THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 48 December 1969 Number 10

On Communication of Analog Data from a Bounded Source Space

By AARON D. WYNER and JACOB ZIV

(Manuscript received June 5, 1969)

We consider the problem of the transmission of discrete-time analog data with a variety of fidelity criteria. The outputs of the analog source are assumed to belong to a bounded set. Bounds on the minimum achievable average distortion for memoryless sources are derived both for the case where the coding delay is infinite (an extension of the Shannon Theory) and also for some cases where the coding delay is finite. Several examples are given, for which the upper and lower bounds coincide.

Further, we discuss the case where the assumption of the existence of a probabilistic model for the source is dropped. We adopt as our fidelity criterion the supremum over all possible source-output n-sequences $\mathbf{x}$*, of the conditional expectation of the distortion given* $\mathbf{x}$ *("guaranteed distortion"). The Shannon Theory is not directly applicable in determining the minimum guaranteed distortion. We do obtain results for two important cases. Some generalizations and applications are also discussed.*

I. INTRODUCTION

In this paper we are concerned with communication of discrete-time analog data over a communication channel with a variety of fidelity criteria. The central assumption about the analog source is that its

outputs belong to a bounded set, typically the interval $[-A/2, A/2]$. We begin with a rough outline of our results, leaving the precise formulation and statement to Section II. Proofs are found in Section III.

Suppose that we have a data source which emits a sequence of symbols $x_1, x_2, \cdots \in \mathcal{X}$ (an arbitrary set) at a rate of ρ_S per second. This sequence is fed into an "encoder" which assigns to each successive block of n source symbols, say $\mathbf{x} = (x_1, x_2, \cdots, x_n)$, a channel input of duration $n/\rho_S = T$ seconds. At the receiving end of the channel, the T-second output is transformed by a "decoder" into an n-sequence, say $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_n)$, which is delivered to the destination. The "distortion" between the source output sequence $\mathbf{x}$ and the received sequence $\hat{\mathbf{x}}$ is defined as $d^{(n)}(\mathbf{x}, \hat{\mathbf{x}}) = n^{-1} \sum_{k=1}^{n} d(x_k, \hat{x}_k)$, where $d(x, \hat{x}) \geqq 0$ is an arbitrary function.

The classical problem is that of a "memoryless" source, where successive source outputs are statistically independent with identical probability distribution. In this case it is meaningful to let the system performance criterion (fidelity criterion) be the statistical expectation of the distortion $d^{(n)}(\mathbf{x}, \hat{\mathbf{x}})$. A quantity of interest is $\bar{d}^*(T)$, the smallest attainable value of the fidelity criterion when the coding delay is T seconds. The Shannon Theory gives the asymptotic behavior of $\bar{d}^*(T)$ as $T \to \infty$. In many cases this limit is difficult to evaluate analytically. Theorem 1 (in Section 2.2) considers the case where the source output set $\mathcal{X} = [-A/2, A/2]$, and the function $d(x, \hat{x})$ depends only on the difference $\hat{x} - x$. This theorem gives a lower bound on $\lim_{T\to\infty} \bar{d}^*(T)$. The examples which follow this theorem illustrate the applicability and utility of the bound.

There are two cases in which we are particularly interested. In the first, the source set $\mathcal{X} = \{0, 1, \cdots, K-1\}$ with a uniform distribution, and $d(x, \hat{x}) = 0$ or 1 according as $x = \hat{x}$ or $x \neq \hat{x}$. Thus the fidelity criterion is the error-rate. For this case let $\bar{d}^*(T) = P_e(K, T)$. In the second case, $\mathcal{X} = [-A/2, A/2]$ with a uniform distribution, and $d(x, \hat{x}) = 0$ or 1 according as $|x - \hat{x}| < \delta$ or $|x - \hat{x}| \geqq \delta$ (where $\delta > 0$). In this case let $\bar{d}^*(T) = Q(T, A, \delta)$. It turns out that P_e and Q are intimately related. In fact it is a consequence of Theorem 2 (Section 2.2) that if $A/(2\delta) = K_0$, an integer, then $Q(T, A, \delta) = P_e(T, K_0)$. This result is valid for all values of the delay parameter T. From this result it can be deduced that the optimal encoder for the analog source $\mathcal{X} = [-A/2, A/2]$ is a "uniform" quantizer followed by an optimal "digital" encoder. This is the only known case for which analog-to-digital conversion is known to be optimal for finite T for the transmission of analog data from a memoryless source.

We now drop the assumption of a memoryless source. In fact we do

not even assume that there is a probabilistic model for the source. Instead of the expectation of the distortion, we adopt as our fidelity criterion, the supremum, over all possible source output n-sequences $\mathbf{x}$, of the conditional expectation of the distortion given $\mathbf{x}$. We call this criterion the "guaranteed distortion". Let $\hat{d}^*(T)$ be the minimum attainable guaranteed distortion for a system with delay parameter T. The Shannon Theory is not directly applicable in determining $\hat{d}^*(T)$. We do obtain results for the two interesting cases discussed below.

In the first, $\mathfrak{X} = \{0, 1, \cdots, K-1\}$ and $d(x, \hat{x}) = 0$ or 1, respectively, when $x = \hat{x}$ or $x \neq \hat{x}$. For this case let $\hat{d}^*(T) = \hat{P}_e(T, K)$. It is a consequence of Theorem 3 (Section 2.3) that $\text{limit}_{T\to\infty} \hat{P}_e(T, K) = \text{limit}_{T\to\infty} P_e(T, K)$, which is known from the Shannon Theory.

In the second case, $\mathfrak{X} = [-A/2, A/2]$ and $d(x, \hat{x}) = 0$ or 1, respectively, when $|x - \hat{x}| < \delta$ or $|x - \hat{x}| \geqq \delta$. For this case, let $\hat{d}^*(T) = \hat{Q}(T, A, \delta)$. Theorem 4 (Section 2.3) relates $\hat{P}_e$ and $\hat{Q}$ by

$$\hat{Q}(T, A, \delta) = \hat{P}_e(T, M),$$

where M is the unique integer satisfying $(M-1) \leqq A/(2\delta) < M$. Here too, we can deduce the optimality of analog-to-digital conversion. Theorem 4 is generalized by Theorem 5 (Section 2.4) to apply to an arbitrary set $\mathfrak{X}$ with a distance-like measure defined on it (replacing $|x - \hat{x}|$).

In Section 2.5, we give some applications of the above results. In particular we obtain some results for the distortion $d(x, \hat{x}) = |x - \hat{x}|^\alpha$.

In order to state our results completely and precisely, it is unfortunately necessary to give a rather large collection of definitions and to introduce a large number of symbols. In order to ease the reader's burden somewhat, we have included a glossary of symbols in the appendix.

II. STATEMENT OF THE PROBLEM AND PRINCIPAL RESULTS

In Section 2.1 we define a "channel" (and its "capacity") in a very general and abstract way. We do this because the nature of the channel does not figure explicitly in our results (except for the channel capacity), and we want our results to apply as broadly as possible. In Section 2.2 we describe the communication system which we shall consider, and state our results for the case of a "memoryless" information source. The remainder of the results follows in Sections 2.3–2.5.

2.1 *Channel and Channel Capacity*

A *channel* is defined as follows. For every $T > 0$ we have a set $\mathcal{W}_T$ of "allowable" inputs and a set $\mathcal{Z}_T$ of possible outputs. Every T

seconds some $w \in \mathcal{W}_T$ is transmitted through the channel, and the channel output z is a member of $\mathcal{Z}_T$. The output is related to the input $w \in \mathcal{W}_T$ by a probability measure μ_w on the set $\mathcal{Z}_T$. Thus given that $w \in \mathcal{W}_T$ is transmitted, the probability that $z \in B$ [where B is a (measurable) subset of $\mathcal{Z}_T$] is $\mu_w(B)$. For example $\mathcal{W}_T$ and $\mathcal{Z}_T$ may be the set of binary sequences of length $[T]^{-}$†. The measure μ_w is then a discrete conditional probability distribution. Another example is the case where $\mathcal{W}_T$ and $\mathcal{Z}_T$ are sets of real valued functions defined on the interval $[0, T]$, and the members of $\mathcal{W}_T$ have "energy" not exceeding PT.

With T specified, a *block code* with parameter N is a set of N pairs $\{(w_i, B_i)\}_{i=1}^{N}$, where $w_i \in \mathcal{W}_T$ are called *code words* and the collection of B_i is a set of disjoint (measurable) subsets of $\mathcal{Z}_T$ called *decoding sets*. If code word $w_i (1 \leqq i \leqq N)$ is transmitted, the resulting error probability is

$$\lambda_i = \Pr \{z \notin B_i \mid w_i \text{ is transmitted}\} = 1 - \mu_{w_i}(B_i). \tag{1}$$

The *word error probability* for the code is

$$\lambda = \max_{1 \leqq i \leqq N} \lambda_i . \tag{2}$$

Let $\lambda^*(T, N)$ be the smallest attainable word error probability for a code with parameters T and N. The *channel capacity* C is defined as the supremum of those numbers $R \geqq 0$, for which

$$\lambda^*(T, [e^{RT}]^{-}) \to 0, \quad \text{as} \quad T \to \infty .$$

Let us define the *average word error probability* by

$$\bar{\lambda} = \frac{1}{N} \sum_{i=1}^{N} \lambda_i . \tag{3}$$

Thus $\bar{\lambda}$ is the resulting average error probability which results when each of the N code words are equally likely to be transmitted. Let us define $\bar{\lambda}^*(T, N)$ as the smallest attainable value of $\bar{\lambda}$ for a code with parameters T and N. Since $\bar{\lambda} \leqq \lambda$ for any code, it follows from the above definition of channel capacity that for any $R < C$,

$$\bar{\lambda}^*(T, [e^{RT}]^{-}) \to 0, \quad \text{as} \quad T \to \infty .$$

Further it is known that for a large class of channels including the memoryless gaussian channel and discrete memoryless channels,[1]

$$\bar{\lambda}^*(T, [e^{CT}]^{-}) \to \tfrac{1}{2}, \quad \text{as} \quad T \to \infty . \tag{4}$$

[It is also true that for many of these same channels if $R > C$,

† Throughout this paper we denote by $[x]^{-}$ and $[x]^{+}$ the largest integer $\leq x$ and the smallest integer $\geq x$ respectively $(0 \leq x < \infty)$.

$\bar{\lambda}^*(T, [e^{RT}]^-)$ tends to 1 as $T \to \infty$, but we do not need this fact here.]

Let us remark here that for a large class of channels (including "memoryless" channels and "finite state channels"), the capacity C is known to be the supremum of a quantity called the "information". In fact this equivalence is the essence of the Fundamental Theorem of Information Theory. It will not be necessary, however, to explore this equivalence further.

2.2 *Memoryless Source and Communication With a Fidelity Criterion*

Consider the communication system of Figure 1. The output of the *source* is a sequence of random variables $X_1, X_2, \cdots$ from an arbitrary subset $\mathfrak{X}$ of Euclidean p-space. Assume that these random variables are statistically independent and identically distributed with probability density function $P_S(x)$, $x \in \mathfrak{X}$. If we allow impulses in the density function, then the X_k can be discrete random variables. Say that the source outputs appear at a rate of ρ_S per second. The *encoder* waits T seconds (called the "delay") during which time $n = \rho_S T$ symbols, say $X_1, X_2, \cdots, X_n \in \mathfrak{X}$, have appeared at its input. (Assume that $\rho_S T$ is an integer.) Denote the T-second output of the source by the random n-vector $\mathbf{X} = (X_1, X_2, \cdots, X_n) \in \mathfrak{X}^n$.

The *channel* is defined as above (Section 2.1), so that during the T seconds which it takes for the n-vector $\mathbf{X}$ to appear, the channel can process an input belonging to the channel input set $\mathcal{W}_T$. It is the task of the encoder to assign to each possible source output n-vector $\mathbf{X} = \mathbf{x}$, a channel input $f_E(\mathbf{x}) \in \mathcal{W}_T$. The channel output is a member Z of the channel output set $\mathcal{Z}_T$, and it is the task of the *decoder* to assign to each possible $Z = z$ an n-vector $\hat{\mathbf{X}} = f_D(z) \in \mathfrak{X}^n$. Note that the source and channel statistics define a joint probability density on the random n-vectors $\mathbf{X}$ and $\hat{\mathbf{X}}$.

Now ideally we would like $\mathbf{X} \equiv \hat{\mathbf{X}}$. But this is most often not possible due to imperfections (for example, noise) in the channel. Thus we define a *fidelity criterion* which we use as a measure of the reliability of the system. Suppose we are given a non-negative *distortion* function $d(x, \hat{x})$ defined on $\mathfrak{X} \times \mathfrak{X}$. Typical choices of the distortion function are $d(x, \hat{x}) = |x - \hat{x}|^s (s > 0)$ when $\mathfrak{X}$ is a subset of the reals (that is, the dimensionality $p = 1$), or the "Hamming" distance

Fig. 1 — Communication system.

$$d(x, \hat{x}) = d_H(x, \hat{x}) = \begin{cases} 0, & x = \hat{x}, \\ 1, & x \neq \hat{x}, \end{cases} \tag{5}$$

where $\mathfrak{X}$ is a discrete (that is, countable) set.

The distortion between the n-vectors, $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_n)$ is

$$d^{(n)}(\mathbf{x}, \hat{\mathbf{x}}) = n^{-1} \sum_1^n d(x_k, \hat{x}_k).$$

Our system performance (fidelity) criterion, which we seek to minimize, is

$$\bar{d} = Ed^{(n)}(\mathbf{X}, \hat{\mathbf{X}}),$$

where E denotes expectation (with respect to the joint probability distribution of $\mathbf{X}$ and $\hat{\mathbf{X}}$). For a given delay T, which corresponds to $n = \rho_S T$, let $\bar{d}^*(T)$ denote the infimum (with respect to all encoder-decoder pairs) of the attainable values of $\bar{d}$ (for given ρ_S and source-channel statistics). Although we usually do not know $\bar{d}^*(T)$ exactly, we do know its asymptotic behavior as $T \to \infty$. We proceed as follows.

For $0 \leqq \beta \leqq \infty$, define $\mathfrak{M}(\beta)$ as the set of probability density functions $p(x, \hat{x})$ defined on $\mathfrak{X} \times \mathfrak{X}$ which satisfy

(*i*) $\int_{\mathfrak{X}} p(x, \hat{x})\, d\hat{x} = P_S(x)$, the source output probability density function,

(*ii*) $\int_{\mathfrak{X}} \int_{\mathfrak{X}} d(x, \hat{x}) p(x, \hat{x})\, dx\, d\hat{x} \leqq \beta$.

The *information* corresponding to the density $p(x, \hat{x}) \in \mathfrak{M}(\beta)$ is defined as

$$I\{p(x, \hat{x})\} = \int_{\mathfrak{X}} \int_{\mathfrak{X}} p(x, \hat{x}) \log \frac{p(x, \hat{x})}{P_S(x) p_2(\hat{x})}\, dx\, d\hat{x}, \tag{6}$$

where $p_2(\hat{x}) = \int_{\mathfrak{X}} p(x, \hat{x})\, dx$. It is easy to show that $I \geqq 0$ with equality if and only if $p(x, \hat{x}) = P_s(x) p_2(\hat{x})$. Finally define the *equivalent rate* of the source

$$R_{eq}(\beta) = \inf_{p(x, \hat{x}) \in \mathfrak{M}(\beta)} I\{p(x, \hat{x})\}. \tag{7}$$

$R_{eq}(\beta)$ is usually called the "rate-distortion function". Note that $R_{eq}(\beta)$ depends only on β and $P_s(x)$.

Let us now return to the quantity $\bar{d}^*(T)$. Shannon's well known theorems tell us the following.[2] For a given communication system (as in Fig. 1),

$$(i)\ \bar{d}^*(T) \geqq \bar{d}_0\,, \qquad \text{for all } T, \qquad (ii)\ \bar{d}^*(T) \to \bar{d}_0\,, \qquad \text{as } T \to \infty\,, \tag{8}$$

where $\bar{d}_0$ is the smallest solution of

$$\rho_s R_{eq}(\bar{d}_0) \leqq C,$$

and C is the capacity of the channel.

Some intuitive insight into the meaning of Shannon's theorem can be gained by thinking of $\rho_s R_{eq}(\beta)$ as the equivalent rate in nats per second of the source (when reproduced with distortion β). It is reasonable then to suppose that the minimum attainable distortion $\bar{d}_0$ is that distortion for which the source rate is just equal to the channel capacity C.

There are two well-known cases for which $R_{eq}(\beta)$ is known explicitly. The first is the case where $\mathfrak{X}$ = the reals, $P_S(x) = (2\pi)^{-\frac{1}{2}} \exp(-x^2/2\sigma^2)$, and $d(x, \hat{x}) = (x - \hat{x})^2$. In this case, $R_{eq}(\beta) = \frac{1}{2} \log \sigma^2/\beta^2$, so that $\bar{d}_0 = \sigma^2 \exp(-2C/\rho_S)$.

The second case (which is important in the sequel) is $\mathfrak{X} = \{0, 1, 2, \cdots, K-1\}$ $(K = 2, 3, \cdots)$, $P_S(x) = \sum_{k=0}^{K-1} (1/K)\delta(x - k)$ [$\delta(x)$ is the unit impulse], and $d(x, \hat{x})$ is given by equation (5). In other words, the source output is a sequence of independent random variables, each equally distributed on the K-ary alphabet $\{0, 1, \cdots, K-1\}$. The quantity $\bar{d}$ is the average fraction of symbols received in error, and is often called the "error-rate". In this case, we write $\bar{d}^*(T) = P_e(T, K)$, where the dependence of P_e on K as well as T is indicated explicitly. For this case it is known that[2]

$$R_{eq}(\beta) = \begin{cases} \log K - h(\beta) - \beta \log (K-1), & \beta \leqq \dfrac{K-1}{K}\,, \\ 0, & \beta \geqq \dfrac{K-1}{K}\,, \end{cases} \tag{9a}$$

where

$$h(\beta) = -\beta \log \beta - (1-\beta) \log (1-\beta), \quad (0 \leqq \beta \leqq 1). \tag{9b}$$

Shannon's theorem, equation (8), tells us that

$$P_e(T, K) \to \gamma(K, \rho_S, C), \qquad T \to \infty, \tag{10a}$$

where $\gamma(K, \rho_S, C)$ is the smallest solution of

$$\rho_s R_{eq}(\gamma) \leqq C, \tag{10b}$$

and C is the channel capacity. A graph of $\gamma(K, \rho_S, C)$ versus C/S_S for

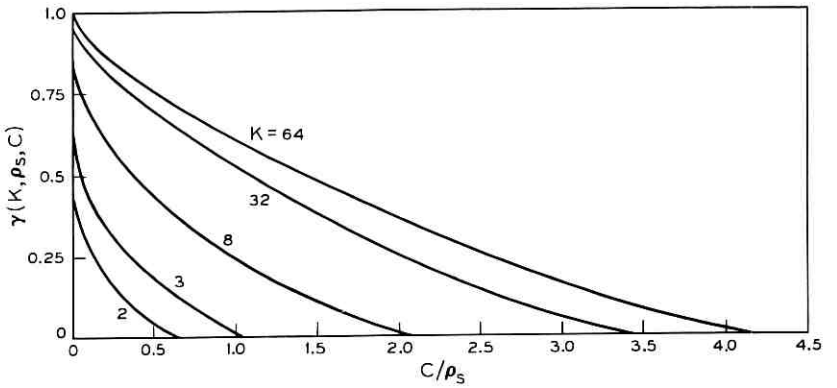

Fig. 2 — $\gamma(K, \rho_s, C)$ versus C/ρ_s (K-a parameter).

various values of K is given in Fig. 2. Notice that $\gamma(K, \rho_S, C)$ decreases from $(K-1)/K$ to zero as C/ρ_S increases from zero to log K.

Let us also remark that the quantity $P_e(T, K)$ is related to $\bar{\lambda}^*(T, M)$ (the smallest attainable average *word* error probability). In fact it is easy to show that

$$\frac{1}{n}\bar{\lambda}^*(T, K^n) \leqq P_e(T, K) \leqq \bar{\lambda}^*(T, K^n) \tag{11}$$

where $n = \rho_S T$ (assumed to be an integer).

Now, in the general case [arbitrary $P_S(x)$ and $d(x, \hat{x})$], it is usually not possible to obtain a closed form expression for $R_{eq}(\beta)$. Theorem 1, which is stated below, gives a useful bound on $R_{eq}(\beta)$ for the case where $P_S(x)$ is a density and $\mathfrak{X}$ is a bounded set. This theorem is an extension of a result of Shannon.[2] The proof is given in Section 3.1.

Let $\mathfrak{X}$ be the interval $[-A/2, A/2]$, where $A(0 < A < \infty)$ is arbitrary. Let the source outputs X have density $P_S(x)$, and let $d(x, \hat{x}) = r(x - \hat{x})$, where $r(u)$ satisfies

$$\begin{aligned} &(i)\ r(u) = r(-u), \\ &(ii)\ r(u) \geqq 0, \text{ with equality at } u = 0, \\ &(iii)\ r(u) \text{ is continuous at } u = 0. \end{aligned} \tag{12}$$

Then it can be shown (see Ref. 3, Appendix A) that for $0 < \beta \leqq 1/A \int_{-A/2}^{A/2} r(u)\, du$, there exists a unique $\lambda_0(\beta)$ which satisfies

$$\int_{-A/2}^{A/2} r(u) e^{-\lambda_0(\beta) r(u)}\, du = \beta \int_{-A/2}^{A/2} e^{-\lambda_0(\beta) r(u)}\, du. \tag{13}$$

Define the probability density $g_\beta(x)$ on $\mathfrak{X}$ by

$$g_\beta(x) = \left[\int_{-A/2}^{A/2} e^{-\lambda_0(\beta) r(x)}\, dx\right]^{-1} e^{-\lambda_0(\beta) r(x)}, \tag{14a}$$

[note that $\int r(x) g_\beta(x)\, dx = \beta$], and let

$$H_1(\beta) = -\int_{-A/2}^{A/2} g_\beta(x) \log g_\beta(x)\, dx, \tag{14b}$$

be the corresponding "entropy".

For $A = \infty$, equation (13) has a solution in many cases. In particular, when $r(u) = |u|^s$ $(s > 0)$, equation (10) has a solution for $0 < \beta < \infty$. Thus $g_\beta(x)$ and $H_1(\beta)$ are meaningful for $A = \infty$ also.

We now state the lower bound on $R_{eq}(\beta)$ as Theorem 1.

Theorem 1: *For the source defined above, for* $0 < \beta \leqq A^{-1} \int_{-A/2}^{A/2} r(u)\, du$,

$$R_{eq}(\beta) \geqq H_S - H_1(\beta), \tag{15a}$$

where

$$H_S = -\int_{-A/2}^{A/2} P_S(x) \log P_S(x)\, dx \tag{15b}$$

is the entropy of the source density $P_S(x)$, *and* $H_1(\beta)$ *is defined in equations* (13) *and* (14). *Inequality* (15a) *also holds for* $A = \infty$, *when* $r(u) = |u|^s$ $(s > 0)$.

Examples:

(*i*) Say $\mathfrak{X}$ = the reals, and $d(x, \hat{x}) = r(x - \hat{x}) = |x - \hat{x}|^s$, where $s > 0$ is arbitrary. Theorem 1 is applicable with $A = \infty$. Solving equation (13), yields $\lambda_0(\beta) = (s\beta)^{-1}$ and

$$g_\beta(x) = \frac{s^{(s-1)/s}}{2\beta^{1/s}\Gamma\left(\frac{1}{s}\right)} \exp\left[-|x|^s/(s\beta)\right],$$

so that

$$R_{eq}(\beta) \geqq H_S - H_1(\beta), \tag{16a}$$

where

$$H_1(\beta) = \frac{1}{s} \log\left[\frac{2^s e \Gamma^s\left(\frac{1}{s}\right)\beta}{s^{s-1}}\right], \tag{16b}$$

and H_S is given by equation (15b).

(*ii*) Quadratic Distortion: Let $\mathfrak{X}$ = the reals, and $d(x, \hat{x}) = (x - \hat{x})^2$. Then from example (*i*), with $s = 2$,

$$R_{eq}(\beta) \geqq H_S - \tfrac{1}{2} \log 2\pi e\beta. \tag{17a}$$

Further Shannon[4] has given the following upper bound to $R_{eq}(\beta)$:

$$R_{eq}(\beta) \leqq \tfrac{1}{2} \log \frac{\sigma^2}{\beta}, \qquad \beta \leqq \sigma^2, \tag{17b}$$

where $\sigma^2 = \int x^2 P_S(x)\, dx$. Note that when $P_S(x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-x^2/2)$, the upper and lower bounds of inequalities (17) coincide for $\beta \leqq \sigma^2$. [Since $R_{eq}(\beta)$ is non-increasing, $R_{eq}(\beta) = 0$ for $\beta > \sigma^2$.]

Another case of interest is $\mathfrak{X} = [-A/2, A/2] (A < \infty)$, $P_S(x) = A^{-1}$, and $d(x, \hat{x}) = (x - \hat{x})^2$. In this case Theorem 1 (applied for finite A) provides a lower bound on $R_{eq}(\beta)$ which is tighter than that of inequality (17a) and can be evaluated numerically. An upper bound can be found by computing $I[p_0(x, \hat{x})]$, where $p_0(x, \hat{x})$, a joint probability density for X and $\hat{X}$, is defined by the following: The variate X has density $P_S(x)$. The variate $\hat{X} = \alpha(X + Y)$, where the Y is a Gaussian variate, independent of X, with

$$EY = 0 \quad \text{and} \quad EY^2 = \beta A^2/(A^2 - 12\beta),$$

and

$$\alpha = (A^2 - 12\beta)/A^2.$$

Note that $E(X - \hat{X})^2 = \beta$. The information $I[p_0(x, \hat{x})]$ corresponding to $p_0(x, \hat{\mathbf{x}})$ can also be evaluated numerically and is an upper bound to $R_{eq}(\beta)$. Figure 3 is a graph of these bounds on $R_{eq}(\beta)$, and also of $\bar{d}_0$, the solution of $\rho_S R_{eq}(\bar{d}_0) = C$.

(*iii*) Say $\mathfrak{X} = [-A/2, A/2]$. Let $P_S(x) = A^{-1}$ and $d(x, \hat{x}) = r(x - \hat{x})$ where, in addition to satisfying conditions (12), $r(u)$ satisfies

$$r(u) = r(v) \quad \text{if} \quad u \equiv v \pmod{A}. \tag{18}$$

[If, for example, $A = 2\pi$ and $\mathfrak{X}$ represents an angle, then equation (18) must hold.] For $r(u)$ satisfying condition (18), the bound (15a) on $R_{eq}(\beta)$ of Theorem 1 holds with equality, namely, $R_{eq} = H_S - H_1(\beta)$. (Section 3.1)

(*iv*) Threshold Distortion: Let $\mathfrak{X} = [-A/2, A/2]$ and let $d(x, \hat{x})$ be the "threshold" distortion defined by

$$d(x, \hat{x}) = d_\delta(x, \hat{x}) = r_\delta(x - \hat{x}), \tag{19a}$$

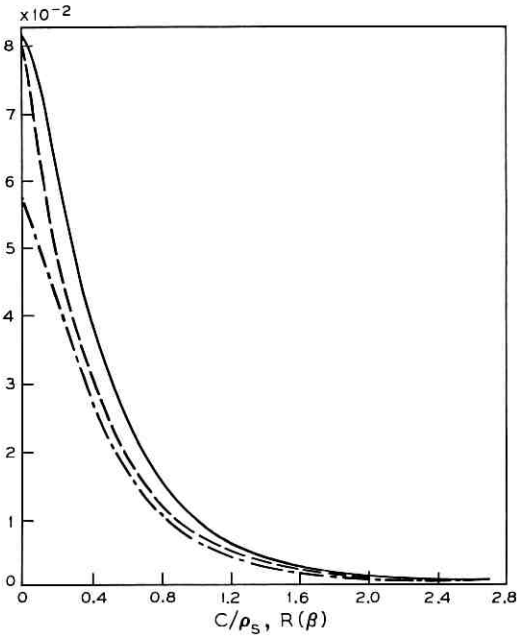

Fig. 3 — Bounds on β/A^2 versus $R_{eq}(\beta)$ or d_0/A^2 versus C/ρ_S. (*i*) ——— upper bound; (*ii*) — — — lower bound (Theorem 1); (*iii*) — - — - — lower bound (17a).

where

$$r_\delta(u) = \begin{cases} 1, & |u| \geq \delta, \\ 0, & |u| < \delta. \end{cases} \tag{19b}$$

In this case, the bound (15) of Theorem 1 is

$$R_{eq}(\beta) \geqq H_S - h(\beta) - \log 2\delta - \beta \log (A/2\delta - 1), \tag{20}$$

where $h(\beta)$ is defined in equation (9b). There is a case where inequality (20) is satisfied with equality, namely $P_S(x) = A^{-1}$ and $A/(2\delta) = K_0 = 1, 2, \cdots$. For this case, we show in Section 3.1 that

$$R_{eq}(\beta) = \begin{cases} \log K_0 - h(\beta) - \beta \log (K_0 - 1), & 0 \leqq \beta \leqq \dfrac{K_0 - 1}{K_0}, \\ 0, & \beta \geqq \dfrac{K_0 - 1}{K_0}. \end{cases} \tag{21}$$

Notice the striking similarity of equations (21) and (9) for the discrete K-ary source. We will have more to say about this later.

When $A/(2\delta)$ is not an integer, we show (in Section 3.1) the right member of (21) is an upper bound to $R_{eq}(\beta)$ with K_0 replaced by $[A/2\delta]^+ = K_+$. Thus with inequality (20),

$$\log \left[\frac{A}{2\delta}\right] - h(\beta) - \beta \log \left[\frac{A}{2\delta} - 1\right] \leqq R_{eq}(\beta) \leqq \log K_+ - h(\beta) - \beta \log (K_+ - 1). \quad (22)$$

A Result for Finite T for the Threshold Distortion is as follows. Let $\mathfrak{X} = [-A/2, A/2]$, $P_S(x) = A^{-1}$, and $d(x, \hat{x}) = d_\delta(x, \hat{x})$, the threshold distortion given in equations (19), as in example (*iv*) above. In the system of Fig. 1, let $\bar{d}^*(T) = Q(T, A, \delta)$, where the dependence on A and δ is indicated explicitly. The results in example (*iv*) [equation (21)] and equation (10) imply that for $A/2\delta = K_0$, $\lim_{T\to\infty} Q(T, A, \delta) = \lim_{T\to\infty} P_e(T, K_0) = \gamma(K_0, \rho_S, C)$. This correspondence between Q and P_e is extended to finite T in the Theorem 2 (proved in Section 3.2.).

Theorem 2: *Let* $K_+ = [A/2\delta]^+$, $K = [A/2\delta]^-$. *For all* T,

$$P_e(T, K_-) \leqq Q(T, A, \delta) \leqq P_e(T, K_+). \quad (23)$$

The quantities P_e *and* Q *are defined, of course, for the same channel and source output rate* ρ_S.

A case of particular interest is $A/2\delta = K_0$, an integer, so that $K_+ = K_- = K_0$ and Theorem 2 yields

$$P_e(T, K_0) = Q(T, A, \delta), \quad \text{all} \quad T. \quad (24)$$

For this case we deduce from equation (24) that (for all T) the optimal encoder for the analog source is a K_0-level "uniform" quantizer with quantization levels $[(2i - K_0 - 1)\delta]_{i=1}^{K_0}$ followed by an optimal "digital" encoder. This is the only known case for which analog-to-digital conversion is known to be optimal for $T < \infty$ for the transmission of analog data.

2.3 *Case Where The Source Has No Statistics*

Suppose that the source output is, as in Section 2.2, a sequence of symbols from the source alphabet $\mathfrak{X}$, which appear at a rate of ρ_S per second. However, in this case, as distinct from above, we assume that there is no known statistical model for the source. Say that, as in Section 2.2, the encoder waits T seconds during which time $n = \rho_S T$ source symbols $\mathbf{x} = (x_1, x_2, \cdots, x_n) \in \mathfrak{X}^n$ have appeared. Again, as above,

the encoder output is $f_E(\mathbf{x}) \in \mathcal{W}_T$, the channel output is $Z \in \mathcal{Z}_T$ and the decoder output is $\hat{\mathbf{X}} = f_D(Z) \in \mathcal{X}^n$. The encoder-decoder pair and the channel statistics induce a probability density for $\hat{\mathbf{X}}$ on $\mathcal{X}^n$ which depends on $\mathbf{x}$ (the source output). Denote this density by $f(\hat{\mathbf{x}} \mid \mathbf{x})$. Assuming, as in Section 2.3, that a non-negative distortion function $d(x, \hat{x})$ on $\mathcal{X} \times \mathcal{X}$ is given, then the average distortion when the source output is $\mathbf{x}$ is

$$\bar{d}(\mathbf{x}) = \int_{\mathcal{X}^n} \left[\frac{1}{n} \sum_{k=1}^{n} d(x_k , \hat{x}_k) \right] f(\hat{\mathbf{x}} \mid \mathbf{x}) \, d\hat{\mathbf{x}}, \tag{25}$$

where $\mathbf{x} = (x_1 , x_2 , \cdots , x_n)$ and $\hat{\mathbf{x}} = (\hat{x}_1 , \hat{x}_2 , \cdots , \hat{x}_n)$. Since we cannot take a meaningful statistical average over $\mathbf{x}$, we adopt as our fidelity criterion, the "guaranteed" distortion

$$\hat{d} = \sup_{\mathbf{x} \in \mathcal{X}^n} \bar{d}(\mathbf{x}). \tag{26}$$

Let $\hat{d}^*(T)$ be the smallest attainable value of $\hat{d}$ for a given delay T (which corresponds to $n = \rho_S T$).

For the special case where $\mathcal{X} = \{0, 1, \cdots , K - 1\}$ and $d(x, \hat{x}) = d_H(x, \hat{x})$ [given by equations (5)] let $\hat{d}^*(T) = \hat{P}_e(T, K)$ where the dependence on K is made explicit. Consider $P_e(T, K)$ (the average error-rate in Section 2.2). Clearly,

$$\hat{P}_e(T, K) \geqq P_e(T, K).$$

The following theorem [taken together with equations (10)] shows that as $T \to \infty$, P_e and $\hat{P}_e$ are asymptotically equal. The proof is in Section 3.4.

Theorem 3: *For the communication system described above with K-ary source alphabet, source output rate* ρ_S, *and channel capacity* C,

$$\lim_{T \to \infty} \hat{P}_e(T, K) = \gamma(K, \rho_S , C), \tag{27}$$

where $\gamma(K, \rho_S , C)$ *is given by inequality* (10b).

A second important special case is $\mathcal{X} = [-A/2, A/2]$, and $d(x, \hat{x}) = d_\delta(x, \hat{x})$, the threshold distortion given by equations (19). In this case let $\hat{d}^*(T) = \hat{Q}(T, \delta, A)$. The quantity $\hat{Q}$ can be related to $\hat{P}_e$, and Theorem 4 (proved in Section 3.3) is analogous to Theorem 2, though somewhat sharper.

Theorem 4: *For* $0 < \delta \leqq A$, *let* $M(\delta)$ *be the integer satisfying*

$$M - 1 \leqq \frac{A}{(2\delta)} < M. \tag{28a}$$

Then for all T,

$$\hat{Q}(T, A, \delta) = \hat{P}_e[T, M(\delta)]. \tag{28b}$$

The quantities $\hat{P}_e$ and $\hat{Q}$ are defined, of course, for the same channel and source output rate ρ_S .

In constrast to Theorem 2, this theorem asserts the equality of corresponding values of $\hat{Q}$ and $\hat{P}_e$ for *all* values of $A/(2\delta)$. Also as in Theorem 2, this theorem implies that the optimal encoder for the source $\mathfrak{X} = [-A/2, A/2]$, with $d = d_\delta$ [with a fidelity criterion as in equation (26)] is a uniform quantizer [with $M(\delta)$ levels] followed by an optimal digital encoder (see part (i) of the proof of Theorem 4).

Theorems 3 and 4 can be combined to obtain the following.

Corollary: For $0 < \delta \leqq A$, *let* $M(\delta)$ *be as in Theorem* 4. *Then*

$$\lim_{T\to\infty} \hat{Q}(T, A, \delta) = \gamma[M(\delta), \rho_S, C], \tag{29}$$

where γ *is given by inequality* (10b).

2.4 *Generalization to Arbitrary Source Alphabets*

In this section we consider the case where the source alphabet $\mathfrak{X}$ is an arbitrary space with an arbitrary metric or metric-like function defined on it. We then give a generalization of Theorem 4. First we give some preliminary definitions.

Let $\mathfrak{X}$ be a set and let $\rho_0(x, \hat{x})$ be real-valued function defined on $\mathfrak{X} \times \mathfrak{X}$ with the properties

$$(i)\ \rho_0(x, \hat{x}) = \rho_0(\hat{x}, x) \tag{30a}$$

$$(ii)\ \rho_0(x, \hat{x}) \geqq 0 \text{ with equality when } x = \hat{x}. \tag{30b}$$

If in addition $\rho_0(x, \hat{x})$ satisfies

$$(iii)\ \rho_0(x, \hat{x}) \leqq \rho_0(x, y) + \rho_0(y, \hat{x}), \tag{30c}$$

then $\rho_0(x, \hat{x})$ is a metric; but we will not require inequality (30c) to hold. For $x \in \mathfrak{X}$ and $\Delta > 0$, let $S_x(\Delta) = \{\hat{x} \in \mathfrak{X} : \rho_0(x, \hat{x}) < \Delta\}$ be the (open) sphere of radius Δ about x.

A set $A \subseteq \mathfrak{X}$ is called a "Δ-covering" (of $\mathfrak{X}$) if $\bigcup_{x\in A} S_x(\Delta)$ contains $\mathfrak{X}$, and A is called a "Δ-packing" (of $\mathfrak{X}$) if $S_x(\Delta) \cap S_{\hat{x}}(\Delta)$ is empty for all $x, \hat{x} \in A, x \neq \hat{x}$. Let $M_C(\Delta)$ be the minimum number of points which can constitute a Δ-covering of $\mathfrak{X}$, and let $M_P(\Delta)$ be the maximum number of points which can constitute a Δ-packing. These quantities are related by the following lemma (proved in Section 3.4).

Lemma 1: Let $\eta = sup_{x,y,z \in \mathfrak{X}} \rho_0(x, y)/[\rho_0(x, z) + \rho_0(z, y)]$. *Then for* $\Delta > 0$,

$$M_C(2\eta\Delta) \leqq M_P(\Delta). \tag{31}$$

In particular, if ρ_0 is a metric, $\eta \leqq 1$. Inequality (31) is of course meaningful only if $\eta < \infty$.

Now consider the communication system discussed in Section 2.3 with an arbitrary source space $\mathfrak{X}$†. Let ρ_0 satisfy expressions (30a) and (30b), and define the "threshold" distortion $d_\delta(x, \hat{x})$ by

$$d_\delta(x, \hat{x}) = \begin{cases} 1, & \rho_0(x, \hat{x}) \geqq \delta, \\ 0, & \rho_0(x, \hat{x}) < \delta. \end{cases} \tag{32}$$

Let $\hat{d}$ be the guaranteed distortion defined by equation (26) with the distortion $d(x, \hat{x}) = d_\delta(x, \hat{x})$ [given by equation (32)]. Finally, let $\hat{G}(T, \delta)$ be the smallest attainable value of $\hat{d}$ for a system with delay T. (The dependence of $\hat{G}$ on δ is made explicit.) Of course $\hat{G}(T, \delta)$ also depends on ρ_S as well as the channel characteristics. The special case treated in Section 2.3 is $\mathfrak{X} = [-A/2, A/2]$, $\rho_0(x, \hat{x}) = |x - \hat{x}|$. In this case $\hat{G}(T, \delta) = \hat{Q}(T, A, \delta)$.

The following is a generalization of Theorem 4 and is proved in Section 3.3.

Theorem 5: Let $M_C(\Delta)$ *and* $M_P(\Delta)$ *be as defined above for the source alphabet* $\mathfrak{X}$ [*with a* $\rho_0(x, \hat{x})$]. *Then* $G(T, \delta)$ *satisfies*

$$\hat{P}_e[T, M_P(\delta)] \leqq \hat{G}(T, \delta) \leqq \hat{P}_e[T, M_C(\delta)], \tag{33}$$

where $\hat{P}_e$ *is defined in Section* 2.3. *Note that* $\hat{P}_e$ *and* $\hat{G}$ *are defined for the same channel and source output rate* ρ_S.

Theorem 5 reduces to Theorem 4 on noting that for $\mathfrak{X} = [-A/2, A/2]$ and $\rho_0(x, \hat{x}) = |x - \hat{x}|$,

$$M_P(\delta) = M_C(\delta) = M(\delta), \tag{34}$$

where $M(\delta)$ is defined by inequality (28a). Let us remark that although $M_P \equiv M_C$, the maximum δ-packing is not in general identical to their minimum δ-covering. For example, when $\delta = A/4$, $M(\delta) = 3$, and the maximum δ-packing is unique, namely

$$\left\{-\frac{A}{2}, 0, \frac{A}{2}\right\},$$

† To be precise, we must assume that the space $\mathfrak{X}$ and the encoder and decoder functions are measurable.

which is not a δ-covering. There are many δ-coverings, for example

$$\left\{-\frac{A}{3}, 0, \frac{A}{3}\right\}.$$

2.5 *Some Applications*

2.5.1 *Rate at Which $Q(T, A, \delta)$ Approaches its Limit*

Consider again the source with $\mathfrak{X} = [-A/2, A/2]$, $P_S(x) = A^{-1}$, and distortion $d(x, \hat{x}) = d_\delta(x, \hat{x})$ [defined by expressions (19)]. Suppose further that $A/(2\delta) = K_0$, an integer and that the channel capacity $C = \rho_S \log K_0$. In this case $\gamma(K_0, \rho_S, C) = 0$ [see expressions (9) and (10)], so that from expressions (24) and (10a)

$$\lim_{T\to\infty} Q(T, A, \delta) = 0.$$

We will now obtain a lower bound on the rate at which this limit is approached. From the first inequality in inequality (11), using $n = \rho_S T$,

$$P_e(T, K) \geqq \frac{1}{\rho_S T} \bar{\lambda}^*(T, e^{CT}). \tag{35}$$

For those channels for which expression (4) holds, the right member of inequality of (35) $\sim (2\rho_S T)^{-1}$. Combining expressions (24) and (35) we have that

$$Q(T, A, \delta) \geqq \frac{1}{2\rho_S T}[1 + \xi(T)], \tag{36}$$

where $\xi(T) \to 0$ as $T \to \infty$. Thus for the class of channels for which expression (4) holds and these parameter values, $Q(T, A, \delta)$, approaches its limit no faster than T^{-1}. Determination of the similar bounds on the rate of approach of Q to its limit for other parameters is an open question.

2.5.2 *The sth-Mean Distortion*

Consider the case where $\mathfrak{X} = [-A/2, A/2]$, and the distortion $d(x, x) = |x - \hat{x}|^s$ $(s > 0)$. When $P_S(x) = A^{-1}$, let the smallest attainable average distortion $\bar{d}^*(T) \triangleq \bar{\epsilon}^s(T)$. For the case of no source statistics (as in Section 2.3), let the smallest attainable guaranteed distortion $\hat{d}^*(T) \triangleq \hat{\epsilon}^s(T)$. We establish some properties of $\bar{\epsilon}^s$ and $\hat{\epsilon}^s$ below.

For any random variable Y (such that $|Y| \leqq A$), and any δ_1, $\delta_2 (0 \leqq \delta_1, \delta_2 \leqq A)$,

$$\delta_1^s \Pr \{| Y | \geqq \delta_1\} \leqq E | Y |^s \leqq \delta_2^s \Pr \{| Y | < \delta_2\} + A^s \Pr \{| Y | \geqq \delta_2\}. \quad (37)$$

It follows from inequality (37) that for arbitrary δ_1, δ_2 $(0 \leqq \delta_1, \delta_2 \leqq A)$,

$$\delta_1^s Q(T, A, \delta_1) \leqq \bar{\epsilon}^s(T) \leqq \delta_2^s[1 - Q(T, A, \delta_2)] + A^s Q(T, A, \delta_2), \quad (38a)$$

and

$$\delta_1^s \hat{Q}(T, A, \delta_1) \leqq \hat{\epsilon}^s(T) \leqq \delta_2^s[1 - \hat{Q}(T, A, \delta_2)] + A^s \hat{Q}(T, A, \delta_2), \quad (38b)$$

where Q and $\hat{Q}$ are defined in Sections 2.2 and 2.3 respectively. Applications of Theorems 2 and 4 (and Q, $\hat{Q} \geqq 0$) yields

$$\delta_1^s P_e(T, K_-) \leqq \bar{\epsilon}^s(T) \leqq \delta_2^s + A^s P_e(T, K_+), \quad (39a)$$

and

$$\delta_1^s \hat{P}_e[T, M(\delta_1)] \leqq \hat{\epsilon}^s(T) \leqq \delta_2^s + A^s \hat{P}_e[T, M(\delta_2)], \quad (39b)$$

where $K_+ = [A/2\delta_2]^+$, $K_- = [A/2\delta_1]^-$, and $M(\delta)$ is defined by inequality (28a). Thus $\bar{\epsilon}^s$ and $\hat{\epsilon}^s$ too are related to the digital error rates P_e and $\hat{P}_e$. Of course, δ_1 and δ_2 may be chosen to yield the tightest bounds.

Examples

(*i*) Since we know the asymptotic value of P_e and $\hat{P}_e$ as $T \to \infty$, we can apply inequalities (39) to obtain estimates of the limiting values $\bar{\epsilon}_0^s = \lim_{T\to\infty} \bar{\epsilon}^s(T)$ and $\hat{\epsilon}_0^s = \lim_{T\to\infty} \hat{\epsilon}^s(T)$. For example, when the channel capacity C is large, setting $A/2\delta_1 = \exp[(C/\rho_S)(1 + \Delta_1)]$ and $A/2\delta_2 = \exp[(C/\rho_S)(1 - \Delta_1)](\Delta_1, \Delta_2 > 0)$, yields, after some computation,

$$\bar{\epsilon}_0^s = \exp\left\{-\frac{sC}{\rho_S}[1 + \xi_1(C)]\right\}, \quad (40a)$$

$$\hat{\epsilon}_0^s = \exp\left\{-\frac{sC}{\rho_S}[1 + \xi_2(C)]\right\}, \quad (40b)$$

where $\xi_1, \xi_2 \to 0$ as $C \to \infty$. Thus for large C, $\bar{\epsilon}_0^s$ and $\hat{\epsilon}_1^s$ decay roughly exponentially in C.

Let us remark that parts of inequalities (40) are obtainable by other means. Specifically, $\bar{\epsilon}_0^s \geqq K_1(s) \exp[-sC/\rho_S]$ follows from inequality (16). Further, $\epsilon_0^s \leqq \exp[-(sC/\rho_S)(1 + \xi_1)]$ and $\hat{\epsilon}_0^s \leqq \exp[-(sC/\rho_S) \cdot (1 + \xi_2)]$ can be deduced from the work of Panter and Dite on quantization.[5] Finally the bound $\hat{\epsilon}_0^s \geqq \exp[-(sC/\rho_S)(1 + \xi_2)]$ is new.

(*ii*) In this example, we apply the first inequality of (39a) to show the possible gains (with the sth mean criterion) obtainable by using coding in a particular (though quite typical) case.

Suppose that the channel is the additive white Gaussian noise channel with average power P_0, one-sided spectral density N_0, with no bandwidth constraint.[6] To begin with, suppose $T = 1/\rho_S$, so that $n = 1$ and there is no "coding", that is, each T-second channel input depends on exactly one source output. When the source is the K-ary digital source (with equi-distributed symbols), it is known that the minimum attainable error rate is lower bounded by†

$$P_e(T, K) \geqq \tfrac{1}{2}\Phi\left\{\left[\frac{KP_0T}{(K-2)(2N_0)}\right]^{\frac{1}{2}}\right\}, \tag{41}$$

where

$$\Phi(\alpha) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\alpha} e^{-u^2/2}\, du,$$

is the cumulative error function.

We now apply the lower bound of inequality (39a) together with inequality (41) to obtain a lower bound on $\bar{\epsilon}^s(T)$ when the channel signal power P_0 made large, while $T = 1/\rho_S$ is held fixed. Setting $\delta_1 = P_0^{-1}$, we obtain from inequalities (39a) and (41) and $\Phi(\alpha) \sim (2\pi\alpha^2)^{-1} \exp(-\alpha^2/2)$ (as $\alpha \to \infty$), that (with $T = \rho_S^{-1}$ held fixed)

$$\bar{\epsilon}^s(T) = \bar{\epsilon}^s\left(\frac{1}{\rho_S}\right) \geqq \exp\left\{-\frac{P_0}{2N_0\rho_S}\left[1 + \xi_3(P_0)\right]\right\} \tag{42}$$

where $\xi_3(P_0) \to 0$ as $P_0 \to \infty$.

Now suppose that for a given channel (and a given P_0) we allow T to become large. In other words, we permit "source coding" in blocks of length $n = \rho_S T$. Since the channel capacity $C = P_0/N_0$, we have from equation (40a) that

$$\lim_{T\to\infty} \bar{\epsilon}^s(T) = \bar{\epsilon}_0^s = \exp\left\{-\frac{sP_0}{2N_0\rho_S}\left[1 + \xi_4(P_0)\right]\right\}, \tag{43}$$

where $\xi_4(P_0) \to 0$ as $P_0 \to \infty$.

Now let $\theta > 0$ be arbitrary, and let P_1 be sufficiently large so that for $P_0 \geqq P_1$,

$$|\, \xi_3(P_0)\,|, \qquad |\, \xi_4(P_0)\,| < \theta.$$

Then from inequality (42), with $P_0 \geqq P_1$, the best attainable mean sth

† This bound follows from Ref. 1 [equation (82)] when the signal energy nP in that reference is replaced by P_0T our signal energy, and M is replaced by K.

error with no coding is bounded by

$$\bar{\epsilon}^{s}\left(\frac{1}{\rho_S}\right) \geqq \exp\left[-\frac{P_0}{2N_0\rho_S}(1+\theta)\right]. \tag{44}$$

The best attainable sth error with infinite delay T is from equation (43) with $P_0 \geqq P_1$, bounded by

$$\bar{\epsilon}_0^{s} \leqq \exp\left[-\frac{sP_0}{N_0\rho_S}(1-\theta)\right]. \tag{45}$$

We conclude that coding with large delay offers a saving of at least a factor of $(2s)$ in power P_0 or rate ρ_S (when $P_0 \geqq P_1$). This of course is interesting when $s > \frac{1}{2}$. Similar results for $s = 2$ have been derived by Ziv and Zakai.[7] This result can be generalized to arbitrary n (here we studied $n = 1$) and arbitrary channels simply by using appropriate bounds on $P_e(T, K)$.

III. PROOFS OF THEOREMS

3.1 *Proof of Theorem* 1 *and Related Examples*

3.1.1 *Proof of Theorem* 1

Shannon [Ref., 2, pp. 155–156] has shown that for a difference distortion measure $d(x, \hat{x}) = r(x - \hat{x})$, that

$$R_{eq}(\beta) \geqq H_S - \Phi(\beta), \tag{46}$$

where H_S is given by equation (15b) and $\Phi(\beta)$ is the maximum attainable entropy $H\{f(x)\}$ for a probability density $f(x)$ which satisfies

$$\int_{-\infty}^{\infty} r(x)f(x)\,dx \leqq \beta. \tag{47}$$

The entropy $H\{f(x)\}$ is defined by

$$H\{f(x)\} = -\int_{-\infty}^{\infty} f(x)\log f(x)\,dx. \tag{48}$$

A trivial modification of Shannon's argument shows that when $\mathfrak{X} = [-A/2, A/2]$, inequality (46) remains valid if $f(x)$ is further restricted to satisfy

$$f(x) = 0, \qquad |x| > \frac{A}{2}. \tag{49}$$

Now the density $g_\beta(x)$ [defined by expressions (13) and (14a)] satisfies

conditions (47) and (49) and has entropy $H_1(\beta)$ [defined by equation (14)]. We prove Theorem 1 by showing that if the density $f(x)$ satisfies conditions (47) and (49), then $H\{f(x)\} \leqq H_1(\beta)$.

Let us write $g_\beta(x) = Be^{-\lambda r(x)}$ where $\lambda = \lambda_0(\beta)$ and where

$$B = \left[\int_{-A/2}^{A/2} e^{-\lambda_0(\beta) r(x)}\, dx\right]^{-1}.$$

Then

$$\begin{aligned} H_1(\beta) &= -\int_{-A/2}^{A/2} g_\beta(x) \log g_\beta(x)\, dx \\ &= -\log B + \lambda \int_{-A/2}^{A/2} r(x) g_\beta(x)\, dx = -\log B + \lambda\beta. \end{aligned}$$

Since $f(x)$ satisfies condition (47),

$$\begin{aligned} H_1(\beta) &\geqq -\log B + \lambda \int_{-A/2}^{A/2} r(x) f(x)\, dx \\ &= -\int f(x) \log Be^{-\lambda r(x)}\, dx = -\int f(x) \log g_\beta(x)\, dx. \end{aligned}$$

Thus

$$\begin{aligned} H\{f(x)\} - H_1(\beta) &\leqq -\int_{-A/2}^{A/2} f(x) \log f(x)\, dx + \int f(x) \log g_\beta(x)\, dx \\ &= \int_{-A/2}^{A/2} f(x) \log \frac{g_\beta(x)}{f(x)}\, dx \leqq \int_{-A/2}^{A/2} f(x)\left[\frac{g_\beta(x)}{f(x)} - 1\right] dx = 1 - 1 = 0, \end{aligned}$$

where the second inequality follows from $\log u \leqq u - 1$. Theorem 1 follows.

Note that Theorem 1 will hold for $A = \infty$ as long as we can find $g_\beta(x)$. Examination of the derivation which establishes the existence of $g_\beta(x)$ (Ref. 3, Appendix A) shows that Theorem 1 is valid in particular for $A = \infty$ and $r(x) = |x|^s$, $s > 0$.

3.1.2 *Determination of* $R_{eq}(\beta)$ *in Example (iii)*

For $\mathfrak{X} = [-A/2, A/2]$, $P_S(x) = A^{-1}$, and $d(x, \hat{x}) = r(x - \hat{x})$, where $r(u)$ satisfies conditions (12); $H_S = \log A$. Theorem 1 implies

$$R_{eq}(B) \geqq \log A - H_1(\beta). \tag{50}$$

We now show that if, in addition, $r(u)$ satisfies equation (18), then inequality (50) is satisfied with equality. Let X and $\hat{X}$ be random varia-

bles such that the density for X is $P_S(x) = A^{-1}(|\ x\ | \leqq A/2)$, and $\hat{X} = X + Y$ where the random variable Y is independent of X and has density $g_\beta(y) = Be^{-\lambda_0 r y}$ [defined by equations (13) and (14a)]. The information of $p(x, \hat{x})$, the joint density for X, $\hat{X}$, is

$$I\{p(x, \hat{x})\} = H\{p_2(\hat{x})\} - \int_{-A/2}^{A/2} P_S(x) H\{p(\hat{x} \mid x)\}\, dx,$$

where $p_2(\hat{x})$ is the density for $\hat{X}$, $p(\hat{x} \mid x)$ is the conditional density for $\hat{X}$ given that $X = x$, and $H\{\ \}$ is the entropy defined in equation (48).

Now $p(x, \hat{x}) = P_S(x)p(\hat{x} \mid x) = A^{-1}Be^{-\lambda_0 r(\hat{x}-x)}$, so that

$$p_2(\hat{x}) = A^{-1}B \int_{-A/2}^{A/2} e^{-\lambda_0 r(\hat{x}-x)}\, dx = A^{-1}B \int_{\hat{x}-A/2}^{\hat{x}+A/2} e^{-\lambda_0 r(u)}\, du.$$

when $\hat{x} \geqq 0$ this becomes, letting $v = u - A$ and using equation (18)

$$\begin{aligned} p_2(\hat{x}) &= A^{-1}B \int_{\hat{x}-A/2}^{A/2} e^{-\lambda_0 r(u)}\, du + A^{-1}B \int_{A/2}^{\hat{x}+A/2} e^{-\lambda_0 r(u)}\, du \\ &= A^{-1}B \int_{\hat{x}-A/2}^{A/2} e^{-\lambda_0 r(u)}\, du + A^{-1}B \int_{-A/2}^{\hat{x}-A/2} e^{-\lambda_0 r(v)}\, dv. \end{aligned}$$

Hence, since $\int g_\beta(x) = 1$,

$$p_2(\hat{x}) = A^{-1}B \int_{-A/2}^{A/2} e^{-\lambda_0 r(u)}\, du = A^{-1}.$$

For $\hat{x} < 0$, a similar proof yields $p_2(\hat{x}) = A^{-1}$. Thus $H\{p_2(\hat{x})\} = \log A$. Further $p(\hat{x} \mid x) = g_\beta(\hat{x} - x)$, and a similar use of equation (18) yields $H\{p(\hat{x} \mid x)\} = H_1(\beta)$, independent of x. Thus we conclude that $I\{p(x, \hat{x})\} = \log A - H_1(\beta)$. Since $p(x, \hat{x}) \in \mathfrak{M}(\beta)$, this and inequality (50) imply $R_{eq}(\beta) = \log A - H_1(\beta)$.

3.1.3 *Proof for Example (iv)*

We first verify equation (21) for the case $A/(2\delta) = K_0$, an integer. That $R_{eq}(\beta)$ is greater than or equal to the right member of equation (21) follows from inequality (20) (since $H_S = \log A$) and from $R_{eq}(\beta) \geqq 0$. To show that $R_{eq}(\beta)$ is less than or equal to the right member of expression (21) we produce a density $p_0(x, \hat{x})$ for which $I\{p_0(x, \hat{x})\}$ equals the right member of equation (21). But first we digress to define "entropy" for a discrete random variable.

Consider a discrete probability density $f(x) = \sum_i a_i\, \delta(x - x_i)$. Then the "discrete entropy of $f(x)$ is defined by

$$H_D\{f(x)\} = -\sum_i a_i \log a_i\ . \tag{51}$$

Now, say that $p(x, \hat{x})$ is the probability density for two random variables X, $\hat{X}$, such that $\hat{X}$ takes values at a countable number of points. Then the marginal density for $\hat{X}$, denoted $p_2(\hat{x})$ and the conditional density for $\hat{X}$ given $X = x$ [denoted $p(\hat{x} \mid x)$] are discrete densities. It is easy to show that the information can be written

$$I\{p(x, \hat{x})\} = H_D\{p_2(\hat{x})\} - \int p_1(x) H_D\{p_2(\hat{x} \mid x)\}\, dx, \tag{52}$$

where $p_1(x)$ is the marginal density for X.

