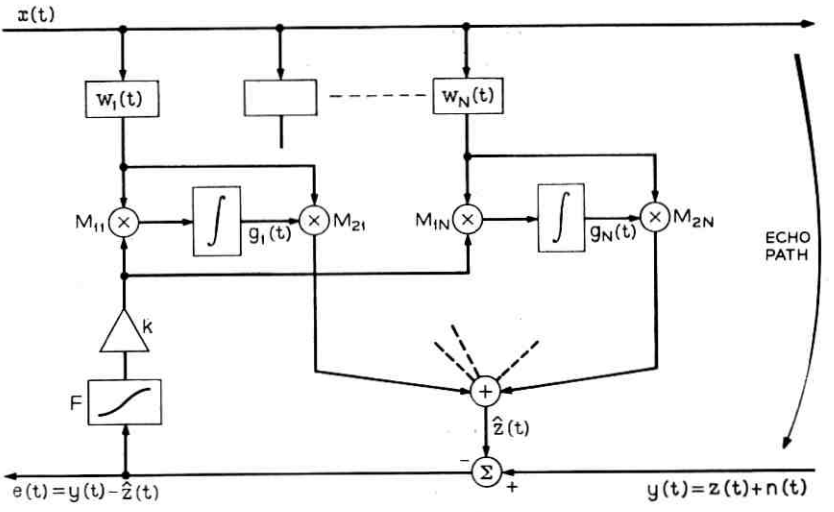


Fig. 3 — Generalization of the system of Fig. 2 using an orthonormal set of impulse responses.

As telephone speech is limited to about 3 kHz the choice of α appropriate in the present case is approximately $2\pi \times 2000$ radians per second, although it is not critical.

Tests by computer simulation, to be described in the next section, indicate that Laguerre networks are at least as satisfactory as a tapped delay line for the simulation of the echo path. However, a cascade of RC sections would be much cheaper than a delay line. The properties and synthesis of Laguerre networks is described in the literature.[6]

VII. COMPUTER SIMULATION

For a computer simulation of the system described by (2), it was converted to a difference equation. Thus, if a subscript n on a quantity is used to denote its value at the nth sampling instant, then the equation simulated on the computer is

$$\mathbf{G}_{n+1} = \mathbf{G}_n + KF(\mathbf{R}_n^T\mathbf{X}_n + n_n)\mathbf{X}_n.$$

In one class of simulations we used filtered Gaussian noise as the signal and computer-simulated echo paths. Equations (8) and (9) appear to be very good approximations if the time constant of the convergence (which we may define as the time taken for $l_R^2 = \mathbf{R}^T\mathbf{R}$ to become, say 30 percent of its initial value) is large compared to the

reciprocal of the bandwidth of the input signal. If uncorrelated white noise is added to the echo before the system has converged then no convergence takes place if the noise level is about 15 dB *above* the echo. If the same noise is introduced after the system has converged, however, the balance is only slightly disturbed. A typical example will illustrate the orders of magnitude of these effects. The nonlinearity F was chosen to be an infinite clipper, and the level of the signal (which was a white gaussian noise) and the constant K were such that in the absence of noise $\mathbf{R}'\mathbf{R}$ converged to about 55 dB below its initial value in 0.7 sec. The following two tests were performed:

(*i*) The same input signal, initial conditions, etc. were used as before, but an uncorrelated noise was added to the return signal at a level 18 dB above the echo.

(*ii*) Same as (*i*) except the noise was added after the system had converged for 0.6 sec so that l_R^2 (hence, the echo power) was about 23 dB below its initial value.

In case (*i*) no convergence took place and l_R^2 hovered around its initial value. In case (*ii*), after the onset of noise, l_R^2 increased slowly by about 3 dB in 1.5 seconds.

For comparison the same simulations were repeated with F replaced by a linear function and the constant K adjusted to give a time constant of the exponential decay of about 0.3 sec. The noiseless case and test (*i*) gave about the same results, except, of course, that the decay was exponential. In test (*ii*) the noise was introduced after 1.1 sec instead of 0.6, to allow l_R^2 to converge to about 30 dB. However, in this case, after the introduction of the noise l_R^2 rose by about 20 dB within 0.2 seconds. Thereafter it stayed at about this level.

The same kind of behavior was obtained when speech signals were used both as input and as interfering noise and when the echos were generated by the computer. Of course, it is much more difficult to estimate time constants of the convergence when speech signals are used.

We also used as inputs, digitized tape recordings of sentences spoken over an N2 carrier system. Two-track tapes were made of the input signal and the echo. Double-talking situations were also recorded and tested. For tests with these tapes, 40 to 50 delay line taps spaced 0.1 ms apart were used.

With very strong echos (0-dB return loss) the system provided a reduction of about 20 to 25 dB as measured on a VU meter. This reduction took place in 0.5 to 5 seconds depending on signal level as discussed in Section IV. The larger variation of convergence rate

with signal level when the clipper is removed, was apparent in these tests also.

When a typical hybrid and a return loss of 6 dB were used the echo was reduced to the point of being unintelligible and almost inaudible even under quiet conditions. As in the case of tests with noise signals, the introduction of double-talking at a very high level after convergence had taken place produced little change in the balance.

The fact that in the case of these recordings over the N-carrier system the echo could not be reduced by more than about 20 dB or so is undoubtedly due to the compandors used in the N-system. The echo canceller can provide only a linear approximation to the transmission path of the echo, which in the case of the N-system has nonlinearities.

We have also simulated the echo canceller using the Laguerre expansion. In this case, the digital equivalent of the Laguerre impulse responses were used. In terms of the delay parameter, $z^{-1} = \exp(-j\omega T_s)$ where T_s is the sampling interval, these are

$$L_n(z^{-1}) = \frac{1}{1 - az^{-1}} \left(\frac{z^{-1} - a}{1 - az^{-1}} \right)^n,$$

where $a = (2 - \alpha T_s)/(2 + \alpha T_s)$, α being as defined in Section V. As mentioned earlier α is chosen so that the cut off frequency of the Laguerre function is about 2 or 2.3 kHz.

VIII. DISCUSSION AND CONCLUSIONS

We have described a new method of cancelling echos in telephone connections. From our theoretical discussion and simulations it appears that the method is feasible and can yield echo cancellation of about 20 dB or so with a convergence time of about 0.2 to 0.5 second for average speech levels. This convergence time increases to 10 times its value for a 20-dB decrease in signal level. Convergence much faster than this is not possible, as then the system becomes too sensitive to noise and behaves erratically with normal noise level to be expected on a telephone connection.

We have shown that the system would not appreciably depart from equilibrium even upon the incidence of double-talking. However, it needs, initially, a period of time in which only the echo is present in the return signal (of course low-level noise may be present also, but there should be no double-talking in this period). This initial interval can be as small as 0.5 seconds if the input signal is loud, but would have to be proportionately longer for weaker input signals.

It would be advisable to break the control loop during bursts of loud double-talking. This need not be done by very sophisticated means. The following simple method should be satisfactory. Assume that the maximum level of an echo is 6 dB below the input signal. Clearly if the return signal is much larger than this it indicates double-talking. Rectification and integration with a time constant of about 0.5 second gives a reasonably good estimate of the levels of the input and return signals. A switch could then be adjusted to open the feedback path in Fig. 3 immediately following the nonlinearity F, whenever the input level is less than, say, 3 dB above that of the return signal. It is important to note that this merely prevents the gain setting **G** from changing. It *does not* interrupt the return path.

We have tested our system only on an N2 carrier system (besides on artificially generated echos). This is a double-sideband modulation system in which compandors are used. There are also in use single-sideband carrier systems, in which no compandors are used but in which carrier frequency variation (during the round-trip transmission time) would introduce a time-varying nonlinearity. The degree to which this type of variation exists and its effect require further investigation.

The delay between the input signal and its echo must be compensated for. This delay may be as large as 60 ms. The problem of automatically determining this delay and compensating for it is a challenging problem and is being investigated by a systems group at Bell Telephone Laboratories. They are also collecting a large sample of impulse responses of various connections. This information would be very useful in the final design of the system. For example, this will enable us to decide upon the optimum number of taps. Also, a fixed weighting of the gain vector **G** depending upon the statistical distribution of the impulse responses would improve the average performance of the system.

The ultimate test of the system's performance and usefulness is its actual use during normal long distance telephone conversations. For this, actual hardware must be built. Two, rather different, instrumentations have been recently completed, one by A. J. Presti[7] and the other by F. K. Becker and H. R. Rudin,[8] and it should be soon possible to carry out such tests.

REFERENCES

1. Helder, G. K., Customer Evaluation of Telephone Circuits with Delay, B.S.T.J., *45*, September, 1966, pp. 1157–1191.
2. Emling, J. W. and Mitchell, D., Effects of Time Delay and Echoes on Telephone Conversations, B.S.T.J., *42*, November, 1963, pp. 2869–2891.
3. Brady, P. T. and Helder, G. K., Echo Suppressor Design in Telephone Communications, B.S.T.J., *42*, November, 1963, pp. 2893–2917.

4. Riesz, R. R. and Klemmer, E. T., Subjective Evaluation of Delay and Echo Suppressors in Telephone Communications, B.S.T.J., *42*, November, 1963, pp. 2919–2941.
5. Flanagan, J. L. and deBarbeyrac, J., unpublished memorandum.
6. Lee, Y. W., *Statistical Theory of Communication*, John Wiley & Sons, Inc., 1960.
7. Sondhi, M. M. and Presti, A. J., A Self-Adaptive Echo Canceller, B.S.T.J., *45*, December, 1966, pp. 1851–1854, also presented at the Acoustical Society of America, 72nd meeting, November 3, 1966.
8. Becker, F. K. and Rudin, H. R., Application of Automatic Transversal Filters to the Problem of Echo Suppression, B.S.T.J., *45*, December, 1966, pp. 1847–1850, also presented at the Acoustical Society of America, 72nd meeting, November 3, 1966.

Temperature Dependence of Inversion-Layer Frequency Response in Silicon

By A. GOETZBERGER and E. H. NICOLLIAN

(Manuscript received October 28, 1966)

Conductance-voltage and capacitance-voltage curves of metal-oxide semiconductor (MOS) capacitors on n-type silicon were investigated in the temperature range between room temperature and 200°C. Plots of the inversion-layer conductance versus reciprocal temperature show a sequence of two activation energies: one corresponding to the temperature dependence of the intrinsic carrier density n_i , the other to that of n_i^2 . The low-temperature range is characterized by recombination-generation in the space-charge region, the high-temperature range by diffusion current from the bulk. The technique permits measurement of bulk lifetime for the two regimes and determination of room temperature cutoff frequency for the channel.

I. INTRODUCTION

Theoretical calculations of metal-oxide semiconductor (MOS) capacitance show a total capacitance approaching oxide capacitance in strong accumulation and strong inversion.[1] Experimentally, it has been found that response time of the inversion layer can be very long.[2] The response time can be drastically shortened, however, by lateral ac current flow in an extended inversion layer.[2,3] The lateral current flow mode requires equilibrium surface inversion beyond the metal contact. This condition is usually found in p-type silicon because of the preponderance of positive surface charge in thermally oxidized silicon. Channel cutoff frequencies are then typically in the MHz range.

In n-type silicon, charge in the inversion layer can communicate with the bulk under steady-state conditions only by means of generation-recombination processes.[2] Inversion-layer cutoff frequencies in n-type silicon are normally below 100 Hz, sometimes below 1 Hz. These low frequencies make it difficult to measure cutoff frequencies and to determine the mechanism of generation of minority carriers.

In this study, measurements with n-type silicon were carried out at elevated temperature where generation is more rapid. It is thus possible to study the generation mechanisms and confirm the theory for calculating response time. This theory was derived by Hofstein and Warfield.[2] They consider three different generation mechanisms for minority carriers. These are: bulk diffusion current, space-charge generation, and surface-state generation. Fig. 1 shows a simplified equivalent circuit proposed by Hofstein and Warfield for strong inversion. The inversion capacitance is fed by three parallel conductances corresponding to the three generation mechanisms. Because inversion capacitance is large compared to oxide capacitance with which it is in series, it can be neglected as done in Fig. 1.

The conductances are given for n-type bulk material by the following relations.[2]

For surface-state response

$$G_{g,s} = q\beta N_s N_D e^{\beta\psi_s} \sigma_p v_p , \tag{1}$$

where q = electronic charge in coulombs, N_s = surface-state density/cm^2, N_D = donor density in the bulk in cm^{-3}, $\beta = q/kT$, σ_p = capture cross section for holes in cm^2, v_p = average thermal velocity of holes in cm/sec, and ψ_s = surface potential in volts. Relation (1) was originally derived for a single level close to midgap. Because only levels in this range contribute to recombination, it is also valid for a continuum of surface states as is generally encountered in oxidized surfaces.

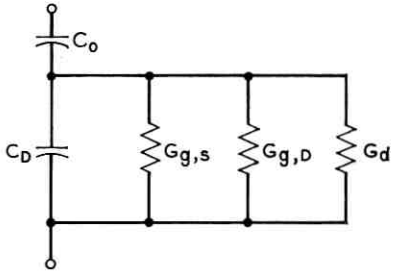


Fig. 1—Equivalent circuit of MIS capacitor in strong inversion proposed in Ref. 2. C_o is the oxide layer capacitance per cm^2, C_D is the depletion layer capacitance per cm^2, $G_{g,s}$ is the conductance arising from generation-recombination through surface states, mhos/cm^2, $G_{g,D}$ is the conductance arising from generation-recombination through states in the silicon space-charge region, mhos/cm^2, and G_d is the conductance due to the diffusion of minority carriers from the quasi neutral region in the silicon to its surface, mhos/cm^2.

Space-charge generation response:

$$G_{g,D} = \frac{qn_i\, d}{\tau_o \psi_s}, \tag{2}$$

where n_i = intrinsic carrier density in cm^{-3}, τ_o = bulk lifetime in seconds, and d = space-charge layer thickness in cm.

Diffusion response

$$G_d = \frac{q\mu_p n_i^2}{L_p N_D}, \tag{3}$$

where μ_p = hole mobility in cm^2/volt-sec, and L_p = diffusion length for holes in cm. We have further

$$L_p = (\tau_o \mu_p / \beta)^{\frac{1}{2}}. \tag{4}$$

By measuring temperature dependence of the inversion-layer response, it is possible to determine which mechanism is dominant. Surface-state generation should go with an activation energy of ψ_s. It has to be considered here that ψ_s is itself a function of temperature. Space-charge generation has the activation energy of n_i, and diffusion current that of n_i^2. In the present investigation, surface-state density was made very small, so that only $G_{g,D}$ and G_d had to be considered. This was also done because surface-state density can reach high values close to the band edges.[4,5] This, in turn, causes considerable uncertainty of the value of surface potential. In the absence of surface-state effects, the experiments reported here showed that at low temperature space-charge generation dominates while at higher temperature diffusion current takes over.

II. EXPERIMENTAL TECHNIQUE

Samples used for the measurements consisted of expitaxial layers of 1.5×10^{16} cm^{-3} doping, 10 μ thick, on low-resistivity substrates of [100] orientation. Use of epitaxial samples was advantageous because the measurements were not affected by series resistance in the substrate. Because epitaxial layers are not as perfect as regular crystals, rather low lifetimes were encountered. Samples were thermally oxidized in steam to a thickness of 1000 Å. The previously described bias oxidation technique[6] was used. In order to reduce surface-state density, the samples were subjected to a 30-minute annealing treatment in N_2 at 350°C after an aluminum film had been evaporated.[7,8] After annealing, circular areas of 3.75×10^{-2} cm diameter were etched out for MOS measurements. Capacitance and conductance were measured versus voltage

at 100 KHz and 6 KHz at various temperatures. For this purpose, the entire wafer was placed on a heated stage and contact was made to one capacitor with a wire probe. Temperature was controlled to $\pm 2°C$. Next, depletion-layer capacitance and inversion-layer conductance were extracted from the raw data by correcting for oxide capacitance as described in Ref. 5 and 3.

III. RESULTS

A family of capacitance versus voltage curves and conductance versus voltage curves at 6 KHz are shown in Figs. 2 and 3. Figs. 4 and 5 contain 100-KHz curves for the same sample. It is seen that both capacitance and conductance saturate in the inversion range at negative voltage. Due to the influence of the residual surface-state density small bumps appear in the depletion region. In Figs. 6 and 7, Arrhenius plots of the computed inversion conductance G_I are presented. These curves were obtained from the conductance curves of Figs. 3 and 5

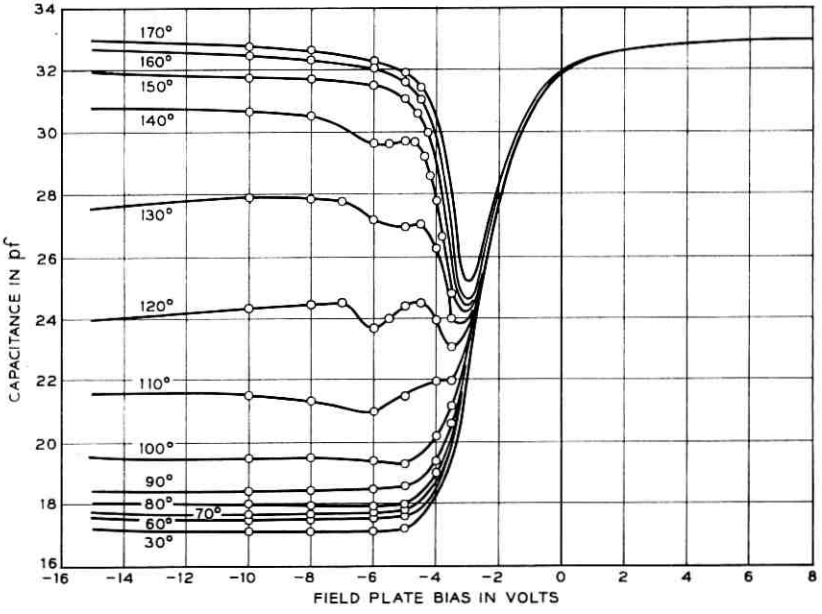


Fig. 2 — Capacitance vs field plate bias measured at 6 kHz with temperature in °C as parameter. Sample is n-type silicon oriented in the [100] direction. Field plate diameter is 370 μ, donor density is 1.17×10^{16} cm^{-3}, and oxide layer capacitance is 2.84×10^{-7} farads/cm^2.

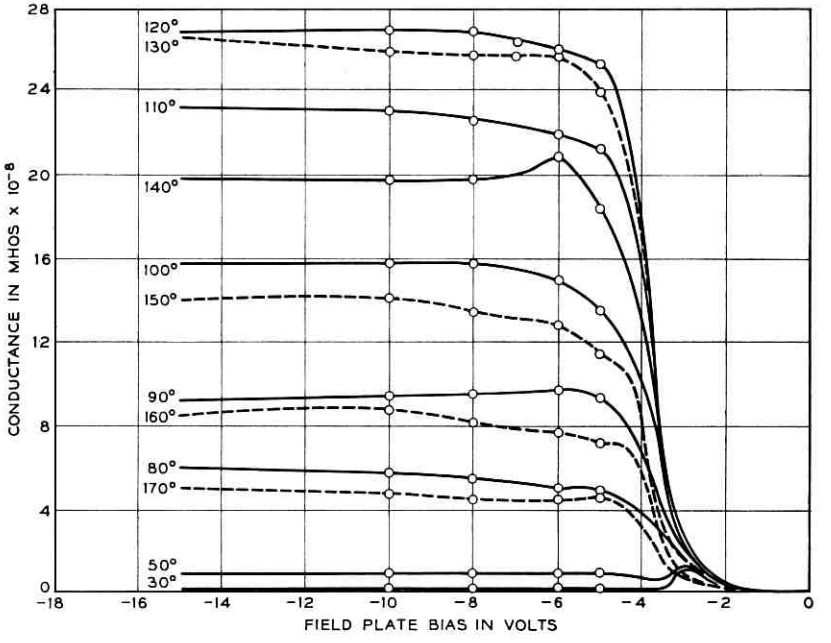


Fig. 3—Equivalent parallel conductance vs field plate bias measured at 6 kHz with temperature in °C as parameter. Sample is the same as in Fig. 2. Conductance is peaked at 120°C. Dotted lines are on high and solid lines on low-temperature side of peak.

at high negative voltage. The fact that both plots agree within the accuracy of the measurement indicates that the equivalent circuit of Fig. 1 is valid. The values of the activation energies also prove that in the surface studied here there is no noticeable influence from surface states. Fig. 8 contains room temperature capacitance-voltage curves at various frequencies.

IV. DISCUSSION

Hofstein and Warfield[2] showed that the dominant effect is most likely space-charge recombination (2). Surface recombination may also be important at relatively high surface-state densities. Because the sample investigated here contained very few surface states, it can be expected that space-charge recombination dominates. From Fig. 6 it is seen that this is the case up to temperatures around 140°C. In this range, the activation energy is 0.56 eV for Fig. 6, curve (a), and

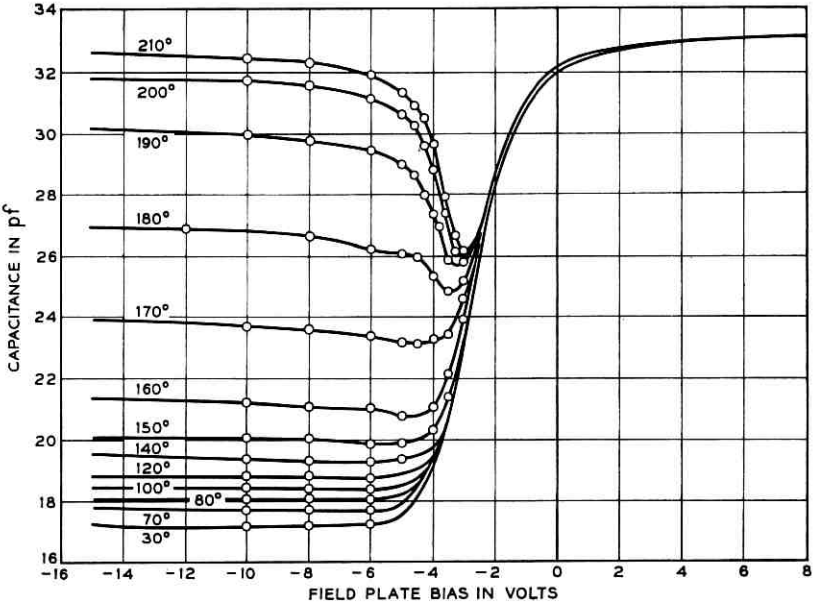


Fig. 4—Capacitance vs field plate bias measured at 100 kHz with temperature in °C as parameter. Sample is same as in Fig. 2.

0.620 eV for Fig. 7, curve (a). The expected activation energy[9] for n_i is 0.605 eV. Equation (2) can now be used to calculate bulk lifetime, τ_o, under certain simplifying assumptions given in Ref. 2. We obtain $\tau_o = 4.19 \times 10^{-9}$ seconds. This rather low lifetime is explained by the fact that it refers to an epitaxial layer.

Above 140°C a new process dominates as shown by the break in the $1/T$ curves. This process could be either surface-state generation or diffusion current from the neutral part of the bulk. It can be shown that surface-state generation is very unlikely in this case. Surface-state density as determined by the conductance technique[5] is varying between 6.1×10^{10} and 3.3×10^{11} states per cm^2 and eV. This density would, according to (1), give a conductance orders of magnitude lower than the measured G_I. It is also expected that the activation energy of surface-state processes should decrease because surface potential at constant voltage decreases considerably with increasing temperature.

If the high-temperature points in Figs. 6 and 7 are connected by a straight line, they give an activation energy of 0.908 eV for 6 KHz and 0.935 eV for 100 KHz. This energy is lower than the expected energy of 1.21 eV. The discrepancy can be resolved by correcting the

high temperature points by subtracting the influence of space-charge generation as indicated in Fig. 6, curve (*c*). If this is done, the high-temperature activation energy in Fig. 6, curve (*c*), is 1.17 eV which is very close to the expected value. Fig. 7 did not contain sufficient experimental points to carry out the correction.

Using (3) and (4), the high-temperature lifetime and diffusion length can be calculated. We find, $L_p = 20.1\ \mu$ and $\tau_o = 1.8 \times 10^{-7}$ seconds. Because the calculated diffusion length is of the order of the epitaxial layer thickness, it is possible that the actual diffusion length might be longer. In calculating the above values, a temperature dependence of the mobility μ_p of $T^{\frac{1}{2}}$ was used as is necessary for highly-doped samples.

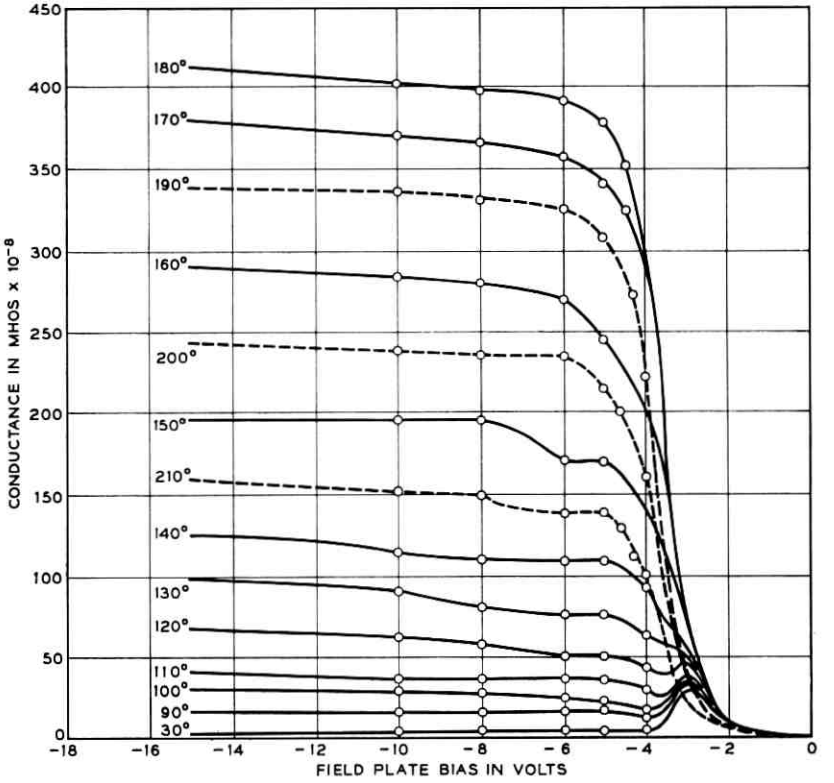


Fig. 5—Equivalent parallel conductance vs field plate bias measured at 100 kHz with temperature in °C as parameter. Sample is same as in Fig. 2. Conductance is peaked at 180°C. Dotted lines are on high and solid lines are on low-temperature side of peak.

The two lifetimes calculated from the two temperature ranges are actually expected to be equal. The linearity of the plots in curve (*b*) of Figs. 6 and 7 indicates that there is no great temperature dependence of τ_o. A possible explanation for the discrepancy of lifetimes is that they are measured in different parts of the crystal. Space-charge recombination occurs within 0.5 μ from the surface, while diffusion lifetime is determined in the entire epitaxial layer. It is likely that a thin surface layer contains a higher concentration of recombination centers.

An alternative explanation is that electron and hole lifetime are

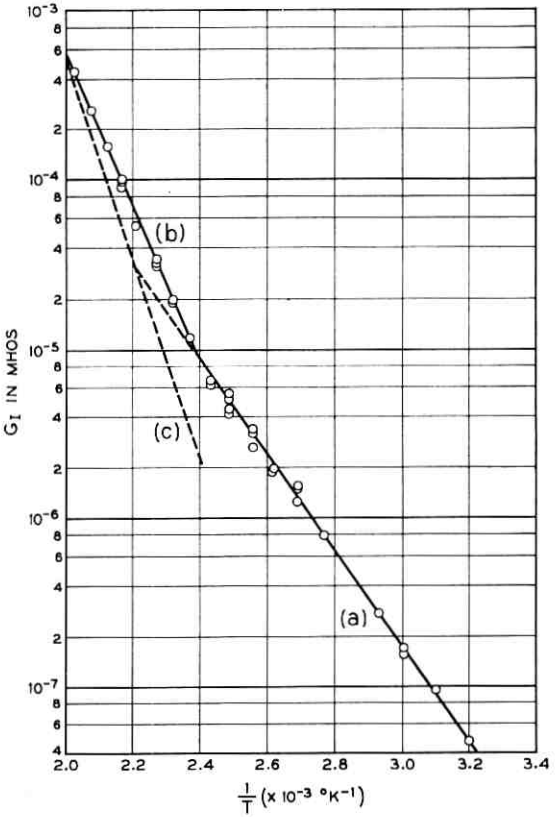


Fig. 6—Equivalent parallel conductance measured at 6 kHz and a bias of −15 volts as a function of reciprocal degrees Kelvin. The experimental points indicated by the circles were obtained from Fig. 3. Multiple circles at a given temperature represent several runs. The solid lines are the best fit to the experimental points. Curve (*c*) is obtained by subtracting the values of G_I in curve (*b*) from the extrapolation of curve (*a*) at each temperature.

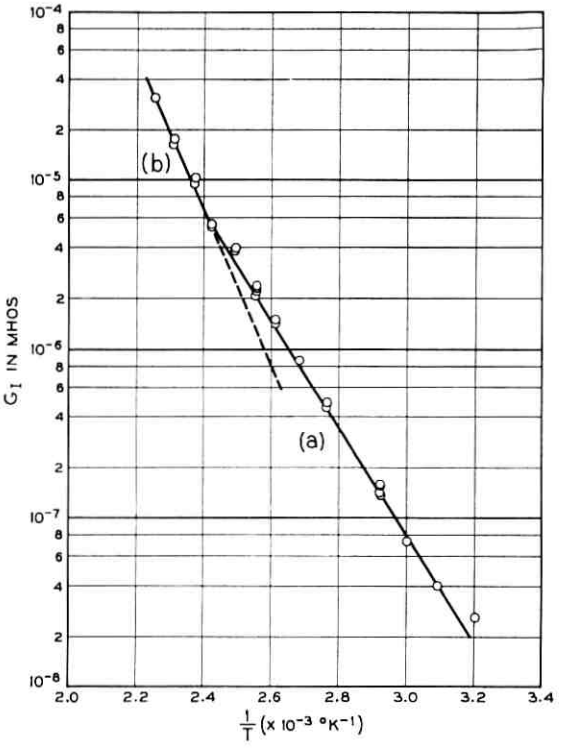


Fig. 7 — Equivalent parallel conductance measured at 100 kHz and a bias of −15 volts as a function of reciprocal degrees Kelvin. The experimental points indicated by the circles were obtained from Fig. 5. Multiple circles at a given temperature represent several runs. The solid lines are the best fit to these points.

significantly different. In this case, $\tau_o = (\tau_{no}\tau_{po})^{\frac{1}{2}}$ would have to be used in (2) and $\tau_o = \tau_{po}$ in (3). Under this assumption τ_{no} is calculated to be 10^{-10} second.

By taking inversion conductance from the curves in Fig. 6 at room temperature, inversion-layer time constant can be accurately calculated. This time constant[2] is $\tau_I = C_D/G_I = 2.25 \times 10^{-3}$ second leading to a cutoff frequency of 71 Hz. Fig. 8 demonstrates that a cutoff frequency in this neighborhood is indeed observed.

V. CONCLUSIONS

By measuring inversion conductance, it could be shown that the equivalent circuit and theory by Hofstein and Warfield is valid. The

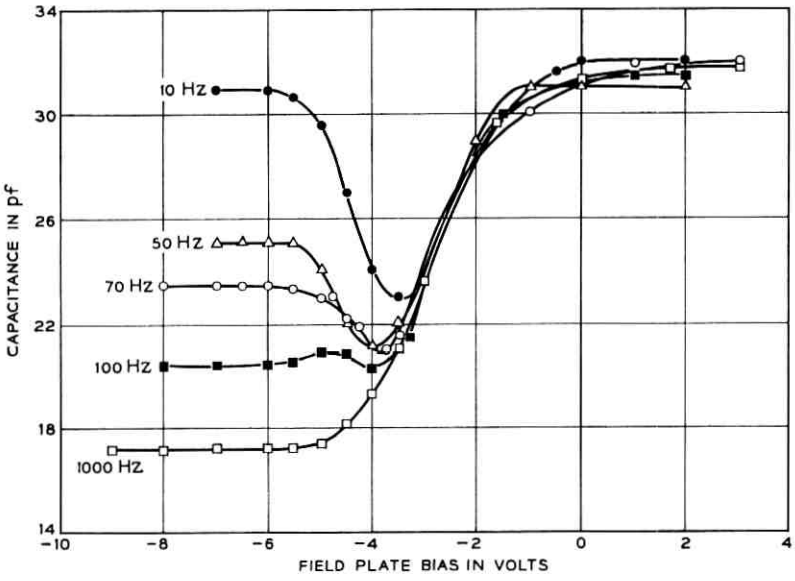


Fig. 8—Capacitance vs field plate bias measured at 27°C with frequency as parameter. Sample is the same as in Fig. 2.

technique applied here permits an estimate of the room-temperature time constant of an inversion layer by extrapolating the high-temperature curves. This way, even very long time constants may be estimated. In samples having low surface-state density, like the one described here, only bulk generation processes are important. The temperature range up to 140°C is characterized by space-charge generation, above this range diffusion current which has a higher activation energy becomes more important.

VI. ACKNOWLEDGMENT

We wish to thank R. V. Terio for his help in making the measurements.

REFERENCES

1. Lindner, R., Semiconductor Surface Varactor, B.S.T.J., *41*, 1962, p. 803.
2. Hofstein, S. R. and Warfield, G., Solid-State Electron, *8*, 1965, p. 321.
3. Nicollian, E. H. and Goetzberger, A., IEEE Trans., *ED-12*, 1965, p. 108.
4. Grey, P. V. and Brown D. M., Appl. Phys. Letters, *8*, 1966, p. 31.
5. Nicollian, E. H. and Goetzberger, A., to be published.
6. Goetzberger, A., J. Electrochem. Soc., *113*, 1966, p. 138.
7. Balk, P., Electrochem. Soc. (Extended Abstracts, Electronics Division), *14*, 1965, p. 237,
8. Goetzberger, A. and Nigh, H. E., Proc. IEEE, *54*, 1966, p. 1454.
9. Morin, F. J. and Maita, J. P., Phys. Rev., *94*, 1954, p. 1525.

The Charge-Control Concept in the Form of Equivalent Circuits, Representing a Link Between the Classic Large Signal Diode and Transistor Models

By DANKWART KOEHLER

(Manuscript received November 2, 1966)

It is shown in this paper that the charge-control concept can be conceived as a special form of the Linvill model for semiconductors. Instead of mathematical tools, charge-control models become equivalent circuits amenable to ordinary network analysis techniques. In the simplest form, the charge-control equivalent circuit for the junction transistor is fully equivalent to the Linvill and the Beaufoy-Sparkes model. For all practical purposes, it is also equivalent to the Ebers-Moll model.

The charge-control junction transistor equivalent circuit combines those features of the other models that are important for electrical engineering applications. It also permits the conversion between the three basic types of models. Because of its close relationship to the physical processes governing a device, it can readily be extended to higher-order phenomena. This is usually done by expressing a Linvill-type lumped model in terms of charge parameters. The charge-control equivalent circuit can be useful for modeling a variety of semiconductor devices.

I. INTRODUCTION

Three basic approaches are generally used to obtain descriptive large-signal models for transistors and diodes, the Ebers-Moll model,[1] the Linvill model[2] and the charge-control concept[3] after Beaufoy and Sparkes.

The *Ebers-Moll* transistor model[1,4] is based on the idea of superimposing a "normal" and an "inverse" transistor. Semiconductor junctions are represented by means of diodes and capacitors, whereas the properties of the transistor base are represented by frequency-dependent current sources. The Ebers-Moll transistor model is the

most popular of all transistor models since it lends itself most readily to simple "rule-of-thumb calculations." The current relations are described in the frequency domain, whereas the junction voltages are described as functions of current in the time domain, or, as in the original paper, only at dc. The model simulates only the effect which minority carrier storage exercises on the relations among the various device currents, but not the effect on current-voltage relations. Since the diode is a one-port device, no diode model of the Ebers-Moll type exists that could simulate carrier storage.*

The *Linvill* model[2,5-13] is almost a direct representation of the continuity and diffusion equations for semiconductor materials. It uses physical rather than circuit parameters and is superior to any other model when it comes to incorporating second-order physical effects or symbolizing new structures.

The *charge-control* concept[3,14-33] stands about halfway between physics and circuit considerations. It has proven in the past to be very useful for studying storage effects in diodes and transistors, but appeared to be entirely a mathematical tool. Certain equivalent circuits have been presented[14,15] to illustrate charge control, but, as Linvill phrased it, "they have little more meaning than a symbolic model useful for the purposes of visualizing only."

Hamilton, Lindholm and Narud compared the three models for the transistor in a well-written tutorial paper.[9,10] They discussed the approximations used in deriving each model from the same physical background. [See also Ref. 34] In contrast to this parallel treatment of the three models, the following study dwells on the interrelations and conversions between the various models. This is illustrated symbolically in Fig. 1.

We may call the Linvill model a physical model, the Beaufoy-Sparkes charge-control model a mathematical model, and the Ebers-Moll model an electrical model. The link between the three models is accomplished through a modified approach to charge-control theory: instead of deriving, from device physics by means of integration, mathematical charge-control expressions, the charge-control concept can be treated entirely as an equivalent circuit tool.[27] The transistor model, for example, is in such a form readily comparable with, and convertible into the Linvill and the Ebers-Moll model, provided all of these models are at the same level of approximation. In its simplest form, the charge-

* Diode models that simulate storage and use neither the charge control nor the Linvill concept are usually extensions of small-signal models towards incorporating certain nonlinear properties.

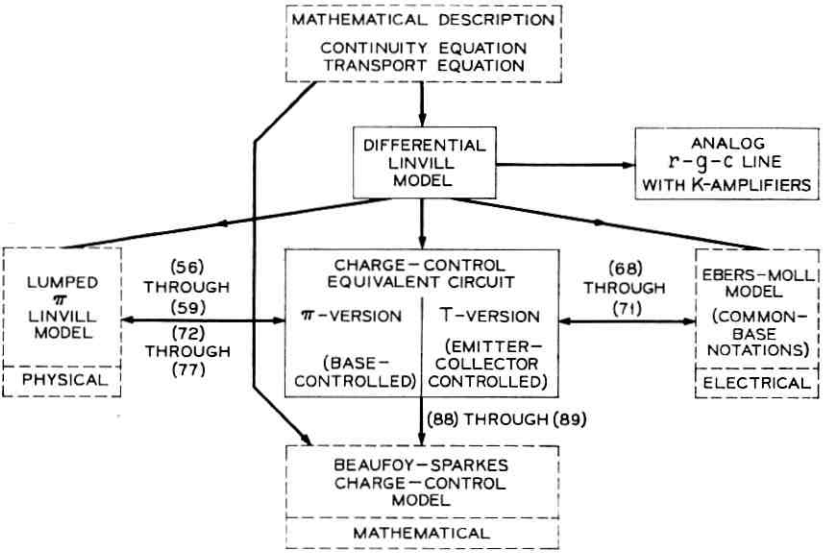


Fig. 1 — Principle of derivation of transistor models and their interrelations (heavy lines indicate main aspect of this paper; numbers refer to conversion equations in the text).

control equivalent circuit model is fully equivalent with the standard form of the Beaufoy-Sparkes charge-control model. But equivalency is usually lost, as extensions to higher-order approximations are made in each model.

In this paper, we shall review the derivation of the above-mentioned types of models for diodes and transistors. This will be done with the help of a differential transmission line model. The equivalent circuit type charge-control concept will then be derived for diodes and transistors. This will be followed by a discussion of higher-order approximations, the inclusion of drift fields, and possible applications to other semiconductor devices.

II. DIODE MODELS

2.1 *Mathematical Description*

As a starting point for our discussion it is assumed that the reader is familiar with the continuity and transport equations, describing current flow and carrier density in a semiconductor material.
Continuity equations

$$-\text{div } j_p(t) = e\frac{\partial p(t)}{\partial t} + e\frac{p(t) - p_0}{\tau_p} \tag{1a}$$

$$+\text{div } j_n(t) = e\frac{\partial n(t)}{\partial t} + e\frac{n(t) - n_0}{\tau_n}. \tag{1b}$$

Transport equations

$$j_p(t) = e\mu_p Ep(t) - eD_p \text{ grad } p(t) \tag{2a}$$

$$j_n(t) = e\mu_n En(t) + eD_n \text{ grad } n(t). \tag{2b}$$

j_p and j_n are the hole and electron current densities, respectively. p and n are the hole and electron carrier densities with p_0 and n_0 being their equilibrium values at a given temperature. E is the electric field intensity. D_p and D_n are the hole and electron diffusion constants, and μ_p and μ_n are the respective carrier mobilities. $e = +|\ e\ |$ is the value of the electronic charge.

2.1.1 *p-n Junction*

A p-n junction is described in a first-order approximation by the transport equation (2). The well-justified assumption is made that both j_p and j_n are numerically small compared with the mutually opposing diffusion and drift currents. With the help of the Einstein relations

$$D_p = \frac{kT}{e}\mu_p \tag{3a}$$

$$D_n = \frac{kT}{e}\mu_n \tag{3b}$$

and the appropriate boundary conditions one obtains the Boltzmann equations that express carrier densities as functions of the applied junction voltage v_{ext} :

$$p_n(0, t) = p_{n0} \exp\left[\frac{e}{kT} v_{ext}(t)\right] \tag{4a}$$

$$n_p(0, t) = n_{p0} \exp\left[\frac{e}{kT} v_{ext}(t)\right]. \tag{4b}$$

Here, $p_n(0,t)$ and $n_p(0,t)$ are the carrier densities on both sides of the junctions; p_{n0} and n_{p0} are the densities for $v_{ext} = 0$ or, in other words, at points away from the junction, previously called p_0 and n_0 in (1). The definitions of these notations are illustrated in Fig. 2.

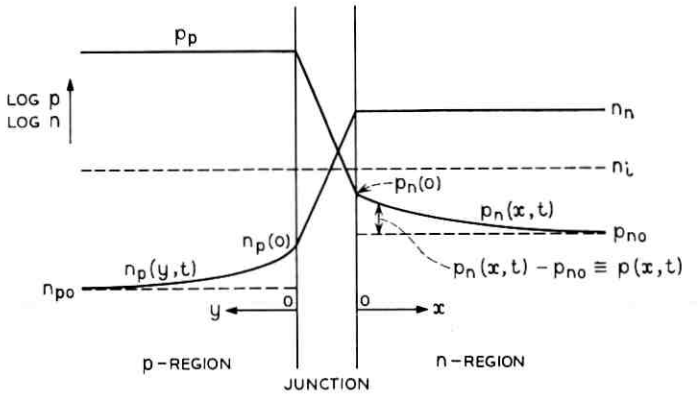


Fig. 2 — Carrier density distributions in the vicinity of a p-n junction and explanation of notations used.

In terms of excess carrier densities, (4) transform into the following expressions

$$p_{excess}(t) = p_n(0,t) - p_{n0} = p_{n0}\left[\exp\left\{\frac{e}{kT}v_{ext}(t)\right\} - 1\right] \tag{5a}$$

$$n_{excess}(t) = n_p(0,t) - n_{p0} = n_{p0}\left[\exp\left\{\frac{e}{kT}v_{ext}(t)\right\} - 1\right]. \tag{5b}$$

Together with the reasonable approximation that the hole and electron currents pass through the junction unchanged,* (5) uniquely characterizes the junction.

2.1.2 *p and n Regions*

The following assumptions are implied in the analysis presented for a p-n diode:

(*i*) The p-region is so heavily doped that the electron current can be neglected and appreciable carrier injection occurs only in the n-region.

(*ii*) The problem is reduced to one-dimensional variations along the x axis.

(*iii*) Drift fields are neglected. (Their inclusion will be briefly discussed later in Section 7.3.4.)

(*iv*) Space charge neutrality is assumed.

* This is not quite true for silicon diodes at low forward currents and in the reverse direction where recombination in the space charge layer cannot be neglected. With respect to some of the diode properties, especially the current-versus-voltage relationship, the discrepancy can be accounted for by changing the exponent to $ev_{ext}/2kT$.[11]

With these assumptions the continuity and transport equations reduce to

$$-\frac{\partial j_p(x,t)}{\partial x} = e\frac{\partial p_n(x,t)}{\partial t} + e\frac{p_n(x,t) - p_{n0}}{\tau_p} \tag{6}$$

$$j_p(x,t) = -eD_p\frac{\partial p_n(x,t)}{\partial x}. \tag{7}$$

We shall now express (6) and (7) in terms of the excess carrier densities $p_{n_{excess}}$ which we shall denote for simplicity as p, i.e.,

$$p(x,t) = p_{n_{excess}}(x,t) \equiv p_n(x,t) - p_{n0}\ .$$

Multiplying by the cross section A we obtain

$$-\frac{\partial i_p(x,t)}{\partial x} = eA\frac{\partial p(x,t)}{\partial t} + eA\frac{p(x,t)}{\tau_p} \tag{8}$$

$$i_p(x,t) = -eAD_p\frac{\partial p(x,t)}{\partial x}. \tag{9}$$

These are the two equations describing the n-region.

2.2 *Differential Diode equivalent circuits*

Equations (8) and (9) become transmission line equations if i_p and p are taken as currents and voltages, respectively. (Mathematically, one may think of p as an analog voltage representing carrier density.) Fig. 3 illustrates the resulting *r-g-c* transmission line.

The currents in the network branches are true currents but the voltages associated with the nodes are analog voltages. As a reminder, we have labeled the nodes with encircled "p's". The series and shunt elements are accordingly analog resistors, conductors and capacitors per unit length.

If the diode is forward biased, the junction injects carriers into the n-region. They diffuse across the n-region gradually recombining until, at $x \to \infty$, all hole current is converted into electron current. Fig. 4 shows the carrier distribution across the n-region which is equal to the voltage distribution along the infinitely long *r-g-c* line. It can be derived easily from (8) and (9) that, under steady-state conditions, the shape of the charge distribution is proportional to

$$\exp(-x/L_p),$$

where

$$L_p = \sqrt{D_p\tau_p}\ . \tag{10}$$

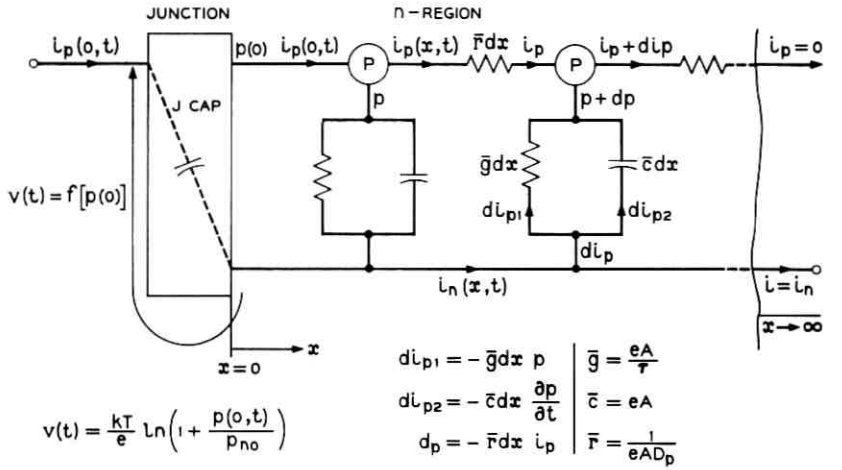


Fig. 3 — Analog differential transmission line representation of diode model. (The bars indicate dimension per unit length.)

L_p is called the diffusion length. τ_p is the hole recombination time constant, or hole "lifetime".

Since the analog voltage distribution on the capacitors of the *r-g-c* transmission line is identical with the physical charge density distribution, and since many engineers have a much better feel for the charging and discharging processes of such a line than for the physical process, the *r-g-c* line representation may be quite helpful as an illustration of the carrier injection process. In early semiconductor work, such *r-g-c* transmission lines were frequently used.[26,35,36,37] No attempt was made, however, to attribute the physical meaning of carrier density to the network nodes; the junctions were represented by so-called *K*-amplifiers. These amplifiers transform the internal voltage at $x = 0$

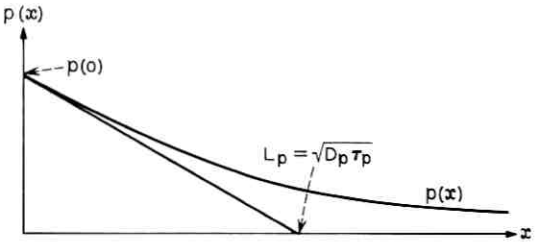


Fig. 4 — Excess carrier distribution in diode n-region.

to the external voltage with the appropriate exponential relationship, while not transforming current at all.

Linvill[2,5,6,7] has introduced new symbols for the network elements which relate current to carrier density. These new notations avoid possible confusion between analog and physical circuit parameters, especially voltages, and hence enable us to combine current/carrier-density with current/voltage networks. Fig. 5 shows such a Linvill model in differential form. Again we have added bars over the letters as it was done with the $\bar{r}$, $\bar{g}$, and $\bar{c}$ in Fig. 3 to denote their dimensions as being "units per length".