Return now to Example (*iv*). Let $0 \leqq \beta \leqq (K_0 - 1)/K_0$, and let $p_0(x, \hat{x})$ be the density for X, $\hat{X}$, where X has density $P_S(x) = A^{-1}$ and $\hat{X}$ has conditional density $p_0(\hat{x} \mid x)$ given as follows. Partition the interval $[-A/2, A/2]$ into K_0 subintervals $\{I_i\}_0^{K_0-1}$ of width 2δ. Let x_i be the midpoint of I_i ($i = 0, 1, 2, \cdots, K_0 - 1$). Then for $x \in I_i$

$$p_0(\hat{x} \mid x) = (1 - \beta)\, \delta(\hat{x} - x_i) + \frac{\beta}{(K_0 - 1)} \sum_{j \neq i} \delta(\hat{x} - x_j).$$

In other words, $\hat{X}$ is an imperfectly quantized version of X. With probability $(1 - \beta)$, $\hat{X}$ is the midpoint of the subinterval in which X lies, and with probability β, $\hat{X}$ is uniformly distributed among the remaining $(K_0 - 1)$ midpoints. Note that $P_S(x)$ and $p_0(\hat{x} \mid x)$ together determine $p_0(x, \hat{x})$, and that $p_0(x, \hat{x}) \in \mathfrak{M}(\beta)$.

Further, by symmetry, $\hat{X}$ is uniformly distributed on the K_0 midpoints, so that

$$H_D\{p_{02}(\hat{x})\} = \log K_0 ,$$

where $p_{02}(\hat{x})$ is the marginal density for $\hat{X}$ [corresponding to $p_0(x, \hat{x})$]. Also

$$H_D\{p_0(\hat{x} \mid x)\} = h(\beta) + \beta \log (K_0 - 1),$$

independent of x. Thus equation (52) yields

$$I\{p_0(x, \hat{x})\} = \log K_0 - h(\beta) - \beta \log (K_0 - 1),$$

the right member of expression (21). This establishes equation (21) for $0 \leqq \beta \leqq (K_0 - 1)/K_0$. Since $R_{eq}[(K_0 - 1)/K_0] = 0$ and $R_{eq}(\beta)$ is nonincreasing, we have $R_{eq}(\beta) = 0$ for $\beta \geqq (K_0 - 1)/K_0$, establishing expression (21).

It remains to verify the upper bound of expressions (22). But this follows immediately on noting that for fixed A and β, $R_{eq}(\beta)$ is a decreasing function of δ. Thus decreasing δ to $\delta' = A/2[A/2\delta]^+$ results in an

increase in $R_{eq}(\beta)$. Since $A/(2\delta')$ is an integer, we can apply expression (21) to obtain the upper bound of expression (22).

3.2 *Proof of Theorem 2*

Theorem 2 relates the attainable distortions for a digital source and an analog source when connected to a given channel. The proof is in two parts [corresponding to the two inequalities in expression (23)], the second of which uses a bounding technique introduced by Ziv and Zakai.[7]

In part (*i*) we are given an encoder and decoder for the digital source (with appropriate parameters), which when connected to the channel as in Fig. 1 results in an average Hamming distortion $\bar{d} = \bar{d}_H$. We show how to quantize the outputs of the analog source (with appropriate parameters) to essentially simulate the digital source. When this quantizer is connected to the digital encoder, we show that we attain an average distortion for the analog source $\bar{d}_\delta \leqq \bar{d}_H$. This leads us directly to the second inequality of expression (23).

In part (*ii*) we establish the first inequality of expression (23) in an essentially dual way. We begin by assuming the existence of an analog encoder and decoder. We then show how to modulate the outputs of the digital source to virtually simulate the analog source. Unfortunately, this is not as easy as the quantization in part (*i*), and we have to make use of an "averaging" argument in the course of the proof.

(*i*) Let us denote by $\mathcal{S}_a$, the analog source whose output is a sequence $X_1, X_2, \cdots$ of independent random variables, each uniformly distributed on the source space $\mathcal{X}_a = [-A/2, A/2]$. The random variables appear at a rate of ρ_S per second. For this source we use the distortion $d(x, \hat{x}) = d_\delta(x, \hat{x})$ defined by equations (19). Assume first that $A/(2\delta) = K_0$ an integer, and consider the following (uniform) quantizer. Partition the interval $[-A/2, A/2]$ into K_0 subintervals $\{I_i\}_0^{K_0-1}$ of width (2δ) where

$$I_i = (e_i, e_{i+1}], \qquad i = 0, 1, \cdots, K_0 - 1, \tag{53a}$$

and

$$e_i = (2\delta)\left[\left(i - \frac{K_0}{2}\right)\right], \qquad i = 0, 1, \cdots, K_0. \tag{53b}$$

To be precise, the first interval I_0 should be closed on the left. The quantizer q is defined by

$$q(x) = i, \quad \text{if} \quad x \in I_i \qquad \left(-\frac{A}{2} \leqq x \leqq \frac{A}{2}\right). \tag{54}$$

Let us now consider the digital source $\mathcal{S}_d$ whose output is a sequence $S_1, S_2, \cdots$ of independent discrete random variables, each uniformly distributed on the K_0-ary set $\mathfrak{X}_d = \{0, 1, \cdots, K_0 - 1\}$. These random variables also appear at ρ_S per second. (Note that we use S_k instead of X_k as in Section II to distinguish the outputs of $\mathcal{S}_d$ from those of $\mathcal{S}_a$.) Say that the distortion $d = d_H$ as defined in equation (5).

Suppose that $\mathcal{S}_d$ can be connected with delay T to a channel as in Fig. 1 with (digital) encoder $f_E^{(d)}$ and decoder $f_D^{(d)}$, and average distortion $\bar{d}_H$. We now show how to connect the "analog" source $\mathcal{S}_a$ to the channel [with the help of $f_E^{(d)}$ and $f_D^{(d)}$] to attain an average distortion $\bar{d}_\delta \leqq \bar{d}_H$. Consider the system in Fig. 4. In T seconds the output of the analog source is an n-vector $(n = \rho_S T)\mathbf{X} = (X_1, \cdots, X_n)$. The "quantizer" output is the n-vector $\mathbf{S} = (S_1, S_2, \cdots, S_n)$, where $S_k = q(X_k)$ $(k = 1, 2, \cdots, n)$. Note that the S_k are independent and uniformly distributed on $\{0, 1, \cdots, K_0 - 1\}$, as are the outputs of the digital source $\mathcal{S}_d$. The digital encoder and decoder $f_E^{(d)}$ and $f_D^{(d)}$ are as given above, and the output of the latter is the K_0-ary vector $\hat{\mathbf{S}} = (\hat{S}_1, \cdots, \hat{S}_n)$. Thus

$$Ed^{(n)}(\mathbf{S}, \hat{\mathbf{S}}) = \bar{d}_H .$$

The "converter" output is the n-vector $\hat{\mathbf{X}} = (\hat{X}_1, \hat{X}_2, \cdots, \hat{X}_n)$ where

$$\hat{X}_k = (2\hat{S}_k - K_0 + 1)\delta.$$

In other words if $\hat{S}_k = i$, then $\hat{X}_k$ is the midpoint $(e_i + e_{i+1})/2$ of the ith subinterval. Disregarding the case when X_k is equal to one of the endpoints e_i of the subintervals, (an event with zero probability), it is clear that $|X_k - \hat{X}_k| \geqq \delta$ if and only if $S_k \neq \hat{S}_k$ $(k = 1, 2, \cdots, n)$. Thus

$$\bar{d}_\delta = Ed_\delta^{(n)}(\mathbf{X}, \hat{\mathbf{X}}) = Ed_H^{(n)}(\mathbf{S}, \hat{\mathbf{S}}) = \bar{d}_H .$$

It follows that

$$Q(T, A, \delta) \leqq P_e\left(T, \frac{A}{2\delta}\right), \tag{55}$$

when $A/(2\delta)$ is an integer. The second inequality of expression (23)

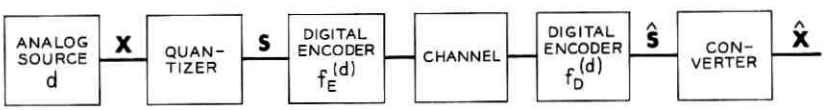

Fig. 4 — An analog communication scheme.

follows on noting that $Q(T, A, \delta)$ is a nonincreasing function of δ. Thus decreasing δ to $\delta' = A/2K_+$ does not result in a decrease in $Q(T, A, \delta)$. Since $A/(2\delta')$ is an integer, we can apply inequality (55) to obtain the second inequality of expression (23). This completes part (i).

(ii) Let us suppose that the analog source $\mathcal{S}_a$ defined in part (i) is connected with delay T to a channel as in Fig. 1. The T-second source output is the n-vector $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ and the decoder output is the n-vector $\hat{\mathbf{X}} = (X_1, X_2, \cdots, X_n)$. Say we attain an average distortion

$$\bar{d}_\delta = Ed_\delta^{(n)}(\mathbf{X}, \hat{\mathbf{X}}).$$

Letting $E[d_\delta^{(n)}(\mathbf{X}, \hat{\mathbf{X}}) \mid \mathbf{X} = \mathbf{x}]$ be the conditional expectation of $d_\delta^{(n)}(\mathbf{X}, \hat{\mathbf{X}})$ given $\mathbf{X} = \mathbf{x}$, we can write

$$\bar{d}_\delta = \int_{[-A/2, A/2]^n} E[d_\delta^{(n)} \mid \mathbf{X} = \mathbf{x}] \frac{1}{A^n} d\mathbf{x}. \tag{56}$$

Suppose that $(A/2\delta) = K_0$, an integer. Let us partition the interval $[-A/2, A/2]$ into K_0 subintervals of width 2δ as in equations (53). Let $\mathcal{E}$ be the set of left end-points of these subintervals, that is,

$$\mathcal{E} = \{e_i\}_{i=1}^{K_0 - 1}. \tag{57}$$

Now consider the n-cube $[-A/2, A/2]^n$. Note that the random n-vector $\mathbf{X}$ is uniformly distributed on this cube. The partition of the interval $[-A/2, A/2]$ defines a partition of the n-cube into K_0^n subcubes, each the product of n subintervals. Let the members of $\mathcal{E}^n$ be denoted by the n-vectors ξ_j, $j = 1, \cdots, K_0^n$, and let C_j be the corresponding subcube. (That is, C_j is the product of the subintervals whose left end-points are the coordinates of ξ_j.) Then clearly,

$$\left[-\frac{A}{2}, \frac{A}{2}\right]^n = \sum_{j=1}^{K_0^n} C_j,$$

where $\sum$ denotes disjoint union. Thus we can rewrite equation (56) as

$$\begin{aligned} \bar{d}_\delta &= \sum_{j=1}^{K_0^n} \int_{C_j} \frac{1}{A^n} E[d_\delta^{(n)} \mid \mathbf{X} = \mathbf{x}] \, d\mathbf{x} \\ &= \sum_{j=1}^{K_0^n} \frac{1}{K_0^n} \int_{[0, 2\delta]^n} \frac{1}{(2\delta)^n} E[d_\delta^{(n)} \mid \mathbf{X} = \xi_j + \boldsymbol{\alpha}] \, d\boldsymbol{\alpha}, \end{aligned} \tag{58}$$

where the second equality follows from the change of variable of integration to $\boldsymbol{\alpha} = \mathbf{x} - \xi_j$, and the fact that $A = 2\delta K_0$.

Some insight into what we have done may be gained by considering

the special case where $K_0 = 2$ and $n = 2$. In this case the n-cube $[-A/2, A/2]^n$ is a square, and there are $K_0^n = 2^2 = 4$ members of $\mathcal{E}^n$ denoted ξ_1, ξ_2, ξ_3, and ξ_4. (See Fig. 5.) The subcubes are C_1, C_2, C_3, and C_4 as indicated.

Let us consider now the digital source $\mathcal{S}_d$ defined in part (*i*) whose output is the sequence $S_1, S_2, \cdots$. We would like to transmit the outputs of $\mathcal{S}_d$ through a channel (as in Fig. 1) with delay T, so that the source output must be an n-vector $(n = \rho_S T)\mathbf{S}$, and the decoder output an n-vector $\hat{\mathbf{S}}$. The fidelity criterion is

$$\bar{d}_H = Ed_H^{(n)}(\mathbf{S}, \hat{\mathbf{S}}).$$

Now suppose that we are given an encoder-decoder, $f_E^{(a)}$, $f_D^{(a)}$, for the analog source $\mathcal{S}_a$ [for which $A/(2\delta) = K_0$], connected with delay T, to a given channel. Say this encoder-decoder attains an average distortion $\bar{d}_\delta$. We show that there exists an encoder-decoder for the K_0-ary digital source $\mathcal{S}_d$, connected with delay T, to the same channel such that the average distortion $\bar{d}_H \leqq \bar{d}_\delta$. From this we deduce immediately that for $A/(2\delta) = K_0$,

$$P_s(T, K_0) \leqq Q(T, A, \delta). \tag{59}$$

The digital encoder is given schematically in Fig. 6a. The analog encoder which we are given is $f_E^{(a)}(\mathbf{x})$, $\mathbf{x} \in [-A/2, A/2]^n$, and is realized in the right box of Fig. 6a. The function of the "modulator" is to assign to each n-vector $\hat{\mathbf{s}} \in \{0, 1, \cdots, K_0 - 1\}^n$, a member of $[-A/2, A/2]^n$. This is done as follows. Let $\mathcal{E}$ be the set defined by equation (57). For $s \in \{0, 1, 2, \cdots, K_0 - 1\}$, let

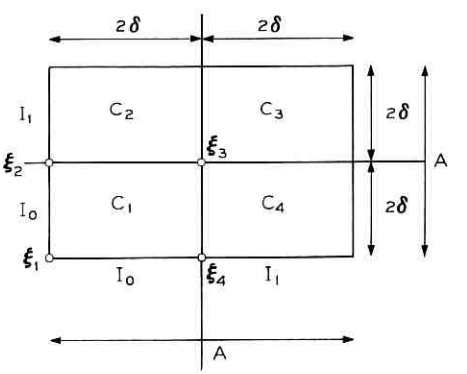

Fig. 5 — A digital encoder.

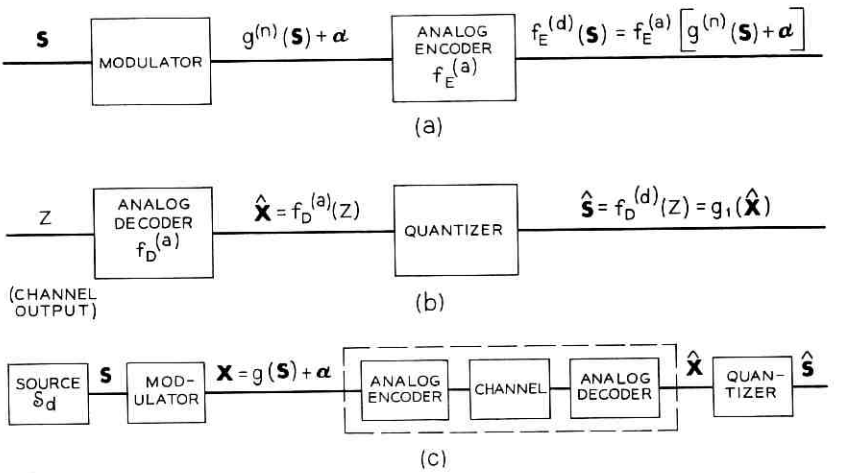

Fig. 6 — (a) Digital to analog encoder. (b) Analog to digital decoder. (c) Digital communication scheme.

$$g(s) = (2\delta)\left[s - \frac{K_0}{2}\right]$$

be the sth member of $\mathcal{E}$. For $\mathbf{s} = (s_1, s_2, \cdots, s_n) \in \{0, 1, \cdots, K_0 - 1\}^n$, let

$$g^{(n)}(\mathbf{s}) = [g(s_1), g(s_2), \cdots, g(s_n)].$$

When the input to the modulator is $\mathbf{s}$, its output is

$$\alpha + g^{(n)}(\mathbf{s}),$$

where $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n) \in [0, 2\delta]^n$ is a fixed vector. Thus the digital encoder is

$$f_E^{(d)}(\mathbf{s}) = f_E^{(a)}[\alpha + g^{(n)}(\mathbf{s})].$$

The digital decoder is given schematically in Fig. 6b. The left box is the analog decoder $f_D^{(a)}$ which we are given. Its output $\hat{\mathbf{x}}$ is a real n-vector. The right box is a quantizer. When its input is $\hat{\mathbf{x}} = (\hat{x}_1, \cdots, \hat{x}_n)$, its output is $g_1(\hat{\mathbf{x}}) = \hat{\mathbf{s}} = (\hat{s}_1, \cdots, \hat{s}_n)$, where $\hat{s}_k (k = 1, 2, \cdots, n)$ is a member of $\{0, 1, \cdots, K_0 - 1\}$ which minimizes $| g_k(\hat{s}_k) + \alpha_k - \hat{x}_k |$.

When the digital source $\mathcal{S}_d$ is connected to the channel with this encoder-decoder pair, the result is schematized in Fig. 6c. (Upper case X's and S's are used to signify random variables.) The portion of the system in the dotted lines is precisely the analog encoder-channel-

decoder which would produce an average distortion $\bar{d}_\delta$ given by equation (56), *if* the analog input, $\mathbf{X}$, where uniformly distributed on the n-cube $[-A/2, A/2]^n$. But this is not the case here. In fact, $\mathbf{X}$ takes only one of K_0^n possible values. However, the quantity $E[d_\delta^{(n)}(\mathbf{X}, \hat{\mathbf{X}}) \mid \mathbf{X} = \mathbf{x}]$ is exactly the same in the system of Fig. 6c as in equation (56), for $\mathbf{x} = g^{(n)}(\mathbf{s}) + \boldsymbol{\alpha}(\mathbf{s} \in \{0, \cdots, K_0 - 1\}^n)$.

Let us write an expression for the average distortion $\bar{d}_H$ for the digital source. Note that $S_k \neq \hat{S}_k$, only if $|\hat{X}_k - [g(S_k) + \alpha_k]| \geqq \delta$. Thus

$$d_H^{(n)}(\mathbf{S}, \hat{\mathbf{S}}) \leqq d_\delta^{(n)}[g^{(n)}(\mathbf{S}) + \boldsymbol{\alpha}, \hat{\mathbf{X}}],$$

and

$$\begin{aligned} \bar{d}_H &= E d_H(\mathbf{S}, \hat{\mathbf{S}}) \\ &\leqq \sum_{\mathbf{s}} \frac{1}{K_0^n} E[d_\delta^{(n)}(\mathbf{X}, \hat{\mathbf{X}}) \mid \mathbf{X} = g^{(n)}(\mathbf{s}) + \boldsymbol{\alpha}], \end{aligned} \tag{60}$$

where $\sum_{\mathbf{s}}$ is the sum over the K_0^n equally-likely values of $\mathbf{s}$. Let us now average the right member of expression (60) over all $\boldsymbol{\alpha}$ in $[0, 2\delta]^n$, with $\boldsymbol{\alpha}$ assumed to be uniformly distributed. That average is

$$\int_{[0,2\delta]^n} \frac{d\mathbf{x}}{(2\delta)^n} \sum_{\mathbf{s}} \frac{1}{K_0^n} E[d_\delta(\mathbf{X}, \hat{\mathbf{X}}) \mid \mathbf{X} = g^{(n)}(\mathbf{s}) + \boldsymbol{\alpha}].$$

If we note that the set $\{g^{(n)}(\mathbf{s})\}$ are in one-to-one correspondence with the K_0^n members $\boldsymbol{\xi}_i$ of $\mathcal{E}^n$, this quantity may be written as

$$\sum_{i=1}^{K_0^n} \frac{1}{K_0^n} \int_{[0,2\delta]^n} \frac{1}{(2\delta)^n} E[d_\delta(\mathbf{X}, \hat{\mathbf{X}}) \mid \mathbf{X} = \boldsymbol{\xi}_i + \boldsymbol{\alpha}] \, d\boldsymbol{\alpha},$$

which equals $\bar{d}_\delta$ by equation (58). Since there must be at least one value of $\boldsymbol{\alpha}$ for which the right member of expression (60) is as small as the average, we have proved inequality (59).

The first inequality of expression (23) follows from inequality (59) on noting as in part (*i*) that $Q(T, A, \delta)$ is a decreasing function of δ.

3.3 *Proof of Theorems* 4 *and* 5

Since Theorem 5 includes Theorem 4 as a special case we need only give a proof of Theorem 5. Our task is further simplified since the basic idea of the proof of Theorem 5 is the same as in Theorem 2 (Section 3.2). Here too we break the proof into two parts. In part (*i*) we assume that we are given an encoder-decoder for the digital source and deduce the existence of an encoder-decoder for the general source (which plays the

part of the analog source in Theorem 2). In part (*ii*) we do the opposite. However we do not have the complications here which necessitated an averaging argument in Section 3.2.

(*i*) We prove here that $\hat{G}(T, \delta) \leqq \hat{P}_e[T, M_C(\delta)]$, the second inequality of expression (33). The proof parallels that of part (*i*) in Section 3.2. Instead of the analog source space $\mathfrak{X}_a$ we have here a general space $\mathfrak{X}$. The distortion is $d_\delta(x, \hat{x})$ with $| x - \hat{x} |$ replaced by $\rho_0(x, \hat{x})$.

To transmit the source outputs which belong to $\mathfrak{X}$ we use the system in Fig. 4. The digital encoder-decoder is for a K_0-ary source where $K_0 = M_C(\delta)$. We assume that it attains a guaranteed distortion $\hat{d}_H$. The quantizer is defined as follows. Let $\{\beta_i\}_0^{K_0-1}$ be a minimum δ-covering of $\mathfrak{X}$. For $x \in \mathfrak{X}$, let $q(x)$ be the smallest $i (0 \leqq i \leqq K_0 - 1)$ such that $x \in S_{\beta_i}(\delta)$. Then if $\mathbf{x} = (x_1, x_2, \cdots, x_n) \in \mathfrak{X}^n$ is the source output, the quantizer output is $\mathbf{s} = q^n(\mathbf{x}) = [q(x_1), q(x_2), \cdots, q(x_n)]$. The output of the digital decoder is $\hat{\mathbf{S}} = (\hat{S}_1, \hat{S}_2, \cdots, \hat{S}_n)$ and the converter output is $\hat{\mathbf{X}} = (\hat{X}_1, \cdots, \hat{X}_n)$, where $\hat{X}_k = \beta_i$ when $\hat{S}_k = i$. Clearly, if $S_k = \hat{S}_i$, then $\rho_0(X_k, X_k) < \delta$. Thus for any source output $\mathbf{x}$,

$$\bar{d}_\delta(\mathbf{x}) \leqq \bar{d}_H[q^{(n)}(\mathbf{x})] \leqq \hat{d}_H ,$$

so that the overall guaranteed distortion $\hat{d}_\delta \leqq \hat{d}_H$, from which part (*i*) follows.

(*ii*) We prove here that $\hat{P}_e[T, M_P(\delta)] \leqq \hat{G}(T, \delta)$, the first inequality of expression (33). As in part (*i*), the proof of part (*ii*) parallels that in Section 3.2. Again $\mathfrak{X}_a$ is replaced by $\mathfrak{X}$ and $| x - \hat{x} |$ by $\rho_0(x, \hat{x})$.

As in Section 3.2, we assume that we are given an encoder-decoder for the general source with guaranteed distortion $\hat{d}_\delta$. We set $K_0 = M_P(\delta)$ and use the system of Fig. 6 to transmit the outputs of the K_0-ary digital source. The modulator is defined as follows. Let $\{\beta_i\}_{i=0}^{K_0-1}$ be a minimum δ-packing of $\mathfrak{X}$. If source output is $\mathbf{s} = (s_1, s_2, \cdots, s_n)$, then the modulator output is $g^{(n)}(\mathbf{s}) = (\beta_{s_1}, \beta_{s_2}, \cdots, \beta_{s_n})$. The output of the decoder is $\hat{\mathbf{X}} = (\hat{X}_1, \cdots, \hat{X}_n)$, and the quantizer output is $\hat{\mathbf{S}} = (\hat{S}_1, \hat{S}_2, \cdots, \hat{S}_n)$, where $\hat{S}_k = i$ when $\hat{X}_k \in S_{\beta_i}(\delta)$. If $X_k \notin S_{\beta_i}(\delta)$ for all i $(0 \leqq i \leqq K_0 - 1)$, then $\hat{S}_k = 0$.

Clearly, if $\rho_0(X_k, \hat{X}_k) < \delta$, then $S_k = \hat{S}_k$. Thus for any source output $\mathbf{s}$, the conditional expectation

$$\bar{d}_H(\mathbf{s}) \leqq \bar{d}_\delta[g^{(n)}(\mathbf{s})] \leqq \hat{d}_\delta .$$

Thus the overall guaranteed distortion is $\hat{d}_H \leqq \hat{d}_\delta$, completing the proof of part (*ii*) and the theorem.

3.4 *Proofs on Packing and Covering*

In this section we give a proof of Theorem 3, the main part of which is a lemma on covering of the K-ary n-cube. We also prove Lemma 1 relating packing and covering in Section 2.4.

3.4.1 *Proof of Theorem* 3

We first establish the following lemma.

Lemma 2: Let $\theta(0 < \theta < (K-1)/K)$ *be arbitrary, and let* r *satisfy*

$$R_{eq}(\theta) < \mathrm{r} < \log K,$$

where $\mathrm{R}_{eq}(\lambda)$ *is the equivalent rate for the* K*-ary source given by expressions* (9). *Using the terminology of Section 2.4, let* $\mathfrak{X} = \{0, 1, \cdots, \mathrm{K}-1\}^{\mathrm{n}}$ *(the K-ary n-cube) and* $\rho_0(\mathbf{x}, \hat{\mathbf{x}}) = d_H^{(n)}(x, \hat{\mathbf{x}})$. *Then for* n *sufficiently large, there exists a* θ*-covering of* $\mathfrak{X}$ *with* $M = e^{rn}$ *points.*

Proof: Let $\{\mathbf{x}_i\}_1^M$ be a set of K-ary n-vectors. Let $F(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M)$ be the number of members $\hat{\mathbf{x}}$ of $\mathfrak{X}$ such that $d_H^{(n)}(\mathbf{x}_i, \hat{\mathbf{x}}_i) \geqq \theta$ for all $i = 1, 2, \cdots, M$. If $F = 0$, then $\{\mathbf{x}_i\}_1^M$ is a θ-covering of $\mathfrak{X}$. We can write

$$F(\mathbf{x}_1, \cdots, \mathbf{x}_M) = \sum_{\hat{\mathbf{x}} \in \mathfrak{X}} \Phi(\hat{\mathbf{x}}, \mathbf{x}_1, \cdots, \mathbf{x}_M),$$

where

$$\Phi(\hat{\mathbf{x}}, \mathbf{x}_1, \cdots, \mathbf{x}_M) = \begin{cases} 1 & \text{if } d_H^{(n)}(\mathbf{x}_i, \hat{\mathbf{x}}) \geqq \theta, \text{ all } i = 1, 2, \cdots, M, \\ 0 & \text{otherwise.} \end{cases}$$

Now consider an experiment in which $M = e^{rn}$ n-vectors $\{\mathbf{X}_i\}_1^M$ are chosen at random from $\mathfrak{X}$ independently with identical (uniform) distribution

$$\Pr\{\mathbf{X}_i = \mathbf{x}\} \equiv K^{-n}.$$

Then $F(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_M)$ is random variable with expectation

$$EF = \sum_{\hat{\mathbf{x}} \in \mathfrak{X}} E\Phi(\hat{\mathbf{x}}, \mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_M),$$

where, as indicated, $E\Phi$ is computed with $\hat{\mathbf{x}}$ held fixed. Now for a given $\hat{\mathbf{x}}$,

$$\begin{aligned} E\Phi(\hat{\mathbf{x}}, \mathbf{X}_1, \cdots, \mathbf{X}_M) &= \Pr\{\Phi = 1\} \\ &= \Pr \bigcap_{i=1}^{M} \{d_H^{(n)}(\mathbf{x}, \mathbf{X}_i) \geqq \theta_i\} \\ &= [\Pr\{d_H^{(n)}(\hat{\mathbf{x}}, \mathbf{X}_i) \geqq \theta\}]^M, \end{aligned}$$

where the last equality follows from the independence and identical distribution of the random vectors $\{\mathbf{X}_i\}$. Letting $a_n = \sum_{0 \leqq i < \theta n} \binom{n}{j}$. $(K-1)^i K^{-n}$ be the probability that $d_H(\hat{\mathbf{x}}, \mathbf{X}_i) < \theta n$, we have

$$E\Phi(\hat{\mathbf{x}}, \mathbf{X}_1, \cdots, \mathbf{X}_M) = (1 - a_n)^M \leqq e^{-a_n M},$$

independent of $\mathbf{x}$. Thus

$$EF \leqq Me^{-a_n M}.$$

Now it is well known (see for example, Ref. 8, p. 173) that for $0 < \theta < (K-1)/K$, as $n \to \infty$,

$$a_n = e^{-nR_{eq}(\theta) + o(n)}.$$

Thus since $M = 2^{rn}$ and $r > R_{eq}(\theta)$,

$$E(F) \leqq Me^{-a^n M} = e^{rn} \exp\{-e^{[r - R_{eq}(\theta)]n + o(n)}\} \to 0, \quad \text{as} \quad n \to \infty.$$

Now, there must be at least one particular set $\{\mathbf{x}_i\}_1^M$ such that

$$F(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_M) \leqq EF.$$

Thus if we choose n large enough so that $E(F) < 1$, $F(\mathbf{x}_1, \cdots \mathbf{x}_M) = 0$ (since F is an integer valued function). Thus $\{\mathbf{x}_i\}_1^M$ is the required covering.

The proof of Theorem 3 now follows the standard proof of a source-channel coding theorem, with Lemma 2 playing the role of the source coding theorem. (See Ref. 2.) Roughly speaking the proof is as follows. When $\gamma(K, \rho_S, C) = (K-1)/K$, the entire theorem is trivial, since we can attain a guaranteed distortion of $(K-1)/K$ without even using the channel by simply letting the decoder outputs take the value $i (0 \leqq i \leqq K-1)$ with probability $1/K$. Thus assume that $0 \leqq \gamma < (K-1)/K$.

The channel can transmit e^{RT} (where $R < C$, the channel capacity) in T seconds with arbitrarily high reliability (see Section 2.1). By the definition of $\gamma = \gamma(K, \rho_S, C)$ [expression (10b)],

$$R_{eq}(\gamma) \leqq C/\rho_S. \tag{61}$$

Let $\epsilon > 0$ be arbitrary. In Lemma 2, let $r = C/\rho_S - \epsilon_1$, where $\epsilon_1 > 0$ will be chosen below. Then approximate the T-second source output (a K_0-ary n-vector, $n = \rho_S T$) by a (covering) set with $e^{rn} = e^{r\rho_S T}$ members. Since $r\rho_S < C$ we can transmit these n-vectors through the channel with arbitrarily high reliability. Further, with $\epsilon > 0$ arbitrary, and if

$$r > R_{eq}(\gamma + \epsilon), \tag{62}$$

we have from Lemma 2 that the error in making the approximation will always be less than or equal to $(\gamma + \epsilon)$ for T sufficiently large. In fact, if we set

$$\epsilon_1 = R_{eq}(\gamma) - R_{eq}\left(\gamma + \frac{\epsilon}{2}\right) > 0$$

[since $R(\gamma)$ as defined in equation (9) is strictly decreasing for $\gamma < (K - 1)/K$], then [using inequality (61)]

$$r = \frac{C}{\rho_S} - \epsilon_1 = \frac{C}{\rho_S} - R_{eq}(\gamma) + R_{eq}\left(\gamma + \frac{\epsilon}{2}\right)$$

$$\geqq R_{eq}\left(\gamma + \frac{\epsilon}{2}\right) > R_{eq}(\gamma + \epsilon)$$

and condition (62) is satisfied.

We conclude that for T sufficiently large, we can make

$$\hat{d}_H \leqq \gamma + \epsilon$$

for arbitrary $\epsilon > 0$. Thus

$$\limit_{T\to\infty} \hat{P}_e(T, K) \leqq \gamma + \epsilon \to \gamma, \quad \text{as} \quad \epsilon \to 0,$$

which is Theorem 3.

3.4.2 *Proof of Lemma* 1

We say that $A \subseteq \mathfrak{X}$ is a "maximal Δ-packing" if A is a Δ-packing, and if for all $v \notin A$, the union $\{v\} \cup A$ is not a Δ-packing. We establish Lemma 1 by showing that every maximal Δ-packing is a $(2\eta\Delta)$-covering. Let A be a maximal Δ-packing. If A is not a $(2\eta\Delta)$-covering, then there exists a $v_0 \in \mathfrak{X}$ such that $\rho_0(v_0, u) > 2\eta\Delta$, for all $u \in A$. From condition (30b), $v_0 \notin A$. We claim that $\{v_0\} \cup A$ is a Δ-packing, contradicting the maximality of A. If $w \in S_{v_0}(\Delta)$, then for all $u \in A$ (using the definition of η)

$$\rho_0(v_0, u) \leqq \eta[\rho_0(v_0, w) + \rho_0(w, u)],$$

so that

$$\rho_0(w, u) \geqq \frac{\rho_0(v_0, u)}{\eta} - \rho_0(v_0, w) > \frac{2\eta\Delta}{\eta} - \Delta = \Delta.$$

Thus $w \notin S_u(\Delta)$ and $\{v_0\} \cup A$ is a Δ-packing, establishing the lemma.

APPENDIX

List of Symbols

$\mathfrak{X}$		the source output space
$P_S(x)$		the source probability density function
ρ_S		source output rate (symbols per second)
x_i		$(x_i \in \mathfrak{X})$ the ith output of the source
$\mathbf{x}$		$(x_1, x_2, \cdots x_n) \in \mathfrak{X}^n$
$\mathfrak{W}_T$		set of "allowable" channel inputs
$\mathfrak{d}_T$		the set of all channel outputs
T		the coding delay
n	$=$	$\rho_S T$
$f_E(\mathbf{x})$		the encoding function, $f_E(\mathbf{x}) \in \mathfrak{W}_T$
$f_D(z)$		the decoding function, $f_D(z) \in \mathfrak{X}^n$
$\hat{\mathbf{x}}$		the decoded n-vector, $\hat{\mathbf{x}} = f_D(z) \in \mathfrak{X}^n$
N		number of code words in a code
R	$=$	$1/T \log N$, the rate of a code
λ		the word probability of error
$\lambda^*(T, N)$		smallest attainable word error probability for a code with parameters N and T
$\bar{\lambda}$		average probability of error
$d(x, \hat{x})$		the distortion function
$d_H(x, \hat{x})$	$=$	$\begin{cases} 0, & x = \hat{x} \\ 1, & x \neq \hat{x} \end{cases}$
$d^n(\mathbf{x}, \hat{\mathbf{x}})$	$=$	$1/n \sum_{K=1}^{n} d(x_K, \hat{x}_K)$
$\bar{d}$	$=$	$Ed^{(n)}(\mathbf{x}, \hat{\mathbf{x}})$
$\bar{d}^*(T)$		the smallest attainable $\bar{d}$ for a given delay T
$d_\delta(x, \hat{x})$	$=$	$\begin{cases} 1 & \lvert x - \hat{x} \rvert < \delta \\ 0 & \lvert x - \hat{x} \rvert \geqq \delta \end{cases}$
$\bar{d}(\mathbf{x})$		the expectation of $d^n(\mathbf{x}, \hat{\mathbf{x}})$ given $\mathbf{x}$
$\hat{d}$	$=$	$\sup_{\mathbf{x} \in \mathfrak{X}^n} \bar{d}(\mathbf{x})$
$\hat{d}^*(T)$		the smallest attainable value of $\hat{d}$ for a given delay T
$Q(T, A, \delta)$		$\bar{d}^*(T)$ for $d_\delta(x, \hat{x})$ and $x \in [-A/2, A/2]$
$\hat{Q}(T, \delta, A)$		$\hat{d}^*(T)$ for $d_\delta(x, \hat{x})$ and $x \in [-A/2, A/2]$
$P_e(T, K)$		the minimum attainable per symbol error rate for an equiprobable K-ary memoryless source
$\gamma(K, \rho_S, C)$	$=$	$\lim_{T \to \infty} P_e(T, K)$

C	the channel capacity
$\hat{P}_e(T, K)$	the minimum attainable guaranteed per symbol error rate for a K-ary source
$G(T, \delta)$	generalization of $\hat{Q}$, defined in Section 2.4

REFERENCES

1. Shannon, C. E., "Probability of Error for Optimal Codes in a Gaussian Channel," B.S.T.J., *38*, No. 3 (May 1959), pp. 611–656.
2. Shannon, C. E., "Coding Theorems for a Discrete Source with a Fidelity Criterion," IRE Nat. Conv. Record, March 23–26, 1959, part 4, pp. 142–163.
3. Wyner, A. D., "Capabilities of Bounded Discrepancy Decoding," B.S.T.J., *44*, No. 6 (July-August 1965), pp. 1061–1022.
4. Shannon, "A Mathematical Theory of Communication," B.S.T.J., *27*, No. 3 (July 1948), pp. 379–423; No. 4 (October 1948), pp. 623–656.
5. Panter, P. F., and Dite, W., "Quantization Distortion in Pulse-Count Modulation with Non-Uniform Spacing of Levels," Proc. IRE, *39*, No. 1 (January 1951), pp. 44–48.
6. Wozencraft, J. M., and Jacobs, I. M., *Principles of Communication Engineering*, New York: Wiley, 1965.
7. Ziv, J., and Zakai, M. "Some Lower Bounds on Signal Parameter Estimation," IEEE Trans. Inform. Theory, *IT-15*, No. 3 (May 1969), pp. 386–391.
8. Ash, R. B., *Information Theory*, New York: Interscience Publishers, 1965.

On Digital Communication Over a Discrete-Time Gaussian Channel with Noisy Feedback

By AARON D. WYNER

(Manuscript received April 21, 1969)

We consider the problem of transmission of digital data over a discrete-time Gaussian channel with the use of a Gaussian feedback channel. We are particularly interested in the case where the signal-to-noise ratio in the feedback channel is finite. By making use of simple extension of P. Elias' scheme for transmitting analog data over this channel with feedback, we show that it is possible at some transmission rates to increase the error-exponent (reliability) compared to the error-exponent found by C. E. Shannon for the one-way channel. In particular at transmission rate zero, we show that the error-exponent can be improved by a factor of $1 + [\hat{\rho}/(1 + \rho)]$, *where* ρ *and* $\hat{\rho}$ *are the forward and feedback signal-to-noise ratios respectively.*

I. INTRODUCTION

We consider the problem of transmission of digital data over a discrete-time Gaussian channel with the use of a Gaussian feedback channel. We are particularly interested in the case where the signal-to-noise ratio in the feedback channel is finite.

In Sections II and III we consider Elias' scheme and a simple extension for transmitting analog data over this channel with feedback.[1,2] In Section IV we apply this extended Elias scheme to the digital transmission problem. The main result is that for any rate $R < R^*$, a number less than the channel capacity, it is possible to transmit digital data at a rate R with error probability

$$P_e = \exp [-E^* n_o + o(n_o)], \quad \text{as} \quad n_o \to \infty,$$

where n_o is the encoding-decoding delay, and $E^* > E_1$, the "one-way" exponent estimated by Shannon.[3] In particular, when $R = 0$, $E_1 = \rho/4$ and $E^* = (\rho/4)[1 + \hat{\rho}(1 + \rho)^{-1}]$, where ρ and $\hat{\rho}$ are the forward and

feedback signal-to-noise ratios respectively. Finally, we suggest a modfication of this scheme which will probably permit extending R^* to capacity.

Stimulated by the work of Schalkwijk and Kailath, a great deal of research has been done on this problem (see for example Refs. 4–11). To the present author's knowledge, however, the result in this paper is the first to show that a noisy feedback channel can improve the error-exponent for digital communication on a band-limited channel. (References 4 and 8 treat the infinite band case.) Like the optimal coding schemes for the one-way channel, our scheme is not constructive. Let us remark here that this discrete-time channel is a model for the continuous-time Gaussian channel with a bandwidth constraint. (See Ref. 12 or 13.)

II. STATEMENT OF ELIAS' PROBLEM

We define a Gaussian channel as follows. The input is a real number x and the output is a number $y = x + z$, where the "noise" z is a Gaussian variate with mean zero and variance σ^2 and is independent of x. We assume here that the channel input x is a random variable, and require that the expectation $Ex^2 \leq P$, the "signal power".

To begin with, let us suppose that we wish to transmit the value of a random variable θ with the use of N transmissions over a Gaussian channel (with parameters P and σ^2). Assume also that a feedback Gaussian channel (with parameters $\hat{P}$ and $\hat{\sigma}^2$) is available which we may use $(N - 1)$ times alternating with the N forward uses. We assume nothing about the statistical nature of θ except that the expectation $E\theta^2 = \sigma_\theta^2$. Our goal is to obtain an unbiased estimate $\hat{\theta}$ of θ with minimum possible mean-squared error. Further, we restrict ourselves to linear processing of all data. We now state the problem and constraints precisely.

The forward and feedback channels are memoryless Gaussian channels with signal power P and $\hat{P}$ respectively and noise power σ^2 and $\hat{\sigma}^2$ respectively. Thus for the nth use of the forward channel the input is x_n and the output is $y_n = x_n + z_n$, where $Ex_n^2 = P$ and z_n is a Gaussian variate (independent of x_n) with mean zero and variance σ^2. For the nth use of the feedback channel the input is $\hat{x}_n$ and the output is $\hat{y}_n = \hat{x}_n + \hat{z}_n$, where $E\hat{x}_n^2 = \hat{P}$ and $\hat{z}_n$ is a Gaussian variate (independent of $\hat{x}_n$) with mean zero and variance $\hat{\sigma}^2$. We assume that the random variables $\{\theta, z_n, \hat{z}_n\}$ are independent. The condition requiring "linear processing" means the following. The input x_n to the forward channel (at the nth use) is given by

$$x_1 = a_1\theta, \qquad x_n = a_n\theta + \sum_{k=1}^{n-1} b_{nk}\hat{y}_k , \qquad n = 2, 3, \cdots , N. \tag{1}$$

The input to the feedback channel $\hat{x}_n$ (at the nth use) is given by

$$\hat{x}_n = \sum_{k=1}^{n} c_{nk}y_k , \qquad n = 1, 2, \cdots , N - 1. \tag{2}$$

Finally, the receiver's estimate after N uses of the forward channel (and $N - 1$ uses of the feedback channel) is

$$\hat{\theta} = \sum_{n=1}^{N} d_n y_n . \tag{3}$$

We require that $\hat{\theta}$ be unbiased, that is, that given that $\theta = \theta_o$, the conditional expectation of $\hat{\theta}$ is

$$E(\hat{\theta} \mid \theta = \theta_o) = \theta_o . \tag{4}$$

The mean squared-error, which we wish to minimize is

$$\gamma^2 = E(\theta - \hat{\theta})^2. \tag{5}$$

Let γ^2_{OPT} be the minimum attainable value of γ^2 (over all choices of the coefficients a_n, b_{nk}, c_{nk}, d_n). It is easy to show that

(*i*) γ^2_{OPT} depends on P and σ^2 only through their ratio $\rho \triangleq P/\sigma^2$ (the forward "signal-to-noise" ratio), and on $\hat{P}$ and $\hat{\sigma}^2$ only through $\hat{\rho} \triangleq \hat{P}/\hat{\sigma}^2$.

(*ii*) for a given N, ρ, and $\hat{\rho}$, γ^2_{OPT} is proportional to σ_θ^2.

Thus we can write

$$\gamma^2_{\text{OPT}} = \sigma_\theta^2\epsilon^2_{\text{OPT}}(\rho, \hat{\rho}, N),$$

and our problem reduces to the determination of $\epsilon^2_{\text{OPT}}(\rho, \hat{\rho}, N)$ (which can be thought of as a noise-to-signal ratio).

Let us observe that from the linearity assumptions (equation 1, 2, and 3) it follows that

$$\hat{\theta} = a\theta + \xi, \tag{6}$$

where a is a constant and ξ is a Gaussian variate independent of θ. From equation (4) it follows that $a = 1$ and $E\xi = 0$, and from equation (5) $E\xi^2 = \gamma^2$. Thus we can rewrite equation (6) as

$$\hat{\theta} = \theta + \xi, \tag{7}$$

where ξ is a Gaussian variate (independent of θ) with mean zero and variance γ^2. The important point here is that the entire process may

be thought of as reducing the N uses of the forward channel (and the $N-1$ uses of the feedback channel) to a single one-way Gaussian channel with signal-to-noise ratio $(E\theta^2)/(E\xi^2) = \sigma_\theta^2/\gamma^2$.

III. ELIAS' RESULT

Elias solved our problem for the special case $N = 2$, where two uses of the forward channel and one of the feedback channel are permitted.[1,2] In his solution Elias admits the possibility that for the two uses of the forward channel, the signal-to-noise ratios are ρ_1 and ρ_2 respectively, where ρ_1 is not necessarily equal to ρ_2. His result is that the smallest attainable mean-squared error is given by

$$\gamma_E^2 = \sigma_\theta^2\left[\rho_1 + \rho_2 + \frac{\rho_1\rho_2\hat{\rho}}{(1+\rho_1)(1+\rho_2)+\hat{\rho}}\right]^{-1}. \tag{8}$$

As discussed at the end of Section II, we can consider the entire process as a single one-way gaussian channel with signal-to-noise ratio $\sigma_\theta^2/\gamma_E^2$. We now turn to our problem, and note that we can obtain a (suboptimal) solution by applying Elias' technique recursively. For $N = 2$ we can, by setting $\rho_1 = \rho_2 = \rho$ in equation (8), obtain a signal-to-noise ratio $S_2 = \{2\rho + \rho^2\hat{\rho}/[(1+\rho)^2 + \hat{\rho}]\}$. For $N = 3$ we can, by setting $\rho_1 = S_2$ and $\rho_2 = \rho$, obtain a signal-to-noise ratio S_3 given by

$$S_3 = \left[S_2 + \rho + \frac{\rho\hat{\rho}S_2}{(1+S_2)(1+\rho)+\hat{\rho}}\right],$$

and for arbitrary N we can obtain a signal-to-noise ratio S_N given by the recurrence

$$S_N = S_{N-1} + \rho + \frac{\rho\hat{\rho}S_{N-1}}{(1+S_{N-1})(1+\rho)+\hat{\rho}}, \tag{9a}$$

with initial condition

$$S_1 = \rho. \tag{9b}$$

Although equation (9) is difficult to solve explicitly we can obtain an approximate solution valid for large N. From equation (9a)

$$S_{N-1} + \rho \leqq S_N \leqq S_{N-1} + \rho + \frac{\hat{\rho}\rho}{(1+\rho)}, \tag{10}$$

so that

$$\rho N \leqq S_N \leqq \left(\rho + \frac{\hat{\rho}\rho}{1+\rho}\right)N. \tag{11}$$

We will show that as $N \to \infty$, S_N is asymptotic to the right member of inequality (11). Let us rewrite equation (9a)

$$S_{N+1} = S_N + \rho + \frac{\rho\hat{\rho}}{(1+\rho)}\left[1 + \frac{1+\rho+\hat{\rho}}{(1+\rho)S_N}\right]^{-1}. \tag{12}$$

Let $S_N = [\rho + \hat{\rho}\rho/(1+\rho)]N + \delta_N$, and expand the last term in equation (12) into a power series in $(1+\rho+\hat{\rho})/[(1+\rho)S_N]$. We then obtain, after cancelling terms,

$$\delta_{N+1} = \delta_N + \frac{\rho\hat{\rho}}{(1+\rho)}\left[-\frac{(1+\rho+\hat{\rho})}{(1+\rho)S_N} + \frac{(1+\rho+\hat{\rho})^2}{(1+\rho)^2 S_N^2} + \cdots\right]. \tag{13}$$

From equation (11) we have that $S_N = O(N)$, so that equation (13) becomes

$$\delta_{N+1} - \delta_N = -O(1/N), \tag{14}$$

and therefore

$$\delta_N = -O(\log N). \tag{15}$$

Thus we conclude that

$$S_N = \left[\rho + \frac{\rho\hat{\rho}}{(1+\rho)}\right]N - O(\log N). \tag{16}$$

An exact solution for S_N for various values of ρ, $\hat{\rho}$, and N is given in Table I. S_N^{-1} provides an upper bound to ϵ_{OPT}^2.

Elias also found a lower bound to ϵ_{OPT}^2,

$$\epsilon_{\text{OPT}}^2 \geqq 1/[\rho N + \hat{\rho}(N-1)]. \tag{17}$$

This is the mean-squared error which results when the feedback channel is reversed and used in the forward direction, and we are allowed to use the forward channel N times and the feedback channel $(N-1)$ times. Combining these results we have that

$$[(\rho+\hat{\rho})N]^{-1} \leqq \epsilon_{\text{OPT}}^2 \leqq \left[\left(\rho + \frac{\rho\hat{\rho}}{1+\rho}\right)N - O(\log N)\right]^{-1}. \tag{18}$$

Let us remark here that the recurrence (9) can be solved exactly for the special case $\hat{\rho} = \infty$. In this case equation (9a) becomes

$$S_N = S_{N-1}(1+\rho) + \rho, \tag{19}$$

and the solution is

$$S_N = (1+\rho)^N - 1. \tag{20}$$

TABLE I—THE EXTENDED ELIAS SCHEME

FORWARD SNR = ρ, FEEDBACK SNR = $\hat{\rho}$

ASYMP. SNR = $\lim_{N\to\infty} \frac{S_N}{N} = \rho + \frac{\rho\hat{\rho}}{1+\rho}$, ASYMP. $E(0) = E^*(0)$

FORWARD SNR = 0.01 FEEDBACK SNR = 0.01
ASYMP. SNR = 0.010099 ASYMP. $E(0)$ = 2.52475E-03†

N	EQ. SNR = S_N	EQ. $E(0) = E_N(0)$	CAPACITY = c_N
1	0.01	0.0025	4.97516E-03
2	0.020001	2.50012E-03	4.95089E-03
3	3.00029E-02	2.50024E-03	4.92693E-03
4	4.00058E-02	2.50036E-03	4.90328E-03
5	5.00095E-02	2.50048E-03	4.87992E-03
6	6.00142E-02	2.50059E-03	4.85686E-03
7	7.00197E-02	2.50071E-03	4.83408E-03
8	8.00262E-02	2.50082E-03	4.81158E-03
9	9.00334E-02	2.50093E-03	4.78935E-03
10	0.100042	2.50104E-03	4.76740E-03

FORWARD SNR = 0.01 FEEDBACK SNR = 0.1
ASYMP. SNR = 1.09901E-02 ASYMP. $E(0)$ = 2.74752E-03

N	EQ. SNR = S_N	EQ. $E(0) = E_N(0)$	CAPACITY = c_N
1	0.01	0.0025	4.97516E-03
2	2.00089E-02	2.50112E-03	4.95284E-03
3	3.00266E-02	2.50222E-03	4.93078E-03
4	0.040053	2.50331E-03	4.90895E-03
5	5.00878E-02	2.50439E-03	4.88738E-03
6	6.01309E-02	2.50546E-03	4.86604E-03
7	7.01823E-02	2.50651E-03	4.84493E-03
8	8.02417E-02	2.50755E-03	4.82405E-03
9	9.03091E-02	2.50859E-03	4.80340E-03
10	0.100384	2.50961E-03	4.78297E-03
20	0.20154	2.51924E-03	4.59009E-03
30	0.303354	2.52795E-03	4.41569E-03
40	0.40574	2.53587E-03	4.25704E-03
50	0.508622	2.54311E-03	4.11197E-03
100	1.02871	2.57178E-03	3.53701E-03
150	1.55532	2.59219E-03	3.12725E-03
200	2.0861	2.60762E-03	2.81727E-03
250	2.61979	2.61979E-03	2.57283E-03
300	3.1556	2.62967E-03	2.37410E-03
350	3.69305	2.63789E-03	2.20869E-03
400	4.23179	2.64487E-03	2.06844E-03
450	4.77156	2.65087E-03	1.94771E-03
500	5.31219	2.65609E-03	1.84248E-03

† The notation "3E-5" means 3×10^{-5}.

TABLE I—(Continued)

FORWARD SNR = 0.01 FEEDBACK SNR = 1
ASYMP. SNR = 0.019901 ASYMP. $E(0) = 4.97525\text{E-}03$

N	EQ. SNR = S_N	EQ. $E(0) = E_N(0)$	CAPACITY = c_N
1	0.01	0.0025	4.97516E-03
2	2.00495E-02	2.50619E-03	4.96279E-03
3	3.01483E-02	2.51235E-03	4.95045E-03
4	0.040296	2.51850E-03	4.93816E-03
5	5.04925E-02	2.52463E-03	4.92591E-03
10	0.102197	2.55493E-03	4.86529E-03
15	0.155082	2.58470E-03	4.80572E-03
20	0.209115	2.61394E-03	4.74722E-03
40	0.436077	2.72548E-03	4.52394E-03
60	0.678846	2.82852E-03	4.31755E-03
80	0.935508	2.92346E-03	4.12731E-03
100	1.20434	3.01085E-03	3.95214E-03
300	4.31176	3.59313E-03	2.78320E-03
500	7.79234	3.89617E-03	2.17388E-03
700	11.4295	4.08196E-03	1.80005E-03
900	15.1502	4.20840E-03	1.54552E-03

FORWARD SNR = 1 FEEDBACK SNR = 0.1
ASYMP. SNR = 1.05 ASYMP. $E(0) = 0.2625$

N	EQ. SNR = S_N	EQ. $E(0) = E_N(0)$	CAPACITY = c_N
1	1	0.25	0.346574
2	2.02439	0.253049	0.276677
3	3.05731	0.254776	0.23342
4	4.09453	0.255908	0.203521
5	5.13433	0.256716	0.18139
6	6.17584	0.257327	0.164227
7	7.21857	0.257806	0.150457
8	8.26222	0.258194	0.139122
9	9.30658	0.258516	0.129599
10	10.3515	0.258788	0.121468

FORWARD SNR = 1 FEEDBACK SNR = 1
ASYMP. SNR = 1.5 ASYMP. $E(0) = 0.375$

N	EQ. SNR = S_N	EQ. $E(0) = E_N(0)$	CAPACITY = c_N
1	1	0.25	0.346574
2	2.2	0.275	0.290788
3	3.4973	0.291441	0.250579
4	4.84722	0.302951	0.220746
5	6.22905	0.311453	0.197811
6	7.63202	0.318001	0.179623
7	9.04989	0.32321	0.164826
8	10.4788	0.327462	0.152531
9	11.9162	0.331005	0.142138
10	13.3603	0.334007	0.133223

TABLE I—(Continued)

FORWARD SNR = 1 ASYMP. SNR = 51		FEEDBACK SNR = 100 ASYMP. $E(0) = 12.75$	
N	EQ. SNR = S_N	EQ. $E(0) = E_N(0)$	CAPACITY = c_N
1	1	0.25	0.346574
2	2.96154	0.370192	0.344158
3	6.70566	0.558805	0.340326
4	13.5159	0.844743	0.334405
5	24.9907	1.24954	0.325774
6	42.434	1.76808	0.31427
7	66.142	2.36222	0.300486
8	95.3736	2.98042	0.285515
9	128.952	3.58201	0.270398
10	165.782	4.14455	0.255834
50	2073.46	10.3673	7.63746E-02
90	4079.34	11.3315	4.61885E-02
200	9646.14	12.0577	0.022936

FORWARD SNR = 100 ASYMP. SNR = 100.99		FEEDBACK SNR = 1 ASYMP. $E(0) = 25.2475$	
N	EQ. SNR = S_N	EQ. $E(0) = E_N(0)$	CAPACITY = c_N
1	100	25	2.30756
2	200.98	25.1225	1.32704
3	301.965	25.1638	0.95227
4	402.952	25.1845	0.750162
5	503.94	25.197	0.622444
6	604.928	25.2053	0.533897
7	705.916	25.2113	0.468637
8	806.905	25.2158	0.418403
9	907.894	25.2193	0.378457
10	1008.88	25.2221	0.345879

Since the capacity, $\frac{1}{2}\log(1 + S_N)$, of the equivalent Gaussian channel (with signal-to-noise ratio S_N), cannot exceed N times the capacity, $\frac{1}{2}\log(1 + \rho)$, of a single channel (with signal-to-noise ratio ρ), S_N as given by equation (20) is in fact optimal. Thus

$$\epsilon^2_{\text{OPT}}(\rho, \infty, N) = [(1 + \rho)^N - 1]^{-1}, \tag{21}$$

which is an exponential in N.

IV. APPLICATION TO DIGITAL COMMUNICATION

4.1 *Schalkwijk–Kailath Technique*

Suppose we wish to transmit one of M equally likely messages over a Gaussian forward channel with signal-to-noise ratio ρ with the aid of

TABLE I—(Continued)

FORWARD SNR = 100 FEEDBACK SNR = 100
ASYMP. SNR = 199.01 ASYMP. $E(0) = 49.7525$

N	EQ. SNR = S_N	EQ. $E(0) = E_N(0)$	CAPACITY = c_N
1	100	25	2.30756
2	297.078	37.1347	1.42434
3	495.429	41.2858	1.03457
4	694.043	43.3777	0.817997
5	892.77	44.6385	0.679545
6	1091.56	45.4816	0.583023
7	1290.39	46.0853	0.511677
8	1489.25	46.539	0.456669
9	1688.12	46.8923	0.412887
10	1887.02	47.1754	0.377164

FORWARD SNR = 100 FEEDBACK SNR = 1000
ASYMP. SNR = 1090.1 ASYMP. $E(0) = 272.525$

N	EQ. SNR = S_N	EQ. $E(0) = E_N(0)$	CAPACITY = c_N
1	100	25	2.30756
2	1092.78	136.597	1.74935
3	2173.1	181.091	1.28073
4	3258.25	203.641	1.01116
5	4345.00	217.253	0.837702
10	9786.78	244.669	0.459444
15	15232.7	253.878	0.321.042
20	20680.1	258.501	0.248424
40	42474.8	265.467	0.133209
60	64272.6	267.803	9.22575E-02
80	86071.7	268.974	7.10184E-02
100	107871.	269.679	5.79435E-02
120	129672.	270.149	4.90532E-02
140	151472.	270.486	4.26006E-02
160	173273.	270.738	3.76957E-02
180	195073.	270.935	3.38365E-02
200	216874.	271.093	3.07177E-02

a Gaussian feedback channel with signal-to-noise ratio $\hat{\rho}$, using the forward channel N times and the feedback channel $N - 1$ times. Following Schalkwijk and Kailath,[10] we assign to message i ($i = 1, 2, \cdots, M$) the number $\theta = \theta_i = i - (M + 1)/2$. Thus the M messages are equally spaced on the interval $[-(M - 1)/2, (M - 1)/2]$ at distance 1 apart. We can now apply the results of Sections III and IV to transmit θ. The expectation $E\theta^2 = \sigma_\theta^2$ is

$$\sigma_\theta^2 = (M + 1)(M - 1)/12, \quad M = 1, 2, 3 \cdots . \tag{22}$$

When message i is transmitted, the output of the system is $\hat{\theta} = \theta_i + \xi$, where ξ is a zero mean Gaussian random variable with variance γ^2. We select as the decoder output, that $j(1 \leqq j \leqq M)$ which minimizes $|\hat{\theta} - \theta_j|$, so that we make an error only when $|\xi| \geqq \frac{1}{2}$. This event has probability

$$P_e = 2\Phi(-1/2\gamma), \tag{23}$$

where $\Phi(x) = 1/(2\pi)^{\frac{1}{2}} \int_{-\infty}^{x} \exp(-x^2/2)\, dx$ is the cumulative error function. Thus the smallest error probability attainable using this scheme (with parameters N, ρ, $\hat{\rho}$) is

$$P_{e,\,\mathrm{OPT}} = 2\Phi\left[-\frac{1}{2\epsilon_{\mathrm{OPT}}(\rho, \hat{\rho}, N)\sigma_\theta}\right] \tag{24}$$

where σ_θ is given by equation (22) and ϵ_{OPT} in Section II. The bounds on $\epsilon^2_{\mathrm{OPT}}$ in Section III immediately yield bounds on $P_{e,OPT}$.

Let us assume that every T seconds, a digital message source emits one of $M = e^{RT}$ equally likely messages (R is the message "rate"). Further assume that $N = \alpha T$ (for example, if the "physical" channel has bandwidth W cps, then $\alpha = 2W$). Consider two cases: $\hat{\rho} = \infty$, $\hat{\rho} < \infty$.

(*i*) When $\hat{\rho} = \infty$, it follows immediately from equation (21) and (22) that as $T \to \infty$

$$\frac{1}{2\epsilon_{\mathrm{OPT}}(\rho, \infty, N)\sigma_\theta} \sim \sqrt{3}\,\frac{(1+\rho)^{\alpha T}}{e^{RT}} = \sqrt{3}\, e^{(C-R)T}, \tag{25}$$

where $C = (\alpha/2)\log(1+\rho)$ is the channel capacity in nats per second. Thus, provided $R < C$, as $T \to \infty$ the argument of Φ in equation (24) becomes infinite and $P_{e,\mathrm{OPT}} \to 0$. In fact, (since $\Phi(x) \sim (2\pi x^2)^{-\frac{1}{2}} \exp(-x^2/2)$, as $x \to \infty$)

$$P_{e,\,\mathrm{OPT}} = \exp\left[e^{2(C-R)T + o(T)}\right], \quad \text{as} \quad T \to \infty, \tag{26}$$

a double exponential decay. This is the celebrated result of Schalkwijk and Kailath.[10,11]

(*ii*) If we try to apply the same scheme when the feedback signal-to-noise $\hat{\rho} < \infty$, then from equation (18) $(2\epsilon_{\mathrm{OPT}}\sigma_\theta)^{-1} \to 0$ as $T \to \infty$. Thus it is not possible using this scheme to obtain vanishingly small error probability as $T \to \infty$ with fixed signal-to-noise ratios in the forward and feedback channel. This is so no matter how large $\hat{\rho}$ may be, provided it is finite. For finite T however, equations (18) and (24) yield useful estimates of attainable error probabilities.

4.2 *Improving the One-Way Error Exponent*

Suppose that, as in Section 4.1, we wish to transmit one of $M = e^{RT}$ equally likely messages in T seconds. Suppose that we use only a forward Gaussian channel (with signal-to-noise ratio ρ) $n_o = \alpha T$ times. Then it is well known that one can attain an error probability

$$P_e = \exp\left[-E_1\left(\frac{R}{\alpha}, \rho\right)\alpha T + o(T)\right], \quad \text{as} \quad T \to \infty, \tag{27}$$

where $E_1(R/\alpha, \rho) > 0$, if $R < \alpha/2 \log (1 + \rho) = C$, the channel capacity. As indicated, the quantity $E_1(R/\alpha, \rho)$ depends on R and α only through their ratio. Although E_1 is not known exactly, estimates are given in Ref. 3.† In particular, $E_1(0, \rho) = \rho/4$ and $E_1(C/\alpha, \rho) = 0$.

Now suppose we have a Gaussian feedback channel available with signal-to-noise ratio $\hat{\rho}$. Let us divide the n_o forward channel uses into $\nu = n_o/N$ groups of N forward channel uses. In each of these groups we use the extended Elias scheme, (of Sections III and IV, with N uses of the forward channel and $N - 1$ uses of the feedback channel) to generate an equivalent forward Gaussian channel with signal-to-noise ratio S_N given by the recurrence (9). We then use a one-way coding scheme with $\nu = n_o/N = (\alpha/N)T$ uses of the equivalent forward channel. With N held fixed as $T \to \infty$, we can attain an error probability as in equation (27) with α replaced by (α/N) and ρ replaced by S_N—namely,

$$P_e = \exp\left[-E_1\left(\frac{RN}{\alpha}, S_N\right)\frac{\alpha T}{N} + o(T)\right]. \tag{28}$$

Thus the new error-exponent is

$$E_N(R, \rho, \hat{\rho}) = \frac{1}{N} E_1\left(\frac{RN}{\alpha}, S_N\right). \tag{29}$$

Since N is arbitrary, we can state our result:

Theorem: Given a forward and feedback Gaussian channels which can each process α inputs (independently) per second, with signal-to-noise ratio ρ and $\hat{\rho}$ respectively. Then it is possible to transmit digital data at a rate R nats per second with error probability

$$P_e = \exp\left[-E^*\alpha T + o(T)\right], \quad \text{as} \quad T \to \infty, \tag{30a}$$

† The conventional power constraint for a one-way channel is that the time average of the square of the inputs must not exceed P. The power constraint used here is that the *statistical* expectation of the square of *each* input not exceed P. Nither of these constraints imply the other. However, it is not hard to show that the estimates of E_1 (in Ref. 3) are valid for both constraints.

where

$$E^* = E^*\left(\frac{R}{\alpha}, \rho, \hat{\rho}\right) = \sup_{1 \leq N < \infty} E_N = \sup_{1 \leq N < \infty} \frac{1}{N} E_1\left(\frac{RN}{\alpha}, S_N\right), \tag{30b}$$

S_N is the solution to the recurrence (9), and E_1 is the reliability (error-exponent) for the one-way Gaussian channel as in equation (27), and T is the encoding-decoding delay.

Remarks:

(i) Since $S_1 = \rho$ and $E_1(R/\alpha, \rho) > 0$ for $R < \alpha/2 \log (1 + \rho) = C$, then $E^*(R/\alpha, \rho, \hat{\rho}) > 0$ for $R < C$.

(ii) Since $E_1(0, \rho) = \rho/4$,

$$E_N(0, \rho, \hat{\rho}) = \frac{1}{N} E_1\left(\frac{RN}{\alpha}, S_N\right)\Big|_{R=0} = \frac{S_N}{4N},$$

so that from equation (16),

$$E^*(0, \rho, \hat{\rho}) \geqq \frac{S_N}{4N} \to \frac{\rho}{4}\left(1 + \frac{\hat{\rho}}{1 + \rho}\right), \quad \text{as} \quad N \to \infty.$$

In fact, since S_N/N can be shown to be non-decreasing, $E^*(0, \rho, \hat{\rho})$ is in fact equal to this quantity. Thus the use of the feedback channel represents an improvement of a factor of $[1 + \hat{\rho}/(1 + \rho)]$ in the error-exponent at zero rate.

(iii) We can get a rough idea of the behavior of $E^*(R/\alpha, \rho, \hat{\rho})$ as follows. Let $r = R/\alpha$ be the rate in nats per channel use. Let us *crudely* approximate the one-way exponent $E_1(r, \rho)$ as r varies from 0 to $c = C/\alpha$ (the capacity in nats per channel use) by a straight line connecting $(r = 0, E_1 = \rho/4)$ and $(r = c, E_1 = 0)$. See Fig. 1.
Then E_2 has $r = 0$ intercept at

$$E_2(0, \rho, \hat{\rho}) = \frac{S_2}{2 \cdot 4} = \frac{\rho}{4} + \frac{\hat{\rho}\rho^2}{8[(1 + \rho)^2 + \hat{\rho}]} > \frac{\rho}{4},$$

and $E_2(r, \rho, \hat{\rho}) = 0$ at $r = c_2 \triangleq (1/2)(1/2) \log (1 + S_2)$. Similarly, E_N has $r = 0$ intercept at

$$E_N(0, \rho, \hat{\rho}) = \frac{SN}{4N}, \frac{SN}{4(N - 1)},$$

and $E_N(r, \rho, \hat{\rho}) = 0$ at $r = c_N \triangleq (1/2N) \log (1 + S_N)$. From Fig. 1, we see that for each value of $r > 0$, there is a value of $N(1 \leqq N < \infty)$ which maximizes $E_N(r, \rho, \hat{\rho})$ to achieve $E^*(r, \rho, \hat{\rho})$. Values of $E_N(0, \rho, \hat{\rho})$ and c_N are tabulated for various values of ρ, $\hat{\rho}$, and N in Table I.

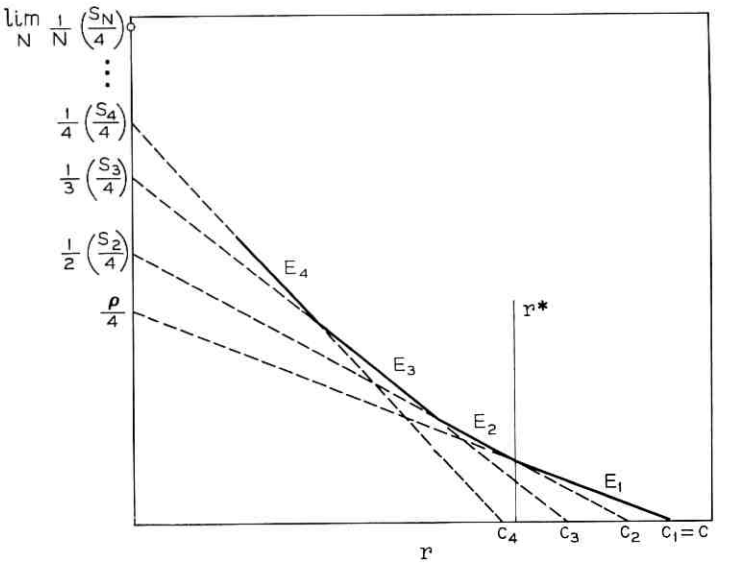

Fig. 1 — $E^*(r, \rho, \hat{\rho})$ vs. r (an approximation).

(*iv*) We see from Fig. 1, that the feedback scheme offers no improvement over the one-way scheme (that is, $E_N < E_1$, for $N > 2$) for $r^* \leqq r < c$ where r^* is the solution of $E_2 = E_1$, that is,

$$\tfrac{1}{2}E_1(2r^*, S_2) = E_1(r^*, \rho).$$

However, the rate $r^* \to c$ as $\hat{\rho} \to \infty$.

Actually, it is probably possible to improve on our results substantially and in particular bring about an increase in the error-exponent for all $r < c$. Let $\{N_1, N_2, \cdots, N_k\}$ be a set of positive integers (not necessarily equal). Then divide the $n_o = \alpha T$ forward channel uses into $\nu = n_o/(N_1 + N_2 + \cdots N_k)$ uses of an equivalent channel which is the parallel combination of k Gaussian channels with signal-to-noise ratios $S_{N_1}, S_{N_2}, \cdots, S_{N_k}$. These k Gaussian channels are generated by $N_1, N_2, \cdots, N_k$ iterations, respectively, of the Elias scheme. One must then compute the error-exponent for a parallel combination of channels to obtain a new improved exponent.[14] We leave this task as an open problem.

(*v*) Let us finally remark that although the expectation of the channel input power x^2 is constrained, the quantity x^2 is in fact a random variable distributed on the interval $[0, \infty)$. This is in contrast to the one-way

schemes where the channel input is bounded. This point is discussed in Ref. 13.

REFERENCES

1. Elias, P., "Channel capacity without coding," *Lectures on Communication System Theory,* New York: McGraw Hill, 1961, pp. 363–366.
2. Elias, P., "Networks of Gaussian channels with application to feedback systems," IEEE Trans. Inform. Theory, *IT-13,* No. 7 (July 1967), pp. 493–501.
3. Shannon, C. E., "Probability of error for optimal codes in a Gaussian channel," B.S.T.J., *38,* No. 3 (May 1959), pp. 611–656.
4. Butman, J., "Phase in coherent feedback communication," paper presented at 1969 Int. Symp. on Inform. Theory, Ellenville, New York.
5. Butman, S., "Optimum linear coding for additive noise systems using information feedback," Technical Rep. No. 1, Communication Theory Laboratory, California Institute of Technology, Pasadena, California, May 1967.
6. Kashyap, R. L., "Feedback coding schemes for an additive noise channel with a noisy feedback link," IEEE Trans. Inform. Theory, *IT-14,* No. 5 (May 1968), pp. 471–486.
7. Kramer, A. J., "Improving communication reliability by use of an intermittent feedback link," IEEE Trans. on Information Theory, *IT-15,* No. 1 (January 1969), pp. 52–60.
8. Lavenberg, S., "Feedback communication using orthogonal signals," IEEE Trans. on Information Theory, *IT-15,* No. 4 (July 1969), pp. 478–483.
9. Omura, J. K., "Optimum linear transmission of analog data for channels with feedback," IEEE Trans. Inform. Theory, *IT-14,* No. 1 (January 1968), pp. 38–43.
10. Schalkwijk, J. P. M., and Kailath, T., "A coding scheme for additive channels with feedback, part I," IEEE Trans. Inform. Theory, *IT-12,* No. 4 (April 1966), pp. 172–182.
11. Schalkwijk, J. P. M., "A coding scheme for additive channels with feedback, part II," IEEE Trans. Inform. Theory, *IT-12,* No. 4 (April 1966), pp. 183–189.
12. Wozencraft, J. M., and Jacobs, I. M., *Principles of Communication Engineering,* New York: Wiley, 1965, pp. 294–297.
13. Wyner, A. D., "On the Schalkwijk–Kailath coding scheme with a peak energy constraint," IEEE Trans. Inform. Theory, *IT-14,* No. 1 (January 1968), pp. 71–78.
14. Gallager, R., *Information Theory and Reliable Communication,* New York: Wiley, 1968, pp. 149–150.

Mode Conversion Caused by Surface Imperfections of a Dielectric Slab Waveguide

By DIETRICH MARCUSE

(Manuscript received May 8, 1969)

This paper contains a perturbation theory which is applicable to the scattering losses suffered by guided modes of a dielectric slab waveguide as a consequence of imperfections of the waveguide wall. The development of the theory occupies the bulk of the paper. Numerical results appear in Sections VI and VIII to which a reader less interested in the theory is referred.

The theory allows us to conclude that random deviations of the waveguide wall in the order of 1 percent, for guides designed to guide an optical wave of $\lambda_0 = 1\mu$ wavelength, can cause scattering losses of 10 percent per centimeter or 0.46 dB per centimeter. A systematic sinusoidal deviation of the waveguide wall can cause total exchange of energy from the lowest order to the first order guided mode in a distance of approximately 1 cm if the amplitude of the sinusoidal deviation from perfect straightness is only 0.5 percent of the thickness of the guide. An rms deviation of one of the waveguide walls of 9Å causes a radiation loss of 10 dB per kilometer (index difference 1 percent, guide width 2.5μ).

I. INTRODUCTION

The problem of how to transmit laser light over large distances or carry it short distances inside the laboratory has renewed the interest in dielectric waveguides.[1-5] Such waveguides usually used in the form of clad fibers or as strips of a medium of larger dielectric constant embedded in another dielectric medium are capable, in principle, of guiding electromagnetic radiation. By proper dimensioning, a dielectric waveguide can be made to transmit only one guided mode. In this respect mode guidance by dielectric waveguides resembles mode guidance by hollow metallic waveguides. Hollow metallic tubes can be constructed to allow only one mode to propagate so that mode conver-

sion (except for conversion to the reflected dominant mode) becomes impossible. Such truly single mode operation is impossible for dielectric waveguides since these guides can always lose electromagnetic energy to the continuous spectrum of unguided modes.

The possible solutions of Maxwell's equations for a dielectric waveguide consist of a discrete spectrum of a finite number of guided modes plus a continuum of waveguide modes.[6] The guided modes have field configurations which concentrate the electromagnetic energy inside and in the immediate vicinity of the structure. The continuum of unguided modes extends to infinite distances from the waveguide and consists of a superposition of incident and reflected waves. A convenient way of visualizing the physical significance of the continuum of unguided modes is as follows. If a plane wave is incident on the dielectric waveguide at an arbitrary angle, part of it penetrates the dielectric structure while some portion is reflected. The resulting superposition field of incident and reflected waves satisfies Maxwell's equations and the boundary conditions at the dielectric waveguide and as such can be viewed as a mode of the structure, but the energy of this mode is not concentrated near the waveguide and there are no specific restrictions on the projection of the propagation vector in the direction of the guide axis.

A perfect dielectric waveguide can transmit any of its guided modes without converting energy to any of the other possible guided modes or to the continuous spectrum. But any imperfection of the guide, such as a local change of its index of refraction or a deviation from perfect straightness or an imperfection of the interface between two regions with different index of refraction, couples the particular guided mode to all other guided modes as well as to all the modes of the unguided continuum. Imperfections of this type are unavoidable. They transfer energy from the desired guided mode to unwanted guided modes and the radiation field of the continuum of unguided modes, thus increasing the loss of the desired guided mode.

This paper gives a simple, approximate theory of the losses of dielectric waveguides, caused by imperfections of the boundary between the inner region of higher dielectric index and the surrounding outer region of the dielectric waveguide. Even though the method of analysis used here can be used to describe any arbitrary dielectric waveguide, we limit the discussion to a simple case. We describe the effects of mode conversion for a dielectric slab surrounded by vacuum, assuming for simplicity, that there is no variation of the dimensions or properties of the rod as well as the field distribution in one co-ordinate direction. The

restriction of demanding $\partial/\partial_y = 0$ for one of the co-ordinates y is no limitation on the method of analysis but is imposed strictly for convenience. It simplifies the analysis considerably without drastically changing the conclusions. The tolerance requirements based on our analysis are rather stringent. They show the order of magnitude of the losses which can be expected from deviations from perfect geometry. Additional variations in the direction considered perfect in this paper is unlikely to improve any of the loss predictions.

II. TE MODES OF A DIELECTRIC SLAB

Let us consider the transverse electric modes of the dielectric slab of Fig. 1. True to the simplifying assumption discussed in Section I, we assume

$$\frac{\partial}{\partial_y} = 0 \tag{1}$$

with y being the co-ordinate perpendicular to the x and z directions, but parallel to the slab. The only nonvanishing field components are E_y, H_x, and H_z.

Leaving the z and time dependence

$$e^{i(\omega t - \beta z)} \tag{2}$$

understood, we obtain the following modes of the ideal structure as a solution of Maxwell's equations satisfying the boundary conditions.

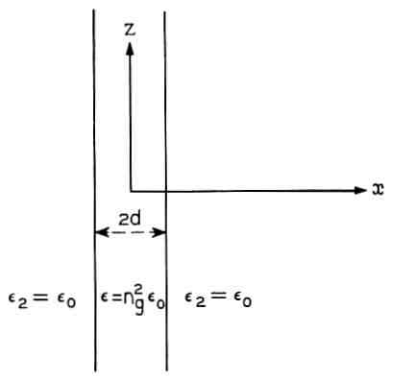

Fig. 1 — Geometry of a dielectric slab waveguide.

2.1 *Even Guided Modes*

For even guided modes

$$\mathcal{E}_y = A^{(e)} \cos \kappa x \quad \text{for} \quad |x| \leqq d, \tag{3a}$$

$$\mathcal{E}_y = A^{(e)} \cos \kappa d e^{-\gamma(x-d)} \quad \text{for} \quad x \geqq d, \tag{3b}$$

$$\mathcal{H}_x = -\frac{i}{\omega\mu} \frac{\partial \mathcal{E}_y}{\partial z}, \tag{4}$$

$$\mathcal{H}_z = \frac{i}{\omega\mu} \frac{\partial \mathcal{E}_y}{\partial x}, \tag{5}$$

The field component $\mathcal{E}_y$ satisfies the wave equation

$$\frac{\partial^2 \mathcal{E}_y}{\partial x^2} + \frac{\partial^2 \mathcal{E}_y}{\partial z^2} + n_0^2 k^2 \mathcal{E}_y = 0. \tag{6}$$

The value of the index of refraction n_0 is different inside and outside of the dielectric slab. For simplicity, we assume

$$n_0 = 1 \quad \text{for} \quad |x| > d. \tag{7}$$

The other constants are related as follows

$$k^2 = \omega^2 \epsilon_0 \mu_0, \tag{8}$$

$$\kappa = (n^2 k^2 - \beta^2)^{\frac{1}{2}}, \tag{9}$$

$$\gamma = (\beta^2 - k^2)^{\frac{1}{2}}. \tag{10}$$

The propagation constant β is obtained as a solution of the eigenvalue equation

$$\tan \kappa d = \frac{\gamma}{\kappa}. \tag{11}$$

The mode amplitude A can be expressed in terms of the power P carried by the mode.

$$P = \tfrac{1}{2} \operatorname{Re} \int_{-\infty}^{\infty} (-\mathcal{E}_y \mathcal{H}_x^*)\, dx = \frac{\beta}{\omega\mu} \int_0^{\infty} |\mathcal{E}_y|^2\, dx. \tag{12}$$

P is the power per unit length (unit length in y-direction) flowing along the z-axis. We obtain for the amplitude coefficient

$$A^{(e)2} = \frac{2\omega\mu}{\beta d + \dfrac{\beta}{\gamma}} P. \tag{13}$$

2.2 *Even Modes of the Continuum*

The continuum of unguided modes of even symmetry is given by the equations:

$$\mathcal{E}_y = B^{(e)} \cos \sigma x \quad \text{for} \quad |x| \leqq d, \tag{14a}$$

$$\mathcal{E}_y = C^{(e)} e^{i\rho x} + D^{(e)} e^{-i\rho x} \quad \text{for} \quad x \geqq d. \tag{14b}$$

The other field components follow again from equations (4) and (5) and $\mathcal{E}_y$ is a solution of equation (6). The constants are related to each other by the equations

$$\sigma = (n^2 k^2 - \beta^2)^{\frac{1}{2}}, \tag{15}$$

$$\rho = (k^2 - \beta^2)^{\frac{1}{2}}. \tag{16}$$

The radial propagation constant ρ can assume all values from 0 to ∞. The continuous mode spectrum starts at $\beta = k$ and continuous to $\beta = 0$ at which point we have $\rho = k$. Larger values of ρ are obtained for imaginary values of β corresponding to modes of the continuum exhibiting a cutoff behavior.

The boundary conditions do not lead to an eigenvalue equation for β but they determine $C^{(e)}$ and $D^{(e)}$ in relation to $B^{(e)}$.

$$C^{(e)} = \tfrac{1}{2} B^{(e)} e^{-i\rho d} \left(\cos \sigma d + i \frac{\sigma}{\rho} \sin \sigma d \right), \tag{17}$$

$$D^{(e)} = C^{(e)*}, \tag{18}$$

(the asterisk indicates the complex conjugate quantity).

The normalization of the modes of the continuum involves a δ-function. Instead of equation (12) we use

$$P\, \delta(\rho - \rho') = \frac{\beta}{\omega\mu} \int_0^\infty \mathcal{E}_y(\rho)\, \mathcal{E}_y^*(\rho')\, dx. \tag{19}$$

With this normalization we get

$$B^{(e)^2} = \frac{2\omega\mu P}{\pi\beta \left(\cos^2 \sigma d + \frac{\sigma^2}{\rho^2} \sin^2 \sigma d \right)}. \tag{20}$$

2.3 *Odd Guided Modes*

In a manner similar to that for obtaining the preceding equations we obtain the equations for the odd guided modes

$$\mathcal{E}_y = A^{(0)} \sin \kappa x \quad \text{for} \quad x \leqq d, \tag{21a}$$

$$\mathcal{E}_y = A^{(0)} \sin \kappa d e^{-\gamma(x-d)} \quad \text{for} \quad x \geqq d. \tag{21b}$$

Equations (4) through (10) apply to the odd modes unaltered. The eigenvalue equation is given by

$$\tan \kappa d = -\frac{\kappa}{\gamma}, \tag{22}$$

and the mode normalization is

$$A^{(0)2} = \frac{2\omega\mu}{\beta d + \frac{\beta}{\gamma}} P. \tag{23}$$

2.4 *Odd Modes of the Continuum*

As in Section 2.3 we obtain the equations for the odd modes of the continuum

$$\mathcal{E}_y = B^{(0)} \sin \sigma x \quad \text{for} \quad |x| \leqq d, \tag{24a}$$

$$\mathcal{E}_y = C^{(0)} e^{i\rho x} + D^{(0)} e^{-i\rho x} \quad \text{for} \quad x \geqq d, \tag{24b}$$

$$C^{(0)} = \tfrac{1}{2} B^{(0)} e^{-i\rho d}\left(\sin \sigma d - i \frac{\sigma}{\rho} \cos \sigma d\right), \tag{25}$$

$$D^{(0)} = C^{(0)*}, \tag{26}$$

$$B^{(0)2} = \frac{2\omega\mu P}{\pi\beta\left(\sin^2 \sigma d + \frac{\sigma^2}{\rho^2} \cos^2 \sigma d\right)}. \tag{27}$$

All these modes are orthogonal to one another. The even modes are orthogonal to all the odd modes, the guided modes are orthogonal to all the modes of the continuum, and all guided modes as well as all modes of the continuum are orthogonal among each other. The orthogonality of the modes of the continuum among each other was already expressed by equation (19). Labeling the discrete modes by indices and dropping the vector component label y we can express the orthogonality of the discrete modes by the equation

$$P\ \delta_{nm} = \frac{\beta_m}{2\omega\mu} \int_{-\infty}^{\infty} \mathcal{E}_n \mathcal{E}_m^* \, dx. \tag{28}$$

III. MODE COUPLING CAUSED BY IMPERFECTIONS

We want to study the losses which the lowest order guided mode suffers because of imperfections of the waveguide wall. A dielectric waveguide with wall imperfections is shown in Fig. 2.

Fig. 2 — Dielectric slab waveguide with wall distortions.

The waveguide with wall imperfections is mathematically described by a refractive index distribution

$$n^2(x, z) = n_0^2(x, z) + \Delta n^2(x, z). \tag{29}$$

The index distribution

$$n_0^2(x, z) = \begin{cases} n_g^2 & |x| < d \\ 1 & |x| > d \end{cases} \tag{30}$$

describes the ideal dielectric waveguide whose TE modes were given in the Section II. The additional term Δn^2 describes how the guide deviates from its perfect shape. Consider a deviation shown in Fig. 3. The corresponding distribution Δn^2 is (n_g = index of refraction of the dielectric material of the guide)

$$\Delta n^2 = \begin{cases} 0 \begin{cases} x < d & \text{if } d < f(z) \\ x < f(z) & \text{if } d > f(z) \end{cases} \\ n_g^2 - 1 \quad d < x < f(z) \quad \text{if } d < f(z) \\ -(n_g^2 - 1) \quad f(z) < x < d \quad \text{if } d > f(z) \\ 0 \begin{cases} x > f(z) & \text{if } d < f(z) \\ x > d & \text{if } d > f(z) \end{cases} \end{cases}. \tag{31}$$

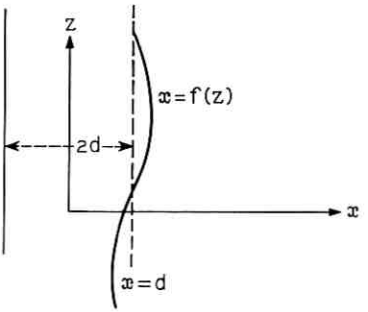

Fig. 3 — Illustration of the wall distortion function $f(z)$.

The field distribution E_y of this waveguide is a solution of

$$\frac{\partial^2 E_y}{\partial x^2} + \frac{\partial^2 E_y}{\partial z^2} + (n_0^2 + \Delta n^2)k^2 E_y = 0 \tag{32}$$

with H_x and H_z given by equations (4) and (5). The modes of the perfect waveguide form a complete orthogonal set for all TE modes with no variation in the y-direction. It is, therefore, possible to express any field distribution on the waveguide with imperfect walls by the expansion

$$E_y = \sum_n C_n(z)\,\mathcal{E}_n + \sum \int_0^\infty g(\rho, z)\,\mathcal{E}(\rho)\,d\rho. \tag{33}$$

The first summation extends over all even and odd modes of the discrete spectrum of guided modes. The integral extends over all modes of the continuum, and the summation sign in front of the integral indicates summation over even and odd modes. The expansion coefficients C_n and $g(\rho)$ are unknown functions of z.

To obtain a coupled system of differential equations for the expansion coefficients we substitute equation (33) into equation (32). Multiplying the resulting equation by

$$\frac{\beta_m}{2\omega\mu}\,\mathcal{E}_m^*,$$

integrating over x from $-\infty$ to $+\infty$, and using the orthogonality relations and the fact that $\mathcal{E}_n$ and $\mathcal{E}(\rho)$ are the (discrete and continuous) modes of the perfect guide leads to

$$\frac{\partial^2 C_m}{\partial z^2} - 2i\beta_m \frac{\partial C_m}{\partial z} = F_m(z) \tag{34}$$

with

$$F_m(z) = -\frac{\beta_m k^2}{2\omega\mu P}\left[\sum_n C_n(z)\int_{-\infty}^{\infty} \mathcal{E}_m^*\,\Delta n^2\,\mathcal{E}_n\,dx + \sum\int_0^{\infty} d\rho g(\rho, z)\int_{-\infty}^{\infty}\mathcal{E}_m^*\,\Delta n^2\,\mathcal{E}(\rho)\,dx\right]. \tag{35}$$

Similarly multiplying by

$$\frac{\beta'}{2\omega\mu}\,\mathcal{E}^*(\rho')$$

leads to

$$\frac{\partial^2 g(\rho')}{\partial z^2} - 2i\beta'\frac{\partial g(\rho')}{\partial z} = G(\rho' z) \tag{36}$$

with

$$G(\rho', z) = -\frac{\beta' k^2}{2\omega\mu P}\left[\sum_n C_n(z)\int_{-\infty}^{\infty}\mathcal{E}^*(\rho')\,\Delta n^2\,\mathcal{E}_n\,dx + \sum\int_0^{\infty} d\rho g(\rho, z)\int_{-\infty}^{\infty}\mathcal{E}^*(\rho')\,\Delta n^2\,\mathcal{E}(\rho)\,dx\right]. \tag{37}$$

No n-label on the power term P is necessary since we assume that all the normal modes are normalized to the same amount of power. The actual power carried by each mode relative to the power of the other modes is given by the C_n coefficients. Solutions of equations (34) and (35) with appropriate initial conditions provide us with exact solutions of the imperfect waveguide. It is interesting to note that this method of solution does not require the consideration of boundary conditions.

The normal modes $\mathcal{E}_n$ and $\mathcal{E}(\rho)$ were assumed to have the time and z-dependence of equation (2); this means they represent waves traveling in the positive z-direction. However, the solutions of equations (34) and (36) introduce waves traveling in positive as well as negative z-direction. To see this, let us assume that $\Delta n^2 = 0$ so that $F_n(z) = 0$. The equation

$$\frac{\partial^2 C_n}{\partial z^2} - 2i\beta_n\frac{\partial C_n}{\partial z} = 0 \tag{38}$$

has the solution

$$C_n(z) = A + Be^{2i\beta_n z} \tag{39}$$

with constant A and B. The product of A with $\mathcal{E}_n$ results in a wave traveling in the positive z-direction but the product of B exp $(2i\beta_n z)$

with ε_n results in a wave traveling in the negative z-direction. So, even though we started out with waves traveling in the positive z-direction the expansion (33) contains partial waves traveling in positive as well as negative z-direction.

For the purpose of obtaining perturbation solutions of equations (34) and (36), an integral form of these equations is more useful. Treating equations (34) and (36) as inhomogeneous differential equations, we can immediately write the following integral equations

$$C_m = A_m + B_m e^{2i\beta_m z} + \frac{1}{2i\beta_m}\int_0^z [e^{2i\beta_m(z-\zeta)} - 1]F_m(\zeta)\,d\zeta, \tag{40}$$

$$g(\rho', z) = C(\rho') + D(\rho')e^{2i\beta' z} + \frac{1}{2i\beta'}\int_0^z [e^{2i\beta'(z-\zeta)} - 1]G(\rho', \zeta)\,d\zeta. \tag{41}$$

It is important to know which part of equations (40) and (41) is associated with waves traveling in the positive or negative z-direction. Therefore, we introduce the notation.

$$C_m = C_m^{(+)} + C_m^{(-)} \tag{42}$$

with

$$C_m^{(+)}(z) = A_m - \frac{1}{2i\beta_m}\int_0^z F_m(\zeta)\,d\zeta, \tag{43}$$

$$C_m^{(-)}(z) = \left\{B_m + \frac{1}{2i\beta_m}\int_0^z e^{-2i\beta_m\zeta}F_m(\zeta)\,d\zeta\right\}e^{2i\beta_m z}. \tag{44}$$

The superscript $(+)$ indicates the coefficient which after substitution into equation (33) produces waves traveling in positive z-direction, while $(-)$ indicates the part which produces waves traveling in negative z-direction. A similar notation and resulting equations is used for $g(\rho', z)$; however, the corresponding equations are obvious and are therefore omitted.

The constants A_m, B_m, and so on, occurrng in equations (43), (44), and the corresponding equations for $g(\rho', z)$ must be determined from initial conditions. We always assume that the lowest order guided mode is incident on the imperfect waveguide at $z = 0$. Using the subscript 0 for this incident mode we get immediately from equation (43)

$$C_m^{(+)} = 0 \quad \text{for} \quad m \neq 0 \qquad \text{at} \quad z = 0$$

or

$$A_m = 0 \quad \text{for} \quad m \neq 0, \tag{45}$$

but

$$A_0 = 1. \tag{46}$$

We imagine that at $z = L$ the waveguide is connected to a perfect guide so that at that point there are no waves traveling in negative z-direction. This leads to the condition

$$B_m = -\frac{1}{2i\beta_m} \int_0^L e^{-2i\beta_m \zeta} F_m(\zeta)\, d\zeta \tag{47}$$

for all values of m. The power loss ΔP of the incident mode due to mode conversion is given by

$$\frac{\Delta P}{P} = \sum_{n=1}^{\infty} [\,|\, C_n^{(+)}(L) \,|^2 + |\, C_n^{(-)}(0) \,|^2] + \sum \int_0^{\infty} [\,|\, g^{(+)}(\rho, L) \,|^2 + |\, g^{(-)}(\rho, 0) \,|^2]\, d\rho. \tag{48}$$

Equation (48) states that the total power lost by mode conversion from the incident mode escapes at $z = L$ in spurious modes traveling in position z-direction and at $z = 0$ in spurious modes traveling in negative z-direction. The factor P is the normalized power factor of equations (12) and (19); it is the power incident in mode 0. Notice that because of equations (45) and (47) only the integral terms of equations (43) and (44) (taken from $z = 0$ to $z = L$) enter into equation (48).

The integral equations (43) and (44) can only be solved approximately. We perform first order perturbation theory by using $C_m(0)$ instead of $C_m(z)$ and $g(\rho, 0)$ instead of $g(\rho, z)$ in equations (35) and (37). Furthermore, we realize that $C_m^{(-)}(0)$ for all m is a quantity of first order and will therefore be neglected in equations (35) and (37). The same is true for $C_m^{(+)}(0)$ with $m \neq 0$. In the spirit of first order perturbation theory we use therefore

$$C_m = \delta_{0m} \tag{49}$$

and

$$g(\rho) = 0 \tag{50}$$

in equations (35) and (37).

The perturbation theory is feasible not only when $n_g^2 - 1 \ll 1$ but also when $n_g^2 - 1$ is arbitrarily large but the geometrical deviation of the guide walls from perfect straightness is slight. In either case we obtain from equations (35) and (37) the simple approximations

$$F_m(z) = -\frac{\beta_m k^2}{2\omega\mu P}(n_g^2 - 1)\{[f(z) - d]\,\mathcal{E}_0(d, z)\,\mathcal{E}_m^*(d, z) - [h(z) + d]\,\mathcal{E}_0(-d, z)\,\mathcal{E}_m^*(-d, z)\}, \quad (51)$$

$$G(\rho, z) = -\frac{\beta k^2}{2\omega\mu P}(n_g^2 - 1)\{[f(z) - d]\,\mathcal{E}^*(\rho, d, z)\,\mathcal{E}_0(d, z) - [h(z) + d]\,\mathcal{E}^*(\rho, -d, z)\,\mathcal{E}_0(-d, z)\}. \quad (52)$$

The function $f(z)$ describes the dielectric-air interface in the vicinity of $x = d$, while $h(z)$ describes it near $z = -d$. We assumed that $f(z)$ and $h(z)$ depart so little from $x = d$ and $x = -d$ that the functions $\mathcal{E}(x, z)$ could be replaced by $\mathcal{E}(\pm d, z)$.

IV. EVALUATION OF THE SPURIOUS MODE AMPLITUDES

We begin the discussion of the consequences of our scattering theory by calculating the coefficients $C_m^{(+)}$ and $g^{(+)}$. We obtain [from equations (43) and (51) with the help of equations (3a) and (13) for the even modes] the following

$$C_{me}^{(+)}(L) = \frac{Lk^2}{2i}(n_g^2 - 1)\frac{\cos\kappa_0 d\,\cos\kappa_m d}{\left[\left(\beta_0 d + \frac{\beta_0}{\gamma_0}\right)\left(\beta_m d + \frac{\beta_m}{\gamma_m}\right)\right]^{\frac{1}{2}}}(\varphi_m - \psi_m). \quad (53)$$

The coefficients φ_m and ψ_m are defined by

$$\varphi_m = \frac{1}{L}\int_0^L [f(z) - d]e^{-i(\beta_0 - \beta_m)z}\,dz \quad (54)$$

and

$$\psi_m = \frac{1}{L}\int_0^L [h(z) + d]e^{-i(\beta_0 - \beta_m)z}\,dz. \quad (55)$$

These are the Fourier coefficients of the functions $f(z) - d$ and $h(z) + d$ which are expanded in a domain

$$0 \leqq z \leqq L.$$

The amplitude of the mth even mode depends on the Fourier components of the wall function whose "spatial frequency" Γ is

$$\Gamma_m = \frac{2\pi}{\Lambda_m} = \beta_0 - \beta_m\,. \quad (56)$$

The corresponding expression for the even modes of the continuous

spectrum is:

$$g_e^{(+)}(\rho, L) = \frac{Lk^2}{2i(\pi)^{\frac{1}{2}}}(n_g^2 - 1)\frac{\cos \kappa_0 d \cos \sigma d[\varphi(\beta) - \psi(\beta)]}{\left[\left(\beta_0 d + \frac{\beta_0}{\gamma_0}\right)\beta\left(\cos^2 \sigma d + \frac{\sigma^2}{\rho^2}\sin^2 \sigma d\right)\right]^{\frac{1}{2}}} \quad (57)$$

with [$\beta = \beta(\rho)$ see equation (16)]

$$\varphi(\beta) = \frac{1}{L}\int_0^L [f(z) - d]e^{-i(\beta_0 - \beta)z}\, dz, \quad (58)$$

$$\psi(\beta) = \frac{1}{L}\int_0^L [h(z) + d]e^{-i(\beta_0 - \beta)z}\, dz. \quad (59)$$

The corresponding expressions for the odd modes are

$$C_{n0}^{(+)}(L) = \frac{Lk^2}{2i}(n_g^2 - 1)\frac{\cos \kappa_0 d \sin \kappa_n d}{\left[\left(\beta_0 d + \frac{\beta_0}{\gamma_0}\right)\left(\beta_n d + \frac{\beta_n}{\gamma_n}\right)\right]^{\frac{1}{2}}}(\varphi_n + \psi_n), \quad (60)$$

$$g_0^{(+)}(\rho, L) = \frac{Lk^2}{2i(\pi)^{\frac{1}{2}}}(n_g^2 - 1)\frac{\cos \kappa_0 d \sin \sigma d[\varphi(\beta) + \psi(\beta)]}{\left[\left(\beta_0 d + \frac{\beta_0}{\gamma_0}\right)\beta\left(\sin^2 \sigma d + \frac{\sigma^2}{\rho^2}\cos^2 \sigma d\right)\right]^{\frac{1}{2}}}. \quad (61)$$

The Fourier coefficients φ and ψ are given by equations (54), (55), (58), and (59) except that β_n and β are now the propagation constants of the odd modes.

The corresponding expressions for $C^{(-)}$ and $g^{(-)}$ are obtained by replacing β_m with $-\beta_m$ and β with $-\beta$ in equations (54), (55), (56), (58), and (59).

V. SINUSOIDAL WALL DEFLECTIONS

As a specific example, let us assume that the wall imperfections have sinusoidal shape. Then

$$f(z) - d = a \sin \theta z \quad (62)$$

and

$$h(z) + d = -a \sin (\theta z + \alpha). \quad (63)$$

The phase factor α allows us to consider either a waveguide whose width varies sinusoidally

$$\alpha = 0, \quad (64)$$

or one whose direction changes sinusoidally

$$\alpha = \pi. \tag{65}$$

We obtain from equation (54) with

$$\theta = \beta_0 - \beta_m \tag{66}$$

the Fourier component

$$\varphi_m = \frac{a}{2i} \tag{67}$$

and from equation (55)

$$\psi_m = -\frac{a}{2i} e^{i\alpha}. \tag{68}$$

A term of the order $a/L \ll 1$ was omitted in equations (67) and (68). It is apparent that only one spurious mode is excited by the sinusoidal wall deflection since condition (66) can be satisfied for only one value of β_m. If condition (66) is not satisfied, φ_m and ψ_m are of the order of $a/L \ll 1$. The fractional power scattered into one spurious guided mode due to a sinusoidal wall irregularity is [from equations (48), (53), (67) and (68)]

$$\left(\frac{\Delta P}{P}\right)_{eo} \quad \frac{L^2a^2k^4}{4}(n_g^2 - 1)^2 \frac{\cos^2 \kappa_0 d \cos^2 \kappa_m d}{\left(\beta_0 d + \frac{\beta_0}{\gamma_0}\right)\left(\beta_m d + \frac{\beta_m}{\gamma_m}\right)} \cos^2 \frac{\alpha}{2} \tag{69}$$

for even modes or [from equations (48) and (60)]

$$\left(\frac{\Delta P}{P}\right)_{oo} \quad \frac{L^2a^2k^2}{4}(n_g^2 - 1)^2 \frac{\cos^2 \kappa_0 d \sin^2 \kappa_m d}{\left(\beta_0 d + \frac{\beta_0}{\gamma_0}\right)\left(\beta_m d + \frac{\beta_m}{\gamma_m}\right)} \sin^2 \frac{\alpha}{2} \tag{70}$$

for odd modes. However only one even or one odd mode can be excited by one particular sinusoidal wall deviation since it is impossible to satisfy the "resonance" condition (66) for more than one mode simultaneously.

If $\alpha = 0$, that is if the width of the guide changes sinusoidally, only even modes can be excited while sinusoidal deviations from straightness ($\alpha = \pi$) couple the even fundamental mode only to odd spurious modes. It must also be noticed that for a long period length

$$\Lambda = \frac{2\pi}{\theta} \tag{71}$$

equation (66) can be satisfied only for forward scattering modes. To couple to backward scattering modes, the period length D must be approximately equal to half the wavelength of the guided modes. The fact that only one spurious mode is coupled to the incident mode by sinusoidal wall imperfections (it can be shown that the coupling to the continuous mode spectrum is also weak if one guided mode can couple strongly) allows us to give a much better description of the coupling process.

Since the mode amplitudes C_m can change only slowly in the distance of one wavelength we can neglect the second derivative of C_m in equation (34). Labeling the incident mode 0 and the one coupled spurious mode 1 we can write the equation system (34) in the following form

$$\frac{\partial C_0}{\partial z} = -\kappa_{01} C_1 , \tag{72}$$

$$\frac{\partial C_1}{\partial z} = \kappa_{01}^* C_0 , \tag{73}$$

with

$$\kappa_{01} = \frac{k^2 a}{2}(n_v^2 - 1)\frac{\cos \kappa_0 d \cos \kappa_1 d}{\left[\left(\beta_0 d + \frac{\beta_0}{\gamma_0}\right)\left(\beta_1 d + \frac{\beta_1}{\gamma_1}\right)\right]^{\frac{1}{2}}} \exp\left(i\frac{\alpha}{2}\right)\cos\frac{\alpha}{2}. \tag{74}$$

The coupling coefficient κ_{01} of equations (74) holds for coupling from an even mode 0 to an even mode 1. The case of coupling from an even mode 0 to an odd mode 1 can be treated similarly. In fact, except for an unimportant phase factor, we get it from $(1/L)[(\Delta P/P)_{oo}]^{\frac{1}{2}}$ of equarion (70). In equation (72) we omitted a term with C_0 on the right-hand side, and similarly a term with C_1 was omitted in equation (73). These terms would be multiplied by sinusoidally varying functions and would describe the local change of phase velocity as the guide dimensions vary. These terms give no contribution if we use an average over C_0 and C_1 over the mechanical period length of equation (71).

Assuming $C_0 = 1$, $C_1 = 0$ at $z = 0$ the equation system (72) and (73) has the solution

$$C_0 = \cos |\kappa_{01}| z, \tag{75}$$

$$C_1 = \left(\frac{\kappa_{01}^*}{\kappa_{01}}\right)^{\frac{1}{2}} \sin |\kappa_{01}| z. \tag{76}$$

Total exchange of energy is possible between the two coupled modes.

The distance D over which all the energy is exchanged is given by

$$D = \frac{\pi}{2 \mid \kappa_{01} \mid}. \tag{77}$$

Finally, we need the power loss to the modes of the continuous spectrum. From equations (48), (57), (61), (62), and (63) we obtain

$$\left(\frac{\Delta P}{P}\right)_c = \frac{a^2k^4}{\pi} (n_g^2 - 1)^2 \frac{\cos^2 \kappa_0 d}{\beta_0 d + \frac{\beta_0}{\gamma_0}}$$

$$\cdot \int_0^\infty \left[\frac{\cos^2 \sigma d \cos^2 \frac{\alpha}{2}}{\beta\left(\cos^2 \sigma d + \frac{\sigma^2}{\rho^2} \sin^2 \sigma d\right)} + \frac{\sin^2 \sigma d \sin^2 \frac{\alpha}{2}}{\beta\left(\sin^2 \sigma d + \frac{\sigma^2}{\rho^2} \cos^2 \sigma d\right)} \right]$$

$$\cdot \frac{\sin^2 [\theta - (\beta_0 - \beta)] \frac{L}{2}}{[\theta - (\beta_0 - \beta)]^2} d\rho. \tag{78}$$

The integration can be performed easily if one realizes that for large values of L only a very narrow region in the β range near $\beta - \beta_0 - \theta$ contributes to the integral. We consider all functions in the integrand as constant in this very narrow range and take them out of the integral with the exception of

$$\left\{ \frac{\sin [\theta - (\beta_0 - \beta)] \frac{L}{2}}{\theta - (\beta_0 - \beta)} \right\}^2.$$

This remaining integral can easily be performed if we use equation (16) to obtain

$$d\rho = -\frac{\beta}{\rho} d\beta.$$

Following this procedure yields

$$\left(\frac{\Delta P}{P}\right)_c = \frac{La^2k^4}{2} (n_g^2 - 1)^2 \frac{\cos^2 \kappa_0 d}{\beta_0 d + \frac{\beta_0}{\gamma_0}}$$

$$\cdot \left[\frac{\rho \cos^2 \sigma d \cos^2 \frac{\alpha}{2}}{\rho^2 \cos^2 \sigma d + \sigma^2 \sin^2 \sigma d} + \frac{\rho \sin^2 \sigma d \sin^2 \frac{\alpha}{2}}{\rho^2 \sin^2 \sigma d + \sigma^2 \cos^2 \sigma d} \right]. \tag{79}$$

The parameters σ and ρ follow from equations (15) and (16) with

$$\beta = \beta_0 - \theta. \tag{80}$$

Equation (79) holds only for $\beta < k$; we get $\Delta P/P = 0$ for $\beta > k$. The most interesting aspect of equation (79) is its linear dependence on L. The scattering loss due to the modes of the continuous spectrum acts like a true loss process. By contrast, the corresponding equation (69) for the loss to guided modes is proportional to L^2 because coupling to a guided mode does not result in loss of energy but results in energy exchange between the two coupled modes. Energy loss to one of the guided modes is followed by energy gain when the energy exchange has reversed itself.

VI. NUMERICAL EXAMPLES FOR SINUSOIDAL IMPERFECTIONS

A few numerical examples resulting from equations (74) and (77) are listed in Table I. Two different values of the index of refraction n_g have been assumed, and for each value of the index three different values of $kd = 2\pi(d/\lambda_0)$ have been chosen so that one, two, or three guided modes can exist simultaneously. The mode with β_0 is the lowest

TABLE I—NUMERICAL EXAMPLES FOR SINUSOIDAL IMPERFECTIONS

n_g	kd	$\beta_0 d$	$\beta_1 d$	$\beta_2 d$	$\frac{aD}{d^2}$	Remarks
1.5	1.3	1.729	—	—	—	Single mode operation
	1.8	2.495	1.916	—	6.98	0 − 1 coupling $\alpha = \pi$
	3.0	4.336	3.831	3.051	6.17	0 − 1 coupling $\alpha = \pi$
					5.52	0 − 2 coupling $\alpha = 0$
1.01	8.0	8.041	—	—	—	Single mode operation
	15.0	15.113	15.022	—	42.54	0 − 1 coupling $\alpha = \pi$
	23.0	23.199	23.112	23.002	36.28	0 − 1 coupling $\alpha = \pi$
					43.69	0 − 2 coupling $\alpha = 0$

order even guided mode which is assumed to be incident on the waveguide with sinusoidal wall imperfections. This mode couples to the first odd mode with β_1 or the next even mode with β_2. The values for the normalized, dimensionless quantitity $(aD)/d^2$ [a = amplitude of the sinusoidal wall deviation according to equation (62) and (63), d = half width of the guide, and D = energy exchange length] have been obtained with the assumption that equation (66) is satisfied for the two modes which are coupled together. Coupling from mode 0 to mode 1 is considered only for the case of sinusoidal straightness deviations of the waveguide ($\alpha = \pi$) while coupling between even modes 0 to 2 is considered only for sinusoidal changes of the thickness of the waveguide ($\alpha = 0$). It is immediately apparent from Table I that the energy exchange length D is shorter for a guide with larger values of the refractive index.

To obtain a feeling for the numbers involved in this mode coupling phenomenon, let us assume that $n_g = 1.5$ and that the free space wavelength is $\lambda_0 = 1\mu$. The value of $kd = 1.8$ corresponds to $d = 0.286\mu$. To achieve total exchange of energy between modes 0 and 1 in $D = 1$ cm requires the extremely small amplitude $a = 5.72\ 10^{-5}\mu$ or $a = 0.572$ Å![†] The length of the mechanical period in this example is $\Lambda = 3.1\mu$.

Next, let us assume that the index of refraction is $n_g = 1.01$. Using again, $\lambda_0 = 1\mu$, we obtain from $kd = 15.0$ the value $d = 2.39\mu$ for the half width of the waveguide. Requiring again, $D = 1$ cm, we find $a = 243$ Å.

We can look at this problem in a different way. It is unlikely that any optical waveguide has a strictly sinusoidal deviation from perfect straightness. In fact, the numbers just presented show that it would be impossible to produce such a waveguide intentionally. However, we have seen [equation (53)] that the mode conversion between two guided modes is produced by a Fourier component of the actual deviation function. It is therefore not necessary to have a strictly sinusoidal straightness deviation. Any arbitrary deviation from straightness can be decomposed into a Fourier series and the Fourier component at the mechanical frequency which satisfies equation (66) is responsible for the coupling. In the more general case of arbitrary straightness deviations, there can be no complete exchange of energy between any two modes since power loss to other guided modes and the continuous

† A mechanical period of a fraction of an Angstrom is somewhat unphysical due to the granular nature of matter. However, this result can be restated to say that complete power conversion occurs in 0.1 mm if the amplitude is $a = 57.2$ Å.

spectrum of modes compete with each other since all of them are coupled simultaneously.

We can now ask the question: What amplitude of the mechanical straightness deviation is required to transfer 10 percent of power from mode 0 to mode 1 in a distance of $L = 1$ cm? Again, we use the previous examples. From equation (76) [or directly from equations (53) and (77)] we obtain

$$\frac{\Delta P}{P} = |\kappa_{01}|^2 L^2 = \frac{\pi^2}{4}\frac{L^2}{D^2}.$$

For the first example we obtain with $n_g = 1.5$, $\Delta P/P = 0.1$, $d = 0.286\mu$, and $aD/d^2 = 6.98$ the value $a = 0.115$ Å.† This result shows that if the Fourier component of the mechanical straightness deviation with a period length of 3.1μ is $a = 0.12$ Å (measured over a distance of 1 cm) the power loss caused by mode conversion to the first odd mode is 10 percent.

For the second example, we use again $n_g = 1.01$, $\Delta P/P = 0.1$, $d = 2.39\mu$, and $aD/d^2 = 42.54$ and obtain $a = 48.8$ Å. The important Fourier component in this case has a period of $\Lambda = 135\mu$. The power loss to the modes of the continuous spectrum caused by a sinusoidal change in thickness of the waveguide (which is very similar to its effect as a straightness deviation) can be calculated from equation (79) with $\alpha = 0$.

Let us consider only one case, $n_g = 1.01$, $kd = 15$, $\Lambda/d = 25$. For these values we obtain from equation (79)

$$\frac{d^3}{a^2 L}\frac{\Delta P}{P} = 4.6 \times 10^{-2}.$$

Assuming again $\Delta P/P = 0.1$ for a guide length $L = 1$ cm, we obtain with $d = 2.39\mu$

$$a = 5.46 \times 10^{-2}\mu = 546 \text{ Å}.$$

This number can be compared to the value $a = 48.8$ Å which gave 10 percent loss by conversion to one guided mode. However, for a meaningful comparison, we must remember that all the Fourier components of a Fourier expansion of the guide imperfections scatter power into the modes of the continuous spectrum. The total loss would have to be obtained by integrating the scattering loss over the spectral dis-

† Again it is more reasonable to restate this example to say that 10 percent loss occurs over a distance of $L = 0.1$ mm if $a = 12$ Å.

tribution of the Fourier components of the mechanical Fourier spectrum. Instead of doing this integration we use a different approach in Section VII.

VII. STATISTICAL TREATMENT OF WALL IMPERFECTIONS†

Equation (48) gives the relative loss of a guided mode caused by a definite (deterministic) distortion of the boundary of a dielectric waveguide. A quantity that may be even more interesting is the average of equation (48) taken over an ensemble of statistically identical systems.

For simplicity, let us assume that one wall of the waveguide is perfect while the other is randomly distorted. If both walls are randomly distorted, with no correlation between the distortions on opposite walls the loss value doubles compared to the case of only one wall being distorted. If the distortions on opposite sides of the waveguide are perfectly correlated the amount of loss is at most increased four times. So to simplify the discussion we assume

$$h(z) + d = 0. \tag{81}$$

In order to be able to calculate $\langle \Delta P/P \rangle_{av}$, we must evaluate

$$\langle | \varphi_m |^2 \rangle_{av} = \frac{1}{L^2} \int_0^L dz \int_0^L dz' R(z - z') e^{-i(\beta_0 - \beta_m)(z - z')} \tag{82}$$

We assumed that the correlation function

$$R(z - z') = \langle [f(z) - d][f(z') - d] \rangle_{av} \tag{83}$$

depends only on the difference between the coordinates z and z' but not on their individual values.

A change of integration variables allows us to write

$$\langle | \varphi_m |^2 \rangle_{av} = \frac{2}{L^2} \int_0^L (L - u) R(u) \cos (\beta_0 - \beta_m) u \, du. \tag{84}$$

To obtain equation (84) we made use of the fact that $R(u)$ is an even function.

The particular form of $R(u)$ depends on the statistics of the wall imperfections. However, all correlation functions have two features in common. They all have their maximum value at $u = 0$ and decrease to zero as $u \to \infty$. If $R(u)$ would not become 0 as $u \to \infty$ there would be a

† An excellent statistical treatment of random coupling effects in metallic waveguides can be found in Ref. 7.

systematic distortion of the waveguide boundary instead of the assumed random behavior. To get an idea of what one might expect, we assume the following form for the correlation function

$$R(u) = A^2 \exp\left(-\frac{|u|}{B}\right). \tag{85}$$

A is the rms deviation of the wall from perfect straightness and B is the correlation length. Using equation (85) we obtain from equation (84)

$$\langle |\varphi_m|^2 \rangle_{av} = \frac{2A^2}{L} \frac{1}{(\beta_0 - \beta_m)^2 + \frac{1}{B^2}} \left\{ \frac{1}{B} + \frac{(\beta_0 - \beta_m)^2 - \frac{1}{B^2}}{L\left((\beta_0 - \beta_m)^2 + \frac{1}{B^2}\right)} \right\} \tag{86}$$

where we neglected terms with $\exp(-L/B)$ assuming that L/B is sufficiently large. In fact if

$$L \gg B, \tag{87}$$

equation (86) can be simplified further:

$$\langle |\varphi_m|^2 \rangle_{av} = \frac{2A^2}{BL} \frac{1}{(\beta_0 - \beta_m)^2 + \frac{1}{B^2}}. \tag{88}$$

Using equation (88) we obtain, from equation (53) for the ensemble average of the square magnitudes of the even guided modes,

$$\langle |C_{me}|^2 \rangle_{av} = \frac{A^2 k^4 L}{2B} (n_g^2 - 1)^2 \cdot \frac{\cos^2 \kappa_0 d \cos^2 \kappa_m d}{\left((\beta_0 - \beta_m)^2 + \frac{1}{B^2}\right)\left(\beta_0 d + \frac{\beta_0}{\gamma_0}\right)\left(\beta_m d + \frac{\beta_m}{\gamma_m}\right)}. \tag{89}$$

The corresponding expression for the odd modes is very similar except that $\cos^2 \kappa_m d$ is replaced by $\sin^2 \kappa_m d$ and β_m, κ_m, and γ_m are the parameters of the odd modes.