The symbols in the models are defined as follows:

$$di_{p1}(x,t) = -\overline{H_c}\, dx\, p(x,t) \tag{11a}$$

$$di_{p2}(x,t) = -\overline{S}\, dx\, \frac{\partial p(x,t)}{\partial t} \tag{11b}$$

$$dp(x,t) = -\overline{(1/H_d)}\, dx\, i_p(x,t), \tag{11c}$$

where

$$\overline{H_c} = \text{combinance per length} = eA/\tau_p \tag{12a}$$

$$\overline{S} = \text{storance per length} = eA \tag{12b}$$

$$\overline{(1/H_d)} = \frac{1}{\text{diffusance}} \text{ per length} = 1/eAD_p\,. \tag{12c}$$

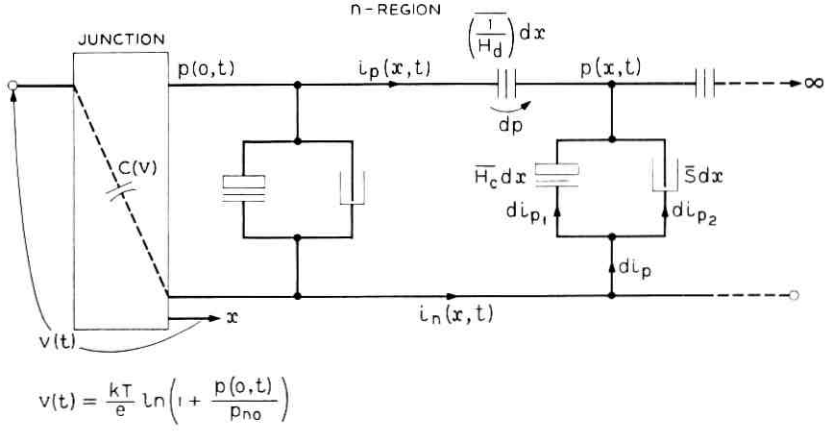


Fig. 5—Differential Linvill diode model. Note that, in consistency with common transmission line notations, the reciprocal of diffusance must be used.

This model can be extended to include majority carriers, drift fields etc. The reader is referred to the literature.[2,5–13]

2.3 *Integrated Diode Models*

2.3.1 *Mathematical Integration*

In order to arrive at an expression for the external diode current from the continuity equation (8) we can integrate this expression with respect to x. Choosing $x = 0$ and $x = \infty$ as the limits of integration, we obtain

$$-\int_0^\infty \frac{\partial i_p(x,t)}{\partial x}\,dx = \int_0^\infty eA\,\frac{\partial p(x,t)}{\partial t}\,dx + \frac{1}{\tau_p}\int_0^\infty eAp(x,t)\,dx. \qquad (13)$$

The third integral represents the total charge in the bulk material. With the appropriate boundary conditions $i_p(0) = i$, $i_n(0) = 0$, $i_p(\infty) = 0$, $i_n(\infty) = i$, the well-known charge-control equation[3] can readily be obtained as

$$i(t) = \frac{dq(t)}{dt} + \frac{q(t)}{\tau_p}. \qquad (14)$$

To obtain (14) from (13) the assumption must be made that A and τ_p are constant. Note that no approximations or restrictions to specific charge distributions are implied in (14). (They must be made, however, when relating the current to the junction voltage.)

2.3.2 *Lumped Linvill Diode Model*

The crudest approximation to the distributed Linvill model of Fig. 5 is to replace the "line" by just one storance and one combinance[2,7] as shown in Fig. 6. These two elements are obtained by summing, i.e.,

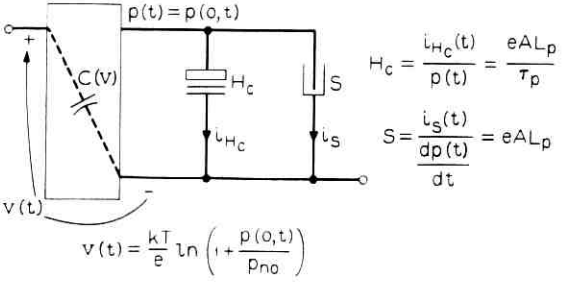


Fig. 6—Lumped Linvill diode model showing single-pole approximation for minority carrier storage. Chosen values: $\Delta x = L_p$, $p(t) = p(0,t)$.

integrating, all differential storances and combinances from $x = 0$ to some value Δx. The value of Δx is usually chosen to equal the diffusion length L_p . This may seem arbitrary,[12] but has no effect on the terminal properties of the first-order model, as long as $p(t)$ is chosen such as to maintain the same amount of total charge.

The values of the circuit elements follow from (11), (12), and (5a) as

$$H_c = \overline{H_c}\,\Delta x = \overline{H_c} L_p = \frac{eAL_p}{\tau_p} \tag{15a}$$

$$S = \overline{S}\,\Delta x = \overline{S} L_p = eAL_p \tag{15b}$$

$$v(t) = \frac{1}{\lambda} \ln\left(1 + \frac{p(0,t)}{p_{n0}}\right), \tag{15c}$$

where λ is an abbreviated notation, used hereafter for

$$\lambda = \frac{e}{kT}. \tag{16}$$

The meaning of such lumping with respect to the carrier distribution is illustrated in Figs. 7 and 8. The solid lines in Fig. 7 present the actual carrier distribution in a switching example in which a current pulse is assumed. As required by the transport equation, the slope at $x = 0$ is, at any time, proportional to the current. Under steady-state conditions, an exponential distribution is obtained. To assume such exponential distributions at any instant of time (dashed lines in Fig. 7) represents a simplifying assumption. The corresponding

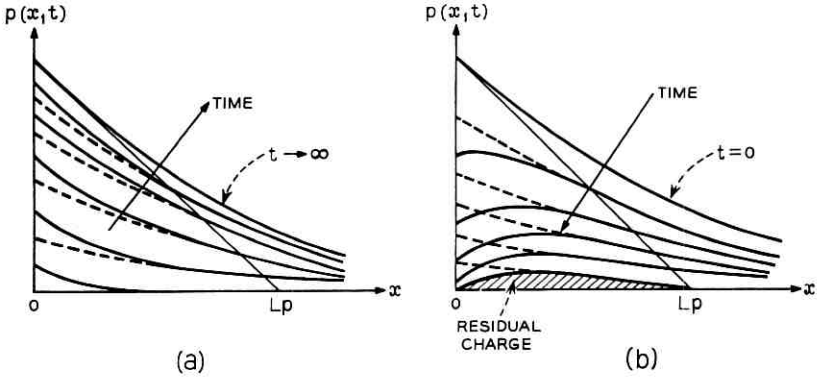


Fig. 7 — Illustration of the (a) charging and (b) discharging process in the neutral bulk material. The applied signals are assumed to be forward and reverse current pulses. The solid lines represent the actual shape for current pulse drive; the dashed lines represent exponential model approximations.

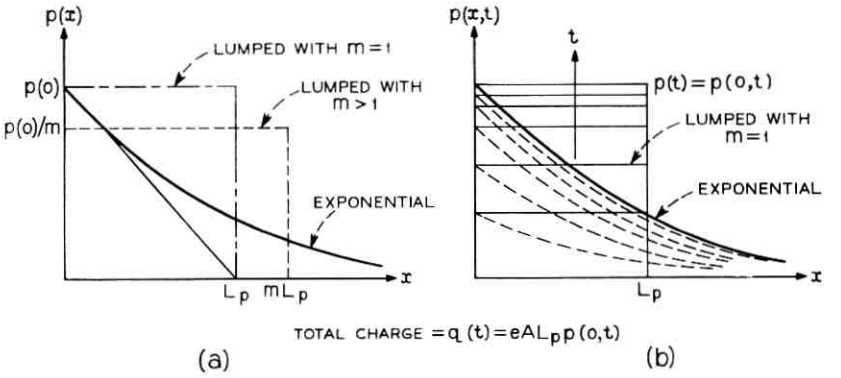


Fig. 8—Exponential and corresponding lumped distribution of excess minority carriers in the bulk material of a diode. (a) Illustration of the choice of lumping length. (b) Time variation for $m = 1$; $m = 1$ is generally preferred in the Linvill model, and is irrelevant in charge models or circuit applications of the Linvill model.

errors are negligible in all those applications where the switching times are large compared with the carrier redistribution times (= diffusion times τ_a and τ_b in Fig. 12).

In the lumped Linvill model, it is assumed that the carrier density is, at any instant of time, constant from $x = 0$ to $x = L_p$, and that it is 0 for all $x > L_p$. Any information on the distribution of the charge, especially of the slope at $x = 0$, as expressed in the transport equation, has been lost since all series elements (diffusances) are neglected. The only parameter of importance left is the total number of minority carriers and hence, the total charge. The approximation used is therefore equivalent to the dashed line exponential distribution in Fig. 7.

As mentioned above and illustrated in Fig. 8(a), the length Δx over which p is nonzero, is most conveniently chosen to equal L_p. But it is permissible to choose $\Delta x \neq L_p$ if the constant value $p(x)$ is recognized to be different from $p(0,t)$; for $\Delta x = mL_p$, we must choose $p(t) = p(0,t)/m$ such as to yield the same total charge

$$q = eAL_p p(0,t). \tag{17}$$

Fig. 8(b) shows, for $m = 1$, the time variation of the carrier distributions for the lumped model (solid lines) and the exponential distribution (dashed lines).

As the external voltage $v(t)$ varies, the carrier density $p(0,t)$ changes accordingly. The relation between $v(t)$ and $p(0,t)$ has been given above in (15c). We shall see below that the approximation made in the lumping process, as discussed above, effects only $v(t)$ but not the current. Little

or no error is made whenever, and as long as the external driving source impedance is large.

Since, in this and any other lumped Linvill model, all circuit parameters are functions of total charges in the various sections of the semiconductor we can transform the model of Fig. 6 into a charge-controlled model, in which all carrier densities are replaced by charges. This will be shown in Section 2.4.1.

2.4 *Charge-Control Diode Model*

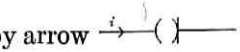
Without invoking any of the approximations introduced in Section 2.3.2 and illustrated in Figs. 7 and 8, the diode is completely described by (14) and (5a). These two equations are the basis for the classic charge-control theory after Beaufoy-Sparkes as applied to diodes. Throughout the charge-control literature, only the current appears as a function of the total charge but not the voltage. If we want to relate the junction voltage to the charge rather than to $p_n(0,t)$ as it was done in (5a), we must make some approximation: The simplest possible approximation is the assumption that $p(0,t)$ is proportional to $q(t)$. This is, for example, satisfied if the shape $f(x)$ of the carrier distribution never changes, i.e., if the carriers redistribute themselves instantaneously. $p(x,t)$ is then of the form

$$p(x,t) = f(x)g(t).$$

The shape of $f(x)$ does not matter as long as the integral $\int f(x)\,dx$ yields the proper proportionality constant. Examples of this are the exponential distribution or the lumped distribution (with any arbitrary value of m) in Fig. 8(a). This shows the equivalency between the postulate of instantaneous carrier redistribution in classic charge-control theory and carrier density lumping in the first order Linvill model.

If we now want to establish an equivalent charge-control circuit we must first represent (14) by corresponding circuit symbols. This is done in the n-region part of Fig. 9. S is the store originally introduced by Beaufoy and Sparkes.[11,12] To account for directionality, we have added a vertical bar to the store symbol in the manner of the standard diode symbol. The properties of S, as defined in this paper, are:

(*i*) charge stored = $q(t)$

(*ii*) current in direction indicated by arrow 

$$i = \frac{dq(t)}{dt}$$

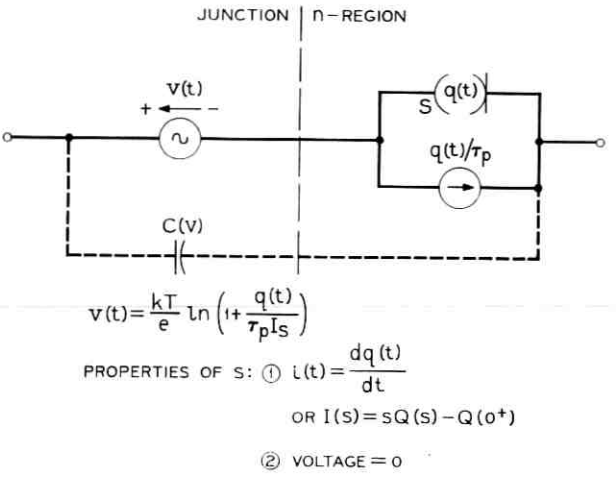


Fig. 9—Complete first-order charge-control equivalent circuit (this circuit is a charge representation of Fig. 6).

(*iii*) voltage across store = 0.

S is often interpreted as an infinite capacitor for which $i = dq/dt = d(Cv)/dt$ = finite, but $C \to \infty$ and $v \to 0$.

It follows from (15c) and (17) that the junction voltage is of the form

$$v(t) = \frac{1}{\lambda} \ln [1 + Kq(t)],$$

where K is a proportionality factor. If we denote the steady-state reverse current (flowing through the diode when $v(t)$ is very large and negative) by I_S we can evaluate the constant: For $v \to -\infty$ we obtain

$$K \cdot Q = -1$$

and from Fig. 9

$$-I_S = Q/\tau_p .$$

Hence,

$$K = \frac{1}{I_S \tau_p}$$

and thus,

$$v(t) = \frac{1}{\lambda} \ln \left[1 + \frac{q(t)}{I_S \tau_p} \right]. \tag{18}$$

Under steady-state conditions where $Q/\tau_p = I$, this equation becomes the well-known diode equation.

For the possibilities of incorporating the junction capacitances see the discussion in Section 7.3.2.

2.4.1 *Derivation of the Charge-Control Model from the Lumped Linvill Model*

It represents not merely an additional proof of equivalency but also a good preparation for the derivation of more complex models, if we show[27] that we can derive the charge-control model from the Linvill model. A somewhat related modification of the Linvill model was more recently proposed by Beddoes.[12] To this end we calculate the currents through the elements H_c and S in Fig. 6:

$$i_{H_c}(t) = H_c p(0,t) \tag{19a}$$

$$i_S(t) = S \frac{dp(0,t)}{dt}. \tag{19b}$$

Substitution of the values for H_c, S, and $p(0,t)$ from (15a), (15b), and (17) yields

$$i_{H_c}(t) = \frac{eAL_p}{\tau_p} p(0,t) = \frac{q(t)}{\tau_p} \tag{20a}$$

$$i_S(t) = eAL_p \frac{dp(0,t)}{dt} = \frac{dq(t)}{dt}. \tag{20b}$$

This result shows that the current source in the charge-control model of Fig. 9 represents the current i_{H_c} through the combinance, and that the store S represents the current i_S through the storance.

To find the expression for the junction, we can express $p(0,t)$ in terms of $q(t)$ by means of (17). p_{n0} can again be obtained from the case, where $V \to -\infty$, and where $p(0,t) = P(0) = -p_{n0}$:

$$I_{H_c}]_{V\to-\infty} = -I_S = \frac{eAL_p}{\tau_p} P(0) \Big]_{V\to-\infty} = -\frac{eAL_p}{\tau_p} p_{n0}.$$

Thus, we find

$$\frac{p(0,t)}{p_{n_0}} = \frac{q(t)}{eAL_p} \Big/ \frac{I_S \tau_p}{eAL_p} = \frac{q(t)}{I_S \tau_p}. \tag{21}$$

With this we can make the transition from (15c) to (18).

2.4.2 *Evaluation of the Charge-Control Model*

The charge-control model is completely equivalent with the lumped Linvill model in Fig. 6; in fact, it may be considered a circuit oriented

form of the Linvill model. In almost all instances[7,11] where the Linvill model is being used for circuit applications the conversion of carrier density into charge must be made anyhow. The charge-control equivalent circuit in Fig. 9 uses current and voltage sources plus one lesser known circuit element described by the simple relations

$$v(t) = 0 \tag{22a}$$

$$i(t) = \frac{dq(t)}{dt} \tag{22b}$$

or, in Laplace notation

$$I(s) = sQ(s) - Q(0^+). \tag{22c}$$

Ordinary circuit analysis techniques can be used in working with the model. No restriction exists with respect to the external waveforms. Q appears as an additional circuit parameter with additional complexity comparable to that of an additional branch current. From a topological viewpoint it is a branch current. This is the price to be paid for inclusion of the first-order dynamic storage properties.

Junction and n-region are clearly separated in the model. Thus, little difficulty should arise in adding junction capacitors (dashed in Fig. 9), series path resistors, and leakage resistors, provided, physical knowledge of such effects exist.

2.4.3 *Charge-Control Model for Short-Base Diodes*

Diodes with extremely short bases do not show the exponential minority carrier distribution represented in Fig. 4, but rather a practically linear fall-off (like in a transistor base except that the collector is now a nonrectifying contact). With reference to Figs. 3 or 5, this means that the distributed "transmission line" is so short that the effect of the series diffusances H_d dominates over that of the shunt combinances H_c. The metallic contact behaves like a short circuit at the end of the line.

The analogy with the *r-g-c* line of Fig. 3 may help the reader visualize the difference between the long base and the short base diode: The first-order approximation for the infinitely long line with respect to currents and input voltage is the parallel connection of the *shunt* resistor and the shunt capacitor; the first-order approximation for a very short line is the parallel connection of the *series* resistor and the shunt capacitor. In terms of the Linvill model, the short base diode model is obtained by replacing H_c in Fig. 6 by $H_d = eAD_p/L$ and

L_p by L. (Note that H_d increases as L becomes small.) In the charge control model, of Fig. 9, the term τ_p which represents the recombination time constant for the long base diode, now becomes the diffusion time constant. The new value τ_p' can be derived most easily from the Linvill model as follows:

$$\tau_p' = \frac{Q}{i_{H_d}} = \frac{Q}{p(0)H_d/W} = \frac{\frac{1}{2}p(0)eAW}{p(0)eAD_p/W} = \frac{W^2}{2D_p}. \tag{23}$$

Apart from this numerical change, the model in Fig. 9 for the normal diode is equally valid for the short base diode.

2.4.4 *Piecewise Linear Charge-Control Diode Model*

For many practical purposes the logarithmic voltage source relation can be approximated by a switch as illustrated in Fig. 10. The switch opens when q becomes negative and closes when q is able to charge up to $q > 0$. A threshold voltage V_{th} is connected in series with the forward path. If desired, the slope of the logarithmic curve

$$\frac{dV}{dI} = \frac{dV}{d(q/\tau)} = \frac{\tau}{\lambda q} = \frac{\tau}{\lambda q_{\text{average}}} = \frac{1}{\lambda I_{\text{av}}} \approx \frac{2}{\lambda I_{\text{Max}}} \tag{24}$$

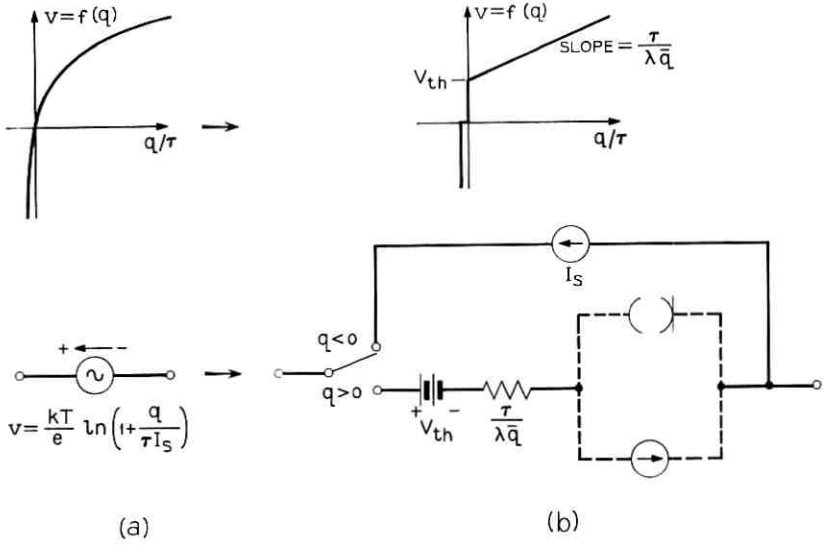


Fig. 10—Piecewise linear approximation for the semiconductor junction. (a) Theoretical logarithmic curve. (b) Approximated curve (the dashed lines indicate the completion of the diode model).

can be added as a resistor, where I_{av} is an average current, which may, in long hand calculations, be assumed to be $\frac{1}{2}I_{max}$. The saturation current I_S must now be represented by an external current source.

2.4.5 *Application of the Model*

The above discussion of diode models serves two purposes: First, they form a basic understanding for deriving transistor models. Secondly, the diode models can be very useful in simulating dynamic effects due to carrier storage in diodes.

With the piecewise linear junction approximation of Fig. 10 applied to the charge-control model in Fig. 9, storage time equations can be derived easily using Laplace transform concepts. The model has proven to be very useful in the analysis of step-recovery diode circuits. In the piecewise linear form, it can be handled without a computer, whereas, for the more complex models with various parasitics added, computers soon become mandatory.

Switching times for *step-recovery diodes* are derived in Appendix A.1 as an example of the use of the charge-control model. The equations obtained have been found by many authors to agree well with actual measurements. The normalized storage time for recovery from an infinite ON-pulse according to (97) is plotted in curve *a* of Fig. 11 as a function of the reverse-to-forward current ratio according to the relation

$$T = \tau \ln\left(1 + \frac{I_F}{I_R}\right). \tag{25}$$

When applying this result to an *ordinary diode* with homogeneous doping profile, one must be aware of the implied approximations: (*i*) The single-section approximation in the model does not affect any mutual relationships between currents and charges, but represents approximations with respect to the junction voltage. As the amount of stored charge is reduced considerably in the diode, the junction voltage decreases noticeably. (*ii*) As the carrier density near the junction becomes extremely small, the voltage reverses sign and the diode impedance, at some point, becomes comparable with the external source impedance. The ideal current source assumed in (97) ceases to exist, and instead of the step-recovery, as given by the model, a long tail in the current response results.

From either one of the two differential models in Figs. 3 or 5, we can calculate the time in which the carrier density at $x = 0$, and hence the junction voltage, reaches zero. Such a calculation yields the relation

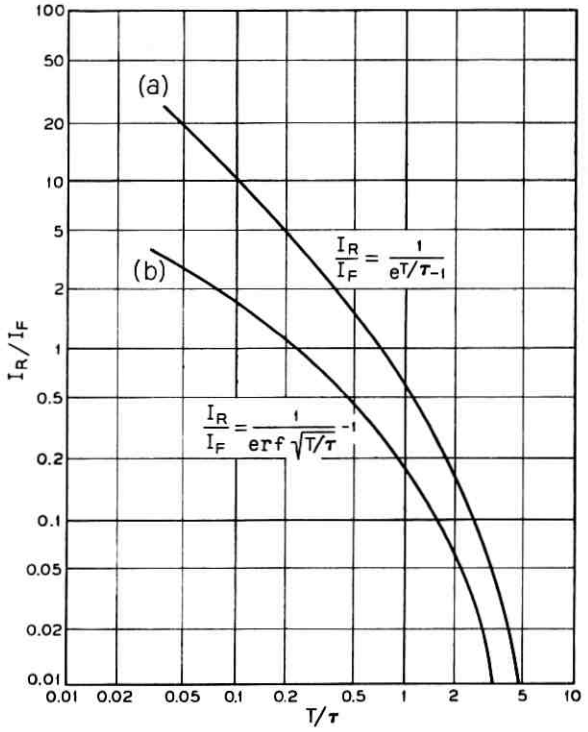


Fig. 11—Comparison of diode storage times as functions of the driving ratios. (a) Single lump model; T = time when charge is fully depleted. (b) Differential model; T = time when excess carrier density at $x = 0$ reaches zero.

originally derived by Lax and Neustadter[15]

$$\text{erf}\sqrt{\frac{T}{\tau}} = \frac{1}{1 + \frac{I_R}{I_F}}. \tag{26}$$

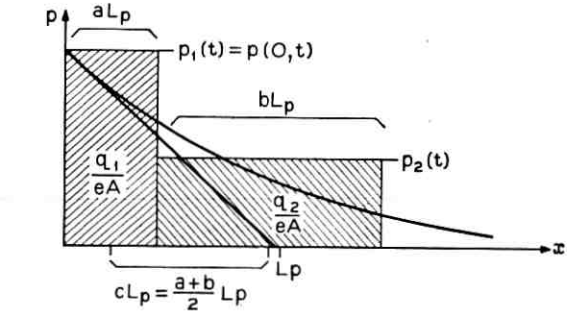
This relationship is illustrated in Fig. 11(b). Since curve (b) represents only the storage phase but not the very long tail of the recovery, the values are much smaller than those in curve (a) in which some sort of "effective total recovery" is represented. The difference is most remarkable at strong relative reverse drives where the carrier distributions on the lines differ most from the steady-state distributions.

If it becomes necessary to incorporate the tail of the recovery into a lumped diode model, the double π-extension described below may prove adequate for most applications.

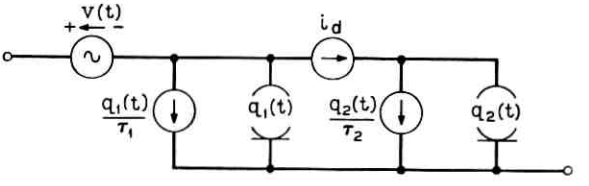
2.4.6 *Higher-Order Approximations*

Bearing in mind how the model originated as an approximation to the differential transmission line or as just another form of the lumped Linvill model we can now understand how higher-order approximations are to be obtained.

Fig. 12 shows the example of a π-approximation for a diode. The



(a)



WHERE $i_d(t) = \dfrac{q_1(t)}{acL_p^2/D_p} - \dfrac{q_2(t)}{bcL_p^2/D_p} \equiv \dfrac{q_1(t)}{\tau_a} - \dfrac{q_2(t)}{\tau_b}$

$$v(t) = \frac{kT}{e} \ln\left(1 + \frac{q_1(t)}{\tau_v I_s}\right)$$

τ_v (FROM DC CONSIDERATIONS) $= \dfrac{\tau_1 \tau_2}{\tau_1\left(\frac{1}{a} - 1\right) + \tau_2}$

CHARGE CONSERVATION CONSTRAINT ON a, b, c:

$$\frac{cb(1-a)L_p^2}{(a+b-1)D_p\tau_2} = 1$$

WHERE c IS MOST APPROPRIATELY CHOSEN TO BE $c = \dfrac{a+b}{2}$

(b)

Fig. 12 — Higher-order, π-Approximation of diode charge-control model. (a) Charge approximation. (b) Corresponding model.

charge is broken up into two parts q_1 and q_2. The diffusance between the two stores controls the redistribution of the charge. Such a structure provides a better representation of the junction at the higher frequencies or at higher speeds than the model of Fig. 9, since the junction voltage is now a function of only that part of the total charge which is close to the junction. The model simulates recovery tails. It also permits the simulation of variations in recombination time along the x-axis. Fig. 12 assumes two different recombination times τ_1 and τ_2. The $i = f(q)$ relation then becomes

$$i = \frac{d(q_1 + q_2)}{dt} + \frac{q_1}{\tau_1} + \frac{q_2}{\tau_2} \tag{27}$$

[which reduces to (14), if one assumes $\tau_1 = \tau_2$].

Three additional time constants τ_a, τ_b and τ_s appear in Fig. 12. They depend on the choice of the sections aL_p and bL_p over which the shunt elements are integrated and on the choice of the section cL_p over which the diffusances are integrated. The three degrees of freedom reduce to one, however, if one considers that (*i*) the total charge must be conserved by the lumped approximation, and (*ii*) in a multisectional approximation the diffusances are most appropriately lumped over sections cL_p which extend between the centers of the charge sections. The corresponding relations are given in Fig. 12; derivations have been omitted.

III. LARGE-SIGNAL TRANSISTOR MODELS

In complete analogy to the diode models, we shall now compare the various junction transistor models and establish the charge-control model in the form of equivalent circuits. The Ebers-Moll concept, which was found not to be applicable to dynamic diode description, will now enter the "competition".

In order to dwell on the philosophies underlying each concept we shall, at first, limit ourselves to diffusion type junction transistors, neglecting again drift currents and secondary effects such as base-width modulation. All derivations will be carried out for pnp transistors; but, of course, everything will be correspondingly valid for npn transistors.

3.1 *Differential Transistor Model*

The most rigorous of all the equivalent circuits describing a junction transistor, as defined by (5), (8), and (9), is the differential model shown

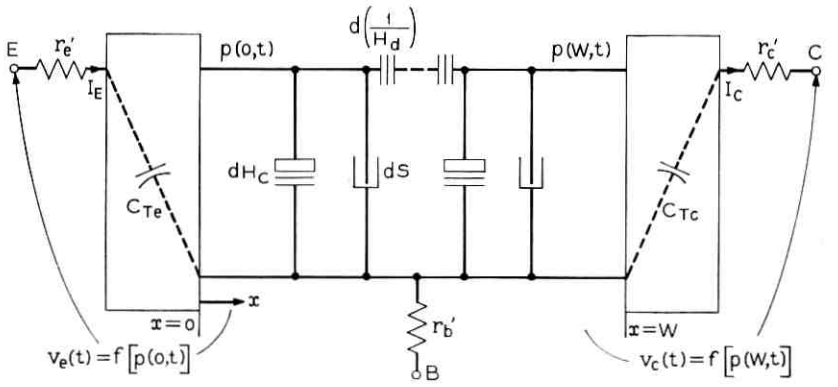


Fig. 13 — Differential Linvill model for the transistor with drift fields neglected, *pnp* version shown.

in Fig. 13. Linvill notations comparable to the diode model in Fig. 5 were chosen. (If so wanted, the model could also be drawn with the notations used in Fig. 3 resulting in an *r-g-c* line and two *K*-amplifiers at both ends.)

The base section of the transistor model is only a very short "transmission" line when compared with the "infinitely long" diode n-region of the normal diode. Instead of 100 percent recombination, as found in the diode, the transistor must have as little recombination as possible in order to achieve high gain. Fig. 14(a) shows a steady-state charge distribution under normal forward operation, and Fig. 14(b) shows the distribution for the case where both junctions are emitting, i.e.,

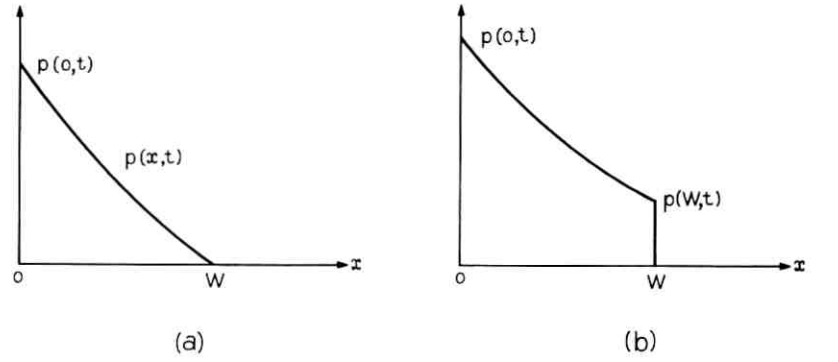


Fig. 14 — Excess minority carrier distribution in the transistor base under (a) normal and (b) saturated operation.

in saturation. Under normal operation the collector acts as a "charge short circuit" for the line. For high-gain units, the slope is almost a straight line; at $x = 0$ it is proportional to I_E and at $x = W$ proportional to I_C.

The general case is that of Fig. 14(b) where both junctions are emitting and $p(W) \neq 0$. Any section of the base region can be described analogously to a four-pole using the definitions given in Fig. 15. Note that there are no nonlinearities in the base section.

$$I_E(s) = A_{11}P_1(s) + A_{12}P_2(s) \tag{28a}$$

$$I_C(s) = A_{21}P_1(s) + A_{22}P_2(s). \tag{28b}$$

By using complete analogy to standard transmission line theory, it can be shown that with the use of (10) and (11) one obtains for a homogeneous section Δx

$$I_E(s) = \frac{P_1(s)}{Z} \coth \gamma \, \Delta x - \frac{P_2(s)}{Z} \operatorname{cosech} \gamma \, \Delta x \tag{29a}$$

$$I_C(s) = \frac{P_1(s)}{Z} \operatorname{cosech} \gamma \, \Delta x - \frac{P_2(s)}{Z} \coth \gamma \, \Delta x, \tag{29b}$$

where

$$Z = \frac{1}{eA} \sqrt{\frac{\tau}{D_p}} \sqrt{\frac{1}{1 + s\tau}} \tag{30}$$

$$\gamma = \sqrt{\frac{1}{D_p\tau}} \sqrt{1 + s\tau}. \tag{31}$$

In the general case, the base is not homogeneous, which means that Z and γ will vary along the line.

The junctions are described by the time relations

$$p_1(0,t) = p_{n0}[\exp \{\lambda v_e(t)\} - 1] \tag{32a}$$

$$p_2(W,t) = p_{n0}[\exp \{\lambda v_e(t)\} - 1]. \tag{32b}$$

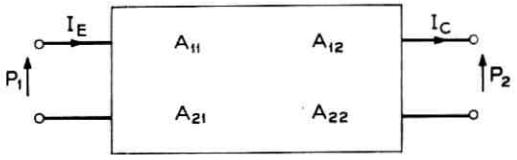
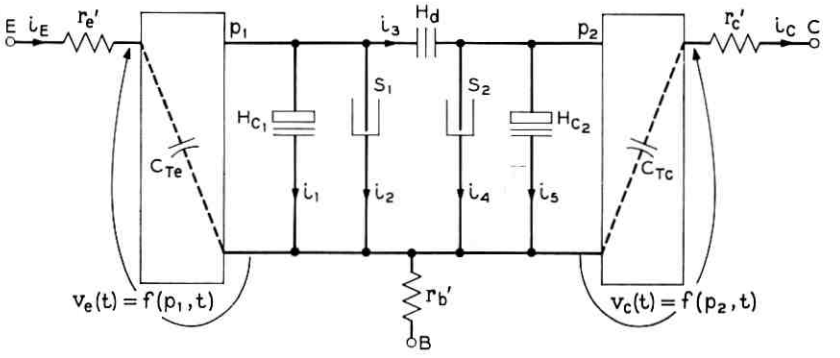


Fig. 15—Symbols and polarity conventions defining the four-pole description of the transistor base.



$$v_e(t) = \frac{kT}{e} \ln\left(1 + \frac{p_1(t)}{p_{no}}\right) \qquad H_{C_1} = \frac{i_1(t)}{p_1(t)} = \frac{eA_1W_1}{\tau_1} \qquad S_1 = \frac{i_2(t)}{dp_1(t)/dt} = eA_1W_1$$

$$v_c(t) = \frac{kT}{e} \ln\left(1 + \frac{p_2(t)}{p_{no}}\right) \qquad H_{C_2} = \frac{i_5(t)}{p_2(t)} = \frac{eA_2W_2}{\tau_2} \qquad S_2 = \frac{i_4(t)}{dp_2(t)/dt} = eA_2W_2$$

$$H_d = \frac{i_3(t)}{p_1(t) - p_2(t)} = \frac{eAD_p}{W}$$

Fig. 16 — First-order lumped Linvill transistor model.

IV. THE LUMPED LINVILL π-MODEL

In the literature, any lumped approximation to the transmission line model presented in Fig. 13 is referred to as a "lumped Linvill model" or simply "lumped model". The most common form is the π-model. With respect to its current properties and steady-state voltages, this form will prove to be equivalent to the commonly known "Ebers-Moll model", if both models are taken to be in the form of first-order approximations.

By integrating all differential diffusances over a length $\Delta x = W$, all emitter sided differential combinances and storances over a length $\Delta x = W_1$, and all collector sided differential combinances and storances over a length $\Delta x = W_2 (= W - W_1)$ one obtains the circuit shown in Fig. 16. Nonsymmetry has been taken into account by using different recombination times τ_1 and τ_2 and different cross-sectional areas A_1 and A_2 on the two sides. Note that the latter represents an extension from the one-dimensional carrier flow and as such an example for the reduction of multidimensional effects to a one-dimensional model. Area A is some average cross section effective for the diffusion process. H_d is the diffusance, the H_c's are the combinances, and the S's are the two storances.

The four-pole equations describing the transistor base are obtained as

$$I_E(s) = \frac{eAD_p}{W} P_1(s)\left[1 + \frac{A_1WW_1}{A\tau_1 D_p}(1 + s\tau_1)\right] - \frac{eAD_p}{W} P_2(s)$$

$$= H_d P_1(s)\left[1 + \frac{H_{c1}}{H_d}\left(1 + s\frac{S_1}{H_{c1}}\right)\right] - H_d P_2(s) \tag{33a}$$

$$I_C(s) = \frac{eAD_p}{W} P_1(s) - \frac{eAD_p}{W} P_2(s)\left[1 + \frac{A_2WW_2}{A\tau_2 D_p}(1 + s\tau_2)\right]$$

$$= H_d P_1(s) - H_d P_2(s)\left[1 + \frac{H_{c2}}{H_d}\left(1 + s\frac{S_2}{H_{c2}}\right)\right]. \tag{33b}$$

The junctions are described as

$$p_1(t) = p(0,t) = p_{n0}[\exp\{\lambda v_e(t)\} - 1] \tag{34a}$$

$$p_2(t) = p(W,t) = p_{n0}[\exp\{\lambda v_c(t)\} - 1]. \tag{34b}$$

A constraint has to be satisfied: Under equilibrium conditions, the total charge in the base must equal that in the two sections, i.e.,

$$eA_1W_1P_1 + eA_2W_2P_2 = eA\int_0^W P(x)\,dx \approx \tfrac{1}{2}eAW[P(0) + P(W)]. \tag{35}$$

The approximation holds for high-gain units. For this case the base volume sections are equal, i.e., $A_1W_1 = A_2W_2 = \frac{1}{2}AW$. For low-gain units (34) must be modified: The terms $p_1(t)$ or $p_2(t)$, or both, must be replaced by $p_1(t)/m_1$ and $p_2(t)/m_2$, respectively, whereby the m's are constants >1, similar to m in Fig. 8.

Equation (33) represents one of several possible approximations to (29) with the additional property of nonsymmetry being added.

Higher-order approximations of a lumped linear model are obtained by representing the base of width W by more than the two sections W_1 and W_2.

V. THE EBERS-MOLL TRANSISTOR MODEL

The focal point of the Ebers-Moll model is the two-port description of the base. Such a description has been given in (28) and (29), and is permissible because of the linearity which exists between currents and carrier densities in the base. Nonlinearity exists, however, in the relationship between carrier densities and external voltages according to (32). Since linearity allows the use of the superposition principle,

the total current can be conceived as consisting of the superimposed contributions of the currents injected by the two junctions.

When put into the form of an equivalent circuit, the Ebers-Moll model shows the superposition of a normal transistor (subscript N) and an inverse transistor (subscript I). In Fig. 17(a) the lower diode and current source represent the normal transistor and the upper elements represent the inverse transistor. Each junction is represented by a diode, a fraction of the diode current is collected by the other electrode. The ratios of collected currents to emitted currents are called α_N and α_I for normal and inverse operation, respectively. The general frequency behavior of the α's can be calculated for a homogeneous base from (29), (30), and (31) as

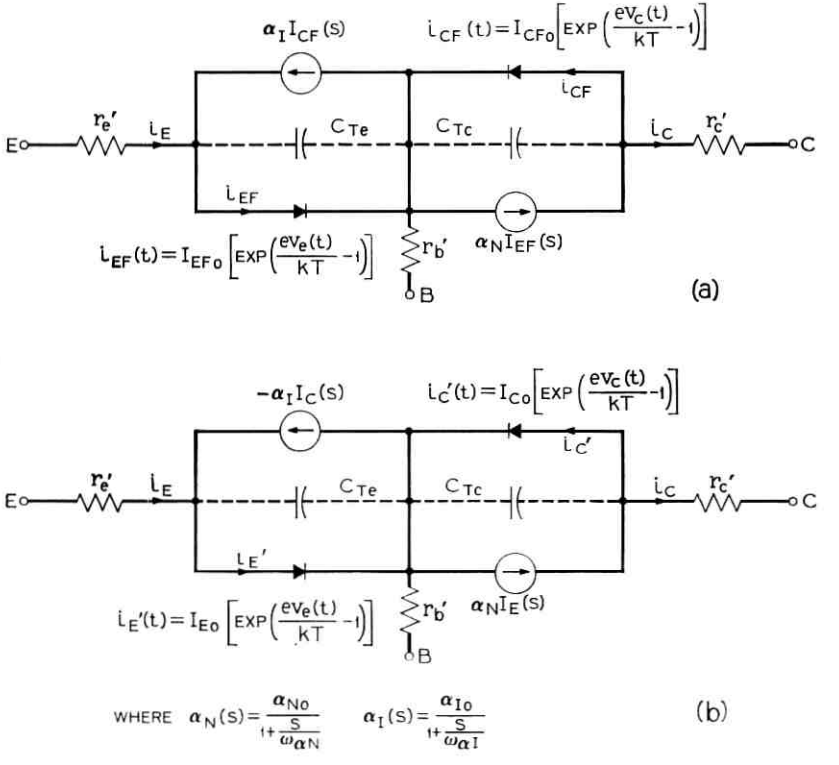


Fig. 17 — The two forms of the Ebers-Moll transistor model: (a) direct representation of the idea of superimposing a normal and an inverse transistor, (b) collecting current sources as functions of the electrode currents. The junction saturation currents in (b) are identical with the open-electrode diode saturation currents.

$$\alpha(s) = \frac{I_{\text{output}}(s)}{I_{\text{input}}(s)}\Bigg]_{P_{\text{output}}=0} = \frac{\operatorname{cosech} \gamma\, \Delta x}{\coth \gamma\, \Delta x} = \frac{1}{\cosh \gamma\, \Delta x}$$

$$= \frac{1}{1 + \dfrac{\gamma^2\, \Delta x^2}{2}} = \frac{2D_p\tau}{2D_p\tau + \Delta x^2 + s\tau\, \Delta x^2}. \tag{36}$$

But symmetry does not exist in a practical transistor. The constants in (36) are therefore, different for α_N and for α_I. Equation (36) can be rewritten under this consideration in the well-known form

$$\alpha_N(s) = \frac{\alpha_{N0}}{1 + s/\omega_{\alpha N}} \tag{37a}$$

$$\alpha_I(s) = \frac{\alpha_{I0}}{1 + s/\omega_{\alpha I}}. \tag{37b}$$

The relations between the constants in (37) and the physical parameters (corresponding to the constants in (36) modified for the nonsymmetrical case) will be derived in Section 5.1.

On account of their nonlinearity, the junction diodes must be described in the time domain. In their original paper, Ebers and Moll defined only a dc relationship between voltages and currents. This would restrict the use of their model to piecewise linear analysis. But the model can be made more general[38] by postulating that the $v = f(i)$ relation be valid at all times, as indicated in Fig. 17.

In either case, an important property of the semiconductor junction is lost: Voltages and currents appear as being directly related instead of being related indirectly through current density or charge. This can best be illustrated by an example. If a forward current through a junction is suddenly replaced by a reverse current the voltage actually does not reverse sign until the excess carrier density at the junction is reduced to zero. According to the Ebers-Moll model, voltage and current always change polarity together. As mentioned before, it is for this reason that for a diode, no dynamic model of the Ebers-Moll type exists that would represent charge storage effects. In addition to this shortcoming, the feature of mixed time and frequency domain characterization is undersirable if the model is to be used in its nonlinear form, say on a computer.

The Ebers-Moll model was originally presented in a form, shown in Fig. 17(b), which differs slightly from that in Fig. 17(a). Both versions have been used throughout the literature over the past years and very few authors[39,40] have clearly pointed out the difference between

them. In Fig. 17(b) the collecting currents are α times as large as the total emitter and collector currents, respectively. A simple calculation shows that the two versions are formally equivalent, if the relations

$$I_{EF}(s) = \frac{I'_E(s)}{1 - \alpha_N(s)\alpha_I(s)} \tag{38a}$$

and

$$I_{CF}(s) = \frac{I'_C(s)}{1 - \alpha_N(s)\alpha_I(s)} \tag{38b}$$

are satisfied. A glance at the equations for the voltage sources in Fig. 17 reveals that the two versions could not be completely equivalent, unless either I_{EF0} and I_{CF0} or I_{E0} and I_{C0} would be considered frequency dependent. Due to the approximative nature of both models, this is normally not done.

From both Fig. 17(a) and (b) the respective four-pole equations, on which the model is based, can readily be derived in terms of electrical parameters:

$$I_E(s) = I_{EF}(s) - \alpha_I(s)I_{CF}(s) = \frac{I'_E(s) - \alpha_I(s)I'_C(s)}{1 - \alpha_N(s)\alpha_I(s)} \tag{39a}$$

$$I_C(s) = \alpha_N(s)I_{EF}(s) - I_{CF}(s) = \frac{\alpha_N(s)I'_E(s) - I'_C(s)}{1 - \alpha_N(s)\alpha_I(s)}. \tag{39b}$$

After substituting the expressions for the junctions one obtains for the *steady-state* case the well-known Ebers-Moll equations

$$I_E = \frac{I_{E0}}{1 - \alpha_{N0}\alpha_{I0}} [\exp(\lambda V_e) - 1] - \frac{\alpha_{I0}I_{C0}}{1 - \alpha_{N0}\alpha_{I0}} [\exp(\lambda V_e) - 1] \tag{40a}$$

$$I_C = \frac{\alpha_{N0}I_{E0}}{1 - \alpha_{N0}\alpha_{I0}} [\exp(\lambda V_e) - 1] - \frac{I_{C0}}{1 - \alpha_{N0}\alpha_{I0}} [\exp(\lambda V_e) - 1]. \tag{40b}$$

5.1 *Comparison Between the Ebers-Moll and the Linvill Model*

Comparing (40) with (33) and (34) for the steady-state solution leads to the following relations:

$$\frac{eAD_p p_{n0}}{W} = \frac{\alpha_{N0}I_{E0}}{1 - \alpha_{N0}\alpha_{I0}} = \frac{\alpha_{I0}I_{C0}}{1 - \alpha_{N0}\alpha_{I0}}. \tag{41}$$

A corresponding comparison for the ac case would yield the same expression as in (41), except that α_{N0} and α_{I0} would have to be replaced by their frequency dependent forms. Since the left side term of (41)

is frequency independent, no rigorous equality exists between the Linvill model and any of the two versions of the Ebers-Moll model under ac conditions. In the Ebers-Moll model, the junction voltage is a function of the total diode current [being different in the two versions of Fig. 17(a) and (b)]; in the Linvill model it is only a function of the resistive component of the diode current in Fig. 17(a); this component equals the current through the combinance which is proportional to the carrier density p. It can be shown that the correct solution in which the junction voltage is a function of the carrier density directly at the junction, lies between these two cases but much closer to the lumped Linvill simulation. The discrepancy, mentioned here, affects only the junction voltages and does not appear in many analyses that use piecewise linearity.

$\alpha_N(s)$ and $\alpha_I(s)$ can be expressed in terms of the physical parameters by comparing (39) and (33) separately for the normal operation ($I_{CF}=0$) and for the inverse operation ($I_{EF} = 0$). Subsequent conversion of the α's into β's yields

$$\beta_N(s) = \frac{\alpha_N(s)}{1 - \alpha_N(s)} = \frac{\beta_{N0}}{1 + \frac{s}{\omega_{\beta N}}} = \frac{A\tau_1 D_p/A_1 W W_1}{1 + s\tau_1} \tag{42}$$

$$\beta_I(s) = \frac{\alpha_I(s)}{1 - \alpha_I(s)} = \frac{\beta_{I0}}{1 + \frac{s}{\omega_{\beta I}}} = \frac{A\tau_2 D_p/A_2 W W_2}{1 + s\tau_2}, \tag{43}$$

where

$$\omega_{\beta N} = \frac{\omega_{\alpha N}}{1 + \beta_{N0}} = \frac{1}{\tau_1} \tag{44}$$

and

$$\omega_{\beta I} = \frac{\omega_{\alpha I}}{1 + \beta_{I0}} = \frac{1}{\tau_2}. \tag{45}$$

By definition we shall call in later sections

$$\tau_1 \equiv \tau_{BN} \tag{46}$$

and

$$\tau_2 \equiv \tau_{BI} \tag{47}$$

5.2 *A Better Approximation for the α Frequency Dependence in the Ebers-Moll Model*

Pritchard[41] has first suggested that a better approximation for the 3-dB cut-off points of the α's or β's are obtained if one inserts a factor

1.22 into the corresponding equations, i.e.,

$$\alpha_N = \frac{1}{\cosh \gamma \, \Delta x} \approx \frac{1}{1 + \dfrac{1.22 j\omega}{\omega_{\text{cut-off measured}}}}. \tag{48}$$

This can readily be calculated from the fall-off behavior of the cosh expression while assuming $\beta_{N0} \gg 1$.

The same factor 1.22 appears in the corresponding expressions for α_I, β_N and β_I. It is evident from (48) that this problem can be reduced to a matter of defining $\omega_{\alpha N}$. For less ideal transistors the factor is usually between 1 and 1.22.

Higher-order approximations to the hyperbolic function commonly use two pole expressions or delay-producing excess phase terms.

VI. THE CHARGE-CONTROL TRANSISTOR MODEL

In analogy to the diode charge-control model we can establish a charge-control equivalent circuit for the transistor. To that end, we want to express all parameters in terms of the charge in the base.

Three approaches appear feasible: A lumped Linvill model can be labeled in such a way that all elements appear as functions of charges rather than integrated charge densities of the form $p\Delta x$. The two are proportional; the proportionality factors are of the form "electron charge times area". Most of the special circuit components of the Linvill model become current or voltage sources in the charge-control version. This procedure of converting a Linvill model into a charge control model can readily be applied to higher-order Linvill models.

A second approach is to use the Ebers-Moll principle of superposition whereby two charge-control diode models plus the corresponding collecting currents can be joined to form the transistor model. This approach is essentially limited to the first order of approximation. Two seemingly different, but fully equivalent and easily convertible models result.

The third and classic approach to charge-control theory, originated by Beaufoy and Sparkes,[3] is basically mathematical. Through integration of the continuity equation the carrier density as a variable is replaced by the total charge in the base. Certain simplifying assumptions have to be made to obtain a relation between currents and charges. In essence, these assumptions are equivalent to the approximations implied in the first-order Linvill and Ebers-Moll models as well as in the first-order charge-control equivalent circuits to be described below.