The total loss caused by coupling to all guided modes supported by the dielectric waveguide is the sum over all $\langle |C_m|^2 \rangle_{av}$ for even as well as odd modes traveling in positive ($\beta_m = +|\beta_m|$) as well as negative ($\beta_m = -|\beta_m|$) z-direction. It is noteworthy that equation (89) is proportional to L and not to L^2. The conversion to spurious guided modes by random imperfections appears as a true loss to the incident mode.

The losses due to the modes of the continuous spectrum are obtained

from equations (48), (57), (61), (81) and (88) (with $\beta_m = \beta$):

$$\left\langle \frac{\Delta P}{P} \right\rangle_{av} = \frac{A^2 k^4 L}{2\pi B} (n_g^2 - 1)^2 \int_{-k}^{k} \left[\frac{\rho \cos^2 \kappa_0 d}{\left((\beta_0 - \beta)^2 + \frac{1}{B^2}\right)\left(\beta_0 d + \frac{\beta_0}{\gamma_0}\right)} \cdot \left(\frac{\cos^2 \sigma d}{\rho^2 \cos^2 \sigma d + \sigma^2 \sin^2 d} + \frac{\sin^2 \sigma d}{\rho^2 \sin^2 \sigma d + \sigma^2 \cos^2 \sigma d} \right) \right] d\beta. \quad (90)$$

The relation between β, σ, and ρ is given by equations (15) and (16) while β_0, κ_0, and γ_0 are related by equations (9) and (10) and their value is obtained by solving equation (11). The integral in equation (90) is extended over β from $-k$ to k, the range of real values of the propagation constant (in z-direction) of the modes of the continuous spectrum. Equation (90) thus includes the losses due to forward as well as backward scattered radiation. The radiation modes with imaginary values of β can carry power away from the waveguide only strictly perpendicular to its axis. This power loss, if any, is not included in equation (90).

VIII. NUMERICAL RESULTS FOR THE STATISTICAL CASE

Figures 4 through 9 show numerical evaluations of equations (89) and (90). These figures can be grouped into two classes. Figures 4 through 6 are drawn for a dielectric waveguide whose index of refraction is $n_g = 1.01$. Figures 7 through 9 apply to a waveguide with $n_g = 1.5$. Within each of these two classes, the kd value was chosen to allow for three different cases. Figures 4 and 7 apply to waveguides which can support only the lowest order guided mode. In this case there is power lost only to the modes of the continuous spectrum. Figures 5 and 8 apply to waveguides supporting two guided modes and Figs. 6 and 9 apply to waveguides supporting three guided modes. Each figure shows the normalized loss caused by scattering into modes of the continuous spectrum as solid lines and the loss to the possible guided modes as dotted lines. Also shown are the ratios of backward to forward scattered power as solid lines for the modes of the continuum and as dotted lined for the guided modes. The total power lost to the lowest order guided modes is the sum of the losses to the continuum and the spurious guided modes.

Several remarkable features of these loss curves are worthy of a comment. The losses caused by the modes of the continuum as well as by the guided modes peak at certain values of the correlation length B. The location of these peaks are different, however, for the continuum and guided modes.

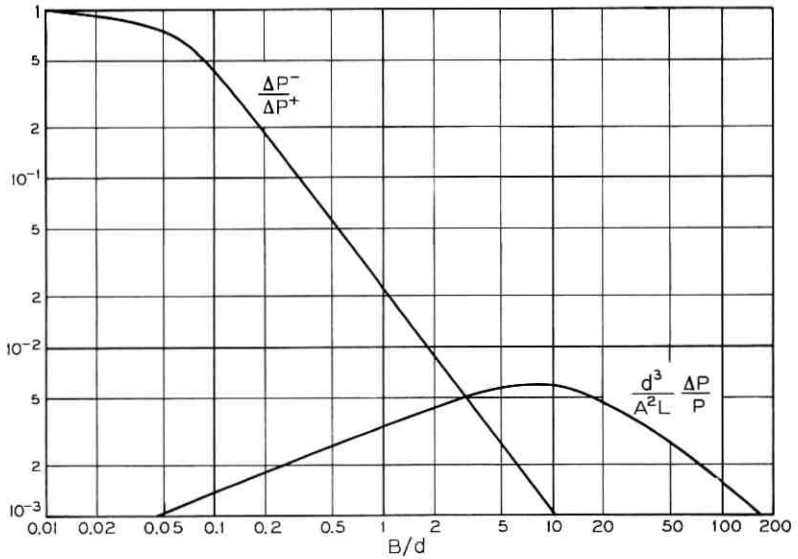

Fig. 4 — Normalized radiation loss (d^3/A^2L) $(\Delta P/P)$ and ratio of backward to forward scattered power $\Delta P^-/\Delta P^+$ as functions of the normalized correlation length B/d for $n_g = 1.01$ and $kd = 8.0$. Single guided mode operation (d = half width of waveguide, A = rms deviation of one waveguide wall, L = Length of waveguide section, n_g = index of refraction of waveguide, k = free space propagation constant).

The losses to the guided modes increase with increasing number of guided modes supported by the waveguide. However, the losses caused by the continuum of modes also increase as an increasing number of guided modes can be supported. This increase is less rapid, however, as one might expect because of the dependence of equation (90) on the fourth power of k. The fourth power dependence on frequency (or inverse wavelength) is typical for Rayleigh scattering by small particles, and it is not surprising that we encounter it here.

Finally, it is apparent from the curves showing the ratio of back-scattered to forward scattered power that forward scattering is predominant for large values of the correlation length. The ratio of $\Delta P^-/\Delta P^+$ levels off for large values of B. In some of the curves the leveling of the $\Delta P^-/\Delta P^+$ curves occurs out of the diagram but it is a common feature of all the curves. For small values of the correlation length there is as much scattering in the forward as in the backward direction.

For many practical applications, a waveguide supporting only one

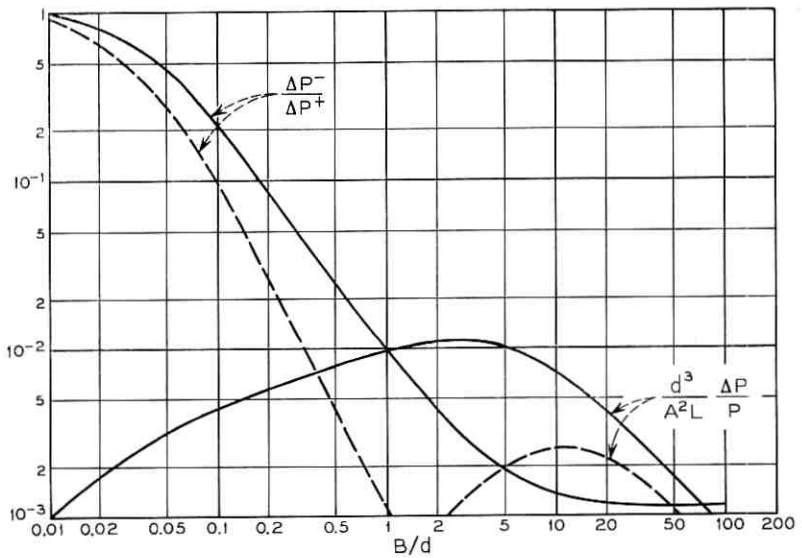

Fig. 5 — Normalized power loss and ratio of foreward to backward scattered power for radiation (solid curves) and spurious guided modes (dashed curves). Two guided modes ($n_g = 1.01$, $kd = 15$).

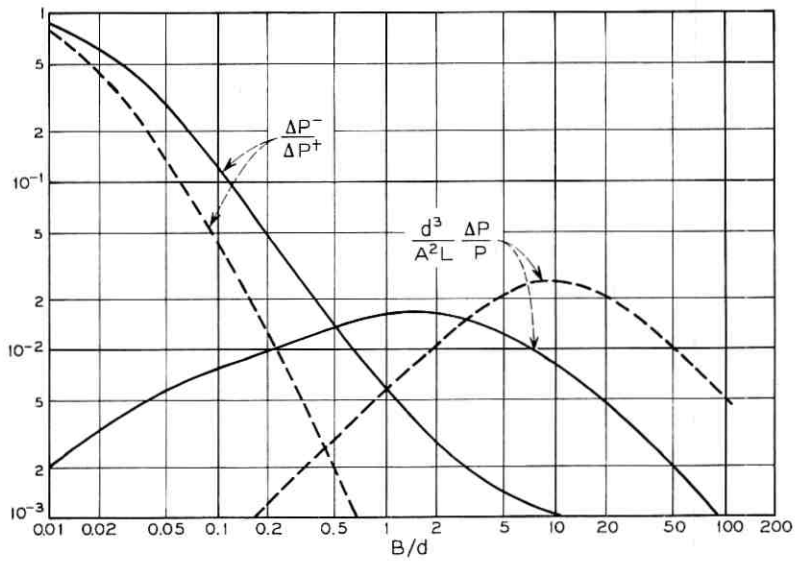

Fig. 6 — Similar to Fig. 5. Three guided modes ($n_g = 1.01$, $kd = 23$).
— — — — two guided mode loss; ——— continuum loss.

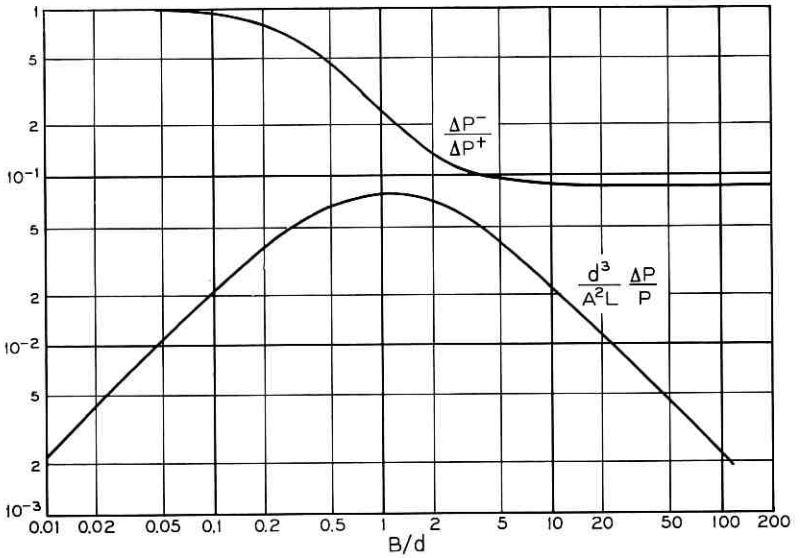

Fig. 7 — Similar to Fig. 4. One guided mode ($n_g = 1.5$, $kd = 1.3$).

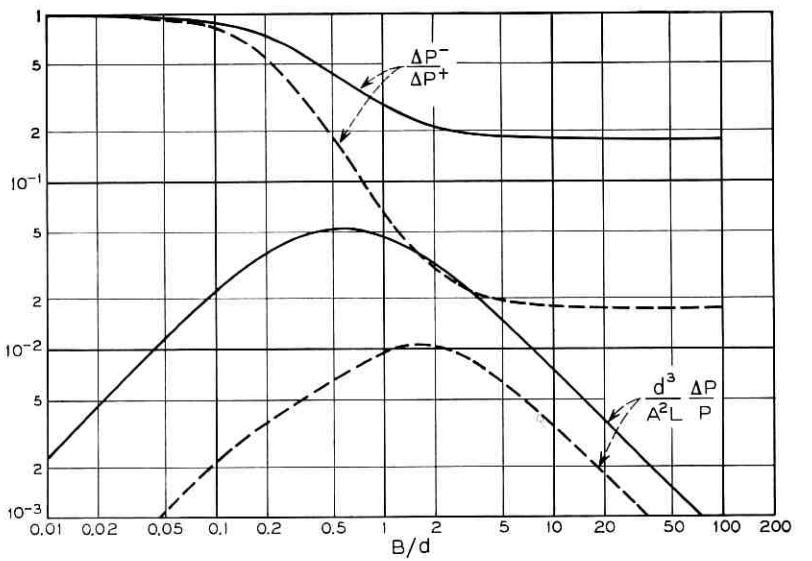

Fig. 8 — Similar to Fig. 5. Two guided modes ($n_g = 1.5$, $kd = 1.8$).
— — — — one guided mode loss; ——— continuum loss.

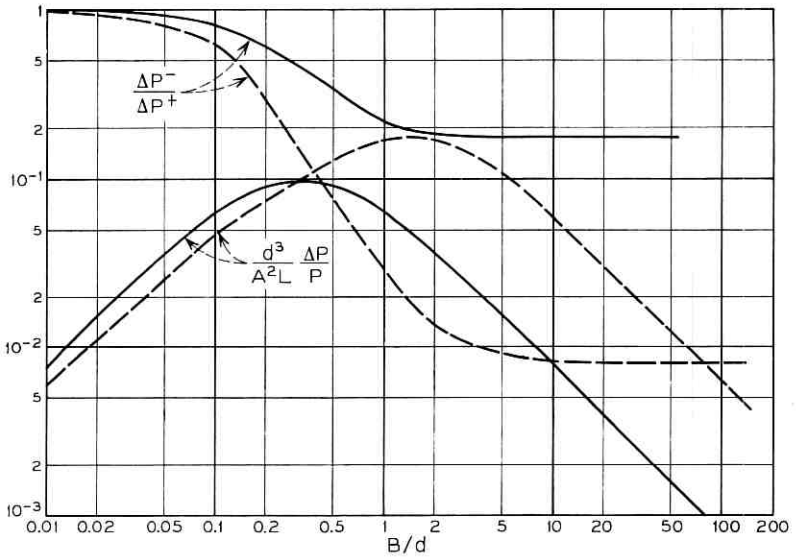

Fig. 9 — Similar to Fig. 5. Three guided modes ($n_g = 1.5$, $kd = 3$).
— — — — two guided mode loss; continuum loss.

guided mode may be of most interest. Let us assume $\lambda_0 = 1\mu$. Figure 4, holding for $kd = 8.0$ and $n_g = 1.01$, applies to a waveguide whose half width is $d = 1.27\mu$. Taking the worst possible case of $B/d = 9$ or $B = 11.4\mu$, we find from Fig. 4

$$\frac{d^3}{A^2L}\frac{\Delta P}{P} = 6 \times 10^{-3}.$$

If we want to know how much rms deviation A of one wall of the guide would be required to cause a 10 percent loss ($\Delta P/P = 0.1$) in one centimeter of waveguide ($L = 1$ cm) we find $A = 5.85 \times 10^{-2}\mu = 585$ Å. The ratio of A over d gives an idea of the relative tolerance requirements:

$$\frac{A}{d} = 4.6 \times 10^{-2} = 4.6\%.$$

If the waveguide were to conform to the conditions of Fig. 6, we would have for $\lambda_0 = 1\mu$ a half width $d = 3.66\mu$. The losses caused by the two spurious modes are of the same order of magnitude as the radiation losses caused by the continuous spectrum. For $B/d = 10$ or $B = 36.6\mu$ we get a total loss of

$$\frac{d^3}{A^2L}\frac{\Delta P}{P} = 3.4 \times 10^{-2}.$$

To cause $\Delta P/P = 0.1$ for $L = 1$ cm requires that

$$A = 1.2 \times 10^{-1}\mu \quad \text{or} \quad \frac{A}{d} = 3.28\%.$$

The relative tolerance requirements are, therefore, approximately the same in both examples.

As a last example let us use Fig. 9 corresponding ($\lambda_0 = 1\mu$) to a waveguide with $n_g = 1.5$ and a half width $d = 0.477\mu$. For $B/d = 1.3$ or $B = 6.2\mu$ we find for the total loss

$$\frac{d^3}{A^2L}\frac{\Delta P}{P} = 2.3 \times 10^{-1}.$$

We get $\Delta P/P = 0.1$ with $L = 1$ cm for

$$A = 2.18 \times 10^{-3}\mu = 21.8 \text{ Å} \quad \text{or} \quad \frac{A}{d} = 0.457\%.$$

The perturbation theory, strictly speaking, holds only for small values of $\Delta P/P$. However, it is reasonable to expect that the power scattered into the radiation modes escapes sufficiently rapidly so that no appreciable amount of power reconversion from the radiation field to the guided mode occurs. The incremental power loss, $\Delta P/P = -\alpha L$, is therefore the same for any section of the guide so that we obtain the total scattering loss into the continuum of radiation modes $P = P_0 e^{-\alpha L}$. We may now ask how much rms deviation is required to cause a radiation loss of 10 dB/km or $\alpha = 2.3 \text{ km}^{-1} = 2.3 \times 10^{-5} \text{ cm}^{-1}$. Using $B/d = 10$, corresponding to the top of the loss curve of Fig. 4, we obtain the equation

$$\frac{d^3}{A^2} \times 2.3 \times 10^{-5} = 6 \times 10^{-3}$$

so that ($\lambda = 1\mu$, $n_g = 1.01$, $kd = 8.0$, $d = 1.27 \times 10^{-4}$ cm)

$$\frac{A}{d} = 6.98 \times 10^{-4} \quad \text{or} \quad A = 8.86 \times 10^{-8} \text{ cm} = 8.86 \text{ Å}.$$

This figure dramatizes the stringent tolerance requirements of dielectric waveguides for long distance optical communications. In fact, such tolerances seem impossible to obtain. One can only hope that the correlation length can be kept far from the worst possible value of $B/d = 10$

(in this example) so that these extremely stringent tolerance requirements might be eased.

IX. CONCLUSION

We have analyzed the losses suffered by the lowest order symmetric mode propagating on a dielectric slab waveguide caused by imperfections of the waveguide boundaries. The analysis was simplified by assuming that there is no change in either the dielectric slab or the guided and unguided fields in one direction parallel to the slab. This assumption causes all our conclusions to be optimistic since variation of the slab in this direction can only cause additional losses. However, we expect that the results of this analysis give at least the correct order of magnitude of the actual scattering losses.

The statistical analysis was limited to a study of the effects which an exponential correlation function might have on the waveguide losses. The actual form of the correlation function may be quite different from this assumed exponential shape.† Conclusions regarding loss predictions are further hampered by a lack of knowledge of the expected correlation length.

However, our analysis does lead one to conclude that scattering losses suffered by optical fibers or other dielectric waveguide structures may be very serious. Deviations of the waveguide wall in the order of a few percent can cause a power loss of 10 percent or 0.46 dB/cm if the wall imperfection can be described by an exponential correlation function with a correlation length to guide half width ratio of approximately $B/d = 10$. An rms deviation of $A = 9$ Å causes a radiation loss of 10 dB/km if the free space wavelength is $\lambda_0 = 1\mu$ and the guide has an index of refraction of $n_g = 1.01$ (with vacuum on the outside). The width of the slab in this last example is $2d = 2.54\mu$.

The mode coupling and radiation loss theory has been experimentally confirmed at microwave frequencies. A report on these measurements is given in Ref. 8.

---

† Several other correlation functions have been tried and it was found that the results are insensitive to the particular choice of the function for values of B/d less than the value corresponding to the loss peak. In particular, the maximum loss value and the position of this loss peak were the same for different correlation functions. However, the loss values for B/d larger than the value corresponding to the maximum of the curve are very strongly dependent on the choice of the correlation function.

REFERENCES

1. Kapany, N. S., "Fiber Optics," New York: Academic Press, 1967.
2. Jones, A. L., "Coupling of Optical Fibers and Scattering in Fibers," J. Opt. Soc., *55*, No. 3 (March 1965), pp. 261–271.
3. Marcatili, E. A. J., unpublished work.
4. Miller, S. E., unpublished work.
5. McKenna, J., "The Excitation of Planar Dielectric Waveguides at p-n Junctions, I," B.S.T.J., *46*, No. 7 (September 1967), pp. 1491–1526.
6. Collin, R. E., "Field Theory of Guided Waves," New York: McGraw-Hill, 1960.
7. Rowe, H. E., and Warters, W. D., "Transmission in Multimode Waveguide with Random Imperfections," B.S.T.J., *41*, No. 3 (May 1962), pp. 1031–1170.
8. Marcuse, D., and Derosier, R. M., "Mode Conversion caused by Diameter Charges of a Round Dielectric Waveguide," B.S.T.J., this issue, pp. 3217–3232.

Mode Conversion Caused by Diameter Changes of a Round Dielectric Waveguide

By DIETRICH MARCUSE and RICHARD M. DEROSIER

(Manuscript received July 8, 1969)

This paper presents the theory of mode conversion and radiation losses of the lowest order circular electric mode in a dielectric rod (fiber) waveguide and its confirmation by a microwave experiment. The theoretical results were obtained from a theory whose detailed development has been presented in an earlier paper.

The microwave experiment was carried out at approximately 50 GHz. The optical fiber with imperfect walls was simulated by a teflon rod of 1 cm diameter and 1 m length with a periodically corrugated wall.

Mode conversion was observed in excellent agreement with theory. The observed radiation losses are somewhat less than the prediction of the perturbation theory, but the agreement is quite good. The direction and width of the far-field radiation pattern was observed in agreement with theory.

I. INTRODUCTION

A theory of mode conversion and radiation losses of a guided mode in a dielectric slab was described in Ref. 1. The power conversion to spurious guided modes as well as to the continuum of unguided radiation modes was assumed to be caused by deviations from perfect straightness of the air-dielectric interface of the slab. The model of the dielectric slab waveguide was chosen for its simplicity.

Even though the dielectric slab exhibits all the relevant features of mode conversion caused by surface roughness and allows one to draw conclusions as to the order of magnitude of the losses suffered by guided modes in dielectric waveguides of other geometries, it is desirable to report the calculations for a round dielectric rod. The results of calculations for the dielectric rod are directly applicable to light transmission along optical fibers. Furthermore, we wanted to test the predictions of the theory at microwave frequencies where a controlled

experiment, to check the effect of surface imperfections on mode guidance, is feasible. We present in this paper the theoretical treatment of the round dielectric waveguide with wall imperfections and its confirmation by a microwave experiment.

The mode conversion theory of round dielectric waveguides is only sketched in this paper since the basic method of calculation has already been described elsewhere.[1] The theory is simplified by limiting the discussion to circular electric modes. In order to avoid coupling between the circular symmetric and other modes, we assume that the symmetry of the rod is such that all derivatives with respect to the angle φ of a cylindrical polar coordinate system (r, φ, z) vanish $(\partial/\partial\varphi = 0)$.

We conclude again (as in Ref. 1) that the radiation and mode conversion losses caused by deviation of the waveguide walls from perfect straightness are extremely severe, imposing strict tolerance requirements on the fabrication of low loss optical fiber transmission lines.

To confirm the basic aspects of our theory we conducted a microwave experiment. Because of the ready availability of equipment, the frequency range of 50 GHz was chosen. Two teflon rods were used to simulate optical fibers. Both rods had 1 cm diameters and a length of 1 m. One rod was smooth and was used for calibration and reference purposes, while the other rod was machined with periodic grooves to simulate an optical fiber with wall imperfections (Fig. 1).

The periodic wall perturbations cause two guided modes to be coupled together. In fact, it is possible to obtain complete power conversion between these two coupled modes. We have observed complete power conversion in agreement with our theory.

In a certain frequency interval, the periodic grooves cause coupling to the continuous spectrum of radiation modes of the dielectric rod. The measured results are somewhat lower than the theoretical prediction. The reason for this discrepancy can be partly explained by a

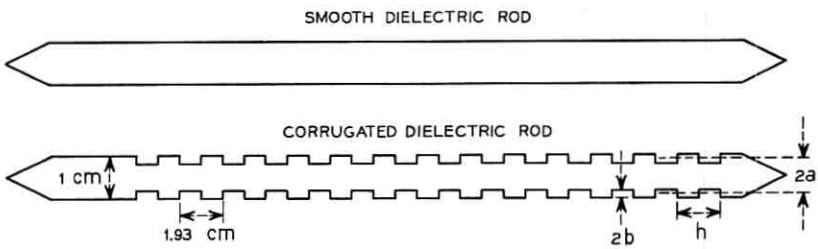

Fig. 1 — The smooth and corrugated teflon rods used for the microwave experiment ($n_\varrho^2 = 2.05$).

certain ambiguity in the value of the effective radius of the corrugated rod. If we make the assumption that the effective radius is either the largest or smallest radius of our rod, we obtain two curves which bracket our experimental results. However, our experimental values are consistently lower than the theoretical predictions based on an average diameter which is the arithmetic mean of the largest and smallest rod diameter. It is more likely, therefore, that the loss prediction of the perturbation theory is slightly too large for losses which are as high as those which occurred in our experiment.

Our theory also predicts the far-field radiation pattern caused by a strictly periodic wall perturbation.[2] We have observed the peak of the far-field radiation lobe and its width in agreement with theory.

II. TE MODES OF THE DIELECTRIC ROD[3]

Imposing the condition

$$\frac{\partial}{\partial \varphi} = 0, \tag{1}$$

the transverse electric field is composed of the components

$$E_\varphi , H_r , H_z . \tag{2}$$

The guided modes have the following form (normal modes of the perfect waveguide are indicated by script letters)

$$\mathcal{E}_\varphi = A_n J_1(\kappa_n r) e^{i(\omega t - \beta_n z)} \quad \text{for} \quad r < a \tag{3a}$$

$$\mathcal{E}_\varphi = A_n \frac{J_1(\kappa_n a)}{H_1^{(1)}(i\gamma_n a)} H_1^{(1)}(i\gamma_n r) e^{i(\omega t - \beta_n z)} \quad \text{for} \quad r > a. \tag{3b}$$

The two magnetic field components are obtained from the E_φ component

$$\mathcal{H}_r = -\frac{i}{\omega \mu} \frac{\partial E_\varphi}{\partial z} \tag{4a}$$

$$\mathcal{H}_z = \frac{i}{\omega \mu} \frac{1}{r} \frac{\partial}{\partial r} (r E_\varphi). \tag{4b}$$

The various symbols used in these equations have the meanings:

a = radius of the dielectric rod,
β_n = propagation constant of mode n,
$\kappa_n = (n_\varrho^2 k^2 - \beta_n^2)^{\frac{1}{2}}$, (5)
$\gamma_n = (\beta_n^2 - k^2)^{\frac{1}{2}}$, (6)
$k = 2\pi/\lambda_0$ = free space propagation constant,

n_g = index of refraction of the waveguide (rod),
ω = radian frequency,
J_1 = Bessel function of order 1, and
$H_1^{(1)}$ = Hankel function of first kind and order 1.

The boundary conditions, requiring that the field components $\mathcal{E}_\varphi$ and $\mathcal{H}_z$ are continuous at $r = a$, lead to the eigenvalue equation for β

$$\frac{\gamma_n}{\kappa_n}\frac{J_1(\kappa_n a)}{J_0(\kappa_n a)} = -i\frac{H_1^{(1)}(i\gamma_n a)}{H_0^{(1)}(i\gamma_n a)}. \tag{7}$$

The subscript 0 designates the Bessel and Hankel functions of zero order. It is convenient to express the mode amplitude A_n by the actual power carried by each mode:

$$P_n = -\frac{1}{2}\int_0^\infty dr \int_0^{2\pi} d\varphi r \mathcal{E}_\varphi \mathcal{H}_r^* = \pi\frac{\beta_n}{\omega\mu}\int_0^\infty r \mid \mathcal{E}_\varphi \mid^2 dr. \tag{8}$$

The modes will be normalized to the same amount of power (1 watt, for example) so that we write

$$P_n = P. \tag{9}$$

The mode amplitude can now be expressed as

$$A_n^2 = \frac{2\omega\mu}{\pi a^2 \beta_n}\frac{P}{\left(1 + \frac{\kappa_n^2}{\gamma_n^2}\right) \mid J_0(\kappa_n a)J_2(\kappa_n a) \mid}. \tag{10}$$

The modes of the continuous spectrum are given by the expressions

$$\mathcal{E}_\varphi = BJ_1(\sigma r)e^{i(\omega t - \beta z)} \qquad r < a \tag{11a}$$

$$\mathcal{E}_\varphi = [CJ_1(\rho r) + DN_1(\rho r)]e^{i(\omega t - \beta z)} \qquad r > a. \tag{11b}$$

The two magnetic components are again obtained from equations (4a) and (4b). N_1 is the Neumann function of order 1 and the parameters σ and ρ are defined:

$$\sigma = (n_g^2 k^2 - \beta^2)^{\frac{1}{2}}, \qquad \rho = (k^2 - \beta^2)^{\frac{1}{2}}. \tag{12}$$

The normalization of the continuous modes involves the Dirac δ-function

$$P\delta(\rho - \rho') = \pi\frac{\beta}{\omega\mu}\int_0^\infty rE_\varphi(\rho)E_\varphi^*(\rho')\, dr. \tag{13}$$

The boundary conditions at $r = a$ determine the relations between the constants C, D, and B

$$\frac{C}{B} = \frac{\pi}{2} \rho a \left(J_1(\sigma a) N_0(\rho a) - \frac{\sigma}{\rho} J_0(\sigma a) N_1(\rho a) \right) \tag{14a}$$

$$\frac{D}{B} = -\frac{\pi}{2} \rho a \left(J_1(\sigma a) J_0(\rho a) - \frac{\sigma}{\rho} J_0(\sigma a) J_1(\rho a) \right), \tag{14b}$$

and these coefficients can be expressed in terms of the power carried by the mode

$$P = \pi \frac{\beta}{\rho \omega \mu} (C^2 + D^2). \tag{15}$$

The actual field of a dielectric rod with imperfect walls can be expanded in terms of the normal modes of the perfect rod:

$$E_\varphi = \sum_{n=0}^{\infty} C_n \mathcal{E}_n + \int_0^\infty g(\rho) \mathcal{E}(\rho) \, d\rho. \tag{16}$$

The remaining caluclation of the power loss to radiation and guided modes, as well as the energy exchange phenomena between different guided modes, are exactly analogous to those developed in Ref. 1 so that their derivation need not be repeated here. In Section III we simply quote the results of the corresponding calculations.

III. SINUSOIDAL WALL PERTURBATION

It was pointed out in Ref. 1 that a sinusoidal wall perturbation can couple only those two modes whose beat wavelength

$$\Lambda_n = \frac{2\pi}{\beta_0 - \beta_n} \tag{17}$$

coincides with the mechanical period h of the wall perturbation. It is therefore possible to consider the coupling phenomenon between only two modes with the result that the coefficient C_0 of the incident mode and the coefficient C_1 of one of the spurious modes obey the relations

$$C_0(z) = \cos | \kappa_{01} | z \tag{18a}$$

$$C_1(z) = \left(\frac{\kappa_{01}^*}{\kappa_{01}} \right)^{\frac{1}{2}} \sin | \kappa_{01} | z \tag{18b}$$

with

$$a\kappa_{01} = \frac{(n_g^2 - 1)\frac{A_0}{a}(ka)^2}{2a(\beta_0\beta_1)^{\frac{1}{2}}} \cdot \frac{J_1(\kappa_0 a)J_1(\kappa_1 a)}{\left[\left(1 + \frac{\kappa_0^2}{\gamma_0^2}\right)\left(1 + \frac{\kappa_1^2}{\gamma_1^2}\right)J_0(\kappa_0 a)J_0(\kappa_1 a)J_2(\kappa_0 a)J_2(\kappa_1 a)\right]^{\frac{1}{2}}}. \tag{19}$$

Here, A_0 is the amplitude of the sinusoidal wall deflection

$$\left.\begin{aligned} r(z) &= a - A_0 \sin \theta z \\ \theta &= \beta_0 - \beta_1 \end{aligned}\right\}. \tag{20}$$

The microwave experiment was conducted with a teflon rod with meandering grooves cut into it. The depth of the grooves is given by $2b$ as shown in Fig. 1. The amplitude of the fundamental Fourier component of the periodic wall deflection of Fig. 1 is given by

$$A_0 = \frac{4b}{\pi}. \tag{21}$$

The two modes exchange their power completely over a distance

$$D = \frac{\pi}{2 \mid \kappa_{01} \mid}. \tag{22}$$

The radiation loss of the dielectric waveguide of Fig. 1 can be calculated by the methods of Ref. 1 resulting in the following equation.

$$\frac{\Delta P}{P} = \frac{L}{a}\frac{4(n_g^2 - 1)^2\left(\frac{b}{a}\right)^2(ka)^4}{\pi\beta_0 a} \cdot \frac{J_1^2(\kappa_0 a)}{\left(1 + \frac{\kappa_0^2}{\gamma_0^2}\right) \mid J_0(\kappa_0 a)J_2(\kappa_0 a) \mid} \sum_{m=0}^{N} \frac{J_1^2(\sigma_m a)}{(2m+1)^2\left[\left(\frac{C_m}{B_m}\right)^2 + \left(\frac{D_m}{B_m}\right)^2\right]}. \tag{23}$$

ΔP is the power lost to radiation modes on a section of the waveguide of length L, and P is the power of the incident lowest order circular electric mode. The meaning of a and b is explained in Fig. 1. The sum in equation (23) takes account of the contributions of each component of the Fourier expansion of the distorted wall profile. The Fourier amplitudes of the function shown in Fig. 2 are

$$A_m = \frac{4b}{\pi(2m+1)} \tag{24}$$

[the zero component of this expansion appeared already in equation

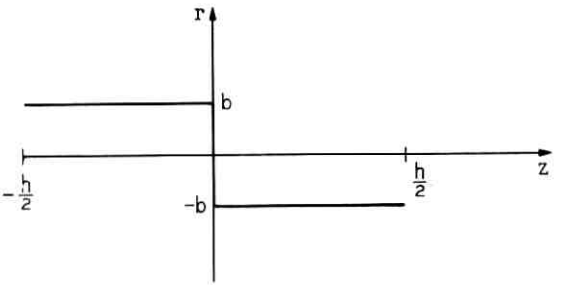

Fig. 2 — The wall distortion function with Fourier expansion:

$$r = \sum_{m=0}^{\infty} \frac{4b}{(2m+1)\pi} \sin\left[(2m+1)\frac{2\pi}{h}z\right]$$

(21)]. The index m, which has been added to the coefficients B, C, and D appearing in equations (14a and b) indicates that they must be evaluated for the following values of

$$\beta_m = \beta_0 - (2m+1)\frac{2\pi}{d}, \tag{25a}$$

$$\sigma_m = (n_v^2 k^2 - \beta_m^2)^{\frac{1}{2}}, \tag{25b}$$

$$\rho_m = (k^2 - \beta_m^2)^{\frac{1}{2}}. \tag{25c}$$

The physical reason for the occurence of these discrete values of the propagation constant β in the continuous spectrum of modes is the requirement (derived in Ref. 1) that only those values of β are appreciably coupled to the incident guided mode which satisfy the relation

$$\beta_0 - \beta = \frac{2\pi}{\Lambda_m} \tag{26}$$

where Λ_m is the period length of a Fourier component of the wall distortion function.

IV. THE STATISTICAL CASE

To first order of perturbation theory, the expansion coefficient $g(\rho, z)$ appearing in equation (16) is given by

$$g(\rho, z) = L \frac{k^2(n_v^2 - 1)(\rho)^{\frac{1}{2}}}{(2)^{\frac{1}{2}} i(\beta_0\beta)^{\frac{1}{2}}} \cdot \frac{\varphi J_1(\kappa_0 a) J_1(\sigma a)}{\left\{\left[\left(\frac{C}{B}\right)^2 + \left(\frac{D}{B}\right)^2\right]\left(1 + \frac{\kappa_0^2}{\gamma_0^2}\right) \mid J_0(\kappa_0 a) J_2(\kappa_0 a) \mid\right\}^{\frac{1}{2}}} \tag{27}$$

with

$$\varphi = \frac{1}{L}\int_0^L [f(z) - a]e^{-i(\beta_0-\beta)z}\,dz. \tag{28}$$

It was pointed out in Ref. 1 that the average power loss caused by scattering into the radiation field is given by

$$\left\langle \frac{\Delta P}{P} \right\rangle_{av} = \int_{-k}^{k} \langle |\, g\,|^2 \rangle_{av} \frac{\beta}{\rho}\,d\beta. \tag{29}$$

The symbol $\langle\ \rangle_{av}$ indicates an ensemble average. The ensemble average of $|\,\varphi\,|^2$ is given by

$$\langle |\,\varphi\,|^2 \rangle_{av} \approx \frac{2}{L}\int_0^L R(u)\cos(\beta_0 - \beta_m)u\,du \tag{30}$$

with the correlation function

$$R(u) = \langle [f(z) - a][f(z+u) - a] \rangle_{av}. \tag{31}$$

The relative power loss caused by radiation from the rod is obtained from equations (27) and (29)

$$\frac{1}{L}\left\langle \frac{\Delta P}{P} \right\rangle_{av} = \frac{k^4(n_g^2 - 1)^2}{2\beta_0\left(1 + \frac{\kappa_0^2}{\gamma_0^2}\right)} \frac{J_1^2(\kappa_0 a)}{|\, J_0(\kappa_0 a)J_2(\kappa_0 a)\,|} \int_{-k}^{k} \frac{[\langle |\,\varphi\,|^2 \rangle_{av} L]J_1^2(\sigma a)}{\left(\frac{C}{B}\right)^2 + \left(\frac{D}{B}\right)^2}\,d\beta. \tag{32}$$

V. NUMERICAL RESULTS FOR THE STATISTICAL CASE

To be able to make numerical predictions, let us assume that the correlation function is given by

$$R(u) = A^2 \exp\left(-\frac{|\,u\,|}{B}\right) \tag{33}$$

so that we obtain

$$L\langle |\,\varphi\,|^2 \rangle_{av} = \frac{2A^2}{B}\frac{1}{(\beta_0 - \beta)^2 + \frac{1}{B^2}}. \tag{34}$$

Figure 3 shows a plot of $(a^3/LA^2)(\Delta P/P)$ as a function of B/a for $n_g = 1.01$, $ka = 23.0$ and $n_g = 1.5$, $ka = 3.0$. Both conditions are chosen so that only the lowest order circular electric mode can propagate in the dielectric rod.

To get a feeling for the magnitude of the losses to be expected from random variations of the rod's radius, we calculate the rms deviation

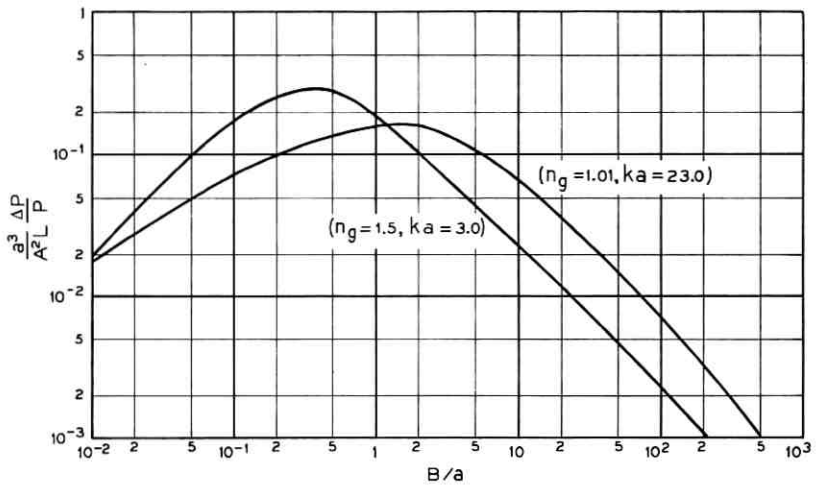

Fig. 3 — Normalized radiation loss caused by random wall perturbations with exponential correlation function, a = radius of fiber, A = rms value of wall deviation, L = length of waveguide section, k = free space propagation constant. The dimensions shown in the figure were chosen to ensure single guided mode operation.

A required to cause $\Delta P/P = 0.1$ for a rod length of $L = 1$ cm for $n_g = 1.01$ and the worst possible value of $B/a = 2$. Assuming $\lambda = 1\mu$ we get from $ka = 23$ the value $a = 3.66\mu$ for the guide radius. With (from Fig. 3)

$$\frac{a^3}{A^2L}\frac{\Delta P}{P} = 0.16$$

we find

$$\frac{A}{a} = 1.5 \times 10^{-2} = 1.5\%$$

or

$$A = 550 \text{ Å}.$$

As discussed in Ref. 1, it may be permissible to apply the perturbation calculation of the radiation loss repeatedly so that from

$$\frac{\Delta P}{P} = -\alpha z, \tag{35}$$

$$P = P_0 e^{-\alpha L} \tag{36}$$

can be obtained. We can then ask for the rms deviation A of the rod's radius which causes a loss of 10 dB/km. With the numerical values used above we find

$$A = 8.4 \text{ Å}.$$

Almost the same figure was obtained for the rms deviation of the half width of the dielectric slab which causes a 10 dB/km radiation loss of the lowest order (even) guided mode. However, in the case of the slab, one wall was assumed to be perfect.

VI. THE MICROWAVE EXPERIMENT

The experimental setup is shown in Fig. 4. The microwave signal is generated by a reflex klystron whose rectangular waveguide output is fed into a round waveguide by means of a rectangular-to-round waveguide transducer. The round waveguide is connected to a section of round helix waveguide which serves as a mode filter suppressing all but the circular electric $TE_{01}^{(m)}$ mode. Transition between the $TE_{01}^{(m)}$ mode of the round waveguide and the corresponding TE_{01} mode of the dielectric rod waveguide is achieved by inserting the rod into the waveguide. This mode launcher is not perfect since a small amount of TE_{02} mode of the dielectric waveguide is excited. The $TE_{01}^{(m)}$ mode of the round waveguide cannot excite the pure TE_{01} mode of the dielectric rod since the field configurations of the two modes are slightly different. In addition to some residual TE_{02} mode, small amounts of asymmetric modes of the dielectric rod are also excited because of imperfect centering of the rod inside the round waveguide.

To probe the field outside of the dielectric rod and detect the conversion of power from the TE_{01} to the TE_{02} mode, we used a probe which

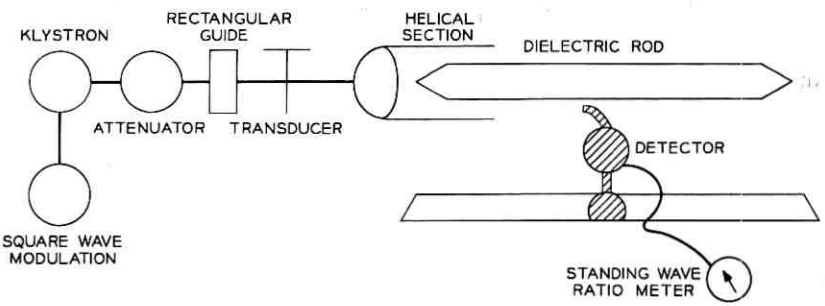

Fig. 4 — Block diagram of the microwave experiment.

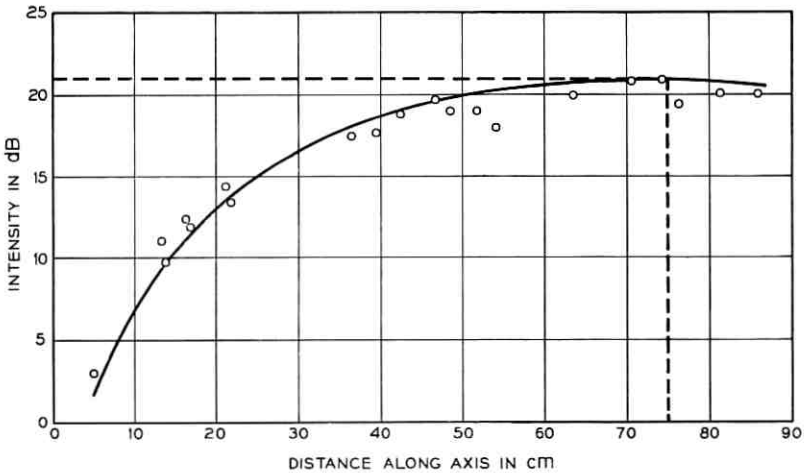

Fig. 5 — Buildup of the TE_{02} mode along the corrugated rod. Groove depth $= 7.6 \times 10^{-3}$ cm.

consisted simply of an L-shaped piece of RG 98U waveguide which was mounted on an optical rail, which made it possible to move the detector parallel to the dielectric rod. The receiver attached to the L-shaped probe consisted of a single diode detector followed by an amplifier which was tuned to 250 Hz. The klystron was amplitude modulated at that same frequency. The periodicity of the grooves of the corrugated dielectric rod (Fig. 1) was chosen equal to the beat wavelength between the TE_{01} and TE_{02} modes of the dielectric rod as given by equation (17).

Mode conversion from TE_{01} to the TE_{02} mode can easily be observed with our detector arrangement because the TE_{02} mode extends much farther away from the rod than the more tightly confined TE_{01} mode. Moving the detector to approximately 4 mm from the surface of the rod made it impossible to observe any trace of the TE_{01} mode, while the TE_{02} mode could easily be detected.

That the corrugation does indeed serve to transfer power from the TE_{01} to the TE_{02} mode is shown in Fig. 5. The measured values of TE_{02} power are shown as dots on this figure. Also shown is a plot of the $\sin^2 x$ function which gives the theoretical law of the power increase according to equation (18b). The slight scatter of the measured points is caused by interference between the TE_{02} mode and some other residual mode which is unintentionally generated by the mode launcher. From equation (22) we calculate $D = 80$cm for our particular experiment. From Fig. 5 we see that the experimental value of the total energy ex-

change length is approximately 75 cm. The remaining discrepancy between the theoretical and experimental values can easily be attributed to the machining accuracy of the rod which was no better than 2.5×10^{-3} cm. Striking proof of the identity of the mode whose buildup is shown in Fig. 5 is provided by Fig. 6.

Figure 6 was obtained by moving the L-shaped detector transversely at the end of either the smooth or the corrugated rod. The detector is thus probing the near field radiation pattern which results as the guided mode leaves the end of the rod and radiates into space. This near field radiation pattern is a faithful reproduction of the shape of the guided mode inside of the waveguide. The solid curve shown in Fig. 6 was obtained by probing the transverse field pattern of the smooth rod. This field pattern shows clearly the TE_{01} mode. There is a slight distortion in the wings of this mode which is caused by interference between the TE_{01} mode and a small amount of TE_{02} power launched by the transducer. The dotted curve in Fig. 6 was obtained by placing the detector at the end of the corrugated rod. We took care to insert the corrugated

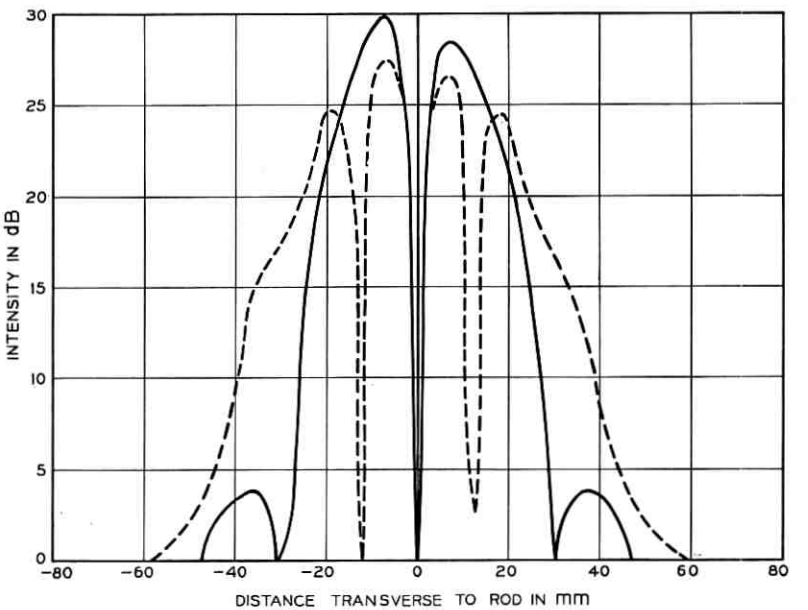

Fig. 6 — The near-field radiation patterns of the guided modes (transverse field distribution). Solid line = TE_{01} mode at end of smooth rod; dotted line = TE_{02} mode at end of corrugated rod.

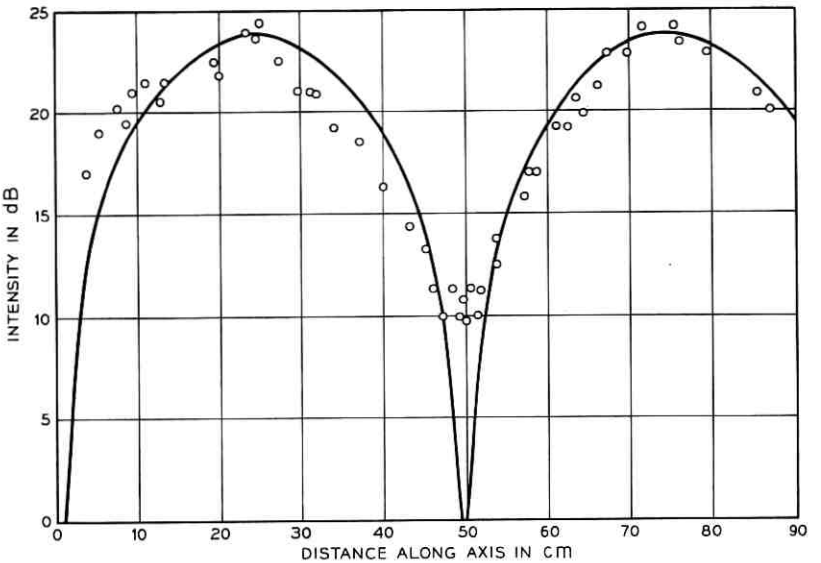

Fig. 7 — Buildup of TE_{02} mode along the corrugated rod. Groove depth = 2.3×10^{-2} cm.

rod so far into the launcher that the section protruding from the launcher was equal to the total power exchange length shown in Fig. 5. It is apparent that the TE_{02} mode (instead of the TE_{01} mode generated by the launcher) is present at the output end of the corrugated rod. It is also apparent that almost complete mode conversion has taken place. Figures 5 and 6 were obtained from a corrugated rod whose grooves had a depth of 7.6×10^{-3} cm. In order to be able to observe radiation losses, we deepened the grooves in this rod to a depth of 2.3×10^{-2} cm. The power buildup as a result of mode conversion from TE_{01} to TE_{02} on the rod with deeper grooves is shown in Fig. 7. The TE_{02} mode is shown to go through two complete power exchanges. The exchange length is now 25 cm in agreement with theory.

Finally, we observed the radiation of power from the corrugated rod with the deeper grooves. Equation (26) indicates the relation between the z-component of the propagation vector of those radiation modes that couple to the TE_{01} mode and the period of the periodic corrugation of the rod. It is clear that the basic Fourier component with length Λ_0 of the corrugated wall distortion function will contribute predominantly to radiation loss. Furthermore, since $\beta < k$ is required for all radiation modes, we see that only very little power can be lost to radia-

tion unless the relation

$$\beta_0 - k \leqq \frac{2\pi}{\Lambda_0} \tag{37}$$

is satisfied. It follows from equation (37) that above $f = 51$GHz very little radiation loss is to be expected. Indeed we see in Fig. 7 that complete energy exchange between two guided modes is taking place which would be impossible if substantial amounts of power had been lost to radiation. However, below 51 GHz, equation (23) predicts considerable radiation loss.

The applicability of the radiation loss theory to our experiment is somewhat questionable. We must not forget that equation (23) was derived from a perturbation theory under the assumption that only very little power is lost from the original guided mode. If the radiation detaches itself from the rod over a distance for which the power loss of the guided mode due to radiation is only slight, we may be justified in making the transition to equation (36). However, this procedure becomes more and more questionable as the radiation losses increase. Furthermore, the transition to equation (36) is less likely to be accurate if the radiation is directed forward along the rod. It is shown in Ref. 2 that forward radiation results close to the region where the equal sign of equation (37) applies.

Finally, there is some uncertainty what value "a" for the rod's radius should be used in equation (23). Since the radius of the corrugated rod is variable, some suitable average value must be taken. Figure 8 shows three theoretical curves. The two dotted curves were calculated using the largest and smallest value of the radius in equation (23). The solid curve was obtained by using the average value of the radius. The crosses in Fig. 8 show the results of our loss measurements. It is apparent that most of these points fall within the two dotted curves. However, all points lie below the solid curve. These loss measurements were obtained by comparing the output power at the end of the smooth and corrugated rod. The accuracy of these measurements is no better than approximately $\pm\frac{1}{2}$ dB. In view of the discussion of the applicability of the perturbation theory to high radiation losses, the agreement between theory and experiment must be considered as good.

Figure 9 shows the angle of the far-field pattern of the radiation lobes caused by power loss due to the corrugated wall. The dots are measured values, while the curve is a result of the theory of Ref. 2. Again we see good agreement between experiment and theory.

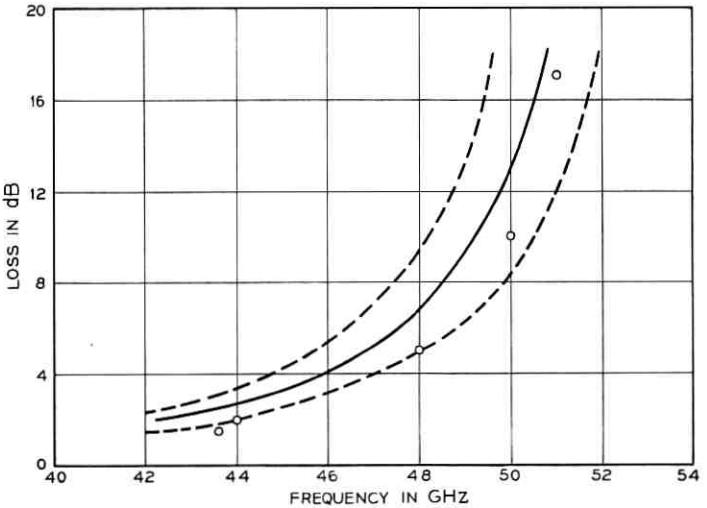

Fig. 8 — Radiation loss as a function of frequency. Dotted lines represent theoretical loss assuming that guide radius is either the maximum or the minimum value. Solid curve shows theoretical loss based on average radius of corrugated rod. The dots are the measured points.

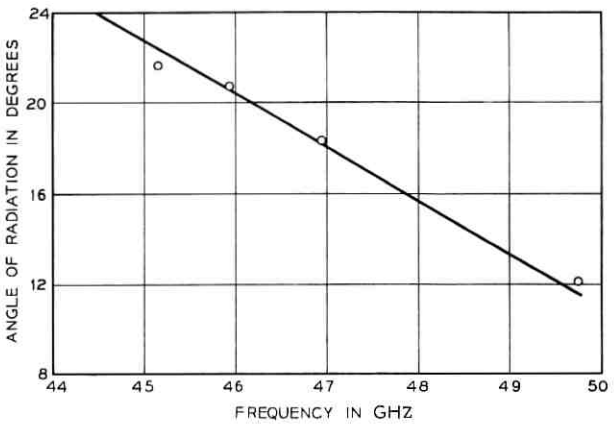

Fig. 9 — The angle of the far-field radiation lobe as a function of frequency. Solid-line represent theory; the dots are measured points.

VII. CONCLUSION

This paper contains a perturbation theory of mode conversion effects and radiation losses of a round dielectric waveguide. This theory is applicable to light transmission in optical fibers. The theory developed here is limited to the circular electric modes of round dielectric waveguides. However, the order of magnitude of the losses for other modes is expected to be similar.

The theory has been checked by a scaled experiment at microwave frequencies. The dielectric fiber with wall imperfections was simulated by a teflon rod of 1 cm diameter which was provided with periodic grooves. Mode conversion from the TE_{01} mode of the dielectric rod to the TE_{02} mode was observed in excellent agreement with experiment. The observed radiation losses are in reasonable agreement with theory. An existing discrepancy can be attributed to the limitations of the perturbation theory to predict correctly the high losses encountered in this experiment.

The conclusion to be drawn from our theory for the operation of optical fibers is a need for very strict tolerance requirements. For example, the radiation losses caused by surface roughness of a fiber designed for single mode operation at 1μ wavelength can be as high as 10 dB/km for an rms variation of the fiber wall of as little as 8 Å.

REFERENCES

1. Marcuse, D., "Mode Conversion Caused by Surface Imperfections of a Dielectric Slab Waveguide," B.S.T.J., this issue, pp. 3187–3215.
2. Marcuse, D., "Radiation Losses of Dielectric Waveguides in Terms of the Power Spectrum of the Wall Distortion Function," B.S.T.J., this issue, pp. 3233–3242.
3. Collin, R. E., *Field Theory of Guided Waves,* New York: McGraw-Hill, 1960.

Radiation Losses of Dielectric Waveguides in Terms of the Power Spectrum of the Wall Distortion Function

By DIETRICH MARCUSE

(Manuscript received July 23, 1969)

In an earlier paper I described a perturbation theory of the radiation losses of a dielectric slab waveguide. The statistical treatment of the radiation losses was based on the correlation function of the wall distortion. This paper discusses the results of the radiation loss theory in terms of the power spectrum of the function describing the thickness of the slab. We found that only those mechanical frequencies θ of the power spectrum contribute to the radiation loss that fall into the range $\beta_0 - k < \theta < \beta_0 + k$. ($\beta_0$ = propagation constant of guided mode, k = free space propagation constant.) The mechanical frequencies near both end points of this mechanical frequency range contribute more to the radiation loss than the region well inside of this range.

We also discuss the far-field radiation pattern caused by a strictly sinusoidal wall distortion.

I. INTRODUCTION

In an earlier paper I developed a perturbation theory of the mode conversion effects between guided modes and of the radiation losses of a given guided mode caused by deviations from perfect straightness of the waveguide wall.[1] For simplicity, the discussion had been limited to a waveguide in the form of an infinitely extended dielectric slab.

The statistical discussion had been based on the description of the wall distortion by means of a correlation function. In Ref. 1 an exponential correlation function had been assumed. However, it has been established that the shape of the correlation function has little influence on the radiation losses.

It is possible to base the discussion of radiation losses not on correlation functions, but on the mechanical power spectrum of the wall distortion function. This study provides information as to how the various

mechanical frequencies of the wall distortion function contribute to the radiation losses.

The analysis of Ref. 1 was based on the use of radiation modes of the dielectric slab which represent standing waves in directions transverse to the propagation direction of the guided modes. The question naturally arises how a superposition of these standing waves can result in radiation flowing away from the rod. This question is answered by examining the far field radiation pattern caused by a sinusoidal distortion of one wall of the dielectric waveguide. This paper gives the relation between the length of the mechanical period, the wavelength of the guided mode, and the direction of the main lobe of the radiation.

II. RADIATION LOSS AND POWER SPECTRUM

The amplitudes of the modes of the continuous spectrum were derived in Ref. 1, equations (65) and (69). We have

$$g_e(\rho, L) = \frac{Lk^2}{2i(\pi)^{\frac{1}{2}}}(n_v^2 - 1)\frac{\rho(\cos \kappa_0 d \cos \sigma d)[\varphi(\theta) - \psi(\theta)]}{\left[\beta\left(\beta_0 d + \frac{\beta_0}{\gamma_0}\right)(\rho^2 \cos^2 \sigma d + \sigma^2 \sin^2 \sigma d)\right]^{\frac{1}{2}}} \tag{1}$$

for the even modes, and

$$g_0(\rho, L) = \frac{Lk^2}{2i(\pi)^{\frac{1}{2}}}(n_v^2 - 1)\frac{\rho(\cos \kappa_0 d \sin \sigma d)[\varphi(\theta) + \psi(\theta)]}{\left[\beta\left(\beta_0 d + \frac{\beta_0}{\gamma_0}\right)(\rho^2 \sin^2 \sigma d + \sigma^2 \cos^2 \sigma d)\right]^{\frac{1}{2}}} \tag{2}$$

for the odd modes. The functions

$$\varphi(\theta) = \frac{1}{L}\int_0^L [f(z) - d]e^{-i\theta z}\, dz, \tag{3}$$

$$\psi(\theta) = \frac{1}{L}\int_0^L [h(z) + d]e^{-i\theta z}\, dz, \tag{4}$$

with

$$\theta = \beta_0 - \beta \tag{5}$$

are the Fourier transforms of the wall distortion functions $f(z) - d$ and $h(z) + d$. [$x = f(z)$ is the boundary of the dielectric-air interface, $x = d$ describes the wall of the perfect guide, and $x = h(z)$ is the distorted boundary near $x = -d$.]

The meaning of the constants appearing in equations (1) to (5) is:

β_0 = propagation constant of guided mode (propagating in z-direction),
β = component of the propagation constant of the continuum mode in z-direction,
k = propagation constant in free space,
L = length of guide section with wall distortions,
n_g = dielectric constant of slab,

$$\rho = (k^2 - \beta^2)^{\frac{1}{2}} \tag{6}$$

$$\sigma = (n_g^2 k^2 - \beta^2)^{\frac{1}{2}}, \tag{7}$$

$$\kappa_0 = (n_g^2 k^2 - \beta_0^2)^{\frac{1}{2}}, \tag{8}$$

$$\gamma_0^2 = (\beta_0^2 - k^2)^{\frac{1}{2}}. \tag{9}$$

The y-component of the electric radiation field caused by the wall distortions is given by

$$E_y = \int_0^\infty [g_e(\rho, L)\mathcal{E}_e(\rho, z) + g_0(\rho, L)\mathcal{E}_0(\rho, z)]\, d\rho. \tag{10}$$

The functions $\mathcal{E}_e$ and $\mathcal{E}_o$ are the even and odd radiation modes. The ratio of scattered power to incident guided mode power is obtained from

$$\frac{\Delta P}{P} = \int_{-k}^{k} (|\, g_e(\rho, L)\,|^2 + |\, g_0(\rho, L)\,|^2)\, \frac{\beta}{\rho}\, d\beta. \tag{11}$$

For simplicity we assume that one wall of the slab is perfect

$$h(z) = -d, \tag{12}$$

so that

$$\psi(\theta) = 0, \tag{13}$$

the relative scattering loss, follows from equations (1), (2), and (10)

$$\frac{\Delta P}{P} = \int_{-k}^{k} \frac{1}{d^2} L\, |\, \varphi(\theta)\,|^2\, I(\beta)\, d\beta \tag{14a}$$

with

$$I(\beta) = \frac{(kd)^4}{4\pi} (n_g^2 - 1)^2 \frac{\cos^2 \kappa_0 d}{\beta_0 d + \frac{\beta_0}{\gamma_0}} (\rho d) \left[\frac{\cos^2 \sigma d}{(\rho d)^2 \cos^2 \sigma d + (\sigma d)^2 \sin^2 \sigma d} + \frac{\sin^2 \sigma d}{(\rho d)^2 \sin^2 \sigma d + (\sigma d)^2 \cos^2 \sigma d} \right]. \tag{14b}$$

Since $\varphi(\theta)$ is the Fourier component of the wall distortion function its absolute square value

$$| \varphi(\theta) |^2 \tag{15}$$

is the "power spectrum" of $f(z) - d$. It is apparent from equation (14) that $\Delta P/P$ depends on the power spectrum of the wall distortion function. Incidentally, equation (14) is not a statistical expression, but holds for a specific dielectric slab waveguide. We entered the power spectrum in the combination $L | \varphi |^2$ in equation (14) since this combination is independent of L for a randomly varying function $f(z) - d$.

Equation (14) allows us immediately to determine the range of mechanical frequencies θ which contribute to the radiation loss. The integral in equation (14) is extended from $-k$ to k, the β range of continuous radiation modes. The range of mechanical frequencies contributing to the scattering loss is therefore given by

$$\beta_0 - k < \theta < \beta_0 + k. \tag{16}$$

This is an important result since it states that those parts of the power spectrum which lie outside of the range, equation (16), do not contribute to radiation loss.

This last statement must not be misconstrued to mean that a waveguide with a sinusoidal wall distortion extending over length L

$$f(z) = d + a \sin \theta' z \qquad 0 \leqq z \leqq L \tag{17}$$

with θ' lying outside the range of equation (16) does not lose power by radiation. The power spectrum of equation (17) is

$$| \varphi(\theta) |^2 = \left[\frac{a}{L} \frac{\sin (\theta' - \theta) \frac{L}{2}}{\theta' - \theta} \right]^2. \tag{18}$$

A term with $\theta' + \theta$ in the denominator has been neglected in equation (18). The accuracy of this approximation improves with increasing values of L.

It is apparent from equation (18) that $| \varphi(\theta) |^2$ has non-vanishing values for $\theta \neq \theta'$ so that there is some small contribution to radiation loss even if θ' lies outside of the range of equation (16).

However, if we consider the limit $L \to \infty$ we can approximate the power spectrum, equation (18), by a δ-function:

$$\lim_{L \to \infty} | \varphi(\theta) |^2 = \frac{\pi a^2}{2L} \delta(\theta - \theta'). \tag{19}$$

In this special case the expression (14a) for the scattered power becomes

$$\frac{\Delta P}{P} = \frac{\pi}{2}\left(\frac{a}{d}\right)^2 I(\beta_0 - \theta'). \tag{20}$$

The scattering from a dielectric waveguide with a wall distortion function whose power spectrum is a δ-function is proportional to $I(\beta_0 - \theta')$.

The function $I(\beta)$ is plotted in Fig. 1 for $n_g = 1.01$, $kd = 8.0$, and $\beta_0 d = 8.041$. The scattering caused by a wall distortion with a δ-function spectrum (a sinusoidal wall distortion of infinite length) is nearly independent of the value of $\beta = \beta_0 - \theta'$ over most of the β-range. There are two sharp peaks at $\beta \approx k$ and $\beta \approx -k$. The physical reasons for the sharp increase in loss at these values is easy to understand if we consider the direction of the radiation pattern as a function of θ'. We show in Section III [equation (35)] that the angle α between the waveguide and the main radiation lobe is given by

$$\cos \alpha = \frac{\beta}{k} = \frac{\beta_0 - \theta'}{k}. \tag{21}$$

The two peaks of the function $I(\beta)$, or correspondingly of the radiation loss, are associated with

$$\alpha \approx 0 \quad \text{and} \quad \alpha \approx \pi. \tag{22}$$

This shows that the radiation loss is high when the radiation pattern is

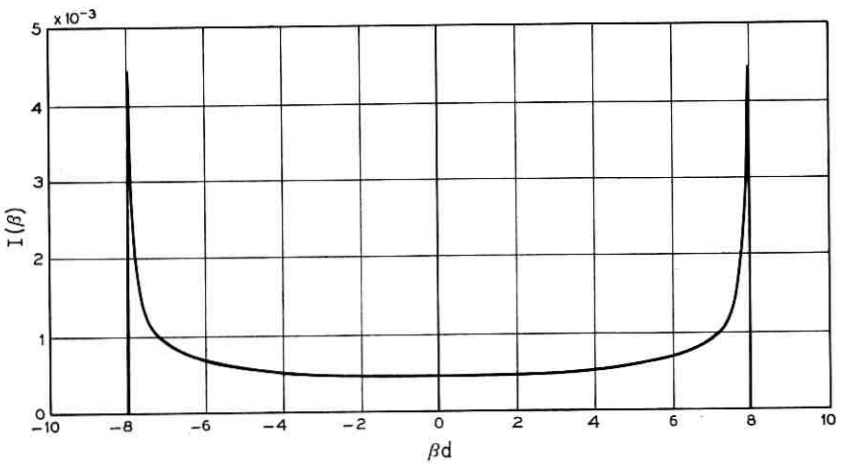

Fig. 1 — Graphical representation of the function $I(\beta)$ [eq. (14b)]. $n_g = 1.01$, $kd = 8.0$, $\beta_0 d = 8.041$.

directed very nearly parallel to the surface of the waveguide. The radiation modes gain more power if the guided mode can interact with them over a longer distance. An observation of this loss peak is reported in Ref. 2.

A power spectrum with sharp peaks much like that of equation (18) or (19) is not likely to occur for dielectric waveguides with random imperfections of the dielectric interface. It is much more reasonable to expect that such waveguides may have spectral distributions which are nearly independent of θ over a certain range of θ values. In the limit of a "white" spectrum,

$$| \varphi(\theta) |^2 = \text{constant}, \tag{23}$$

the scattering loss is proportional to the integral over the function $I(\beta)$ shown in Fig. 1. The two peaks contribute very little to this integral. Numerical integration of $I(\beta)$ of Fig. 1 including and excluding the peaks resulted in the values:

$$\int_{-8}^{8} I(\beta)\, d\beta = 0.011, \int_{-7.8}^{7.8} I(\beta)\, d\beta = 0.0096,$$

$$\text{and} \quad \int_{-7.5}^{7.5} I(\beta)\, d\beta = 0.0087.$$

This result is reassuring for the use of the perturbation theory which was used to derive equation (14). The perturbation theory is based on the assumption that power is converted from the guided mode to the radiation field but that no power is converted back from the radiation field to the guided mode. This approximation is certain to yield better results if the radiation pattern is directed away from the rod. In other words, the perturbation theory will work poorest in the region of the peaks of Fig. 1. However, for spectra that do not particularly favor the regions of these peaks, the contribution of those regions (which at the same time give the least reliable results) to the total radiation loss is only slight.

III. THE FAR FIELD RADIATION PATTERN

The far field pattern of the radiation field (that is excited by the lowest order even guided mode traveling in the dielectric slab with sinusoidal perturbation of one wall) can easily be calculated from equation (10). The even and odd radiation modes were given in Ref. 1 (for $|x| > d$)

$$\mathcal{E}_y^{(e)} = \left[\frac{2\omega\mu P}{\pi\beta(\rho^2\cos^2\sigma d + \sigma^2\sin^2\sigma d)}\right]^{\frac{1}{2}}$$

$$\times [\rho\cos\rho(|x| - d)\cos\sigma d - \sigma\sin\rho(|x| - d)\sin\sigma d]e^{i(\omega t-\beta z)} \quad (24)$$

$$\mathcal{E}_y^{(0)} = \frac{x}{|x|}\left[\frac{2\omega\mu P}{\pi\beta(\rho^2\sin^2\sigma d + \sigma^2\cos^2\sigma d)}\right]^{\frac{1}{2}}$$

$$\times [\rho\cos\rho(|x| - d)\sin\sigma d + \sigma\sin\rho(|x| - d)\cos\sigma d]e^{i(\omega t-\beta z)}. \quad (25)$$

With $\psi(\theta) = 0$ and

$$\varphi(\theta) \approx \frac{a}{iL}\exp\left[i(\theta' - \theta)\frac{L}{2}\right]\frac{\sin(\theta' - \theta)\frac{L}{2}}{\theta' - \theta} \quad (26)$$

and with the help of equations (1) and (2) we get from equation (10)

$$E_y = -\frac{ak^2}{(2)^{\frac{1}{2}}\pi}(\omega\mu P)^{\frac{1}{2}}(n_y^2 - 1)\frac{\cos\kappa_0 d}{\left(\beta_0 d + \frac{\beta_0}{\gamma_0}\right)^{\frac{1}{2}}}$$

$$\times \int_0^\infty \frac{\rho}{\beta}\left\{\frac{\cos\sigma d[\rho\cos\rho(x - d)\cos\sigma d - \sigma\sin\rho(x - d)\sin\sigma d]}{\rho^2\cos^2\sigma d + \sigma^2\sin^2\sigma d}\right.$$

$$\left. + \frac{\sin\sigma d[\rho\cos\rho(x - d)\sin\sigma d + \sigma\sin\rho(x - d)\cos\sigma d]}{\rho^2\sin^2\sigma d + \sigma^2\cos^2\sigma d}\right\}$$

$$\times \exp\left[i(\theta' - \theta)\frac{L}{2}\right]\frac{\sin(\theta' - \theta)\frac{L}{2}}{\theta' - \theta} \times e^{i(\omega t-\beta z)}\, d\rho. \quad (27)$$

In the far field with $x \to \infty$ and $z \to \infty$ (but L finite) we can obtain an approximate solution of the integral in equation (27) by the method of stationary phase.[3] The sine and cosine functions of argument $\rho(x - d)$ can be expressed as sums of exponential functions. The most important terms of the integrand of equation (27) are, therefore, of the form

$$\exp[-i(\beta z \pm \rho x)]. \quad (28)$$

This exponential term is an extremely rapidly varying function of ρ as $x \to \infty$ and $z \to \infty$. All other terms in the integrand vary slowly by comparison. According to the method of stationary phase the contribution to the integral comes predominantly from a region that is determined by

$$\frac{\partial}{\partial\rho}(\beta z \pm \rho x) = 0. \quad (29)$$

With the help of equation (6), equation (29) leads to the condition

$$\frac{x}{z} = \pm\frac{\rho_0}{\beta} \tag{30}$$

or

$$\rho_0 = k \sin \alpha \tag{31a}$$

$$\beta = k \cos \alpha \tag{31b}$$

with

$$\cos \alpha = \frac{z}{(x^2 + z^2)^{\frac{1}{2}}} = \frac{z}{r}. \tag{32}$$

For $x > 0$ and $z > 0$ only the $+$ sign in equation (30) is possible. This is an important point. It shows that even though the radiation modes, equations (24) and (25), represent standing wave patterns in x-direction only, the outward traveling part of the decomposition of the standing wave into traveling waves makes a contribution to the radiation field, equation (27).

All terms of the integrand with the exception of equation (28) can be taken out of the integral. The remaining integration can be carried out using the expansion

$$\beta z + \rho x = k(x \sin \alpha + z \cos \alpha) - \frac{1}{2}\frac{z}{k \cos^3 \alpha}(\rho - \rho_0)^2 + \cdots$$

$$\cdot \int_0^\infty e^{-i(\beta z + \rho x)}\, d\rho = (1 + i)(\pi)^{\frac{1}{2}} \frac{(k)^{\frac{1}{2}} \cos \alpha}{(r)^{\frac{1}{2}}} e^{-ik(x \sin \alpha + z \cos \alpha)}. \tag{33}$$

The far field is therefore obtained in the form

$$E_y = \frac{1}{(\pi)^{\frac{1}{2}}} \exp\left(i\frac{\pi}{4}\right) ak^{\frac{1}{2}}(\omega\mu P)^{\frac{1}{2}}(n_y^2 - 1) \frac{\cos \kappa_0 d}{\left(\beta_0 d + \frac{\beta_0}{\gamma_0}\right)^{\frac{1}{2}}}$$

$$\cdot \frac{\rho_0^2 \sin 2\sigma_0 d - i\rho_0\sigma_0 \cos 2\sigma_0 d}{(\rho_0^2 + \sigma_0^2) \sin 2\sigma_0 d - 2i\rho_0\sigma_0 \cos 2\sigma_0 d} \frac{\sin(\theta' - \theta)\frac{L}{2}}{\theta' - \theta}$$

$$\cdot \exp\left[i(\theta' - \theta)\frac{L}{2}\right] e^{i\rho_0 d} \frac{1}{(r)^{\frac{1}{2}}} e^{i[\omega t - k(x \sin \alpha + z \cos \alpha)]}. \tag{34}$$

The index zero was added to σ to indicate that it must be evaluated from equations (7) and (8) using ρ_0 of equation (31a).

Equation (34) reveals several important features of the far field of

radiation. This field is essentially a plane wave traveling in the direction of α ($\tan \alpha = x/z$, and x and z are the coordinates of the point of observation).

The field intensity is inversely proportional to the square root of the distance r from (the sinusoidally distorted) waveguide section. The dependence on distance is inversely proportional to $(r)^{\frac{1}{2}}$ rather than r because the waveguide is infinitely extended in y-direction (see Ref. 1).

The main radiation lobe occurs at the maximum value of $[\sin (\theta' - \theta)L/2]/(\theta' - \theta)$ that is at $\theta = \theta'$ or from equations (5) and (31b) at

$$\cos \alpha_m = \frac{\beta_0 - \theta'}{k} \tag{35}$$

(β_0 = propagation constant of guided mode).

The width of the main lobe depends on the length L of the sinusoidally distorted waveguide section. The difference in angle between the peak of the lobe and the first null determines the half width of the main lobe

$$\Delta\alpha = \frac{2\pi}{Lk \sin \alpha} \text{ for } \alpha \neq 0. \tag{36a}$$

The width of the main radiation lobe is inversely proportional to L. The lobe is narrowest for $\alpha = \pi/2$ and becomes wider as α decreases toward zero. If the peak of the main lobe is at $\alpha = 0$, we obtain

$$\Delta\alpha = \left(\frac{4\pi}{Lk}\right)^{\frac{1}{2}} \text{ for } \alpha = 0. \tag{36b}$$

The peak amplitude of the main radiation lobe is not strongly dependent on α. The increase in radiated power in forward direction ($\alpha = 0$) which is apparent from Fig. 1 is caused by the broadening of the radiation lobe with decreasing angle.

IV. CONCLUSION

The radiation loss of dielectric waveguides caused by deviations from perfect straightness of the waveguide walls depends on the "power spectrum" of the wall deviation function. A sinusoidal wall perturbation gives rise to radiation into a particular direction in space. Each Fourier component of the Fourier expansion of the wall distortion function is responsible for radiation into a particular direction. The width of the radiation lobes is wide for scattering directions parallel to the rod so that those Fourier components responsible for forward and backward scattering contribute more to the radiation loss than those causing scat-

tering in other directions. However, this preferential loss behavior is not very pronounced, so that the Fourier components responsible for forward and backward scattering contribute only a small amount of the total radiation loss caused by a broad power spectrum.

The coupling between two guided modes of the dielectric waveguide is also governed by equation (5). Only one component of the power spectrum of the wall distortion function influences the coupling between two guided modes, while the entire range of mechanical frequencies, equation (16), determines the radiation loss.

The general predictions of this theory have been experimentally verified. Microwave experiments on a periodically corrugated teflon rod have shown that the radiation losses are negligibly small if the period of the corrugation is such that θ lies outside of the interval indicated by equation (16).[2] However, if θ falls inside of the interval, equation (16), considerable radiation losses do occur. The peak of the radiation losses shown in Fig. 1 and the direction and width of the radiation lobes have also been observed in agreement with this theory.

REFERENCES

1. Marcuse, D., "Mode Conversion Caused by Surface Imperfections of a Dielectric Slab Waveguide," B.S.T.J., this issue, pp. 3187–3215.
2. Marcuse, D., and Derosier, R. M., "Mode Conversion Caused by Diameter Changes of a Round Dielectric Waveguide," B.S.T.J., this issue, pp. 3217–3232.
3. Mathews, J., and Walker, R. L., "Mathematical Methods of Physics," New York: W. A. Benjamin, 1965, pp. 85–86.

Amplitude Distributions of Telephone Channel Noise and a Model for Impulse Noise

By J. H. FENNICK

(Manuscript received June 30, 1969)

The noise waveforms found on voice bandwidth telephone channels are generally recognized to be non-gaussian in their amplitude distribution. This paper presents data which suggests that a simple exponential is a good function to describe amplitude densities in the extreme tails.

A comprehenisve model of impulse noise as viewed on trunk groups is then presented. The model relates the distributions of impulse noise levels and impulse noise counts.

I. INTRODUCTION

Noise on telephone channels has been measured for years with instruments which are constructed to enable reasonably good correlations between the reading obtained and the annoyance of the noise during a telephone conversation.[1] Fluctuations of the meter pointer during a measurement are either ignored or mentally averaged by the observer, depending upon their frequency of occurrence and their magnitude. With the introduction of data transmission on the telephone network, the relatively frequent high amplitude excursions of the noise waveform were viewed as a "new" kind of noise, primarily because they were generally not annoying in voice communication and it was recognized that no meaningful measure of them could be obtained with the standard noise measuring sets. The term "impulse noise" was applied to these high excursions and new instruments were designed to measure them.[2]

The significance of impulse noise in data transmission has given rise to a great deal of effort devoted to its measurement, characterization, and evaluation as a transmission impairment.[3-6] (For an extensive bibliography, see Ref. 3.) Several models have been suggested to describe the erratic behavior and clustering phenomena associated with

this type of noise. The Pareto model of Berger and Mendelbrot and the generalized hyperbolic model proposed by Mertz appear to be the best presented to date.[3,7] A more mathematically tractible (than the hyperbolic) model has recently been applied to error rate data by Fritchman. He proposed a partitioned Markov chain model which would seem to show promise in this area although it does not seem to have been applied to impulse noise data as yet.[8] The model presented here does not deal specifically with the intervals between occurrences of noise pulses but is concerned directly with the number of occurrences per unit time above any threshold (in decibel) of observation. Extrapolation of occurrences of noise pulses to errors created in data transmission is a function of many parameters besides the occurrence of noise and will not be discussed here although good prediction techniques exist.[5]

In order to set the background for the discussion of impulse noise as a separate phenomenon, as opposed to the background noise or as a part of the composite noise waveform on a channel, data are first presented on the amplitude probability density function of the noise as observed and comparisons made with gaussian noise. The data reflect only the range of variables encountered and should not be considered as statistically describing the amplitude distributions of noise on telephone channels.

II. IMPULSE NOISE AS A DISTINCT PROCESS

Typical oscillograph noise waveforms from a random noise generator and from a telephone channel are shown in Fig. 1. Each trace is 200 ms long and both have the same rms value. The upper one is from the noise generator, the lower one from a telephone channel. The occurrence of two "impulses" are shown near the left end of the lower trace. It is primarily the occurrence of such "pulses" that make real channel noise decidedly different from band-limited white gaussian distributed noise (the upper trace).

Figures 2a and b show two such impulses extracted from a noise recording, sampled at a 15 kHz rate and analyzed to determine their amplitude and phase characteristics in the frequency domain. In both cases, the phase characteristic is shown to be relatively smooth, but the frequency content highly variable. Similar analyses on about 2000 noise pulses verified these observations. However, if a large sample of pulses, on the order of 200, is taken from a given channel, the average spectrum appears to be approximately the shape of the channel gain-frequency characteristic—not a very surprising result. Such an averaging is shown in Fig. 3.

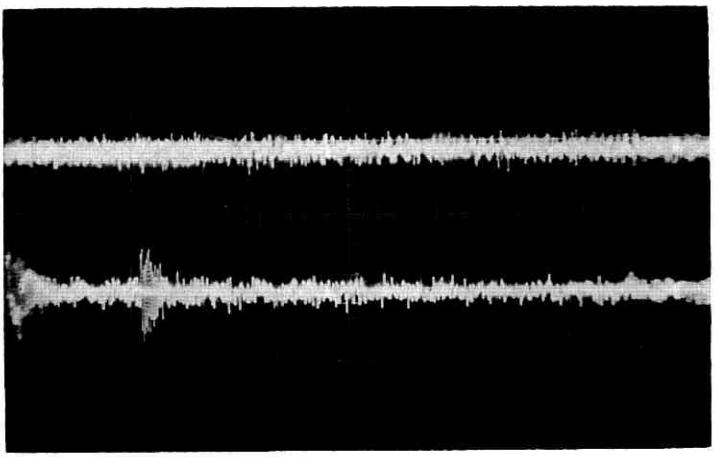

Fig. 1 — 200 ms samples of random noise and telephone channel noise with equal rms levels.

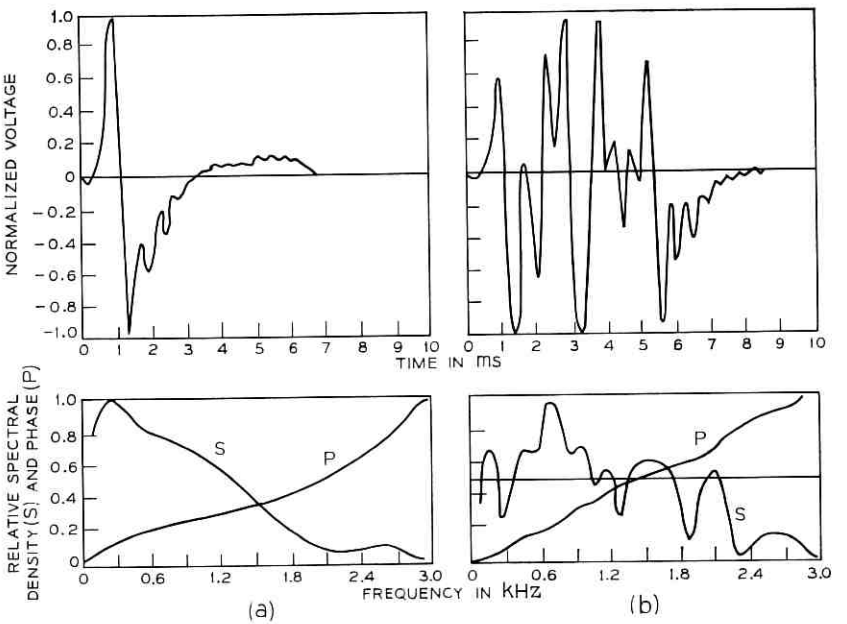

Fig. 2 — Samples of impulses extracted from telephone channel noise with their amplitude and phase characteristics in the frequency domain.

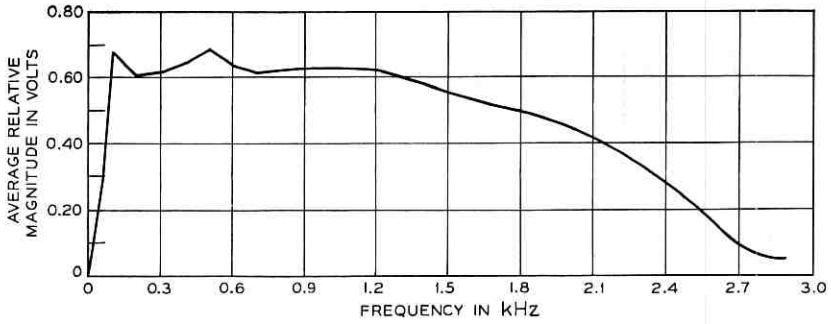

Fig. 3 — Average spectral content of about 200 impulses from a single telephone channel.

Figures 1 and 2 serve as partial justification for treating impulse noise as a separate phenomenon. The pulses shown in Fig. 1 do not rise to strikingly high amplitudes compared to the rest of the noise waveform. Those in Fig. 2, however, are so large that the scale prohibits viewing the background noise waveform which continues beyond that shown. This extreme peaking will become more apparent in Section III.

III. PERCENT OF TIME WAVEFORM IS WITHIN AN INTERVAL

The percentage of time that the noise is within a given interval (± 0.5 dB in this case) is a useful means of describing a random waveform. Data in the form of histograms were obtained by sampling, at a 10 kHz rate, 30 minute tape recordings of telephone channel noise. Equipment limitations imposed a usable dynamic range of 30 dB, so the apparatus was adjusted to examine only the extreme peaks of the noise. In practice this usually required that the noise be examined at levels corresponding to percentages of 10^{-2} or less. This approach was also consistent with the nature of the problem—the relatively high noise amplitudes were of greatest interest. Logarithmic compression and decibel scaling were used and resulted in a unique presentation of the data. Instead of the usual scaling in voltage, the abscissa is scaled in decibels removed from the rms value of the noise. A negative sign preceding an abscissa value refers simply to one polarity of noise waveform, a positive sign refers simply to the opposite polarity. Zero on the abscissa corresponds to the rms value of the noise waveform. For convenient comparison, the equivalent data for a gaussian distribution are also shown in each of the figures presented. The ordinate, proportion

of time the waveform is within $\pm\frac{1}{2}$ dB of the indicated level, is presented in powers of 10 from 10^{-2} to 10^{-8}.

Figures 4 and 5 show the histograms as measured on two different channels. Figure 4 was taken from data recorded on a coaxial cable system and Fig. 5 from a microwave radio system. The striking departure from a stationary gaussian process is obvious. The sampling rate of 10 kHz over a 30 minute period resulted in 18×10^6 samples. Values on Fig. 4 of 5.5×10^{-8} represent one sample in 18 million and can hardly be considered significant. The values of 10^{-8} shown on Fig. 4 represent voids in the data. Figure 6 shows a histogram constructed by combining seven 30 minute recordings, and so represents an "average" histogram over 3.5 hours of real time. The result is surprisingly linear for values below about 5×10^{-5} and suggests that the tails of the amplitude distribution of real channel noise are approximated quite well by a simple exponential.

A total of 37 half hour recordings were analyzed in this fashion. Seventeen of these were taken from microwave radio channels and 20

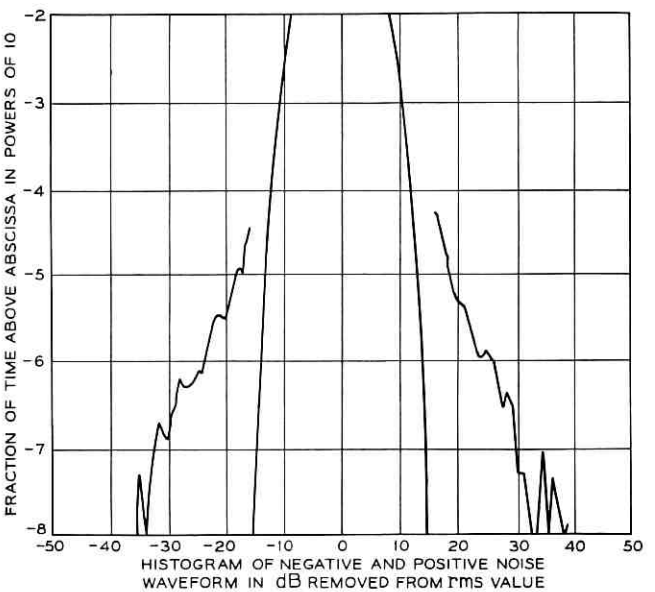

Fig. 4 — Histogram of telephone channel noise amplitudes compared with gaussian distribution. Sample from a coaxial transmission system. Proportion of time that the noise waveform is within ±0.5 dB of abscissa value. Abscissa is decibel removed from rms.

from various types of cable or coaxial carrier systems. The variability observed on the microwave systems is much greater than that on cable systems so the two sets of data are treated separately.

Since no data are available on the amplitude histograms at values in excess of 10^{-4}, it is assumed here that the histogram for such values is represented by a truncated normal function. The observed data suggest that, if the noise is stationary, its amplitude density function then may be written:

$$p(x) = \begin{cases} 0; & x < -b \\ ce^{kx}; & -b \leqq x \leqq -a \\ \hat{\Phi}; & -a < x < a \\ ce^{-kx}; & a \leqq x \leqq b \\ 0; & x > b \end{cases} \tag{1}$$

where

$\pm b$ = realistic bounds on the voltage waveform (channel saturation),

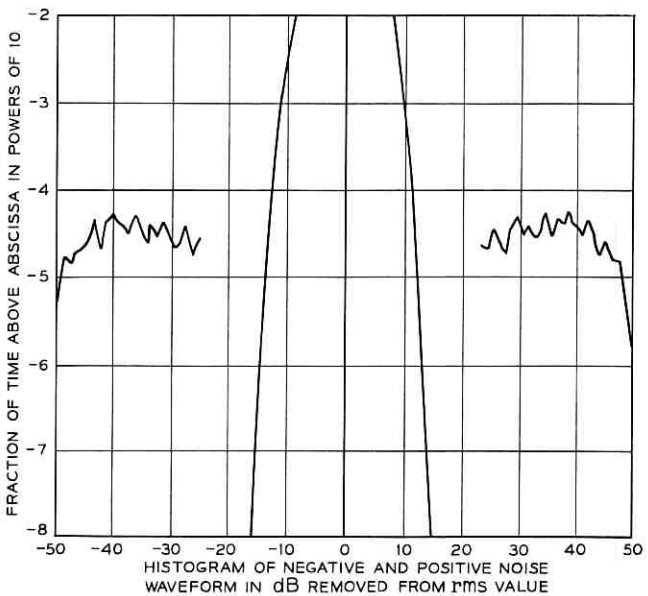

Fig. 5 — Histogram of telephone channel noise amplitudes compared with gaussian distribution. Sample from a microwave radio system. Constructed as in Fig. 4.

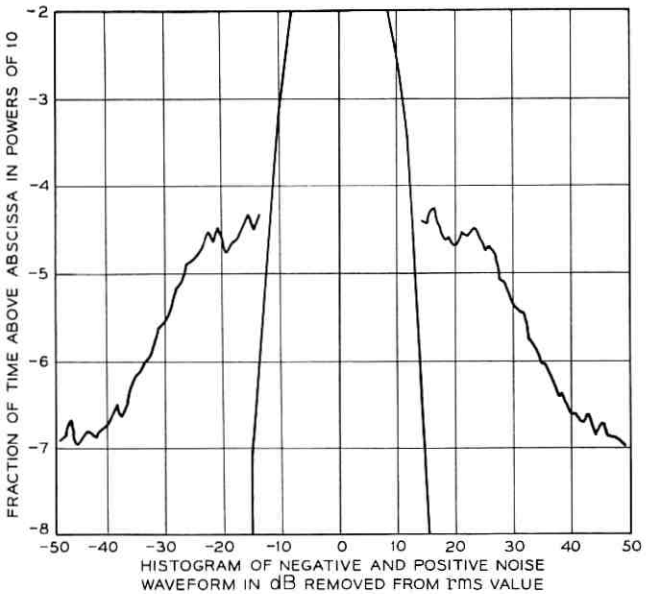

Fig. 6 — Histogram of noise amplitude from cable and coaxial systems taken over 3½ hours. Constructed as in Fig. 4.

$\pm a$ = points of departure of the density function from an assumed underlying Gaussian,

$\hat{\Phi}$ = gaussian density truncated at $\pm a$.

c, k = parameters describing the exponential density function.

The value of b ranges from about 30 to 50 dB*, the value of a ranges from about 10 to 15 dB*, and k may be negative as illustrated in Fig. 5. The variable c ranges over 14 orders of magnitude from 10^{-7} to 10^{7}. No significant correlations were found between any of the variables in the data analyzed. The point of departure from the assumed underlying truncated gaussian distribution a is given by the positive solution of the quadratic

$$a = k \pm \{k^2 - 2 \ln [c(2\pi)^{\frac{1}{2}}]\}^{\frac{1}{2}}.$$

Some values of k and c are given below.

3.1 *Histograms on Cable and Coaxial Carrier Systems*

As stated earlier, the histograms on cable and coaxial carrier systems showed less variability than those on microwave radio systems. In fact,

* That is, above the rms value.

two of the 20 observations tracked the assumed gaussian distribution to within less than $\frac{1}{2}$ dB over the entire range from 10^{-3} to 10^{-7}. These two observations lend credence to the assumption of an underlying gaussian process and also show that at least two channels had no impulse noise in the sense of the term as defined in Section I.

The data are summarized in two ways. First, the values of the variables k and c were examined, and then the intercepts (abscissa values) for various values of proportion were studied.

For cable and coaxial carrier systems (20 samples)* the mean of k was 0.45 and the estimated standard deviation s was 0.46. Because of the extreme range of c, as mentioned in Section III, only the median appears to be of interest; it was found to be 0.0028. A probability density function using the mean value of k and the median of c is shown in Fig. 7. The resultant exponential departs from the guassian distribution at about 4.5×10^{-6} on the ordinate.

The second method of examining the data is considered to be more meaningful in terms of a representative average. The intercepts at proportion values of 10^{-4} to 10^{-7} were studied. The mean $\langle x \rangle_{av}$ (in dB), estimated standard deviation s, median, and 90 percent confidence intervals (CI) about the mean, are shown in Table I. The average and median functions so derived are also shown in Fig. 7. The distributions of intercepts were found to be very nearly log-normal for all four proportion values (10^{-4} through 10^{-7}). This explains the differences between the means and medians as in Table I and Fig. 7. A skew distribution of the intercepts is of the form to be expected. A lower bound on the intercept is imposed by the gaussian assumption and a gradual tailing off of the values at the high range might be expected.

Taking the median value of the exponential distribution as being a representative value of conditions on cable carrier systems, it is of interest to compare tail values of the resultant cumulative distribution function (CDF) with the gaussian distribution. The median exponential intercepts the gaussian at an x value 12.6 dB above the rms. This corresponds to the $\log^{-1}(12.6/20) = 4.26\sigma$ point. If the noise amplitude were truly gaussian, only 0.004 percent of the waveform would lie beyond the $\pm 4.26\sigma$ points. However, 0.0134 percent of the area lies below the exponential portion of the density function, nearly a full order of magnitude difference. This sheds some light on the predicted performance of data systems, for instance, in the presence of gaussian noise, a typical analysis situation, and that actually observed in a working version of the system over real channels.

* Each "sample" is 30 minutes of time.

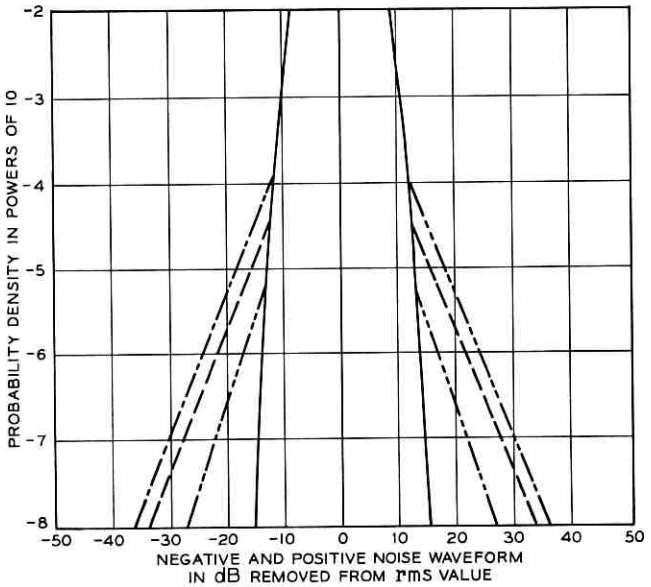

Fig. 7 — Amplitude probability density functions. Estimated average and median taken over constant density values and derived from mean parameter k and median parameter c, compared to gaussian. Averages taken over 10 hours of noise on cable carrier systems. — gaussian; — · — average; — — — median, — · · — mean k, median C.

3.2 *Histograms on Microwave Radio Systems*

The data for the microwave systems were analyzed in the same way as the second method for the cable systems. The individual values of k and c were not computed because of the dubious value of such an effort. Averaging over the intercepts for constant values of density (method 2) illustrates the greater variability in the microwave systems. The results, presented in Table II, show this by the larger estimated standard deviations and wider 90 percent confidence intervals about the

TABLE I—ESTIMATED PROBABILITY DENSITIES FOR CABLE AND COAXIAL CARRIER SYSTEMS

Probability Density	$\langle x \rangle_{av}$ (dB)	s (dB)	Median (dB)	90% CI (dB)
10^{-5}	19.3	4.4	15	1.5
10^{-6}	24.6	4.9	23	1.8
10^{-7}	30.2	5.5	28	2.0

TABLE II—ESTIMATED PROBABILITY DENSITIES FOR MICROWAVE RADIO SYSTEMS

Probability Density	$\langle x \rangle_{av}$ (dB)	s (dB)	Median (dB)	90% CI (dB)
10^{-4}	17.7	10.7	12.6	5.5
10^{-5}	23.2	12.8	18.5	5.2
10^{-6}	24.7	9.7	22	4.1
10^{-7}	29.6	9.4	27	4.1

estimated means (compare with Table I). The median function so derived is shown in Fig. 8. The distributions of intercepts were again found to be closely approximated by the log-normal distribution and the median curve examined as for the cable carrier systems. The median exponential intercepts the gaussian distribution at 11.6 dB above the rms value. This corresponds to $\log^{-1}(11.6/20) = 3.8\sigma$, or 0.0165 percent of the noise waveform that would lie beyond $\pm 3.8\sigma$ of a gaussian distribution. The values of k and c for the median curve on Fig. 8 are 0.48 and 0.042. Integration of the resultant exponential function over the

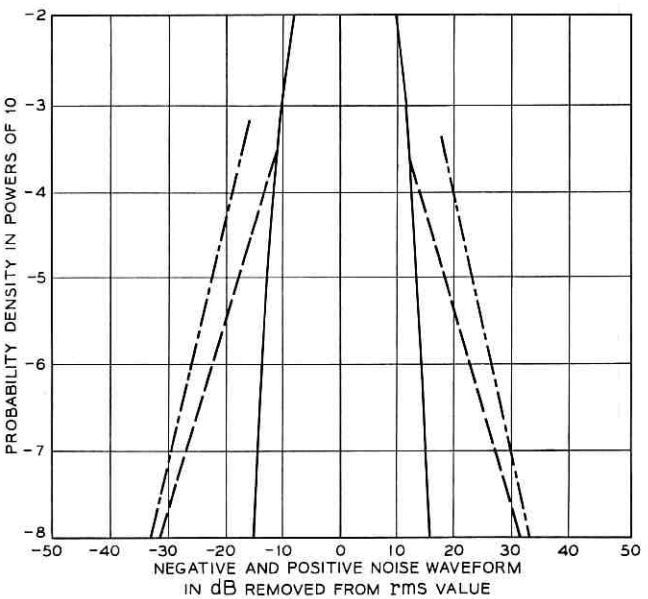

Fig. 8 — Amplitude probability density functions. Estimated median taken over constant density values from 8½ hours of noise from microwave radio systems. — gaussian, — - — average, — — — median.

appropriate intervals yields 0.068 percent of the median waveform in excess of the 3.8σ points of an assumed gaussian. In this case, the median difference is about a factor of four.

IV. FORMAL DEFINITION OF AN IMPULSE AND SOME PULSE LENGTH DATA

In the process of impulse noise analysis a formal definition is required. The definition is illustrated in Fig. 9 and was first proposed by Kaenel, and others.[9] The waveform illustrated in Fig. 9 represents an ideally rectified noise waveform being sampled by an A/D converter. All portions of the noise waveform that remain below a variable slicing level, designated level 2, are considered as part of the underlying band-limited white gaussian or background noise until level 2 is exceeded. Once level 2 is exceeded, the noise pulse, or impulse, is measured starting at the point where level 1 was exceeded as indicated in the figure until it returns below level 1 and remains for a specified amount of time referred to as a guard interval. The function of the guard interval is to distinguish between nodes of a single impulse and two impulses which occur close together in time. Various guard intervals have been used in the analysis of voiceband impulse noise, from 0.3 ms to 0.8 ms. The choice is somewhat arbitrary, but on the basis of the author's unpublished interpulse time distributions, his choice is 0.6 ms as an optimum value. This is preferred because interpulse gap length histograms commonly exhibit a null at about 0.6 ms. The adjustment of levels 1 and 2 vary, but level 1 is typically 10 dB above the rms value, and level 2

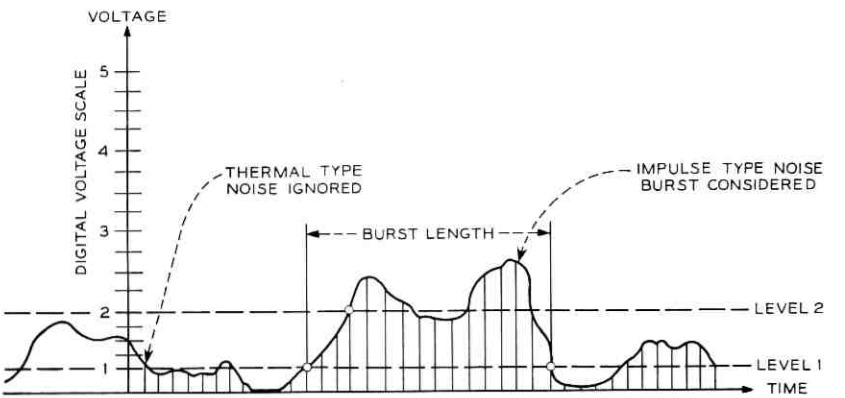

Fig. 9 — Ideally rectified noise waveform illustrating definition of pulse length.

from 13 to 16 dB above rms. In the context of this formal definition, the impulse has been referred to as a burst.[9]

Under the rules of the definition, just given, frequency functions were constructed for the lengths of several thousand impulses. A set of these are shown in Fig. 10; the set is shown as "envelopes of all the observed frequency functions." Only two points appear to be significant. The modes of the functions occur at about 1.2 ms, and lengths in excess of 10 ms are almost never observed. The remainder of this paper discusses an impulse noise model.

V. A MODEL FOR IMPULSE NOISE ON TELEPHONE CHANNELS

This section describes impulse noise as viewed on a trunk group as it is used by a switched network subscriber. The distributions of the peak amplitudes of individual impulses have been of interest for some time, and extensive data concerning them have been collected.[2-4,6] Methods of relating such distributions to data system performance have also been derived.[5]

The data are most frequently collected by means of simple threshold detectors. Excursions of the noise waveform above the threshold are recorded on electromechanical counters.[2] Such measuring devices have finite counting rates which may be exceeded at times by the rate of occurrence of impulses in clusters. For this reason, from this point on, "counts" referring to values recorded by the instruments will be used instead of the word "impulse." The count process necessarily differs in some respects from the impulse noise process for the reasons just cited.

5.1 *Terminology and Definitions*

Some jargon has accumulated in the area of impulse noise studies; it is sometimes conflicting as well as confusing. The following terminology is adopted here.

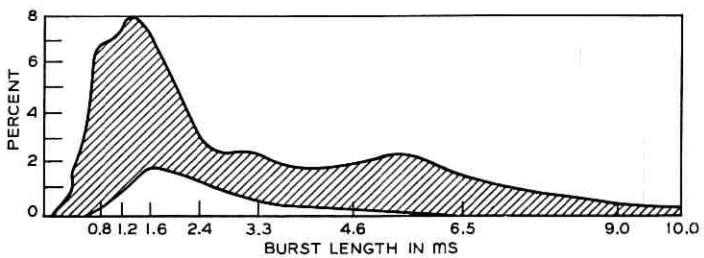

Fig. 10 — Envelopes of length density functions derived under definition of Fig. 9.

(*i*) Count—Refers to a number registered on the counter of an impulse noise threshold detecting type of measuring instrument, set at a specified level, during a specified measurement interval. (The count may be less than the actual number of impulses which exceeded the measurement threshold during the interval because of the finite maximum counting rate of the instrument.) Upper case C denotes the random variable count.

(*ii*) Impulse Noise Level—A level, expressed in decibels, at which the recorded count in a specified measurement interval is equal to some specified count denoted C_o.

(*iii*) Level Distribution—A distribution of levels, expressed in decibels (dBm, dBrn, and so on), taken across a number of channels, at which a specified count C_o is recorded in a specified measurement interval. Script "$\mathscr{l}$" denotes the random variable level.

(*iv*) Count Distribution—A distribution of counts observed in measurements on a number of channels taken at a specified level.

(*v*) Log-Count Distribution—A distribution of the logarithms of counts, expressed in decibels. Upper case "D" denotes the random variable log-count and is defined: $D = -10 \log_{10} (C/C_o)$, where C_o is an arbitrarily "specified reference count" greater than zero. C_o is arbitrary, but once picked it must be held constant for its associated level distribution.

(*vi*) Amplitude Distribution—A cumulative distribution of the peak amplitudes of individual impulses on a single channel. The average complementary distribution is linear on semi-log paper for counts in the range of interest; that is; $C < \approx 300$ in 30 minutes.

(*vii*) Slope—When spelled with "S", Slope refers to the slope of the peak amplitude distribution. Through common usage, the number assigned to Slope is the negative reciprocal of the slope of the peak amplitude distribution, designated "m", and expressed in decibels per decade of counts.

5.2 *General Comments on Level and Count Distributions*

Sample level distributions are constructed from data obtained through the use of multilevel impulse counters which record the number of counts at several levels, each separated by 2 to 6 dB, occurring during a prescribed measurement period. Level distributions may be constructed from the data depending upon the specific number of counts C_o in which one is interested. Suitable interpolation between the levels actually observed permits one to estimate the level at which some specified number of counts C_o actually occurred. Thus each level distribution has a number C_o associated with it, as well as a specific meas-

urement interval. The primary data used in this study consists of level distributions of 15 counts in 15 minutes and 90 counts in 30 minutes.[4,6]

The number of counts observed in a multilevel measurement tends to decrease exponentially as the level, in decibels, increases. Some departures from this rule are observed in individual measurements, but the average amplitude distribution taken over a large number of measurements in a single class of trunks appears to be exponential.[6] The number of counts C, at any level ℓ, may be estimated from the number of counts C', at level ℓ', by the empirically derived relation

$$C = C' \exp [(\ell' - \ell)/(Mm)]. \tag{2}$$

where $M = (\log_e 10)^{-1}$. Because different types of transmission facilities exhibit different impulse noise properties, the average noise level and average Slope vary over an appreciable range as facilities change. However, within a given type of facility or within a class of trunks, greater homogeneity is observed.[6] The model is therefore directed at a description of the noise as observed within populations of transmission channels on a single type of facility which are common to some larger grouping such as a trunk group.

5.3 *Assumptions*

The following two assumptions, supported by studies of available noise data, are basic to the model which is presented in Section 5.4.

(*i*) Level distributions for a specified count C_o are normal with mean ℓ_o and standard deviation σ_ℓ.

(*ii*) σ_ℓ is independent of C_o within a given trunk class. The first assumption is the most reasonable in view of the data; there are conflicting data concerning the second and it appears to be more valid for compandored facilities than for noncompandored facilities.[4,6] Under these assumptions, and one more stated below, it is shown below that the count and level distributions are completely described by the parameters associated with one level distribution: C_o, ℓ_o, σ_ℓ, and the Slope m. The Slope is estimated by the straight line connecting the mean of the level distributions for different choices of C_o.[6]

5.4 *The Model*

Any number of level distributions may be obtained from the data by choosing different values of C_o. As C_o increases, the corresponding level ℓ_o will decrease and trace a path in the count-level plane given by equation (2). A family of such level distributions form a probability

density surface above the count-level plane with normal cross sections parallel to the level axis. Such a surface is illustrated in Fig. 11. Under assumption (*ii*), lines parallel to the mean Slope are projections of constant probability density with the same functional form as equation (2). One of two cross sections may be taken which will define a probability density function. If the cross section is parallel to the count axis, a count distribution results. To see this more clearly, consider an experiment where impulse noise measurements are made on a group of similar trunks. A value for C_o is chosen and the associated level distribution with mean ℓ_o is found. The distribution will be normal with standard deviation σ_ℓ. Another value of C_o is chosen and a second level distribution is constructed. It will have the same standard deviation as the first. The experiment may be repeated any number of times to construct the family of distributions illustrated in Fig. 11.

In the experiment just described, the noise level ℓ associated with C_o was the random variable. Now suppose one wishes to let the count, or log-count, be the random variable while holding ℓ fixed. It is noted that equation (2) is the relationship between the means ℓ_o and C_o. Assume for the moment that equation (2) holds completely and is indeed a fixed relation between the two possible random variables, ℓ and C. Equation (2) may be rewritten, with $\ell' = \ell_o = 0$ and $C' = C_o$ as this constitutes an arbitrary shift in the decibel scale to define $\ell' = 0$:

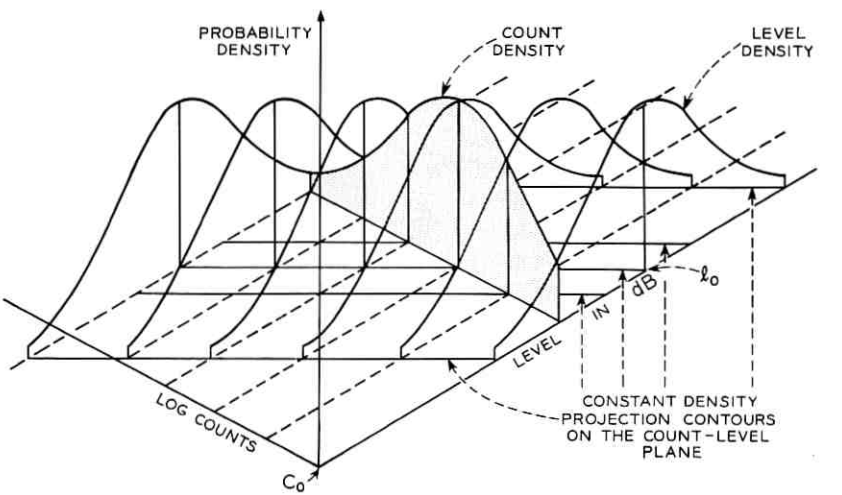

Fig. 11 — Probability density surface for the impulse noise count process on trunk groups.

$$\ell = -m \log_{10} (C/C_o). \tag{3}$$

Define $D = -10 \log (C/C_o)$. Then $\ell = Dm/10$, and the log-count distribution, the probability that D is $\leqq$ some value x, is by assumption 1,

$$P[D \leqq x] = P[\ell \leqq xm/10] = \frac{1}{\sigma_\ell (2\pi)^{\frac{1}{2}}} \int_{-\infty}^{xm/10} \exp (-\ell^2/2\sigma_\ell^2)\, d\ell. \tag{4}$$

The density function $f(x)$ is found to be

$$f(x) = \frac{m}{10\sigma_\ell (2\pi)^{\frac{1}{2}}} \exp [-m^2x^2/200\sigma_\ell^2]; \qquad -\infty \leqq x \leqq \infty. \tag{5}$$

Thus D is approximated by a normal distribution with mean zero and standard deviation $\sigma_D = 10\sigma_\ell/m$.*

In the previous derivation, equation (2), a relationship between expected values was assumed to hold as a mapping between the random variables ℓ and C or ℓ and D. To check the validity of this assumption a second experiment can be performed on the data collected in the first. The level ℓ can be held fixed at ℓ_o, and the count distribution at ℓ_o obtained by interpolation as described earlier. The observed log-count distribution may be compared with that derived in equation (5). This is done in Section 5.5.

5.5 *A Check on the Model Using Count Distribution Data*

Figure 13 is an example of count distributions derived in three different ways from a set of data consisting of 127 measurements on non-compandored carrier facility trunks 1,000 to 2,000 miles in length. The level distribution for these data, with $C_o = 15$, is slightly skew, the mean is 6.129 dBrn and the median 61.8 dBrn. The count distribution at 61.8 dBrn, obtained by interpolation between levels measured, is shown by the circled points on the figure. A point-by-point mapping from the level distribution by use of equation (2) is shown, as well as the log-normal one predicted by equation (5). The coincidence of all three sets of data is striking.

VI. THE TIME VARIABILITY OF IMPULSE NOISE

An additional check on the validity of this model is provided by its implications in the time variability of the noise. To see this, one additional assumption is made, and predicted and observed results serve to

* As a matter of interest, values of σ_D calculated from the 1964 Intertoll Trunk Survey (Ref. 6), are shown in Fig. 12.

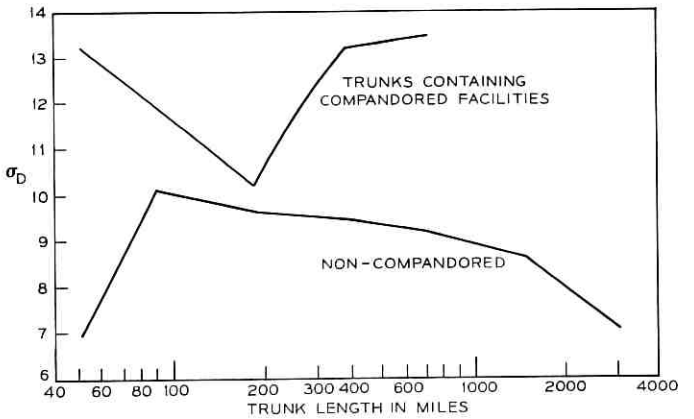

Fig. 12 — Values of standard deviation of log-count distributions $F(D)$ as observed on Bell System trunks.

validate both this additional assumption and the preceding model.

Consider making impulse noise measurements on a large number of channels at a fixed level ℓ_o and recording the cumulative count on the ith channel C_{ni} at times nT, $n = 1, 2, \cdots$. Now assume that the accrued count is a linear function of time so that each total count C_{ni} after time nT may be estimated by $C_{ni} = nC_{1i}$, where C_{1i} is the count in the first interval T on the ith channel. If the same reference count C_o is retained in the definition of D (log-counts), for all time intervals, then the mean value of D will increase as log (n) but the variance of the count distribution (as opposed to the log-count) behaves differently however, as shown by the following.