Some equivalent circuits have been presented in the literature, but they were less rigorous than the circuits described below in the sense that they cannot be used as complete networks. Additional knowledge of the physics of the device is required to use these models. Extension to higher-order models in the Beaufoy-Sparkes approach is accomplished through increased physical and mathematical complexity and not through more complex network topology as in the Linvill model or the charge-control model to be described.

6.1 *The π-Version (Base-Controlled Version) of the Charge-Control Equivalent Circuit*

In the lumped Linvill π-model of Fig. 16, the base charge distribution is approximated by two levels of carrier density. This is illustrated in Fig. 18. $p_1(t)$ is constant over the length $\Delta x = W_1$, and $p_2(t)$ is constant over the length W_2, where $W_1 + W_2 =$ basewidth W. The total charge in the two sections follows with (34a) and (34b) as

$$q_1(t) = p_1(t)W_1eA_1 = p_{n0}W_1eA_1[\exp\{\lambda v_e(t)\} - 1] \tag{49}$$

$$q_2(t) = p_2(t)W_2eA_2 = p_{n0}W_2eA_2[\exp\{\lambda v_e(t)\} - 1]. \tag{50}$$

Using the definitions of the elements given in Fig. 16, one can calculate from the Linvill model in Fig. 16 the currents through H_{e1}, H_{e2}, and H_d and obtains

$$i_1(t) = H_{e1}p_1(t) = \frac{eA_1W_1}{\tau_1}p_1(t) = \frac{q_N(t)}{\tau_1} \equiv \frac{q_N(t)}{\tau_{BN}} \tag{51}$$

$$i_5(t) = H_{e2}p_2(t) = \frac{eA_2W_2}{\tau_2}p_2(t) = \frac{q_I(t)}{\tau_2} \equiv \frac{q_I(t)}{\tau_{BI}} \tag{52}$$

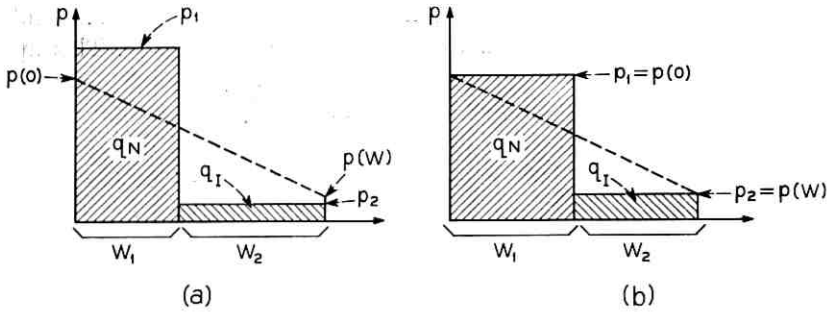


Fig. 18—Excess carrier distribution in the transistor base as used or implied in all first-order transistor models. (a) General case $p_1 \neq p(0)$, $p_2 \neq p(W)$. (b) Commonly used choice $p_1 = p(0)$, $p_2 = p(W)$.

$$i_3(t) = H_d[p_1(t) - p_2(t)]$$

$$= \frac{eAD_p}{W}(p_1(t) - p_2(t)) = \frac{D_pA}{W}\left(\frac{q_N(t)}{W_1A_1} - \frac{q_I(t)}{W_2A_2}\right). \quad (53a)$$

Using the more familiar Ebers-Moll notations and the relations found earlier in (42) through (47), i_3 can be expressed as

$$i_3(t) = \frac{\beta_{N0}}{\tau_{BN}} q_N(t) - \frac{\beta_{I0}}{\tau_{BI}} q_I(t). \quad (53b)$$

Thus, the three current sources in the charge-control model are found and related to the Linvill model by means of (51) through (53).

The remaining two branch currents i_2 and i_4 are obtained from Fig. 16 as

$$i_2(t) = eA_1W_1 \frac{dp_1(t)}{dt} = \frac{dq_N(t)}{dt} \quad (54)$$

$$i_4(t) = eA_2W_2 \frac{dp_2(t)}{dt} = \frac{dq_I(t)}{dt}. \quad (55)$$

These equations describe two stores S_N and S_I, whose properties have been described in Section 2.4.

The conversion between the two models will be summarized and further discussed in Section 6.4.

The voltage sources for the junctions follow from (34), (51), and (52) as

$$\lambda v_e(t) = \ln\left(1 + \frac{p(0,t)}{p_{n0}}\right) = \ln\left(1 + \frac{q_N(t)}{p_{n0}W_1eA_1}\right) \quad (56)$$

$$\lambda v_c(t) = \ln\left(1 + \frac{p(W,t)}{p_{n0}}\right) = \ln\left(1 + \frac{q_I(t)}{p_{n0}W_2eA_2}\right). \quad (57)$$

With the help of (41) through (47) that link the constants used in the Ebers-Moll model to those in the Linvill Model, (56) and (57) can be rewritten as

$$\lambda v_e(t) = \ln\left[1 + \frac{q_N(t)}{\tau_{BN}} \times \frac{1 + \beta_{N0}}{I_{E0}/(1 - \alpha_{N0}\alpha_{I0})}\right] \quad (58)$$

$$\lambda v_c(t) = \ln\left[1 + \frac{q_I(t)}{\tau_{BI}} \times \frac{1 + \beta_{I0}}{I_{C0}/(1 - \alpha_{N0}\alpha_{I0})}\right]. \quad (59)$$

(The reader may prefer to derive the constants directly from the steady-state Ebers-Moll model in (40) by considering the limiting

cases $v_e = 0$ and $v_c = 0$.) With the addition of (58) and (59) the equivalent circuit in Fig. 19(a) is completely defined.

It is customary and useful in charge-control work to define additional parameters τ_{EN}, τ_{CN}, τ_{EI}, τ_{CI}. We define their relationship as follows:

$$\frac{\tau_{BN}}{\beta_{N0}} \equiv \frac{\tau_{EN}}{\alpha_{N0}} \equiv \tau_{CN} \tag{60}$$

$$\frac{\tau_{BI}}{\beta_{I0}} \equiv \frac{\tau_{CI}}{\alpha_{I0}} \equiv \tau_{EI} . \tag{61}$$

Since, as usual,

$$\alpha_{N0} = \frac{\beta_{N0}}{1 + \beta_{N0}} \tag{62}$$

and

$$\alpha_{I0} = \frac{\beta_{I0}}{1 + \beta_{I0}} , \tag{63}$$

it also follows that

$$\frac{1}{\tau_{EN}} = \frac{1}{\tau_{BN}} + \frac{1}{\tau_{CN}} \tag{64}$$

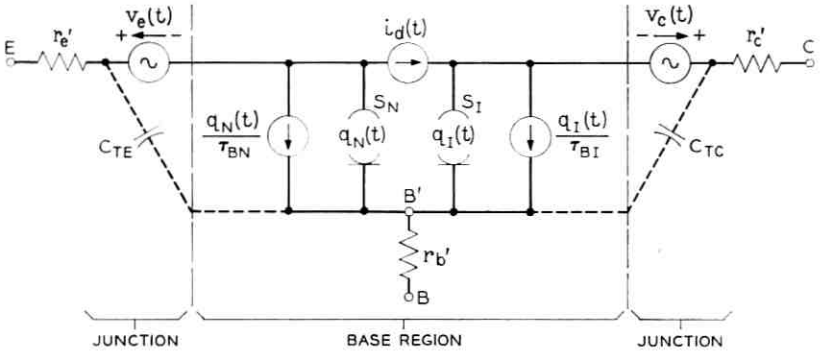
and

$$\frac{1}{\tau_{CI}} = \frac{1}{\tau_{BI}} + \frac{1}{\tau_{EI}} . \tag{65}$$

(The classic definition of these time constants after Beaufoy-Sparkes will be discussed in Section 6.5.) The subscripts B, E, and C stand for base, emitter, and collector, respectively. The subscripts N and I on the time constants and on the charges have been chosen to indicate the normal and inverse transistor operation. Many authors[9,10,11] use F (forward) and R(reverse) instead of N and I. Since F and R are commonly reserved for diode forward and reverse currents, and since such currents can flow in each of the two junctions, the different notations N and I, as proposed by Ebers and Moll, appear more appropriate.

In Appendix B, the notations for the stored charges and the time constants used in this paper are related to those used in a recent book published by the Semiconductor Electronics Education Committee;[11] they are also compared with the notations and definitions used by Beaufoy and Sparkes.

The additional time constants do not add any additional degree of freedom. But it is advantageous to use "base" notations when controlling base current, i.e., in common-emitter or common-collector



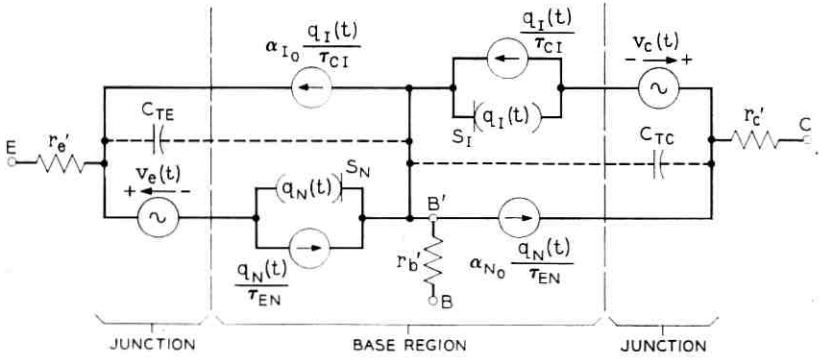
WHERE:

$$i_d(t) = \beta_{N_0}\frac{q_N(t)}{\tau_{BN}} - \beta_{I_0}\frac{q_I(t)}{\tau_{BI}} = \frac{q_N(t)}{\tau_{CN}} - \frac{q_I(t)}{\tau_{EI}}$$

$$v_e(t) = \frac{kT}{e}\ln\left[1 + \frac{q_N(t)\left(1+\beta_{N_0}\right)}{\tau_{BN} I_{E_0}/\left(1-\alpha_{N_0}\alpha_{I_0}\right)}\right]$$

$$v_c(t) = \frac{kT}{e}\ln\left[1 + \frac{q_I(t)\left(1+\beta_{I_0}\right)}{\tau_{BI} I_{C_0}/\left(1-\alpha_{N_0}\alpha_{I_0}\right)}\right]$$

(a)



WHERE:

$$v_e(t) = \frac{kT}{e}\ln\left[1 + \frac{q_N(t)}{\tau_{EN} I_{E_0}/\left(1-\alpha_{N_0}\alpha_{I_0}\right)}\right]$$

$$v_c(t) = \frac{kT}{e}\ln\left[1 + \frac{q_I(t)}{\tau_{CI} I_{C_0}/\left(1-\alpha_{N_0}\alpha_{I_0}\right)}\right]$$

(b)

Fig. 19 — Charge-control equivalent circuit for transistor in first-order approximation, shown in two equivalent and convertible forms: (a) π-version, Linvill type, (b) T-version, Ebers-Moll type.

connection, and to use "emitter" and "collector" notations when emitter and collector forward currents are injected, such as in common-base connection.

The model obtained in Fig. 19(a) maintains the most valuable property found in the Linvill model, namely the close relationship between the physical processes and the circuit elements. For example, τ_{BN} and τ_{BI} are the recombination times on the emitter and collector side, respectively, and τ_{CN} and τ_{EI} are the diffusion time constants for the charges injected from the two junctions. Junctions and base are represented by individual sections within the equivalent circuit. This separation makes it easy to expand the model and to take other effects into account.

6.2 *The T-Version (Emitter-Collector Controlled Version) of the Charge-Control Equivalent Circuit*

In complete analogy with the derivation of the Ebers-Moll model in Fig. 17(a) we can take two diode charge-control models back to back and add current sources on the collector and emitter side, which are α_{N0} and α_{I0} times the diode currents.

For the simulation of the junctions, we are left with two alternatives: One is to convert the corresponding expressions in the Ebers-Moll model in Fig. 17(a) into charge functions; the other is to use the expressions in the charge-control π-model (which are equivalent to the Linvill model), but replace the β-notations by α-notations according to (60) through (63). The first-mentioned alternative for simulating the voltage sources would amount to simply substituting the diodes from Fig. 17(a) for the voltage sources in Fig. 19(b). The property of charge control would not be simulated. The second procedure is therefore chosen; it yields

$$v_e(t) = \frac{1}{\lambda} \ln \left(1 + \frac{q_N(t)}{\tau_{EN} I_{E0}/(1 - \alpha_{N0}\alpha_{I0})}\right) \tag{66}$$

$$v_c(t) = \frac{1}{\lambda} \ln \left(1 + \frac{q_I(t)}{\tau_{CI} I_{C0}/(1 - \alpha_{N0}\alpha_{I0})}\right). \tag{67}$$

Thus, the equivalent circuit in Fig. 19(b) is obtained.

As far as the current relations in the models are concerned, the main difference between the charge-control T-model and the Ebers-Moll model is that the frequency dependence is simulated by a mathematical expression in the Ebers-Moll model, and by an additional network branch in the charge-control model. This is analogous to the option existing in small signal models where one can represent the frequency dependence either with an appropriate RC circuit, holding α_0 frequency

independent, or alternatively, with a frequency dependent α in the collecting current source.

The equivalency of the charge-control model with the Ebers-Moll model exists only for the relations between the currents. It can be shown readily that the following relations must be satisfied to establish equivalency:

$$(i) \qquad \tau_{EN} = \frac{1}{\omega_{\alpha N}} \tag{68}$$

$$(ii) \qquad \tau_{CI} = \frac{1}{\omega_{\alpha I}} \tag{69}$$

$$(iii) \qquad \tau_{BN} = \frac{1}{\omega_{\beta N}} = \frac{1 + \beta_{N0}}{\omega_{\alpha N}} \tag{70}$$

$$(iv) \qquad \tau_{BI} = \frac{1}{\omega_{\beta I}} = \frac{1 + \beta_{I0}}{\omega_{\alpha I}}. \tag{71}$$

All ω's must be replaced in these equations by the corresponding $\omega/1.22$ if the ω's correspond to the measured 3-dB gain fall-off points, and if the better approximation mentioned in Section 5.2 is to be included in the Ebers-Moll model, provided the particular transistor follows the underlying theory well enough.

6.3 *Conversion Between the Two Proposed Charge-Control Models*

The identity between the two charge-control models, presented in Figs. 19(a) and (b) can best be proven by converting one model into the other.

To convert the π-model into the T-model one first adds a branch current i_d both into and out of the base point B' and splits i_d up into its two components. The resulting circuit diagram is shown in Fig. 20. The two parallel current sources proportional to $q_N(t)$ on the left side can then be combined into one current source. Likewise, the two current sources proportional to $q_I(t)$ on the right side can be combined. If with the help of (60) through (63), one now relabels all current sources in terms of α_N and α_I instead of β_N and β_I and extends the upper current sources beyond the voltage sources, one obtains the model in Fig. 19(b).

6.4 *Summary of the Conversion Equations between the Linvill and the Charge-Control Model*

6.4.1 *Conversion Equations for the First-Order Transistor Model*

In (49) through (59), the charge-control π-model was derived from the Linvill model. With the help of the defining equations for the

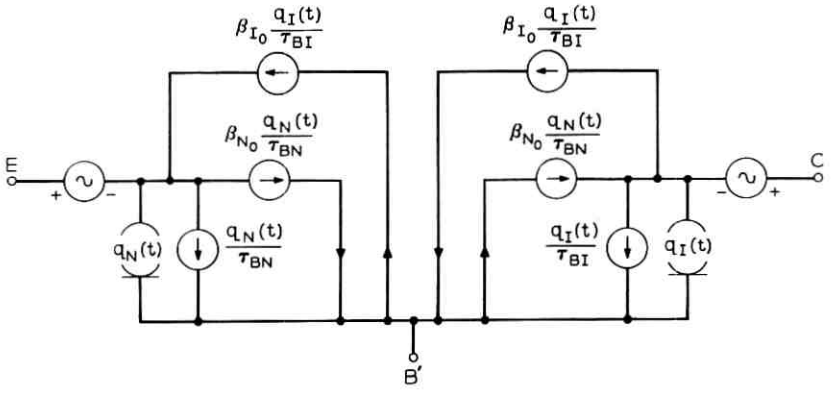


Fig. 20 — Intermediate step used in the conversion from the π to the T charge-control model, demonstrating equivalency between these two models.

Linvill model elements, the constants in the charge-control model can be calculated as a function of the Linvill combinances, storances, and diffusances. For the relation between the Linvill π-model of Fig. 16 and the charge-control π-model of Fig. 19(a), such calculations yield

$$\tau_{BN} = \frac{S_1}{H_{c1}} \tag{72}$$

$$\tau_{BI} = \frac{S_2}{H_{c2}} \tag{73}$$

$$\tau_{CN} = \frac{S_1}{H_d} \tag{74}$$

$$\tau_{EI} = \frac{S_2}{H_d} \tag{75}$$

$$I_{E0} = p_{N0}\frac{H_{c1}H_{c2} + (H_{c1} + H_{c2})H_d}{H_{c2} + H_d} \tag{76}$$

$$I_{C0} = p_{N0}\frac{H_{c1}H_{c2} + (H_{c1} + H_{c2})H_d}{H_{c1} + H_d}. \tag{77}$$

Note that (72) through (75) reveal that the five parameters in the Linvill model lead to only four parameters in the charge-control model. The one degree of freedom that is lost in the charge-control model is the conversion factor from current to carrier density; conversion of the charge-control model into a Linvill model is only possible, if one of the five Linvill parameters is known. This is tantamount to saying that one needs some information on the geometry of the device

such as the value of one, or in low-gain units, both of the two base volume sections A_1W_1 and A_2W_2.

6.4.2 *Conversion Between the Linvill and the Charge-Control Model for an Arbitrary Number of Base Sections*

In higher-order approximations for diodes or transistors, the parameters of the Linvill and the charge-control models, as defined in Fig. 21, are related by the equations

$$\tau_1 = \frac{S_1}{H_{c1}}, \qquad \tau_2 = \frac{S_2}{H_{c2}} \qquad \tau_\mu = \frac{S_\mu}{H_{c\mu}} \tag{78}$$

$$\tau_{12a} = \frac{S_1}{H_{d12}} \qquad \tau_{\mu\nu a} = \frac{S_\mu}{H_{d\mu\nu}} \tag{79}$$

$$\tau_{12b} = \frac{S_2}{H_{d12}} \qquad \tau_{\mu\nu b} = \frac{S_\nu}{H_{d\mu\nu}} \tag{80}$$

$$q_{10} = S_1 p_{n0} \tag{81}$$

$$q_{m0} = S_m p_{n0}\,. \tag{82}$$

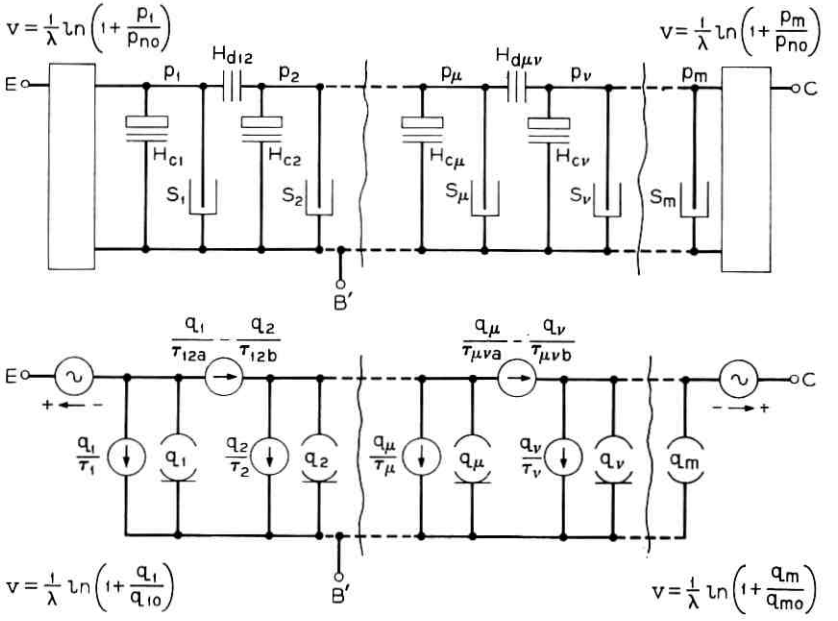


Fig. 21 — A Linvill and a charge-control equivalent circuit for a junction and part of a multisectional n-region, with indication of the notations used in converting one model into the other.

6.5 *The Transistor in Saturation*

In all lumped transistor models (Linvill, Ebers-Moll, or charge-control type) the charge in the base is explicitly or implicitly broken up into the charge q_N injected from the emitter under normal operation and the charge q_I injected from the collector under inverse operation, e.g., in saturation. This was illustrated in Fig. 18.

When the transistor is overdriven into saturation with a base current larger than $I_{C\,sat}/\beta_{N0}$, the two stores q_N and q_I do not change by exactly equal amounts, i.e.,

$$\frac{dq_N}{di_{B\,excess}} \neq \frac{dq_I}{di_{B\,excess}}$$

where, by definition,

$$i_{B\,excess} \equiv i_B - \frac{I_{C\,sat}}{\beta_{N0}}. \tag{83}$$

This is illustrated in Fig. 22. It can be calculated from any of the two models of Fig. 19 that, under steady-state conditions, the excess charges in the two stores are related to the excess base current by the expressions

$$\Delta Q_I = Q_I = \frac{\alpha_{N0}\tau_{CI}}{1 - \alpha_{N0}\alpha_{I0}} I_{B\,excess} \tag{84}$$

$$\Delta Q_N = \frac{\tau_{EN}}{1 - \alpha_{N0}\alpha_{I0}} I_{B\,excess}. \tag{85}$$

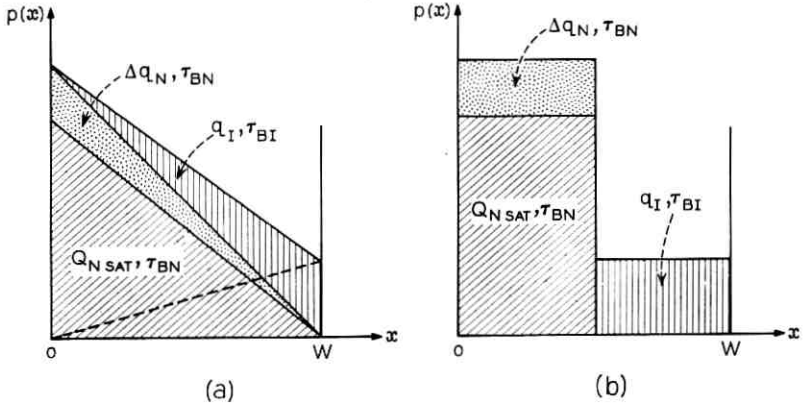


Fig. 22—The transistor in saturation. (a) Actual distribution of excess minority carriers. (b) Lumped approximation. The π and the T models use q_N and q_I with lifetimes τ_{BN} and τ_{BI}; the Beaufoy-Sparkes model uses $Q_{N\,SAT}$ and $q_{BS} = \Delta q_N + q_I$ with lifetimes τ_{BN} and τ_S, where τ_S is as defined in (88).

From this it follows that

$$\frac{\Delta Q_N}{Q_I} = \frac{\tau_{EN}}{\alpha_{N0}\tau_{CI}} = \frac{\tau_{CN}}{\alpha_{I0}\tau_{EI}} = \frac{\omega_{\alpha I}}{\alpha_{N0}\omega_{\alpha N}} \neq 1. \tag{86}$$

Since $\alpha_{I0} < 1$, the charge-up ratio is somewhat larger than the ratio of the diffusion times of the normal and inverse transistor.

The rate at which the two stores charge and discharge in saturation because of external step disturbances is described by the eigenfunction of the system

$$s^2 + s\left[\frac{1 + \beta_{N0}}{\tau_{BN}} + \frac{1 + \beta_{I0}}{\tau_{BI}}\right] + \frac{1 + \beta_{N0} + \beta_{I0}}{\tau_{BN}\tau_{BI}} = 0. \tag{87}$$

If $\beta_{N0}\tau_{BI}/\tau_{BN} \gg 1$, the two poles are far apart in frequency. Furthermore, the high frequency pole contributes in most nonoscillatory cases little to the overall response. The higher pole or, alternatively, the s^2 term can then be neglected and a single time constant results described by

$$\tau_S = \frac{(1 + \beta_{N0})\tau_{BI} + (1 + \beta_{I0})\tau_{BN}}{1 + \beta_{N0} + \beta_{I0}} = \frac{\tau_{EN} + \tau_{CI}}{1 - \alpha_{N0}\alpha_{I0}}. \tag{88a}$$

Using ω-notation, one obtains the expression given by Ebers and Moll

$$\tau_S = \frac{\omega_{\alpha N} + \omega_{\alpha I}}{\omega_{\alpha N}\omega_{\alpha I}(1 - \alpha_{N0}\alpha_{I0})}. \tag{88b}$$

For large β_N and small β_I, τ_S is approximately equal to

$$\tau_S \approx \tau_{BI}\left(1 + \frac{\tau_{EN}}{\tau_{CI}}\right). \tag{88c}$$

If $\tau_{EN} \ll \tau_{CI}$, i.e., if the carriers diffuse more easily from the emitter to the collector than vice versa, then the recombination rate τ_{BI} on the collector side is mainly responsible for the overall decay of the excess base charge.

6.5.1 *Storage Time Calculations*

For first-order storage time calculations with the transistor driven into a steady-state saturation condition by means of an excess base current $I_{B\,\text{excess}}$, one can simplify the charge-control model to the one shown in Fig. 23. Storage time is the time it takes to deplete the store which is charged to a value of

$$Q_{BS} = Q_I + \Delta Q_N = I_{B\,\text{excess}}\tau_S = I_{B\,\text{excess}}\frac{\tau_{EN} + \alpha_{N0}\tau_{CI}}{1 - \alpha_{N0}\alpha_{I0}} \tag{89}$$

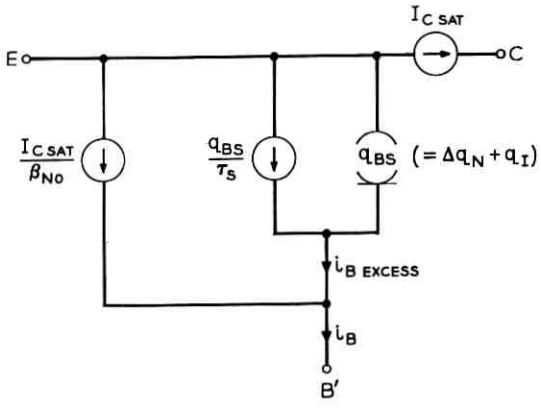


Fig. 23 — Single-pole equivalent circuit for saturated transistor after Beaufoy-Sparkes.

while, at the same time, this charge is exposed to an effective recombination time of τ_S as given in (88). In this form the model is identical with the classic Beaufoy-Sparkes model for the saturated transistor.

In the general case one must refer to the complete model.

6.6 *The Beaufoy-Sparkes Charge-Control Model*

In the classic approach to charge-control theory, the starting point is, like in the diode case, the integration of the continuity equation (8). In comparison with the integration performed for the diode in Section 2.3.1, the upper limit of integration has to be changed to $x = W$. The expression

$$i_p(0, t) - i_p(W, t) = \frac{dq_N(t)}{dt} + \frac{q_N(t)}{\tau_{BN}}$$

obtained from the integration becomes that for the base current under normal, nonsaturated operation:

$$i_B(t) = \frac{dq_N(t)}{dt} + \frac{q_N(t)}{\tau_{BN}}\ ; \tag{90}$$

$q_N(t)$ is the total charge in the base. The next step being made is again the approximative assumption that the carriers redistribute themselves so quickly, that we can always assume steady-state distribution. (See also Section 2.4.) Mathematically, this means that in normal transistor operation both $p(0,t)$ and $i_C(t)$ are proportional to the base charge $q_N(t)$. It can be seen from Fig. 19 that the same assumption

is implied in the two charge-control models presented there, despite the fact that they were derived through entirely different procedures. (Instantaneous redistribution is, however, not implied in models that use more than one π or T section for representing the base.)

The time constants are defined in the classic charge-control theory on the basis of the above-mentioned assumption of instantaneous carrier redistribution, i.e., in steady state

$$\tau_{BN} = \frac{Q_N}{I_{BN}}, \qquad \tau_{CN} = \frac{Q_N}{I_{CN}}, \qquad \tau_{EN} = \frac{Q_N}{I_{EN}} \tag{91a}$$

and dynamically

$$i_{BN} = \frac{q_N}{\tau_{BN}} + \frac{dq_N}{dt}, \qquad i_{CN} = \frac{q_N}{\tau_{CN}}, \qquad i_{EN} = i_{BN} + i_{CN}. \tag{91b}$$

The remaining three time constants can be defined likewise for the inverse transistor. Narud, et al[9] have used such definitions in an equivalent circuit for the charge-to-current relations in the transistor. Beaufoy and Sparkes discussed this possibility in their original paper[3] but chose to present two separate charge-control models, one for the normal active operation and one for saturation. In normal operation, the charge q_N called "q_B" is bounded by the value reached at the edge of saturation:

$$q_B \leqq Q_{N\,\mathrm{sat}}, \quad \text{where} \quad Q_{N\,\mathrm{sat}} = \frac{I_{C\,\mathrm{sat}}}{\beta_{N0}} \tau_{BN}.$$

In their saturated model, all excess charge which exceeds $Q_{N\,\mathrm{SAT}}$ is lumped into one store rather than two; this charge "q_{BS}" has a lifetime $\tau_S = q_{BS}/i_{B\,\mathrm{excess}}$ which is identical with τ_S as defined in (88).

By lumping $\Delta q_N \equiv q_N - Q_{N\,\mathrm{SAT}}$ and q_I into q_{BS}, the Beaufoy and Sparkes arrangement provides only a minor short cut for calculating storage time, while sacrificing not only some of the physical understanding, but also the possibility of mutual conversion with the other models. No relations have been given that would express the junction voltages in terms of the charges in the stores, and recourse must be taken to the Boltzmann equation to find expressions for the voltages.

Throughout the literature the charge-control concept has been used primarily as a mathematical-physical tool. Extensions to higher-order effects are usually made by improving the simple continuity and transport equations stated in (8) and (9) and then carrying out the corresponding integration for the specific application.

VII. SOME REMARKS ABOUT THE EQUIVALENT CIRCUIT TYPE CHARGE-CONTROL APPROACH

7.1 *Use of the Charge-Control Models*

It is believed that the first-order approximation to a charge-control model in the form presented for the transistor in Fig. 19, combines the main advantages of the three basic approaches to modeling. The π and the T-models are as easy to handle from an equivalent circuit point of view as the Ebers-Moll model. Instead of frequency dependent α's and β's, one additional current branch exists for each side of the transistor. Circuit problems are solved in the usual way by means of loop and node equations. The charges q_N and q_I appear as circuit parameters which can either be calculated, if so desired, or else, eliminated in the algebraic process. The store elements in the circuit are clearly defined by the circuit properties given in (22).

The model provides all the features that have made the charge-control concept attractive in the past: quick estimates of switching times by integrating the base current and equating with the charges needed to fill and deplete the stores. The general base current equations of charge control are directly read from Fig. 19(a) as

$$i_B = \frac{q_N}{\tau_{BN}} + \frac{dq_N}{dt} + \frac{q_I}{\tau_{BI}} + \frac{dq_I}{dt} + \frac{C_{TE}\, dv_e}{dt} + \frac{C_{TC}\, dv_c}{dt}. \tag{92}$$

Of course, there is no restriction to step inputs. The chore of calculating responses to a nonstep input is transformed through the model into a circuit problem. In complex cases the help of a computer will be required.

Due to its direct relationship to the Linvill model, the charge-control model lends itself quite readily to extensions based on the physics of the device. This will be discussed in Section 7.3.

7.2 *Piecewise Linear Approximation of the Logarithmic Voltage Function*

The logarithmic voltage functions for the junctions are of the form

$$v = \frac{1}{\lambda} \ln \left[1 + \frac{q/\tau}{I_0/(1 - \alpha_{N0}\alpha_{I0})} \right]. \tag{93}$$

For most practical cases, this can be approximated by piecewise linear functions, like in the diode case of Fig. 10. Except for small values of q/τ, i.e., q/τ not $\gg I_0/(1 - \alpha_{N0}\alpha_{I0})$, one obtains

$$\frac{dv}{d(q/\tau)} = \frac{\tau}{\lambda q}. \tag{94}$$

Thus, the slope can be represented by a resistor $\tau/\lambda q$, which may, like in Section 2.4.4, be taken as the average value

$$\frac{dv}{d\left(\frac{q}{\tau}\right)} \approx \frac{\tau}{\lambda \bar{q}} \approx \frac{\tau}{\frac{1}{2}\lambda q_{\max}} = \frac{1}{\frac{1}{2} I_{\max} \lambda}, \tag{95}$$

where $I_{\max}$ is the maximum forward junction current.

It can be shown that in models which use the exponential relationship, the expressions of the form $\ln(1 + x)$ can be replaced by just $\ln x$ if one simulates the majority carrier currents by special current sources as follows:

$$\frac{1 - \alpha_{I0}}{1 - \alpha_{N0}\alpha_{I0}} I_{C0} \quad (\approx I_{C0}) \quad \text{from internal base to collector}$$

and

$$\frac{1 - \alpha_{N0}}{1 - \alpha_{N0}\alpha_{I0}} I_{E0} \quad (\approx 0) \quad \text{from internal base to emitter.}$$

This transformation is rigorous only at dc. However, in a piecewise linear analysis, as discussed above, the addition of one current source, namely I_{C0}, becomes mandatory if the model is to be valid at very small collector currents.

7.3 *Extensions of the Model*

7.3.1 *Path Impedances, Leakage Resistors*

Like in the Linvill model, junctions and base material are clearly separated in the charge-control model. Therefore, it is a straightforward procedure to add series path resistors, series inductances, or leakage resistances to models like the ones in Figs. 9 or 19.

7.3.2 *Junction Capacitors*

It has been indicated by the dashed lines in Figs. 9 and 19 how the junction capacitances are to be incorporated into the model. They are properties of the junction, but their currents flow as majority carrier currents through the bulk material. Hence, in Fig. 9, for example, they must be connected across the whole n-region. (Connecting directly across the voltage source would have no effect on the external properties.) In Fig. 19 they lead to the internal base point.

7.3.3 *Higher Order than π-Transistor Models*

Another desirable expansion may be to replace the π structure of Fig. 19(a) by a double π or by some other higher-order approximation

to the original differential "transmission line". This is of special importance, if emphasis is to be placed on charge redistributions in the base. By extending the model in such a way, the restricting assumption of instantaneous carrier redistribution is no longer implied. A qualitative example of an elaborate planar or mesa transistor model is given in Fig. 24.

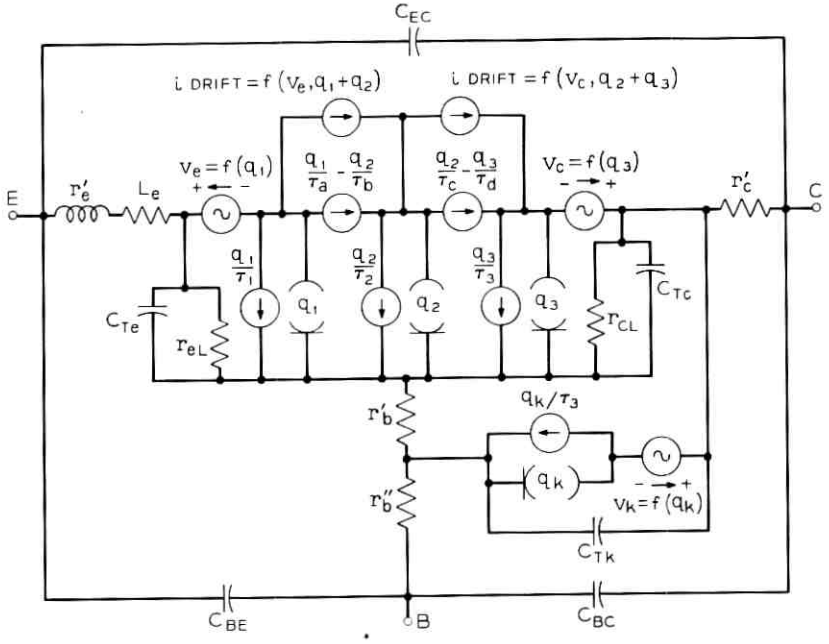


Fig. 24 — Example of an elaborate high-frequency planar or mesa transistor equivalent circuit.

7.3.4 *Drift Fields*

If the charge-control model under consideration is being developed on the basis of physical phenomena such as in the model of Fig. 19(a), the contribution from drift effects may be represented in the same way as has been proposed by Linvill [Ref. 7, Sections 2, 3]. As a direct consequence of the transport equation (2), drift can be represented by a current source added in parallel to the diffusion current source. In terms of the *r-g-c* transmission line representation, discussed earlier, drift consideration amounts to a resistor in parallel with the series diffusance resistor *r*. This was used in a recent paper by Bloodworth.[26]

Alternative methods of representing drift effects in conjunction with

conductivity modulation are presently being investigated; results will be published later.

7.3.5 *Base Width Modulation*

Base width modulation can be taken into account by replacing the basewidth, especially the collector section W_2, by an expression $W_2(1 + \Delta)$, where Δ is some function of the junction voltage. Equations (52) and (53) show the dependence of the branch currents on W_2, from which we can readily derive the required modification of the charge-control model in Fig. 19(a).

7.3.6 *Multiple-Layer Devices, Multiple Storage*

In accordance with Linvill's proposal, storage in more than one region can be simulated by considering that the minority carrier current on one side of a junction becomes the majority carrier current in the adjacent region. An example is shown in Fig. 25. This figure represents the charge-control model for an npnp device. Avalanche

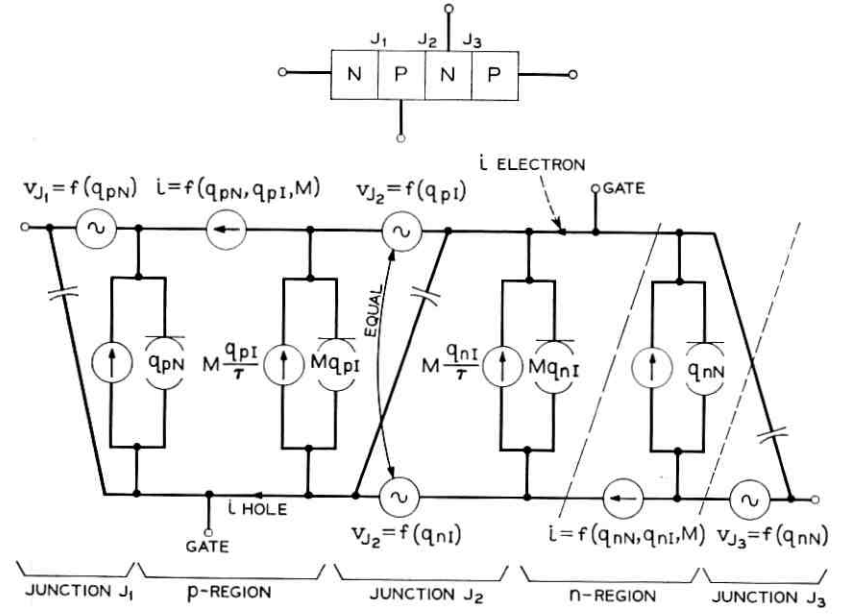


Fig. 25 — Charge-control model for npnp device. The model for an npn-transistor with storage in the collector can be obtained from this model by omitting the part to the right of the dashed or the dotted line.

multiplication may be considered by adding a multiplication factor M to all hole and electron currents flowing through the junction of interest (usually the center junction J_2), as indicated in Fig. 25. M is a function of the voltage across the junction.

By omitting the last electrode, an npn transistor with charge storage in the collector is obtained.

7.4 *Establishment of a Large-Signal Model*

The question naturally arises as to how one arrives at a numerical model. There is no clear-cut answer to this question, since the procedure to be taken depends on whether the informations available are predominantly physical or electrical in nature, whether a computer is available or not, etc. The following outline can, therefore, only be considered as a typical example.

(*i*) Obtain dc measurements which yield information on junction characteristics and electrode resistances. All measurements must be made under widely differing drive and load conditions.

(*ii*) Add information from device manufacturer to establish first-order dc model. (If necessary, convert to Linvill model.)

(*iii*) Add dynamic parameters, such as capacitances, as far as they are known and establish first-order dynamic model.

(*iv*) Use computer to improve numerical parameter values by matching frequency response curves or switching data in the active region with the model.

(*v*) Use computer to match large-signal nonlinear switching data.

(*vi*) Check model with switching measurements under different conditions, such as extremely low, extremely high and medium input and output impedance levels for various drive conditions. Improve model basically and numerically as necessary.

For purposes of *device* design, more emphasis is generally placed on the simulation of higher-order effects than in model building for *circuit* design where, especially in the case of integrated circuits, it is necessary to trade accuracy for simplicity.

VIII. CONCLUSIONS

The differential Linvill model stands out among all models as the most perfect one. Whereas the lumped Linvill model is the most suitable model for the device physicist, the circuit engineer usually prefers a more circuit oriented approach. It is felt that the charge-control *equivalent circuit* approach is well suited to combine the main advantages

of the various models: It is as easy to handle as the Ebers-Moll model, yet bears the close relationship to the physical phenomena of the device inherent in the Linvill model. It can also be extended easily to include higher-order physical phenomena.

Despite the difference in basic philosophy underlying the creation of each of the three classic modeling concepts (such as lumping, superposition, and integration), they are equivalent with respect to their current relations and to all dc properties. When compared at the same level of complexity, equivalency with respect to time dependency of the junction voltages exists between the two charge-control models and the Linvill model, but not between these models and the Ebers-Moll model.

In the Ebers-Moll model the effect which storage exercises on voltage cannot be included. The hydrib use of both time and frequency domains in the model may also be felt as a disadvantage in some applications.

At the first-order level of approximation, the charge-control equivalent circuit can be converted into the Ebers-Moll model, the Beaufoy-Sparkes model, and into the Linvill model (in the latter case the base volume is a constant which must also be known). Thus, the charge-control model serves as a bridge between the various models. This can be very useful in establishing a model, since both physical and electrical information can be incorporated easily into the model.

The diode charge-control model has been found very useful for analyzing storage effects in diodes.

Because of the close relationship to the physical phenomena in the device, extensions to larger complexity can readily be accomplished. We may interpret the charge control equivalent circuit as simply a circuit-oriented form of the Linvill model. The basic ideas and procedures that are used in converting diode and transistor linear models into equivalent charge-control models can be applied to many other semiconductor devices.

APPENDIX A

Switching Time Calculations for Ideal Charge-Storage-Step-Recovery Diodes

(Example for use of charge-control model)

A.1 *Equivalent Circuit* (See Fig. 26)

A.2 *Generator Source Current* (See Fig. 27)

A.3 *Diode Model* (See Fig. 28)

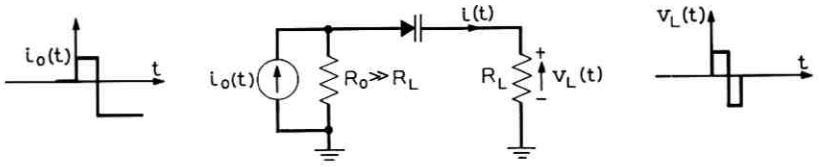


Fig. 26 — Equivalent circuit for charge-control model.

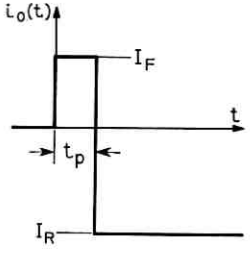


Fig. 27 — Generator source current.

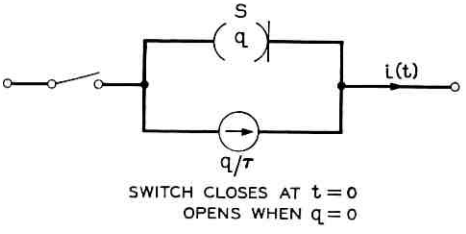


Fig. 28 — Diode model.

A.4 *Forward Operation*

$$I(s) = \frac{Q(s)}{\tau} + sQ(s) = \frac{I_F}{s}.$$

From this follows

$$Q(s) = \frac{I_F}{s\left(s + \frac{1}{\tau}\right)}$$

$$q(t) = \tau I_F[1 - \exp(-t/\tau)]$$

$$Q(t_p) = \tau I_F[1 - \exp(-t_p/\tau)]$$

A.5 *Reverse Operation*

For simplicity of writing, $t = t_p$ will now be referred to as $t = 0$:

$$I(s) = \frac{Q(s)}{\tau} + sQ(s) - Q(t_p) = -\frac{I_R}{s}.$$

From this follows

$$Q(s) = \frac{sQ(t_p) - I_R}{s\left(s + \frac{1}{\tau}\right)}$$

$$q(t) = Q(t_p) \exp(-t/\tau) - \tau I_R[1 - \exp(-t/\tau)].$$

Step recovery occurs at $t = T_x$, when $q = 0$

$$\exp(-T_x/\tau)[Q(t_p) + \tau I_R] = \tau I_R$$

$$T_x = \tau \ln\left[1 + \frac{Q(t_p)}{\tau I_R}\right] = \tau \ln\left(1 + \frac{I_F}{I_R}[1 - \exp(-t_p/\tau)]\right). \quad (96)$$

A.6 *Graphical Representation* (See Fig. 29)

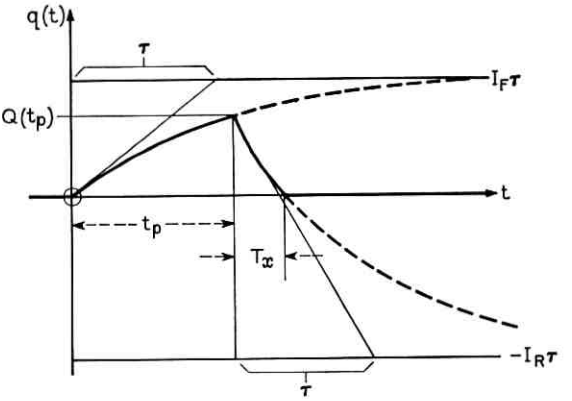


Fig. 29 — Graphical representation.

A.7 *Special Cases*

(*i*) $t_p \to \infty$: $Q(t_p) = I_F\tau$

$$T_x = \tau \ln\left(1 + \frac{I_F}{I_R}\right) \quad (97)$$

(*ii*) $I_R \gg I_F$: $T_x = \frac{Q(t_p)}{I_R}$ (98)

(*iii*) $I_R \gg I_F$ and $t_p \to \infty$: $T_x = \frac{I_F}{I_R}\tau$. (99)

APPENDIX B

Comparison of Notations

Table I lists comparisons of the notations used in this article. The first column lists the notations used in this article while column A lists those used by Beaufoy and Sparkes.[3] Column B lists the notations used in *Physical Electronics and Circuit Models* by P. E. Gray, et al;[11] SEEC Series, *2*.

TABLE I—COMPARISON OF NOTATIONS

This paper	A Beaufoy-Sparkes	B *SEEC*
q_N	q_B(for $q_N < Q_{NSAT}$ only)	q_F
q_I		q_R
$\Delta q_N + q_I$ where $\Delta q_N = q_N - Q_{NSAT}$	q_{BS}	
τ_{BN}	T_B	τ_{BF}
τ_{CN}	T_C	τ_F
τ_{EN}	T_E	$1 \Big/ \left(\frac{1}{\tau_{BF}} + \frac{1}{\tau_F}\right)$
τ_{BI}		τ_{BR}
τ_{CI}		$1 \Big/ \left(\frac{1}{\tau_{BR}} + \frac{1}{\tau_R}\right)$
τ_{EI}		τ_R
τ_S	T_S	τ_{SL}

LIST OF SYMBOLS

Lower-case letters are used for time variables, capital letters are used for steady-state values or Laplace transforms of values.