Under the assumption that equation (2) holds as a mapping between ℓ and C, the distribution of C (counts) may also be derived:

$$P[C \leqq y] = P[\ell \geqq -m \log (y/C_o)]$$

$$= \frac{1}{\sigma_\ell (2\pi)^{\frac{1}{2}}} \int_{-m \log y/C_o}^{\infty} \exp (-\ell^2/2\sigma_\ell^2)\, d\ell, \tag{6}$$

and the density of C is approximated by the log-normal:

$$f(y) = \frac{Mm}{\sigma_\ell (2\pi)^{\frac{1}{2}}}\, y^{-1} \exp \left[-\frac{M^2 m^2}{2\sigma_\ell^2} \ln^2 (y/C_o) \right];$$

$$0 \leqq y \leqq \infty,$$

$$\ln \equiv \log_e .$$

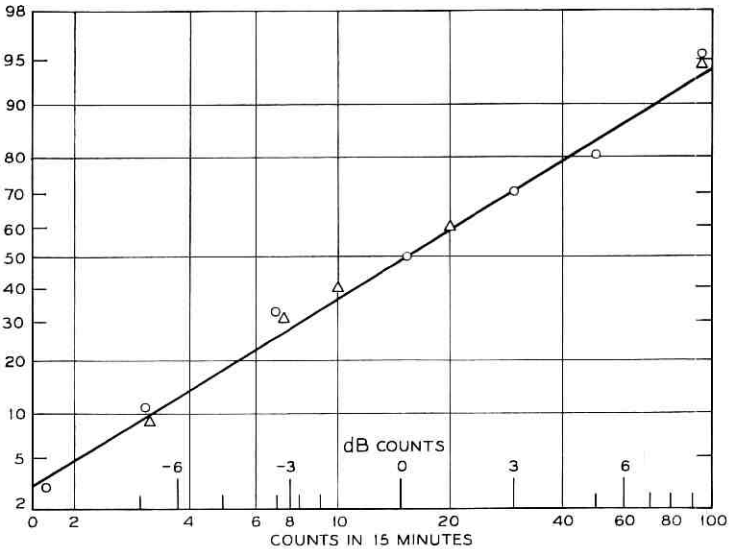

Fig. 13 — Empirical verification of the distribution of counts as derived from distribution of levels of constant counts. Distribution of counts in 15-minute measured periods on one class of trunks. ○ as measured; △ constructed from level distribution; — predicted from equation (5).

The rth moment of C is then found to be

$$E[C^r] = C_o^r \exp [r^2/(4a)]; \qquad a = \frac{M^2 m^2}{2\sigma_t^2} \approx \frac{9.5}{\sigma_D^2}, \tag{7}$$

and the variance, $\sigma_C^2 = C_o^2(e^{1/a} - e^{1/2a}) = C_o^2 A$. Now, as the measurement interval is increased as above, C_o is replaced by nC_o and $\sigma_C^2(nC_o) = n^2 C_o^2 A$. The variance of the count distribution increases as the square of time if the mean increases linearly.

Note from equation (7), that the mean of the count distribution is not equal to the reference count C_o which is associated with the level distribution. The two are related as*

$$\langle C \rangle_{av} = C_o \exp [\sigma_t^2/(2m^2 M^2)]$$
$$\approx C_o(1.027)^{\sigma_D^2}. \tag{8}$$

Thus, the mean of the count distribution at level ℓ_o is always greater than the reference count C_o. Furthermore, from the definitions of the

* Note from Fig. 12 that σ_D may be as large as 13.5 so $\langle C \rangle_{av}$ may be as large as 130 times C_o.

level distribution and the quantity D, C_o is equal to the median of the count distribution. Solving equation (8) for σ_D yields

$$\sigma_D \approx 8.7(\log \langle C\rangle_{av}/C_o)^{\frac{1}{2}},$$

and an estimate of the variance of the log-count distribution may be made from the mean and median of the count distribution. This relation should be very useful in practice.

Now consider measurements of length K_2T taken on a number of channels with the counts recorded after K_1T and K_2T, $K_2 > K_1$. Let x be a random variable that takes the value of the count at K_1T, and y one that takes the value of the count at K_2T. If it were true that the count on each channel is a linear function of time, then for the ith channel measurement, $y_i = (K_2/K_1)x_i$ and the coefficient of correlation $\rho_{xy} = 1$. Such correlation coefficients were calculated for several sets of data. The results are presented in Table III. T was equal to 5 minutes in all cases. The notation ρ_{ij} indicates the correlation between the counts at the end of i 5-minute intervals with that after j 5-minute intervals. The mean ratio of the count after j intervals to the count after i intervals and the ratio of the variance after j and i intervals is also given. The expected values, derived from the model, are given in each case (in parentheses), as well as the observed values. While the correlation coefficients are not all as close to unity as one might hope, especially for the 5-minute versus 30-minute measurements ($i = 1$, $j = 6$), the mean and variance do appear to increase directly and as the square of time respectively.

On the basis of the data shown in Table III and Fig. 13, the model appears to be an adequate description of the observed behavior of the impulse noise on transmission facilities as viewed through impulse noise measuring sets.

TABLE III—CORRELATION COEFFICIENTS AND RATIOS OF MEANS AND VARIANCES*

i, j	ρ_{ij}	μ_j/μ_i	s_j^2/s_i^2	Sample Size
1,2	(1) 0.87	(2) 2.04	(4) 3.70	87
1,2	(1) 0.92	(2) 1.97	(4) 5.00	76
1,2	(1) 0.90	(2) 1.98	(4) 3.98	216
1,3	(1) 0.96	(3) 3.10	(9) 9.90	161
1,3	(1) 0.98	(3) 2.90	(9) 8.50	168
2,4	(1) 0.97	(2) 2.06	(4) 3.60	93
1,6	(1) 0.58	(6) 6.76	(36) 46	161

* For counts observed after i and j 5-minute intervals. Predicted values in parentheses are followed by observed values.

VII. SUMMARY

The following relations and conclusions come from the model presented and the data upon which it is based.

(*i*) Level distributions are normal with mean ℓ_o and variance σ_ℓ^2.

(*ii*) Count distributions are log-normal with mean $\langle C \rangle_{av}$ which is linearly related to the length of the measurement interval, and variance, σ_C^2, which is proportional to the square of the interval. Equivalently, log-count distributions are normal with mean proportional to the logarithm of the measurement interval and variance, σ_D^2, independent of interval.

(*iii*) σ_ℓ is dependent upon the class of trunk but is independent of C_o, an arbitrary reference count greater than zero.

(*iv*) $\sigma_D = 10\sigma_\ell/m$, m is a measure of the slope of the distribution of noise peak amplitudes.

(*v*) $\sigma_D \approx 8.7(\log_{10} \langle C \rangle_{av}/C_0)^{\frac{1}{2}}$.

(*vi*) $\langle C \rangle_{av} \approx C_o(1.027)^{\sigma_D{}^2}$

(*vii*) The mean of the count distribution, $\langle C \rangle_{av} = C_o e^{1/(4a)}$ and the variance

$$V(C) = C_o^2 e^{1/(2a)}[e^{1/(2a)} - 1]; \qquad a = \frac{m^2 M^2}{2\sigma_\ell^2}; \qquad \frac{1}{M} = \log_e 10.$$

(*viii*) The median of a count distribution, taken at level ℓ_o, is equal to C_o and the mean $\langle C \rangle_{av}$, may be 100 times C_o. Expected count by itself is accordingly a very poor statistic for describing impulse noise. However, the mean and the median completely describe the count or log-count distributions.

The model helps to explain the apparent erratic behavior of impulse noise measurements. Any measurement is a sample taken from the bivariate sample space illustrated in Fig. 11. The fact that the distribution of counts is log-normal also accounts for the great fluctuation in the count observed on successive measurements on a given channel. It is shown however that the average rate of occurrence is reasonably constant with time for intervals from 5 to 30 minutes.

VIII. ACKNOWLEDGMENTS

Numerous people have contributed to the collection and analysis of the data that has resulted in the preparation of this paper. Special accord is to be given to Messrs. D. A. Lewinski and I. Nasell, whose efforts of several years ago gave rise to the system-wide statistically sound surveys

which provided the fundamental data; and to the many people from Bell Telephone Laboratories, the American Telephone and Telegraph Company, and associated companies who participated in those surveys. My appreciation of the efforts of Messrs. D. P. Dumas, C. R. Ellison, and B. F. Hellrigel must also be noted for their significant contributions in the data processing, reduction, and analysis. I wish to express my thanks for the many stimulating discussions with N. A. Marlow during the course of this work and for a critical review of the first draft.

REFERENCES

1. Aikens, A. J., and Lewinski, D. A., "Evaluation of Message Circuit Noise," B.S.T.J., *39*, No. 4 (July 1960), pp. 911–931.
2. Favin, D. L., "The 6A Impulse Noise Counter," Bell Laboratories Record, *41*, No. 3 (March 1963), pp. 100–102.
3. Mertz, P., "Impulse Noise and Error Performance in Data Transmission," AD 614 416, U. S. Department of Commerce, Clearing House for Scientific and Technical Information, April 1965.
4. Fennick, J. H., and Nasell, I. E. O., "The 1963 Survey of Impulse Noise on Bell System Carrier Facilities," 1966 IEEE Int. Conv. Record, Part I, Wire and Data Communication, March 1966, pp. 149–156.
5. Fennick, J. H., "A Method for the Evaluation of Data Systems Subject to Large Noise Pulses," 1965 IEEE Int. Conv. Record, Part I, Communications 1, Wire and Data Communication, March 1965, pp. 106–110.
6. Ellison, C. R., Holmstrom, J. R., and Nasell, I. E. O., "The Transmission Performance of Bell System Intertoll Trunks," B.S.T.J. *47*, No. 8 (October 1968), pp. 1561–1614.
7. Berger, J. M., and Mandelbrot, B., "A New Model for Error Clustering in Telephone Circuits," IBM J. Res. and Development, *7*, (July 1963), pp. 222–236.
8. Frichtman, B. D., "A Binary Channel Characterization Using Partitioned Markov Chains," IEEE Trans. Inform. Theory, *IT-13*, No. 2 (April 1967), pp. 221–227.
9. Kaenel, R. A., and others, "Burst Measurements in the Time Domain," Trans. Audio and Electroacoustics, *AU-14*, No. 3 (September 1966), pp. 115–121.

The Determinability of Classes of Noisy Channels

By LEONARD J. FORYS

(Manuscript received June 30, 1969)

This paper is concerned with the identification of a fairly general class of nonlinear operators using corrupted measurements. A precise mathematical definition of identification is presented and the relationship between a priori information and identification is studied. The a priori information is represented as a subset of a metric space of nonlinear operators. Necessary and sufficient conditions are developed to answer the question "When is identification possible?"

I. INTRODUCTION

A large body of literature already exists for the problem of identifying a control system or communication channel with noisy measurements. In the usual identification problems, a certain structure is assumed at the outset in order to reduce the identification problem to one of parameter estimation. The absence of such parametrization increases the difficulty of the problem substantially. It is often not clear if identification is even possible.

In this paper we are concerned with the determinability (identifiability) of quite general nonlinear operators whose outputs are corrupted by additive gaussian noise. We introduce a norm on this space of nonlinear operators and define precisely what we mean by determinability. Loosely speaking, we say that we can determine an operator H if we can choose a finite observation interval $[0, T]$, a test signal with constrained peak value over this interval, a finite set of linear measurements over $[0, T]$, and an estimate $\hat{H}$ of H which is a continuous function of our measurements such that $\hat{H}$ is close to H in norm with high probability.

The question of determinability is of course intimately related to the kind of *a priori* knowledge one has of the operator. We represent this *a priori* information by saying that the operator H belongs to a subset

$\mathfrak{D}$ of possible operators. We derive conditions on $\mathfrak{D}$ which are sufficient for determinability. We also show that most of these conditions are in fact necessary for the determination of H.

Our results are motivated by the work on the determinability of noiseless channels done by Root, Prosser, and Varaiya.[1-4] They derive necessary and sufficient conditions to estimate a noiseless channel closely with a "one-shot" experiment. These conditions are similar to those presented here. Some work on the noisy problem has been done by Root.[5] His approach and results are fundamentally different than those presented in this paper. Root investigated a class of stochastic nonlinear operators represented by a Volterra series whose kernels are gaussian random variables. He derived necessary and sufficient conditions for the second moments of the kernels to be determinable.

II. PRELIMINARIES

The types of channels to be considered can be described as follows. The input signal x and observed signal w are related via the operator equation

$$w(t) = [Hx](t) + z(t) \qquad t \in [0, \infty) \tag{1}$$

where H is an operator and z is zero mean white gaussian noise† with covariance $Ez(t)z(\tau) = \delta(t - \tau)$. (The colored noise case will be treated separately in Section V.)

We constrain our input functions x to have peak value less than s,

† The noise term $z(t)$ in equation (1) must be interpreted symbolically since white noise cannot be parametrized with a time variable, but must properly be parametrized with an element of a space of "testing functions." However, we deal only with functionals of $w(t)$ of the form

$$\int_0^b w(t)\phi(t)\,dt,$$

where $\phi \in L_2(0, b)$, or with quantities derivable from these functionals. Hence we can define

$$\int_0^b z(t)\phi(t)\,dt$$

to mean

$$\int_0^b \phi(t)\,d\zeta(t)$$

where $\zeta(t)$ is Brownian motion and the operations to be performed are readily justified.

that is,

$$x \in L_\infty(s) = \{x \mid x \text{ is a real valued measurable function on } [0, \infty) \text{ and } |x(t)| \leqq s \text{ for all } t \in [0, \infty)\}.$$

If we let $\| \ \|_2$ denote the norm on $L_2[0, \infty)$ and define the projection operator P_T by

$$[P_T x](t) = x(t) \quad \text{for} \quad t \leqq T$$
$$= 0 \quad \text{for} \quad t > T$$

then

$$\| P_T x \|_2 = \left(\int_0^T x^2(t)\, dt \right)^{\frac{1}{2}} \leqq s(T)^{\frac{1}{2}} \quad \text{for all} \quad x \in L_\infty(s).$$

The types of operators which we consider are assumed to belong to the space $\mathfrak{H}$. The space $\mathfrak{H}$ is defined: if $H \in \mathfrak{H}$ then

(*i*) $H : L_\infty(s) \to L_{2e}$

where $L_{2e} = \{y \mid y$ is a real valued, measurable function on $[0, \infty), \| P_T y \|_2 < \infty$ for all $T > 0\}$,

(*ii*) H is causal; that is, for all $T > 0$, $x \in L_\infty(s)$, $P_T H x = P_T H P_T x$,

(*iii*) $\| H \| < \infty$.

Using the usual definitions of addition of operators and multiplication by scalars, the norm of H, $\| H \|$ is defined as:

$$\| H \| = \sup_{\substack{T>0 \\ x \in L_\infty(s) \\ \| P_T x \|_2 \neq 0}} \frac{\| P_T H x \|_2}{\| P_T x \|_2}.$$

We consider H to be the zero operator† if $\| H \| = 0$. It is then easy to show that $\| \ \|$ satisfies the norm axioms. Obviously $\| H \| \geqq 0$ for all $H \in \mathfrak{H}$ and $\| \lambda H \| = |\lambda| \, \| H \|$ for all scalars λ. The triangle inequality is also satisfied since

$$\| H + K \| = \sup \frac{\| P_T(Hx + Kx) \|_2}{\| P_T x \|_2} = \sup \frac{\| P_T H x + P_T K x \|_2}{\| P_T x \|_2}$$

† The equivalence classes defined in this manner are not unreasonable. In fact, if $\| H \| = 0$ then $\| P_T H x \|_2 = 0$ for all $x \in L_\infty(s)$, $\| P_T x \|_2 \neq 0$ and all $T > 0$. As far as we are concerned this is the zero operator since Hx is then the zero function in the $L_2(0, \infty)$ sense.

$$\leqq \sup \left[\frac{|| P_T H x ||_2 + || P_T K x ||_2}{|| P_T x ||_2} \right]$$

$$\leqq \sup \frac{|| P_T H x ||_2}{|| P_T x ||_2} + \sup \frac{|| P_T K x ||_2}{|| P_T x ||_2} = || H || + || K ||$$

where the supremums are taken over all $T > 0$, $x \in L_\infty(s)$, $|| P_T x ||_2 \neq 0$.

If we consider the metric induced by the norm $|| \ ||$ then $\mathfrak{IC}$ is a complete metric space. The proof of this proposition is contained in the appendix. The completeness property is crucial to Theorem 2 of this paper.

The space $\mathfrak{IC}$ includes many types of operators familiar to those in communication and control theory. Linear time invariant convolution operators whose kernels are either in $L_1(0, \infty)$ or $L_2(0, \infty)$ are in $\mathfrak{IC}$. If these operators are cascaded with a memoryless nonlinearity having bounded slope, the composite operators are also in $\mathfrak{IC}$. Operators described by certain nonlinear dynamical systems are also in $\mathfrak{IC}$. Let $x \in L_\infty(s)$ be the input to the following dynamical system and let y be the output:

$$\dot{\underline{q}}(t) = \underline{f}(\underline{q}(t), x(t), t), \qquad \underline{q}(0) = 0$$

$$f : R^n \times R \times R \to R^n$$

$$y(t) = g(\underline{q}(t))$$

$$g : R^n \to R$$

with

$$|g(\underline{q})| \leqq K_1 |\underline{q}|, \qquad |f(\underline{q}, x, t)| \leqq K_2 |\underline{q}| + K_3 |x|$$

for all $\underline{q} \in R^n$, $|x| < s$, $t > 0$. Assume also that for each $x \in L_\infty(s)$ there exists a solution to the differential equation. Then, via the Bellman-Gronwall inequality we see that

$$|\underline{q}(t)| \leqq K_3 \int_0^t e^{K_2(t-\tau)} |x(\tau)| \, d\tau.$$

Hence,

$$\left(\int_0^T |\underline{q}(t)|^2 \, dt \right)^{\frac{1}{2}} \leqq K_2 K_3 \left(\int_0^T x^2(t) \, dt \right)^{\frac{1}{2}}$$

and

$$\left(\int_0^T |y(t)|^2 \, dt \right)^{\frac{1}{2}} \leqq K_1 K_2 K_3 \, || P_T x ||_2 .$$

Thus, the operator described is in $\mathfrak{IC}$ with norm bounded by $K_1 K_2 K_3$.

Subsets of $\mathfrak{H}$ will be used to represent the a priori information in an identification problem. We call a subset $\mathfrak{D}$ of $\mathfrak{H}$ determinable if every member of $\mathfrak{D}$ can be identified. The determinability of a subset $\mathfrak{D}$ depends of course on our definition of identification. We would like to consider only those identification procedures which could theoretically be implemented in real time. The identification procedures which we are concerned with must have the following properties. To identify H we must be able to

(*i*) choose a finite observation interval,
(*ii*) select an input function with constrained peak value,
(*iii*) perform linear measurements on the noisy observations generated by this input, and
(*iv*) operate on these measurements to yield an estimate of H which is a continuous function of these measurements,

so that our estimate of H is close to H with high probability.

The properties of such an identification procedure are physically very appealing. We obviously must be able to identify within a finite period of time. The peak value restriction is the usual kind of input constraint used in communication theory. Linear measurements are easily implemented and tend to reduce the sensitivity to unknown biases as does the continuity requirement on the estimate. Finally, we are usually satisfied to identify to within a small tolerance.

For $H \in \mathfrak{H}$ and channel model given by equation (1) we may specify our definition of identification even further. A *linear measurement* over the time interval $[0, T]$ is a finite collection of bounded linear functionals† (p_i, w), $i = 1, 2, \cdots N$, $p_i \in L_2[0, T]$ defined when $P_T w \in L_2[0, T]$ and

$$w(t) = [Hx](t) + z(t), \qquad 0 \leqq t \leqq T$$

is the received waveform with $H \in \mathfrak{H}$, $x \in L_\infty(s)$. We say that a class $\mathfrak{D} \subset \mathfrak{H}$ of channel operators is *determinable* if given arbitrary positive constants ϵ and η, there exists a finite observation interval $[0, T]$, an input (test) signal $x \in L_\infty(s)$, a linear measurement $[(p_1, w), (p_2, w), \cdots, (p_N, w)]$ over $[0, T]$, and a continuous function $g : R^N \to \mathfrak{H}$ such that for each $H \in \mathfrak{D}$,

$$\text{Probability } (\| H - \hat{H} \| > \epsilon) < \eta$$

where $\hat{H} = g[(p_1, w), (p_2, w), \cdots, (p_N, w)]$. Thus, if $\mathfrak{D}$ is determinable,

† The symbol (f, h) is used to represent the inner product in $L_2[0, T]$; that is, $(f, h) = \int_0^T f(t)h(t)\, dt$.

we can "identify" any element of $\mathfrak{D}$ to within any specified accuracy with sufficient processing and long enough observation time.

The bulk of this paper is related to answering the following question. What structure must $\mathfrak{D}$ have in order to be determinable? Theorem 1 derives sufficient conditions on $\mathfrak{D}$ in order to be determinable. The key condition is compactness. Theorem 2 indicates that this condition is in fact necessary for determinability. A number of corollaries are given which interpret these results for the case where $\mathfrak{D}$ is composed of linear convolution operators.

III. SUFFICIENT DETERMINABILITY CONDITIONS

Despite the generality of our class of operators and the rather rigid nature of allowable identification schemes only two conditions guarantee the determinability of a subset of operators. Both conditions are somewhat obvious. One condition insures that the class may be approximated closely by a finite number of elements; the other insures that a test signal exists which will produce sufficiently dissimilar responses for dissimilar channels. These conditions are rigorously stated in Theorem 1.

Theorem 1: *Let $\mathfrak{D}$ be a subset of $\mathfrak{K}$ having the following properties*:

(*i*) *the closure of $\mathfrak{D}$ is compact* (*thus $\bar{\mathfrak{D}}$ is also bounded*; *that is, there exists a constant* $R > 0$, *such that* $\| H - K \| < R$ *for all* $H, K \in \bar{\mathfrak{D}}$)

(*ii*) *given any* $\delta > 0$ *there exists an unbounded sequence* $\{T_i\}$, *a sequence of inputs* $x_i \in L_\infty(s)$ *and a positive number* r *such that*

$$\| P_{T_i}(Hx_i - Kx_i) \|_2^2 > rT_i$$

for all pairs $H, K \in \bar{\mathfrak{D}}$ *for which* $\| H - K \| \geqq \delta$. *Then $\mathfrak{D}$ is a determinable subset of $\mathfrak{K}$.*

Proof: Since the proof of this theorem is lengthy, we give here a brief, rough description of the key steps involved which the reader may use as a guide through the mathematical details.

(*i*) Using (*ii*) of Theorem 1 we select an input x_i to give sufficient separation of outputs over $[0, T_i]$ for sufficiently dissimilar channels.

(*ii*) We then approximate the class $\mathfrak{D}$ to within a judiciously chosen accuracy by a finite number of elements.

(*iii*) The actual received signal due to the input selected in (*i*) of this proof is correlated over $[0, T_i]$ with the calculated outputs of the channels selected in step (*ii*) of this proof.

(*iv*) If one of these correlations is larger than the others by some amount we select as our estimate the corresponding element of the

approximating class that yielded this correlation. If there is no such correlation we assign an arbitrary rule so as to make the identification procedure a continuous function of the correlated values.

(*v*) We finally show that as i (and hence T_i) increases, the probability that there will not be a correlation larger than the others by some prescribed amount goes to zero. In addition, we show that the probability that our identification procedure yields an estimate which is further apart in norm from the actual channel than is desired is vanishingly small as i increases.

The formal statement of the proof follows below.

We may assume that $\mathfrak{D}$ is closed, since subsets of a determinable set of channels are determinable. Using assumption (*ii*) of Theorem 1 with $\delta = 3\epsilon/4$ we have that there exists an unbounded sequence $\{T_i\}$, a positive number r, and for each i an input signal $x_i \in L_\infty(x)$ such that for all pairs $H, K \in \mathfrak{D}$ with $\| H - K \| > 3\epsilon/4$

$$\| P_{T_i}(Hx_i - Kx_i) \|_2^2 \geqq rT_i . \tag{2}$$

In what follows we will denote the operator which we wish to identify by H. Since $\mathfrak{D}$ is closed, by assumption (*i*) of Theorem 1, it is also compact and hence totally bounded (see for example Ref. 6, p. 22). Therefore, given any $T_i \in \{T_i\}$ we can choose a finite number of balls of radius $r_0 = \min \{r^3/2s, \epsilon/4\}$ with centers $H_\alpha \in \mathfrak{D}$, $\alpha = 1, 2, \cdots, M$ to cover $\mathfrak{D}$. There may be operators $H_l, H_k \in \{H_\alpha\}$ for which $\| P_{T_i}(H_l x_i - H_k x_i) \|_2 = 0$, in which case retain only the H_α's with the lowest subscript. Thus we have a subset of $\{H_\alpha\}$ which we label $\{H_\beta\}$ for which $\| P_{T_i}(H_l x_i - H_k x_i) \|_2 > \theta_i > 0$ for some θ_i and all $H_l, H_k \in \{H_\beta\}$. For convenience order the $\{H_\beta\}$ so that $\| H - H_\beta \| \leqq 3\epsilon/4$ for $\beta = 1, 2, \cdots, N_0 - 1$ and $\| H - H_\beta \| > 3\epsilon/4$ for $\beta = N_0, N_0 + 1, \cdots, N$, $N \leqq M$.

We can now choose an appropriate linear measurement over the interval $[0, T_i]$. We define the linear measurement $\underline{m}(w) = \{f(w, 1), f(w, 2), \cdots f(w, N)\}: f(w, \beta) = (w, 2H_\beta x_i)$, $\beta = 1, 2, \cdots N$ where the inner product is defined over the interval $[0, T_i]$. Thus for each received waveform $w(t)$, the linear measurement gives us a point in R^N. From this measurement we will determine an estimator function $g : R^N \to \mathfrak{C}$. We first partition R^N into $N + 1$ disjoint subsets: $A_1, A_2, \cdots, A_N, B$, with

$$A_j = \{\underline{a} = (a_1, a_2, \cdots a_N): a_j - a_k > (H_j x_i, H_j x_i) - (H_k x_i, H_k x_i) + \theta_i/T_i, \quad k = 1, 2, \cdots N, \quad k \neq j\}$$

and B the remainder of R^N,

$$B = \left(\bigcup_{j=1}^{N} A_j\right)^c.$$

The disjointness of the above subsets of R^N is easily verified by making use of the fact that $\theta_i/T_i > 0$. The estimator function is defined in terms of this partition:

$$g(\underline{m}) = H_j \qquad \text{if} \quad \underline{m}(w) \in A_j$$

$$g(\underline{m}) = \sum_{i=1}^{N} \alpha_i(w) H_i \quad \text{if} \quad \underline{m}(w) \in B$$

where†

$$\alpha_i(w) = \frac{\prod_{j \neq i} d(\underline{m}(w), A_j)}{d(\underline{m}(w), A_i) + \prod_{j \neq i} d(\underline{m}(w), A_j)}$$

and

$$d(x, A) = \inf_{y \in A} | x - y |.$$

It is not difficult to show that g is a continuous mapping from R^N into $\mathfrak{IC}$. Having given the identification scheme we now show that for any $H \in \mathfrak{IC}$, $\epsilon > 0$

$$P\{|| H - \hat{H} || > \epsilon\} \underset{T_i \to \infty}{\to} 0.$$

Recalling the definition of B, A_i and the labeling convention we have used, we see that

$$P\{|| H - g(\underline{m}(w)) || > \epsilon\} \leqq P\{\underline{m}(w) \in B\} + P\left\{\underline{m}(w) \in \bigcup_{j=N_0}^{N} A_j\right\}$$

$$= P\left\{\underline{m}(w) \in \bigcap_{j=1}^{N} A_j^c\right\} + P\left\{\underline{m}(w) \in \bigcup_{j=N_0}^{N} A_j\right\}. \qquad (3)$$

Let us first concentrate on obtaining bounds for the first term on the right side of equation (3). We rewrite A_j as $A_j = (\bigcup_{k \neq j} F_{jk})^c$ where

$$F_{jk} = \{\underline{a} = (a_1, a_2, \cdots, a_N) : a_j - a_k \leqq (H_j x_i, H_j x_i) - (H_k x_i, H_k x_i) + \theta_i/T_i\}.$$

Thus

† It turns out that the form of $\alpha_i(w)$ is irrelevant since we show that $P[m(\underline{w}) \in B]$ vanishes as T_i increases. It is merely included to make the estimator function continuous.

$$\bigcap_{i=1}^{N} A_i^c = \bigcap_{j=1}^{N} \left(\bigcup_{k \neq j} F_{jk}\right). \tag{4}$$

Applying DeMorgan's rules to equation (4), and after some thought, we see that

$$\bigcap_{j=1}^{N} A_j^c = \bigcup_{l=1}^{(N-1)^N} D_l \tag{5}$$

where D_l has the form

$$D_l = F_{1l_1} \cap F_{2l_2} \cap \cdots \cap F_{Nl_N}$$

with $l_i \neq j$ for all j. We can upper bound $P\{\underline{m}(w) \in D_l\}$ by

$$\sup_{\substack{k,j \\ k \neq j}} P\left\{-\frac{N\theta_i}{T_i} + (H_i x_i , H_i x_i) - (H_k x_i , H_k x_i) \leqq f(w, k) - f(w, j) \right.$$

$$\left. \leqq \frac{N\theta_i}{T_i} + (H_k x_i , H_k x_i) - (H_i x_i , H_i x_i)\right\}. \tag{6}$$

To see this, define $q(w, k) = f(w, k) - (H_k x_i , H_k x_i)$. Then $P\{\underline{m}(w) \in D_l\}$ is the probability of the N events $q(w, 1) - q(w, l_1) \leqq \theta_i/T_i$, $q(w, 2) - q(w, l_2) \leqq \theta_i/T_i \cdots q(w, N) - q(w, l_N) \leqq \theta_i/T_i$ occurring simultaneously. Suppose $l_1 = k$. Then consider the two events $q(w, 1) - q(w_1 , l_1) = q(w, 1) - q(w, k) \leqq \theta_i/T_i$ and $q(w, k) - q(w, l_k) \leqq \theta_i/T_i$. If $l_k = 1$, then these two events are contained in the event $-\theta_i/T_i \leqq q(w, 1) - q(w, k) \leqq \theta_i/T_i$. If $l_k = j \neq 1$ then consider the three events

$$q(w, 1) - q(w, k) \leqq \theta_i/T_i ,$$

$$q(w, k) - q(w, j) \leqq \theta_i/T_i ,$$

$$q(w, j) - q(w, l_j) \leqq \theta_i/T_i .$$

If $l_j = 1$, then these three simultaneous events are contained in the event $-\theta_i/T \leqq q(w, 1) - q(w, j) \leqq 2\theta_i/T_i$. If $l_j = k$, then these three simultaneous events are contained in the event $-\theta_i/T \leqq q(w, k) - q(w, j) \leqq \theta_i/T_i$. Continuing in this fashion we obtain the bound in equation (6).

Since $q(w, k) - q(w, j)$ is gaussian, we can bound the value of the expression in equation (6) quite easily.

Let

$$a_{kj} = E[q(w, k) - q(w, j)]$$

$$= || P_{T_i}(Hx_i - H_j x_i) ||_2^2 - || P_{T_i}(Hx_i - H_k x_i) ||_2^2 \tag{7}$$

and

$$\sigma_{kj}^2 = \text{Var}\,[q(w,k) - q(w,j)] = 4\,||P_{T_i}(H_k x_i - H_j x_i)\,||_2^2 > 4\theta_i^2\,. \quad (8)$$

Hence,

$$P\left\{-\frac{N\theta_i}{T_i} \leqq q(w,k) - q(w,j) \leqq \frac{N\theta_i}{T_i}\right\}$$

$$= (2\pi)^{-\frac{1}{2}} \int_{-(N\theta_i/T_i\sigma_{kj}) - a_{kj}/\sigma_{kj}}^{N\theta_i/T_i\sigma_{kj} - a_{kj}/\sigma_{kj}} \exp\left(-\frac{z^2}{2}\right) dz$$

$$\leqq (2\pi)^{-\frac{1}{2}} \int_{-(N\theta_i/T\sigma_{kj})}^{N\theta_i/T_i\sigma_{kj}} \exp\left(-\frac{z^2}{2}\right) dz$$

$$\leqq (2\pi)^{-\frac{1}{2}} \int_{-(N/2T_i)}^{N/2T_i} \exp\left(-\frac{z^2}{2}\right) dz \leqq (2\pi)^{-\frac{1}{2}} \int_{-(M/2T_i)}^{M/2T_i} \exp\left(-\frac{z^2}{2}\right) dz \quad (9)$$

(recall that $N \leqq M$).

Using equations (9) and (5) we see that

$$P\{\underline{m}(w) \in \bigcap_{i=1} A_i^c\} \leqq \sum_{l=1}^{(N-1)N} P\{\underline{m}(w) \in D_l\}$$

$$< (M-1)^M (2\pi)^{-\frac{1}{2}} \int_{-(M/T_i)}^{M/T_i} \exp\left(-\frac{z^2}{2}\right) dz. \quad (10)$$

Since the right side of equation (10) goes to zero as T_i increases we can choose a $T \in \{T_i\}$ large enough so that this term is less than $\eta/2$. We now bound the second term on the right side of equation (3):

$$P\left\{\underline{m}(w) \in \bigcup_{j=N_0}^{N} A_j\right\} = \sum_{j=N_0}^{N} P\{\underline{m}(w) \in A_j\}. \quad (11)$$

Recall that

$$A_j = (\bigcup_{k \neq j} F_{jk})^c = \bigcap_{k \neq j} F_{jk}^c\,.$$

Hence

$$P\left\{\underline{m}(w) \in \bigcup_{j=N_0}^{N} A_j\right\} = \sum_{j=N_0}^{N} P\{\underline{m}(w) \in \bigcap_{k \neq j} F_{jk}^c\}. \quad (12)$$

Observe that for all $k \neq j$

$$P\{\underline{m}(w) \in \bigcap_{k \neq j} F_{jk}^c\} \leqq P\{\underline{m}(w) \in F_{jk}^c\} \quad (13)$$

$$= P\left\{q(w,j) - q(w,k) > \frac{\theta_i}{T_i}\right\} \quad (14)$$

$$= \int_{\theta_i/T_i - a_{jk}/\sigma_{jk}}^{\infty} \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{z^2}{2}\right) dz. \tag{15}$$

Since $\mathfrak{D}$ was covered by balls of radius r_0, there exists at least one integer $\tilde{k} < N_0$ such that $\| H - H_{\tilde{k}} \| < r_0 \leqq \epsilon/4$ and hence $\| P_{T_i}(Hx_i - H_{\tilde{k}}x_i)\|^2 < r_0^2 s^2 T_i$. Note also that since $\| H_j - H \| > 3\epsilon/4$ for $j \geqq N_0$, $\| P_{T_i}(Hx_i - H_j x_i) \|^2 > rT_i$. Hence,

$$\begin{aligned} -a_{j\tilde{k}} &= \| P_{T_i}(Hx_i - H_j x_i) \|_2^2 - \| P_{T_i}(Hx_i - H_{\tilde{k}}x_i) \|_2^2 \\ &\geqq rT_i - r_0^2 s^2 T_i \geqq \left(r - \frac{r}{4}\right) T_i = \tfrac{3}{4} r T_i . \end{aligned} \tag{16}$$

Recalling that $\mathfrak{D}$ was bounded,

$$\sigma_{j\tilde{k}}^2 = 4 \| P_{T_i}(H_j x_i - H_{\tilde{k}} x_i) \|_2^2 \leqq 4R^2 s^2 T_i . \tag{17}$$

Using equations (16) and (17) in equation (15) we see that

$$P\{\underline{m}(w) \in \bigcap_{k \neq j} F_{jk}^c\} \leqq P\{\underline{m}(w) \in F_{j\tilde{k}}^c\} \leqq \int_{3rT_i^{\frac{1}{2}}/16Rs}^{\infty} (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{z^2}{2}\right) dz. \tag{18}$$

Hence from equation (11) we see that

$$P\left\{\underline{m}(w) \in \bigcup_{j=N_0}^{N} A_j\right\} \leqq M \int_{3rT_i^{\frac{1}{2}}/16Rs}^{\infty} (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{z^2}{2}\right) dz. \tag{19}$$

Thus we can select a $T \in \{T_i\}$ so that this term is less than $\eta/2$. This T makes $P\{\| H - \hat{H} \| > \epsilon\} < \eta$ for all $H \in \mathfrak{D}$.

The identification technique proposed in the above proof is not necessarily a practical technique. Our intent is to indicate the possibility of identification rather than to derive easily implementable techniques. Notice, however, that since the measurements are linear functionals on $L_2(0, T)$ they are iterative in nature because of the integral representation of such functionals.

Theorem 1 gives sufficient conditions for determinability. Theorem 2 indicates that some of these conditions are in fact necessary for identification.

IV. NECESSARY DETERMINABILITY CONDITIONS

In this section we show that the approximability condition given by condition (*i*) of Theorem 1 is in fact necessary. We also show that a type of separation property is necessary, although it is not as strong as that given by condition (*ii*) of Theorem 1.

Theorem 2: *Let $\mathfrak{D}$ be a bounded, determinable subset of $\mathfrak{K}$, then*

(i) the closure of $\mathfrak{D}$ is compact

(ii) given any $\delta > 0$ there exists an $\hat{x} \in L_\infty(s)$, $\hat{T} > 0$ and a positive number $r(\delta)$ such that

$$|| P_{\hat{T}}(H\hat{x} - K\hat{x}) ||_2^2 > r(\delta) \quad \textit{for all} \quad H, K \in \bar{\mathfrak{D}}$$

satisfying $|| H - K || \geqq \delta$.

Proof: (*i*) Given $\epsilon > 0$, choose $\hat{T}$, $\hat{x} \in L_\infty(s)$, $\hat{N}$ linear measurements and an estimator $g(\underline{m}(H, \omega))$ so that†

$$P\{|| H - g(\underline{m}(H, \omega)) || < \epsilon/2\} > \tfrac{3}{4} \quad \text{for all} \quad H \in \mathfrak{D}.$$

Since $\mathfrak{D}$ is bounded and the measurements are linear, there exists a compact ball $B_{\hat{T}} \in R^{\hat{N}}$ so that

$$P\{\underline{m}(H, \omega) \in B_{\hat{T}}^c\} < \tfrac{1}{4} \quad \text{for all} \quad H \in \mathfrak{D}.$$

Thus, since g is continuous, $g(B_{\hat{T}})$ is compact. We can therefore cover $B_{\hat{T}}$ by a finite number of balls of radius $\epsilon/2$. If $g(B_{\hat{T}}) \supset \mathfrak{D}$ we could also cover $\mathfrak{D}$ by the balls. We don't have enough information to verify that $g(B_{\hat{T}}) \supset \mathfrak{D}$. Notice however that

$$\begin{aligned} &P\{[\omega: || H - g(\underline{m}(H, \omega)) || > \epsilon/2] \cap [\omega: \underline{m}(H, \omega) \in B_{\hat{T}}]\} \\ &\quad = P\{\omega: || H - g(\underline{m}(H, \omega)) || > \epsilon/2\} + P\{\omega: \underline{m}(H, \omega) \in B_{\hat{T}}\} \\ &\quad - P\{[\omega: || H - g(\underline{m}(H, \omega)) || > \epsilon/2] \cup [\omega: \underline{m}(H, \omega) \in B_{\hat{T}}]\} \\ &\quad \geqq \tfrac{3}{4} + \tfrac{3}{4} - 1 = \tfrac{1}{2}. \end{aligned} \tag{20}$$

We conclude that there exists an ω_0 so that $\underline{m}(H, \omega_0) \in B_{\hat{T}}$ and $|| H - g(\underline{m}(H, \omega_0)) || < \epsilon/2$. We can repeat this argument for each $H \in \mathfrak{D}$. Therefore, $\mathfrak{D}$ must lie within an $\epsilon/2$ neighborhood of $g(B_{\hat{T}})$. By expanding the balls of radius $\epsilon/2$ which cover $g(B_{\hat{T}})$ by a factor of two, the expanded balls will also cover $\mathfrak{D}$. Since this argument holds for any $\epsilon > 0$, $\mathfrak{D}$ is shown to be totally bounded. Since $\mathfrak{K}$ is complete, $\bar{\mathfrak{D}}$ is complete; and hence $\bar{\mathfrak{D}}$ is compact (see Ref. 6, p. 22).

(*ii*) If $\mathfrak{D}$ is determinable, then the closure of $\mathfrak{D}$, $\bar{\mathfrak{D}}$, is also determinable. This is easily shown by noting that any channel in $\bar{\mathfrak{D}}$ can be approximated arbitrarily closely by a channel in $\mathfrak{D}$. Hence the measurements will be arbitrarily close and because of the continuity of the estimate, the estimate will be close with high probability.

† Since the measurements are gaussian random variables we have included the dependence on the sample points ω of the corresponding sample space Ω.

Since $\mathfrak{D}$ is determinable, for every $\epsilon > 0$ there exists an observation interval $[0, \hat{T}]$, a test signal $\hat{x} \in L_\infty(s)$, and an estimator $g(\underline{m}(\cdot, \omega))$ so that

$$P\{|| H - g[\underline{m}(H, \omega)] || < \delta/2\} > \tfrac{3}{4} \quad \text{for all} \quad H \in \mathfrak{D}. \tag{21}$$

Suppose that $|| P_{\hat{T}}(H\hat{x} - K\hat{x}) ||_2 = 0$. Then, the measurements obtained will be the same irrespective of whether H or K were used and therefore the estimates for K and H will be identical. Since

$$P\{\omega: || H - g[\underline{m}(H, \omega)] || < \delta/2\} > \tfrac{3}{4}$$

and

$$P\{\omega: || K - g[\underline{m}(H, \omega)] || < \delta/2\} > \tfrac{3}{4},$$

we see that

$$\begin{aligned} P\{[\omega: || K - g[\underline{m}(H, \omega)] || < \delta/2\} & \\ \cap \{\omega: || H - g[\underline{m}(H, \omega)] || < \delta/2]\} & \\ = P\{\omega: || K - g[\underline{m}(H, \omega)] || < \delta/2\} & \\ + P\{\omega: || H - g[\underline{m}(H, \omega)] || < \delta/2\} & \\ - P\{[\omega: || K - g[\underline{m}(H, \omega)] || < \delta/2\} & \\ \cup \{\omega: || H - g[\underline{m}(H, \omega)] || < \delta/2]\} & \\ \geqq \tfrac{3}{4} + \tfrac{3}{4} - 1 = \tfrac{1}{2}. & \end{aligned}$$

Thus there exists at least one sample point ω_0 such that

$$|| K - g(\underline{m}(H, \omega_0)) || < \delta/2$$

and

$$|| H - g(\underline{m}(H, \omega_0)) || < \delta/2$$

which together imply that $|| H - K || < \delta$. If $H, K \in \mathfrak{D}$ and $|| H - K || > \delta$ then $|| P_{\hat{T}}(H\hat{x} - K\hat{x}) ||_2 > 0$.

Note that $\mathfrak{D} \times \mathfrak{D}$ is compact in the product topology and hence $C(\delta) = \{(H, K): || H - K || \geqq \delta, H, K \in \mathfrak{D}\}$ is also compact. The function $f(H, K) = || P_{\hat{T}}(H\hat{x} - K\hat{x}) ||_2$ is a continuous map of $C(\delta)$ into the real line and hence it has a minimum value. This minimum value cannot be zero because we have already shown that $f(H, K) > 0$ for $(H, K) \in C(\delta)$. As a consequence, there exists a positive number $r(\delta)$ such that

$$|| P_{\hat{T}}(H\hat{x} - K\hat{x}) ||_2^2 > r(\delta) \quad \text{for all} \quad H, K \in \mathfrak{D}$$

satisfying $|| H - K || \geqq \delta$.

V. LINEAR CONVOLUTION OPERATORS

When we specialize the results of Theorems 1 and 2 to linear convolution operators, it is possible to obtain the characterization of the determinable sets in terms of the kernels of these operators. These results are given in Corollaries 1, 2 and 3 below. We note that the resulting conditions are similar to those obtained by Root and Prosser for the deterministic identification problem.[1]

Corollary 1: *If* $\mathfrak{C}$ *is composed only of causal linear time invariant convolution operators* H, $[\mathrm{Hx}](t) = \int_0^t \mathrm{h}(t - \tau)x(\tau)\, d\tau$, $\mathrm{h} \in \mathrm{L}_1(0, \infty)$ *and if*

(*i*) $\tilde{\mathfrak{D}} = \{\mathrm{h} \mid \mathrm{h} \in \mathrm{L}_1(0, \infty), \mathrm{H} \in \mathfrak{D}\}$ *has a compact closure in* $\mathrm{L}_1(0, \infty)$, *and*

(*ii*) *for each* $\delta > 0$ *there exists an* $\mathrm{x} \in \mathrm{L}_\infty(\mathrm{s})$, $\mathrm{T} > 0$ *such that* $|| \mathrm{P_T Hx} - \mathrm{P_T Kx} ||_2 > 0$ *for all* h, $\mathrm{k} \in \tilde{\mathfrak{D}}$ *for which* $|| \mathrm{h} - \mathrm{k} ||_1 = \int_0^\infty | \mathrm{h}(t) - \mathrm{k}(t) | \, dt \geqq \delta$ *then* $\mathfrak{D}$ *is determinable.*

Necessary and sufficient conditions for $\tilde{\mathfrak{D}}$ to have a compact closure are (see Ref. 6, pp. 298–299):

(*i*) $\tilde{\mathfrak{D}}$ is a bounded subset of $L_1(0, \infty)$,
(*ii*) $\lim_{\tau \to 0} \int_0^\infty | h(t + \tau) - h(t) | \, dt = 0$ uniformly for $h \in \tilde{\mathfrak{D}}$, and
(*iii*) $\lim_{T \to \infty} \int_T^\infty | h(t) | \, dt = 0$ uniformly for $h \in \tilde{\mathfrak{D}}$.

Proof: We first show that if the closure of $\tilde{\mathfrak{D}}$ is compact then the closure of $\mathfrak{D}$ is compact in the respective topologies. Let $|| H ||^*$, $H \in \mathfrak{C}$ denote the usual operator norm, that is,

$$|| H ||^* = \sup_{\substack{x \in L_2(0,\infty) \\ x \neq 0}} \frac{|| Hx ||_2}{|| x ||_2}.$$

Given any $\epsilon > 0$ there exists $T^* > 0$, $x^* \in L_\infty(s)$ such that

$$|| H || \leqq \epsilon + \frac{|| P_{T^*} Hx^* ||_2}{|| P_{T^*} x^* ||_2} \leqq \epsilon + \frac{|| P_{T^*} H P_{T^*} x^* ||_2}{|| P_{T^*} x^* ||_2}. \tag{23}$$

Note however that $P_{T^*} x^* \in L_2(0, \infty)$; hence $|| H || \leqq \epsilon + || H ||^*$ for arbitrary $\epsilon > 0$, so

$$|| H || \leqq || H ||^*. \tag{24}$$

Using the linearity of H and Holder's inequality we see that

$$|| H ||^* = \sup_{x \in L_2(0,\infty)} \frac{|| Hx ||_2}{|| x ||_2} \leqq \sup_{x \in L_2(0,\infty)} \frac{|| h ||_1 \, || x ||_2}{|| x ||_2} = || h ||_1 . \tag{25}$$

Thus compactness in $L_1(0, \infty)$ implies compactness in $\mathfrak{C}$ and condition (i) of Theorem 1 is satisfied.

Given $\delta > 0$, choose x_o, T^o so that condition (ii) of Corollary 1 is satisfied. We have already used the fact that $|| HP_{T^o}x_o ||_2 \leqq || h ||_1 \cdot || P_{T^o}x_o ||_2$. Hence $HP_{T^o}x_o$ is a continuous linear mapping (that is, mapping the kernels into time functions) from $L_1(0, \infty)$ into $L_2(0, \infty)$. Thus the image of $\mathfrak{D}$ under this mapping has a compact closure. We can therefore choose a number $\hat{T} > T^o$ so that

$$\int_{\hat{T}}^{\infty} (HP_{T^o}x_o - KP_{T^o}x_o)^2(t)\, dt < \tfrac{1}{4} \quad \text{for all} \quad H, K \in \mathfrak{D}. \tag{26}$$

Define $\hat{x}$ as follows:

$$\begin{aligned}
\hat{x}(t) &= x_0(t) && \text{for} \quad 0 < t \leqq T^o \\
&= 0 && \text{for} \quad T^o < t \leqq \hat{T} \\
&= x_0(t - \hat{T}) && \text{for} \quad \hat{T} < t \leqq \hat{T} + T^o \\
&= 0 && \text{for} \quad \hat{T} + T^o < t \leqq 2\hat{T} \\
&\vdots \\
&= x_0(t - n\hat{T}) && \text{for} \quad n\hat{T} < t \leqq n\hat{T} + T^o \\
&= 0 && \text{for} \quad n\hat{T} + T^o < t \leqq (n+1)\hat{T} \\
&\vdots
\end{aligned} \tag{27}$$

Note that $\hat{x} \in L_\infty(s)$. Following the same line of reasoning as in the proof of condition (ii) of Theorem 2 we can show that there exists an $r(\delta) > 0$ so that $|| P_{T^o}(Hx_0 - Kx_0) ||_2^2 > r(\delta)$ for all $H, K \in \mathfrak{D}$ for which $|| h - k ||_1 > \delta$. We now proceed to show that

$$|| P_{n\hat{T}}(H\hat{x} - K\hat{x}) ||_2^2 \geq \tilde{r}(\delta) n\hat{T} \tag{28}$$

where $\tilde{r}(\delta) = r(\delta)/4\hat{T}$. Let $y_0(t) = [HP_{T^o}x_0 - KP_{T^o}x_0](t)$ and $y_i(t) = y_0(t - i\hat{T})$. Then, by linearity and time invariance,

$$\int_{(i-1)\hat{T}}^{i\hat{T}} (H\hat{x} - K\hat{x})^2(t)\, dt = \int_{(i-1)\hat{T}}^{i\hat{T}} [y_0(t) + y_1(t) + \cdots + y_{i-1}(t)]^2\, dt \tag{29}$$

and

$$\int_{i\hat{T}}^{(i+1)\hat{T}} y_i^2(t)\, dt = \int_{(i-j)\hat{T}}^{(i+1-j)\hat{T}} y_0^2(t)\, dt \quad \text{for} \quad j \leqq i. \tag{30}$$

Using these relationships we see that

$$\begin{aligned}
&\int_{i\hat{T}}^{(i+1)\hat{T}} (y_0 + \cdots + y_i)^2\, dt \\
&\quad \geqq \int_{i\hat{T}}^{(i+1)\hat{T}} y_i^2\, dt - 2\int_{i\hat{T}}^{(i+1)\hat{T}} |\, y_i\, |\,(|\, y_0\, | + \cdots + |\, y_{i-1}\, |)\, dt \\
&\quad \geqq \int_{i\hat{T}}^{(i+1)\hat{T}} y_i^2\, dt\left[1 - 2\int_{i\hat{T}}^{(i+1)\hat{T}} (y_0^2 + \cdots + y_{i-1}^2)\, dt\right] \\
&\quad = \int_{i\hat{T}}^{(i+1)\hat{T}} y_i^2\, dt\left[1 - 2\int_{\hat{T}}^{(i+1)\hat{T}} y_0^2\, dt\right] \geqq r(\delta)/2. \qquad (31)
\end{aligned}$$

Hence

$$\begin{aligned}
||\, P_{n\hat{T}}(H\hat{x} - K\hat{x})\, ||_2^2 &= \int_0^{\hat{T}} y_0^2\, dt + \int_{\hat{T}}^{2\hat{T}} (y_0 + y_1)^2\, dt + \cdots \\
&\quad + \int_{(n-1)\hat{T}}^{n\hat{T}} (y_0 + \cdots + y_{n-1})^2\, dt \\
&\geqq nr(\delta)/4 = r'(\delta)n\hat{T}. \qquad (32)
\end{aligned}$$

We see that this relation implies that condition (*ii*) of Theorem 1 is satisfied; thus $\mathfrak{D}$ is determinable.

When $\mathfrak{IC}$ is composed only of causal linear time invariant convolution operators we can also strengthen the conclusion of Theorem 2. This result is given in the following corollary.

Corollary 2: *If* $\mathfrak{IC}$ *is composed only of causal linear time invariant convolution operators and if* $\mathfrak{D}$ *is a determinable subset of* $\mathfrak{IC}$ *then*

(*i*) *given any* $\delta > 0$ *there exists an unbounded sequence* T_i, *a sequence of inputs* $x_i \in L_\infty(s)$ *and a positive number* $r(\delta)$ *such that* $||\, P_{T_i}(Hx_i - Kx_i)^2\, ||_2 > r(\delta)T_i$ *for all pairs* $H, K \in \mathfrak{D}$ *for which* $||\, H - K\, || \geq \delta$.

Proof: As a consequence of Theorem 2 we know that for any $\delta > 0$ there exists an $\hat{x} \in L_\infty(s)$, $\hat{T} > 0$ and a positive number $r(\delta)$ such that $||\, P_{\hat{T}}(H\hat{x} - K\hat{x})^2\, ||_2 > r(\delta)$ for all $H, K \in \mathfrak{D}$ satisfying $||\, H - K\, || \geq \delta$.

Obviously, $||\, HP_{\hat{T}}\hat{x}\, ||_2 \leq ||\, H\, ||\, ||\, P_{\hat{T}}\hat{x}\, ||_2$. Hence $HP_{\hat{T}}\hat{x}$ is a continuous linear mapping from $\mathfrak{IC}$ into $L_2(0, \infty)$. Thus the image of $\mathfrak{D}$ under this mapping has a compact closure. We can therefore choose a positive number $\bar{T} > \hat{T}$ so that

$$\int_{\bar{T}}^{\infty} (HP_{\hat{T}}\hat{x} - HP_{\hat{T}}\hat{x})^2(t)\, dt < \tfrac{1}{4} \quad \text{for all} \quad H, K \in \mathfrak{D}. \qquad (33)$$

Proceeding as in the proof of corollary 1 we can easily establish (i) of Corollary 2.

Corollary 3: *If* $\mathfrak{K}$ *is composed only of causal Hilbert-Schmidt operators* H, $[\mathrm{Hx}](t) = \int_0^t h(t, \tau)x(\tau)dt$, $\int_0^\infty \int_0^\infty | h(t, \tau) |^2 \, dt \, d\tau < \infty$, $h(t, \tau) = 0$ *for* $\tau > t$ *and if*

(i) $\hat{\mathfrak{D}} = \{h \mid H \in \mathfrak{D}\}$ *has compact closure in the Hilbert-Schmidt metric* ($|| h - k ||_2^2 = \int_0^\infty \int_0^\infty | h(t, \tau) - k(t, \tau) |^2 \, dt \, d\tau$)

(ii)*for each* $\delta > 0$ *there exists an unbounded sequence* T_i, *a sequence of* $x_i \in L_\infty(s)$ *and a positive constant* $r(\delta)$ *so that* $|| P_{T_i}(Hx_i - Kx_i) ||_2 \geq r(\delta)T_i$ *for all* h, $k \in \hat{\mathfrak{D}}$ *for which* $|| h - k ||_2 > \delta$.

Then $\mathfrak{D}$ *is determinable.*

Proof: As in the proof of Corollary 1 we can show that $|| H || \leq || H ||^*$ where

$$|| H ||^* = \sup_{x \in L_2(0,\infty)} \frac{|| Hx ||_2}{|| x ||_2}.$$

From the Schwartz inequality we see that

$$\begin{aligned} || Hx ||_2^2 &= \int_0^\infty \left(\int_0^t h(t, \tau)x(\tau) \right)^2 dt = \int_0^\infty \left(\int_0^\infty h(t, \tau)x(\tau) \, d\tau \right)^2 dt \\ &\leqq \int_0^\infty \left[\int_0^\infty | h(t, \tau) |^2 \, d\tau \int_0^\infty | x(\tau) |^2 \, d\tau \right] dt \\ &\leqq || h ||_2^2 \, || x ||_2^2 , \end{aligned} \tag{34}$$

which implies that

$$|| H || \leq || h ||_2 . \tag{35}$$

Hence, compactness of $\hat{\mathfrak{D}}$ implies that $\mathfrak{D}$ is compact and condition (i) and (ii) of Theorem 1 are easily verified to hold.

VI. COLORED NOISE

Theorems 1 and 2 were derived for the case when $z(t)$ the additive noise was a zero mean white stochastic process. The situation when $Ez(t)z(\tau) = R(t, \tau)$ can be handled in a similar fashion. The only additional assumptions are:

(i) $R(t, \tau)$ is positive definite; that is,

$$\int_0^\infty \int_0^\infty R(t, \tau)w(t)w(\tau) \, dt \, d\tau > 0 \quad \text{for all} \quad w \in L_2(0, \infty)$$

satisfying $\int_0^\infty | w(t) |^2 dt > 0$, and either

(*ii*) $R(t, \tau)$ is Hilbert-Schmidt; that is,

$$\int_0^\infty \int_0^\infty | R(t, \tau) |^2 dt\, d\tau = C^2 < \infty , \quad \text{or}$$

(*iii*) if $R(t, \tau) = R_0(t - \tau)$ then

$$\int_{-\infty}^{\infty} | R_0(t) |^2 dt = C_0^2 < \infty .$$

Inspecting the proof of Theorem 1, one sees that the whiteness assumption was only used in equations (8) and (17). If $Ez(t)z(\tau) = R(t, \tau)$, then equation (8) becomes

$$\begin{aligned}\sigma_{ki}^2 &= \text{Var } [q(w, k) - q(w, j)] \\ &= 4 \int_0^{T_i} \int_0^{T_i} R(t, \tau)(H_k x_i - H_i x_i)(t) \cdot (H_k x_i - H_i x_i)(\tau)\, dt\, d\tau . \quad (36)\end{aligned}$$

Since H_k and H_j were chosen so that $|| P_{T_i}(H_k x_i - H_i x_i) || > 0$, we see that since $R(t, \tau)$ is positive definite, $\sigma_{ki}^2 > 0$. If we choose θ_i to be less than $\min_{i,k} \sigma_{ik}^2$ instead of $|| P_{T_i}(H_k x_i - H_i x_i) ||_2^2$, inequality (9) will remain true.