A, A_1, A_2	cross-sectional areas
A_{11}, A_{12}, A_{21}, A_{22}	four-pole parameters
a, b, c, K	constants
c; $\bar{c}$	analog capacitance; same per unit length
C_{Te}, C_{Tc}	emitter and collector junction capacitance, respectively
D_p	hole diffusion constant
D_n	electron diffusion constant
e	magnitude of electronic charge
E	electric field intensity
g; $\bar{g}$	analog shunt conductance; same per unit length

H_c , H_{c1} , H_{c2}	lumped combinances
$\overline{H_c}$	combinance per unit length
H_d	lumped diffusance
$\overline{1/H_d}$	reciprocal of diffusance, per unit length
i_1 , i_2 , i_3 , i_4 , i_5	network branch currents
i_B	base current
$i_{B\,\text{excess}}$, $I_{B\,\text{excess}}$	excess base current in saturation
i_C , I_C	collector current
i_E , I_E	emitter current
I_F , I_R	forward and reverse diode switching current, respectively
i_n	electron current
i_p	hole current
I_S	diode saturation current
i'_C , I'_C , i'_E , I'_E	network branch currents as defined in Figs. 17(a) and 17(b)
i_{CF} , I_{CF} , i_{EF} , I_{EF}	
I_0 , I_{C0} , I_{E0} , I_{CF0} , I_{EF0}	dc junction saturation currents
$I_{C\,\text{Sat}}$	collector current in saturation
i_{H_c} , I_{H_d} , i_S	currents through combinance, diffusance and storance, respectively
i_{BN} , i_{EN} , i_{CN}	base, emitter, and collector current in normal transistor operation
j_n	electron current density
j_p	hole current density
k	Boltzmann constant
L_p	diffusion length for holes
m, m_1 , m_2	constants relating lumped carrier density to carrier density at junction boundary
n	electron density; excess electron density
n_p	excess electron density in p-region
n_p , n_{p0}	values of n and n_p in thermal equilibrium
p, $P(s)$	hole density or excess hole density
p_n	excess hole density in n-region
p_0 , p_{n0}	value of p and p_n in thermal equilibrium
q, Q	charge
q_1 , q_2 , q_m	lumped charges in base region
q_N , Q_N	charge in normal store
q_I , Q_I	charge in inverse store
Δq_N , ΔQ_N , Δq_I , ΔQ_I , Q_{BS}	additional charges stored due to saturation
$Q_{N\,\text{SAT}}$	limiting value of Q_N, reached at edge of saturation

q_{10} , q_{m0}	total minority carrier charge in equilibrium
r; $\bar{r}$	analog resistor, same per unit length
r_e	small signal junction resistance
s	Laplace operator
S, S_1 , S_2 , S_m	stores = storances
$\bar{S}$	storance per unit length
t, t_x , T	time
T	absolute temperature
v, V	voltage
v_{ext}	externally applied junction voltage (excluding resistive drops)
v_c , v_e	collector and emitter junction voltages
W_1 , W_2	lengths denoting sections in neutral region
W	base width
x	neutral region length variable
Z	characteristic impedance
α_N , α_I	normal and inverse ac current gain in common-base connection
α_{N0} , α_{I0}	dc values of α_N and α_I
β_N , β_I	normal and inverse ac current gain in common-emitter connection
β_{N0} , β_{I0}	dc values of β_N and β_I
γ	transmission line propagation constant
λ	short for e/kT
μ_n , μ_p	electron and hole mobility, respectively
τ, τ_p	recombination time constant in p-region
τ_p' , τ_a , τ_b , τ_{12}	diffusion time constants
τ_s	approximative effective recombination time constant for excess charge in saturation
$\tau_1 \equiv \tau_{BN}$	recombination time in base under normal operation
$\tau_2 \equiv \tau_{BI}$	recombination time in base under inverse operation
τ_{CN} , τ_{EI}	normal collector and inverse emitter diffusion time constants, respectively
τ_{EN} , τ_{CI}	normal emitter time constant ($= 1/\omega_{\alpha N}$) and inverse collector time constant ($= 1/\omega_{\alpha I}$), respectively
$\omega_{\alpha N}$, $\omega_{\alpha I}$	common-base angular cut-off frequencies
$\omega_{\beta N}$, $\omega_{\beta I}$	common-emitter angular cut-off frequencies

REFERENCES

1. Ebers, J. J. and Moll, T. L., Large Signal Behavior of Junction Transistors, Proc. IRE, *42*, December, 1954, pp. 1761–1772.
2. Linvill, J. G. and Gibbons, J. F., *Transistors and Active Circuits*, McGraw-Hill Book Co., New York, 1961.
3. Beaufoy, R. and Sparkes, J. J., The Junction Transistor as a Charge-Controlled Device, ATE Journal, *B*, October 1957, pp. 310–327.
4. Moll, J. L., Large Signal Transient Response of Junction Transistors, Proc. IRE, *42*, December, 1954, pp. 1773–1783.
5. Linvill, J. G. and Wunderlin, W., Transient Response of Junction Diodes, IEEE Trans. Circuit Theor., *10*, June 1963, pp. 191–197; Technical Report No. 1513–1, August, 1962, Stanford University.
6. Linvill, J. G. and Wunderlin, W., Untersuchung von Schaltvorgaengen in Halbleiterdioden mittels Modellen mit konzentrierten Ersatzelementen, AEÜ, *17*, 1963, pp. 35–40.
7. Linvill, J. G., *Models of Transistors and Diodes*, McGraw-Hill Book Co., New York, 1963.
8. Hamilton, D. J., Lindholm, F. A., and Narud, J. A., Large Signal Models for Junction Transistors, Engineering Research Laboratories College of Engineering, University of Arizona, Tucson, Arizona.
9. Narud, J. A., Hamilton, D. J., and Lindholm, F. A., Large Signal Models for Junction Transistors, 1963, ISSCC Philadelphia, Digest, pp. 56–57.
10. Hamilton, D. J., Lindholm, F. A., and Narud, J. A., Comparison of Large Signal Models for Junction Transistors, Proc. IEEE, *52*, March, 1964, pp. 239–248.
11. Gray, P. E., et al, *Physical Electronics and Circuit Models of Transistors*, Semiconductor Electronics Education Committee, *2*, John Wiley & Sons, Inc., New York, 1964.
12. Beddoes, M. P., Linvill's Lumped Models and the Simplified Model, Proc. IEEE, *53*, Correspondence May, 1965, pp. 552–554.
13. Melchior, H. and Strutt, M. J. O., Small Signal Equivalent Circuit of Unsymmetrical Diodes at High Current Densities, IEEE Trans. Electron Devices, *ED-12*, February, 1965, pp. 47–55.
14. Boothroyd, A. R., Charge Definition of Transistor Properties, 1962 ISSCC Philadelphia, Digest, pp. 30–31.
15. Lax, B. and Neustadter, S. F., Transient Response of a *P-N* Junction, J. Appl. Phys. *25*, September, 1954.
16. Sparkes, J. J., A Study of the Charge Control Parameters of Transistors, Proc. IRE, October, 1960, pp. 1696.
17. Ekiss, J. A. and Simmons, C. D., Junction Transistor Transient Response Characterization, Solid-State J., *2*, January, 1961, pp. 17–24.
18. Ekiss, Spiegel, Simmons, and Blank, Characterization of Switching Transistors, Armed Serv. Techn. Inf. Agency, Philco, No. R-113, AD 271/122, 275/510.
19. Ekiss, J. A., Applications of the Charge-Control Theory, IRE Trans. Electron Computers, *EC-11*, June, 1962, pp. 374–381.
20. Bader, C. J., Charge-Step-Derived Transfer Functions for the Junction Transistor, IEEE Trans. Commun. Electron, *66*, May, 1963, pp. 179–185.
21. Schmeltzer, R. A., Transient Characteristic of Alloy Junction Transistors Using a Generalized Charge Storage Model, IRE Trans. Electron Devices, *10*, May, 1963, pp. 164–170.
22. Den Brinker, C. S., Fairbairn, D., and Norris, B. L., An Analysis of the Switching Behavior of Graded Base Transistors, Electron. Eng., August, 1963, pp. 500–505.
23. Singhakowinta, A., Some Effects of Transit Time Through the Collector Depletion Layer of Junction Transistors, IEEE Trans. Circuit Theor., *CT-10*, September, 1963, p. 445.
24. Cho, Y., A Method of Theoretical Analysis of High Speed Junction Diode Logic Circuits, IEEE Trans. Electron Computers, *EC-12*, October, 1963, pp. 492–502.
25. Hegedus, C. L., Charge Model of Fast Transistors and the Measurement of

Charge Parameters by High Resolution Electronic Integrator. Solid-State Design, *5*, August, 1964, pp. 23–36.
26. Bloodworth, G. G., The Significance of the Excess Charge Product in Drift Transistors, Radio Electron. Eng., *28*, November, 1964, pp. 304–312.
27. Koehler, D., A New Charge Control Equivalent Circuit for Diodes and Transistors and Its Relation to Other Large Signal Models, 1965 International Solid-State Circuits Conference, Philadelphia, Digest of Technical Papers, pp. 38–39.
28. Bassett, H. G. and Greenaway, P. E., Electrical Properties of High-Frequency Transistors, Post Office Elec. Engrs. *J57*, April, 1964, pp. 54–59.
29. Nanavati, R. P., Charge Control Analysis of Transistor Storage Time Dependence on Input "On" Pulse Width, IRE Trans. Electron Devices, *10*, July, 1963, pp. 290–291.
30. Thiney, A., Rise and Fall Times of Transistors in Switching Operation Regardless of the Driving Source Impedance, IEEE Trans. Electron Computers, *EC-12*, February, 1963, p. 23.
31. Simmons, C. D., High-Speed Microenergy Switching, Solid-State J., *1*, September-October, 1960, pp. 31–36.
32. Nanavati, R. P. and Wilfinger, R. J., Predicting Transistor Storage Time for Non-Step, Quasi-Voltage Inputs, IRE Trans. Electron. Devices, *ED-9*, November, 1962, pp. 492–499.
33. Kuno, H. J., Rise and Fall Time Calculations of Junction Transistors, IEEE Trans. Electron Devices, *11*, April, 1964, pp. 151–55.
34. Lindholm, F. A. and Hamilton, D. J., Systematic Modeling of Solid-State Devices and Integrated Circuits, 1965 International Solid-State Circuits Conference, Philadelphia, Digest of Technical Papers, pp. 36–37.
35. Lo, A. W., et al., *Transistor Electronics*, Prentice-Hall, 1955.
36. Gärtner, W. W., *Transistors, Principles, Design and Applications*, D. van Nostrand Company, Inc., Princeton, 1960.
37. Lindmayer, J. and Wrigley, C. Y., *Fundamentals of Semiconductor Devices*, D. Van Nostrand Co., Inc., Princeton, 1965.
38. Narud, J. A., Seelbach, W. C., and Meyer, C. S., Microminiaturized Logic Circuits: Their Characterization, Analysis, and Impact Upon Computer Design, IEEE Conv., March, 1963.
39. Searle, S. C., et al., *Elementary Circuit Properties of Transistors*, Semiconductor Electronics Education Committee, *3*, John Wiley & Sons, Inc., New York, 1964, Section 2.1.
40. Lloyd, R. H. F., A Simpler Transistor Model, Proc. IEEE, *53*, Correspondence, May, 1965, pp. 527–528.
41. Pritchard, R. L., Frequency Variations of Current Amplification Factor of Junction Transistors, Proc. IRE, *40*, November, 1952, pp. 1476–1481.
42. Geller, S. B., Mantek, P. A., and Boyle, D. R., A General Junction Transistor Equivalent Circuit for Use in Large-Signal Switching Analysis, IRE Trans. Electron Computers, December, 1961, pp. 670–679.
43. Beale, J. R. A. and Beer, A. F., The Study of Large Signal High-Frequency Effects in Junction Transistors Using Analog Techniques, Proc. IRE, January, 1962, pp. 66–77.

Generalized Optimum Receivers of Gaussian Signals

By T. T. KADOTA

(Manuscript received October 28, 1966)

Optimum reception of two zero-mean Gaussian signals is accomplished by comparing a quadratic form $\int\int x(s)H(s,t)x(t)\,ds\,dt$ *in the observable waveform* $x(t)$ *with a predetermined threshold, if the symmetric kernel* $H(s,t)$ *can be given as a square-integrable solution of*

$$\iint R_1(s,u)H(u,v)R_2(v,t)\,du\,dv = R_2(s,t) - R_1(s,t),$$

where $R_1(s,t)$ *and* $R_2(s,t)$ *are the covariances of the two signals. In this paper, we generalize this result so that* $\sum_{l,m}\int\int x^{(l)}(s)H_{lm}(s,t)x^{(m)}(t)\,ds\,dt$ *is the quadratic form to be used and* $\{H_{lm}(s,t)\}$ *is given as a formal solution of*

$$\sum_{l,m}\iint \frac{\partial^l}{\partial u^l}R_1(s,u)H_{lm}(u,v)\frac{\partial^m}{\partial v^m}R_2(v,t)\,du\,dv = R_2(s,t) - R_1(s,t).$$

In other words, the generalized quadratic form is in the derivatives of $x(t)$ *as well as* $x(t)$ *itself and the kernels* $H_{lm}(s,t)$ *consist of two-dimensional* δ*-functions in addition to square-integrable functions. This result is extended to the case of two nonzero-mean signals and then to the case of* M *Gaussian signals in noise.*

I. INTRODUCTION

Consider the problem of discriminating between two zero-mean Gaussian signals by observing the sample function $x(t)$, $0 \leqq t \leqq 1$. We assume that their covariances $R_1(s,t)$ and $R_2(s,t)$ are continuous and positive-definite on $[0,1] \times [0,1]$. According to previous results,[1,2,3] if the integral equation

$$\int_0^1 \int_0^1 R_1(s,u)H(u,v)R_2(v,t)\,du\,dv = R_2(s,t) - R_1(s,t), \quad 0 \leqq s,t \leqq 1, \qquad (1)$$

has a symmetric and square-integrable solution $H(s,t)$, then the following decision scheme minimizes the error probability:

$$\text{choose } R_1(s,t) \quad \text{if} \quad \int_0^1 \int_0^1 x(s)H(s,t)x(t)\,ds\,dt < c, \tag{2}$$

$$\text{choose } R_2(s,t) \text{ otherwise,}$$

where

$$c = 2 \log \frac{\alpha_1}{\alpha_2} - \sum_{i=0}^{\infty} \log \lambda_i , \tag{3}$$

in which α_1 and α_2 are the *a priori* probabilities associated with the two signals, and $\lambda_i > 0$, $i = 0, 1, 2, \cdots$, are the eigenvalues of an operator $R_1^{-\frac{1}{2}}R_2R_1^{-\frac{1}{2}}$.*

Unfortunately, existence of a square-integrable solution of (1) is too restrictive a condition. Thus, relaxation of the condition, which amounts to generalization of the quadratic form of (2), is desirable. In this paper, we accomplish this in two ways: one is to allow $H(s,t)$ to contain δ-functions as well as square-integrable functions, resulting in the generalization of the structure of the quadratic form; the other is to consider the derivatives of $x(t)$ as well as $x(t)$ itself, thus generalizing the elements of the quadratic form. The result is extended to the case where the means of the two signals are nonzero, and is further extended to the case of M Gaussian signals in noise.

II. GENERALIZED OPTIMUM RECEIVER OF TWO ZERO-MEAN GAUSSIAN SIGNALS

Consider the following generalization of the quadratic form of (2):

$$Q(x) = \sum_{l,m=0}^{r} \int_0^1 \int_0^1 x^{(l)}(s)H_{lm}(s,t)x^{(m)}(t)\,ds\,dt, \tag{4}$$

where $x^{(l)}(t)$ is the lth derivative of $x(t)$, and

$$H_{lm}(s,t) = \sum_{i,k=1}^{n_1} a_{iklm}\,\delta(s - s_i)\,\delta(t - s_k)$$

---

* More precisely, λ_i, $i = 0, 1, 2, \ldots$, are the eigenvalues of the extension of $R_1^{-\frac{1}{2}}R_2R_1^{-\frac{1}{2}}$ to the whole of $\mathfrak{L}_2$, where R_1 and R_2 denote the integral operators with the kernels $R_1(s,t)$ and $R_2(s,t)$, and $\mathfrak{L}_2$ the space of all square-integrable functions on [0,1]. We recall that existence of a symmetric, square-integrable solution of (1) implies that $R_1^{-\frac{1}{2}}R_2R_1^{-\frac{1}{2}}$ has a unique bounded extension to the whole of $\mathfrak{L}_2$ having eigenvalues $\{\lambda_i\}$ such that $0 < \Pi_{i=0}^{\infty} \lambda_i < \infty$.

$$+ \sum_{j=1}^{n_2} [\delta(s - t_j)h_{jlm}(t) + \tilde{h}_{jlm}(s)\ \delta(t - t_j)]$$
$$+ \hat{h}_{lm}(s)\ \delta(s - t) + \tilde{H}_{lm}(s,t), \tag{5}$$

in which a_{jklm} are real constants, and $0 \leqq s_j\ ,\ t_j \leqq 1$, and $h_{jlm}(t)$, $h_{jlm}(t)$ and $\hat{h}_{lm}(t)$ are square-integrable functions on [0, 1] while $\tilde{H}_{lm}(s, t)$ are square-integrable functions on [0, 1] $\times$ [0, 1]. In writing (4), we have assumed that almost all sample functions of both signals have rth derivatives.* Note that the nonsquare-integrable part of $H_{lm}(s,t)$ consists of three types of two-dimensional δ-functions: (*i*) those at points and their mirror images with respect to the diagonal $s = t$, (*ii*) those along horizontal lines (t = constant) and their mirror images (s = constant), and (*iii*) those along the diagonal. By formally substituting (5) into (4), we obtain an explicit form of $Q(x)$, namely,

$$Q(x) = \sum_{l,m=0}^{r} \Bigg[\sum_{j,k=1}^{n_1} a_{jklm}x^{(l)}(s_j)x^{(m)}(s_k)$$
$$+ \sum_{j=1}^{n_2} x^{(l)}(t_j) \int_0^1 [h_{jlm}(t) + \tilde{h}_{jlm}(t)]x^{(m)}(t)\ dt$$
$$+ \int_0^1 x^{(l)}(t)\hat{h}_{lm}(t)x^{(m)}(t)\ dt + \int_0^1 \int_0^1 x^{(l)}(s)\tilde{H}_{lm}(s,t)x^{(m)}(t)\ ds\ dt \Bigg]. \tag{6}$$

As the corresponding generalization of the integral equation (1), we consider the following:

$$\sum_{l,m=0}^{r} \int_0^1 \int_0^1 \frac{\partial^l}{\partial u^l} R_1(s,u)H_{lm}(u,v) \frac{\partial^m}{\partial v^m} R_2(v,t)\ du\ dv$$
$$= R_2(s,t) - R_1(s,t), \qquad 0 \leqq s,t \leqq 1. \tag{7}$$

Again, through formal substitution of (5), (7) becomes

$$\sum_{l,m=0}^{r} \Bigg\{ \sum_{j,k=1}^{n_1} a_{jklm} \frac{\partial^l}{\partial t^l} R_1(s,t)|_{t=s_j} \frac{\partial^m}{\partial u^m} R_2(s,t)|_{s=s_k} + \sum_{j=1}^{n_2} \int_0^1$$
$$\cdot \Bigg[\frac{\partial^l}{\partial t^l} R_1(s,t)|_{t=t_j} h_{jlm}(u) \frac{\partial^m}{\partial u^m} R_2(u,t)$$
$$+ \frac{\partial^l}{\partial u^l} R_1(s,u)\tilde{h}_{jlm}(u) \frac{\partial^m}{\partial s^m} R_2(s,t)|_{s=t_j} \Bigg]\ du$$

* A simple sufficient condition for this is existence of $(\partial^{2r+2}/\partial s^{r+1}\partial t^{r+1})\ R_i(s,t)$, $i = 1, 2$.[4]

$$+\int_0^1 \frac{\partial^l}{\partial u^l} R_1(s,u)\hat{h}_{lm}(u)\,\frac{\partial^m}{\partial u^m} R_2(u,t)\,du + \int_0^1\int_0^1 \frac{\partial^l}{\partial u^l} R_1(s,u)\tilde{H}_{lm}(u,v)$$

$$\cdot\frac{\partial^m}{\partial v^m} R_2(v,t)\,du\,dv\Big\} = R_2(s,t) - R_1(s,t), \qquad 0 \leqq s,t \leqq 1, \tag{8}$$

where we have assumed that $(\partial^{2r}/\partial s^r \partial t^r)R_1(s,t)$ and $(\partial^{2r}/\partial s^r \partial t^r)R_2(s,t)$ exist and are continuous on $[0,1] \times [0,1]$.

Unlike $H(s,t)$ of (2), which is uniquely given as the symmetric, square-integrable solution of (1),* the defining elements of $Q(x)$ (i.e., $\{a_{jklm}\}$, $\{s_i\}$, $\{t_i\}$, $\{h_{jlm}(t)\}$, $\{\tilde{h}_{jlm}(t)\}$, $\{\hat{h}_{lm}(t)\}$, $\{H_{lm}(s,t)\}$) cannot be uniquely determined by (8) in general for a given pair of covariances $R_1(s,t)$ and $R_2(s,t)$. Nevertheless, we can establish the following:

If (*i*) $R_1(s,t)$ and $R_2(s,t)$ are positive-definite,

(*ii*) $(\partial^{2r}/\partial s^r \partial t^r)R_1(s,t)$ and $(\partial^{2r}/\partial s^r \partial t^r)R_2(s,t)$ are continuous,

(*iii*) for almost all sample functions both signals have rth derivatives,† and

(*iv*) there exist some set of finite sequences $\{a_{jklm}\}$, $\{s_i\}$, $\{t_i\}$, $\{h_{jlm}(t)\}$, $\{\tilde{h}_{jlm}(t)\}$, $\{\hat{h}_{lm}(t)\}$ and $\{\tilde{H}_{lm}(s,t)\}$ which satisfy (8), then the decision scheme (2) with $\int_0^1\int_0^1 x(s)H(s,t)x(t)dsdt$ replaced by $Q(x)$ of (6) is optimum.

The proof is based on two measure theoretical facts: (*i*) two probability measures P_1 and P_2 corresponding to two Gaussian signals are either equivalent or singular,‡[5,6] and (*ii*) if they are equivalent then there is a special random variable called the Radon-Nikodym derivative $(dP_2/dP_1)(x)$, in terms of which the optimum decision scheme is specified as follows:[1]

$$\text{choose} \quad R_1(s,t) \quad \text{if} \quad \frac{dP_2}{dP_1}(x) < \frac{\alpha_1}{\alpha_2},$$

$$\text{choose} \quad R_2(s,t) \quad \text{otherwise.}$$

Hence, in the Appendix, we first prove that existence of $\{a_{jklm}\}$, $\{s_i\}$, $\{t_i\}$, $\{h_{jlm}(t)\}$, $\{\tilde{h}_{jlm}(t)\}$, $\{\hat{h}_{lm}(t)\}$ and $\{\tilde{H}_{lm}(s,t)$ satisfying (8) implies equivalence of P_1 and P_2. Then, it follows that the eigenvalues λ_i, $i =$

* The uniqueness of $H(s,t)$ follows from positive-definiteness of $R_i(s,t)$, $i = 1,2$, and square integrability of $H(s,t)$.

† Continuity of $(\partial^{2r}/\partial s^r \partial t^r)\,R_i(s,t)$, $i = 1, 2$, and existence of $x^{(r)}(t)$ for almost all $x(t)$ may be replaced by a simpler but stronger condition that $(\partial^{2r+2}/\partial s^{r+1}\partial t^{r+1})$ $R_i(s,t)$, $i = 1, 2$, exist.

‡ From the communication theoretical point of view, singularity corresponds to the case of "perfect reception" where error probability vanishes. For the mathematical definition, see Ref. 7.

$0, 1, 2, \cdots$, exist.[8,3] Next, we explicitly obtain λ_i from (8) and show that $0 < \prod_{i=0}^{\infty} \lambda_i < \infty$. Thus, the threshold c of (3) is well defined. Lastly, we prove that

$$\frac{dP_2}{dP_1}(x) = \left(\prod_{i=0}^{\infty} \lambda_i\right)^{-\frac{1}{2}} \exp\left[\tfrac{1}{2}Q(x)\right] \tag{9}$$

for almost all $x(t)$ of both signals. Then, by substituting (9) into the above decision scheme and taking the logarithm of both sides, the assertion is immediately proved.

III. EXTENSION TO TWO NONZERO-MEAN GAUSSIAN SIGNALS

The preceding result can be extended to the case where the means of the two Gaussian signals are no longer zero.* Let P_{11} and P_{22} be two probability measures corresponding to two Gaussian signals with means $m_1(t)$, $m_2(t)$, $0 \leqq t \leqq 1$, and covariances $R_1(s,t)$, $R_2(s,t)$. $m_1(t)$ and $m_2(t)$ are assumed square-integrable while the assumptions on $R_1(s,t)$ and $R_2(s,t)$ remain the same. Introduce a third measure P_{21} corresponding to a Gaussian signal with mean $m_2(t)$ and covariance $R_1(s,t)$. Then, P_{11} and P_{22} are equivalent and

$$\frac{dP_{22}}{dP_{11}}(x) = \frac{dP_{22}}{dP_{21}}(x)\,\frac{dP_{21}}{dP_{11}}(x)$$

for almost all $x(t)$ of all three signals, if and only if P_{22} is equivalent to P_{21}, which in turn is equivalent to P_{11}. According to a previous result,[9] if there exist finite sequences of real numbers $\{\hat{a}_{il}\}$, and $\{\hat{t}_i\}$, $0 \leqq \hat{t}_i \leqq 1$, and square-integrable functions $\{\tilde{g}_l\}$ which satisfy

$$\sum_{l=0}^{r}\left[\sum_{j=1}^{n_3} \hat{a}_{jl}\,\frac{\partial^l}{\partial s^l} R_1(s,t)\Big|_{s=\hat{t}_j} + \int_0^1 \frac{\partial^l}{\partial s^l} R_1(s,t)\tilde{g}_l(s)\,ds\right] = m_2(t) - m_1(t), \qquad 0 \leqq t \leqq 1, \tag{10}$$

for almost all $x(t)$ of the two signals, then P_{11} and P_{21} are equivalent and $(dP_{21}/dP_{11})(x) = \exp[L(x)]$ for almost all $x(t)$ of the two signals, where

$$L(x) = \sum_{l=0}^{r}\left\{\sum_{j=1}^{n_3} \hat{a}_{jl}\,\frac{d^l}{dt^l}\left[x(t) - \frac{m_1(t) + m_2(t)}{2}\right]_{t=\hat{t}_j} + \int_0^1 \frac{d^l}{dt^l}\left[x(t) - \frac{m_1(t) + m_2(t)}{2}\right]\tilde{g}_l(t)\,dt\right\}. \tag{11}$$†

* This extension follows the development in Ref. 3, pp. 1628–1629 and pp. 1636–1637.

† This is the "sure signals-in-noise" counterpart of the result in Section II, namely, the generalized optimum receiver of two sure signals in Gaussian noise.

The remaining half of showing the equivalence of P_{21} and P_{22} and obtaining $(dP_{22}/dP_{21})(x)$ is accomplished simply by replacing $x(t)$ with $x(t) - m_2(t)$ in the result in Section II. Thus, upon combination, we conclude that, if there exist a set of finite sequences $\{\hat{a}_{jl}\}$, $\{\hat{t}_j\}$, $\{\tilde{g}_l(t)\}$ satisfying (10) and another set of sequences $\{a_{jklm}\}$, $\{s_j\}$, $\{t_j\}$, $\{h_{jlm}(t)\}$, $\{\tilde{h}_{jlm}(t)\}$, $\{\hat{h}_{lm}(t)\}$, $\{\tilde{H}_{lm}(s,t)\}$ satisfying (8), then the optimum decision scheme for this case is specified as follows:

$$\text{choose} \quad m_1(t),\ R_1(s,t) \quad \text{if} \quad 2L(x) + Q(x - m_2) < c,$$

$$\text{choose} \quad m_2(t),\ R_2(s,t) \quad \text{otherwise.}$$

IV. EXTENSION TO M GAUSSIAN SIGNALS IN NOISE

The above result can be further extended to the problem of discriminating among M Gaussian signals in Gaussian noise.* Let $m_i(t)$, $R_i(s,t)$ and α_i, $i = 1, 2, \cdots, M$, be the means, covariances and *a priori* probabilities of the signals, and $R_o(s,t)$ the noise covariance where the noise mean is assumed zero. The assumptions concerning $m_i(t)$, $R_i(s,t)$ and $R_o(s,t)$ are the same as in Section III.† Denote by P_{ii} the probability measure corresponding to the ith signal plus the noise, and by P_o the measure corresponding to the noise alone. Then, according to the theory of the generalized maximum likelihood test,[11] if each P_{ii} is equivalent to P_o,‡ then the optimum decision is to choose that $m_i(t)$ and $R_i(s,t)$ for which $\alpha_i(dP_{ii}/dP_o)(x)$ is maximum as a function of i.§ Observe that, if the ith signal plus the noise and the noise alone are interpreted as the two Gaussian signals of Section III with means $m_i(t)$ and zero, and covariances $R_o(s,t) + R_i(s,t)$ and $R_o(s,t)$, then the condition for equivalence of P_{ii} and P_o and the expression for $(dP_{ii}/dP_o)(x)$ are obtained simply by the following changes: $m_1(t) \to 0$, $m_2(t) \to m_i(t)$, $R_1(s,t) \to R_o(s,t)$, $R_2(s,t) \to R_o(s,t) + R_i(s,t)$. Thus, we conclude that, if for each i there exist a set of finite sequences $\{\hat{a}_{iil}\}$, $\{\hat{t}_{ij}\}$ and $\{\tilde{g}_{il}(t)\}$ satisfying

$$\sum_{l=0}^{r}\left[\sum_{j=1}^{n_{ii}} \hat{a}_{iil}\frac{\partial^l}{\partial s^l}R_o(s,t)|_{s=\hat{t}_{ij}} + \int_0^1 \frac{\partial^l}{\partial s^l}R_0(s,t)\tilde{g}_{il}(s)\,ds\right] = m_i(t), \qquad 0 \leqq t \leqq 1,$$

---

* This extension follows the development in Ref. 10, pp. 2192–2194.

† $R_i(s,t)$ need not be strictly positive-definite.

‡ Equivalence of P_{ii} and P_o corresponds to the condition that the ith Gaussian signal cannot be detected perfectly in the presence of this noise.

§ If $\alpha_i(dP_{ii}/dP_o)$ (x) becomes maximum at more than one value of i, choose the lowest of such i-values. See Ref. 11.

and another set of finite sequences $\{a_{ijklm}\}$, $\{s_{ij}\}$, $\{t_{ij}\}$, $\{h_{ijlm}(t)\}$, $\{\tilde{h}_{ijlm}(t)\}$, $\{\hat{h}_{ilm}(t)\}$ and $\{\tilde{H}_{ilm}(s,t)\}$ satisfying

$$\sum_{l,m=0}^{r}\left\{\sum_{j,k=1}^{n_{ik}} a_{ijklm}\frac{\partial^l}{\partial t^l}R_o(s,t)|_{t=s_{ij}}\frac{\partial^m}{\partial s^m}[R_o(s,t)+R_i(s,t)]_{s=s_{ik}}\right.$$

$$+\sum_{j=1}^{n_{i2}}\int_0^1\left[\frac{\partial^l}{\partial t^l}R_o(s,t)|_{t=t_{ij}}h_{ijlm}(u)\frac{\partial^m}{\partial u^m}(R_o(u,t)+R_i(u,t))\right.$$

$$\left.+\frac{\partial^l}{\partial u^l}R_o(s,u)\tilde{h}_{ijlm}(u)\frac{\partial^m}{\partial s^m}(R_o(s,t)+R_i(s,t))|_{s=t_{ij}}\right]du$$

$$+\int_0^1\frac{\partial^l}{\partial u^l}R_o(s,u)\hat{h}_{ilm}(u)$$

$$\cdot\frac{\partial^m}{\partial u^m}[R_o(s,t)+R_i(s,t)]\,du+\int_0^1\int_0^1\frac{\partial^l}{\partial u^l}R_o(s,u)\tilde{H}_{ilm}(u,v)$$

$$\left.\cdot\frac{\partial^m}{\partial v^m}[R_o(v,t)+R_i(v,t)]\,du\,dv\right\}=R_i(s,t),\qquad 0\leqq s,t\leqq 1,$$

then the optimum decision is to choose that signal $(m_i(t), R_i(s,t))$ for which $2L_i(x)+Q_i(x-m_i)+c_i$ is maximum as a function of i, where $L_i(x)$ and $Q_i(x)$ are defined by (11) and (6) with $\hat{a}_{jl}$, $\hat{t}_{ij}$, $\tilde{g}_l(t)$, and a_{jklm}, s_{ij}, t_{ij}, $h_{ilm}(t)$, $\tilde{h}_{ilm}(t)$ $\hat{h}_{lm}(t)$, $\tilde{H}_{lm}(s,t)$ replaced by $\hat{a}_{ijl}$, $\hat{t}_{ij}$, $\tilde{g}_{il}(t)$, and a_{ijklm}, s_{ij}, t_{ij}, $h_{ijlm}(t)$, $\tilde{h}_{ijlm}(t)$ $\hat{h}_{ilm}(t)$, $\tilde{H}_{ilm}(s,t)$, respectively, and

$$c_i=2\log\alpha_i-\sum_{n=0}^{\infty}\log\lambda_n^{(i)},$$

where $\lambda_n^{(i)}$, $n=0, 1, 2, \cdots$, are the eigenvalues of the extension of $I+R_o^{-\frac{1}{2}}R_iR_o^{-\frac{1}{2}}$ to the whole of $\mathfrak{L}_2$.

APPENDIX

Let P_1 and P_2 be two Gaussian measures associated with a separable and measurable process $\{x(t),\ 0\leqq t\leqq 1\}$ with means zero and covariances $R_1(s,t)$ and $R_2(s,t)$.

Theorem: Suppose $R_1(s,t)$ *and* $R_2(s,t)$ *are (strictly) positive-definite,* $(\partial^{2r}/\partial s^r\partial t^r)R_1(s,t)$ *and* $(\partial^{2r}/\partial s^r\partial t^r)R_2(s,t)$, $0\leqq r<\infty$, *exist and are continuous on* $[0,1]\times[0,1]$, *and almost all sample functions have the rth derivatives with respect to* P_1 *and* P_2. *If there exist a set of finite sequences** $\{a_{iklm}\}$ $\{s_i\}$, $\{t_i\}$, $\{h_{ilm}(t)\}$, $\{\tilde{h}_{ilm}(t)\}$ $\{\hat{h}_{lm}(t)\}$ *and* $\{\tilde{H}_{lm}(s,t)\}$ *which satisfy* (8), *then*

* The definitions of these sequences are given after (5).

(*i*) P_1 *and* P_2 *are equivalent, i.e.,* $P_1 \equiv P_2$,
(*ii*) (10) *holds a.s.,* $[P_1 , P_2]$.*

Proof: For simplicity, we introduce the following notations:

$$R_{is^l}(u,t) = \frac{\partial^l}{\partial s^l} R_i(s,t)|_{s=u} , \qquad R_{it^m}(s,v) = \frac{\partial^m}{\partial t^m} R_i(s,t)|_{t=v} ,$$

$$R_{is^lt^m}(u,v) = \frac{\partial^{l+m}}{\partial s^l \, \partial t^m} R_i(s,t)|_{s=u,\, t=v} , \qquad i = 1, 2,$$

$$K_1(s,t) = \sum_{l,m=0}^{r} \sum_{j,k=1}^{n_1} a_{jklm} R_{1t^l}(s,s_j) R_{2s^m}(s_k , t),$$

$$K_2(s,t) = \sum_{l,m=0}^{r} \sum_{j=1}^{n_2} [R_{1t^l}(s,t_j)(R_{2t^m} h_{jlm})(t) + (R_{1t^l} \tilde{h}_{jlm})(s) R_{2s^m}(t_j , t)],$$

$$K_3(s,t) = \sum_{l,m=0}^{r} \int_0^1 R_{1t^l}(s,u) \hat{h}_{lm}(u) R_{2s^m}(u,t) \, du,$$

$$K_4(s,t) = \sum_{l,m=0}^{r} \int_0^1 \int_0^1 R_{1t^l}(s,u) \tilde{H}_{lm}(u,v) R_{2s^m}(v,t) \, du \, dv.$$

Note $K_i(s,t)$, $i = 1, 2, 3, 4$, are square-integrable. Again, we delete the arguments s and t of the kernels to denote the corresponding integral operators. Thus, (8) becomes

$$\sum_{i=1}^{4} K_i = R_2 - R_1 , \tag{12}$$

hence,

$$R_1^{-\frac{1}{2}} R_2 R_1^{-\frac{1}{2}} - I = \sum_{i=1}^{4} R_1^{-\frac{1}{2}} K_i R_1^{-\frac{1}{2}}. \tag{13}$$

(*i*) To establish $P_1 \equiv P_2$, it suffices to prove that $R_1^{-\frac{1}{2}} R_2 R_1^{-\frac{1}{2}}$ is densely defined on $\mathfrak{L}_2$ and $R_1^{-\frac{1}{2}} R_2 R_1^{-\frac{1}{2}} - I$ is of Hilbert-Schmidt type, i.e., $||R_1^{-\frac{1}{2}} R_2 R_1^{-\frac{1}{2}} - I|| < \infty$.[8,2,3]† The principal tool to be used for this proof is the following expansion:[12]

$$R_{1s^lt^m}(s,t) = \sum_i \mu_i f_i^{(l)}(s) f_i^{(m)}(t), \qquad 0 \leqq l, m \leqq r, \tag{14}$$

uniformly on $[0,1] \times [0,1]$, where $\mu_i > 0$ and $f_i(t)$, $i = 0, 1, 2 \cdots$, are the eigenvalues and orthonormalized eigenfunctions of R_1 .

To prove that $R_1^{-\frac{1}{2}} R_2 R_1^{-\frac{1}{2}}$ is densely defined on $\mathfrak{L}_2$, it sufficies to show

* "a.s. $[P_1 , P_2]$" is the abbreviation of "almost surely with respect to P_1 and P_2".
† $||A||$ denotes the Hilbert-Schmidt norm of an operator A.

that $R_1^{-\frac{1}{2}}R_2$ is bounded since $R_1^{-\frac{1}{2}}$ is densely defined. Now, by applying the formula $||A||^2 = \operatorname{tr} A^*A = \sum_i (f_i , A^*Af_i)$ to the individual terms of $R_1^{-\frac{1}{2}}K_1$ first, we obtain

$$
\begin{aligned}
|| R_1^{-\frac{1}{2}}K_1 || &\leqq \sum_{l,m=0}^{r} \sum_{j,k=1}^{n_1} | a_{jklm} | \\
&\cdot \left[\sum_i \left| \int_0^1 (R_1^{-\frac{1}{2}}f_i)(u) R_{1t^l}(u,s_j) \right|^2 \int_0^1 R_{2s^m}(s_k , u) R_{2t^m}(u,s_k)\, du \right]^{\frac{1}{2}} \\
&= \sum_{l,m=0}^{r} \sum_{j,k=1}^{n_1} | a_{jklm} | \, | \sum_i \mu_i \, | f_i^{(l)}(s_j) |^2 R^2_{2s^m t^m}(s_k , s_k) \, |^{\frac{1}{2}} \\
&= \sum_{l,m=0}^{r} \sum_{j,k=1}^{n_1} | a_{jklm} | \, | R_{1s^l t^l}(s_j,s_j) R^2_{2s^m t^m}(s_k , s_k) \, |^{\frac{1}{2}},
\end{aligned}
$$

where (14) is used for the last two equalities. Similarly,

$$
\begin{aligned}
|| R_1^{-\frac{1}{2}}K_2 || &\leqq \sum_{l,m=0}^{r} \sum_{j=1}^{n_2} \{ [R_{1s^l t^l}(t_j , t_i)(h_{jlm} , R^2_{2s^m t^m} h_{jlm})]^{\frac{1}{2}} \\
&\qquad + [(h_{jml} , R_{1s^l t^l} \tilde{h}_{jlm}) R^2_{2s^m t^m}(t_i , t_j)]^{\frac{1}{2}} \}, \\
|| R_1^{-\frac{1}{2}}K_3 || &\leqq \sum_{l,m=0}^{r} | \operatorname{tr} (\hat{R}_{1s^l t^l , m} R^2_{2s^m t^m}) \, |^{\frac{1}{2}}, \\
|| R_1^{-\frac{1}{2}}K_4 || &\leqq \sum_{l,m=0}^{r} | \operatorname{tr} (\tilde{R}_{1s^l t^l , m} R^2_{2s^m t^m}) \, |^{\frac{1}{2}},
\end{aligned}
$$

where $R_{1s^l t^l , m}(s,t) = \hat{h}_{lm}(s) R_{1s^l t^l}(s,t) \hat{h}_{lm}(t)$, $\tilde{R}_{1s^l t^l , m} = \tilde{H}_{ml} R_{1s^l t^l} \tilde{H}_{lm}$. Hence, from (13), $||R_1^{-\frac{1}{2}}R_2|| < \infty$.

To prove $||R_1^{-\frac{1}{2}}R_2R_1^{-\frac{1}{2}} - I|| < \infty$, we apply the formula $||A||^2 = \sum ||Af_i||^2$ to the individual terms of $R_1^{-\frac{1}{2}}K_1R_1^{-\frac{1}{2}}$ first. Thus, we obtain

$$
\begin{aligned}
|| R_1^{-\frac{1}{2}}K_1R_1^{-\frac{1}{2}} || &\leqq \sum_{l,m=0}^{r} \sum_{j,k=1}^{n_1} | a_{jklm} | \, | \sum_i || R_1^{-\frac{1}{2}} R_{2t^m}(\cdot , s_k) \mu_i^{\frac{1}{2}} f_i^{(l)}(s_j) \, ||^2 \, |^{\frac{1}{2}} \\
&= \sum_{l,m=0}^{r} \sum_{j,k=1}^{n_1} | a_{jklm} | \, | R_{1s^l t^l}(s_j , s_j) \, |^{\frac{1}{2}} \, || R_1^{-\frac{1}{2}} R_{2t^m}(\cdot , s_k) \, ||,
\end{aligned}
$$

where (14) is used twice, and $R_1^{-\frac{1}{2}}R_{2t^m}(\cdot , s_k)$ denotes the result of $R_1^{-\frac{1}{2}}$ acting on an s-function $R_{2t^m}(s,s_k)$. By differentiating both sides of (8) with respect to t, we obtain

$$
\begin{aligned}
R_{2t^m}(s,s_k) &= R_{1t^m}(s,s_k) + \sum_{l,m=0}^{r} a_{jklm} R_{1t^l}(s,s_j) R_{s^m t^m}(s_k , s_k) \\
&+ \sum_{l,m=0}^{r} \sum_{j=1}^{n_2} [R_{1t^l}(s,t_j)(R_{2s^m t^m} h_{jlm})(s_k) + (R_{1t^l} \tilde{h}_{jlm})(s) R_{2s^m t^m}(t_i , s_k)]
\end{aligned}
$$

$$+ \sum_{l,m=0}^{r} \int_0^1 R_{1s^l}(s,u)\hat{h}_{lm}(u)R_{2s^mt^m}(u,s_k)\,du$$

$$+ \sum_{l,m=0}^{r} \int_0^1 \int_0^1 R_{1s^l}(s,u)\tilde{H}_{lm}(u,v)R_{2s^mt^m}(v,s_k)\,du\,dv. \tag{15}$$

Thus,

$$\begin{aligned}
\| R_1^{-\frac{1}{2}}R_{2t^m}(\cdot,s_k) \| \leqq{}& | R_{1s^mt^m}(s_k,s_k) |^{\frac{1}{2}} \\
&+ \sum_{l,m=0}^{r} \sum_{j,k=1}^{n_1} | a_{jklm} | \, | R_{1s^l}(s_j,s_j) |^{\frac{1}{2}} R_{2s^mt^m}(s_k,s_k) \\
&+ \sum_{l,m=0}^{r} \sum_{j=1}^{n_2} \{ | R_{1s^lt^l}(t_j,t_j) |^{\frac{1}{2}} | (R_{2s^mt^m}h_{jlm})(s_k) | \\
&+ (\tilde{h}_{jlm}, R_{1s^lt^l}h_{jlm})^{\frac{1}{2}} | R_{2s^mt^m}(s_j,s_k) | \} \\
&+ \sum_{l,m=0}^{r} (R_{2s^mt^m}(s_k,\cdot), \hat{R}_{1s^lt^l,m}R_{2s^mt^m}(\cdot,s_k))^{\frac{1}{2}} \\
&+ \sum_{l,m=0}^{r} (R_{2s^mt^m}(s_k,\cdot), \tilde{R}_{1s^lt^l,m}R_{2s^mt^m}(\cdot,s_k))^{\frac{1}{2}} \\
&< \infty.
\end{aligned} \tag{16}$$

Hence,

$$\| R_1^{-\frac{1}{2}}K_1R_1^{-\frac{1}{2}} \| < \infty.$$

Similarly,

$$\begin{aligned}
\| R_1^{-\frac{1}{2}}K_2R_1^{-\frac{1}{2}} \| \leqq \sum_{l,m=0}^{r} \sum_{j=1}^{n_2} [& | R_{1s^lt^l}(t_j,t_j) |^{\frac{1}{2}} \| R_1^{-\frac{1}{2}}R_{2t^m}h_{jlm} \| \\
&+ (\tilde{h}_{jlm}, R_{1s^lt^l}\tilde{h}_{jlm})^{\frac{1}{2}} \| R_1^{-\frac{1}{2}}R_{2t^m}(\cdot,t_j) \|].
\end{aligned}$$

From (15),

$$\begin{aligned}
\| R_1^{-\frac{1}{2}}R_{2t^m}h_{jlm} \| \leqq{}& (h_{jlm}, R_{1s^mt^m}h_{jlm})^{\frac{1}{2}} \\
&+ \sum_{l,m=0}^{r} \sum_{j,k=1}^{n_1} | a_{jklm} | \, | R_{1s^lt^l}(s_j,s_j) |^{\frac{1}{2}} | (R_{2s^mt^m}h_{jlm})(s_k) | \\
&+ \sum_{l,m=0}^{r} \sum_{j=1}^{n_2} [| R_{1s^lt^l}(t_j,t_j) |^{\frac{1}{2}} | (h_{jlm}, R_{2s^mt^m}h_{jlm}) | \\
&+ (\tilde{h}_{jlm}, R_{1s^lt^l}\tilde{h}_{jlm})^{\frac{1}{2}} | (R_{2s^mt^m}h_{jlm})(t_j) |] \\
&+ \sum_{l,m=0}^{r} (h_{jlm}, R_{2s^mt^m}\hat{R}_{1s^lt^l,m}R_{2s^mt^m}h_{jlm})^{\frac{1}{2}} \\
&+ \sum_{l,m=0}^{r} (h_{jlm}, R_{2s^mt^m}\tilde{R}_{1s^lt^l,m}R_{2s^mt^m}h_{jlm})^{\frac{1}{2}} \\
&< \infty,
\end{aligned}$$

and, from (16)

$$|| R_1^{-\frac{1}{2}} R_{2l'}(\cdot, t_j) || < \infty .$$

Hence,

$$|| R_1^{-\frac{1}{2}} K_2 R_1^{-\frac{1}{2}} || < \infty .$$

Similarly,

$$\begin{aligned}
&|| R_1^{-\frac{1}{2}} K_3 R_1^{-\frac{1}{2}} || \\
&\leqq \sum_{l,m=0}^{r} \left[\sum_i \mu_i \left|\left| R_1^{-\frac{1}{2}} \int_0^1 R_{2l'm}(\cdot, u) \hat{h}_{lm}(u) f_i^{(l)}(u)\, du \right|\right|^2 \right]^{\frac{1}{2}} \\
&\leqq \sum_{l,m=0}^{r} \Bigg\{ | \operatorname{tr} (\hat{R}_{1s'l'l',m} R_{1s'm'l'}) |^{\frac{1}{2}} + \sum_{l',m'=0}^{r} \sum_{j,k=1}^{n_1} | a_{jkl'm'} | \\
&\cdot | R_{1s'l'l'}(s_j, s_i)(R_{2s'm'l'm}(s_k, \cdot), \hat{R}_{1s'l'l',m} R_{2s'm'l'm'}(\cdot, s_k) |^{\frac{1}{2}} \\
&+ \sum_{l',m'=0}^{r} \sum_{j=1}^{n_2} [| R_{1s'l'l'}(t_j, t_i)(R_{2s'm'l'm}, h_{jl'm'}, \hat{R}_{1s'l'l',m} R_{2s'm'l'm}, h_{jl'm'}) |^{\frac{1}{2}} \\
&+ | (\bar{h}_{jl'm'}, R_{1s'l'l'}, \bar{h}_{jl'm'})(R_{2s'm'l'm}(t_j, \cdot), \hat{R}_{1s'l'l',m} R_{2s'm'l'm}(\cdot, t_j)) |^{\frac{1}{2}}] \\
&+ \sum_{l',m'=0}^{r} | \operatorname{tr} (\hat{R}_{1s'l'l',m'} R_{2s'm'l'm} \hat{R}_{1s'l'l',m} R_{2s'm'l'm'}) |^{\frac{1}{2}} \\
&+ \sum_{l',m'=0}^{r} | \operatorname{tr} (\tilde{R}_{1s'l'l',m'} R_{2s'm'l'm} \hat{R}_{1s'l'l',m} R_{2s'm'l'm'}) |^{\frac{1}{2}} \Bigg\} \\
&< \infty .
\end{aligned}$$

Similarly,

$$\begin{aligned}
&|| R_1^{-\frac{1}{2}} K_4 R_1^{-\frac{1}{2}} || \\
&\leqq \sum_{l,m=0}^{r} [\sum_i \mu_i || R_1^{-\frac{1}{2}} R_{2l'm} \tilde{H}_{mi} f_i^{(l)} ||^2]^{\frac{1}{2}} \\
&\leqq \sum_{l,m=0}^{r} \Bigg\{ | \operatorname{tr} (\tilde{R}_{1s'l'l',m} R_{1s'm'l'}) |^{\frac{1}{2}} + \sum_{l',m'=0}^{r} \sum_{j,k=1}^{n_1} | a_{jkl'm'} | \\
&\cdot | R_{1s'l'l'}(s_j, s_i)(R_{2s'm'l'm}(s_k, \cdot), \tilde{R}_{1s'l'l',m} R_{2s'm'l'm'}(\cdot, s_k)) |^{\frac{1}{2}} \\
&+ \sum_{l',m'=0}^{r} \sum_{j=1}^{n_2} [| R_{1s'l'l'}(t_j, t_i)(h_{jl'm'}, R_{2s'm'l'm} \tilde{R}_{1s'l'l',m} R_{2s'm'l'm}, h_{jl'm'}) |^{\frac{1}{2}} \\
&+ | (\bar{h}_{jl'm'}, \tilde{R}_{1s'l'l'}, \bar{h}_{jl'm'})(R_{2s'm'l'm}(t_j, \cdot), \tilde{R}_{1s'l'l',m} R_{2s'm'l'm'}(\cdot, t_j)) |^{\frac{1}{2}}] \\
&+ \sum_{l',m'=0}^{r} | \operatorname{tr} (\hat{R}_{1s'l'l',m'} R_{2s'm'l'm} \tilde{R}_{1s'l'l',m} R_{2s'm'l'm'}) |^{\frac{1}{2}}
\end{aligned}$$

$$+ \sum_{l',m'=0}^{r} | \operatorname{tr} (\tilde{R}_{1s^{l'}t^{l'},m'} R_{2s^{m'}t^{m}} \tilde{R}_{1s^{l}t^{l},m} R_{2s^{m}t^{m'}}) |^{\frac{1}{2}} \Big\}$$

$$< \infty .$$

Therefore, from (13),

$$|| R_1^{-\frac{1}{2}} R_2 R_1^{-\frac{1}{2}} - I || < \infty .$$

(*ii*) We have established in (*i*) that $R_1^{-\frac{1}{2}} R_2 R_1^{-\frac{1}{2}}$ is bounded and densely defined on $\mathfrak{L}_2$. Hence, it has a unique extension to the whole of $\mathfrak{L}_2$, which we denote by M. Since $M - I$ is a Hilbert-Schmidt operator, M has eigenvalues and orthonormal eigenfunctions, which we denote by λ_i and $\varphi_i(t)$, $i = 0, 1, 2, \cdots$. Note $0 < \lambda_i \leqq | M |$, where $| M |$ is the norm of M. Then[13]

$$R_{1s^{l}t^{m}}(s,t) = \sum_{i} (R_1^{\frac{1}{2}}\varphi_i)^{(l)}(s)(R_1^{\frac{1}{2}}\varphi_i)^{(m)}(t).$$

$$0 \leqq l, m \leqq r, \qquad (17)$$

$$R_{2s^{l}t^{m}}(s,t) = \sum_{i} \lambda_i (R_1^{\frac{1}{2}}\varphi_i)^{(l)}(s)(R_1^{\frac{1}{2}}\varphi_i)^{(m)}(t),$$

uniformly on $[0,1] \times [0,1]$.

Let $\{\varphi_{in}\}$ be sequences of functions in the domain of $R_1^{-\frac{1}{2}}$ such that $\varphi_i = \text{l.i.m. } \varphi_{in}$ for each i. Multiply both sides of (12) by $(R_1^{-\frac{1}{2}}\varphi_{in})(s)$ and $(R_1^{-\frac{1}{2}}\varphi_{in})(t)$, integrate with respect to s and t, and let $n \to \infty$. Then, the four terms on the left-hand side become

$$(R_1^{-\frac{1}{2}}\varphi_{in} , K_1 R_1^{-\frac{1}{2}}\varphi_{in})$$

$$= \sum_{l,m=0}^{r} \sum_{j,k=1}^{n_1} a_{jklm} (R_1^{-\frac{1}{2}}\varphi_{in} , R_{1t^{l}}(\cdot, s_j))(R_{2s^{m}}(s_k , \cdot), R_1^{-\frac{1}{2}}\varphi_{in})$$

$$= \sum_{l,m=0}^{r} \sum_{j,k=1}^{n_1} a_{jklm} \sum_{\nu} (\varphi_{in} , \varphi_\nu)(R_1^{\frac{1}{2}}\varphi_\nu)^{(l)}(s_j) \sum_{\nu} \lambda_\nu (R_1^{\frac{1}{2}}\varphi_\nu)^{(m)}(s_k)(\varphi_\nu , \varphi_{in}),$$

where (17) is used for the second equality. By virtue of (17) again, we can define an s-function $R_{1t^{l}}^{\frac{1}{2}}(s,u)$ for any $u \in [0,1]$ by

$$R_{1t^{l}}^{\frac{1}{2}}(s,u) = \text{l.i.m.}_{n\to\infty} \sum_{\nu=0}^{n} \varphi_\nu(s)(R_1^{\frac{1}{2}}\varphi_\nu)^{(l)}(u).$$

Then

$$\sum_{\nu} (\varphi_{in} , \varphi_\nu)(R_1^{\frac{1}{2}}\varphi_\nu)^{(l)}(s_j) = (\varphi_{in} , R_{1t^{l}}^{\frac{1}{2}}(\cdot, s_j)),$$

$$\sum_{\nu} \lambda_\nu (R_1^{\frac{1}{2}}\varphi_\nu)^{(m)}(s_k)(\varphi_\nu , \varphi_{in}) = (R_{1s^{m}}^{\frac{1}{2}}(s_k , \cdot), M\varphi_{in}),$$

and

$$\lim_{n\to\infty} (\varphi_{in} , R_{1t}^{\frac{1}{2}}(\cdot ,s_j)) = (\varphi_i , R_{1t}^{\frac{1}{2}}(\cdot ,s_j)) = (R_1^{\frac{1}{2}}\varphi_i)^{(l)}(s_j),$$

$$\lim_{n\to\infty} (R_{1s}^{\frac{1}{2}}(s_k , \cdot), M\varphi_{in}) = (R_{1s}^{\frac{1}{2}}(s_k , \cdot), M\varphi_i) = \lambda_i(R_1^{\frac{1}{2}}\varphi_i)^{(m)}(s_k).$$

Hence,

$$\lim_{n\to\infty} (R_1^{-\frac{1}{2}}\varphi_{in} , K_1R_1^{-\frac{1}{2}}\varphi_{in}) = \lambda_i \sum_{l,m=0}^{r} \sum_{j,k=1}^{n_1} a_{jklm}(R_1^{\frac{1}{2}}\varphi_i)^{(l)}(s_j)(R_1^{\frac{1}{2}}\varphi_i)^{(m)}(s_k).$$

Similarly,

$$\lim_{n\to\infty} (R_1^{-\frac{1}{2}}\varphi_{in} , K_2R_1^{-\frac{1}{2}}\varphi_{in})$$

$$= \lim_{n\to\infty} \sum_{l,m=0}^{r} \sum_{j=1}^{n_2} [(R_1^{-\frac{1}{2}}\varphi_{in} , R_{1t}^{\frac{1}{2}}(\cdot ,t_j))(R_{2t}h_{jlm} , R_1^{-\frac{1}{2}}\varphi_{in})$$

$$+ (R_1^{-\frac{1}{2}}\varphi_{in} , R_{1t}^{\frac{1}{2}}\tilde{h}_{jlm})(R_{2s}(t_j , \cdot), R_1^{-\frac{1}{2}}\varphi_{in})]$$

$$= \lambda_i \sum_{l,m=0}^{r} \sum_{j=1}^{n_2} (R_1^{\frac{1}{2}}\varphi_i)^{(l)}(t_j)((R_1^{\frac{1}{2}}\varphi_i)^{(m)} , h_{jlm} + \tilde{h}_{jlm}),$$

$$\lim_{n\to\infty} (R_1^{-\frac{1}{2}}\varphi_{in} , K_3R_1^{-\frac{1}{2}}\varphi_{in})$$

$$= \lim_{n\to\infty} \sum_{l,m=0}^{r} \int_0^1 (R_1^{-\frac{1}{2}}\varphi_{in} , R_{1t}^{\frac{1}{2}}(\cdot ,u))\hat{h}_{lm}(u)(R_{2s}(u,\cdot), R_1^{-\frac{1}{2}}\varphi_{in})\, du$$

$$= \lambda_i \sum_{l,m=0}^{r} \int_0^1 (R_1^{\frac{1}{2}}\varphi_i)^{(l)}(u)\hat{h}_{lm}(u)(R_1^{\frac{1}{2}}\varphi_i)^{(m)}(u)\, du,$$

$$\lim_{n\to\infty} (R_1^{-\frac{1}{2}}\varphi_{in} , K_4R_1^{-\frac{1}{2}}\varphi_{in})$$

$$= \lim_{n\to\infty} \sum_{l,m=0}^{r} \int_0^1 \int_0^1 (R_1^{-\frac{1}{2}}\varphi_{in} , R_{1t}^{\frac{1}{2}}(\cdot ,u))\tilde{H}_{lm}(u,v)(R_{2s}(v,\cdot),R_1^{-\frac{1}{2}}\varphi_{in})\, du\, dv$$

$$= \lambda_i \sum_{l,m=0}^{r} ((R_1^{\frac{1}{2}}\varphi_i)^{(l)} , \tilde{H}_{lm}(R_1^{\frac{1}{2}}\varphi_i)^{(m)}).$$

On the other hand, the right-hand side becomes

$$\lim_{n\to\infty} (R_1^{-\frac{1}{2}}\varphi_{in} , (R_2 - R_1)R_1^{-\frac{1}{2}}\varphi_{in}) = \lim_{n\to\infty} (\varphi_{in} , (M - I)\varphi_{in}) = \lambda_i - 1.$$

Hence, by equating both sides and dividing by λ_i,

$$1 - \frac{1}{\lambda_i} = \sum_{l,m=0}^{r} \left[\sum_{j,k=1}^{n_1} a_{jklm}(R_1^{\frac{1}{2}}\varphi_i)^{(l)}(s_j)(R_1^{\frac{1}{2}}\varphi_i)^{(m)}(s_k) \right.$$

$$+ \sum_{j=1}^{n_2} (R_1^{\frac{1}{2}}\varphi_i)^{(l)}(t_j)((R_1^{\frac{1}{2}}\varphi_i)^{(m)}, h_{jlm} + \tilde{h}_{jlm})$$

$$+ \int_0^1 (R_1^{\frac{1}{2}}\varphi_i)^{(l)}(u)\hat{h}_{lm}(u)(R_1^{\frac{1}{2}}\varphi_i)^{(m)}(u)\,du + ((R_1^{\frac{1}{2}}\varphi_i)^{(l)}, \tilde{H}_{lm}(R_1^{\frac{1}{2}}\varphi_i)^{(m)})\Big]. \quad (18)$$

Thus,

$$\sum_i \left(1 - \frac{1}{\lambda_i}\right) = \sum_{l,m=0}^{r} \Big[\sum_{j,k=1}^{n_1} a_{jklm} R_{1s^l t^m}(s_j, s_k)$$

$$+ \sum_{j=1}^{n_2} R_{1s^l t^m}(h_{jlm} + \tilde{h}_{jlm})(t_j) + \int_0^1 R_{1s^l t^m}(u,u)\hat{h}_{lm}(u)\,du + \operatorname{tr}(R_{1s^l t^m}\tilde{H}_{ml})\Big]$$

$$< \infty,$$

where (17) is used repeatedly. Hence,*

$$0 < \prod_{i=0}^{\infty} \lambda_i < \infty.$$

$$\frac{dP_2}{dP_1}(x) = \left(\prod_{i=0}^{\infty} \lambda_i\right)^{-\frac{1}{2}} \exp\left[\tfrac{1}{2} \sum_i \left(1 - \frac{1}{\lambda_i}\right)\eta_i^2(x)\right], \quad \text{a.s., } [P_1],$$

where

$$\eta_i(x) = \underset{n\to\infty}{\text{l.i.m.}}\ (x, R_1^{-\frac{1}{2}}\varphi_{in}), \qquad [P_1, P_2] \qquad i = 0, 1, 2, \cdots. \quad (19)$$

Now, $x^{(l)}(t)$ has the following orthogonal expansion:[13]

$$x^{(l)}(t) = \underset{n\to\infty}{\text{l.i.m.}} \sum_{i=0}^{n} \eta_i(x)(R_1^{\frac{1}{2}}\varphi_i)^{(l)}(t), \qquad [P_1], \qquad 0 \leqq l \leqq r,$$

uniformly in t. Hence, there exists a subsequence of the partial sums $\sum_{i=0}^{n_\nu} \eta_i(x)(R_1^{\frac{1}{2}}\varphi_i)^{(l)}(t)$ which converges a.s. $[P_1]$ to $x^{(l)}(t)$, uniformly in t. Therefore, from (18) and (19),

$$\sum_i \left(1 - \frac{1}{\lambda_i}\right)\eta_i^2(x)$$

$$= \lim_{n_\nu\to\infty} \sum_{i=0}^{n_\nu} \left(1 - \frac{1}{\lambda_i}\right)\eta_i^2(x)$$

$$= \sum_{l,m=0}^{r} \Big[\sum_{j,k=1}^{n_1} a_{jklm} x^{(l)}(s_j)x^{(m)}(s_k) + \sum_{j=1}^{n_2} x^{(l)}(t_j)(x^{(m)}, h_{jlm} + \tilde{h}_{jlm})$$

$$+ \int_0^1 x^{(l)}(u)\hat{h}_{lm}(u)x^{(m)}(u)\,du + (x^{(l)}, \tilde{H}_{lm}x^{(m)})\Big], \quad \text{a.s.} \quad [P_1],$$

which completes the proof of (*ii*).

* See Ref. 3, pp. 1653–1654.

REFERENCES

1. Kadota, T. T., Optimum Reception of Binary Gaussian Signals, B.S.T.J., *43*, November, 1964, pp. 2767–2810.
2. Pitcher, T. S., An Integral Expression for the Log Likelihood Ratio of Two Gaussian Processes, SIAM J. on Applied Math., March, 1966, pp. 228–233.
3. Kadota, T. T., Optimum Reception of Binary Sure and Gaussian Signals, B.S.T.J., *44*, October, 1965, pp. 1921–1658.
4. Loeve, M., *Probability Theory*, 2nd ed., Van Nostrand, Princeton, 1960.
5. Feldman, J., Equivalence and Perpendicularity of Gaussian Processes, Pacific J. Math., *8*, No. 4, 1958, pp. 699–708.
6. Hajek, J., On a Property of Normal Distribution of any Stochastic Process, Czechoslovak Math. J., *83*, 1958, pp. 610–618.
7. Halmos, P. R., *Measure Theory*, Van Nostrand, Princeton, 1950.
8. Root, W. L., Singular Gaussian Measures in Detection Theory, Proc. of Symp. on Time Series Analysis, John Wiley, New York, 1963, pp. 292–315.
9. Kadota, T. T., Differentiation of Karhunen-Loève Expansion and Application to Optimum Reception of Sure Signals in Noise, IEEE Trans. Inform. Theor., April, 1967.
10. Kadota, T. T., Optimum Reception of M-ary Gaussian Signals in Gaussian Noise, B.S.T.J., *44*, November, 1965, pp. 2187–2197.
11. Kadota, T. T., Generalized Maximum Likelihood Test and Minimum Error Probability, IEEE Trans., *IT-12*, 1, January 1966, pp. 65–67.
12. Kadota, T. T., Term-by-term Differentiability of Mercer's Expansion, to appear in Proc. Am. Math. Soc.
13. Kadota, T. T., Simultaneous Diagonalization of Two Covariance Kernels and Application to Second-Order Stochastic Processes, submitted for publication in SIAM J. Appl. Math.

Timing Recovery for Synchronous Binary Data Transmission

By BURTON R. SALTZBERG

(Manuscript received November 15, 1966)

This paper analyzes different methods of adjusting the sampling time for detecting synchronous binary data, based on properties of the random data signal itself. The static error and the variance of the jitter of the resultant sampling instant are calculated where the effects of frequency offset, additive noise, signal overlap, and jitter of the reference source are included.

The threshold crossing timing recovery system adjusts the sampling time in response to the times at which the data signal crosses the amplitude threshold. The sampled-derivative system uses the time derivative of the signal at the sampling time to adjust sampling phase. It is shown that both systems lead to approximately the same amount of jitter in the presence of noise and signal overlap for a given bandwidth of the control loop.

An improved timing recovery system is presented which is constructed by adding correction signals to the sampled-derivative system. This system accounts for intersymbol interference in a manner that tends to set the sampling time at the point of maximum eye opening, where the error probability is minimum for the most adverse message sequence.

1. INTRODUCTION AND SUMMARY OF RESULTS

In synchronous polar binary data transmission, information is sent by serially transmitting either a basic signaling waveform or its negative at fixed time intervals. Modulation may be used to better fit the signal to the channel. At the receiving end, the signal is demodulated and filtered. The resultant baseband signal is sampled periodically, and the polarities at the sampling instants determine the output data. The choice of sampling time is critical for minimizing the error probability due to intersymbol interference and noise, particularly when the signal has been subjected to sharp cutoff filtering. The sampling time is best set by using some properties of the data signal itself.

The problem of timing is particularly acute in pulse code modulation (PCM) systems, where the accumulation of jitter in a long chain of regenerators frequently limits the allowable length of such systems. For this reason, previous studies of timing recovery have concentrated on PCM applications.[1-4] The use of a tuned circuit as the memory element is generally assumed, since this is commonly employed in PCM repeaters.

This paper will concentrate on timing recovery for data transmission applications. The effects of multiple regeneration will not be considered. The recovery of timing will be accomplished by a feedback control system, such as a phase-locked loop. Different methods of generating the error signal for the control loop will be compared.

The received signal, after demodulation and filtering, is of the form

$$s(t) = \sum_{k=-\infty}^{\infty} a_k f(t - \beta T - kT) + n(t) \tag{1}$$

where $\{a_k\}$ is a set of independent random variables, each equal to $+1$ or -1 with equal probability. This may be assured by the use of a scrambler if the data source is itself not random. The basic signaling waveform is $f(t)$. The abscissa of $f(t)$ will be adjusted for each system to be studied so that the desired sampling time of $f(t)$ is $t = 0$. The quantity β is an unknown fractional time delay. Since we are not concerned with absolute time delay between transmitter and receiver, we will assume $| \beta | \leqq \frac{1}{2}$. The additive noise is $n(t)$.

The sampling wave which determines the times at which $s(t)$ is sampled may be represented by

$$q(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT - \gamma T), \tag{2}$$

where γ is a phase that is generally time varying.

The output data is determined by

$$\hat{a}_n = \operatorname{sgn} s(nT + \gamma T), \tag{3}$$

where $\operatorname{sgn} v = v/| v |$. Then

$$\hat{a}_n = \operatorname{sgn} [a_n f(\gamma T - \beta T) + \sum_{k \neq 0} a_{n-k} f(kT + \gamma T - \beta T) + n(t)]. \tag{4}$$

For simplicity, the argument of the noise term is not made explicit since it is of no consequence. Assuming that $f(\gamma T - \beta T)$ is positive, then $\hat{a}_n$ will agree with a_n provided that

$$-a_n[\sum_{k \neq 0} a_{n-k} f(kT + \gamma T - \beta T) + n(t)] < f(\gamma T - \beta T). \tag{5}$$

It is readily seen that the probability that (5) holds depends strongly on $\gamma - \beta$. For each timing recovery system, $f(t)$ is defined so that the desired value of $\gamma - \beta$ is zero.

The principal part of a timing recovery system is a phase detector which examines $s(t)$ and $q(t)$ and attempts to generate an error signal proportional to $\beta - \gamma$. It is not possible to determine β exactly from $s(t)$ since the signaling waveform and the noise are unknown. This paper will consider different methods of forming this estimate of $\beta - \gamma$.

Another essential component of the system is the reference source which is used to generate the sampling wave. Its phase or frequency is adjusted by the error signal in order to form a sampling wave of the proper phase. The error signal may be filtered prior to its use in shifting the phase of the reference source. The reference source may be a local oscillator whose natural frequency is set as close as possible to the bit rate. The reference source may instead be derived from transmitted pilot tones, in which case its frequency is exact, but it might have phase jitter of its own due to channel noise.

Section II describes and analyzes a timing recovery system which uses a threshold crossing phase detector. This detector generates an error signal each time the signal crosses zero. The amplitude of this signal is proportional to the difference between the time of occurrence of the zero crossing and the time of the nearest sampling pulse, displaced by half a bit period. This system tends to choose a sampling instant which is midway between the mean transition times.

The sampled-derivative phase detector is discussed in Section III. This device generates an error signal during each bit interval which is proportional to the time derivative of the signal at the sampling time multiplied by the signal polarity at that time. The sampled-derivative timing recovery system attempts to set the sampling time to coincide with the peak of $f(t)$.

The analysis shows that the performance of these two systems is very similar for a given open loop gain function of the control system, $G(\omega)$. Approximations are made based on the assumption that the phase error is small and that $G(\omega)$ is a narrowband low-pass function compared with the bit rate.

The systems fail if $G(0)$ is finite and the reference source does not agree exactly in natural frequency with the bit rate. If the reference source has the correct frequency and a phase θ, then a static phase error

$$\bar{e} = \frac{\theta - \beta}{1 + G(0)} \tag{6}$$

results in the sampling wave.

Much better performance can be achieved if $G(\omega)$ has a pole at the origin. In this case the system is insensitive to the phase of the reference source. In the presence of a frequency error, Δf, the static error is

$$\bar{e} = -j\,\Delta f \frac{d}{d\omega}\left[\frac{1}{1 + G(\omega)}\right]_{\omega=0}. \tag{7}$$

In addition to any static error, the sampling time will also jitter about its mean value. The variance of this phase jitter is the sum of several components, each due to a different cause. Jitter is produced by jitter in the reference source, by additive noise and by signal overlap. In the case of the threshold crossing system, jitter is also introduced whenever there is a static error.

If the reference source has a jitter whose power spectral density is $S_\eta(\omega)$, then the output fractional jitter will have a variance equal to

$$\sigma_{RS}^2 = \frac{1}{\pi}\int_0^\infty \frac{S_\eta(\omega)}{|\,1 + G(\omega)\,|^2}\,d\omega. \tag{8}$$

This indicates that high-frequency noise components must be removed from the reference source prior to its use for timing recovery.

The jitter produced by the additive noise is

$$\sigma_N^2 = \frac{2[R_n(0) - R_n(T)]}{T[f'(-T/2) - f'(T/2)]^2}\,\omega_1 \tag{9}$$

for the threshold crossing system. $R_n(t)$ is the autocorrelation of the noise and ω_1 is the noise bandwidth of the closed control loop.

$$\omega_1 = \frac{1}{\pi}\int_0^\infty \left|\frac{G(\omega)}{1 + G(\omega)}\right|^2 d\omega. \tag{10}$$

For the sampled-derivative system, the noise leads to jitter variance

$$\sigma_N^2 = -\frac{R_n''(0)}{Tf''^2(0)}\,\omega_1\,. \tag{11}$$

In typical data transmission systems, (9) and (11) are similar in magnitude, and not very sensitive to the shape of $f(t)$ if the noise is similarly filtered.

The jitter variance due to signal overlap is of the form

$$\sigma_S^2 = A_1\omega_1 T + A_2(\omega_2 T)^3, \tag{12}$$

where

$$\omega_2^3 = \frac{1}{\pi}\int_0^\infty \omega^2 \left|\frac{G(\omega)}{1 + G(\omega)}\right|^2 d\omega. \tag{13}$$

If $\omega_1 T$ and $\omega_2 T$ are comparable and much less than unity, then the first term is usually much larger than the second.

For the threshold crossing system,

$$A_1 = \frac{1}{T^2[f'(-T/2) - f'(T/2)]^2} \sum_{k=-\infty}^{\infty} [f(kT + T/2) - f(kT - T/2)] \cdot [2f(kT + T/2) + f(-kT + T/2) - f(-kT - T/2)] \quad (14)$$

and

$$A_2 = \frac{1}{2T^2[f'(-T/2) - f'(T/2)]^2} \left\{ -4f^2(T/2) + 2 \sum_{k=-\infty}^{\infty} f(kT + T/2) f(kT - T/2) - \sum_{k=-\infty}^{\infty} k^2 [f(kT + T/2) - f(kT - T/2)][f(-kT + T/2) - f(-kT - T/2)] \right\}. \quad (15)$$

For the sampled-derivative system,

$$A_1 = \frac{1}{T^2 f''^2(0)} \sum_{k=-\infty}^{\infty} f'(kT)[f'(kT) + f'(-kT)] \quad (16)$$

and

$$A_2 = \frac{-1}{2T^2 f''^2(0)} \sum_{k=-\infty}^{\infty} k^2 f'(kT) f'(-kT). \quad (17)$$

In both cases, $A_1 = 0$ if $f(t)$ is an even function, so the timing recovery systems are very sensitive to asymmetry of the basic signaling waveform. The jitter variances are again comparable for both systems. As may be expected, the jitter increases considerably as the filter used to shape $f(t)$ is made sharper.

There is an additional jitter component for the threshold crossing system whenever there is a static error. Its variance is given by

$$\sigma_e^2 = \bar{e}^2 \omega_1 T. \quad (18)$$

An example is provided in Section IV. A typical data transmission system using a distorted signal is studied so as to illustrate the magnitudes of the above quantities and to indicate the narrowness of loop bandwidth required for satisfactory performance.

In this example it is also seen that neither the threshold crossing timing recovery system nor the sampled-derivative system chooses a mean sampling time which is very near to the time at which the eye pattern has its maximum opening.

Section V describes an improved timing recovery system whose mean sampling time coincides with that of the maximum eye opening. This system is constructed by adding correction signals to the sampled-derivative system in order to account for the effects of intersymbol interference on the mean sampling time.

Finally, an outline of some extensions and modifications of these timing recovery systems is presented in Section VI.

II. THE THRESHOLD CROSSING SYSTEM

Most timing recovery systems make use of the instants at which the data signal crosses the threshold to alter the phase of the sampling wave. A block diagram of a typical threshold crossing timing recovery system is shown in Fig. 1.

The principal part in this system is the threshold crossing detector. This device generates an error pulse each time the signal crosses zero. The amplitude of the error pulse is proportional to the difference between the time of occurrence of the threshold crossing and the time of the nearest pulse of the displaced sampling wave. The displaced sampling wave is

$$q_d(t) = q(t - T/2) = \sum_{n=-\infty}^{\infty} \delta(t - nT - T/2 - \gamma T), \qquad (19)$$

where γ is the phase measured in fractional signal periods. If the axis crossing following the mth sampling time is displaced by $\alpha_m T$,

$$s(mT + T/2 + \alpha_m T) = 0, \qquad |\alpha_m + \gamma| < \tfrac{1}{2}, \qquad (20)$$

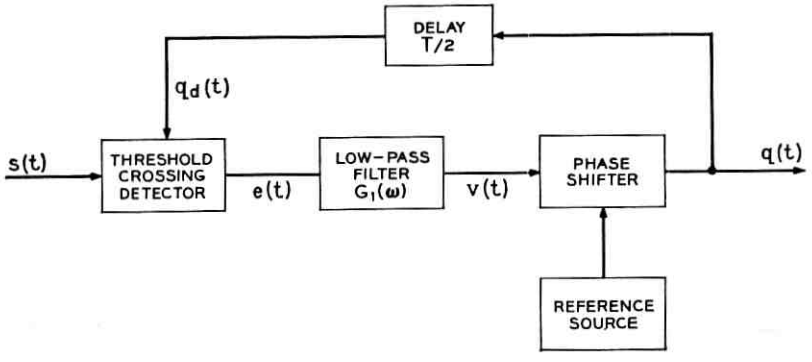


Fig. 1 — Threshold crossing timing recovery system.

then the error signal is

$$e(t) = \sum_m K_1(\alpha_m - \gamma)\ \delta(t - mT), \tag{21}$$

where K_1 is a gain constant. The error pulse has been represented in (21) as an ideal impulse function. Since the error signal will be passed through a narrow low-pass filter, the response will be virtually identical to that when a more realistic pulse of the same area is used. Similarly, the effects of variation of the position of the pulse within the interval may be neglected, since the low-frequency components of the error signal are substantially uneffected.

We will now determine the threshold crossing time α_m as a function of the signal overlap and noise. Substituting (1) into (20) yields

$$\sum_{k=-\infty}^{\infty} a_{m-k} f(kT + T/2 - \beta T + \alpha_m T) + n(t) = 0. \tag{22}$$

If

$$f(T/2) + f(-T/2) > \sum_{k\neq 0,-1} | f(kT + T/2) | + n(t) \tag{23}$$

then a crossing will occur following the mth bit, if and only if $a_m = -a_{m+1}$. When $a_m = -a_{m+1}$ and (23) holds, (22) may be written as

$$a_m[f(T/2 - \beta T + \alpha_m T) - f(-T/2 - \beta T + \alpha_m T)] + \sum_{k\neq 0,-1} a_{m-k} f(kT + T/2 - \beta T + \alpha_m T) + n(t) = 0. \tag{24}$$

If $\alpha_m - \beta$ is small, we may approximate (24) by the first terms of its Taylor series expansion.

$$a_m[f(T/2) - f(-T/2)] + a_m(\alpha_m - \beta)T[f'(T/2) - f'(-T/2)] + \sum_{k\neq 0,-1} a_{m-k} f(kT + T/2) + n(t) \approx 0. \tag{25}$$

Let the abscissa of the function $f(t)$ be adjusted so that

$$f(T/2) = f(-T/2). \tag{26}$$

Then define

$$b = f'(-T/2) - f'(T/2). \tag{27}$$

We may now solve (25) for α_m.

$$\alpha_m \approx \beta + \frac{a_m}{bT}\left[\sum_{k\neq 0,-1} a_{m-k} f(kT + T/2) + n(t)\right]. \tag{28}$$

If we let

$$d_n = \begin{cases} 0, & \text{if } a_n = a_{n+1} \\ 1, & \text{if } a_n = -a_{n+1} \end{cases} \tag{29}$$

then the error signal (21) becomes

$$e(t) = K_1 \sum_{n=-\infty}^{\infty} d_n \Big\{ \beta - \gamma + \frac{a_n}{bT} \Big[\sum_{k \neq 0, -1} a_{n-k} f(kT + T/2) + n(t) \Big] \Big\} \delta(t - nT). \tag{30}$$

The error signal is passed through the filter $G_1(\omega)$ and then shifts the phase of a reference source. The reference source may be a local oscillator whose frequency is tuned as closely as possible to the signaling rate. Alternatively, the reference source may be derived from pilot tones which are transmitted along with the data. In the latter case, there is no error in the average frequency of the reference source, but its phase may be poorly related to that of the data signal and may also be perturbed by noise. In either case, the reference source generates a signal of the form

$$r(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT - \tau T). \tag{31}$$

When a local oscillator is used,

$$\tau \approx \Delta f t + \theta, \qquad \Delta f T \ll 1 \tag{32}$$

where Δf is the frequency offset of the oscillator and θ is an arbitrary constant. When the reference source is derived from pilot tones,

$$\tau = \theta + \eta(t) \tag{33}$$

where $\eta(t)$ is a zero mean random variable.

The sampling wave is formed by shifting the phase of the reference source by an amount proportional to the value of the filtered error signal.

$$q(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT - \tau T - K_2 v T), \tag{34}$$

where v is the filtered version of $e(t)$ and K_2 is the proportionality constant. Comparing (34) with (2),

$$\gamma = \tau + K_2 v = \tau + K_2 \int g_1(t - x) e(x)\, dx. \tag{35}$$

In the phase-locked loop type of control system, the frequency of the reference source is adjusted rather than its phase. This, however, is completely equivalent to integrating the error signal prior to phase adjustment. The block diagram is therefore valid for the phase-locked loop provided that the filter includes a pole at the origin. As will be shown, this pole is highly desirable and, in some cases, absolutely essential.

Substituting (35) into (30)

$$e(t) = K_1 \sum_{n=-\infty}^{\infty} d_n \Big\{ \beta - \tau - K_2 \int g_1(t - x) e(x)\, dx + \frac{a_n}{bT} \Big[\sum_{k \neq 0, -1} a_{n-k} f(kT + T/2) + n(t) \Big] \Big\} \delta(t - nT). \tag{36}$$

Let

$$e_2(t) = \frac{2}{K_1} e(t) \tag{37}$$

$$g_2(t) = \frac{K_1 K_2}{2} g_1(t) \tag{38}$$

$$e(t) = \sum_{n=-\infty}^{\infty} e(n)\, \delta(t - nT) \tag{39}$$

and normalize the time variable so that $T = 1$. Then (36) can be written as

$$e_2(n) = 2d_n \Big\{ \beta - \tau - \sum_{k=-\infty}^{n} g_2(n - k) e_2(k) + \frac{a_n}{b} \Big[\sum_{k \neq 0, -1} a_{n-k} f(k + 1/2) + n(t) \Big] \Big\}. \tag{40}$$

A model of the threshold crossing timing recovery system which conforms with (40) is shown in Fig. 2. This is not a time-invariant linear system because of the presence of the multiplier.

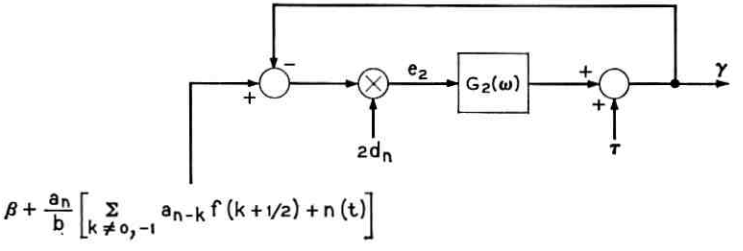


Fig. 2 — Model of threshold crossing system.

In Appendix A, it is shown that

$$\overline{e_2(n)} = \beta - \bar{\tau} - g_2(0)\overline{e_2(n)} - \sum_{k=-\infty}^{n} g_2(n-k)\overline{e_2(k)}. \tag{41}$$

This system may be readily analyzed either by means of the z-transform, or equivalently, by discrete Fourier analysis.[2] The discrete Fourier transform is given by

$$X(\omega) = \sum_{n=-\infty}^{\infty} x(n) \exp(-j\omega n). \tag{42}$$

If $G_2(\omega)$ is bandlimited to $|\omega| < \pi$, which is approximately true in all cases of interest, then these transforms will coincide with the true Fourier transforms.

The solution of (41) in terms of Fourier transforms is

$$\overline{E_2}(\omega) = \frac{\beta(\omega) - \bar{\tau}(\omega)}{1 + g_2(0) + G_2(\omega)}, \tag{43}$$

where

$$g_2(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} G_2(\omega)\, d\omega. \tag{44}$$

The static error can now be found from

$$\bar{\gamma}(\omega) - \beta(\omega) = G_2(\omega)E_2(\omega) + \bar{\tau}(\omega) - \beta(\omega) \tag{45}$$

$$\bar{\gamma}(\omega) - \beta(\omega) = \frac{1 + g_2(0)}{1 + g_2(0) + G_2(\omega)}[\bar{\tau}(\omega) - \beta(\omega)]. \tag{46}$$

In particular, if $\beta(t)$ is a constant, β_0, and $\tau(t)$ is given by (32), then

$$\bar{\gamma}(\omega) - \beta_0\,\delta(\omega) = 2\pi\frac{1 + g_2(0)}{1 + g_2(0) + G_2(\omega)}[j\,\Delta f\,\delta'(\omega) + (\theta - \beta_0)\delta(\omega)] \tag{47}$$

$$\bar{\gamma}(t) - \beta_0 = \frac{1 + g_2(0)}{1 + g_2(0) + G_2(0)}[\Delta f t + \theta - \beta_0] - j\,\Delta f\frac{d}{d\omega}\left[\frac{1 + g_2(0)}{1 + g_2(0) + G_2(\omega)}\right]_{\omega=0}. \tag{48}$$

If $G_2(0)$ is finite, then the first term is a steadily increasing error and the system fails. If $\Delta f = 0$, the system does not fail, but a static error will be present due to the arbitrary phase values, β_0 and θ. It is, therefore, highly desirable that $G(\omega)$ have a pole at the origin. In the presence of frequency offset, this pole is essential. In this case, only

the last term in (48) is nonzero, so that a frequency offset leads to a static error. The system is completely insensitive to the values of the arbitrary phases, β_0 and θ, except during initial start-up.

For the system to be stable, it is required that $1 + g_2(0) + G_2(\omega)$ have no zeros in the half plane $Im(\omega) < 0$. In most cases, $g_2(0) \ll 1$, so that the usual criteria for stability apply to $G_2(\omega)$.

We now wish to calculate the variance of γ, or the mean square jitter. From (40) and (41), and assuming the static error to be constant,

$$e_3(n) = e_2(n) - \bar{e}_2 = 2d_n\Big[\bar{e}_2 + g_2(0)\bar{e}_2 - \sum_{k=-\infty}^{n} g_2(n-k)e_3(k) - \eta(t) + \frac{a_n}{b}n(t)\Big] + z(n) - \bar{e}_2, \tag{49}$$

where

$$z(n) = \frac{2}{b}d_n a_n \sum_{k\neq 0,-1} a_{n-k} f(k + 1/2). \tag{50}$$

Let

$$x(n) = 2d_n\Big[\bar{e}_2 + g_2(0)\bar{e}_2 + \frac{a_n}{b}n(t)\Big] + z(n) - \bar{e}_2. \tag{51}$$

Then (49) may be written as

$$e_3(n) = x(n) - 2d_n\eta(t) - 2d_n\sum_{k=-\infty}^{n} g_2(n-k)e_3(k). \tag{52}$$

The zero mean component of the output phase error is

$$\gamma_1(n) = \gamma(n) - \bar{\gamma} = \eta(t) + \sum_{k=-\infty}^{n} g_2(n-k)e_3(k) \tag{53}$$

$$\gamma_1(n) = \eta(t) + \sum_{k=-\infty}^{n} g_2(n-k)[x(k) - 2d_k\gamma_1(k)]. \tag{54}$$

The autocorrelation of (54) is

$$\sum_k \sum_l E\{[2d_k g(n-k) + \delta_{nk}][2d_l g(n+m-l) + \delta_{n+m,l}]\gamma_k\gamma_l\} = R_\eta(m) + \sum_k \sum_l g_2(n-k)g_2(n+m-l)R_x(k-l), \tag{55}$$

where $E(v)$ denotes the expected value of v and $R_v(m) = E[v(n)v(n+m)]$ is the autocorrelation of v.

In almost all cases of practical interest, $g_2(0) \ll 1$. Then we may approximate

$$\bar{x} = g_2(0)\bar{e}_2 \approx 0 \tag{56}$$

$$\overline{d_n\gamma(n)} = \tfrac{1}{2}\bar{x} \approx 0 \tag{57}$$

and

$$\bar{e}_2 = \frac{\beta_0 - \bar{\gamma}}{1 + g_2(0)} \approx \beta_0 - \bar{\gamma}. \tag{58}$$

Subject to these approximations, we can evaluate the discrete Fourier transform of (55). After some algebraic manipulation, the result obtained is

$$| 1 + G_2(\omega) |^2 S_\gamma(\omega) + | G_2(\omega) |^2 R_\gamma(0) = S_\eta(\omega) + | G_2(\omega) |^2 S_x(\omega), \tag{59}$$

where

$$S_v(\omega) = \sum_{m=-\infty}^{\infty} R_v(m) \exp(-jm\omega) \tag{60}$$

is the power spectral density of v.

The variance of γ is calculated as

$$\sigma_\gamma^2 = R_\gamma(0) = \frac{1}{\pi}\int_0^\infty S_\gamma(\omega)\, d\omega \tag{61}$$

$$\sigma_\gamma^2 = \frac{\frac{1}{\pi}\int_0^\infty \left|\frac{G_2(\omega)}{1 + G_2(\omega)}\right|^2 S_x(\omega)\, d\omega + \frac{1}{\pi}\int_0^\infty \frac{S_\eta(\omega)}{| 1 + G_2(\omega) |^2}\, d\omega}{1 + \frac{1}{\pi}\int_0^\infty \left|\frac{G_2(\omega)}{1 + G_2(\omega)}\right|^2 d\omega}. \tag{62}$$

Since $G_2(\omega)$ is narrowband, we may assume that

$$\int_0^\infty \left|\frac{G_2(\omega)}{1 + G_2(\omega)}\right|^2 d\omega \ll 1 \tag{63}$$

and therefore,

$$\sigma_\gamma^2 \approx \frac{1}{\pi}\int_0^\infty \left|\frac{G_2(\omega)}{1 + G_2(\omega)}\right|^2 S_x(\omega)\, d\omega + \frac{1}{\pi}\int_0^\infty \frac{S_\eta(\omega)}{| 1 + G_2(\omega) |^2}\, d\omega. \tag{64}$$

The second term indicates that low-frequency components of the reference source noise are attenuated while high-frequency components are not. Therefore, if the reference source is derived from transmitted pilot tones, it should be filtered to a narrow bandwidth before being used in the timing recovery system.

In Appendix B, $S_x(\omega)$ is evaluated for $|\omega| \ll 1$. This is the only region of interest when $G_2(\omega)$ is a sufficiently narrow low-pass filter.

$$S_x(\omega) = \bar{e}_2^2 + \frac{2}{b^2}[R_n(0) - R_n(1)] + A_1 + A_2\omega^2, \tag{65}$$

where

$$A_1 = \frac{1}{b^2}\sum_{k=-\infty}^{\infty}[f(k+1/2) - f(k-1/2)] \cdot [2f(k+1/2) + f(-k+1/2) - f(-k-1/2)] \tag{66}$$

and

$$A_2 = \frac{1}{2b^2}\left\{-4f^2(1/2) + 2\sum_{k=-\infty}^{\infty} f(k+1/2)f(k-1/2) - \sum_{k=-\infty}^{\infty} k^2[f(k+1/2) - f(k-1/2)][f(-k+1/2) - f(-k-1/2)\right\}. \tag{67}$$

If we let

$$\omega_1 = \frac{1}{\pi}\int_0^{\infty}\left|\frac{G_2(\omega)}{1+G_2(\omega)}\right|^2 d\omega \tag{68}$$

and

$$\omega_2^3 = \frac{1}{\pi}\int_0^{\infty}\omega^2\left|\frac{G_2(\omega)}{1+G_2(\omega)}\right|^2 d\omega \tag{69}$$

then (64) becomes

$$\sigma_\gamma^2 = \bar{e}_2^2\omega_1 + \frac{2}{b^2}[R_n(0) - R_n(1)]\omega_1 + A_1\omega_1 + A_2\omega_2^3 + \frac{1}{\pi}\int_0^{\infty}\frac{S_\eta(\omega)}{|1+G_2(\omega)|^2}d\omega. \tag{70}$$

This equation is given in the summary with the normalization $T = 1$ removed. An application to a typical data transmission system is given in Section IV.

It should be noted from (69) that ω_2 will be unbounded unless $G_2(\omega)$ has at least two more poles than zeros. Good design of $G_2(\omega)$ requires that the second pole (assuming no zeros) occur somewhere in the vicinity of gain crossover. In this case, ω_2 is approximately equal to ω_1.

The first term of (70) indicates that the standard deviation of the jitter caused by frequency offset will be much less than the mean

of that error. The second term is the jitter produced by the additive noise. In typical data transmission systems, the signal is filtered such that $R_n(1) \approx 0$. In that case, the variance of the jitter is directly proportional to the noise power. Of particular interest is the jitter produced by signal overlap. It can be seen from (66) that $A_1 = 0$ if $f(t)$ is an even function. In that case, the jitter variance is proportional to the cube of the system bandwidth, and can be made quite small by the use of narrow filtering. If $f(t)$ is not an even function, then the third term will usually be much larger than the fourth term. In either case, the jitter will greatly increase as the filter used to shape $f(t)$ is made sharper.

III. THE SAMPLED–DERIVATIVE DETECTOR

An alternative method of adjusting the phase of the sampling wave makes use of the time derivative of the signal at the sampling times. Implementation of a sampled-derivative timing recovery system is about equally complex as a threshold crossing system.

Except for the manner in which the error signals are generated, the control loop is the same for both systems. Fig. 1 may be used to describe the sampled-derivative system if the delay in the feedback path is eliminated and a sampled-derivative detector is substituted for the threshold crossing detector.

The sampled-derivative detector generates an error pulse during each bit interval whose amplitude is proportional to the time derivative of the data signal at the sampling time, multiplied by the polarity of the signal at that time

$$e(t) = K_3 \sum_n \operatorname{sgn} [s(nT + \gamma T)]s'(nT + \gamma T)\ \delta(t - nT). \tag{71}$$

However, the output data is generated by setting

$$\hat{a}_n = \operatorname{sgn} [s(nT + \gamma T)], \tag{72}$$

where $\hat{a}_n$ is the receiver decision on the nth bit. If the error rate is low, $\hat{a}_n = a_n$ with high probability, and (71) may be approximated by

$$e(t) \approx K_3 \sum_n a_n s'(nT + \gamma T)\ \delta(t - nT) \tag{73}$$

if the effect of errors is neglected.

Using (1),

$$e(t) = K_3 \sum_n a_n[\sum_k a_{n-k} f'(kT + \gamma T - \beta T) + n'(t)]\ \delta(t - nT). \tag{74}$$

The abscissa of $f(t)$ in this case is adjusted such that the origin coincides with the peak of $f(t)$.

$$f'(0) = 0. \tag{75}$$

If the phase error $\gamma - \beta$ is small, we may approximate (74) by the first terms of its Taylor series expansion

$$e(t) \approx K_3 \sum_n a_n[a_n(\gamma - \beta)Tf''(0) + \sum_{k\neq 0} a_{n-k}f'(kT) + n'(t)]\ \delta(t - nT). \tag{76}$$

As in the previous case,

$$e(t) = \sum_n e(n)\ \delta(t - nT) \tag{77}$$

$$\gamma = K_2 \int_{-\infty}^{t} g_1(t - x)e(x)\ dx + \tau \tag{78}$$

and we normalize the time variable by setting $T = 1$. Equation (76) may now be written as

$$e(n) = K_3[(\tau - \beta)f''(0) + a_n \sum_{k\neq 0} a_{n-k}f'(k) + a_n n'(t) + K_2 f''(0) \sum_{k=-\infty}^{n} g_1(n - k)e(k)\Big]. \tag{79}$$

Let

$$e_3(t) = -\frac{e(t)}{K_3 f''(0)} \tag{80}$$

and

$$g_2(t) = -K_2K_3f''(0)g_1(t). \tag{81}$$

Then (79) becomes

$$e_3(n) = \beta - \tau - y(n) - \sum_{k=-\infty}^{n} g_3(n - k)e_3(k), \tag{82}$$

where

$$y(n) = \frac{a_n}{f''(0)}\ [\sum_{k\neq 0} a_{n-k}f'(k) + n'(t)]. \tag{83}$$

Unlike the threshold crossing system, the sampled-derivative system is a time-invariant linear one when the phase error is small. A model of the system conforming with (82) is shown in Fig. 3. This model may be readily analyzed because of its linearity.

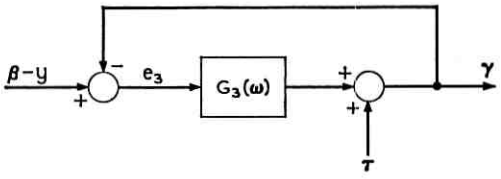


Fig. 3 — Model of sampled-derivative system.

$$\gamma(n) = \sum_{k=-\infty}^{n} g_3(n - k)e_3(k) + \tau \tag{84}$$

$$\gamma(n) = \sum_{k=-\infty}^{n} g_3(n - k)[\beta - y(k) - \gamma(k)] + \tau. \tag{85}$$

The mean error may be found from

$$\gamma(\omega)[1 + G_3(\omega)] = G_3(\omega)\beta(\omega) + \bar{\tau}(\omega) \tag{86}$$

$$\gamma(\omega) - \beta(\omega) = -E_3(\omega) = \frac{\bar{\tau}(\omega) - \beta(\omega)}{1 + G_3(\omega)}. \tag{87}$$

Equation (87) is identical to (46) if $G_3(\omega)$ is substituted for $G_2(\omega)/[1 + g_2(0)]$. All the comments of Section II concerning the static error and the desirability of $G(\omega)$ having a pole at the origin therefore, also apply to the sampled-derivative system.

The variance of γ when the mean error is constant will now be found.

$$\gamma_1(n) = \gamma(n) - \bar{\gamma} = -\sum_{k=-\infty}^{\infty} g_3(n - k)[y(k) + \gamma_1(k)] + \eta(t), \tag{88}$$

where $\eta(t)$ is again the reference source jitter. In terms of the power spectral densities,

$$S_\gamma(\omega) = \left| \frac{G_3(\omega)}{1 + G_3(\omega)} \right|^2 S_v(\omega) + \frac{S_\eta(\omega)}{|1 + G_3(\omega)|^2}. \tag{89}$$

In Appendix C it is shown that, for $|\omega| \ll 1$,

$$S_v(\omega) \approx -\frac{R_n''(0)}{f''^2(0)} + A_3 + A_4\omega^2, \tag{90}$$

where

$$A_3 = \frac{1}{f''^2(0)} \sum_{k=-\infty}^{\infty} f'(k)[f'(k) + f'(-k)] \tag{91}$$

and

$$A_4 = -\frac{1}{2f''^2(0)} \sum_{k=-\infty}^{\infty} k^2 f'(k) f'(-k). \tag{92}$$

The variance of γ is then

$$\sigma_\gamma^2 = \frac{1}{\pi} \int_0^\infty S_\gamma(\omega)\, d\omega \tag{93}$$

$$\sigma_\gamma^2 = -\frac{R_n''(0)}{f''^2(0)} \omega_1 + A_3\omega_1 + A_4\omega_2^3 + \frac{1}{\pi} \int_0^\infty \frac{S_\eta(\omega)}{|\, 1 + G_3(\omega)\,|^2}\, d\omega, \tag{94}$$

where

$$\omega_1 = \frac{1}{\pi} \int_0^\infty \left| \frac{G_3(\omega)}{1 + G_3(\omega)} \right|^2 d\omega \tag{95}$$

and

$$\omega_2^3 = \frac{1}{\pi} \int_0^\infty \omega^2 \left| \frac{G_3(\omega)}{1 + G_3(\omega)} \right|^2 d\omega. \tag{96}$$

There is a very strong similarity between (94) and (70). The last terms are identical, so that it is just as important in the sampled-derivative system as in the threshold crossing system that high-frequency noise components be removed from the referenced source prior to use for timing recovery.

The jitter due to additive noise is proportional to the power of the derivative of the noise. The example in the next section illustrates that this is not serious if the noise is bandlimited to the same frequency range as the signal. However, if any high-frequency noise is allowed to enter the receiver beyond the signal filter, the jitter will be greatly increased.

The jitter due to signal overlap is very similar to that of the threshold crossing system. If $f(t)$ is an even function, then its derivative will be odd, and $A_3 = 0$. Both A_3 and A_4 increase markedly as the spectrum of $f(t)$ is made sharper.

Finally, unlike the threshold crossing system, there is no additional jitter term due to static phase error. This jitter component is eliminated because there is an error pulse generated during each bit interval.