Equation (17) is changed as follows. If $Ez(t)z(\tau) = R(t, \tau)$, then by the Schwartz inequality

$$\begin{aligned}\sigma_{i\bar{k}}^2 &= 4 \int_0^{T_i} \int_0^{T_i} R(t, \tau)(H_i x_i - H_{\bar{k}} x_i)(t)(H_i x_i - H_{\bar{k}} x_i)(\tau)\, dt\, d\tau \\ &\leqq 4 \int_0^{T_i} (H_i x_i - H_{\bar{k}} x_i)(\tau) \left\{ \int_0^{T_i} | R(t, \tau) |^2 dt \right\}^{\frac{1}{2}} \\ &\quad \cdot \left\{ \int_0^{T_i} (H_i x_i - H_{\bar{k}} x_i)^2(t)\, dt \right\}^{\frac{1}{2}} d\tau \\ &\leqq 4 \left\{ \int_0^{T_i} \int_0^{T_i} | R(t, \tau) |^2 dt\, d\tau \right\}^{\frac{1}{2}} || P_{T_i}(H_i x_i - H_{\bar{k}} x_i) ||_2^2 \\ &\leqq 4CR^2 s^2 T_i . \quad (37)\end{aligned}$$

On the other hand, if $Ez(t)z(\tau) = R_0(t - \tau)$, equation (17) is changed as follows.

$$\begin{aligned}\sigma_{i\bar{k}}^2 &= 4 \int_0^{T_i} \int_0^{T_i} R(t - \tau)(H_i x_i - H_{\bar{k}} x_i)(t)(H_i x_i - H_{\bar{k}} x_i)(\tau)\, dt\, d\tau \\ &\leqq 4 \int_0^{T_i} | (H_i x_i - H_{\bar{k}} x_i)(t) | \left[\int_0^{T_i} | R(t - \tau) |^2 d\tau \right]^{\frac{1}{2}}\end{aligned}$$

$$\cdot\left[\int_0^{T_i} (H_i x_i - H_k x_i)^2(\tau)\, d\tau\right]^{\frac{1}{2}} dt$$

$$\leqq 4C_0 Rs(T_i)^{\frac{1}{2}} \int_0^{T_i} |\, (H_i x_i - H_k x_i)(t)\, |\, dt$$

$$\leqq 4C_0 Rs(T_i)^{\frac{1}{2}} \left(\int_0^{T_i} dt\right)^{\frac{1}{2}} \left[\int_0^{T_i} (H_i x_i - H_k x_i)^2(t)\, dt\right]^{\frac{1}{2}}$$

$$\leqq 4C_0 Rs(T_i)^{\frac{1}{2}} \cdot (T_i)^{\frac{1}{2}} \cdot Rs(T_i)^{\frac{1}{2}} = C_0 R^2 s^2 T_i^{\frac{3}{2}} . \tag{38}$$

From equation (37) we see that the limit of integration in equation (19) now becomes $3rT_i^{\frac{1}{2}}/4RsC^{\frac{1}{2}}$. If we use equation (38), this limit becomes $3rT_i^{\frac{1}{2}}/16RsC_0^{\frac{1}{2}}$. In either case this limit diverges as i increases. Thus Theorem 1 is still correct if the noise is colored. One can also see that Theorem 2 is true without any modifications. The whiteness assumption does enter into the proof in any substantial manner.

VII. CONCLUSIONS

In this paper we have attempted to formalize the notion of identification and examined conditions under which the *a priori* information would guarantee that the problem of identification was well formulated. Our purpose has been to indicate when identification was possible and not to specify a given identification procedure. It is hoped that the conditions derived here may motivate researchers to consider larger classes of identification problems than have hitherto been examined and also to indicate for what classes of problems identification is not possible.

VIII. ACKNOWLEDGMENTS

The author wishes to thank J. M. Holtzman and P. P. Varaiya for their helpful discussions and criticisms. He also wishes to acknowledge support for the initial phases of this work from NASA under Grant NsG-354, Supplement 4 while at the University of California at Berkeley.

APPENDIX

Proof that the Space $\mathfrak{IC}$ Is Complete

In this appendix we show that the space $\mathfrak{IC}$ with the metric induced by its norm is a complete space. If $\{H_n\}$ is a Cauchy sentence in $\mathfrak{IC}$, we show that there exists an element $\tilde{H} \in \mathfrak{IC}$ such that $\lim_{n\to\infty} \| \tilde{H} - H_n \| = 0$.

Let $\{H_n\}$ be a Cauchy sequence in $\mathfrak{IC}$. Then given any $\epsilon > 0$ there

exists a number $N(\epsilon)$ such that if $m, n > N(\epsilon)$, $|| H_n - H_m || < \epsilon$. From the definition of the metric,

$$|| H_n - H_m || \geqq \frac{|| P_T H_n x - P_T H_m x ||_2}{|| P_T x ||_2} \tag{39}$$

for all $T > 0$, $x \in L_\infty(s)$, $|| P_T x ||_2 \neq 0$. Using the definition of $L_\infty(s)$,

$$\epsilon s(T)^{\frac{1}{2}} \geq \epsilon || P_T x ||_2 \geq || P_T(H_n x - H_m x) ||_2 \tag{40}$$

for all $n, m > N(\epsilon)$, $T > 0$, $x \in L_\infty(s)$, $|| P_T x ||_2 \neq 0$. Thus, for each $T > 0$, $x \in L_\infty(s)$, $|| P_T x ||_2 \neq 0$, $\{H_n x\}$ is a sequence of functions in L_{2e} and for each $T > 0$, $P_T H_n x$ is a Cauchy sequence in $L_2[0, T]$. Hence, for each T there exists at least one time function $y_T \in L_{2e}$ such that $P_T y_T \in L_2(0, \infty)$ and $\lim_{n\to\infty} || P_T H_n x - P_T y_T ||_2 = 0$. Furthermore, y_T is uniquely (except for a set of measure zero) specified over $[0, T]$. Because of this uniqueness, if $T_1 < T_2$, then $P_{T_1} y_{T_2} = P_{T_1} y_{T_1}$. Hence there exists a unique function $\tilde{y} \in L_{2e}$ such that $P_T \tilde{y} = P_T y_T$ for each $T > 0$. This function can be constructed:

$$\begin{aligned}
\tilde{y}(t) &= y_1(t) \quad \text{for} \quad 0 \leqq t < 1 \\
&= y_2(t) \quad \text{for} \quad 1 \leqq t < 2 \\
&\vdots \qquad\qquad \vdots \\
&= y_n(t) \quad \text{for} \quad n - 1 \leqq t < n \\
&\vdots \qquad\qquad \vdots
\end{aligned} \tag{41}$$

For each $x \in L_\infty(s)$, $x \neq 0$ we have uniquely specified a function $\tilde{y} \in L_{2e}$. For $x = 0$ we arbitrarily put $\tilde{y} = 0$. Call the operator defined by this association $\tilde{H}$; that is, $\tilde{H}x = \tilde{y}$. We now show that $\lim_{n\to\infty} || \tilde{H} - H_n || = 0$.

For each $T > 0$, $x \in L_\infty(s)$, $|| P_T x ||_2 \neq 0$ we can use the triangle inequality to show that

$$\frac{|| P_T(\tilde{H}x - H_n x) ||_2}{|| P_T x ||_2} \leqq \frac{|| P_T \tilde{H}x - P_T H_m x ||_2}{|| P_T x ||_2} + \frac{|| P_T(H_n x - H_m x) ||_2}{|| P_T x ||_2}. \tag{42}$$

If H_n, H_m are members of the Cauchy sequence, from our previous development we know that there exists a number $N(\epsilon/2)$ independent of x and T such that

$$\frac{|| P_T(H_n x - H_m x) ||_2}{|| P_T x ||_2} < \epsilon/2 \quad \text{for} \quad m, n > N(\epsilon/2). \tag{43}$$

Since $\lim_{m\to\infty} || P_T(\tilde{H}x - H_m x) ||_2 = 0$ we can find another number $N^*(\epsilon/2, x, T) > N(\epsilon/2)$ such that

$$\frac{|| P_T(\tilde{H}x - H_m x) ||_2}{|| P_T x ||_2} < \epsilon/2 \quad \text{for} \quad m > N^*(\epsilon/2, x, T). \tag{44}$$

Hence for all $T > 0$, $P_T x \neq 0$

$$\frac{|| P_T(\tilde{H}x - H_n x) ||_2}{|| P_T x ||_2} < \epsilon \quad \text{for} \quad n > N(\epsilon/2), \tag{45}$$

and if $\tilde{H}$ were causal it follows that $\tilde{H} \in \mathfrak{C}$ with $\lim || \tilde{H} - H_n || = 0$. The causality of $\tilde{H}$ is easily established. For each $x \in L_\infty(s)$, $T > 0$:

$$\begin{aligned} &|| P_T\tilde{H}x - P_T\tilde{H}P_T x ||_2 \\ &\quad \leqq || P_T\tilde{H}x - P_T H_n x ||_2 + || P_T\tilde{H}P_T x - P_T H_n x ||_2 \qquad (46) \\ &\quad = || P_T(\tilde{H}x - H_n x) ||_2 + || P_T(\tilde{H}P_T x - H_n P_T x) ||_2 . \qquad (47) \end{aligned}$$

For n sufficiently large each term on the right side may be arbitrarily small, hence $|| P_T\tilde{H}x - P_T\tilde{H}P_T x ||_2 = 0$ for all $x \in L_\infty(s)$, $T > 0$.

If $\mathfrak{C}$ is composed only of linear operators the completeness proof follows as above except to additionally observe that $\tilde{H}$ is linear.

REFERENCES

1. Prosser, R. T., and Root, W. L., "Determinable Classes of Channels," J. Math. and Mechanics, *16*, No. 4 (October 1966), pp. 365–398.
2. Root, W. L., "On System Measurement and Identification," *Proc. Symp. on System Theory*, Polytechnic Institute of Brooklyn, April 1965, pp. 133–157.
3. Root, W. L., "On the Measurement and Use of Time-Varying Communication Channels," Inform. and Control, *8*, No. 4 (August 1965), pp. 390–422.
4. Varaiya, P. P., "On Determinable Classes of Signals and Linear Channels," SIAM J. Appl. Math., *15*, No. 2 (March 1967), pp. 440–449.
5. Root, W. L., unpublished work.
6. Dunford, N., and Schwartz, J. T., *Linear Operators-Part I*, New York: Interscience Publishers, 1964.

The Theory of Cylindrical Magnetic Domains

By A. A. THIELE

(Manuscript received December 26, 1968)

The theory of cylindrical magnetic domains provides conditions governing the size and stability of circular cylindrical magnetic domains in plates of uniaxial magnetic materials together with an estimate of the range of applicability of these conditions. The results of the theory are directly applicable to the design of cylindrical domain devices. Computation to first and second order of the energy variation resulting from general small deviation in the domain shape from an initially circular shape yields the conditions governing domain size and stability. The physical origin of the various terms in the energy expansion is examined in detail. A graph from which many domain size and stability properties may be obtained summarizes the results of the energy variation calculation. The minimum theoretically attainable domain diameter is approximately $\sigma_w/\pi M_s^2$, *where* σ_w *is the wall energy density and* M_s *is the saturation magnetization. For domains to exist, the effective anisotropy field must be greater than* $4\pi M_s$.

I. INTRODUCTION

The recent development of a technique for the propagation of isolated magnetic domains in an arbitrary direction in anisotropic ferromagnetic thin films by P. C. Michaelis created a renewed interest in the use of domain propagation for device purposes.[1] The technique used by Michaelis for propagating domains along the easy axis is quite different from that used for propagation along the hard axis. During discussions on the possible application of these techniques, A. H. Bobeck, U. F. Gianola, R. C. Sherwood, and W. Shockley suggested that for general symmetrical domain propagation the direction of magnetization must lie normal to the plane of the film[2]. The recognition that rare earth orthoferrites have the required properties came in response to this suggestion.[3] Experimental work on the application of this type of

domain motion device was then begun. Although at the present time this work has been largely concentrated on the orthoferrites, there exist other materials, such as the hexagonal ferrites and manganese bismuth, having the required properties.

The present work directs attention to structures in which the properties of the material used require the magnetization to lie normal to the surface of the plate. The modes of operation of devices constructed from such structures are classified according to the effect of wall motion coercivity. In the case of very high wall motion coercivity, the application of shaped applied fields determines the initial domain configuration which is then maintained by coercivity. For very low wall coercivity, on the other hand, the saturation magnetization, wall energy, plate thickness and bias field determine the domain size and shape. Between these two extremes, there is a continuum of intermediate modes. In either extremal mode, a complete set of operations (logic, memory, and transmission) may be performed.[4] The present work concerns only the low coercivity mode and specifically, right circular cylindrical domains in plates of uniform thickness and small variations therefrom. When observed by means of the Faraday effect, cylindrical domains have the appearance (particularly when in motion) of bubbles and therefore are colloquially referred to as "bubbles".

The present work largely treats the theory of cylindrical domains with experiments and applications being considered only briefly. Section II presents the domain model and mode of description. Section III contains the calculation of the energy derivatives used in the investigation of domain size and stability. Section IV contains an interpertation of the energy derivatives in terms of fields and potentials. Section V discusses the solution of the domain size and stability equations. Section VI discusses the range of validity of the domain model used in the previous sections. It is found that several assumptions implicit in the model are related, and a requirement on materials suitable for the production of circular domains is obtained. Appendix A contains a derivation the properties of certain elliptic integrals appearing in the theory of circular domains, Appendix B is a listing of the standard forms and series expansions of the magnetostatic force and stability functions, and Appendix C is a list of mathematical symbols.

II. THE DOMAIN MODEL AND MODE OF DESCRIPTION

Figure 1 shows the magnetic domain structure to be considered here.[5] The isolated magnetic domain is magnetized downward while the remainder of the plate is saturated upward. The domain will be

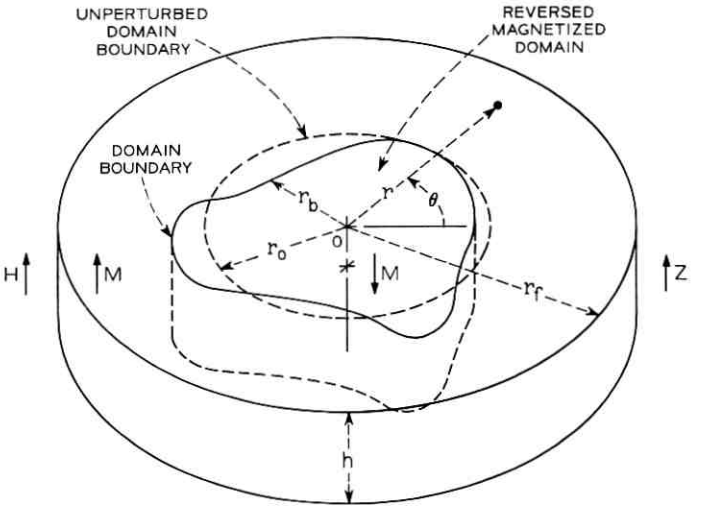

Fig. 1 — Magnetic domain configuration.

considered to be near circular. Examination of the variation of domain energy under a variation of domain shape from the assumed unperturbed shape yields domain stability. Once created, a cylindrical domain continues to exist if the magnetic configuration meets the conditions for stable equilibrium. The stability of a given configuration, however, does not guarantee that it can be produced. The generation of cylindrical domains is a separate problem which is not treated here.

2.1 *Description of the Domain*

A cylindrical (r, θ, z) coordinate system is placed at the center of the domain with its z-axis perpendicular to the plane of the plate. The plate is taken to have planar surfaces and a uniform thickness h. Only the case of a plate of infinite extent, $r_f = \infty$, is considered here. It is assumed that the material constraints allow the magnetization to lie only along the z-axis and the magnitude of the magnetization is independent of the local magnetic field. The boundary between the two regions of magnetization, the domain wall, is assumed to be independent of z (no wall bulging) and to have a width which is negligible in comparison to the domain radius. It is assumed that a wall energy density per unit area σ_w may be assigned independently of either the orientation or curvature of the wall. The assumptions about the detailed magnetic configuration (the magnetization magnitude and orientation and the wall energy and shape) are coupled by the material

properties. Section VI contains a detailed discussion of the validity of these assumptions and the cylindrical wall assumption. Even though the foregoing assumptions appear quite drastic and restrictive, experimentally there does exist a region in which the results obtained under these assumptions are both accurate and useful.

The expansion

$$r_b(\theta) = \sum_{n=0}^{\infty} r_n \cos [n(\theta - \theta_n)] \tag{1}$$

of $r_b(\theta)$ in terms of the Fourier coefficients, r_n and θ_n, describes the domain shape in the plane. The n value is called the "rotational periodicity." The condition

$$| r_0 | \gg \sum_{n=1}^{\infty} n | r_n | \tag{2}$$

assures that the domain is near circular and that the function $r_b(\theta)$ is single valued and smooth.

It is convenient to introduce the finite variations of the r_n and θ_n, $\triangle r_n$ and $\triangle \theta_n$, respectively, in order to describe small variations in domain size and shape from the strictly circular domain of radius $r_0 [r_b(\theta) = r_0]$. In terms of these variations, a small variation of the wall shape from $r_b(\theta) = r_0$ may be written as

$$r_b(\theta) = r_0 + \triangle r_0 + \sum_{n=1}^{\infty} \triangle r_n \cos [n(\theta - \theta_n - \triangle \theta_n)] \tag{3a}$$

where, by assumption,

$$| r_0 | \gg | \triangle r_0 | + \sum_{n=1}^{\infty} n | \triangle r_n |. \tag{3b}$$

Subject to the restrictions stated, equation (3) describes an arbitrary variation because of the completeness of the Fourier expansion.

The externally applied magnetic field, $\mathbf{H}$, is taken to be spatially uniform and to lie in the positive z direction. (The presence of a component of the applied field in the plane of the plate has no effect to the approximation that the magnetization lies only along the z-axis.)

The assumed simple forms of the applied field and magnetic configurations permit the use of simple formal expressions for these quantities. The expression for the externally applied field is

$$\mathbf{H} = H\mathbf{i}_z \tag{4}$$

where H is a constant and $\mathbf{i}_z$ is the unit vector in the z-direction. The magnetization may be written in terms of the unit step function,

$$u(x) \equiv \begin{cases} 0, & x < 0, \\ \frac{1}{2}, & x = 0, \\ 1, & x > 0, \end{cases} \tag{5}$$

as

$$\mathbf{M} = \mathbf{i}_z M_s\{1 - 2u[r_b(\theta) - r]\}u(z + \tfrac{1}{2}h)u(-z + \tfrac{1}{2}h). \tag{6}$$

2.2 *The Energy Variation*

The investigation of domain size and stability proceeds by computing the first and second variations of the total system energy with respect to the r_n and θ_n. The total energy of the domain is

$$E_T = E_W + E_H + E_M, \tag{7}$$

where E_W is the total wall energy, E_H is the interaction energy with the externally applied field, and E_M is the internal magnetostatic energy. The total wall energy, under the previously stated assumptions, is the product of the wall energy density σ_w and the wall area a:

$$E_W = \int_a \sigma_w \, da = h\sigma_w \int_0^{2\pi} \left\{ r_b^2(\theta) + \left[\frac{\partial r_b(\theta)}{\partial \theta} \right]^2 \right\}^{\frac{1}{2}} d\theta. \tag{8}$$

The interaction energy of the magnetization with the externally applied field is

$$E_H = -\int_V \mathbf{M}\cdot\mathbf{H} \, dV = -\int_{-\infty}^{\infty} \int_0^{2\pi} \int_0^{\infty} M_z H r \, dr \, d\theta \, dz, \tag{9}$$

and the internal magnetostatic energy is

$$E_M = \frac{1}{2} \int_V \int_{V'} \frac{\nabla\cdot\mathbf{M}\nabla'\cdot\mathbf{M}'}{|\mathbf{r} - \mathbf{r}'|} \, dV' \, dV$$

$$= \frac{1}{2} \int_{-\infty}^{\infty} \int_0^{2\pi} \int_0^{\infty} \int_{-\infty}^{\infty} \int_0^{2\pi} \int_0^{\infty} \frac{\dfrac{\partial M_z}{\partial z} \dfrac{\partial M_z'}{\partial z'}}{s} rr' \, dr \, d\theta \, dz \, dr' \, d\theta' \, dr' \tag{10a}$$

where

$$s^2 \equiv r^2 + r'^2 - 2rr' \cos(\theta - \theta') + (z - z')^2. \tag{10b}$$

In expressions (9) and (10), V indicates volume and primes indicate quantities in the second coordinate system used in describing the internal magnetostatic interaction.

The variation in the total energy when the r_n and θ_n are varied is

$$\begin{aligned}\Delta E_T = &\sum_{n=0}^{\infty}\left[\left(\frac{\partial E_T}{\partial r_n}\right)_0 \Delta r_n + \left(\frac{\partial E_T}{\partial \theta_n}\right)_0 \Delta\theta_n\right] \\ &+ \frac{1}{2}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\left[\left(\frac{\partial^2 E_T}{\partial r_n\,\partial r_m}\right)_0 \Delta r_n \Delta r_m + 2\left(\frac{\partial^2 E_T}{\partial r_n\,\partial\theta_m}\right)_0 \Delta r_n \Delta\theta_m\right. \\ &\left.+ \left(\frac{\partial^2 E_T}{\partial\theta_n\,\partial\theta_m}\right)_0 \Delta\theta_n \Delta\theta_m\right] + O_3 \end{aligned} \tag{11}$$

where the subscript O refers to evaluation of the partial derivatives at the circular domain state, $r_b(\theta) = r_0$, and O_3 refers to terms of order three and higher in the combination of Δr_n and $\Delta\theta_n$. The first partial derivatives of the energy, $(\partial E_T/\partial r_n)_0$ and $(\partial E_T/\partial\theta_n)_0$, are the generalized forces of the system, while the second derivatives of the total energy form the elements of the stiffness matrix.

Knowledge of the generalized forces and the stiffness matrix completely characterizes domain size and stability. It is shown in Section III that only the energy derivatives, $(\partial E_T/\partial r_0)_0$ and the $(\partial^2 E_T/\partial r_n^2)_0$, are non-zero when $r_b(\theta) = r_0$. The equation obtained by setting the only nonzero generalized force equal to zero is called the "force equation." The expansion (1) is a quasi-normal mode expansion since circular domains are completely metastable with respect to the θ_n and the stiffness matrix is diagonal with respect to the r_n.

III. CALCULATION OF THE ENERGY DERIVATIVES

3.1 *Derivatives of the Wall Energy*

The derivatives of the total wall energy are computed by substituting the wall shape expression (1) into the wall energy expression (8), noting that

$$\frac{\partial r_b}{\partial\theta} = -\sum_{n=1}^{\infty} n r_n \sin\left[n(\theta - \theta_n)\right] \tag{12}$$

and differentiating under the integral sign. There results

$$\begin{aligned}\frac{\partial E_W}{\partial r_n} = h\sigma_w \int_0^{2\pi} &\left\{r_b \cos\left[n(\theta - \theta_n)\right] - \frac{\partial r_b}{\partial\theta} n \sin\left[n(\theta - \theta_n)\right]\right\} \\ &\cdot\left[r_b^2 + \left(\frac{\partial r_b}{\partial\theta}\right)^2\right]^{-\frac{1}{2}} d\theta \end{aligned} \tag{13a}$$

and

$$\frac{\partial^2 E_W}{\partial r_n\,\partial r_m} = h\sigma_w \int_0^{2\pi} \left(\left\{ \cos\left[n(\theta - \theta_n)\right] \cos\left[m(\theta - \theta_m)\right] + nm \sin\left[n(\theta - \theta_n)\right] \sin\left[m(\theta - \theta_m)\right] \right\} \left[r_b^2 + \left(\frac{\partial r_b}{\partial \theta}\right)^2 \right]^{-\frac{1}{2}} - \left\{ r_b \cos\left[n(\theta - \theta_n)\right] - \frac{\partial r_b}{\partial \theta} n \sin\left[n(\theta - \theta_n)\right] \right\} \cdot \left\{ r_b \cos\left[m(\theta - \theta_m)\right] - \frac{\partial r_b}{\partial \theta} m \sin\left[m(\theta - \theta_m)\right] \right\} \cdot \left[r_b^2 + \left(\frac{\partial r_b}{\partial \theta}\right)^2 \right]^{-\frac{3}{2}} \right) d\theta \tag{13b}$$

with analogous expressions for

$$\partial E_W/\partial \theta_n\,, \qquad \partial^2 E_W/\partial \theta_n\,\partial \theta_m\,, \quad \text{and} \quad \partial^2 E_W/\partial r_n\,\partial \theta_m\,.$$

Evaluating equations (13) for a circular domain,

$$r_b(\theta) = r_0 \quad \text{and} \quad [\partial r_b(\theta)/\partial\theta] = 0,$$

the circular domain derivatives are

$$\left(\frac{\partial E_W}{\partial r_0}\right)_0 = 2\pi h \sigma_w \tag{14a}$$

$$\left(\frac{\partial^2 E_W}{\partial r_n^2}\right)_0 = \frac{\pi}{r_0} h\sigma_w n^2, \qquad n \geqq 1 \tag{14b}$$

and all of the first and second derivatives of the total wall energy not explicitly stated are zero.

3.2 *Derivatives of the Applied Field Interaction Energy*

The applied field interaction energy is evaluated by substituting the formal expressions for the applied field (4) and the magnetic configuration (6) into the applied field interaction expression (9), changing the order of integration, and integrating.

$$E_H = -M_s H \int_0^{2\pi} \int_0^{\infty} \int_{-\infty}^{\infty} \{1 - 2u[r_b(\theta) - r]\} \times u(z + \tfrac{1}{2}h)u(-z + \tfrac{1}{2}h) r\, dz\, dr\, d\theta \tag{15a}$$

$$= hM_s H \left[\int_0^{2\pi} r_b^2(\theta) d\theta \right] - \text{constant}. \tag{15b}$$

The infinite constant is independent of the r_n and θ_n and does not

contribute to the derivatives. Differentiating yields

$$\frac{\partial E_H}{\partial r_n} = 2hM_sH \int_0^{2\pi} r_b \cos [n(\theta - \theta_n)] d\theta \tag{16a}$$

and

$$\frac{\partial^2 E_H}{\partial r_n \partial r_m} = 2hM_sH \int_0^{2\pi} \cos [n(\theta - \theta_n)] \cos [m(\theta - \theta_m)] d\theta \tag{16b}$$

with analogous expressions for

$$\partial E_H/\partial \theta_n , \qquad \partial^2 E_H/\partial r_n \partial \theta_m , \quad \text{and} \quad \partial^2 E_H/\partial \theta_n \partial \theta_m .$$

Evaluation of equation (16) for $r_b(\theta) = r_0$ yields

$$\left(\frac{\partial E_H}{\partial r_0}\right)_0 = 4\pi r_0 hM_sH, \tag{17a}$$

$$\left(\frac{\partial^2 E_H}{\partial r_0^2}\right)_0 = 4\pi hM_sH, \tag{17b}$$

$$\left(\frac{\partial^2 E_H}{\partial r_n}\right)_0 = 2\pi hM_sH, \qquad n \geqq 1, \tag{17c}$$

and all the other first and second derivatives of the applied field interaction energy are zero.

3.3 *Derivatives of the Internal Magnetostatic Energy*

The formal expression for the internal magnetostatic energy is obtained by substituting the expression for the magnetic configuration (6) into expression (10). In dealing with the self-interaction energy, it is necessary to use two coordinate systems: an unprimed system and a primed system. Throughout the following calculation functions of the spatial coordinates (r, θ, and z) are written with primes whenever they are of the primed coordinates. Thus $\mathbf{M}$, when considered as a function of the primed coordinates, is written $\mathbf{M}'$. The subscripted r_n and θ_n are independent parameters and are never primed.

The calculation begins with the evaluation of $\partial M_z/\partial z$ by differentiating expression (6) and noting that

$$\frac{d}{dx} u(x) = \delta(x) \tag{18}$$

where $\delta(x)$ is the Dirac delta function. Then

$$\frac{\partial M_z}{\partial z} = M_s kg \tag{19a}$$

where

$$k[r, r_b(\theta)] \equiv 1 - 2u[r_b(\theta) - r] \tag{19b}$$

and

$$g(z) \equiv \delta\left(z + \frac{h}{2}\right) - \delta\left(-z + \frac{h}{2}\right). \tag{19c}$$

After changing the order of integration, the expression for the internal magnetostatic energy becomes

$$E_M = \frac{1}{2} M_s^2 \int_0^\infty \int_0^\infty \int_0^{2\pi} \int_0^{2\pi} \int_{-\infty}^\infty \int_{-\infty}^\infty \frac{kk'gg'rr'}{s} dz\, dz'\, d\theta\, d\theta'\, dr\, dr'. \tag{20}$$

The factor $g(z)g(z')/s$ contains the z and z' dependence of this integral. From expression (19c) it can be seen that this factor consists of four terms. Application of the transformation $(z, z') \rightarrow (-z, -z')$ to two of the terms under the integral sign combines these four terms into two terms. Making the transformation $(z, z') \rightarrow (z, z)$, where

$$z \equiv z - z', \tag{21}$$

on the remaining terms and carrying out the integration over z yields the expression for the internal magnetostatic energy in terms of an integral over surface magnetic charges. This expression is

$$E_M = M_s^2 Z \int_0^\infty \int_0^\infty \int_0^{2\pi} \int_0^{2\pi} \frac{kk'rr'}{s} d\theta\, d\theta'\, dr\, dr' \tag{22}$$

where Z is an operator defined by

$$Z\{\ \} \equiv \int_{-\infty}^\infty dz[\delta(z) - \delta(z - h)]\{\ \} \tag{23}$$

$$s^2 = r^2 + r'^2 - 2rr' \cos(\theta - \theta') + z^2. \tag{24}$$

The factor kk' contains the r_n and θ_n dependence of the integral so that the derivatives of E_M may be calculated by replacing this factor by its derivatives under the integral. Evaluating the first derivatives yields

$$\frac{\partial kk'}{\partial r_n} = k \frac{\partial k'}{\partial r_b(\theta')} \frac{\partial r_b(\theta')}{\partial r_n} + k' \frac{\partial k}{\partial r_b(\theta)} \frac{\partial r_b(\theta)}{\partial r_n}$$

$$= k \frac{\partial k'}{\partial r_b} \cos[n(\theta' - \theta_n)] + k' \frac{\partial k}{\partial r_b} \cos[n(\theta - \theta_n)] \tag{25a}$$

$$\frac{\partial kk'}{\partial \theta_n} = k\frac{\partial k'}{\partial r_b(\theta')}\frac{\partial r_b(\theta')}{\partial \theta_n} + k'\frac{\partial k}{\partial r_n(\theta)}\frac{\partial r_b(\theta)}{\partial \theta_n}$$

$$= -k\frac{\partial k'}{\partial r_b} nr_n \sin\,[n(\theta' - \theta_n)] - k'\frac{\partial k}{\partial r_n} nr_n \sin\,[n(\theta - \theta_n)]. \quad (25b)$$

Substituting these derivatives into the integral and exchanging the primed and unprimed r and θ. The first term becomes identical to the second. The derivatives of the internal magnetic interaction energy are then

$$\frac{\partial E_M}{\partial r_n} = 2M_s^2 Z \int_0^\infty \int_0^{2\pi} \int_0^\infty \int_0^{2\pi} \frac{1}{s} k'\frac{\partial k}{\partial r_b} \cdot \cos\,[n(\theta - \theta_n)]r'r\, d\theta'\, dr'\, d\theta\, dr \quad (26a)$$

$$\frac{\partial^2 E_M}{\partial r_n\, \partial r_m} = 2M_s^2 Z \int_0^\infty \int_0^{2\pi} \int_0^\infty \int_0^{2\pi} \frac{1}{s} \cdot \left\{k'\frac{\partial^2 k}{\partial r_b^2} \cos\,[n(\theta - \theta_n)] \cos\,[m(\theta - \theta_m)] + \frac{\partial k'}{\partial r_b}\frac{\partial k}{\partial r_b} \cdot \cos\,[n(\theta - \theta_n)] \cos\,[m(\theta' - \theta_m)]\right\} r'r\, d\theta'\, dr'\, d\theta\, dr \quad (26b)$$

with analogous expressions for

$$\partial E_M/\partial \theta_n\,, \qquad \partial^2 E_M/\partial r_n\, \partial \theta_m\,, \quad \text{and} \quad \partial^2 E_M/\partial \theta_n\, \partial \theta_m\,.$$

For circular domains ($r_b = r_0$) the factors k, k', $\partial k/\partial r_b$, and $\partial k'/\partial r_b$ are independent of θ and θ' so that the integrands are periodic in θ' with periodicity 2π. The range of integration of θ' may therefore be changed from $[0, 2\pi]$ to $[\theta, 2\pi + \theta]$ so that after making the transformation $(\theta, \theta') \rightarrow (\theta, \zeta)$ where

$$\zeta \equiv \theta' - \theta, \quad (27)$$

the range of integration of both θ and ζ is again $[0, 2\pi]$. Note that now

$$s^2 = r^2 + r'^2 - 2rr' \cos\,\zeta + z^2 \quad (28)$$

depends only on ζ.

Using trigonometric identities, the integrands of the integrals for the various derivatives are written as a sum of terms each of which is the product of a factor depending only on θ and a factor depending on ζ. Carrying out the integration over θ yields, for $r_b(\theta) = r_0$,

$$\left(\frac{\partial E_M}{\partial r_0}\right)_0 = 4\pi M_s^2 Z \int_0^\infty \int_0^\infty \int_0^{2\pi} \frac{1}{s} k'\frac{\partial k}{\partial r_b} r'r\, d\zeta\, dr'\, dr, \quad (29a)$$

$$\left(\frac{\partial^2 E_M}{\partial r_0^2}\right)_0 = 4\pi M_s^2 Z \int_0^\infty \int_0^\infty \int_0^{2\pi} \frac{1}{s}\left(k' \frac{\partial^2 k}{\partial r_b^2} + \frac{\partial k'}{\partial r_b}\frac{\partial k}{\partial r_b}\right) r' r \, d\zeta \, dr' \, dr, \quad (29b)$$

$$\left(\frac{\partial^2 E_M}{\partial r_n^2}\right)_0 = \frac{1}{2} 4\pi M_s^2 Z \int_0^\infty \int_0^\infty \int_0^{2\pi} \frac{1}{s} \cdot \left(k' \frac{\partial^2 k}{\partial r_b^2} + \frac{\partial k'}{\partial r_b}\frac{\partial k}{\partial r_b} \cos n\zeta\right) r' r \, d\zeta \, dr' \, dr, \quad n > 0, \quad (29c)$$

while all the remaining first and second derivatives are zero. Note that by inspection of these integrals and the definitions of k, k', and r_b that

$$\left(\frac{\partial^2 E_M}{\partial r_0^2}\right)_0 = \frac{\partial}{\partial r_0}\left(\frac{\partial E_M}{\partial r_0}\right)_0 \quad (30a)$$

and

$$\left(\frac{\partial^2 E_M}{\partial r_n^2}\right)_0 = \frac{1}{2}\frac{\partial}{\partial r_0}\left(\frac{\partial E_M}{\partial r_0}\right)_0 - \frac{1}{2} 4\pi M_s^2 Z \int_0^\infty \int_0^\infty \int_0^{2\pi} \frac{\partial k'}{\partial r_b}\frac{\partial k}{\partial r_b}\frac{(1 - \cos n\zeta)}{s} \cdot r' r \, d\zeta \, dr' \, dr, \quad n > 0. \quad (30b)$$

Noting that from expressions (18) and (19b)

$$\frac{\partial k}{\partial r_b} = -2\delta(r - r_b) \quad (31)$$

and using the definition of the Z operator given in expression (24), expression (29a) may be integrated with respect to r and z, and the second term of expression (30b) may be integrated with respect to r, r', and z. The result after some rearrangement is

$$\left(\frac{\partial E_M}{\partial r_0}\right)_0 = -(2\pi h^2)(4\pi M_s^2) F(2r_0/h), \quad (32a)$$

$$\left(\frac{\partial^2 E_M}{\partial r_0^2}\right)_0 = -(4\pi h)(4\pi M_s^2)\frac{\partial F(2r_0/h)}{\partial(2r_0/h)}, \quad (32b)$$

$$\left(\frac{\partial^2 E_M}{\partial r_n^2}\right)_0 = -(2\pi h)(4\pi M_s^2)\frac{\partial F(2r_0/h)}{\partial(2r_0/h)} + (h)(4\pi M_s^2)\frac{2r_0}{h}\left[L_n\left(\left(\frac{h}{2r_0}\right)^2\right) - L_n(0)\right] \quad (32c)$$

where renaming r' to r and using expression (19b)

$$F\left(\frac{2r_0}{h}\right) \equiv \frac{2r_0}{\pi h^2}[2B(r_0, r_0, h) - 2B(r_0, r_0, 0) - B(r_0, \infty, h) + B(r_0, \infty, 0)], \quad (33a)$$

$$B(r_0 , r_f , z) \equiv \int_0^{\pi} \int_0^{r_b} (\rho^2 + z^2)^{-\frac{1}{2}} r \, dr \, d\zeta, \tag{33b}$$

$$\rho^2 = r_0^2 + r^2 - 2r_0 r \cos \zeta, \tag{33c}$$

and where

$$L_n[(z/2r_0)^2] \equiv \int_0^{\pi} [(z/2r_0)^2 + \tfrac{1}{2}(1 - \cos \zeta)]^{-\frac{1}{2}}(1 - \cos n\zeta) \, d\zeta. \tag{34}$$

The L_n functions are reduced to standard elliptic integral form, and power series expansions are obtained for both large and small values of the argument in Appendix A. The B function is integrated once after displacing the origin of the cylindrical coordinate system from 0 to 0′ as is shown in Fig. 2. The transformation connecting the (r, ζ) and (ρ, φ) coordinate systems is

$$\rho \sin \varphi = r \sin \zeta \tag{35a}$$

$$\rho \cos \varphi = r \cos \zeta - r_0 . \tag{35b}$$

After the transformation

$$B(r_0 , r_f , z) = \begin{cases} \displaystyle\int_0^{\pi} \int_0^{r=r_f} \frac{\rho \, d\rho \, d\varphi}{(\rho^2 + z^2)^{\frac{1}{2}}}, & r_0 < r_f \qquad (36a) \\ \displaystyle\int_{\pi/2}^{\pi} \int_0^{r_0=r_f} \frac{\rho \, d\rho \, d\varphi}{(\rho^2 + z^2)^{\frac{1}{2}}}, & r_0 = r_f . \qquad (36b) \end{cases}$$

Equations (36a) and (36b) are integrated to obtain, in either case,

$$B(r_0 , r_f , z) = \int_{\zeta=0}^{\zeta=\pi} (\rho_b^2 + z^2)^{\frac{1}{2}} \, d\varphi - \int_{\zeta=0}^{\zeta=\pi} |\, z \,| \, d\varphi \tag{37}$$

where ρ_b is the value of ρ along the boundary $r = r_f$. For $r_f = r_0$, $\rho_b = -2r_0 \cos \varphi$ so that

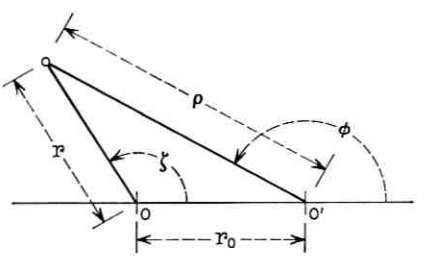

Fig. 2 — The (ζ, r) and (ρ, φ) coordinate systems.

$$B(r_0 , r_0 , h) = \int_{\pi/2}^{\pi} [h^2 + (2r_0)^2 \cos^2 \varphi]^{\frac{1}{2}} \, d\varphi - \frac{\pi}{2} \mid h \mid \qquad (38a)$$

and

$$B(r_0 , r_0 , 0) = 2r_0 . \qquad (38b)$$

The remaining two terms of equation (33a) must be evaluated as a limit

$$\begin{aligned} &\lim_{r_f \to \infty} [B(r_0 , r_f , 0) - B(r_0 , r_f , h)] \\ &\quad = \lim_{r_f \to \infty} \int_0^{\pi} - [(\rho_b^2 + h^2)^{\frac{1}{2}} - \rho_b] \, d\varphi + \pi \mid h \mid \\ &\quad = \pi \mid h \mid \end{aligned} \qquad (39)$$

since ρ_b approaches infinity when r_f approaches infinity. Combining these results yields

$$F(2r_0/h) = \frac{2}{\pi} (2r_0/h)^2 \left\{ \int_0^{\pi/2} [(h/2r_0)^2 + \sin^2 \varphi]^{\frac{1}{2}} \, d\varphi - 1 \right\} \qquad (40)$$

and

$$\frac{\partial F(2r_0/h)}{\partial (2r_0/h)} = (h/2r_0) \left\{ 2F(2r_0/h) - \frac{2}{\pi} \int_0^{\pi/2} [(h/2r_0)^2 + \sin^2 \varphi]^{-\frac{1}{2}} \, d\varphi \right\}. \qquad (41)$$

Appendix B lists the standard elliptic integral form of the force function F and power series expansions for large and small values of the argument. In Fig. 3 the force function is plotted as a function of the domain diameter measured in units of the plate thickness

$$d/h = 2r_0/h. \qquad (42)$$

The stability functions S_n , also shown on this plot, are defined in Section 5.1.

IV. THE ENERGY VARIATION—ORIGIN OF TERMS

Summing the results of the last section according to expression (7), the total energy variation expression (11) is

$$\begin{aligned} \triangle E = {} & [2\pi h \sigma_w + 4\pi r_0 h M_s H - (2\pi h^2)(4\pi M_s^2) F(2r_0/h)] \, \triangle r_0 \\ & + \frac{1}{2} \left[4\pi h M_s H - (4\pi h)(4\pi M_s^2) \frac{\partial F(2r_0/h)}{\partial (2r_0/h)} \right] (\triangle r_0)^2 \\ & + \frac{1}{2} \sum_{n=1}^{\infty} \left\{ \frac{\pi}{r_0} h \sigma_w n^2 + 2\pi h M_s H - (2\pi h)(4\pi M_s^2) \frac{\partial F(2r_0/h)}{\partial (2r_0/h)} \right. \\ & \left. + (h)(4\pi M_s^2) \frac{2r_0}{h} [L_n((h/2r_0)^2) - L_n(0)] \right\} (\triangle r_n)^2 + O_3 \end{aligned} \qquad (43)$$

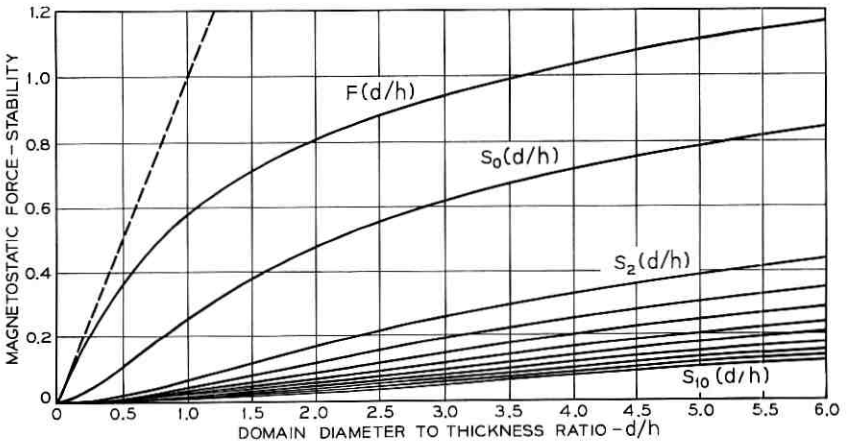

Fig. 3 — The magnetostatic radial force function F and stability functions, $S_0 - S_{10}$, $S_1 = 0$, as functions of domain diameter to thickness ratio, d/h.

where F is defined by expression (33) and plotted in Fig. 3, the L_n are defined by expression (34), and all terms not explicitly stated are equal to zero. The remainder of this section treats the physical origin of the terms in the energy variation expression (43).

4.1 *The Generalized Forces*

The coefficients of the linear variation terms are the negatives of the generalized forces. All forces except the r_0 force are identically zero, which for a circular domain is a consequence of the rotational symmetry of the system. The first term in the coefficient of $\triangle r_0$ is the product of the wall energy density σ_w and the rate of change of wall area with respect to r_0, $2\pi h$. The second term is the product of the external field interaction energy density $2M_sH$ and the rate of change of domain volume with respect to r_0, $2\pi h r_0$. The third term is the rate of change of the internal magnetostatic energy with respect to r_0.

The internal magnetostatic force may be identified in expression (43) and using expressions (32), (33), and (39) may be written in the form

$$-\left(\frac{\partial E_M}{\partial r_0}\right)_0 = 2\pi h^2(4\pi M_s^2)F(2r_0/h) \tag{44a}$$

$$= (2\pi r_0 h)(2M_s)\left\{4\pi M_s + \frac{4M_s}{h}\left[\int_0^{2\pi}\int_0^{r_0}\frac{r\,dr\,d\zeta}{(\rho^2+h^2)^{\frac{1}{2}}} - \int_0^{2\pi}\int_0^{r_0}\frac{r\,dr\,d\zeta}{\rho}\right]\right\} \tag{44b}$$

where ρ^2 is given in expression (33c). In expression (44b) the first factor in parentheses is the domain wall area, the second is the change in magnetization at the wall as the wall moves, and the quantity in braces has the form of an H field whose origin will now be interpreted by superposition of sources. The internally produced field arises from the superposition of the internal field of a plate uniformly magnetized normal to its surface with magnetization magnitude M_s and two disks of magnetic charge of uniform magnetic surface charge density $\pm 2M_s$ and radius r_0. The first term within the braces is thus the demagnetizing field of the infinite plate of uniform magnetization. The second term is the difference in magnetostatic potential between a point on the edge of a disk of magnetic surface charge of uniform density $4M_s$ and a point removed a distance h from this point in a direction normal to the plane of the charge disk divided by the distance h. This is just the z-averaged z-component of the field produced along the wall by the two charge disks since

$$\langle H_z \rangle_{av} \equiv \frac{1}{h} \int_0^h H_z \, dz = \frac{1}{h} \int_0^h - \frac{\partial \Omega}{\partial z} \, dz = - \frac{\Omega(h) - \Omega(0)}{h} \tag{45}$$

where Ω denotes the magnetostatic scalar potential and z is measured from the edge of the disk. Comparing expressions (44a) and (44b), the total internally produced z-averaged z-component of the magnetic field along the domain wall is

$$\langle H_{M_z} \rangle_{av} = -(4\pi M_s)(h/2r_0) F(2r_0/h) \tag{46}$$

so that the total force per unit wall area (averaged over z) is

$$-\frac{1}{2\pi r_0 h} \left(\frac{\partial E_T}{\partial r_0} \right)_0 = -\frac{\sigma_w}{r_0} - 2M_s(H + \langle H_{M_z} \rangle_{av}). \tag{47}$$

The first term is the product of the wall energy density and the wall curvature and always corresponds to an inward directed force. The second term is the change of magnetization at the moving domain wall times the z-averaged z-component of the total field at the wall. [The problem may initially be set up using this fact (Ref. 2, pp. 1922–1925).] The properties of the force function will now be examined in some detail. From expressions (43) or (44a) the first order variation in internal magnetostatic energy, when r_0 is varied, is

$$\Delta E_M = -2(\pi h^2)(4\pi M_s^2) F(2r_0/h) \Delta r_0 . \tag{48}$$

The plot of F in Fig. 3, and the expansions for large and small values

of the argument show that the force function is everywhere positive, is monotonic increasing, and has a negative second derivative. Since the force function is everywhere positive, the internal magnetic interaction energy at all times acts in such a way as to expand the domain. Section 4.2 treats the effect of the slope and curvature properties of the force curve on domain size and stability. Substituting the expansion of the force function for small values of the argument (138d) into expression (48) produces the energy variation for small values for r_0/h,

$$\triangle E_M = \{2\pi h r_0 2M_s(-4\pi M_s) + (4\pi M_s^2)16r_0^2 - (\pi h^2)(4\pi M_s^2)2[\tfrac{1}{4}(2r_0/h)^3 - \tfrac{3}{64}(2r_0/h)^5 + \cdots]\}\triangle r_0 . \quad (49)$$

(In the remainder of this section frequent reference will be made to the properties of F and the L_n given in Appendices A and B.) The interaction of the magnetization with the existing field from the infinite dipole sheet produces the first term in expression (49). This may be seen by comparison with expression (47) and observing that the field internally generated in the infinite dipole sheet with no reversals is $-4\pi M_s$. In Fig. 3 a dashed line through the origin with numerical slope one represents this term and forms the small r_0/h asymptotic of F.

The second term in expression (49) is the only thickness independent term in the expansion and therefore must be identical to the variation of self-energy of the two disks of magnetic charge which form the ends of the reversal when r_0 is varied. Since the interaction with the infinite charge sheet and the self-energy of the disk have been taken into account, the remaining terms are the mutual interaction of the magnetic charge disks.

For large r_0/h, an energy expansion in terms of h/r_0 is appropriate. Substituting the expansion of the force function for large values of the argument (138c) into expression (48) yields

$$\triangle E_M = -h^2(4\pi M_s^2)\left\{[1 + \tfrac{3}{16}(h/2r_0)^2 + 0_4] + [2 - \tfrac{1}{4}(h/2r_0)^2 + 0_4] \ln \left| 4\frac{2r_0}{h} \right| \right\}\triangle r_0 . \quad (50)$$

This expansion obscures the identity of both the infinite sheet magnetic field term and the charge plate self-energy term so that a local (to the wall) magnetic energy lowering per unit line length description appears appropriate. However, the energy reduction per unit line length to lowest order in $2r_0/h$ is

$$\frac{E_M \text{ (domain)} - E_M \text{ (uniform magnetization)}}{2\pi r_0} = -\frac{h^2}{\pi}(4\pi M_s^2) \ln \left| \frac{4}{e^{\frac{1}{2}}} \frac{2r_0}{h} \right| \qquad (51)$$

so that the energy lowering *per unit line length* for the domain of infinite diameter is infinite. [Equation (51) is obtained by integrating equation (50) to lowest order. The integration constant is determined to be zero by term by term integration of the expansions of F for large and small values of the argument and comparing at $2r_0/h = 1$.] The conclusion that the energy lowering per unit line length for an isolated straight line reversal may also be obtained by considering the energy lowering in a strip reversal when the strip width approaches infinity. The author's intention at the outset of this entire calculation was to calculate the numerical value of this magnetic energy reduction per unit wall length. The internal magnetic interaction, however, retains just enough of its global character when the domain is very large so that no finite limiting value for this energy reduction exists.

The internally generated magnetic field at the wall of the domain, for large r_0/h is obtained from expression (46). To lowest order it is

$$\langle H_{M_z} \rangle_{\text{av}} = -\frac{(4\pi M_s)}{\pi} \frac{h}{2r_0} \ln \left| 4e^{\frac{1}{2}} \frac{2r_0}{h} \right| \qquad (52)$$

which approaches zero as the diameter approaches infinity as it must, since for an infinite straight line magnetization reversal, symmetry requires that the z-component of the field be zero along the reversal.

4.2 *The Stiffness Matrix*

The second variation of the energy with respect to the Fourier coefficients describing the domain determines the stability of the domain. Since the stiffness of the domain with respect to externally applied forces is proportional to the coefficient of the bilinear form which is the second variation of the energy, the matrix formed by these coefficients is called the stiffness matrix. The stiffness matrix is composed of three independent submatrices. The second derivatives of the energy with respect to the Fourier amplitudes form the radial stiffness matrix; the second derivatives of the energy with respect to the Fourier phases form the angular stiffness matrix; and the derivatives of the energy with respect to one Fourier amplitude and one Fourier phase form the mixed stiffness matrix. The derivative of the energy with respect to r_n and r_m

are called the (n, m) radial stiffness matrix element, with similar notation for the other submatrices.

All derivatives not explicitly exhibited in expression (43) are zero. Thus, the angular stiffness matrix and the mixed stiffness matrix are zero and the radial stiffness matrix is diagonal so that the system is completely metastable with respect to angle and the amplitudes are normal modes of the system for small amplitudes.

The (0, 0) radial stiffness matrix element is simply the derivative of the negative of the radial generalized force so that no further discussion of it is necessary. It should be noted that the derivative of the internal magnetostatic term with respect to wall position is not directly related to the radial field or potential at the wall since the derivative used in computing the radial field at the wall must be taken with the wall position held fixed.

4.2.1 *The Radial Stiffness Matrix Elements for* $n \geqq 1$

The diagonal radial stiffness matrix elements, for $n \geqq 1$, are the sum of four terms in expression (43). The first term, which always has a stabilizing effect, is the increase in total wall energy due to the lengthening of the wall caused by the deviation from a strictly circular shape. Imposing a sinusoidal variation of amplitude $\triangle s$ onto a straight line produces a relative increase in length of

$$\frac{s + \triangle s}{s} = 1 + \left(\frac{\pi \triangle r_n}{\lambda_n}\right)^2 + \cdots . \tag{53}$$

The corresponding wavelength in expression (43) is

$$\lambda_n = \frac{2\pi r_0}{n}. \tag{54}$$

The wall energy term in expression (43),

$$\triangle E_{Wn} = \sigma_w 2\pi r_0 h \left(\frac{\pi \triangle r_n}{\lambda_n}\right)^2 , \tag{55}$$

is thus the product of the wall energy density, the wall area, and the variation in wall area per unit area. Notice that the relative variation in wall length or area is independent of the wall curvature, $1/r_0$, to lowest order in the amplitude of the variation.

The second order change in volume of the domain interacting with the externally applied field produces the second term in the radial stiffness matrix elements while the rate of change of the internal magnetostatic forces at the wall produces the third term. The sum of the

second and third terms is one-half the (0, 0) radial stiffness matrix element. This factor of one-half relates to the fact that a variation of $\triangle r_n$, $n \geqq 1$ produces only one-half the mean square variation $r_b(\theta)$ as is produced by an equal variation in r_0. This shape-independent, second-order variation in energy arises from the variation in the generalized forces, or fields at the wall, when the domain radius is varied. [See also the steps leading to expression (30b).]

4.2.2 *Translation Invariance*

The requirement of translation invariance in the infinite plate completely determines the (1, 1) radial stiffness matrix element. Consider a cylindrical domain of radius r_0 with a cylindrical coordinate system placed at its center. Under a displacement of the coordinate system of magnitude s in the $\theta = \pi$ direction, the description of the boundary in the new coordinate system is

$$r_b(\theta) = r_0 - \frac{1}{4}\frac{s^2}{r_0} + s\cos\theta + \frac{1}{4}\frac{s^2}{r_0}\cos 2\theta + 0_4 . \tag{56a}$$

Thus, to second order in s, term by term comparison with definition (3) yields

$$\triangle r_0 = -\frac{1}{4}\frac{s^2}{r_0}, \qquad \triangle r_1 = s, \quad \text{and} \quad \triangle r_2 = \frac{1}{4}\frac{s^2}{r_0}. \tag{56b, c, d}$$

The formal change in energy under this displacement (11) is

$$\triangle E = \left(\frac{\partial E}{\partial r_0}\right)_0 \left(-\frac{s^2}{r_0}\right) + \frac{1}{2}\left(\frac{\partial^2 E}{\partial r_1^2}\right)_0 s^2 + 0_3 . \tag{57}$$

Obtaining $(\partial E/\partial r_0)_0$ and $(\partial^2 E/\partial r_1^2)_0$ from expression (43), and substituting expressions (84), (85), (86), (100), and (138a) verifies that

$$\left(\frac{\partial^2 E_T}{\partial r_1^2}\right)_0 = \frac{1}{2r_0}\left(\frac{\partial E_T}{\partial r_0}\right)_0 . \tag{58}$$

The coefficient of s^2 in expression (57) is thus zero as required by translation invariance, and further the (1, 1) stiffness matrix element is zero whenever the total radial generalized force is zero.

4.2.3 *The Magnetostatic Stiffness Terms*

The interpretation of the radial stiffness matrix elements for the higher n values is now considered. As in the case of the generalized forces, examination of the expansions for small r_0 allows the self-interaction energy of the two charge disks which make up the ends of the

domain to be separated from the mutual interaction of these charges. The variation in the internal magnetostatic energy due to a variation in some r_n for a circular domain is in general from expression (43)

$$\triangle E_{Mn} \equiv 4\pi M_s^2\left\{-(\pi h)\frac{\partial F(2r_0/h)}{\partial(2r_0/h)} + r_0\left[L_n\left(\frac{h^2}{4r_0^2}\right) - L_n(0)\right]\right\}(\triangle r_n)^2, \qquad n \geqq 1. \tag{59}$$

Separating the h independent and h dependent terms of the power series representation in powers of $2r_0/h$ uniquely separates the above expression into two parts, one part representing the self-interaction of the charge disks and the other representing the mutual interaction of these disks. The h independent terms then represent the self-interaction forces of the charge disks and the h dependent terms represent the mutual interaction forces. In the expansion of L_n for large $(h/2r_0)^2$, expression (129a), all terms of $L_n[(h/2r_0)^2]$ are h dependent. Using the large $(h/2r_0)^2$ expansion of F, expression (138d), and the expressions for $L_n(0)$, (115) and (116), the thickness independent part of expression (59) is

$$\triangle E_{Mn}(\text{Self}) = 4\pi M_s^2 r_0\left(8 - 4\sum_{j=1}^{n}\frac{1}{2j-1}\right)(\triangle r_n)^2, \qquad n \geqq 1. \tag{60}$$

This energy variation contains a term which results from the variation in the overall size of the disks of charge as well as the shape dependent terms. The size variation term will now be identified and subtracted out so that the shape dependent part of the self-interaction energy may be seen explicitly. From expression (49) and the discussion following it, the ratio of the variation in energy of two isolated disks to the variation in disk area is $(4\pi M_s^2)(16r_0/2\pi)$. The variation in disk area for a variation in r_n for $n \geqq 1$ is $(\pi/2)(\triangle r_n)^2$ so that the change in self-energy of the two disks, other than that due to their mutual interaction or change in overall size, is

$$\triangle E_{Mn}(\text{Self-Shape}) = \begin{cases} 0, & n = 1 \qquad (61a) \\ -(4\pi M_s^2)4r_0\left(\sum_{j=2}^{n}\frac{1}{2j-1}\right)(\triangle r_n)^2, & n > 1. \qquad (61b) \end{cases}$$

It is not surprising that the variation of r_1 produces no shape related energy change, since from expression (56) this variation is to lowest order a displacement with a size change coming in second order. It is seen that the terms which remain after cancellation all come from $L_n(0)$.