IV. AN EXAMPLE

In order to illustrate the results of the previous two sections, the output jitter of both a threshold crossing timing recovery system and

a sampled-derivative system will be calculated for a typical received data transmission signal.

The signal to be considered is one using half raised-cosine amplitude shaping which is distorted by linear delay distortion. For simplicity, it is normalized so that $T = 1$ and the undistorted peak of the signal is 1.

$$F(\omega) = A(\omega) \exp [j\varphi(\omega)] \tag{97}$$

$$A(\omega) = \begin{cases} 1, & 0 < \omega < \frac{\pi}{2} \\ \cos^2 \frac{\omega - \frac{\pi}{2}}{2}, & \frac{\pi}{2} < \omega < \frac{3\pi}{2} \end{cases} \tag{98}$$

$$\varphi(\omega) = \frac{3}{4\pi} \omega^2, \qquad \omega > 0. \tag{99}$$

The outline of the "eye pattern" for this signal is shown in Fig. 4, along with the central portion of $f(t)$. The eye pattern is formed by superimposing the signals of all possible message sequences. Closing of the eye is due to signal overlap.

The time of maximum eye opening is the optimum sampling time in a minimax sense. When such a sampling instant is chosen, then the error

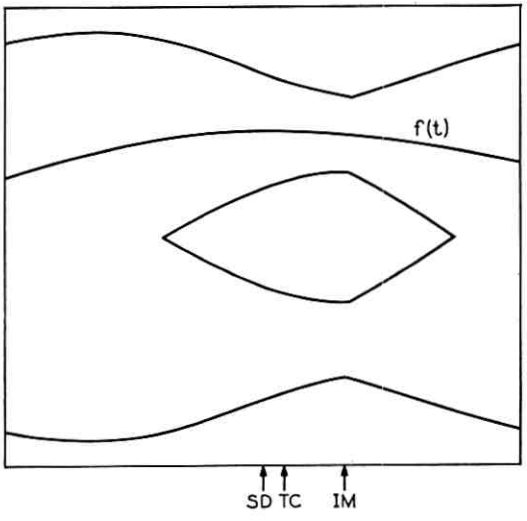


Fig. 4—Eye pattern outline and mean sampling times for a typical distorted data signal.

probability in the presence of additive noise is minimized for the most adverse message sequence, where all adjacent bits overlap in a manner that subtracts from the bit being detected.

The mean sampling times are also shown in Fig. 4 for the case of no static error. The mean sampling time of threshold crossing system, TC, is such that $f(TC + \frac{1}{2}) = f(TC - \frac{1}{2})$. The mean sampling time of the sampled-derivative system, SD, coincides with the peak of $f(t)$. It is seen that the threshold crossing system chooses a better mean sampling time than the sampled-derivative system, yet both systems miss the maximum of the eye opening by a large amount.

In order to compare the jitter due to noise, a particular noise spectrum must be considered. Here it will be assumed that the noise is white noise which has been passed through a receive filter matched to the undistorted signal. In this case, the noise power spectral density is of the form

$$N(\omega) = \frac{\sigma_n^2 A(\omega)}{\frac{1}{\pi}\int_0^\infty A(\omega)\, d\omega} \tag{100}$$

so that

$$R_n(0) = \sigma_n^2 , \tag{101}$$

$$R_n(1) = 0, \tag{102}$$

and

$$R_n''(0) = -\frac{\int_0^\infty \omega^2 A(\omega)\, d\omega}{\int_0^\infty A(\omega)\, d\omega}\,\sigma_n^2 . \tag{103}$$

From (70), the rms jitter due to noise for the threshold crossing system is calculated to be

$$\sigma_\gamma = 0.615\sigma_n\sqrt{\omega_1} . \tag{104}$$

For the sampled-derivative system, this quantity is calculated from (94).

$$\sigma_\gamma = 0.612\sigma_n\sqrt{\omega_1} . \tag{105}$$

The results are virtually identical. In either case, if $\sigma_n = 0.1$ (signal-to-noise ratio of 20 dB) and $\omega_1 = 0.01$, then the rms jitter due to noise alone will be 0.61 percent. It should be mentioned that several other signal pulse shapes were examined, and it was found that the jitter due to noise was not very sensitive to pulse shape.

In order to calculate the jitter due to signal overlap, A_1 and A_3 were computed from (66) and (91). A_2 and A_4 were small enough to have negligible effect on the jitter. The resultant rms jitter is $0.287\sqrt{\omega_1}$ for the threshold crossing system and $0.265\sqrt{\omega_1}$ for the sampled-derivative system. If $\omega_1 = 0.01$, the jitter is 2.87 percent and 2.65 percent, respectively. This is by no means negligible, and illustrates the need for very narrow filtering in the timing recovery control loop. Again there is little difference in the performance of the two systems.

To observe the effects of asymmetry of the signal pulse, let us consider the same signal without phase distortion. Both timing recovery systems will then set a mean sampling time at the best point. The jitter due to signal overlap is greatly reduced since A_1 and A_3 are zero. The computed values of A_2 and A_4 are 0.11 and 0.32, respectively. If $\omega_3 = 0.01$, then the rms jitter due to signal overlap is only 0.01 percent for the threshold crossing detector and 0.03 percent for the sampled-derivative detector. Both values are completely negligible. It may be concluded from this calculation that both timing recovery systems are very sensitive to asymmetry of the signal waveform, both in terms of choosing the average sampling time and the resultant jitter about that time.

V. AN IMPROVED TIMING RECOVERY SYSTEM

It was seen in the previous example that both the threshold crossing timing recovery system and the sampled-derivative system led to average sampling times which differed considerably from the time of maximum eye opening. However, it is possible to modify the sampled-derivative system so that it does seek the time of maximum eye opening as the average sampling time.

At any time t_0, the signal amplitude for the worst message sequence, assuming the current bit is 1, is

$$D(t_0) = f(t_0) - \sum_{k\neq 0} | f(t_0 + kT) |. \tag{106}$$

In the region where the eye is open, $D(t_0) > 0$, and the eye opening is equal to $2D(t_0)$. If a sampling time t_0 is chosen such that $D(t_0) < 0$, then errors will occur for some sequences even in the absence of noise.

An experimental examination of the eye patterns of a large number of actual data transmission systems indicates that $D(t_0)$ is almost always a concave function of t_0. Therefore, if t_0 is adjusted according to the gradient of $D(t_0)$, then the maximum of D will be found.

It is therefore desired to generate an error signal whose average

value is

$$\overline{e(t)} = KD'(t - \beta T + \gamma T). \tag{107}$$

If we again normalize the system so that $T = 1$,

$$\overline{e(t)} = K[f'(t - \beta + \gamma) - \sum_{k \neq 0} f'(t + k - \beta + \gamma) \operatorname{sgn} f(t + k - \beta + \gamma)]. \tag{108}$$

Equation (108) exists and is continuous except at those points t_k where $f(t_k + k - \beta + \gamma) = 0$.

Fig. 5 is a block diagram of a system which generates an error signal whose average is given by (108). The first term of (108) is the average error signal of the sampled-derivative detector discussed in Section III. The improved timing recovery system therefore, will consist of a sampled-derivative system with added correction signals. Enough correction terms are used to account for those adjacent bits which may be expected to overlap significantly into the bit interval under consideration.

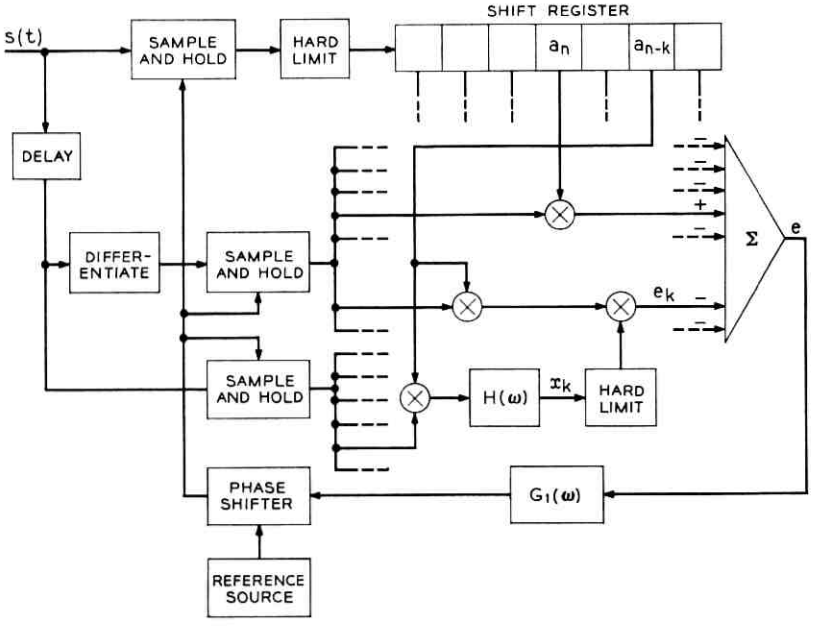


Fig. 5—Improved timing recovery system.

The kth correction signal must have an average value

$$\overline{e_k(t_1)} = Kf'(k + t_1) \operatorname{sgn} f(k + t_1), \tag{109}$$

where

$$t_1 = t - \beta + \gamma. \tag{110}$$

This correction signal is formed by first generating an auxiliary signal $x_k(n)$ whose polarity is expected to agree with that of $f(k + t_1)$. The derivative of the signal at the current sampling time is multiplied by the polarity of the signal at a time displaced from the current time by k bit intervals. The result is multiplied by the polarity of the auxiliary signal to form the correction signal.

$$e_k(n + t_1) = Ks'(n + t_1) \operatorname{sgn} s(n - k + t_1) \operatorname{sgn} x_k(n + t_1). \tag{111}$$

In order to account for overlap into leading pulses as well as lagging pulses, a fixed delay must be built into the system, as indicated in Fig. 5. This delay is equal to half the shift register length, so that the central cell of the shift register stores the polarity of the current bit, while the other cells store the polarities of preceding and succeeding bits.

The auxiliary signal is formed by multiplying the value of the signal at the current sampling time by the polarity of the signal which preceded this signal by k bit intervals. If k is negative, the polarity of a succeeding bit is used. The resultant is filtered by a narrow filter, $H(\omega)$, to form the auxiliary signal x_k.

$$x_k(n) = \sum_m h(n - m)s(m) \operatorname{sgn} s(m - k), \tag{112}$$

where the time displacement t_1 is ignored.

If we assume that the error rate is low, as was done in Section III, then we may approximate sgn $s(n) = a_n$ with little loss of accuracy. Using this approximation and substituting (1) into (112),

$$x_k(n) = \sum_m h(n - m)[a_{m-k}n(t) + f(k) + a_{m-k} \sum_{p \neq k} a_{m-p}f(p)]. \tag{113}$$

The mean of x_k is

$$\overline{x_k} = f(k)H(0). \tag{114}$$

In Appendix D it is shown that the variance of x_k is

$$\sigma_{x_k}^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} |H(\omega)|^2 \left[\sigma_n^2 + \frac{1}{2\pi} \int_{-\infty}^{\infty} |F(u)|^2 \, du + P_k(\omega) \right] d\omega, \tag{115}$$

where $P_k(\omega)$ is the Fourier transform of

$$p_k(t) = f(k - t)f(k + t). \tag{116}$$

If $H(\omega)$ is sufficiently narrowband, then it can be seen from (115) that the variance of x_k will be small. Then, if $f(k)$ is not too small, it may be expected that the polarity of x_k will agree with that of $f(k)$ with high probability.

We will now examine the mean value of the correction signal. Using the assumption of low error probability and substituting (1) into (111), the correction signal is

$$e_k(n + t_1) = Ka_{n-k} \sum_l a_{n-l} f'(n + t_1) \operatorname{sgn} x_k(n + t_1). \tag{117}$$

In order to find the mean of (117), we make the approximation that $x_k(n)$ is independent of a_{n-k} and a_{n-l}. As a justification of this approximation, note from (113) that

$$\overline{x_k} \mid a_{n-k}, a_{n-l} = f(k)H(0) + f(l)h(0) + f(2k - l)h(l - k) \tag{118}$$

$$\overline{x_k} \mid a_{n-k}, a_{n-l} \approx \overline{x_k} \tag{119}$$

since $h(n) \ll H(0)$ for a narrowband filter. Let

$$P_k = \text{Prob } [\operatorname{sgn} x_k = \operatorname{sgn} f(k)]. \tag{120}$$

Then

$$\overline{e_k} \approx K(2P - 1)f'(k + t_1) \operatorname{sgn} f(k + t_1). \tag{121}$$

When the magnitude of $f(k + t_1)$ is sufficiently large, $P \approx 1$ and the mean of the kth correction signal is approximately the desired value. In the vicinity of a zero of $f(k + t_1)$, $\frac{1}{2} < P < 1$, and the correction signal will at least have the correct polarity, although not the correct magnitude. Ideally, the correction term should be a discontinuous function of t_1 at a zero of $f(k + t_1)$. The actual correction signal will have a mean value which is continuous, but the sharpness of change in the vicinity of a zero will increase as $H(\omega)$ is made narrower.

The rms jitter of this timing recovery system is extremely difficult to evaluate because of the presence of many nonlinear operations. However, this jitter may be expected to be much greater than that of a sampled-derivative system, since each correction signal may be expected to introduce jitter of the same order of magnitude as the main error signal. Narrow filtering in the control loop is therefore essential.

The mean sampling for the example of Section IV when this improved timing recovery system is used is shown in Fig. 4 as "*IM*". It is seen that the time of maximum eye opening has been found. For this example, only one leading and one lagging correction term were sufficient to choose this mean sampling time.

VI. EXTENSIONS

Each of the timing recovery systems described here may be modified to work with m-level digital data signals instead of only binary formats. Since in multilevel systems the eye is much narrower, the effects of static timing error and jitter are much more serious.

A threshold crossing detector may be constructed which generates an error pulse whenever the signal crosses any one of the $m - 1$ thresholds. During any signal interval, any number of error pulses between zero and $m - 1$ may be present. Such a system is extremely difficult to analyze, but has been found to work well in practice, provided that some auxiliary means is used to correct the mean sampling time.

The sampled-derivative detector is very easily extended to multilevel systems if the signal derivative is multiplied by the output symbol value in forming the error signal. If the a_n's are scaled so as to form a set of unit variance, then the analysis of Section III applies directly.

The improved timing recovery system is modified for use with multilevel signaling in a manner similar to that of the sampled-derivative system. However, the signal margin against noise for the worst message sequence is now

$$D(t_0) = f(t_0) - (m - 1) \sum_{k \neq 0} | f(t_0 + kT) | . \tag{122}$$

Comparing this criterion with that of (106), it is seen that each of the correction signals must be weighted by the quantity $m - 1$ in order to find the time of maximum eye opening.

Extension of these techniques to partial response systems is also straightforward. Since the modifications depend on the particular partial response system used, a description will not be presented here.

All of the systems analyzed here used linear control loops. This permitted the calculation of jitter variance in terms of loop bandwidth. However, the implementation of these systems may frequently be simplified considerably by using nonlinear control systems. A particular method which has met with practical success uses the polarity of each error pulse to adjust the sampling phase by a small fixed increment.

APPENDIX A

Evaluation of $\overline{e_2}$*

Equation (40) is of the form

$$e_2(n) = 2d_n \left[c_n - \sum_{k=-\infty}^{n} g_2(n - k) e_2(k) \right] , \tag{123}$$

* The approach used here was suggested by J. E. Mazo.

where $d_n = 0$ or 1 with equal probability and are independent. The random variables c_n are uncorrelated with the d_n's.

From (123) and the properties of d_n,

$$e_2(n) = d_n e_2(n) \tag{124}$$

and

$$\overline{d_n e_2(k)} = \tfrac{1}{2}\overline{e_2(k)}, \quad \text{if} \quad n > k. \tag{125}$$

We first find the average of (123) over all c_n and all d_k, $k < n$. This average of a random variable v will be denoted by $\langle v \rangle$, while the average over all c_n and d_n will be shown as $\bar{v}$.

$$\langle e_2(n) \rangle = 2\overline{d_n c_n} - 2 \sum_{k=-\infty}^{n} g_2(n-k)\langle d_n e_2(k) \rangle. \tag{126}$$

Using (124)

$$\langle e_2(n) \rangle = 2\overline{d_n c_n} - 2g_2(0)\langle e_2(n) \rangle - 2 \sum_{k=-\infty}^{n-1} g_2(n-k)\langle d_n e_2(k) \rangle. \tag{127}$$

The only random variable in (127) is d_n. The overall average, $\overline{e_2(n)}$, is therefore the average of (127) over d_n. Using (125), we obtain

$$\overline{e_2(n)} = \overline{c_n} - 2g_2(0)\overline{e_2(n)} - \sum_{k=-\infty}^{n-1} g_2(n-k)\overline{e_2(k)} \tag{128}$$

$$\overline{e_2(n)} = \overline{c_n} - g_2(0)\overline{e_2(n)} - \sum_{k=-\infty}^{n} g_2(n-k)\overline{e_2(k)} \tag{129}$$

which is the result shown in (41).

APPENDIX B

Evaluation of $S_x(\omega)$

From the definition of d_n in (29), we may express $2d_n$ as

$$2d_n = 1 - a_n a_{n+1}. \tag{130}$$

Then (51) may be rewritten as

$$x(n) \approx -a_n a_{n+1}\overline{e_2} + z(n) + \frac{1}{b}(a_n - a_{n+1})n(t), \tag{131}$$

where the term $g_2(0)\overline{e_2}$ has been neglected.

We wish to evaluate the power spectral density of x. It will first be shown that the approximation of x given in (131) is zero-mean. Since the a_n's are zero-mean and independent, and the noise $n(t)$ is zero-

mean and independent of all other random processes, then the first and last terms of (131) are zero-mean. The mean of z is readily found after substituting (130) into (50).

$$z(n) = \frac{1}{b}(a_n - a_{n+1}) \sum_{k \neq 0,-1} a_{n-k} f(k + \tfrac{1}{2}) \tag{132}$$

$$\overline{z(n)} = 0. \tag{133}$$

The mean value of $x(n)$ as given in (131) is therefore zero. It may similarly be shown that the three terms of (131) are mutually uncorrelated. The autocorrelation of x is then

$$R_x(m) = \overline{a_n a_{n+1} a_{n+m} a_{n+m+1}}\, \bar{e}_2^2 + R_z(m) + \frac{1}{b^2}\overline{(a_n - a_{n+1})(a_{n+m} - a_{n+m+1})} R_n(m) \tag{134}$$

$$R_x(0) = \bar{e}_2^2 + R_z(0) + \frac{2}{b^2} R_n(0) \tag{135}$$

$$R_x(\pm 1) = R_z(\pm 1) - \frac{1}{b^2} R_n(\pm 1) \tag{136}$$

$$R_x(m) = R_z(m), \qquad m \neq 0, \quad \neq 1. \tag{137}$$

From (132),

$$R_z(m) = \frac{1}{b^2} \sum_{k \neq 0,-1} \sum_{l \neq 0,-1} \overline{(a_n - a_{n+1})(a_{n+m} - a_{n+m+1}) a_{n-k} a_{n+m-l}} \cdot f(k + \tfrac{1}{2}) f(l + \tfrac{1}{2}) \tag{138}$$

$$R_z(0) = \frac{2}{b^2} \sum_{k \neq 0,-1} f^2(k + \tfrac{1}{2}) \tag{139}$$

$$R_z(\pm 1) = -\frac{1}{b^2}\left[\sum_{k \neq 0,\pm 1} f(k + \tfrac{1}{2}) f(k - \tfrac{1}{2}) + f(-\tfrac{3}{2}) f(\tfrac{3}{2})\right] \tag{140}$$

$$R_z(\pm m) = \frac{1}{b^2}[f(m + \tfrac{1}{2}) - f(m - \tfrac{1}{2})][f(-m + \tfrac{1}{2}) - f(-m - \tfrac{1}{2})], \qquad m \neq 0, \quad \neq 1. \tag{141}$$

The power spectral density of x can now be calculated for $|\omega| < \pi$.

$$S_x(\omega) = \sum_{m=-\infty}^{\infty} R_x(m) \exp(-j\omega m) \tag{142}$$

$$S_x(\omega) = R_x(0) + 2 \sum_{m=1}^{\infty} R_x(m) \cos \omega m. \tag{143}$$

Since x will be passed through a very narrow filter, we are interested in the spectrum of x only in the region $|\omega| \ll 1$. In this region, (143) may be approximated by

$$S_x(\omega) \approx R_x(0) + 2\sum_{m=1}^{\infty} R_x(m) - \omega^2 \sum_{m=1}^{\infty} m^2 R_x(m). \tag{144}$$

This approximation is valid if the third derivative of $S_x(\omega)$ is bounded.[5] This will be true if $R_x(m)$ decreases as $O(m^{-5})$. If the Fourier transform of $f(t)$ is continuous, then $f(m)$ decreases as $O(m^{-2})$. In this case, it can be seen from (137) and (141) that $R_x(m)$ decreases as $O(m^{-6})$, so that the approximation (144) is valid for $|\omega| \ll 1$.

$$\begin{aligned} S_x(\omega) \approx \bar{e}_2^2 + \frac{2}{b^2}\Bigg\{ & R_n(0) - R_n(1) + \sum_{k\neq 0,-1} f^2(k+\tfrac{1}{2}) \\ & - \sum_{k\neq 0,\pm 1} f(k+\tfrac{1}{2})f(k-\tfrac{1}{2}) - f(-\tfrac{3}{2})f(\tfrac{3}{2}) \\ & + \sum_{m=2}^{\infty} [f(m+\tfrac{1}{2}) - f(m-\tfrac{1}{2})][f(-m+\tfrac{1}{2}) - f(-m-\tfrac{1}{2})]\Bigg\} \\ & + \omega^2\Bigg\{ \sum_{k\neq 0,\pm 1} f(k+\tfrac{1}{2})f(k-\tfrac{1}{2}) + f(-\tfrac{3}{2})f(\tfrac{3}{2}) \\ & - \sum_{m=2}^{\infty} m^2[f(m+\tfrac{1}{2}) - f(m-\tfrac{1}{2})][f(-m-\tfrac{1}{2}) - f(-m-\tfrac{1}{2})]\Bigg\}. \end{aligned} \tag{145}$$

After some manipulation, and using (26), (145) may be reduced to

$$\begin{aligned} S_x(\omega) = \bar{e}_2^2 + \frac{2}{b^2}\Bigg\{ & R_n(0) - R_n(1) + \frac{1}{2}\sum_{k=-\infty}^{\infty} [f(k+\tfrac{1}{2}) - f(k-\tfrac{1}{2})] \\ & \cdot [2f(k+\tfrac{1}{2}) + f(-k+\tfrac{1}{2}) - f(-k-\tfrac{1}{2})]\Bigg\} \\ & + \frac{\omega^2}{2b^2}\Bigg\{ -4f^2(\tfrac{1}{2}) + 2\sum_{k=-\infty}^{\infty} f(k+\tfrac{1}{2})f(k-\tfrac{1}{2}) \\ & - \sum_{k=-\infty}^{\infty} k^2[f(k+\tfrac{1}{2}) - f(k-\tfrac{1}{2})][f(-k+\tfrac{1}{2}) - f(-k-\tfrac{1}{2})]\Bigg\}. \end{aligned} \tag{146}$$

APPENDIX C

Evaluation of $S_v(\omega)$

We wish to find the power spectral density of

$$y(n) = \frac{a_n}{f'(0)}\left[\sum_{k\neq 0} a_{n-k} f'(k) + n'(t)\right]. \tag{147}$$

The autocorrelation of y is

$$R_y(m) = \frac{1}{f'^2(0)} \Big[\sum_{k\neq 0} \sum_{l\neq 0} \overline{a_n a_{n+m} a_{n-k} a_{n+m-l}} f'(k) f'(l) + \overline{a_n a_{n+m} n'(t) n'(t+m)}\Big] \quad (148)$$

$$R_y(0) = \frac{1}{f'^2(0)} \Big[\sum_{k\neq 0} f'^2(k) - R_n''(0)\Big]. \quad (149)$$

For $m \neq 0$,

$$R_y(m) = \frac{1}{f'^2(0)} f'(m) f'(-m). \quad (150)$$

Under the same conditions stated in Appendix B, the power spectral density of y may be approximated in the region $|\omega| \ll 1$ by

$$S_y(\omega) \approx R_y(0) + 2\sum_{m=1}^{\infty} R_y(m) - \omega^2 \sum_{m=1}^{\infty} m^2 R_y(m). \quad (151)$$

Using (75) we obtain

$$S_y(\omega) \approx \frac{1}{f'^2(0)} \Big\{-R_n''(0) + \sum_{k=-\infty}^{\infty} f'(k)[f'(k) + f'(-k)] - \tfrac{1}{2}\omega^2 \sum_{k=-\infty}^{\infty} k^2 f'(k) f'(-k)\Big\}. \quad (152)$$

APPENDIX D

Evaluation of σ_{xk}^2

The variance of x_k is the mean square value of the zero-mean random process

$$x_k(n) - \bar{x} = \sum_m h(n-m) a_{m-k}[n(t) + \sum_{m\neq p} a_{m-p} f(p)]. \quad (153)$$

Since the noise and the message are independent, the variances of the two components of (153) will add to form the total variance. The variance of x_k due to the noise is

$$\sigma_1^2 = \sum_m \sum_q h(n-m) h(n-q) \overline{a_{m-k} a_{q-k} n(t_m) n(t_q)} \quad (154)$$

$$\sigma_1^2 = \sum_m h^2(n-m) \overline{n^2(t)} \quad (155)$$

$$\sigma_1^2 = \frac{1}{2\pi} \sigma_n^2 \int_{-\infty}^{\infty} |H(\omega)|^2 \, d\omega. \quad (156)$$

The total variance of x_k is then

$$\sigma_{xk}^2 = \sigma_1^2 + \sum_m \sum_q \sum_{p\neq k} \sum_{r\neq k} h(n-m)h(n-q)\overline{a_{m-k}a_{q-k}a_{m-p}a_{q-r}} \cdot f(p)f(q). \quad (157)$$

Nonzero contributions arise from those terms where $q = m$ and $r = p$ and from those terms where $r = 2k - p$ and $q = m + k - p$.

$$\sigma_{xk}^2 = \sigma_1^2 + \sum_m \sum_{p\neq k} [h^2(n-m)f^2(p) + h(n-m)h(n-m-k+p)f(p)f(2k-p)]. \quad (158)$$

Any one of the terms of (158) is small compared to the sum. We may therefore approximate (158) by including the missing $p = k$ terms.

$$\sigma_{xk}^2 \approx \sigma_1^2 + \frac{1}{4\pi^2}\int_{-\infty}^{\infty} |H(\omega)|^2 d\omega \int_{-\infty}^{\infty} |F(\omega)|^2 d\omega + \sigma_2^2, \quad (159)$$

where

$$\sigma_2^2 = \sum_m \sum_p h(m)h(m-k+p)f(p)f(2k-p) \quad (160)$$

and it is assumed that $H(\omega)$ is bandlimited to $|\omega| < \pi$.

$$\sigma_2^2 = \frac{1}{16\pi^4}\iiiint_{-\infty}^{\infty} H(\omega)H(u)F(v)F(y) \sum_m \exp[jm(\omega+u)] \cdot \sum_p \exp[jp(u+v-y)] \exp[jk(2y-u)]\, d\omega\, du\, dv\, dy \quad (161)$$

$$\sigma_2^2 = \frac{1}{4\pi^2}\iiiint_{-\infty}^{\infty} H(\omega)H(u)F(v)F(y)\,\delta(\omega-u)\,\delta(u-v-y) \cdot \exp[jk(2y-u)]\, d\omega\, du\, dv\, dy \quad (162)$$

$$\sigma_2^2 = \frac{1}{4\pi^2}\int_{-\infty}^{\infty} H(\omega)H(-\omega)\, d\omega \int_{-\infty}^{\infty} F(y-\omega)F(y) \exp[jk(2y-\omega)]\, dy \quad (163)$$

$$\sigma_2^2 = \frac{1}{4\pi^2}\int_{-\infty}^{\infty} |H(\omega)|^2 d\omega \int_{-\infty}^{\infty} F^*(\omega-y) \exp[-jk(\omega-y)] \cdot F(y) \exp(jky)\, dy. \quad (164)$$

The second integral in (164) may be recognized as 2π times the Fourier transform of

$$p_k(t) = f(k-t)f(k+t) \quad (165)$$

so that

$$\sigma_2^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} | H(\omega) |^2 P_k(\omega) \, d\omega. \tag{166}$$

Substituting (166) and (156) into (159)

$$\sigma_{x_k}^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} | H(\omega) |^2 \left[\sigma_n^2 + \frac{1}{2\pi} \int_{-\infty}^{\infty} | F(u) |^2 \, du + P_k(\omega) \right] d\omega. \tag{167}$$

REFERENCES

1. Bennett, W. R., Statistics of Regenerative Digital Transmission, B.S.T.J., *37*, November, 1958, pp. 1501–1542.
2. Rowe, H. E., Timing in a Long Chain of Regenerative Binary Repeaters, B.S.T.J., *37*, November, 1958, pp. 1543–1598.
3. Aaron, M. R. and Gray, J. R., Probability Distribution for the Phase Jitter in Self-Timed Reconstructive Repeaters for PCM, B.S.T.J., *41*, March, 1962, pp. 503–558.
4. Byrne, C. J., Karafin, B. J., and Robinson, D. B., Jr., Systematic Jitter in a Chain of Digital Regenerators, B.S.T.J., *42*, November, 1963, pp. 2679–2714.
5. Woods, F. S., *Advanced Calculus*, Ginn and Company, Boston, 1934, pp. 10–13.

Modulation of Laser Beams by Atmospheric Turbulence – Depth of Modulation*

By M. SUBRAMANIAN and J. A. COLLINSON

(Manuscript received September 26, 1966)

We have studied the fluctuations produced in a laser beam by atmospheric turbulence over transmission paths up to 2400 feet long as a function of size of receiving aperture, range, and atmospheric conditions. The depth of modulation decreases rapidly with increasing size of receiving aperture for apertures smaller than the direct beam. It does not go to zero, however, but rather levels off at an approximately constant, finite value for apertures larger than the direct beam.

When all of the direct beam is collected, the depth of modulation varies approximately with the $\frac{3}{2}$ power of range from about 100 to 2400 feet, the largest range used. At ranges less than about 100 feet, however, the dependence is consistently much less than $\frac{3}{2}$. These results are independent of weather conditions, of time of day, of local conditions along the path, of whether the transmitter is inside or outside a building, of a twofold change in diameter of the launched beam, of whether the range is a straight pass or is multiply-folded, and of mirror separation in the multiply-folded arrangement. The $\frac{3}{2}$ dependence is consistent with near-field scattering theory and leads to an estimate for the lower bound of the effective scale size of turbulence of 5 centimeters.

The depth of modulation, however, depends sensitively on atmospheric conditions; in a time of the order of seconds the value can change as much as an order of magnitude. We have systematically measured depth of modulation of the direct beam simultaneously with wind velocity and variability, temperature gradients, and time of day. No simple dependence on these variables was found.

I. INTRODUCTION

From the point of view of communications, one of the serious effects of the atmosphere on propagation of laser beams is the fluctuations in

* Part of this paper was presented at the 1966 International Quantum Electronics Conference, April 12–15, 1966, Phoenix, Arizona.

the received signal caused by variations in the dielectric constant of the air. An important measure of the fluctuations is their power spectrum. Hogg[1] first measured these spectra and obtained an exponential distribution with a baseband width of the order of a few hundred cycles. Hogg used a multimode 6328 Å laser and a range of 2.6 km. The receiver, 5 cm in diameter, was located in the center of the received beam, which was about 25 cm in diameter. Hogg observed that an increase in the angular beamwidth of the source caused an increase in spectral width.

Hinchman and Buck[2] measured the low-frequency fluctuations in a 6328 Å laser beam at distances of 9 and 90 miles. Their receiver, 3 inches in diameter, collected a very small fraction of the total beam power. They observed a very large depth of modulation. The spectral density of the fluctuations was found to decrease with increasing frequency up to 50 Hz, the highest frequency measured.

Subramanian and Collinson[3] propagated a single-mode, diffraction-limited 6328 Å beam and examined the dependence of the spectrum on a variety of parameters. The transmitted beam diameter was changed from 1 to 38 mm, beam divergence was adjusted by focusing the telescope employed, ranges of 120 and 360 meters were used, and the receiver aperture was varied from much smaller to much larger than the received beam size. The spectrum, which had an exponential distribution (in agreement with Hogg), was independent of these variables within experimental error. The width of spectrum, however, was sensitive to atmospheric conditions. In general, the spectrum became wider as refractive gradients along the path became larger. For example, temperature gradients (caused by the sun) and pressure gradients (caused by turbulent wind) systematically gave broader spectra. The total width of the spectrum (above detector noise) varied over a total range of 60 to 1000 Hz, with typical value of a few hundred hertz. Thus, the spectral width was about the same as Hogg's, although the distance was about an order of magnitude less.

Buck[4] took additional data at more distances on the same 90-mile range, the smallest distance being 550 meters. Although the analytic shape of the power spectra was not given, it is nevertheless significant that the spectra shown all reached cutoff at about 200 Hz, and he stated that there was no systematic variation in the spectra when the detector aperture or path length was changed. Buck commented that, with a very large aperture, a noiseless dc signal is obtained, but this was not found by Subramanian and Collinson.[3] Thus, the three sets of observers at three locations have found a characteristic width of

about a few hundred hertz, rather independent of the experimental arrangement, and, in particular, apparently insensitive to changes in distance from 120 meters to 145 km.

Another important measure of the fluctuations in the received signal is depth of modulation, or the ratio of the rms power in the fluctuations to the average beam power. While spectral width may be independent of the experimental parameters, one expects the depth of modulation to show a strong dependence, especially on distance. The present work was undertaken to establish the nature of the dependence of modulation depth on such variables as distance, receiver aperture, and atmospheric conditions.

II. EXPERIMENTAL ARRANGEMENT

Since beam diameter will vary with distance (due to diffraction as well as atmospheric refraction) and since modulation depth presumably will depend on receiver aperture and on distance, the experiment must be arranged to allow adequate separation of these variables. For changes in distance to be meaningful, the receiver must bear some appropriate, uniform relation to the beam size regardless of distance. The simplest approach is to use apertures always larger than the direct beam, since it is known[3] that the fluctuations do not then disappear.

In order always to collect substantially all of the beam, the distance used should not be too large. Otherwise, the beam will be large and a receiver of an inconvenient size will be needed. With horizontal paths near the ground, it is commonly observed that atmospheric refraction produces angular spreading of the order of 10^{-4} radian, so paths miles long would imply beams feet in diameter. Moreover, it is desirable to be able to change the distance essentially continuously, and ranges where this can be done beyond a few hundred feet are not easily obtained.

Such short distances imply a very low level of atmospheric modulation with some values of modulation depth lower than 0.1 percent. This means, in turn, that the amplitude of the noise of the laser must not exceed about 0.01 percent. (In all cases, percent modulation is defined as 100 times the ratio of the rms power of the fluctuations to the average power). The laser used was designed[5] for high intrinsic frequency stability, and, when properly operated, it has the necessary amplitude stability as well. The RF power supply must be well regulated, and dc (rather than 60 Hz) power must be used on the filaments

of the power supplies. At the frequencies of interest, acoustically-coupled noise is substantial in the average laboratory and must be reduced.

An easy and highly effective method of isolation is to seal the entire laser (at one atmosphere) in a gas-tight container. This was done with a glass bottled fitted with an optical-quality window at one end and a polished flange at the other end. The flange made an O-ring seal to a Lucite plate. (It is clear that pressure fluctuations cause considerable amplitude noise, since enclosures which do not form vacuum quality seals are not as effective.) It is helpful to place a felt pad between the laser and the bottle and Isomode (corrugated rubber) pads between the floor and the legs of the table on which the laser and bottle rest. With this arrangement, the amplitude modulation of the laser could be maintained at or below 0.005 percent.

The laser used[5] was also single-mode and RF-excited in order to further ensure that measurements did not include any spurious noise. Amplitude fluctuations can appear in the output of multimode lasers as a result of mode competition and self-beating effects. Hodara has calculated[6] and measured[7] the excess noise caused by mode-beating in multimode lasers. He observed that many of the reported discrepancies in measurements of laser noise probably arose because multimode lasers sometimes were used.

While our method of mounting the transmitter is not massive, nevertheless no detectable variation in beam pointing occurred during an experiment, and the arrangement has the advantage that it could be readily modified. As will be seen, the variety of necessary experiments required flexibility in both the transmitter and the receiver. The transmitted beam emerged from the room through a selected Lucite window. Ordinary plate glass caused noticeable refraction of the beam, but some parts of new panels of $\frac{1}{4}$-inch thick Lucite produced no measurable distortion of the beam. Many "A-B" experiments were made, and no difference could be found between modulation results with a Lucite pane and the results with an open sash.

The range was located on the flat roof of a two story building at the Whippany location of Bell Telephone Laboratories. The roof surface was asphalt and gravel, but usually most of the path was over a wooden catwalk whose surface was 3 inches above the roof. The beam path was 3 feet above the roof and ran 30° east of north to a total available range of 330 feet.

The receiver took a variety of forms, and it will be best to give each experimental arrangement with the corresponding results. However,

many of the details of a representative setup can be seen in Fig. 1. The equipment cabinet on casters gave the mobility which was required for rapid change of propagation distance. Atop the cabinet is a section of triangular tower used when a long focal length, large aperture lens provided the receiving aperture. In the arrangement shown, a diffraction limited, 6-inch diameter, 8-foot F.L. lens was mounted in the right end of the tower section. The detector, an RCA 7265 photomultiplier, was equipped with a 3 Å wide-interference filter. The detector housing appears in the left end of the tower section.

Just to the left of the tower section is a bank of 6-inch diameter, $\frac{1}{4}$-wave flat mirrors supported by a bench which crosses the catwalk. Such mirrors were used to fold the beam and provide transmission paths up to 2400 feet long. The bench rested directly on the roof, and this provided adequate stability of the multiply-folded optical path.

On the bench in the foreground can be seen an RCA vacuum tube voltmeter, used to measure the dc voltage across the photomultiplier load, and a Hewlett-Packard 403A ac voltmeter, used to measure the rms ac voltage. The bandwidth of the 403A is 1 Hz to 1 MHz. To the best of our knowledge, this is the only ac meter that has such a low frequency response. Such response is important since the fluctuations are exponentially weighted toward the low end of the few-hundred hertz band. Since instantaneous voltage output from the photomulti-

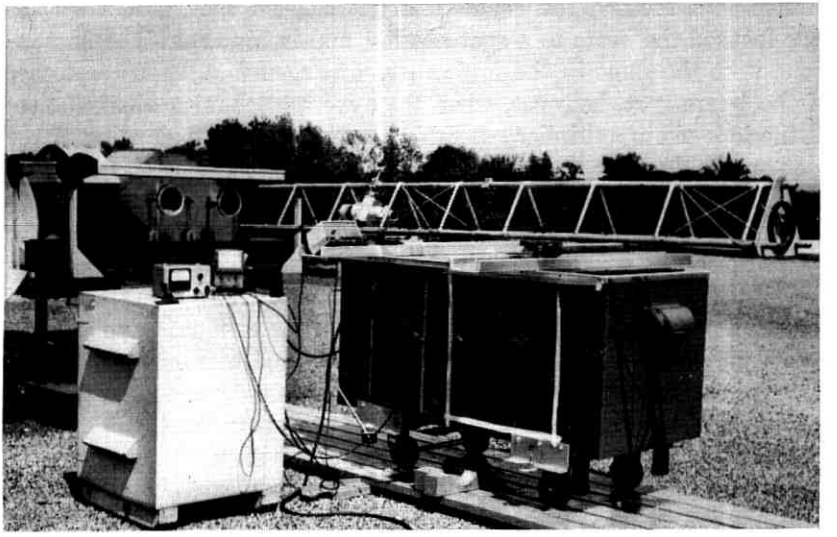


Fig. 1 — Receiving station.

plier is proportional to the instantaneous incident optical power, it follows that percent modulation is given by 100 times the ratio of the ac voltage to the dc voltage.

Since the fluctuations do not disappear[3] when all of the direct beam is collected, it is necessary to show that possible sources of spurious noise are negligible. The laser, for example, was checked periodically to insure that transmitter fluctuations were no larger than about 0.005 percent. Another possibility was that the fluctuations resulted from the product of the time-varying intensity profile of the beam and the uneven sensitivity profile of the photocathode. Also, the noise might have resulted from the dancing of a small beam (order of a millimeter diameter) over the sensitivity profile of the photomultiplier. Two experiments were made to assess these possibilities.

First, a photovoltaic cell (Hoffman 110C) which was one cm square was interchanged in "A-B" fashion with the photomultiplier. When the cell was used, the one-cm beam was focused to about one millimeter in diameter, and all of it therefore was collected by the cell. The spatial variation of sensitivity of the cell was orders of magnitude smaller than that of the photomultiplier, and the measured fluctuations were the same as with the photomultiplier, within experimental error.

In the second check, a comparison was made between the fluctuations obtained when a 1-cm diameter beam fell directly on the photomultiplier, when a 4-inch diameter, 60-inch focal length diffraction limited lens focused the beam to about a $\frac{1}{2}$-cm spot, and when the lens focused the beam to a spot about $\frac{1}{2}$ mm in diameter. The fluctuations were the same in the former two arrangements. In the last case of the $\frac{1}{2}$-mm spot, however, when there were mechanical disturbances of the lens-photomultiplier assembly large enough that dancing of the spot on the detector was visible, the level of fluctuations was measurably higher. Since the photocathode had a sensitivity structure with a scale size of the order of a few millimeters, it seems clear that the excess noise was caused by the random scanning of the small spot over the spatially varying sensitivity of the detector. As a result, in all of the work which employed a collecting lens, the spot was deliberately defocussed to about $\frac{1}{2}$ to 1 cm in diameter, and mechanical disturbances of the receiver were avoided.

The first and dominating result of any measurement of depth of modulation is that the value changes steadily with time. Another way of stating this is that the fluctuation spectrum extends well below 1 Hz, the low-frequency limit of the ac meter, and the ac value changes steadily. The consequence is that this temporal change in modulation

is unavoidably combined with the dependence of modulation on the measured variables (such as distance) when the data are taken at different times. Thus, when distance is changed by moving the receiver, the value obtained changes because of variations both in distance and in time.

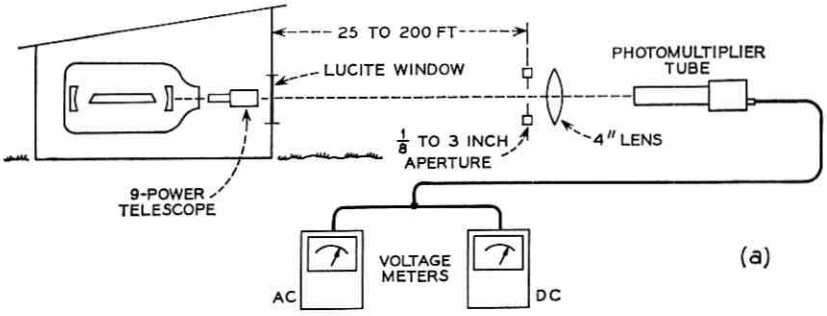
The seriousness of this depended on the degree of temporal instability of the modulation. Typically, the measured value of percent modulation varied by a factor of about two or perhaps three in a period of a few minutes. This was tolerable, but it meant that for a determination of distance or aperture dependence to be meaningful it was necessary to average the results for a large number of runs. For this reason it was important to arrange the experiment so that successive readings could be taken quickly. In general, it was possible to move the receiver and make a reading in about two or three minutes. Thus, a run involving five points took about 10 minutes. On many occasions, however, the temporal instability of modulation was severe, and the value might change by an order of magnitude within a few minutes. This is comparable with the total change produced by the variations in distance and aperture, and useful data could not then be obtained. Most of our results were taken at night, a few hours after the sun had gone down. During this period, changes in weather conditions were relatively small. Besides, the alignment of the receiver could be made very rapidly at night, hence a number of runs could be taken in quick succession.

An alternative which would circumvent this problem would be to divide the beam with beam splitters into receivers at each distance or aperture value. Modulation then would be measured simultaneously at all the receivers. However, this would have required more extensive facilities than were readily available.

III. RESULTS—APERTURE DEPENDENCE

It should once again be emphasized that while the fluctuation spectrum may be insensitive to size of receiving aperture, one expects the depth of modulation to depend rather critically on it. In particular, if the fluctuations are produced entirely by variations in power collected by a finite aperture, one might expect the depth of modulation to decrease as larger apertures are employed. With a large enough receiver, the fluctuations should then go to zero.

Depth of modulation was measured with a beam which appeared to the dark-adapted eye to be about $\frac{1}{2}$ to $\frac{3}{4}$ inch in diameter and with



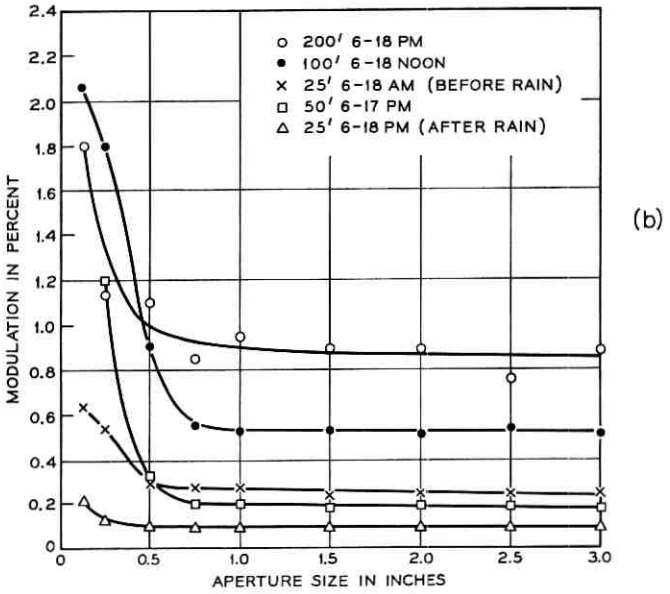


Fig. 2 — Aperture dependence: (a) schematic, (b) data.

receiving apertures from $\frac{1}{8}$ to 3 inches in diameter. Fig. 2(a) shows the schematic of the arrangement. Aperture dependence was thus measured at various distances. At each distance, several successive data runs were taken and averaged. The results are shown in Fig. 2(b).

There are two curves for the 25-foot distance, one with considerably higher percentage modulation than the other. The former was taken before a rain storm and the latter immediately after the rain. Although the rain appeared to have a decided effect in this case, firm conclusions should not be drawn since this was a single observation.

(Nature did not present us with more than one such opportunity.) Dependence on the atmospheric conditions will be discussed more fully below.

IV. RESULTS—RANGE DEPENDENCE

The dependence of depth of modulation on range was measured, using apertures large enough to collect all of the direct beam. The aperture dependence [See Fig. 2(b)] with small apertures is sufficiently strong that a meaningful range dependence would be difficult, or impossible, to obtain using apertures smaller than the beam. Generally, the aperture used was at least twice the size of the direct beam as it appeared at night (i.e., to a dark-adapted eye). Thus, all data for range dependence were obtained in the region in which the value is sensibly independent of aperture size.

The schematic of the first experiment is the same as given in Fig. 2(a) except that a 20-power telescope was used instead of the 9-power one. The averages of all the data are plotted in Fig. 3 on log-log coordinates. (The bars are averages of the mean deviations in the data for each night. The deviations are indicative of the unavoidable uncertainty caused by changing atmospheric conditions).

The results from 100 to 300 feet suggest that depth of modulation

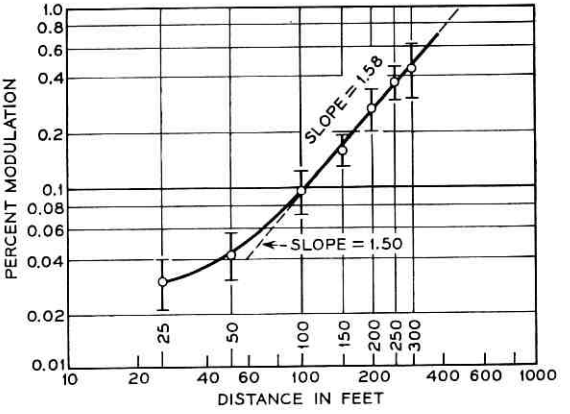


Fig. 3—Range dependence—single pass—(25–300 ft);

—night—4 runs,
5/20/65 night—8 runs,
5/26/65 night—6 runs,
5/28/65 night—6 runs,
6/24/65 night—2 runs.

goes approximately with the 3/2 power of distance. The data at 25 and 50 feet seem to indicate an end effect not intrinsic in the atmosphere. It suggests that the source was noisy, producing spuriously large values at small ranges. To examine this possibility, the data were replotted on linear-linear coordinates, and the curve was extrapolated to the y-intercept. This yielded an apparent zero-distance modulation of 0.024 percent. This is about five times the laser noise of 0.005 percent measured inside a quiet room, so that the laser cannot be the dominating cause of the change in curve shape at short distances.

Another possible explanation is that local conditions affected the atmosphere differently along the beam. Such variations could have been caused by a row of four large exhaust blowers along a line about 30 feet east of the range. On the night of June 24, 1965, we arranged to turn off all the blowers. Two runs were taken, and the results showed the same distance dependence as with the blowers on, so closely, in fact, that the data were simply included in the final curve of Fig. 3.

Another possibility is that local conditions affected the atmosphere differently near the transmitter room than far away from it. For example, the room itself may well have changed the turbulence of the wind. We therefore set up the transmitter outside the room and 30 feet away from the room (which is 8 by 14 feet). If local condi-

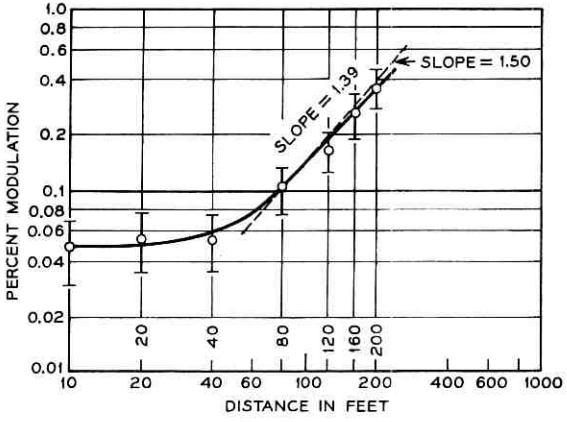


Fig. 4—Range dependence: single pass (10–200 ft) with transmitter located outside the building;
8/3/65—day—6 runs,
8/3/65—night—6 runs.

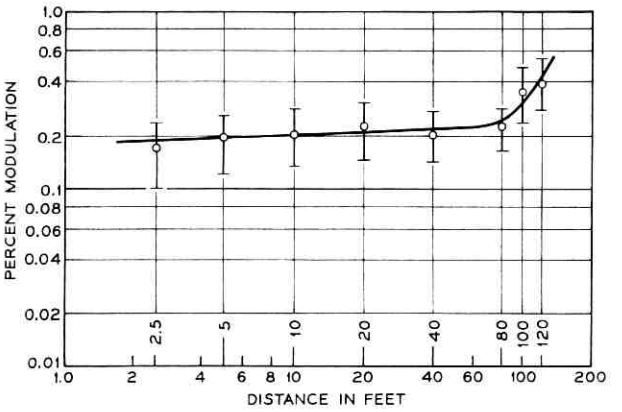


Fig. 5 — Range dependence: single pass (0–120 ft) with transmitter located outside the building and telescope changed to 9 power;

7/20/65—night—4 runs,
7/21/65 day—4 runs,
7/21/65 night—4 runs,
7/22/65 day—5 runs.

tions were the cause, the entire curve should displace to the left on the abscissa by 30 feet.

The schematic of this arrangement is identical to that shown in Fig. 2(a) except that the laser is located outside the building with a 20-power telescope. The results are given in Fig. 4. The curve did not shift along the abscissa and shows generally the same behavior at the same distance from the transmitter. Whatever the cause of the near-distance behavior of the modulation, it seems to have moved 30 feet with the transmitter.

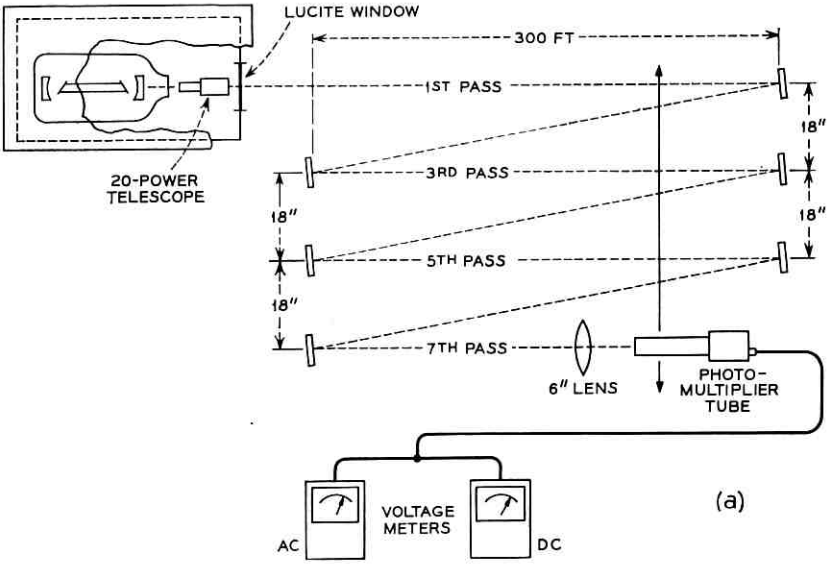
Because of the surprising behavior of the depth of modulation at short ranges, more measurements were made, still with the transmitter outside, but now adding some very short distances. The 20-power telescope was replaced with one of 9 power, giving a reduction in diameter of transmitted beam from about 2 to about 1 cm. This was done to examine what effect beam divergence might have on the distance at which the knee of the curve appears. Data were taken during both day and night. The daytime data were found to be not significantly different from the night data. All values were averaged and are plotted in Fig. 5. It appears that the relatively large values of modulation depth which yield a small slope at short distances persist at distances as small as $2\frac{1}{2}$ feet from the transmitter. (The laser noise

level of 0.005 percent is measured at distances of this order, but in a laboratory where the air turbulence is low.)

There appears in Fig. 5 to have been a significant change from the distance dependence displayed in Figs. 3 and 4. The knee of the curve seems to have moved to larger distances such that the 80-foot value now is aligned with the dependence at short distances. Since the change in beam size was an obvious potential explanation, and since the principal experimental difficulty still was the large temporal variation in modulation, the following experiment was conducted. The laser was mounted outside and 30 feet away from the transmitter room, and the beam was split into two beams of approximately equal intensity. One beam was transmitted with the 20-power telescope, the other with the 9-power telescope. The beams were parallel and 20 inches apart. At each distance, the modulation was measured on both beams before changing distance. Three runs were made on one night. This did not give enough data to define smooth curve shapes, but it was enough to show that the curve shapes were nearly identical for the two beams. The change in Fig. 5 from Figs. 3 and 4, therefore, cannot be attributed to the change in telescopes.

Having explored the behavior of modulation at short distances, we now turned to distances larger than 300 feet and in particular to the question whether the 3/2 power dependence of modulation depth on distance would continue at large distances. Fig. 3 suggests that modulation depth at 300 feet may be a little lower than a 3/2 power dependence would imply. Results for three of the five nights summarized in Fig. 3 showed a rather pronounced reduction in the value expected at 300 feet by extrapolation from shorter distances.

In order to obtain much longer paths in the available space, we now folded the beam back and forth over the range, as shown in Fig. 6(a). The laser again was mounted in the transmitter building, and the 20-power telescope was used in order to reduce beam spreading by diffraction over these longer distances. At the maximum distance of 2100 feet, the spot generally was about two to three inches in diameter. The mirrors used were always substantially larger than the beam. In Fig. 6(a), the first two mirrors were four inches square, and the last four were six inches in diameter. All the mirrors were flat to a quarter-wave and front-aluminized. Fig. 1 is a photograph of the receiving end of this arrangement. Modulation now was measured at distances of 300, 900, 1500, and 2100 feet by moving the receiver laterally in 18-inch increments, placing it in the four appropriate positions to intercept the beam.



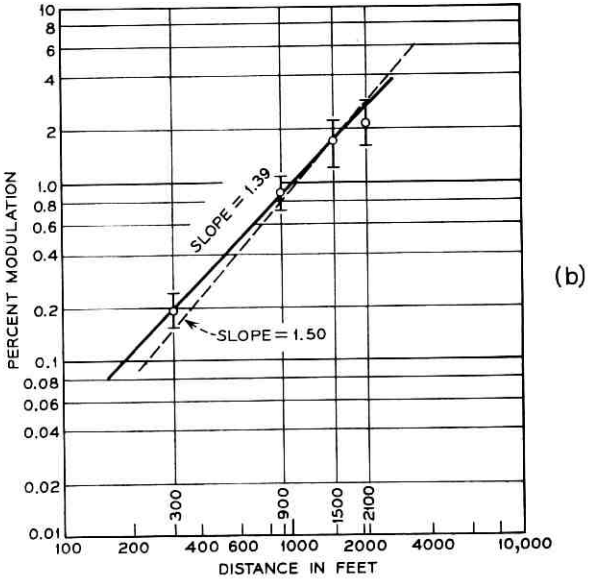


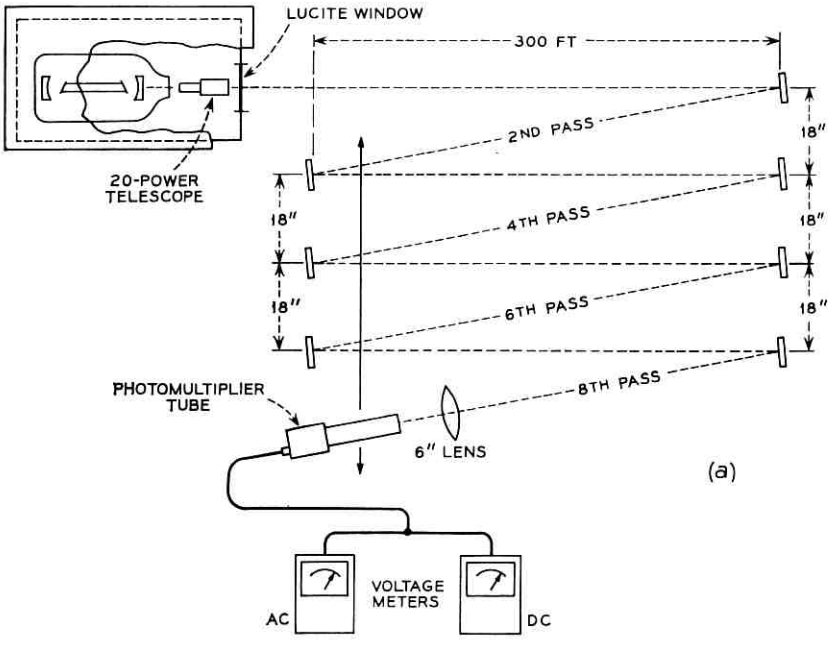
Fig. 6 — Range dependence: multiple pass (300–2100 ft);
(a) schematic,
(b) data 8/6/65 night—8 runs,
8/8/65 night—8 runs,
8/12/65 night—5 runs,
8/17/65 night—10 runs.

On some nights, beam dancing was severe enough to carry the beam off the final mirrors periodically, so that data could not then be taken. In the early part of most summer nights there was a slow drift of the beam upward as the air cooled increasingly below the temperature of the roof and the strength of the inverted "prism" increased with time (the situation is different during winter nights, see Fig. 11). A representative speed of vertical beam movement at 2100 feet was about two inches per hour. This required us not only to align the mirror system for each night's work but to realign each night every 20 to 30 minutes. Mechanical stability of the mirrors was good. When atmospheric conditions were stable enough there was no noticeable movement of the beam, so that building vibrations appeared to have no important effect.

The averages of the data were plotted in Fig. 6(b). It is seen at once that there is no significant change in slope beyond 300 feet. The points at 300, 900, and 1500 feet yield a well-defined straight line of slope 1.39, in good agreement with results below 300 feet. The value at 2100 feet seemed distinctly low, and it was not used in fitting the data. This is reminiscent of the behavior of the last point in Fig. 3 which led to the present measurements. It was not possible to determine any systematic cause of error in the measurements at the largest distance. The fact that there the beam is largest, and most difficult to collect, would explain a modulation which is too large, not too small.

Once again an increase in the distance seemed necessary, and this was accomplished by adding one more 300-foot pass to the folded system and moving the receiver to the transmitter end. This is shown in Fig. 7(a). Here the first three mirrors were four inches square, and the last four were six inches in diameter. The data, averaged in the usual way, are presented in Fig. 7(b). The slope of 1.46 again agrees with previous values, within experimental error. The value at 2400 feet shows that there is no decrease in slope beyond 2100 feet. It would seem that the 3/2 power dependence of modulation depth on distance continues to at least 2400 feet. It would be instructive to continue these measurements at larger distances, but, for this, larger optics would be required than were available.

The folded-path method of obtaining large ranges is both convenient when space is limited and desirable when readings must be made quickly at positions which would be widely spaced in a straight path. However, the question remains whether this is in every sense equivalent to the unfolded, straight path. For the distances considered here,



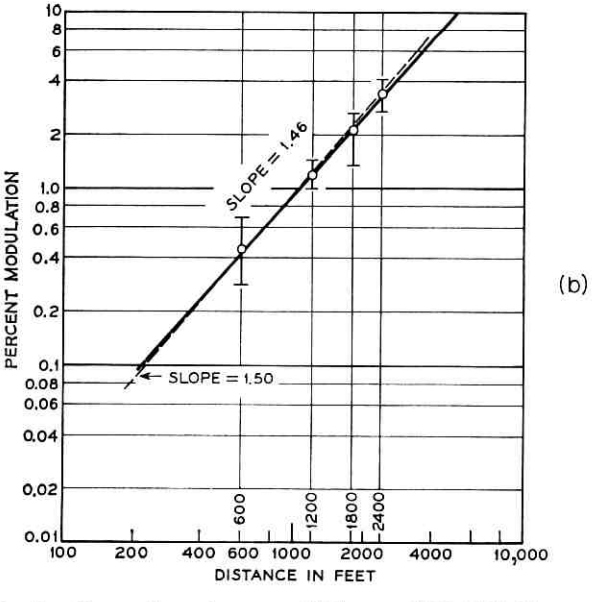


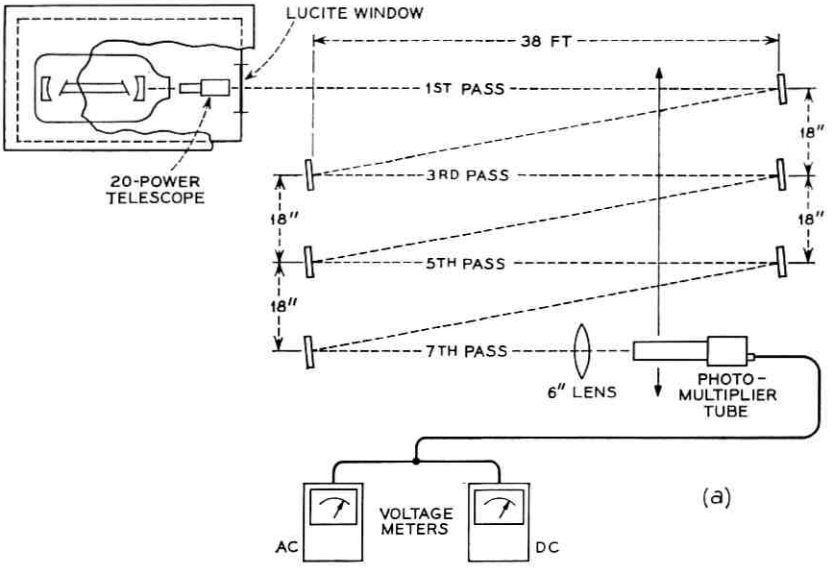
Fig. 7 — Range dependence: multiple pass (600–2400 ft);
(a) schematic,
(b) data 8/18/65 night—8 runs,
8/19/65 night—7 runs,
8/23/65 night—8 runs.

it takes only about 0.3 microsecond for the light to travel each lap and thus a total of 2.4 microseconds for 8 laps of 2400 feet. The characteristic time of turbulence is of the order of milliseconds, and consequently the turbulence can be considered frozen during the time required for the light to travel the 2400-foot folded path. Under the circumstances, not only are there possible correlations between the closely adjoining segments of the folded path but, since the beam travels oppositely along the successive segments, there may be an actual reduction in the net effect of atmospheric turbulence.

To determine this, we now set up a mutiply-folded system of mirrors, in all respects the same as the 2100-foot system in Fig. 6(a) except that now the segment length was reduced from 300 feet to 38 feet. This gave a total, folded distance of 270 feet so that comparison could be made with the results for a straight 300-foot path which were given in Fig. 3. The new arrangement is shown in Fig. 8(a). The results appear in Fig. 8(b). The functional dependence is sensibly unchanged from that in Fig. 3. This therefore not only vindicates the folded-path method used for large distances, but it further confirms the short-distance behavior of the modulation.

Although this experiment provided no evidence of correlation effects, it seemed worthwhile to make a further attempt to detect any correlation effects that may affect depth of modulation. The folded path, therefore, was rearranged as shown in Fig. 9(a). The telescope was focused on the receiver to give the smallest possible spot, which was about one cm in diameter. The first, second, fifth and sixth mirrors were two inches in diameter and quarter-wave flat. They were mounted so that no frame of the mirror or other obstruction extended in front of the six-inch diameter third and fourth mirrors. It was thus possible to position the spots on each bank of mirrors just two inches apart between centers. Hence, the average separation of successive segments of the path was only one inch. However, this made it impossible to intercept the beam with the photomultiplier, as was done before, without obstructing a previous segment of the beam. Consequently, the beam was reflected to the photomultiplier with an elliptical mirror (of the telescope-diagonal type). This mirror was then displaced laterally to intercept the beam. In this experiment, the positioning of the mirror was simple enough that a run could be completed in about 2 minutes, so that many runs were quickly made and exceptionally good averaging should result.

The average values are plotted in Fig. 9(b). Once again, the now familiar features of the curve appear. It is interesting to note that, in



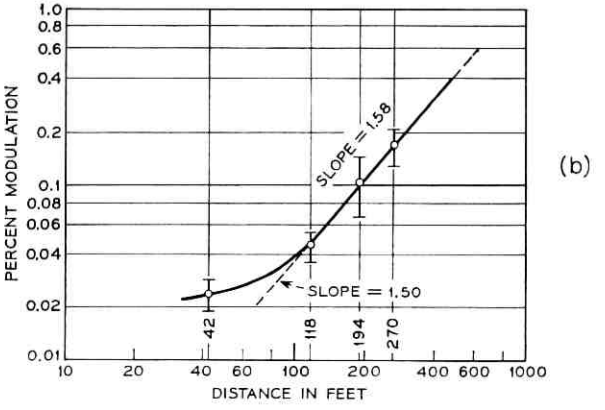
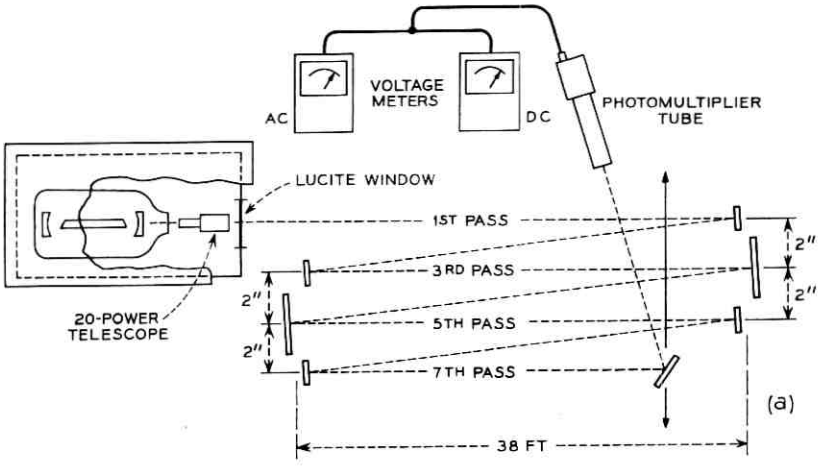


Fig. 8 — Range dependence: multiple pass (42–270 ft);
(a) schematic,
(b) data 10/7/65 day—5 runs,
10/13/65 night—10 runs,
10/14/65 night—7 runs.

the previous experiments, the slope beyond 100 feet has centered around 3/2 and that, in the present experiment in which especially good averaging against changing conditions was expected, the slope was 3/2 as closely as such curve-drawing will allow. The present results again give no evidence of correlation effects.



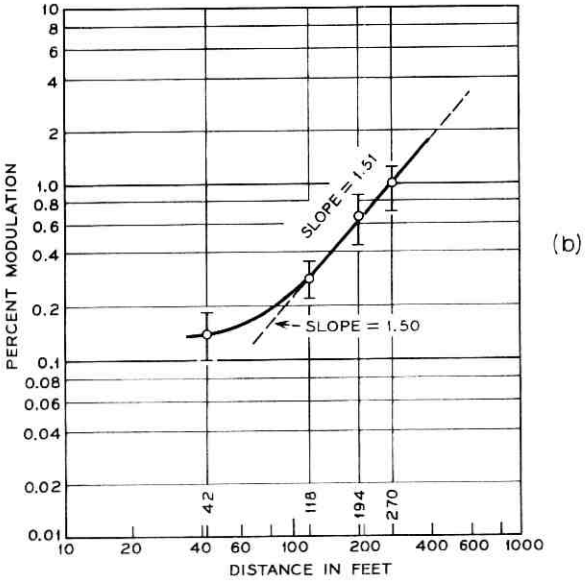


Fig. 9 — Range dependence: multiple pass (42–270 ft) with mirror separation of 2″;
(a) schematic,
(b) data 8/27/65 night—7 runs,
8/30/65 night—11 runs,
9/2/65 night—13 runs.

V. RESULTS—DEPENDENCE ON ATMOSPHERIC CONDITIONS

It was remarked before that the first and dominating result of any measurement of depth of modulation is that the value changes steadily with changing atmospheric conditions. This persists during, and interferes with, all attempts to measure the dependence of modulation on any other parameter. Hence, a great deal of qualitative observation of the effects of conditions was inescapably made during the measurements of aperture and range dependence. And this extensive experience taught simply that the depth of modulation was unpredictable. There never appeared any simple correlation between modulation and qualitative observables such as wind speed or direction, sun conditions, or time of day. The most promising lead which developed was the observation already cited in the aperature work that the depth of modulation was markedly smaller over a 25-foot range after a short heavy rain than before [Fig. 2(b)]. The rain, of course, would have both cooled the roof (which had been heated by the sun) and reduced dust in the air. Since no connection previously was apparent between modulation and temperature conditions, it then was thought that the modulation might be affected strongly by particulate scattering.

Indeed, at night the beam was always well decorated by forward scattering from what appeared to be dust and haze particles. Normally the beam could be seen with the eye placed within about 10^{-2} radian of the forward direction. This was the case with clear or hazy conditions. Only with fog could the beam be seen substantially more than 10^{-2} radian away from the forward direction. Under no conditions, including light fog, could backscattering be detected by eye. The result with the multi-folded system was that, when standing at one end of the range and looking toward the other, only those segments in which the beam was approaching could be seen. This is shown by the photograph of Fig. 10, which was taken looking toward the transmitter. Speed of the film was ASA 125. Exposure time was 15 minutes, so the beams appear well filled-in. There seem to be four separate, parallel beams from four separate sources. There is no trace of the diagonal connections of the returning beams. It seems clear that back scattering was orders of magnitude weaker than forward scattering. The right-hand beam came from the laser, and light scattered at the source caused the saturation of the film.

When measuring depth of modulation, therefore, the level of forward scattering out of the beam was readily observed. Attempts were made to correlate modulation with scattering, and no qualitatively

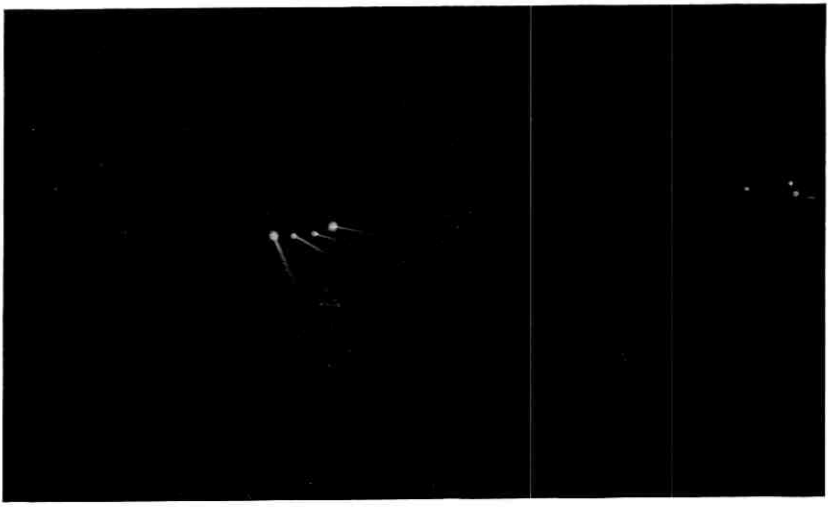


Fig. 10—Photograph of multiple-pass laser beam illustrating the lack of back scattering.

apparent relation could be found. The level of scattering did vary in a pronounced way as the density of particles varied, but modulation did not seem to change correspondingly. With short ranges, such as 25 feet, cigarette smoke was deliberately blown into and along several feet of the beam causing brief and strong decoration of the beam but having no noticeable effect on modulation.

Having thus found no dependence by casual observation, we arranged to measure modulation while also systematically noting wind speed, direction, and variability, temperature of the roof, and temperature of the air at beam level. The arrangement was that shown in Fig. 3 except the distance was 138 feet. Temperatures were measured at the middle of path. Readings were taken every half hour continuously for 24 hours so that any diurnal variation would be detected. In particular, one might expect a change in modulation depth around sunrise and sunset. (Reliable observations of "good-seeing" at sunrise and sunset date at least from 1878 when Michelson measured the velocity of light and found that he could not work at any other time due to excessive "boiling" of the image of his slit.[8])

The 24-hour period began at 5:45 p.m. on November 3, 1965. A broad variety of wind conditions was obtained both at night and during the day, ranging from dead calm to a period which was violently gusty in the late morning. The sky was clear at the beginning, becoming

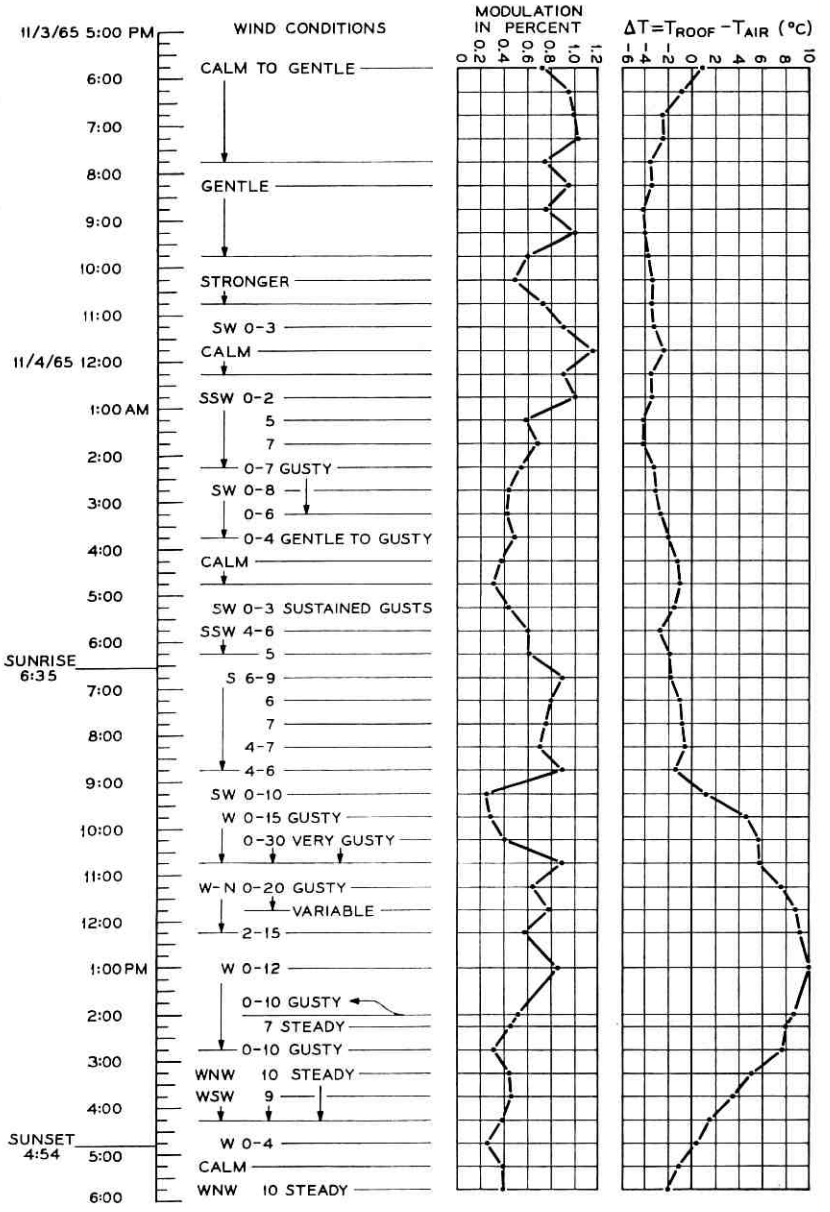


Fig. 11 — Dependence on weather conditions.

gradually overcast until no stars could be seen by 3:45 a.m. on November 4, 1965. This cleared to only a hazy sky by 8:00 a.m., and the daytime then remained clear except for occasional clouds.

The results are assembled in Fig. 11. The difference between roof and air temperature is recorded since one expects temperature gradients to be more significant than temperature level. Wind conditions are given below depth of modulation. It is apparent at once that there was no simple correlation between modulation and temperature difference or time of day. Any distinct reduction or other effect on modulation at sunrise or sunset is conspicuously absent.

The negative results of the previously qualitative observations therefore have been extended by these more quantitative results. It is still clear that modulation depth depends upon atmospheric conditions in a sensitive way, but the measurements so far have not revealed the nature of the dependence.

VI. THEORETICAL BACKGROUND: DISTANCE DEPENDENCE

The amplitude and phase fluctuations of an electromagnetic signal that has propagated through a random medium have been theoretically analyzed using different approaches. Of the more familiar ones are the following: the ray theory, the first Born approximation, and Rytov's method. The results obtained by all three methods agree, though they may hold true only for certain regions of distance.[9] Of the above three approaches, Rytov's method lends itself to a generalized treatment that holds good for both short (Fresnel zone) and long (Fraunhofer zone) distance regions of the scatterer. It should be noted that in the following discussion the Fresnel and Fraunhofer regions are with respect to the scattering medium and not with respect to the transmitter aperature.

The physical configuration for which the calculations are made is shown in Fig. 12. An infinite plane wave is incident upon a semi-infinite ($-\infty < x < +\infty; -\infty < y < +\infty; z \geqq 0$) inhomogeneous and random medium which is assumed to be quasistatic. A point detector is located at $x = L$ which not only sees the unscattered direct wave but also waves scattered from within a scattering volume that is in the form of a cone. This cone has its vertex at the receiver and has an aperture angle of the order of $1/ka$, where k is the propagation constant ($= 2\pi/\lambda$) and a is the scale size of the tubulence. The justification of using this configuration for our measurements will be given at the end of this section.

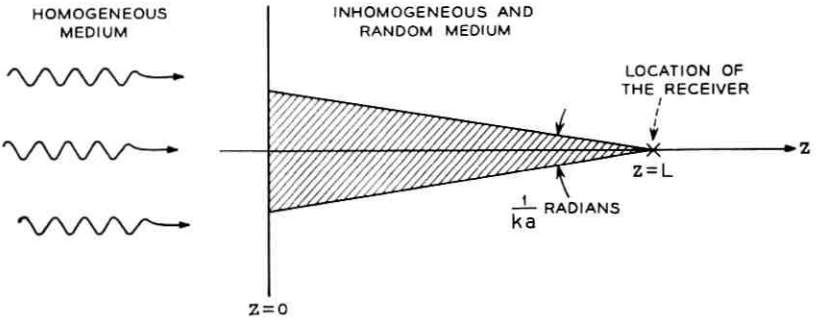


Fig. 12 — Physical configuration for calculation of amplitude fluctuations.

The functional dependence of the amplitude fluctuations will depend on whether the receiver is located in the Fresnel or Fraunhofer zone. The extent of these zones are determined by a dimensionless parameter D, called the wave parameter. It is given by

$$D = \frac{4L}{ka^2}. \tag{1}$$

$D \ll 1$ for the Fresnel region and $\gg 1$ for the Fraunhofer region. The dependence of the mean square fluctuation of the amplitude $\overline{B^2}$ on this distance L is given by[10]

$$\begin{aligned} &\text{For} \quad D \ll 1 \quad \overline{B^2} \alpha L^3 \\ &\qquad\quad D \gg 1 \quad \overline{B^2} \alpha L. \end{aligned} \tag{2}$$

In other words, the root mean square value of the fluctuations will have a functional dependence on distance given by the distance raised to the power 3/2 and 1/2 for the Fresnel and Fraunhofer zones, respectively.

The quantity measured in the experiment is the percentage modulation given by 100 times the ratio of the rms voltage to the average value. This ratio is the same as the ratio of the rms value of the fluctuating light power to the average light power expressed in percentage. For values of ac-to-dc power ratio very small, it can be shown that

$$\frac{V_{\text{rms}}}{V_{\text{av}}} = \frac{P_{\text{rms}}}{P_{\text{av}}} = 2\frac{E_{\text{rms}}}{E_{\text{av}}}, \tag{3}$$

where V's refer to the voltages across the photomultiplier load, P's to the light power and E's the electric field intensities in the light radiation field. From (2) and (3), we see that the ratio of the V_{rms} to V_{av}

should vary as $L^{3/2}$ in the Fresnel zone and as $L^{1/2}$ in the Fraunhofer zone. Even though the configuration used for the theoretical calculation does not seem to represent our experimental arrangement, its validity can be justified by the following argument. The 3/2 power law, which is the one that will be used to compare with our results, can also be derived using ray theory in which fluctuations in an infinitely thin ray of light are considered. The fluctuations in this case can be measured using a point detector. In our case, the ray is of finite diameter (that corresponding to the laser beam diameter), and consequently the detector is also of finite dimension to insure collection of the entire ray.

VII. DISCUSSION

The modulation of a single-mode, single-frequency laser beam at 6328A° by atmospheric turbulence was investigated by varying the propagation distance as the parameter. Unlike the spectral width of modulation, the depth of modulation does depend on distance and varies as 3/2 power of the distance. In making the measurement, it was ensured that all of the direct beam was collected by the receiver. The range of the propagation distance was extended from 0 to 2400 feet. The empirically obtained 3/2 power law agreees well with the theoretical result obtained using Rytov's method, provided the propagation distance is within the Fresnel zone of the scatterer. This assumption leads to an estimation of the scale size of the atmospheric turbulence. The effective scale size in (1) is estimated to be larger than 5 cm in diameter for $L \geqq 2400$ feet and $D = 0.1$.

An interesting feature of the dependence of depth of modulation on distance is its large value at distances lower than about 100 feet. This produces a functional dependence on distance which is other than 3/2 power. We have also noticed that the spectral width which is characteristicaly a few hundred hertz and whose exponential decay with frequency is otherwise independent of distance undergo changes at these short distances. The amplitudes of the low-frequency components decrease more rapidly than those of the high-frequency components as the distance is made shorter. We believe that the short distance variation of the depth of modulation and the spectral width are related and are caused possibly by the same phenomenon. This is under further study.

In comparing the absolute value of depth of modulation with those obtained by others, it has to be borne in mind that we collect all of the direct beam in contrast to previous work in which only part of the

direct beam was collected. Consequently, the magnitude of the depth of modulation measured by us is much less than that of others. For example, Edwards and Steen[12] have observed with a zirconium arc source for a 300-meter path a depth of modulation as high as 80 percent, whereas our value for the same distance is of the order of 1 percent or less. Another measurement by Portman, et al[13] with a partially collected beam at 500 meters yielded a peak to peak percentage modulation of 150 percent.

Although the main goal of this work was the functional dependence of the depth of modulation on distance, two other observations were made which are worth noting. Contrary to the behavior of the spectral width, which depends systematically on weather conditions, the depth of modulation has no clear-cut dependence on the atmospheric conditions which were measured so far. It seems modulation is a sensitive but obscure function of atmospheric conditions. The atmospheric variables which need to be measured evidently are relatively fine. Besides many qualitative observations, the quantitative results of a 24-hour run were a demonstration of this.

Also, we have not been able to observe any back scattering of the laser radiation even under severe weather conditions. Most of our experiments were conducted during night time. Even on very dark nights, the dark adapted eye (of several observers) could not detect any trace of back scattering. This is true in conditions of clear atmosphere with various amounts of particulate matter, haze, fog, and under severe rain storms. From these qualitative observations, we are led to estimate that the back scattering is orders of magnitude lower than the narrow angle forward scattering—a value considerably lower than the 2 percent obtained by Carrier and Nugent.[14] This observation is surprising also in view of the various reports of atmospheric back scattering observed with optical radar systems (e.g., Collis and Ligda[15]).

VIII. ACKNOWLEDGMENT

The authors wish to express their deep gratitude to R. R. Redington for his enthusiastic and energetic cooperation in conducting the experiments at all times of the day and night.

REFERENCES

1. Hogg, D. C., On the Spectrum of Optical Waves Propagated through the Atmosphere, B.S.T.J., *42*, November, 1963, pp. 2967–2969.

2. Hinchman, W. R. and Buck, A. L., Fluctuations in a Laser Beam over 9- and 90- Mile Paths, Proc. IEEE, *52,* March, 1964, pp. 305–306.
3. Subramanian, M. and Collinson, J. A., Modulation of Laser Beams by Atmospheric Turbulence, B.S.T.J., *44,* March, 1965, pp. 543–546.
4. Buck, A. L., Laser Propagation in the Atmosphere, presented at the Conference on Atmospheric Limitations to Optical Propagation, Boulder, Colorado, 18–19 March, 1965.
5. Collinson, J. A., A Stable, Single-Frequency RF-Excited Gas Laser at 6328Å, B.S.T.J., *44,* September, 1965, pp. 1511–1519.
6. Hodara H., Statistics of Thermal and Laser Radiation, Proc. IEEE, *53,* July, 1965, pp. 696–704.
7. Hodara H., Measurements of Laser Excess Photon Noise, 1966 International Quantum Electronics Conference, Phoenix, Arizona, April 12–15, 1966, paper 1A-3.
8. Michelson, A. A., *Velocity of Light,* p. 17 of Michelson's Notebook reproduced by Lund Press, Inc. of Minneapolis for Honeywell, Inc.
9. DeWolf, David A., Wave Propagation through Quasi-Optical Irregularities, J. Opt Soc. Am., *55,* July, 1965, pp. 812–817.
10. Chernov, Lev A., *Wave Propagation in a Random Medium,* Chapter V, McGraw-Hill Book Co. Inc., N. Y., 1960.
11. Consortini, A., et al, Influence of the Atmospheric Turbulence on the Space Coherence of a Laser Beam, Alta Frequenza, Vol. XXXII, N II, 1963, refer to the table on the microscale.
12. Edwards, Byron N. and Steen, Ronald R., Effects of Atmospheric Turbulence on the Transmission of Visible and Near Infrared Radiation, App. Optics, *4,* March, 1965, pp. 311–316.
13. Portman, D. J., Elder F. C., and Ryznar, E., Optical Scintillation in the Surface Layer of the Atmosphere, Conf. on Atmospheric Limitations to Optical Propagation, Boulder, Colorado, 18–19 March, 1965.
14. Carrier, L. W. and Nugent, L. J., Comparison of Some Recent Experimental Results of Coherent and Incoherent Light Scattering with Theory, App. Optics, *4,* November, 1955, pp. 1457–1462.
15. Collis, R. T. H. and Ligda, M. G. H., Laser Radar Echoes from the Clear Atmosphere, Nature, *203,* August 1, 1964, p. 508.

Iterative Solution of Waveguide Discontinuity Problems

By W. J. COLE, E. R. NAGELBERG and C. M. NAGEL

(Manuscript received November 25, 1966)

The application of matrix iterative analysis to the solution of waveguide discontinuity problems is discussed. It is concluded that the "Gauss-Seidel" or "point-single-step" method offers several advantages over more conventional invertive procedures, particularly in the speed of execution. Two examples are presented as illustrations: analysis of an H-plane discontinuity in a rectangular waveguide and conversion from TE_{11} to TM_{11} modes at an abrupt discontinuity in a circular waveguide. The latter results are shown to be in good agreement with measured values obtained in a previous investigation.

I. INTRODUCTION

The analysis of waveguide discontinuities, for application to the design of antennas and microwave networks, continues to offer challenging problems in electromagnetic theory and microwave engineering. Thus far, the solution of these problems has depended to a large extent on various approximate techniques, such as variational and quasi-static methods,[1] which are extremely useful but nevertheless limited in applicability.

The shortcomings of classical analysis have been surmounted to a large extent by our ability to solve electromagnetic boundary value problems by numerical methods, making extensive use of digital computers. Computational techniques are not only an abundant source of engineering data, which might otherwise require elaborate construction and experiment, but they can also provide a unique analytical laboratory in which to evaluate approximate theoretical methods under easily controlled conditions. In this paper, we shall be concerned with these numerical methods as they apply to certain waveguide discontinuity problems.

For the sake of simplicity we shall consider, as an example, the problem of two waveguides with similar cross sections connected together at the plane $z = 0$, as illustrated in Fig. 1. A wave is shown incident from the smaller waveguide impinging on the discontinuity. The result will, of course, be to excite an infinite number of normal modes in each guide, some of which carry real power away from the junction, with the remainder being evanescent and contributing to the electromagnetic field only in the vicinity of the connecting aperture. It must be recognized that these evanescent modes play an important role since they, in part, determine the amplitudes and phases of the propagating modes. It is the fact that an infinite number of waves must, in principle, be considered that makes this type of problem so difficult.

The contents of the paper may be summarized as follows: We begin by establishing an appropriate form of the uniqueness theorem for Maxwell's equations as they apply to boundary value problems of this type. In numerical analysis, the criteria for uniqueness are of more than academic interest since they provide meaningful and practical methods by which to assess the accuracy of results. Next, the normal mode representation of the fields is discussed, the object being to arrive at a matrix equation formulation of the problem in which the components of the unknown vector are the modal coefficients. It is

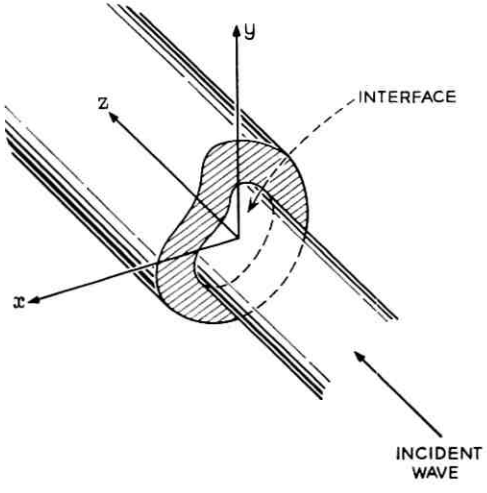


Fig. 1 — Waveguides of similar cross section connected at the plane $z = 0$ by an abrupt discontinuity. A wave is assumed incident from the smaller guide.

suggested that this matrix equation may be solved by an iterative procedure and, upon studying the convergence properties of such methods we find a critical dependence on the particular algorithm used. Two examples will be presented as illustrations, analysis of an H-plane discontinuity in a rectangular waveguide, and conversion from TE_{11} to TM_{11} modes at an abrupt discontinuity in a circular waveguide. The latter results are shown to be in good agreement with measured values obtained in a previous investigation.

Rationalized MKS units and the (suppressed) harmonic time dependence exp $(-i\omega t)$ will be used, unless otherwise specified.

II. UNIQUENESS AND ERROR CRITERIA

A representation of the discontinuity is shown in Fig. 2. It is assumed that the regions to the left (denoted by $-$) and to the right (denoted by $+$) are each filled with homogeneous material, but with possibly different constitutive parameters. Maxwell's curl equations in the respective regions are thus given by

$$\begin{aligned} \nabla \times \mathbf{E}^{\pm} &= -\mu^{\pm} \frac{\partial \mathbf{H}^{\pm}}{\partial t} \\ \nabla \times \mathbf{H}^{\pm} &= \epsilon^{\pm} \frac{\partial \mathbf{E}^{\pm}}{\partial t}. \end{aligned} \tag{1}$$

As usual for uniqueness theorems, we begin with two solutions in each region *presumed* to be correct, and denote the differences respectively by $\mathbf{E}^{\pm}$, $\mathbf{H}^{\pm}$.* Then from the Poynting theorem,[2] it follows that

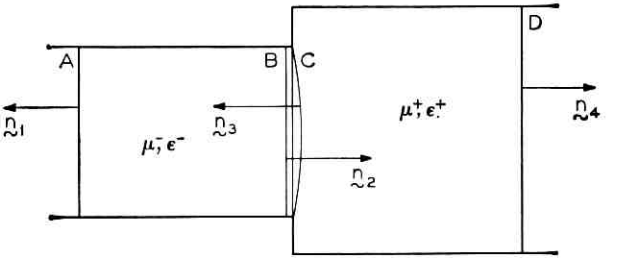


Fig. 2 — Waveguide discontinuity showing boundary surfaces A, B, C, D and respective normals $\mathbf{n}_1$, $\mathbf{n}_2$, $\mathbf{n}_3$, $\mathbf{n}_4$.

* Physically, these fields would correspond to a waveguide discontinuity problem without excitation.

in the $-$ region

$$\iint_A (\mathbf{E}_t^- \times \mathbf{H}_t^-)\cdot\mathbf{n}_1\, dS + \iint_B (\mathbf{E}_t^- \times \mathbf{H}_t^-)\cdot\mathbf{n}_2\, dS$$
$$= -\epsilon^- \frac{\partial}{\partial t}\iiint_{V-} \mathbf{E}^-\cdot\mathbf{E}^-\, dV - \mu^- \frac{\partial}{\partial t}\iiint_{V-} \mathbf{H}^-\cdot\mathbf{H}^-\, dV \qquad (2)$$

and in the $+$ region that

$$\iint_C (\mathbf{E}_t^+ \times \mathbf{H}_t^+)\cdot\mathbf{n}_3\, dS + \iint_D (\mathbf{E}_t^+ \times \mathbf{H}_t^+)\cdot\mathbf{n}_4\, dS$$
$$= -\epsilon^+ \frac{\partial}{\partial t}\iiint_{V+} \mathbf{E}^+\cdot\mathbf{E}^+\, dV - \mu^+ \frac{\partial}{\partial t}\iiint_{V+} \mathbf{H}^+\cdot\mathbf{H}^+\, dV, \qquad (3)$$

where $\partial/\partial t$ denotes differentiation with respect to time and the subscript t denotes the field transverse to the generatrix of the cylinder. The unit normal vectors $\mathbf{n}_1$, $\mathbf{n}_2$, $\mathbf{n}_3$, and $\mathbf{n}_4$ are shown in Fig. 2.

One must also take into account the fact that certain physical considerations will limit the class of admissible solutions. For example, if we let the surfaces A and D recede to infinity, then all evanescent modes will have decayed to zero and the respective surface integrals then represent the power flow away from the discontinuity. Assuming no loss, the total power must vanish. Furthermore, it can be shown from Maxwell's equations that the transverse components of electric and magnetic field at the interface must be continuous. Adding (2) and (3), we find that the following time derivative must vanish,

$$\frac{\partial}{\partial t}\left[\epsilon^- \iiint_{V-} \mathbf{E}^-\cdot\mathbf{E}^-\, dV + \mu^- \iiint_{V-} \mathbf{H}^-\cdot\mathbf{H}^-\, dV\right.$$
$$\left. + \epsilon^+ \iiint_{V+} \mathbf{E}^+\cdot\mathbf{E}^+\, dV + \mu^+ \iiint_{V+} \mathbf{H}^+\cdot\mathbf{H}^+\, dV\right] = 0, \qquad (4)$$

We may, however, regard the quantity in brackets as having had a zero value at some time, say at $t = 0$, the excitation time. The term in brackets therefore, vanishes for all time and, since each of the integrands is positive semi-definite, they must vanish separately. Thus, at each point in the $+$ and $-$ regions,

$$\begin{aligned} \mathbf{E}_1^+ - \mathbf{E}_2^+ &= \mathbf{H}_1^+ - \mathbf{H}_2^+ \equiv 0 \\ \mathbf{E}_1^- - \mathbf{E}_2^- &= \mathbf{H}_1^- - \mathbf{H}_2^- \equiv 0 \end{aligned} \qquad (5)$$

and the solution is thereby shown to be unique. We may now state the following uniqueness theorem for waveguide discontinuity problems.

Theorem: The solution to a waveguide discontinuity problem is uniquely specified if it can be shown to have the following properties:

(i) It satisfies Maxwell's equations and the appropriate boundary conditions in the regions on each side of the discontinuity.

(ii) The components of electric and magnetic fields tangent to the interface are continuous.

(iii) In the case of a lossless discontinuity, energy is conserved.

These three conditions obviously play an important theoretical role in the solution; where numerical methods are used, they also provide fundamental criteria by which the accuracy of computed results can be assessed. Accordingly, we shall define the following quantities to be used as error criteria: First, there is the parameter ε_P, which indicates how well the solution conserves energy, given by

$$\varepsilon_P = \frac{P_r + P_t}{P_{\text{inc}}} - 1, \tag{6}$$

where P_r, P_t, and P_{inc} are the reflected, transmitted, and incident powers, respectively. Second, the mean square error in the tangential electric field is defined by

$$\varepsilon_E = \frac{\iint_{\text{Aperture}} |\, (\mathbf{E}_t^+ - \mathbf{E}_t^-) \,|^2 \, dA}{\iint_{\text{Aperture}} |\, \mathbf{E}_t^{(\text{inc})} \,|^2 \, dA} \tag{7}$$

and third, for the magnetic field,

$$\varepsilon_H = \frac{\iint_{\text{Aperture}} |\, (\mathbf{H}_t^+ - \mathbf{H}_t^-) \,|^2 \, dA}{\iint_{\text{Aperture}} |\, \mathbf{H}_t^{(\text{inc})} \,|^2 \, dA}, \tag{8}$$

where $\mathbf{E}^{(\text{inc})}$ and $\mathbf{H}^{(\text{inc})}$ refer to the incident wave.

The smaller the quantities ε_P, ε_E, and ε_H, the more closely the boundary conditions are satisfied, at least in the mean square sense, and the more accurate we shall consider the solution to be.

III. MATRIX FORMULATION OF THE BOUNDARY VALUE PROBLEM

The most convenient format for numerical solution of waveguide discontinuity problems is a matrix representation, in which the modal coefficients form the unknown column vectors and the discontinuity

is characterized by a square matrix. We recognize that there is also an analogous integral equation in terms of the aperture electric or magnetic field. However, since the numerical solution of the integral equation is generally carried out by reducing it to a matrix equation, we shall proceed to the matrix formulation directly from the physical characterization of the boundary value problem. This matrix equation will then be solved by an iterative method, the theory of which is discussed in Section IV.

It is assumed that in each of the waveguides, the electromagnetic fields may be characterized by a denumerable set of known vector eigenfunctions which may be ordered according to some index. We shall be concerned only with the *transverse* fields,* denoted as follows:

${}^{+}\mathbf{E}_p'(\mathbf{r})$ $(p = 1, 2, 3, \cdots)$ denotes the transverse electric field for the pth TM mode in the $+$ waveguide, with $\mathbf{r}$ as the position vector in the transverse plane.

${}^{+}\mathbf{H}_p'(\mathbf{r})$ = transverse magnetic field for the pth TM mode in the $+$ waveguide.

${}^{+}\mathbf{E}_p''(\mathbf{r})$ = transverse electric field for the pth TE mode in the $+$ waveguide.

${}^{+}\mathbf{H}_p''(\mathbf{r})$ = transverse magnetic field for the pth TE mode in the $+$ waveguide.

By replacing the $+$ by $-$ we have the analogous notation for the other waveguide. An important point concerning sign convention is that the unknown modes in the $-$ waveguide will all be taken to propagate away from the discontinuity, i.e., in the $-z$ direction. Although the electric field does not change sign when the direction of propagation is reversed, the magnetic field does, and this fact must be carefully taken into account.

In order to define the amplitudes of the respective vector wave functions, we adopt the following normalization,[3] written in terms of integrals over the waveguide cross sections:

$$\int_{\pm_A} {}^{\pm}\mathbf{E}_p' \cdot {}^{\pm}\mathbf{E}_q'^{*}\, dA = |\, {}^{\pm}h_p' \,|^2\, \delta_{pq} \tag{9}$$

$$\int_{\pm_A} {}^{\pm}\mathbf{E}_p'' \cdot {}^{\pm}\mathbf{E}_q''^{*}\, dA = \omega^2\mu^2\, \delta_{pq} \tag{10}$$

in which h_p is the respective characteristic wavenumber, and μ is the permeability, which in our case will be the permeability of vacuum,

* It is assumed that the individual waveguides can support pure TE and TM modes, which is the case for applications of interest here.

since both waveguides will be assumed empty.† By introducing the Kronecker delta δ_{pq}, we have also expressed the fact that the transverse fields in the individual waveguides are orthogonal.

Once the normalizations for the electric wave functions are defined, those for the magnetic field are also specified since, for both the TE and TM modes, the transverse electric and magnetic fields are uniquely related. In particular, for a TM mode

$$^{\pm}\mathbf{H}_p' = \pm\frac{\omega\epsilon}{^{\pm}h_p'}\,\mathbf{e}_z \times {}^{\pm}\mathbf{E}_p' \tag{11}$$

and for a TE mode

$$^{\pm}\mathbf{H}_p'' = \pm\frac{^{\pm}h_p''}{\omega\mu}\,\mathbf{e}_z \times {}^{\pm}\mathbf{E}_p'', \tag{12}$$

where $\mathbf{e}_z$ is a unit vector in the z direction. Note again that the sign convention is such that a field in the $-$ waveguide is taken to be a reflected wave, travelling away from the discontinuity. The magnetic field normalization is thus given by

$$\int_{\pm_A} {}^{\pm}\mathbf{H}_p' \cdot {}^{\pm}\mathbf{H}_q'^{*}\, dA = \omega^2\epsilon^2\,\delta_{pq} \tag{13}$$

$$\int_{\pm_A} {}^{\pm}\mathbf{H}_p'' \cdot {}^{\pm}\mathbf{H}_q''^{*}\, dA = |\,{}^{\pm}h_p''\,|^2\,\delta_{pq}\,. \tag{14}$$

Both sets of transverse wave functions have the property of completeness, which is to say that any transverse electric (or magnetic) field can be synthesized from a set of TE and TM vector wave functions, provided that the directions of propagation of the normal modes are known. For the problems to be considered here, this latter information is available from physical considerations, since all modes propagate away from the junction with the exception of the incident wave whose amplitude is known. This amplitude will be taken to be that of a normalized mode.

We now derive the appropriate matrix representation for the discontinuity problem. Assume a dominant (TE) mode wave ($\mathbf{E}_1''$, $\mathbf{H}_1''$) is incident from the $-$ guide, setting up a transverse electric field in the aperture just to the left of the junction. This field, referred to as $^{-}\mathbf{E}_t$, may be synthesized as follows:

$$^{-}\mathbf{E}_t = {}^{-}\mathbf{E}_1'' + \sum_{p=1}^{\infty} {}^{-}A_p\,{}^{-}\mathbf{E}_p' + \sum_{q=1}^{\infty} {}^{-}B_q\,{}^{-}\mathbf{E}_q'' \tag{15a}$$

† The asterisk (*) denotes the complex conjugate.

with the modal coefficients ${}^{-}A_p$ and ${}^{-}B_q$ as yet undetermined. The corresponding transverse electric field on the + side, denoted by ${}^{+}\mathbf{E}_t$, would then be given in terms of normal modes on the + side by

$$ {}^{+}\mathbf{E}_t = \sum_{p=1}^{\infty} {}^{+}A_p {}^{+}\mathbf{E}'_p + \sum_{q=1}^{\infty} {}^{+}B_q {}^{+}\mathbf{E}''_q. \tag{15b} $$

Since the transverse electric field is continuous across the aperture, we have that

$$ {}^{+}\mathbf{E}_t = {}^{-}\mathbf{E}_t \quad \text{on} \quad C $$

$$ {}^{+}\mathbf{E}_t \equiv 0 \quad \text{on} \quad D\text{-}C. $$

As shown in Fig. 2, D-C represents the conducting wall which makes up the remainder of the junction, and on which the transverse electric field must vanish. Expanding (15a) in a Fourier series of modes in the + waveguide, we find that the modal coefficients are related by

$$ {}^{+}A_p = \frac{1}{|\,{}^{+}h'_p\,|^2} \int_C {}^{-}\mathbf{E}''_1 \cdot {}^{+}\mathbf{E}'^{*}_p \, dA + \frac{1}{|\,{}^{+}h'_p\,|^2} \sum_{q=1}^{\infty} {}^{-}A_q \int_C {}^{-}\mathbf{E}'_q \cdot {}^{+}\mathbf{E}'^{*}_p \, dA $$

$$ + \frac{1}{|\,{}^{+}h'_p\,|^2} \sum_{q=1}^{\infty} {}^{-}B_q \int_C {}^{-}\mathbf{E}''_q \cdot {}^{+}\mathbf{E}'^{*}_p \, dA \tag{16a} $$

$$ {}^{+}B_p = \frac{1}{\omega^2\mu^2} \int_C {}^{-}\mathbf{E}''_1 \cdot {}^{+}\mathbf{E}''^{*}_p \, dA + \frac{1}{\omega^2\mu^2} \sum_{q=1}^{\infty} {}^{-}A_q \int_C {}^{-}\mathbf{E}'_q \cdot {}^{+}\mathbf{E}''^{*}_p \, dA $$

$$ + \frac{1}{\omega^2\mu^2} \sum_{q=1}^{\infty} {}^{-}B_q \int_C {}^{-}\mathbf{E}''_q \cdot {}^{+}\mathbf{E}''^{*}_p \, dA \tag{16b} $$

or, more succinctly, using partitioned matrix representations,

$$ \begin{bmatrix} {}^{+}\mathfrak{A} \\ {}^{+}\mathfrak{B} \end{bmatrix} = \begin{bmatrix} {}^{+}\mathfrak{D}' & 0 \\ 0 & \frac{1}{\omega^2\mu^2}\mathfrak{g} \end{bmatrix} \begin{bmatrix} \mathfrak{F}_1 \\ \mathfrak{F}_2 \end{bmatrix} $$

$$ + \begin{bmatrix} {}^{+}\mathfrak{D}' & {}^{+}\mathfrak{D}' \\ \frac{1}{\omega^2\mu^2}\mathfrak{g} & \frac{1}{\omega^2\mu^2}\mathfrak{g} \end{bmatrix} \begin{bmatrix} \mathcal{E}^T\begin{pmatrix} -, + \\ ', ' \end{pmatrix} & \mathcal{E}^T\begin{pmatrix} -, + \\ '', '' \end{pmatrix} \\ \mathcal{E}^T\begin{pmatrix} -, + \\ ', '' \end{pmatrix} & \mathcal{E}^T\begin{pmatrix} -, + \\ '', '' \end{pmatrix} \end{bmatrix} \begin{bmatrix} {}^{-}\mathfrak{A} \\ {}^{-}\mathfrak{B} \end{bmatrix} \tag{17} $$

in which the vectors and submatrices are defined as follows:

$$ {}^{\pm}\mathfrak{A} = \begin{bmatrix} {}^{\pm}A_1 \\ {}^{\pm}A_2 \\ \vdots \\ \vdots \end{bmatrix} \quad {}^{\pm}\mathfrak{B} = \begin{bmatrix} {}^{\pm}B_1 \\ {}^{\pm}B_2 \\ \vdots \\ \vdots \end{bmatrix} \tag{18} $$

$$ \begin{aligned} (\mathfrak{F}_1)_p &= \int_C {}^{-}\mathbf{E}_1'' \cdot {}^{+}\mathbf{E}_p'^{*}\, dA \\ (\mathfrak{F}_2)_p &= \int_C {}^{-}\mathbf{E}_1'' \cdot {}^{+}\mathbf{E}_p''^{*}\, dA \end{aligned} \tag{19} $$

$$ {}^{\pm}\mathfrak{D}' = \begin{bmatrix} \dfrac{1}{|\,{}^{+}h_1'\,|^2} & 0 & 0 & \cdots \\ 0 & \dfrac{1}{|\,{}^{+}h_2'\,|^2} & 0 & \cdots \\ 0 & 0 & & \\ \vdots & \vdots & & \ddots \end{bmatrix} \tag{20} $$

and $\mathfrak{s}$ is the identity matrix. The matrix $\mathfrak{E}$, whose *transpose* appears in (17), is the matrix of coupling coefficients, defined as the scalar products of electric transverse vector wave functions for the waveguides on each side of the discontinuity. The four index notation is interpreted as:

$$ \mathcal{E}_{pq}\begin{bmatrix} -, + \\ ', '' \end{bmatrix} = \int_C {}^{-}\mathbf{E}_p' \cdot {}^{+}\mathbf{E}_q''^{*}\, dA \tag{21} $$

with analogous definitions for other combinations.

The system of equations given in (17) is clearly underdetermined since the number of unknowns is twice the number of equations. However, an additional set can be derived by employing the boundary condition that the transverse magnetic field must also be continuous across the interface. The matrix equation, analogous to (17), but corresponding to this second boundary condition, is given by

$$ \begin{bmatrix} {}^{-}\mathfrak{A} \\ {}^{-}\mathfrak{B} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathfrak{S} \end{bmatrix} + \begin{bmatrix} \dfrac{1}{\omega^2\varepsilon^2}\mathfrak{s} & \dfrac{1}{\omega^2\varepsilon^2}\mathfrak{s} \\ {}^{-}\mathfrak{D}'' & {}^{-}\mathfrak{D}'' \end{bmatrix} \begin{bmatrix} \mathfrak{K}^T\begin{bmatrix} +, - \\ ', ' \end{bmatrix} & \mathfrak{K}^T\begin{bmatrix} +, - \\ ', ' \end{bmatrix} \\ \mathfrak{K}^T\begin{bmatrix} +, - \\ ', '' \end{bmatrix} & \mathfrak{K}^T\begin{bmatrix} +, - \\ '', '' \end{bmatrix} \end{bmatrix} \begin{bmatrix} \mathfrak{A}^{+} \\ \mathfrak{B}^{+} \end{bmatrix} \tag{22} $$

in which

$$\mathcal{G} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ \vdots \end{bmatrix} \tag{23}$$

$$\bar{\mathfrak{D}}'' = \begin{bmatrix} \frac{1}{|\bar{h}_1''|^2} & 0 & 0 & \cdots \\ 0 & \frac{1}{|\bar{h}_2''|^2} & 0 & \cdots \\ 0 & 0 & \cdot & \\ \vdots & \vdots & & \ddots \\ \vdots & \vdots & & \end{bmatrix} \tag{24}$$

and the matrix $\mathfrak{C}$, whose *transpose* appears in (22), is the matrix of scalar products of magnetic transverse vector wave functions. The four-index notation is interpreted in the same way as in (21).

It should be noted once again that all the matrices which appear in (17) and (22) are infinite matrices, corresponding to the fact that in general an infinite number of modes are excited in the neighborhood of the discontinuity. In practice, of course, there must be a truncation and the problem then becomes one of solving a set of matrix equations whose order, N, depends on the accuracy required. Unfortunately there is, as yet, no way in which the number of modes required to produce a given accuracy can be predicted. We can only emphasize the need for meaningful error criteria which will act as a guide in choosing a number of modes which will be large enough to give sufficiently accurate results but at the same time not be so large as to require excess computation. It is expected that the criteria given in Section II will prove very useful in this respect.

IV. MATRIX ITERATIVE METHODS

It was shown in the previous section that the waveguide discontinuity problem of interest here can be formulated in terms of a system of linear algebraic equations. This is a recurrent theme in mathematical physics, so that an extensive theory concerned with the efficient solution of matrix equations has evolved. In this section, we shall be concerned

with some of the elements of this theory, placing particular emphasis on the solution of matrix equations by iteration.

The system of linear algebraic equations which results from satisfying the aperture boundary conditions on the transverse electric and magnetic fields can be written in the matrix forms

$$^{+}\alpha = \mathfrak{U} + \mathfrak{R} \cdot {}^{-}\alpha \tag{25}$$

$$^{-}\alpha = \mathfrak{V} + \mathfrak{S} \cdot {}^{+}\alpha, \tag{26}$$

where

$$^{\pm}\alpha = \begin{bmatrix} ^{\pm}\mathfrak{A} \\ ^{\pm}\mathfrak{B} \end{bmatrix}, \tag{27}$$

the vectors $\mathfrak{U}$ and $\mathfrak{V}$ and the matrices $\mathfrak{R}$ and $\mathfrak{S}$ being correspondingly identified from (17) and (22). Equations (25) and (26) are easily uncoupled to give

$$[\mathfrak{I} - \mathfrak{R}\mathfrak{S}] \cdot {}^{+}\alpha = \mathfrak{U} + \mathfrak{R}\mathfrak{V} \tag{28}$$

$$[\mathfrak{I} - \mathfrak{S}\mathfrak{R}] \cdot {}^{-}\alpha = \mathfrak{V} + \mathfrak{S}\mathfrak{U} \tag{29}$$

both of which are seen to have the general form

$$\mathfrak{M}x = y. \tag{30}$$

In (30), x is an N-dimensional complex vector whose components are the coefficients of the normal modes in the two waveguides, $\mathfrak{M}$ is an $N \times N$ complex matrix characterizing the discontinuity, and y is an excitation vector due to the incident wave.

The obvious method of solving (30) is to compute the inverse of $\mathfrak{M}$ and thus directly obtain

$$x = \mathfrak{M}^{-1}y. \tag{31}$$

However, we should recognize that it is the solution vector x which is required, and that computing the inverse is not always the best equation solving technique. For example, because the modal coefficients may decrease slowly with mode index, an accurate approximation of the physical problem often requires that $\mathfrak{M}$ be a very large matrix, and inversion procedures for large complex matrices require considerable computational effort. An alternate approach is therefore suggested, namely the solution of (30) by a method of iteration.

In an iterative algorithm, we begin with an initial "guess" for the solution and, from this, generate a supposedly improved solution, repeating the process until successive iterations give results which

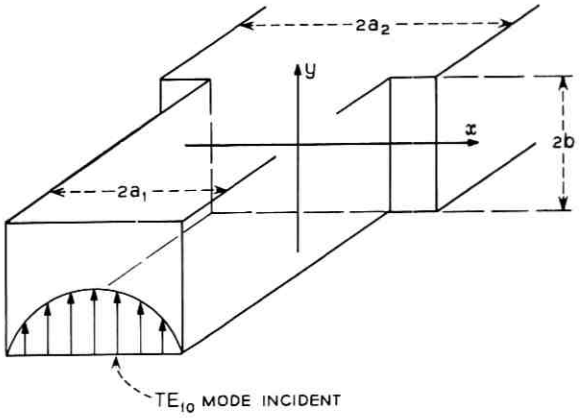


Fig. 3—H-plane discontinuity in a rectangular waveguide. The incident mode is a TE_{10} mode (electric field vertical).

agree to within some prescribed norm. The solution to which the procedure converges must, of course, be *independent of the initial assumption.*

A tempting iterative procedure for the present problem is suggested by writing (28) in the form

$$^{+}\alpha = \mathfrak{U} + \mathfrak{R}\mathfrak{V} + \mathfrak{R}\,\mathfrak{S}\cdot{}^{+}\alpha \tag{32}$$

with an initial assumption

$$^{+}\alpha^{(0)} = \mathfrak{U} + \mathfrak{R}\mathfrak{V}. \tag{33}$$

Physically this corresponds to first assuming the aperture electric field to be that of the unperturbed incident wave, and calculating the

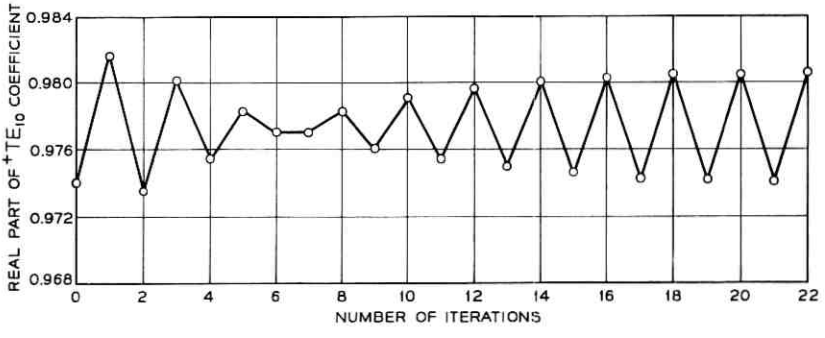


Fig. 4—Transmission coefficient of an H-plane discontinuity.

TE and TM modal coefficients in the + waveguide on this basis. The corresponding magnetic field is then determined on the + side of the aperture and, with the aid of the appropriate continuity condition, is used to find the magnetic field and subsequently an "improved" electric field in the − region. This second guess for the aperture electric field is then used to repeat the process, etc.

As a test, this algorithm was applied to analysis of an H-plane discontinuity in a rectangular waveguide, the incident wave being a TE_{10} mode of normalized amplitude [see (14)] as in Fig. 3. The dimensions were $ka_2 = 4.5$ and $ka_1 = 3.5$ where k is the free space wavenumber. With this choice of parameters, only the TE_{10} modes can propagate in each guide. (This problem is discussed in further detail in Section V.)

Fig. 4 shows the result of calculating the real part of the modal coefficient for the TE_{10} mode in the larger waveguide, plotted as a distribution of points giving the value at each iteration. The Fourier series for this particular example was truncated after twenty five terms. Aside from a small amplitude oscillation of less than one percent rms, the results seem reasonable, especially in view of calculations for the mean-square errors ε_E and ε_H, which are illustrated in Fig. 5. These

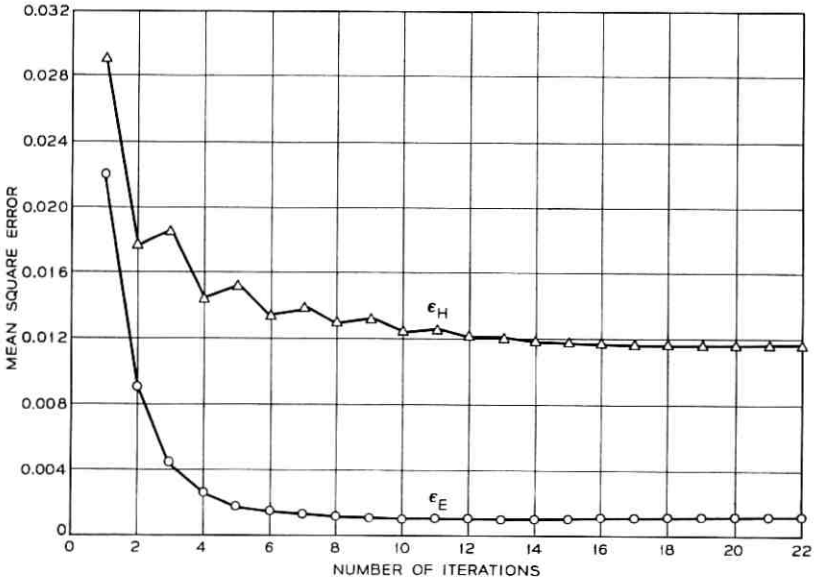


Fig. 5—Mean square errors in transverse electric and magnetic fields for an H-plane discontinuity.

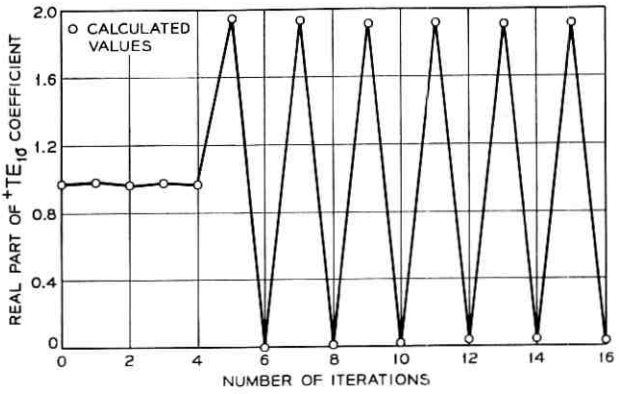


Fig. 6—Oscillatory instability of a nonconvergent algorithm for the H-plane discontinuity analysis.

decrease monotonically with succeeding iterations, ε_E approaching approximately 0.001 and ε_H a value of about 0.012. The larger error in the magnetic field can be attributed to the singular behavior in H near the corner of the discontinuity. The asymptotic value of the energy parameter ε_P is approximately 0.007.

The apparently accurate results obtained using this algorithm are in fact quite deceptive, and may actually be attributed to the propitious initial choice for the aperture electric field. It will be recalled that a very important criterion for validity of an iterative procedure is that the results be independent of the initial assumption. In order to determine whether such a criterion is satisfied for this particular algorithm, the TE_{10} modal coefficient for the larger waveguide was arbitrarily doubled after the fourth iteration, which is equivalent to deliberately assuming a poor initial choice for the aperture field. The effect, shown in Fig. 6, indicates that the algorithm does not relax to the previous values, but continues to oscillate with a large amplitude. Similarly large fluctuations occur in ε_E, ε_H, and ε_P, the conclusion being that this particular precedure is not satisfactory, and can be expected to give reasonable results only if the initial choice is a very good one. The reason for this instability will become apparent after we consider those aspects of matrix-iterative analysis which are relevant to these problems.

In the usual framework for iterative procedures, the matrix equation satisfied by the unknown vector x is written in the form

$$x = \mathfrak{M}x + f, \tag{34}$$

where $\mathfrak{M}$ and f are appropriate to the particular scheme being used.

This leads very naturally to the recursive formula relating the $m + 1$ to the m iteration,

$$x^{(m+1)} = \mathfrak{M}x^{(m)} + f. \tag{35}$$

We denote the error vector at any iteration by $\varepsilon^{(m)}$, where

$$\varepsilon^{(m)} = x^{(m)} - \bar{x}, \tag{36}$$

and $\bar{x}$ is the exact solution to (34). Then, by substituting (36) into (35), we find that $\varepsilon^{(m+1)}$ is related to $\varepsilon^{(m)}$ by

$$\varepsilon^{(m+1)} = \mathfrak{M}\varepsilon^{(m)}. \tag{37}$$

Therefore, the error at the mth iteration is expressible in terms of the initial error by

$$\varepsilon^{(m)} = \mathfrak{M}^m\varepsilon^{(0)}. \tag{37a}$$

For an absolutely converging solution we thus require that

$$\varepsilon^{(m)} \to 0 \quad \text{as} \quad m \to \infty \tag{38}$$

regardless of the initial guess $x^{(0)}$. This is equivalent to

$$\mathfrak{M}^m \to 0 \quad \text{as} \quad m \to \infty \tag{39}$$

where the 0 in (38) and (39) denotes a null vector or matrix, respectively. It can be shown[4] that an $N \times N$ complex matrix $\mathfrak{M}$ is "convergent", in the sense of (39), if and only if all the eigenvalues λ_i of $\mathfrak{M}$ magnitude less than unity, i.e.,

$$|\lambda_i| < 1 \quad \text{all} \quad i. \tag{40}$$

We can easily see why this requirement will guarantee convergence, at least for the special case where the eigenvectors α_i of $\mathfrak{M}$ span the space of N-dimensional complex vectors. The initial error is then expressible as

$$\varepsilon^{(0)} = \sum_{i=1}^{N} C_i\alpha_i , \tag{41}$$

where the C_i are constants. The error at the mth iteration then becomes, from (37),

$$\varepsilon^{(m)} = \sum_{i=1}^{N} C_i\mathfrak{M}^m\alpha_i = \sum_{i=1}^{N} C_i\lambda_i^m\alpha_i \tag{42}$$

and, as $m \to \infty$, each term in the sum approaches zero, provided, of course, that (40) is satisfied.

Equation (42) also reveals useful information concerning the speed of convergence, which is seen to depend on how close the magnitudes of the λ_i are to unity. Clearly if the largest eigenvalue is very near one in magnitude, the convergence will be very slow and hence a large number of iterations will be required. It would be logical, in the light of this reasoning, to evaluate the magnitude of the largest eigenvalue for the H-plane discontinuity problem discussed proviously. Unfortunately, because of the geometrical asymmetry, the matrix $\mathfrak{M}$ is not Hermitian and so the usual computational techniques for determining eigenvalues cannot be used. We can, however, find an upper bound for the modulus of the maximum eigenvalue, $\rho(\mathfrak{M}) = \max \{| \lambda_i |\}$, given by[5]

$$\rho(\mathfrak{M}) \leqq [\rho(\mathfrak{M}^\dagger \mathfrak{M})]^{\frac{1}{2}}. \tag{43}$$

where $\dagger$ denotes the conjugate transpose matrix. Note that $\mathfrak{M}^\dagger\mathfrak{M}$ is Hermitian, so that standard computer programs can be used to evaluate its eigenvalues. We find for the previous H-plane problem that $|\lambda_i| \leqq 1.16$ which, although, not conclusive, shows the possibility of such an oscillatory instability.

One technique which is suggested as a means of obtaining a convergent algorithm is called the "Richardson" or "point-Jacobi" method.[5] In this approach, the matrix $\mathfrak{M}$ is first partitioned as

$$\mathfrak{M} = \mathfrak{D} + \mathfrak{L} + \mathfrak{U}, \tag{44}$$

where $\mathfrak{D}$ is a diagonal matrix containing the diagonal terms of $\mathfrak{M}$, $\mathfrak{L}$ is a strictly lower triangular matrix and $\mathfrak{U}$ is a strictly upper triangular matrix. The system (34) is then written as

$$(\mathfrak{I} - \mathfrak{D})x = (\mathfrak{L} + \mathfrak{U})x + f \tag{45}$$

from which

$$\begin{aligned} x &= (\mathfrak{I} - \mathfrak{D})^{-1}(\mathfrak{L} + \mathfrak{U})x + (\mathfrak{I} - \mathfrak{D})^{-1}f \\ &= \mathfrak{M}_R x + f_R \,, \end{aligned} \tag{46}$$

(46) being the matrix representation of the "Richardson" or "point-Jacobi" method.

A modification of this procedure is referred to as the "Gauss-Seidel" or "point-single-step" iteration method.[6] Note that in solving (46) by iteration, the components of $x^{(m+1)}$ are all computed from the components of $x^{(m)}$. Intuitively, it would seem more attractive to use the latest estimates of x, i.e., in computing $x_i^{(m+1)}$ we should use, wherever

they appear, the components $x_k^{(m+1)}$ $(k < j)$ already computed, and in this way utilize the most accurate information available. This procedure is, in fact, easier to implement in a computer program and, in addition to requiring less storage, often has better convergence properties than Richardson's method. It may be shown that the matrix representation, analogous to (46), for the "Gauss-Seidel" method is

$$x = \mathfrak{M}_G x + f_G \tag{47}$$

$$\mathfrak{M}_G = (\mathfrak{s} - \mathfrak{D} - \mathfrak{L})^{-1}\mathfrak{U} \tag{48}$$

$$f_G = (\mathfrak{s} - \mathfrak{D} - \mathfrak{L})^{-1} f,$$

$\mathfrak{D}$, $\mathfrak{L}$, and $\mathfrak{U}$ having been defined previously.

The eigenvalue condition given in (40) is, of course, a very restrictive one, so that iteration procedures cannot be applied with success to every system of equations. However, when a convergent matrix $\mathfrak{M}$ can be found, the methods which have been discussed offer several distinct advantages over a straightforward matrix inversion. For example, if the order of the system is N, then it can be shown that each iteration requires approximately N^2 multiply-add operations.* On the other hand, an inversion requires *at least* N^3 multiply-adds, so that the relative saving is the ratio of the order N to the number of iterations required. It is often the case that the maximum eigenvalue is so small that the number of iterations required for an accuracy equivalent to that obtained by inversion is considerably less than N.

Iterative methods also have the property that the solution accuracy is "adjustable", in the sense that once the solution has converged to the point where some norm, e.g., ε_E or ε_H defined previously, is less than a specified tolerance, the iteration process can be terminated. This property is especially attractive in view of the fact that truncation errors have already been introduced, and it would therefore be superfluous to accurately invert a system which is itself approximate. By having the option of terminating the iterative procedure, we introduce an additional degree of freedom by which we can optimize the computational program.

V. APPLICATIONS

The iterative techniques discussed in the previous section will now be applied to two problems of interest, namely the H-plane discontinuity

* A multiply-add consists of the multiplication of two complex numbers and the adding of the result to a third complex number. For repetitive computational algorithms, the number of multiply-adds is a measure of the computational effort required.

problem mentioned in Section IV and the analysis of $TE_{11} \rightarrow TM_{11}$ mode conversion at a step discontinuity in a circular waveguide. In both of these examples the required matrix elements may be expressed in convenient closed forms, which considerably reduces the required computational effort.

5.1 *H-plane Discontinuity in a Rectangular Guide*

In Section IV, the H-plane discontinuity of Fig. 3 was analyzed using an iterative algorithm which was observed to exhibit an oscillatory instability when initiated with a poor approximation to the actual solution. It was concluded that this was due to the eigenvalues of the iteration matrix $\mathfrak{M}$ being close to, or perhaps greater than unity. We now consider the same problem using the Gauss-Seidel method which, on the basis of the previous discussion, is expected to improve matters substantially.

We assume a normalized TE_{10} mode incident from the smaller guide. Because of the symmetry of the junction, such a wave excites only TE modes in both the + and − regions. The problem is, of course, to determine the corresponding modal coefficients for the fields on each side of the discontinuity. We shall present results only for the transmitted TE_{10} mode, which is the only mode propagating in the larger waveguide for the present dimensions, $ka_1 = 3.5$, $ka_2 = 4.5$.

It can be shown[9] that the normalized vector wave functions are given by

$$\begin{aligned} {}^{-}\mathbf{E}_p'' &= \mathbf{e}_y \frac{i\omega\mu}{\sqrt{2a_1 b}} \sin\left[\frac{p\pi}{2a_1}(x + a_1)\right] \\ {}^{+}\mathbf{E}_p'' &= \mathbf{e}_y \frac{i\omega\mu}{\sqrt{2a_1 b}} \sin\left[\frac{p\pi}{2a_2}(x + a_2)\right], \end{aligned} \tag{49}$$

where a_1, a_2, and b are the dimensions of the guide as shown and $\mathbf{e}_y$ is a unit vector in the y direction. The corresponding magnetic vector wave functions can be found from (12). The respective propagation constants are

$$\begin{aligned} {}^{-}h_p'' &= \left[k^2 - \left(\frac{p\pi}{2a_1}\right)^2\right]^{\frac{1}{2}} \\ {}^{+}h_p'' &= \left[k^2 - \left(\frac{p\pi}{2a_2}\right)^2\right]^{\frac{1}{2}}. \end{aligned} \tag{50}$$

From (21) and (49) we find that the coupling coefficients for the electric fields are given by

$$\mathcal{E}_{pq}\begin{bmatrix} -, + \\ '' \ , \ '' \end{bmatrix} = 2\omega^2\mu^2\sqrt{\frac{a_1}{a_2}}\sin\frac{p\pi}{2}\sin\frac{q\pi}{2}\left\{\frac{1}{p\pi + \frac{a_1}{a_2}q\pi}\sin\left(\frac{p\pi}{2} + \frac{a_1}{a_2}q\frac{\pi}{2}\right)\right.$$

$$\left. + \frac{1}{p\pi - \frac{a_1}{a_2}q\pi}\sin\left(\frac{p\pi}{2} - \frac{a_1}{a_2}q\frac{\pi}{2}\right)\right\}. \qquad (51)$$

The appropriate magnetic field coupling coefficients are easily found from (51) using the relation

$$\mathcal{H}_{pq}\begin{bmatrix} +, - \\ '' \ , \ '' \end{bmatrix} = -\frac{1}{{}^{+}Z_p'' {}^{-}Z_q''^{*}}\,\mathcal{E}_{qp}\begin{bmatrix} -, + \\ '' \ , \ '' \end{bmatrix}, \qquad (52)$$

where Z_p denotes the modal impedance, equal to

$${}^{\pm}Z_p'' = \frac{\omega\mu}{{}^{\pm}h_p''}. \qquad (53)$$

The results for the TE_{10} modal coefficient, ${}^{+}B_1$, as obtained using the Gauss-Seidel iteration method, are conveniently represented in Table I. Also given are the numerical values for the error parameters ε_P, ε_E and ε_H. Truncation for this example was at 25 modes in each waveguide.

We conclude that for most applications, two iterations would probably have been sufficient, corresponding to a saving of greater than 90 percent in actual execution time, compared to a matrix inversion. The reason for this extremely rapid convergence is, as expected, in the magnitude of the largest eigenvalue, which was found, using (43), to be less than 0.078.

As a means of establishing the convergence of the normal mode solution, we have plotted in Figs. 7 and 8, the mean square errors ε_E and ε_H, respectively, as a function of the number of modes taken.

TABLE I — RESULTS USING THE GAUSS-SEIDEL METHOD FOR ANALYSIS OF THE H-PLANE DISCONTINUITY

Iteration number	${}^{+}B_1$		Energy coefficient ϵ_P	rms error ϵ_P	rms error ϵ_H
	Real	Imaginary			
1	0.97445	0.00951	-0.62×10^{-2}	0.5×10^{-5}	0.01445
2	0.97747	0.00455	-0.74×10^{-6}	0.49×10^{-5}	0.01355
3	0.97747	0.00426	0.85×10^{-5}	0.49×10^{-5}	0.01350
4	0.97747	0.00424	0.71×10^{-6}	0.49×10^{-5}	0.01349
5	0.97747	0.00424	0.51×10^{-7}	0.49×10^{-5}	0.01349

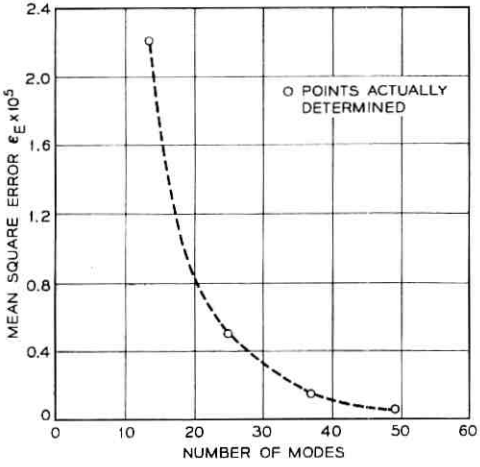


Fig. 7—Mean square error ε_E, as function of the number of modes used, for H-plane discontinuity.

These figures give genuine significance to the term "convergence in mean square", since they indicate that the use of more terms leads to better agreement with the boundary conditions in the mean square sense. Again, the error is uniformly higher for the magnetic field than for the electric field, due to the singularity at the corner of the discontinuity.

It is finally of interest to determine the effect of a poor initial estimate

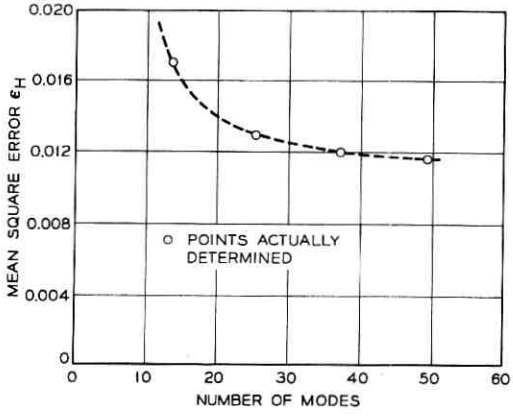


Fig. 8—Mean square error ε_H as function of the number of modes used, for H-plane discontinuity.

TABLE II — RESULTS OF SPOILING ELECTRIC FIELD AT 4TH ITERATION

Iteration number	TE_{10} coefficient	
	Real	Imaginary
3	0.97747	0.00426
4	1.94165	0.02572
5	0.98052	−0.00063
6	0.97747	0.00396
7	0.97746	0.00423

of the solution. We find that, unlike the simple algorithm discussed in the previous section, the Gauss-Seidel procedure is very stable, returning to the correct "steady state" solution within a few iterations after the spoiling was introduced. The results are shown in Table II, again for a truncation of 25 modes in each waveguide.

5.2 *Mode Conversion at a Step Discontinuity in a Circular Waveguide*

The second problem to which these techniques were applied is that of calculating the $TE_{11} \rightarrow TM_{11}$ mode conversion at an abrupt discontinuity in a circular waveguide. Recent studies have indicated that this configuration is a very efficient transducer for use in dual mode conical horns.[10] The discontinuity is illustrated in Fig. 9 which shows the TE_{11} mode, incident from the smaller guide, being converted to a combination of TE_{11} and TM_{11} modes propagating in the larger guide.

The normalized TE and TM vector wave functions are known[11] and, fortunately, it is possible to determine the appropriate coupling coefficients. For the elements of the matrix we find that

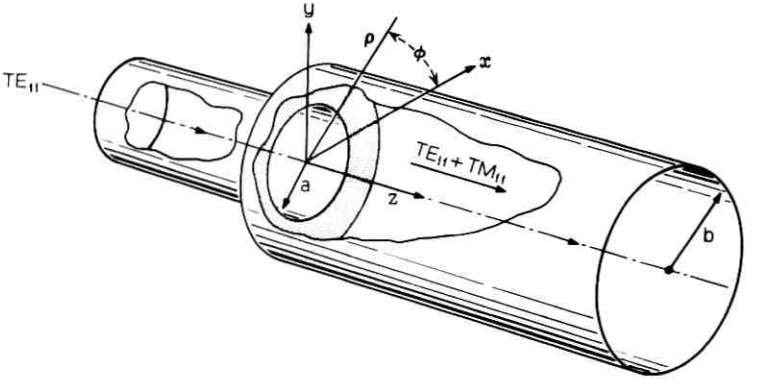


Fig. 9 — $TE_{11} - TM_{11}$ mode conversion at a discontinuity in a circular waveguide.

$$\mathcal{E}_{pq}\begin{bmatrix} -, + \\ ', ' \end{bmatrix} = \frac{2\ {}^{-}h_p'\ {}^{+}h_q'^{*} x_q J_1\left(x_q \frac{a}{b}\right)}{\left(\frac{b^2}{a^2} x_p^2 - x_q^2\right) J_2(x_q)} \tag{54}$$

$$\mathcal{E}_{pq}\begin{bmatrix} -, + \\ '', '' \end{bmatrix} = \frac{2\omega^2\mu^2 b y_p^2 y_q \left[J_0\left(\frac{a}{b} y_q\right) - \frac{b}{a} J_1\left(\frac{a}{b} y_q\right)\right]}{\left(\frac{b^2}{a^2} y_p^2 - y_q^2\right)\sqrt{(y_p^2 - 1)(y_q^2 - 1)}\ J_1(y_q)} \tag{55}$$

$$\mathcal{E}_{pq}\begin{bmatrix} -, + \\ '', ' \end{bmatrix} = \frac{2\omega\mu\ {}^{+}h_q'^{*} J_1\left(x_q \frac{a}{b}\right)}{x_q J_2(x_q)\sqrt{(y_p^2 - 1)}} \tag{56}$$

$$\mathcal{E}_{pq}\begin{bmatrix} -, + \\ ', '' \end{bmatrix} = 0. \tag{57}$$

In (59) through (62) we have used the following notation:

a — radius of smaller waveguide,
b — radius of larger waveguide,
x_p — pth zero of $J_1(x)$,
y_p — pth zero of $J_1'(y)$.

The elements of the matrix $\mathfrak{C}$ can easily be found from the impedance relations of (11) and (12).

One parameter which has been found to be useful in characterizing the mode conversion properties of the discontinuity is the conversion coefficient C, defined as the ratio of the ρ-components of electric field for the two modes evaluated at the wall of the larger waveguide, i.e.,

$$C = 20 \log_{10} \left| \frac{E_\rho^{TM}}{E_\rho^{TE}} \right|_{\rho=b} \text{dB}. \tag{58}$$

This quantity was calculated for the particular discontinuity $a = 1.05''$, $b = 1.4''$ over the frequency range 5.2 to 7.0 kHz, these parameters having been chosen for purposes of comparison with available experimental data. Truncation of the normal mode expansion was made after twenty-five TE and twenty-five TM modes in each waveguide. The iterative sequence was terminated when successive values of the modal amplitudes differed by less than 10^{-6}. It was found that typical values for the error criteria are $\varepsilon_P \approx 10^{-7}$, and ε_E, $\varepsilon_H \approx 0.015$. These results indicate that for a given accuracy, a much lower value can be expected for ε_P, which is a function only of the lower-order modes,

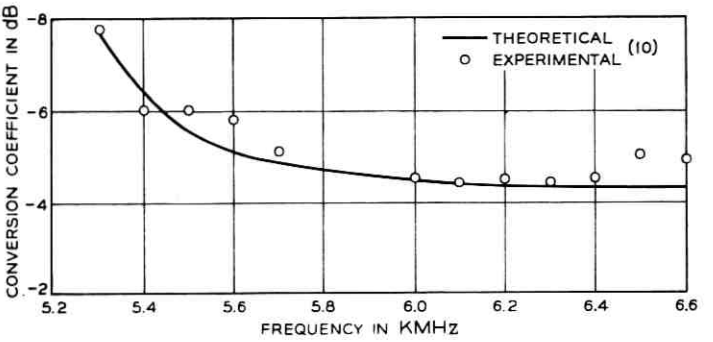


Fig. 10—$TE_{11} \rightarrow TM_{11}$ conversion coefficient of a step discontinuity in a circular waveguide. $a = 1.05''$, $b = 1.4''$.

than for ε_E and ε_H, which depend on the higher-order terms as well.

In Fig. 10 we have plotted the computed values of the conversion coefficient, defined in (58), as a function of frequency. Also shown, as discrete points, are experimental results obtained previously.[10] The theoretical values are seen to be in very good agreement.

VI. SUMMARY AND CONCLUSIONS

In this paper, we have considered the solution of those matrix equations which arise in the analysis of a class of waveguide discontinuity problems. In searching for criteria to estimate the accuracy of computed results, we have found that the uniqueness theorem itself yields a convenient set of error parameters which are easily implemented in the computational program.

It is suggested that an iterative technique, particularly the "Gauss-Seidel" or "point-single-step" method often leads to a rapidly converging solution, thus offering several advantages over the usual invertive procedures, particularly in the speed of execution. Of particular interest is the fact that when this method is applied to the analysis of $TE_{11} \rightarrow TM_{11}$ mode conversion at an abrupt discontinuity in a circular waveguide, it yields a rapidly convergent and accurate solution. This has been established not only on the basis of theoretical error criteria, but also by comparison with experimental results previously obtained.

REFERENCES

1. Lewin, L., *Advanced Theory of Waveguides*, Iliffe and Sons, London, 1951.
2. Stratton, J. A., *Electromagnetic Theory*, McGraw-Hill Book Co., Inc., New York, 1941, p. 131.

3. Borgnis, F. E. and Papas, C. H., Electromagnetic Waveguides and Resonators, *Encyclopedia of Physics*, Volume XVI, Springer, Berlin, 1958, p. 300.
4. Varga, R. S., *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N. J., 1965, p. 13.
5. *Ibid.*, p. 11.
6. *Ibid.*, p. 57.
7. *Ibid.*, p. 58.
8. *Ibid.*, p. 16.
9. Borgnis and Papas, *op. cit.*, p. 315.
10. Nagelberg, E. R. and Shefer, J., Mode Conversion in Circular Waveguides, B.S.T.J., *44*, September, 1965, pp. 1321–1338.
11. Borgnis and Papas, *op. cit.*, p. 322.

Contributors to This Issue

W. James Cole, B.S.E.E., 1963, Lehigh University; M.S.E.E., 1964, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1963—. Mr. Cole has been engaged in research into the numerical solution of electromagnetic boundary value problems. Member, IEEE, SIAM, Tau Beta Pi.

J. A. Collinson, A.B., 1950, Oberlin College; M.S., 1951, Yale University; Ph.D., 1954, Yale University; Bell Telephone Laboratories, 1962—. He has worked on gas lasers and placed emphasis on frequency characteristics and atmospheric transmission of laser beams. Member, American Physical Society, Sigma Xi, Phi Beta Kappa.

A. Goetzberger, Ph.D. in Science, 1955, University of Munich; Bell Telephone Laboratories, 1963—. Mr. Goetzberger is a supervisor in the metal insulator semiconductor group. Prior to 1963, he was with the Shockley Laboratory in Palo Alto, where he worked on junction imperfections and avalanche breakdown phenomena in silicon. He also participated in the development of a power transistor. Member, American Physical Society, IEEE, Electrochemical Society.

T. T. Kadota, B.S., 1953, Yokahama National University (Japan); M.S., 1956, Ph.D., 1960, University of California (Berkeley); Bell Telephone Laboratories, 1960—. Mr. Kadota has been engaged in the study of noise theory with application to optimum detection theory. Member, Sigma Xi.

Dankwart Koehler, Dipl.-Ing, 1955, Technische Hochschule Stuttgart (Germany); M.S., 1955, Georgia Institute of Technology, World Student Fund and Fulbright Student 1953–1954; Dr.-Ing., 1958, Technische Hochschule Stuttgart (Germany); Research Institute Telefunken (Germany) 1957–1960; Assistant Professor, Georgia Institute of Technology, 1960–1961; Bell Telephone Laboratories, 1961—. Mr. Koehler has been engaged in the design of circuits for high-speed

pulse code modulation terminals. His special interest has been devoted to high-speed circuit principles and semiconductor device characterization. He is presently supervising a group responsible for the development of high-speed PCM multiplex terminals. Senior member, IEEE, Sigma Xi.

C. M. Nagel, Jr., B.S., 1964, M.S., 1967, Stevens Institute of Technology; Bell Telephone Laboratories, 1965—. Since joining the Laboratories, Mr. Nagel has been engaged in research in the numerical aspects of various wave guide problems, Faraday rotation in the ionosphere, and electromagnetic scattering. In addition, he has investigated the application of air bearings as supports for acoustical delay lines and participated briefly in computer center operations. He is presently participating in the GSP program in the Detection Systems Laboratory. Member, Tau Beta Pi, Pi Delta Epsilon, American Association for the Advancement of Science.

Elliott R. Nagelberg, B.E.E., 1959, City College of New York; M.E.E., 1961, New York University; Ph.D., 1964, California Institute of Technology; Bell Telephone Laboratories, 1964—. Mr. Nagelberg has been concerned with problems involving microwave antennas and propagation. Member, IEEE, American Physical Society, Eta Kappa Nu, Sigma Xi.

E. H. Nicollian, M.E., 1951, Stevens Institute of Technology; M.A. (Physics) 1956, Columbia University; Bell Telephone Laboratories, 1957—. Mr. Nicollian's work has been in semiconductor device physics. He is currently engaged in research on the electrical properties of semiconductor-insulator interfaces. Member, American Physical Society, Electrochemical Society, RESA, AAAS.

Burton R. Saltzberg, B.E.E., 1954, New York University; M.S., 1955, University of Wisconsin; Eng. Sc.D., 1964, New York University; Bell Telephone Laboratories, 1957—. Mr. Saltzberg has been engaged in the design, development and analysis of data transmission systems. He is currently a member of the Data Theory Department. Member, IEEE, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

Man Mohan Sondhi, B.Sc. (Honours), 1950, Delhi University (Delhi, India); D.I.I.Sc., 1953, Indian Institute of Science (Bangalore,

India); M.S., 1955, Ph.D., 1957, University of Wisconsin; Bell Telephone Laboratories, 1962—. Mr. Sondhi has worked on problems concerning the processing and transmission of speech signals. He is currently interested in similar problems as well as modeling the detection of auditory and visual signals by human beings.

M. SUBRAMANIAN, B.Sc., 1953, Madras University (India); Dip. Madras Inst. Tech. (India) 1956; M.S.E.E., 1961 and Ph.D., 1964, Purdue University; Bell Telephone Laboratories, 1966—. Mr. Subramanian's earlier research in microwaves involved work on receivers, parametric amplifiers and ferroelectric materials. His experience in quantum electronics includes nonlinear optics and cathodoluminescence. Presently he is studying the effect of atmospheric turbulence on laser beam. Member, IEEE, Eta Kappa Nu, Sigma Pi Sigma.