In expression (59) the first term is independent of n and the $-4\pi M_s^2 r_0 L_n(0)(\Delta r_n)^2$ term has been identified with the variation in the self-energy of the charge disks. The term $4\pi M_s^2 r_0 L_n[(h/2r_0)^2](\Delta r_n)^2$ must therefore contain all of the shape dependent part of the charge disk mutual interaction energy. This term also contains a contribution due to the variation in the total amount of charge and contribution due to the shape independent, general smearing out of the charge distribution. Since the second order change in the total amount of charge is independent of n for $n \geqq 1$, these two contributions may be removed from the mutual interaction energy variation by replacing L_n by $L_n - L_\infty$. The remaining mutual interaction energy variation is specifically due to the shape of the variation. This energy variation is

$$\begin{aligned}
&\Delta E_{Mn}(\text{Mutual-Shape}) \\
&\quad = (4\pi M_s^2) r_0 \left[L_n\left(\frac{h^2}{4r_0^2}\right) - L_\infty\left(\frac{h^2}{4r_0^2}\right)\right](\Delta r_n)^2, \qquad n \geqq 1 \\
&\quad = (4\pi M_s^2)\, r_0 \left[M_{n,2n+1}\left(\frac{2r_0}{h}\right)^{2n+1} + M_{n,2n+3}\left(\frac{2r_0}{h}\right)^{2n+3} + \cdots \right], \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad n \geqq 1 \qquad (62)
\end{aligned}$$

where the final form is obtained using the expansion for L_n, equation (131), and the $M_{n,m}$ are the constants of the expansion. The interaction energy of planar multipoles of order n and higher has the form of equation (62), as it must since the variation in the charge distribution for each n may be expressed in terms of such multipoles.

The variation in internal magnetostatic energy due to a variation of r_n, in the infinite sheet, for large $2r_0/h$, to lowest order in $h/2r_0$, is

$$\begin{aligned}
\Delta E_{Mn} &= (4\pi M_s^2)\left(\frac{h^2}{4r_0}\right) \\
&\quad \cdot \left[-2n^2 \ln\left| 4\frac{2r_0}{h}\right| - 2n^2 - 2 + (4n^2 - 1)\sum_{j=1}^{n} \frac{1}{2j-1}\right](\Delta r_n)^2, \\
&\qquad\qquad\qquad\qquad\qquad\qquad n \geqq 1 \qquad (63)
\end{aligned}$$

using equation (59), the large $2r_0/h$ expansion of F (138c) and of $L_n(h^2/4r_0^2) - L_n(0)$, (105), (116) and (125).

The charge-disk self-interaction energy is not evident in this expansion because it is exactly cancelled by the leading term of the mutual interaction energy. In contrast to the energy reduction per unit line length for a straight line reversal in an infinite sheet (which has no

finite value), it is possible in the case of this variation of the domain structure to compute the energy variation per unit line length. In terms of the wavelength of the variation λ, defined in expressions (54) or (127) and in the limit of $r_0 \to \infty$, the total variation in energy when r_n is varied is [using expressions (43) and (55) and the limit (128)]

$$\frac{\triangle E_T}{2\pi r_0} = \left\{\left[\pi^2\sigma_w/h - (4\pi M_s^2)\pi \ln \left| \frac{4e\lambda}{\pi h} \right|\right] \frac{h^2}{\lambda^2} + O_4\left(\frac{h}{\lambda}\right)\right\}(\triangle r_n)^2. \quad (64)$$

Comparison of expression (64) with expression (53) shows that the magnetostatic energy variation per unit line length for a circle of infinite diameter is the product of the magnetostatic energy density constant, the variation in line length, and the logarithm of a maximum effective interaction distance, $4e\lambda/\pi$. (The maximum effective interaction distance for the magnetostatic energy lowering per unit line length is proportional to r_0.) Hagedorn has computed the magnetostatic energy variation per unit line length for the case of a sinusoidal variation imposed on an infinite straight line reversal.[6] The calculation was carried out by considering the energy variation produced by a sinusoidal applied to a strip domain pattern in the limit of infinite strip width. The result of this calculation is

$$\Delta E_M/(\text{unit length}) = -(4\pi M_s^2)\pi \ln | \lambda/(2.111h) | (h/\lambda)^2(\Delta r)^2, \quad (65)$$

which differs from the result for the infinite circle by the constant inside the logarithm.

4.3 *Summary*

The physical origin of terms of the energy variation has thus been traced in the limiting cases of both large and small r_0/h. In either of these limiting cases, it is thus possible to develop intuition with regard to the behavior of the domains. Since, as has been shown, the interpretation of the meaning of the energy terms in the limiting cases is qualitatively different, the development of intuition in the transition region is quite difficult. In many device applications this transition region is the preferred region of operation, making the use of analytical and numerical methods a necessity.

V. THE SIZE AND STABILITY OF CYLINDRICAL DOMAINS

The energy variation expansion (43) in principle contains all cylindrical domain size and stability information. This section treats briefly the use of this expression in the determination of domain size and stability. The only non-zero generalized force in expression (43) is the

uniform radial force. When this force is set equal to zero (the force equation), the system is in equilibrium. Thus the condition that the system be in equilibrium provides, given a material and plate thickness, an equation relating domain size and the applied field. The location of the zeros in expression (43) (all terms not explicitly exhibited are zero) shows that the system is completely metastable with respect to angle and that the radial stiffness matrix is diagonal. The radial amplitudes are thus quasi-normal modes, and the study of stability reduces to the study of the stability of the individual radial amplitudes.

5.1 *Normal Form of the Energy Expansion*

Before proceeding with the discussion, it is appropriate to introduce some new notation and to rearrange the energy variation expansion into what will be called normal form. Since the stiffness matrix is of interest only when the domain is in equilibrium, the applied field H is eliminated from it using the force equation. The geometrical dependences of the various magnetostatic stability terms are then combined and normalized to the wall stiffness term by defining the "stability functions" as

$$S_0(d/h) \equiv F(d/h) - d\frac{\partial}{\partial d}F(d/h) \tag{66a}$$

and

$$S_n(d/h) \equiv -\frac{1}{n^2-1}\left\{S_0(d/h) + \frac{1}{2\pi}(d^2/h^2)[L_n(h^2/d^2) - L_n(0)]\right\}, \quad n \geqq 2. \tag{66b}$$

The S_1 function is undefined or may be taken to be zero since translation invariance in the infinite plate requires that the (1, 1) stiffness matrix element the identically zero whenever the generalized radial force is zero, as is assumed to be the case here. The S_n functions are plotted in Fig. 3 up to S_{10} ; they are given in standard elliptic integral form together with power series expansions for large and small values of the argument in Appendix B. The domain diameter, $d = 2r_0$ represents domain size in this section. The normal form of the energy expansion is written as a function of the ratios of the three fundamental lengths of the system: the plate thickness h, the domain diameter d, and the "characteristic length" defined by

$$l \equiv \frac{\sigma_w}{4\pi M_s^2}. \tag{67}$$

The characteristic length depends only on the type of material used.

Dividing the energy variation expansion (43) by the normalizing energy $2(4\pi M_s^2)(\pi h^3)$ and introducing the notation of the preceding paragraph, the normal form of the energy expansion results:

$$\frac{\triangle E_T}{2(4\pi M_s^2)(\pi h^3)} = \left[\frac{l}{h} + \frac{d}{h}\frac{H}{4\pi M_s} - F\left(\frac{d}{h}\right)\right]\frac{\triangle r_0}{h} + \frac{1}{2}\left\{-\left(2\,\frac{h}{d}\right)\left[\frac{l}{h} - S_0\left(\frac{d}{h}\right)\right]\left(\frac{\triangle r_0}{h}\right)^2 + \sum_{n=2}^{\infty}(n^2 - 1)\left(\frac{h}{d}\right)\left[\frac{l}{h} - S_n\left(\frac{d}{h}\right)\right]\left(\frac{\triangle r_n}{h}\right)^2\right\} + O_3 . \quad (68)$$

In expression (68) the coefficient $-[l/h + (d/h)(H/4\pi M_s) - F(d/h)]$ is the normalized radial force. Setting this force equal to zero yields the normalized force equation. The remaining bracketed quantities $[l/h - S_n(d/h)]$ are proportional to the diagonal elements of the stiffness matrix, and are called "stability coefficients." For uniform radial variation, the stability coefficient has the opposite sign from the (0, 0) element of the radial stiffness matrix; thus this stability coefficient is negative whenever the domain is stable. For the other r_n variations, on the other hand, the stability coefficient has the same sign as the corresponding element in the stiffness matrix, and these stability coefficients are positive whenever the domain is stable.

5.2 *Graphical Solution of the Force Equation*

A graphical solution to the force equation

$$\frac{l}{h} + \frac{d}{h}\frac{H}{4\pi M_s} - F\left(\frac{d}{h}\right) = 0 \quad (69)$$

may be obtained by constructing a straight line on Fig. 3 whose intercept with the vertical axis is l/h and whose numerical slope is $H/4\pi M_s$. The intersections of this straight line with the F curve are then the solutions to the force equation.

As was stated in Section 5.1, (*i*) the force function has a positive first derivative and negative second derivative for all nonzero values of its argument, (*ii*) it is zero and has a first derivative of unity when its argument is zero, and (*iii*) it becomes logarithmic for large values of its argument. From these properties and examination of Fig. 3, several properties of the solutions to the force equation may be appreciated. For negative values of the applied fields, there is only one solution to the force equation. Examination of the sign of the radial force which

results when the diameter is varied about the solution diameter while all other variables are held fixed shows that this solution is unstable. For small positive applied fields, there are two solutions to the force equation, the larger diameter solution being radially stable, the other radially unstable. However, a radially stable solution does not guarantee that the system is stable with respect to all possible deformations, and this must be investigated separately. As the applied field is increased, the two solutions move closer together until they coalesce. When the applied field is increased beyond this point, there are no solutions. Since the function F is asymptotic to a straight line through the origin having unit slope, the solutions will always vanish for a value of the applied field which is greater than $4\pi M_s$. Stable isolated cylindrical domains thus exist only in the presence of an applied field having magnitude between zero and $4\pi M_s$ and polarity tending to collapse the domain.

5.3 *Graphical Determination of Domain Stability*

The stability coefficients are determined graphically by constructing a horizontal line at height l/h on the force stability graph. Metastability for each normal mode of deformation occurs at the intersection of this line with the corresponding stability function. Since the stability functions are monotonic, the diameter of metastability of each normal mode of deformation is uniquely defined and forms the boundary between the regions of stability and unstability. The circular domain will be stable with respect to all variations when its diameter is greater than the radial metastability diameter and less than the metastability diameter for a variation with a rotational periodicity of two. The normal variations with rotational periodicity two are referred to as "elliptical" deformations. When the domain is stable with respect to elliptical deformation, it is necessarily stable with respect to the variations of higher spatial frequency since the stability functions of higher spatial frequency lie progressively (with respect to n) below the elliptical stability function. The radial stability function S_0 and the elliptical stability function S_2 thus form the boundary of the region of total cylindrical domain stability. Therefore, given the magnetic material type and plate thickness, the range of stable domain diameters and the corresponding applied fields may be determined with the aid of these functions.

5.4 *Minimum Domain Diameter*

For any given value of l/h, the minimum domain diameter is the collapse diameter determined by S_0. The domain diameter measured in

units of the characteristic length is $d/l = (d/h)/(l/h)$ which is the inverse of the numerical slope of a line drawn in Fig. 3 from the origin to the operating point. The line of maximum slope, which both passes through the origin and contacts the S_0 curve at at least one point, thus determines the smallest domain diameter attainable in a given material. The coordinates of this contact point are $d/h \approx 1.2$ and $l/h \approx 0.3$, so that the minimum attainable domain diameter is

$$d_{\min} \approx 4l. \tag{70}$$

VI. RANGE OF VALIDITY OF THE MODEL AND THE QUALITY FACTOR

At the present time, no quantitative evaluation of the range of validity of the domain structure model used here has been carried out. The qualitative discussion given here, it is hoped, will provide the reader with an appreciation of the magnitude of the effects produced by the relaxation of the various constraints artificially imposed by the model and the dependence of these effects on the system parameters. It has been assumed that domain walls are cylindrical, have zero width, and have a definite energy per unit area which is independent of wall orientation or curvature, and that the magnetization lies perpendicular to the surface of the plate. Section 6.1 treats the effect of the relaxation of the cylindrical wall approximation only. In Section 6.2, the other assumptions are all shown to be coupled using the simplest uniaxial material model. A single dimensionless material parameter q, which complements the characteristic length l in characterizing circular domain materials, is used to express the results obtained from the simplest material model.

6.1 *The Cylindrical Wall Approximation*

The discussion of the cylindrical wall approximation uses the coordinate system and domain configuration of Fig. 1 except that the walls are allowed to curve as shown in Fig. 4. The radius function, $r_b(\theta, z)$, is determined by the requirement that it minimize the total energy. The Euler equation which results from this two dimensional field variational problem is an integro-differential equation similar to those which appear in Hartree self-consistent field calculations. No solution of this equation, numerical or otherwise, has been attempted or is contemplated at the present time. The Euler equation consists of terms arising from: the wall energy, the interaction of the magnetization with the applied field, the self-interaction of the magnetostatic charges at the surface of the plate, the self-interaction of the charges produced by the slope of the

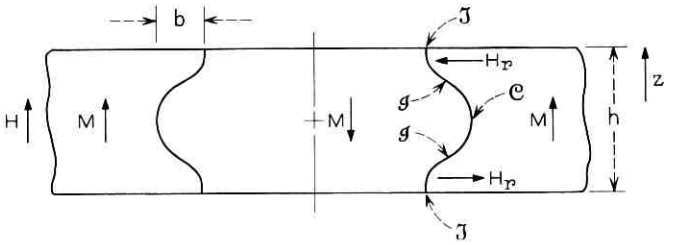

Fig. 4 — Cross section of a noncylindrical, near-circular, domain.

domain wall, and the mutual interaction of the plate surface charges with the domain wall magnetic charges. Boundary conditions (obtained from the appropriate transversality condition[7]) require that the wall surface be perpendicular to the plate surface at all intersection points (𝔍 in Fig. 4). Physically (since in the model used here the crystal is assumed to be strain free) the surface cannot interact with the domain wall, and therefore the wall must intersect the surface at right angles.

Although it is not clear that domains having a roughly conical shape are ruled out, it will be assumed that the domain has reflection symmetry through the central plane of the plate and that the radius is a function of z only, $r_b(z)$. In this case the wall must be vertical at the central plane as indicated at 𝔊 in the figure so that the single parameter b represents the magnitude of the wall bulging. Since the Euler equation requires the curve to be smooth, there must be an inflection point, 𝔤, between 𝔊 and 𝔍. The wall area, and thus total wall energy, is a quadratic increasing function of the wall curvature so that the concentration of the curvature at the center and ends of the wall, produced by the transversality and symmetry conditions, tends to reduce the wall bulging.

The radial field at the domain wall from the charges at the surface of the plate is directed as shown in Fig. 4. The effect of the interaction of the magnetostatic charges due to the slope of the wall with the radial component of the field from the surface charges is destabilizing for either positive or negative bulging. This interaction produces a negative quadratic term in the total energy. However, at the plate surface, where the magnitude of the radial field is greatest, the transversality condition requires that the charge density produced by the wall slope is zero so that the magnitude of this negative term is small. The z component of the field from the charges on the surface of the plate determines the direction of bulging. (The applied field, being uniform by assumption, need not be considered.) Along an initially cylindrical wall the internal

field is everywhere directed, so as to make the domain expand, and attains its greatest magnitude at the center plane of the plate. It therefore provides a linear term in the total energy which tends to bulge the wall in the positive direction as shown in Fig. 4.

Thus, for near cylindrical walls, the bulging is determined by the interaction of this force (tending to bulge the wall) with the wall energy (acting to stabilize the wall) and the radial field (acting to destabilize the wall). The self-interaction of the wall charges enters only as a higher-order term. It should be noted that the transversality condition acts both to strengthen the stabilizing term and weaken the destabilizing term.

The relevant dimensionless wall energy for the wall bulging problem is $l/h = \sigma_w/(h4\pi M_s^2)$. Wall bulging is expected to decrease with increasing wall energy. A second independent effect related to l/h may be appreciated by inspection of the S_0 and S_2 curves in Fig. 3. It can be seen from Fig. 3, equation (68), and the discussion following it that, since the S_0 and S_2 curves bound the region in which stable circular domains exist, d/h must increase with increasing l/h. By symmetry, the z-component of the internally generated magnetic field at a cylindrical wall is zero for a domain of infinite diameter and clearly increases monotonically as the domain diameter to thickness ratio decreases. Thus, as the plate is made thicker, the bulging force becomes stronger and the stabilizing force becomes weaker. Since several independent effects cooperate to increase bulging with increasing plate thickness, the onset may be quite rapid when it does occur. Domain collapse data taken at $d/h \approx 1$ is in good agreement with predictions made on the basis of equation (68) and Fig. 3.[8] This then provides some indication that the cylindrical wall approximation remains valid at this thickness.

6.2 *The Quality Factor*

The discussion of the approximations other than the cylindrical wall approximation uses a polar (M_s, η, ν) coordinate system where η is the polar angle and ν is the aximuthal angle to specify the orientation of M_s (See Fig. 5). The polar axis is taken to be the z-axis of the preceding sections. The domain wall is taken to be planar with its position and orientation specified by a plane at its center. The axis through the origin in the direction of the wall normal is denoted by ξ. The position of the central wall plane is denoted by ξ_0. The orientation angles of the wall normal are denoted by ν_w and η_w, (see Fig. 5).

In the simplest uniaxial material whose easy axis is the z-axis the magnetic energy density for a planar wall is (Ref. 9, pp. 189–192)

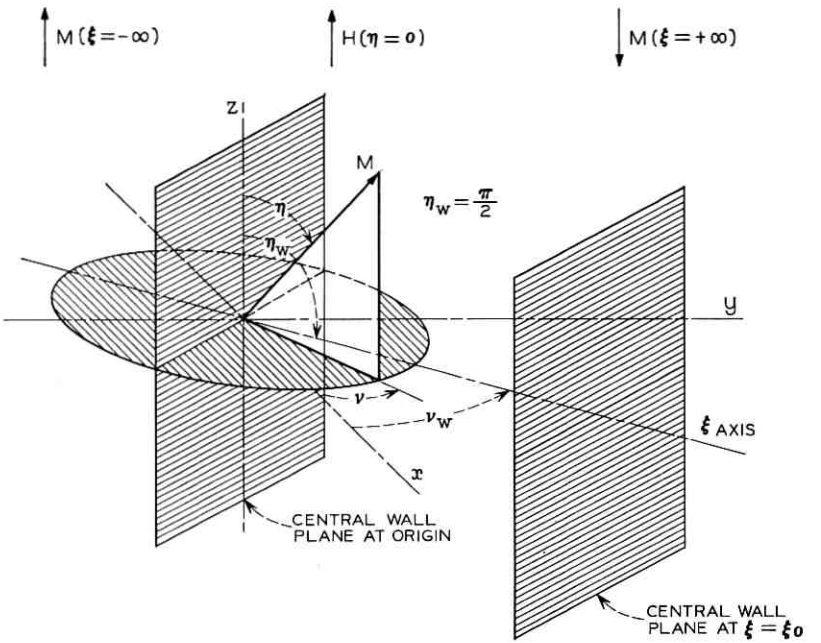

Fig. 5 — Coordinate system for specification of domain walls.

$$\rho_E = A\left[\left(\frac{\partial \eta}{\partial \xi}\right)^2 + \sin^2 \eta \left(\frac{\partial \nu}{\partial \xi}\right)^2\right] + K_u \sin^2 \eta - \mathbf{H}\cdot\mathbf{M} + 2\pi \frac{1}{\xi^2} (\mathbf{M}\cdot\xi)^2 \tag{71}$$

where A is the exchange energy density coefficient, K_u is the anisotropy energy density coefficient, $\mathbf{H}$ is the externally applied field, and the last term is obtained by integrating $\nabla\cdot\mathbf{B} = 0$. For a uniformly magnetized material in the absence of applied or internal fields, this expression reduces to $\rho_E = K_u \sin^2 \eta$ which has absolute minima at $\eta = 0$ and $\eta = \pi$. The z-axis is thus the easy axis as is required for consistency with the preceding sections.

The anisotropy energy density coefficient is sometimes expressed in terms of the effective anisotropy field $H_a \equiv 2k_u/M_s$. The quality factor is now defined as the dimensionless anisotropy energy coefficient or dimensionless anisotropy field

$$q \equiv \frac{K_u}{2\pi M_s^2} = \frac{H_a}{4\pi M_s}. \tag{72}$$

6.2.1 *The Nucleation Field*

When a bias field H is applied in the positive z-direction and demagnetizing fields are neglected the energy density is

$$\rho_E = K_u \sin^2 \eta - HM_s \cos \eta \tag{73}$$

which, for $H < H_a$, has a local minimum at magnetization orientation $\eta = \pi$ and an absolute minimum for $\eta = 0$. When $H > H_a$, only the minimum at $\eta = 0$ remains. In a perfect crystal the effective anisotropy field is thus the field at which the magnetization becomes unstable with respect to reorientation (assuming it is initially oriented in the negative z-direction). If a reorienting field is applied locally (local but over a region whose dimensions are much greater than a wall width so that the effect of exchange forces can be neglected), then H_a is the total local field required for the nucleation of a domain at that locality. If the nucleation field H_N is understood in this sense, then in a perfect crystal q is the nucleation field measured in units of $4\pi M_s$:

$$\frac{H_N}{4\pi M_s} = q. \tag{74}$$

In an imperfect crystal $H_N/4\pi M_s$ may be either larger or smaller than q. If it is larger, the material may be expected to have a high wall motion coercivity.

6.2.2 *Susceptibility*

When a transverse bias field $H_t (\eta = \pi/2)$ is applied and demagnetizing fields are neglected, the energy density becomes

$$\rho_E = K_u \sin^2 \eta - H_t M_s \sin \eta \tag{75}$$

which has stable magnetization orientations

$$\eta = \begin{cases} \sin^{-1}\left(\dfrac{H_t}{H_a}\right) = \sin^{-1}\left(\dfrac{H_t}{4\pi M_s}\dfrac{1}{q}\right), & H_t < H_a; \\ \dfrac{\pi}{2}, & H_t \geqq H_a. \end{cases} \tag{76}$$

The transverse susceptibility is therefore

$$\chi_t = \frac{\partial M_t}{\partial H_t} = \begin{cases} \dfrac{1}{4\pi q} & (H_t < H_a) \\ 0 & (H_t \geqq H_a) \end{cases} \tag{77}$$

where $M_t = M_s \sin \eta$ is the component of the magnetization in the direction of H_t. Thus, the susceptibility to tipping of the magnetization by a transverse field is inversely proportional to q.

6.2.3 *Wall Energy and Wall Width*

Consider now a planar Bloch wall, $\xi \cdot \mathbf{M} = 0$, between two regions whose magnetization at points far from the wall lies along the two easy directions, $\eta = 0$ and π, and again assume that there are no applied fields or fields produced by boundary surfaces. Under these conditions, the magnetic configuration is determined by minimization of the wall energy per unit surface area which in this case is (Ref. 9, pp. 189–192)

$$\sigma_w = \int_{-\infty}^{\infty} \left[A\left(\frac{\partial \eta}{\partial \xi}\right)^2 + K_u \sin^2 \eta \right] d\xi. \tag{78}$$

Carrying out the minimization results in

$$\sigma_w = 4(AK_u)^{\frac{1}{2}} \tag{79}$$

for

$$\xi - \xi_0 = \frac{1}{\pi} l_w \log \tan \left(\frac{\eta}{2}\right) \tag{80}$$

where

$$l_w \equiv \pi \left(\frac{A}{K_u}\right)^{\frac{1}{2}} \tag{81}$$

is the wall width. The definition of wall width is somewhat arbitrary since the wall extends over all space. In this case, following page 191 of Ref. 9, it is chosen so that the magnetization would complete its entire rotation of π radians in a length l_w if the entire rotation took place at its maximum rate, the rate at the center of the wall.

The ratio of the characteristic length, equation (67), to the wall width is

$$\frac{l}{l_w} = \frac{2}{\pi} q, \tag{82a}$$

so that the ratio of the minimum domain diameter, equation (70), to the wall width is

$$\frac{d_{\min}}{l_w} = \frac{4l}{l_w} = \frac{8}{\pi} q. \tag{82b}$$

The approximation of zero wall width thus improves as q becomes larger.

The approximation that the wall energy is independent of wall curvature is clearly related to the wall width. At large distances from the planar wall equation (80) becomes

$$| \eta - \eta_0 | = 2\exp(-\pi | \xi - \xi_0 | / l_w) \tag{83}$$

where η_0 is the appropriate equilibrium orientation of the magnetization at a distance far removed from the wall. Such an exponential relation will hold for the approach to any stable equilibrium orientation in the presence of isotropic exchange. The change in energy of the wall due to overlapping of the tails of the wall as the wall is curved is clearly related to q, becoming larger as q becomes smaller. In order to solve for the dependence of the wall energy on curvature it is necessary to solve the entire (including magnetostatics) micromagnetics problems.[10]

6.2.4 *Summary*

The preceding results may be summarized by noting that the higher the q value, the more closely the simple uniaxial model obeys the constraints of the domain model used in the previous sections. It is clear that, for domains of the type considered to exist at all, q must be greater than one. For device operation, q should probably have a value greater than two.

VII. CONCLUSIONS

The theory of cylindrical magnetic domains yields conditions which predict the size and stability of these domains and provides an estimate of the range of applicability of the model used. The results of theory appear to be accurate in a range useful in the construction of circular domain devices.

The domains considered are isolated right circular cylinders in plates of uniaxial magnetic material of uniform thickness cut so that the plate normal is parallel to the easy axis. The first and second order energy variations which result from a general small deviation from the strictly circular shape determine domain size and stability. The energy method was chosen in preference to the magnetostatic field method because of the uniformity it provides in accounting for the forces in both the equilibrium and stability problems. The integrals arising from the energy method are interpreted physically in terms of fields and interacting charges. The physical interpretation of the integrals is quite different in the limiting cases of very large or very small domains. The integrals are related to special cases of the fields of uniformly charged disks computed by C. Snow and tabulated by N. B. Alexander and A. C.

Downing.[11,12] The present work obtains the needed properties of the integrals (expansions, recursion relations, and others) directly from the definitions.

When the energy variation is described in terms of a Fourier decomposition of the domain radius function, only the generalized force corresponding to a change in domain size is non-zero and the stiffness matrix is completely metastable with respect to angle (phase) and diagonal with respect to the Fourier amplitudes. Since the Fourier amplitude stiffness matrix elements are all found to be distinct, the description is unique and may be described as a quasi-normal mode description.

The normal mode description is summarized by a single graph from which many domain properties may be determined by construction. Cylindrical domains exist only in the presence of a bias field directed so as to tend to collapse the domains and having a magnitude between 0 and $4\pi M_s$. The uniform radial collapse of the domain and the run-out of the domain into an initially elliptical shape bound in the region of stability. The minimum attainable domain diameter in a given material is $d_{\min} \approx 4l$ occurring a plate thickness of $\sim 4l$. It is estimated that the cylindrical wall approximation begins to become doubtful at a plate thickness greater than $4l$. In order for cylindrical domains to exist, $H_a \approx 4\pi M_s$ and in general approximations such as the approximation of zero wall width become more accurate for $H_a \gg 4\pi M_s (d_{\min}/l_w = 8H_a/4\pi^2 M_s$ where l_w is the wall width).

It is interesting to note that since stable cylindrical domains of a definite size exist in the total absence of wall motion coercivity and may be freely moved, they form a relative, easily observable, classical model for illustrating several particle-field concepts. They may be considered a two-dimensional particle which is produced as a singularity of finite extent in an underlying three-dimensional field (the magnetization). Cylindrical domains are particularly useful for demonstrating the concept of identical particles since, while it is possible to put identifying marks on domain locations, it is not possible to mark individual domains. (Cylindrical domains do exist in two species which may be distinguished by the direction of rotation of the spins in the domain wall.[13] All attempts to observe this difference up to the present time have been unsuccessful.)

VIII. ACKNOWLEDGMENTS

The author would like to acknowledge U. F. Gianola, H. E. D. Scovil, and W. Shockley whose original interest in this subject motivated much

of this work; R. D. Pierce for checking the calculation of the energy derivatives; P. I. Bonyhard who showed that the generalized force equation had a convenient graphical solution, and W. J. Tabor and F. B. Hagedorn for reading the entire manuscript and offering suggestions. Special thanks must to go E. Della Torre for assistance in organizing, and checking calculations through many revisions. I am also deeply indebted to A. H. Bobeck whose experiments form the motivational basis of the present work and who has provided many fresh insights in the course of the work.

APPENDIX A

Integrals of Cylindrical Domain Theory

This appendix contains the reduction to standard form of the elliptic integrals which arise in the theory of cylindrical domains in plates of infinite extent and power series expansions of these integrals. All the properties of the functions obtained here are used in either the physical interpretation of the energy variation expansion or in generating the numerical values of the force and stability functions.

It is convenient to define functions U and V which appear repeatedly in cylindrical domain theory. The elliptic integrals which appear in the final results of the theory appear only in the forms U and V, U being a function of only the complete elliptic integral of the second kind and V being only a function of the complete elliptic integral of the first kind. Because of the form of the U and V functions, it has proven easier to obtain the needed properties, (such as the series expansions) directly from the integral definitions rather than deducing them from the tabulated properties of elliptic integrals.

The latter half of this appendix treats the properties of the L_n functions. A recursion relation is obtained and used to reduce the L_n to functions of U and V. Power series expansions of the L_n are obtained directly from the definition (34).

A.1 *Definition of the U and V Functions*

The functions are defined in the alternate forms

$$U(x) \equiv \int_0^{\pi} [x + \tfrac{1}{2}(1 - \cos\alpha)]^{\frac{1}{2}}\, d\alpha \tag{84a}$$

$$= 2\int_0^{\pi/2} (x + \sin^2\beta)^{\frac{1}{2}}\, d\beta \tag{84b}$$

$$= 2\int_0^{\pi/2} (x + 1 - \sin^2\gamma)^{\frac{1}{2}}\, d\gamma \tag{84c}$$

$$= 2(x + 1)^{\frac{1}{2}} E\left(\frac{1}{1 + x}\right) \tag{84d}$$

and

$$V(x) \equiv \int_0^{\pi} [x + \tfrac{1}{2}(1 - \cos\alpha)]^{-\frac{1}{2}}\, d\alpha \tag{85a}$$

$$= 2\int_0^{\pi/2} (x + \sin^2\beta)^{-\frac{1}{2}}\, d\beta \tag{85b}$$

$$= 2\int_0^{\pi/2} (x + 1 - \sin^2\gamma)^{-\frac{1}{2}}\, d\gamma \tag{85c}$$

$$= 2(x + 1)^{-\frac{1}{2}} K\left(\frac{1}{1 + x}\right) \tag{85d}$$

where the dummy variables are related by $\beta = \alpha/2$ and $\gamma = \pi/2 - \alpha/2$ and where K and E are the complete elliptic integrals of the first and second kind respectively. The argument of the elliptic integrals is the parameter m of Abramowitz and Stegun.[14] The parameter m is equal to the parameter k^2 of Jahnke and Emde or Groebner and Hofreiter.[15,16]

A.2 *Differential Equations and the Power Series Expansion of U and V*

From the definitions (84) and (85)

$$\frac{dU}{dx} = \tfrac{1}{2}V. \tag{86}$$

The differential equations obeyed by U and V are

$$\left[(x^2 + x)\frac{d^2}{dx^2} + \frac{1}{4}\right]U(x) = 0 \tag{87a}$$

and

$$\left[(x^2 + x)\frac{d^2}{dx^2} + (2x + 1)\frac{d}{dx} + \frac{1}{4}\right]V(x) = 0. \tag{87b}$$

The U differential equation is verified by substituting in the defining relation (84a) and then reducing the resulting equation to the indentity

$$0 = \int_0^{\pi} \frac{d}{d\alpha}\frac{\sin\alpha}{[x + \frac{1}{2}(1 - \cos\alpha)]^{\frac{1}{2}}}\, d\alpha \tag{88a}$$

$$= \int_0^{\pi} \frac{x\cos\alpha + \frac{1}{2}\cos\alpha - \frac{1}{4} - \frac{1}{4}\cos^2\alpha}{[x + \frac{1}{2}(1 - \cos\alpha)]^{\frac{3}{2}}}\, d\alpha. \tag{88b}$$

The V equation is then easily obtained by differentiation of the U equation.

The roots of the indicial equations of these equations are separated by 1 [they are 0 and 1 in equation (87a) and -1 and 0 in equation (87b)] so that the series expansion of U is of the form

$$U(x) = \sum_{i=0}^{\infty} U_i x^i \quad \text{where} \quad U_i = U_i' + U_i'' \tfrac{1}{2} \ln \left| \frac{16}{x} \right| \qquad (89a, b)$$

and

$$V(x) = \sum_{i=0}^{\infty} V_i x^i \quad \text{where} \quad V_i = V_i' + V_i'' \tfrac{1}{2} \ln \left| \frac{16}{x} \right|. \qquad (90a, b)$$

The form of the logarithmic terms has been chosen with some foresight. Substitution of the expansions into the differential equations and comparing coefficients gives the recursion relations

$$U_{i+1}'' = -\frac{(j - \frac{1}{2})^2}{j(j+1)} U_i'', \qquad j \geqq 1, \qquad (91a)$$

$$U_{i+1}' = -\frac{j - \frac{1}{2}}{j(j+1)} \left[(j - \tfrac{1}{2}) U_i' - \frac{(j + \frac{1}{4})}{j(j+1)} U_i'' \right], \qquad j \geqq 1, \qquad (91b)$$

$$V_{i+1}'' = -\left(\frac{j + \frac{1}{2}}{j+1} \right)^2 V_i'', \qquad j \geqq 0, \qquad (92a)$$

$$V_{i+1}' = -\frac{j + \frac{1}{2}}{(j+1)^2} \left[\left(j + \frac{1}{2} \right) V_i' - \frac{1}{2} \frac{1}{j+1} V_i'' \right], \qquad j \geqq 0. \qquad (92b)$$

The starting values of V_i', V_i'', U_i' and U_i'' are determined directly from the integral definitions of the functions (84) and (85) and the differential equation (86) relating U and V. This is quite straightforward except for V which must be expanded

$$V(x) = 2 \int_0^{\epsilon} \frac{d\beta}{(x + \beta^2)^{\frac{1}{2}}} + 2 \int_{\epsilon}^{\pi/2} \frac{d\beta}{\sin \beta} + O(x^2, \epsilon^2, \tfrac{x}{\epsilon^2}) \qquad (93a)$$

and evaluated as a limit

$$\lim_{x \to 0} V(x) = \ln \left| \frac{16}{x} \right|. \qquad (93b)$$

The limiting value of V may also be obtained quite easily from equation (85d) and the tabulated properties of the complete elliptic integral of the first kind.[15] The expansions of U and V are thus

$$U(x) = \left(2 + \frac{1}{2} x + \frac{3}{32} x^2 - \frac{3}{64} x^3 + \frac{665}{24576} x^4 + \cdots \right)$$

$$+\left(0+x-\frac{1}{8}x^2+\frac{3}{64}x^3-\frac{25}{1024}x^4+\cdots\right)\frac{1}{2}\ln\left|\frac{16}{x}\right| \quad (94)$$

and

$$V(x)=\left(0+\frac{1}{2}x-\frac{21}{64}x^2+\frac{185}{768}x^3+\cdots\right)$$
$$+\left(2-\frac{1}{2}x+\frac{9}{32}x^2-\frac{25}{128}x^3+\cdots\right)\frac{1}{2}\ln\left|\frac{16}{x}\right|. \quad (95)$$

A.3 *Expansion of U and V in Terms of Inverse Powers*

Making a Taylor series expansion of equations (84b) and (85b) respectively and integrating yields the expansions of U and V in terms of the inverse powers of the argument

$$U(x)=2x^{\frac{1}{2}}\int_0^{\pi/2}(1+x^{-1}\sin^2\beta)^{\frac{1}{2}}\,d\beta$$
$$=\sum_{j=0}^{\infty}\frac{-2}{2j-1}\frac{(2j)!}{(-4)^j(j!)^2}x^{-(j-\frac{1}{2})}\int_0^{\pi/2}\sin^{2j}\beta\,d\beta$$
$$=\pi\sum_{j=0}^{\infty}\frac{-1}{(2j-1)(-16)^j}\left[\frac{(2j)!}{(j!)^2}\right]^2x^{-(j-\frac{1}{2})}. \quad (96a)$$
$$=\pi x^{\frac{1}{2}}\left[1+\frac{1}{4}x^{-1}-\frac{3}{64}x^{-2}+\frac{5}{256}x^{-3}+\cdots\right] \quad (96b)$$

and

$$V(x)=2x^{-\frac{1}{2}}\int_0^{\pi/2}(1+x^{-1}\sin^2\beta)^{-\frac{1}{2}}\,d\beta$$
$$=\sum_{j=0}^{\infty}2\frac{(2j)!}{(-4)^j(j!)^2}x^{-(j+\frac{1}{2})}\int_0^{\pi/2}\sin^{2j}\beta\,d\beta$$
$$=\pi\sum_{j=0}^{\infty}\frac{1}{(-16)^j}\left[\frac{(2j)!}{(j!)^2}\right]^2x^{-(j+\frac{1}{2})}. \quad (97a)$$
$$=\pi x^{-\frac{1}{2}}\left[1-\frac{1}{4}x^{-1}+\frac{9}{64}x^{-2}-\frac{25}{256}x^{-3}+\cdots\right]. \quad (97b)$$

A.4 *Definition of the L_n Functions*

The L_n functions are defined in expression (34) by

$$L_n(x)\equiv\int_0^{\pi}\frac{(1-\cos n\alpha)\,d\alpha}{[x+\frac{1}{2}(1-\cos\alpha)]^{\frac{1}{2}}},\qquad x\geqq 0,\ n\geqq 0 \quad (98a)$$

or with the change of variable $\beta = \alpha/2$

$$L_n(x) = 4 \int_0^{\pi/2} \frac{\sin^2 n\beta}{(x + \sin^2 \beta)^{\frac{1}{2}}} \, d\beta, \qquad x \geq 0, n \geq 0. \tag{98b}$$

It can be seen directly from equation (98b) that for a fixed value of n

$$\frac{dL_n(x)}{dx} < 0, \qquad L_n(\infty) = 0, \qquad \text{and} \quad L_0(x) = 0. \tag{99a, b, c}$$

From definitions (84b), (85b), and (98b)

$$L_1(x) = 2[U(x) - xV(x)]. \tag{100}$$

The higher L functions are determined by means of a recursion relation.

A.5 *The* L_n *Recursion Relation*

The L_n recursion relation is

$$L_{n+1}(x) = \frac{1}{2n+1} [4n(2x+1)L_n(x) - (2n-1)L_{n-1}(x) - 8nxV(x)], \qquad n \geqq 1. \tag{101}$$

The recursion relation is verified by substituting in the definitions of L_n and V, equations (98a) and (85a), and reducing the resulting equation to the identity

$$0 = \int_0^{\pi} \frac{d}{d\alpha} \{\sin n\alpha[x + \tfrac{1}{2}(1 - \cos\alpha)]^{\frac{1}{2}}\} \, d\alpha \tag{102a}$$

$$\int_0^{\pi} \frac{n \cos n\alpha[x + \tfrac{1}{2}(1 - \cos\alpha)] + \tfrac{1}{4} \sin n\alpha \sin\alpha}{[x + \tfrac{1}{2}(1 - \cos\alpha)]^{\frac{1}{2}}} \, d\alpha. \tag{102b}$$

The initial functions $L_0(x)$ and $L_1(x)$ are given by equations (99c) and (100). Note that for large values of x the recursion relation is unstable for increasing n.

A.6 *Power Series Expansion of the* L_n *Function*

The function $L_0(x)$ is identically zero, equation (99c). The series expansion for $L_1(x)$, obtained from equations (100), (94), and (95), is

$$L_1(x) = \left(4 + x - \frac{13}{16}x^2 + \frac{9}{16}x^3 - \frac{5255}{12288}x^4 + \cdots\right) + \left(0 - 2x + \frac{3}{4}x^2 - \frac{15}{32}x^3 + \frac{175}{512}x^4 + \cdots\right)\frac{1}{2}\ln\left|\frac{16}{x}\right|. \tag{103}$$

From the recursion relation (101) and the initial functions (99c) and

(100) the general form of $L_n(x)$ is

$$L_n(x) = u_n(x)U(x) + v_n(x)V(x) \tag{104}$$

where $u_n(x)$ and $v_n(x)$ are polynominals of order n or less in x. Because of the form of $U(x)$ and $V(x)$, expressions (89) and (90), an expansion of the form

$$L_n(x) = \sum_{j=0}^{\infty} L_{n,j}x^j \tag{105a}$$

where

$$L_{n,j} = L'_{n,j} + L''_{n,j}\,\tfrac{1}{2}\ln\left|\frac{16}{x}\right| \tag{105b}$$

clearly exists.

Expressions for either the coefficients in the polynominals $u_n(x)$ and $v_n(x)$ or the $L_{n,j}$ may be determined in closed form by similar methods. It has, however, proven more useful to use the recursion relation directly when the complete expression of the form of expression (104) is desired and the expansion (105) when a power series is desired.

To obtain the $L_{n,j}$ the expansion (105) is substituted in the recursion relation (101) and coefficients of x are compared to obtain a hierarchy, in j, of recursion relations, each member of the hierarchy being factorable and depending only on the preceding member. These recursion relations are then factored and successively summed.

The coefficient of x^j is

$$L_{n+1,j} = \frac{1}{2n+1}\{4nL_{n,j} - (2n-1)L_{n-1,j} + 8n[L_{n,j-1} - V_{j-1}]\}, \qquad n \geqq 0 \tag{106a}$$

where

$$L_{n,-1} = 0 \quad \text{and} \quad V_{-1} = 0. \tag{106b, c}$$

With the definition

$$Q_{n,j} \equiv (2n-1)[L_{n,j} - L_{n-1,j}] \tag{107}$$

the second order recursion relation (106) factors into two first order recursion relations

$$Q_{n+1,j} = Q_{n,j} + 8n(L_{n,j-1} - V_{j-1}), \qquad n \geqq 0 \tag{108a}$$

and

$$L_{n+1,j} = L_{n,j} + \frac{1}{2n+1} Q_{n+1,j}, \qquad n \geqq 1. \tag{108b}$$

The recursion relations for $j = 0$ and $j = 1$ will now be summed. From expressions (99c) and (105a)

$$L_{0,j} = 0 \tag{109}$$

so that using expression (107)

$$Q_{1,j} = L_{1,j}. \tag{110}$$

For $j = 0$ using expressions (106b) and (106c), the recursion relation (108b) becomes simply

$$Q_{n+1,0} = Q_{n,0}. \tag{111}$$

By inspection of expression (103) the initial value of $Q_{n,0}$ is

$$Q_{1,0} = L_{1,0} = 4 \tag{112}$$

so that from expression (111)

$$Q_{n,0} = 4, \qquad n \geqq 1. \tag{113}$$

The recursion relation (108b) thus becomes

$$L_{n,0} = L_{n-1,0} + \frac{4}{2n-1}, \qquad n \geqq 0 \tag{114}$$

which with the initial value of expression (109) may be summed to yield

$$L_{n,0} = \begin{cases} 0, & n = 0, \qquad (115a) \\ 4 \sum_{j=1}^{n} \frac{1}{2j-1}, & n > 0. \qquad (115b) \end{cases}$$

From the form of the expansion (105) it can be seen that

$$L_n(0) = L_{n,0}, \tag{116}$$

so that with (99a, b)

$$-4 \sum_{j=1}^{n} \frac{1}{2j-1} \leqq L_n(x) - L_n(0) \leqq 0, \qquad n \geqq 1. \tag{117}$$

For evaluating the $Q_{n,1}$ and $L_{n,1}$ sums, two relations are needed:

$$\sum_{j=1}^{n} \sum_{k=1}^{j} \frac{1}{2k-1} = \sum_{k=1}^{n} \sum_{j=k}^{n} \frac{1}{2k-1}$$

$$= \sum_{k=1}^{n} \frac{n-k+1}{2k-1}$$

$$= \tfrac{1}{2}(2n+1)\left(\sum_{k=1}^{n}\frac{1}{2k-1}\right) - \frac{n}{2} \tag{118}$$

and similarly

$$\sum_{j=1}^{n} j \sum_{k=1}^{j}\frac{1}{2k-1} = \frac{1}{8}(2n+1)^2 \sum_{k=1}^{n}\frac{1}{2k-1} - \frac{1}{8}n^2. \tag{119}$$

Evaluation of the sums of expressions (118) and (119) is analogous to integrating $x^n \log x$ where $n = 0$ and 1. With sufficient patience the sums can clearly be carried out for any finite n.

For $j = 1$, expression (108a) becomes [using expressions (95), (109), and (115)]

$$Q_{n+1,1} = Q_{n,1} + 8n\left[4\sum_{j=1}^{n}\frac{1}{2j-1} - \ln\left|\frac{16}{x}\right|\right], \qquad n \geqq 1 \tag{120}$$

where the initial function is [using expressions (103), (105), and (110)]

$$Q_{1,1} = L_{1,1} = 1 - \ln\left|\frac{16}{x}\right|. \tag{121}$$

Summing [using expression (119)] yields

$$\begin{aligned} Q_{n+1,1} &= 1 - \ln\left|\frac{16}{x}\right| + \sum_{k=1}^{n} 8k\left[4\sum_{j=1}^{k}\frac{1}{2j-1} - \ln\left|\frac{16}{x}\right|\right] \\ &= -(2n+1)^2 \ln\left|\frac{16}{x}\right| + 4(2n+1)^2 \sum_{k=1}^{n}\frac{1}{2k-1} - (4n^2-1), \\ &\qquad n \geqq 1. \end{aligned} \tag{122}$$

The $L_{n,1}$ recursion relation is then

$$L_{n+1,1} = L_{n,1} + (2n+1)\left[-\ln\left|\frac{16}{x}\right| + 4\sum_{k=1}^{n}\frac{1}{2k-1}\right] - 2n + 1, \qquad n \geqq 1 \tag{123}$$

which may be summed using the initial value of expression (121) and the sums (118) and (119):

$$\begin{aligned} L_{n+1,1} &= 1 - \ln\left|\frac{16}{x}\right|\left[1 + \sum_{k=1}^{n}(2k+1)\right] \\ &\quad + \sum_{k=1}^{n} 4(2k+1)\sum_{j=1}^{k}\frac{1}{2j-1} + \sum_{k=1}^{n}(-2k+1) \end{aligned} \tag{124}$$

$$L_{n,1} = \begin{cases} 0, & n = 0. \qquad (125a) \\ -2n^2\left(\frac{1}{2}\ln\left|\frac{16}{x}\right|\right) - 2n^2 + (4n^2 - 1)\sum_{k=1}^{n}\frac{1}{2k-1}, & \\ & n \geqq 1. \qquad (125b) \end{cases}$$

It is clear that the procedure leading to expressions (115) and (125) may be carried onward to lead to an expansion of the form

$$L_n(x) - L_n(0) = +\sum_{j=1}^{\infty}\left[\sum_{k=1}^{j}\left(L^{(1)}(j,k)\,\tfrac{1}{2}\ln\left|\frac{16}{x}\right| + L^{(2)}(j,k) + L^{(3)}(j,k)\sum_{m=1}^{k}\frac{1}{2m-1}\right)n^{2k}\right]x^j \qquad (126)$$

where the $L^{(p)}(j, k)$ are functions of j and k only. It is clear that in the limit $x \to 0$, $n \to \infty$ an expansion in terms of

$$\left(\frac{\lambda}{h}\right)^2 = \frac{\pi^2}{xn^2} \qquad (127)$$

may be made where λ/h is introduced as the finite expansion parameter. Replacing the sum in expression (125) by its approximating integral yields

$$\lim_{x\to 0}\left[L_n(x) - L_n(0)\right] = -2\pi^2\frac{h^2}{\lambda^2}\ln\left|\frac{4e\lambda}{\pi h}\right| + O_4\left(\frac{h}{\lambda}\right) \qquad (128a)$$

where

$$n = \pi h x^{-\frac{1}{2}}\lambda^{-1}. \qquad (128b)$$

A.7 *Expansions of* L *in Terms of Inverse Powers*

The expansion of L in terms of inverse powers of the argument is obtained by Taylor expansion of expression (98b) and integrating. This yields

$$L_n(x) = 4x^{-\frac{1}{2}}\int_0^{\pi/2}\frac{\sin^2 n\beta}{(1 + x^{-1}\sin^2\beta)^{\frac{1}{2}}}\,d\beta$$

$$= 4\sum_{j=0}^{\infty}\frac{(2j)!}{(-4)^j(j!)^2}\,x^{-(j+\frac{1}{2})}\int_0^{\pi/2}\sin^2 n\beta\,\sin^{2j}\beta\,d\beta, \qquad n \geqq 1 \qquad (129a)$$

where for example

$$L_1(x) = \pi\sum_{j=0}^{\infty}\frac{(2j+1)}{(j+1)}\frac{1}{(-16)^j}\left[\frac{(2j)!}{(j!)^2}\right]^2 x^{-(j+\frac{1}{2})}. \qquad (129b)$$

By considering the Fourier decomposition of $\sin^j \beta$ it can be seen that

$$\int_0^{\pi/2} \sin^2 n\beta \sin^{2j} \beta \, d\beta = \frac{1}{2} \int_0^{\pi/2} \sin^{2j} \beta \, d\beta, \qquad n > j$$

$$= \frac{\pi}{4} \frac{(2j)!}{4^j (j!)^2}, \qquad n > j \tag{130}$$

(independent of n) so that

$$L_0(x) = 0,$$

$$L_n(x) = \pi \sum_{j=0}^{n-1} \frac{1}{(-16)^j} \left[\frac{(2j)!}{(j!)^2}\right]^2 x^{-(j+\frac{1}{2})} + O[x^{-(n+\frac{1}{2})}]), \qquad n > 1 \tag{131a}$$

or identifying with expression (97)

$$L_n(x) = V(x) + O[x^{-(n+\frac{1}{2})}], \qquad n > 1. \tag{131b}$$

Evaluating expression (129a) directly in those cases in which expressions (129b) or (131a) cannot be used yields

$$L_0(x) = 0, \tag{132a}$$

$$L_1(x) = \pi x^{-1/2} - \frac{3\pi}{8} x^{-3/2} + \frac{15\pi}{64} x^{-5/2} + O(x^{-7/2}) \tag{132b}$$

$$L_2(x) = \pi x^{-1/2} - \frac{\pi}{4} x^{-3/2} + \frac{15\pi}{128} x^{-5/2} + O(x^{-7/2}), \tag{132c}$$

$$L_n(x) = \pi x^{-1/2} - \frac{\pi}{4} x^{-3/2} + \frac{9\pi}{64} x^{-5/2} + O(x^{-7/2}). \tag{132d}$$

A.8 *The Gaussian Transformation*

In the neighborhood of $x = 1$, the convergence of the power series in x or x^{-1} is rather slow. Either the gaussian or Landen transformations may be used to transform the U, V, or L_1 functions into a region of rapid convergence.[16] In the present case, the gaussian transformation is preferred since it does not introduce incomplete elliptic integrals as does the Landen transformation.

The result of the gaussian transformation is

$$x_1 = 4x^{\frac{1}{2}}(1 + x)^{\frac{1}{2}}[(1 + x)^{\frac{1}{2}} + x^{\frac{1}{2}}]^2, \tag{133a}$$

or inversely

$$x = \frac{x_1^2}{4(1 + x_1)^{\frac{1}{2}}[(1 + x_1)^{\frac{1}{2}} + 1]^2} \tag{133b}$$

for the argument and

$$V(x) = 2TV(x_1), \tag{134}$$

$$TU(x) = U(x_1) - \frac{x_1}{2} V(x_1), \tag{135}$$

$$TL_1(x) = L_1(x_1) + \frac{2x_1}{1 + T^2} V(x_1), \tag{136}$$

for the functions where

$$T \equiv (1 + x_1)^{\frac{1}{2}} = (1 + x)^{\frac{1}{2}} + x^{\frac{1}{2}}. \tag{137}$$

APPENDIX B

The Force and Stability Functions

This appendix is a compilation of expressions for the force F and stability S_n functions. Each of the functions is written in terms of the U and V or L_n functions of Appendix A. Expressions in terms of the complete elliptic integrals of the first and second kind (denoted by K and E respectively) permit the use of tables[14] or numerical computation using the Landen transformation or the gaussian transformation.[16] The gaussian transformation is used in Section A.8. The power series expansions provided are necessary in obtaining numerical values of the functions for either very large or very small values of the argument and also provide the asymptotic forms of the functions. The argument of the functions is the domain diameter to thickness ratio, $d/h = 2r_0/h$.

B.1 *The Force Function*

The force function is written in terms of U by comparing the form of F, expression (40), and the form of U, expression (84b),

$$F\left(\frac{d}{h}\right) = \frac{1}{\pi}\left(\frac{d}{h}\right)^2\left[U\left(\frac{h^2}{d^2}\right) - 2\right]. \tag{138a}$$

This expression is written in terms of the complete elliptic integral of the second kind using expression (84d)

$$F\left(\frac{d}{h}\right) = \frac{2}{\pi}\left(\frac{d}{h}\right)^2\left[\left(1 + \frac{h^2}{d^2}\right)^{\frac{1}{2}} E[(1 + h^2/d^2)^{-1}] - 1\right]. \tag{138b}$$

It is expanded about $h/d = 0$ using expression (94)

$$F\left(\frac{d}{h}\right) = \frac{1}{\pi}\left\{\left[\frac{1}{2} + \frac{3}{32}\left(\frac{h}{d}\right)^2 - \frac{3}{64}\left(\frac{h}{d}\right)^4 + \frac{665}{24576}\left(\frac{h}{d}\right)^6 + \cdots\right]\right.$$

$$+\left[1-\frac{1}{8}\left(\frac{h}{d}\right)^2+\frac{3}{64}\left(\frac{h}{d}\right)^4-\frac{25}{1024}\left(\frac{h}{d}\right)^6+\cdots\right]\ln\left|4\,\frac{d}{h}\right|\Bigg\}. \tag{138c}$$

Additional terms may be generated using expressions (89) and (91). It is expanded about $d/h = 0$ using expression (96b)

$$F\left(\frac{d}{h}\right)=\frac{d}{h}-\frac{2}{\pi}\left(\frac{d}{h}\right)^2+\frac{1}{4}\left(\frac{d}{h}\right)^3-\frac{31}{64}\left(\frac{d}{h}\right)^5+\frac{5}{256}\left(\frac{d}{h}\right)^7+\cdots. \tag{138d}$$

Additional terms may be generated using expression (96a).

B.2 *The Radial Stability Function*

The radial stability function is written in terms of U and V using the definition of the radial stability function [equation (66a)], the expression for F [equation (138a)], and comparing the derivative of F [equation (41)] with the form of V [equation (85b)],

$$S_0\left(\frac{d}{h}\right)=-\frac{1}{\pi}\left(\frac{d}{h}\right)^2\left[U\left(\frac{h^2}{d^2}\right)-\left(\frac{h}{d}\right)^2V\left(\frac{h^2}{d^2}\right)-2\right], \tag{139a}$$

or in terms of L_1 using the expression for L_1 [equation (100)] and the expressions for $L_1(0)$ [equations (115) and (116)]

$$S_0\left(\frac{d}{h}\right)=-\frac{1}{2\pi}\left(\frac{d}{h}\right)^2\left[L_1\left(\frac{h^2}{d^2}\right)-L_1(0)\right]. \tag{139b}$$

Expression (139a) is written in terms of the complete elliptic integrals of the first and second kind using expressions (84d) and (85d),

$$S_0\left(\frac{d}{h}\right)=-\frac{2}{\pi}\left(\frac{d}{h}\right)^2\left[\left(1+\frac{h^2}{d^2}\right)^{\frac{1}{2}}E[(1+h^2/d^2)^{-1}]\right.$$
$$\left.-\left(\frac{h}{d}\right)^2\left(1+\frac{h^2}{d^2}\right)^{-\frac{1}{2}}K[(1+h^2/d^2)^{-1}]-1\right]. \tag{139c}$$

The expansion about $h/d = 0$ is obtained using expressions (139b) and (103).

$$S_0\left(\frac{d}{h}\right)=\frac{1}{\pi}\left\{\left[-\frac{1}{2}+\frac{13}{32}\left(\frac{h}{d}\right)^2-\frac{9}{32}\left(\frac{h}{d}\right)^4+\frac{5255}{24576}\left(\frac{h}{d}\right)^6+\cdots\right]\right.$$
$$\left.+\left[1-\frac{3}{8}\left(\frac{h}{d}\right)^2+\frac{15}{64}\left(\frac{h}{d}\right)^4-\frac{175}{1024}\left(\frac{h}{d}\right)^6+\cdots\right]\ln\left|\frac{4d}{h}\right|\right\}. \tag{139d}$$

Additional terms may be generated using expressions (139a) and (89) through (92). The expansion about $d/h = 0$ is obtained using expressions (139b), (132b), and (103) to obtain L_1 (0)

$$S_0\left(\frac{d}{h}\right) = \frac{2}{\pi}\left(\frac{d}{h}\right)^2 - \frac{1}{2}\left(\frac{d}{h}\right)^3 + \frac{3}{16}\left(\frac{d}{h}\right)^5 - \frac{15}{128}\left(\frac{d}{h}\right)^7 + \cdots . \tag{139e}$$

Additional terms may be generated using expression (129b).

B.3 *The Elliptical Stability Function*

From the general definition of the S_n of expression (66b) the elliptical stability function is

$$S_2\left(\frac{d}{h}\right) = -\frac{1}{3}\left[S_0\left(\frac{d}{h}\right) + \frac{1}{2\pi}\left(\frac{d}{h}\right)^2\left\{L_2\left(\frac{h^2}{d^2}\right) - L_2(0)\right\}\right]. \tag{140a}$$

Using expression (139b) for S_0, the L_n recursion relation (101) to reduce L_2 to L_1, and V and (103) to obtain $L_1(0)$, S_2 is written in terms of L_1 and V:

$$S_2\left(\frac{d}{h}\right) = \frac{1}{18\pi}\left(\frac{d}{h}\right)^2\left\{4 - \left[1 + 8\left(\frac{h}{d}\right)^2\right]L_1\left(\frac{h^2}{d^2}\right) + 8\left(\frac{h}{d}\right)^2 V\left(\frac{h^2}{d^2}\right)\right\}. \tag{140b}$$

The function L_1 is then eliminated using expression (100) to obtain the expression in terms of U and V:

$$S_2\left(\frac{d}{h}\right) = \frac{1}{9\pi}\left(\frac{d}{h}\right)^2\left\{2 - \left[1 + 8\left(\frac{h}{d}\right)^2\right]U\left(\frac{h^2}{d^2}\right) + \left[5\left(\frac{h}{d}\right)^2 + 8\left(\frac{h}{d}\right)^4\right]V\left(\frac{h^2}{d^2}\right)\right\} \tag{140c}$$

which then is written in terms of the complete elliptic integrals of the first and second kind using expression (84d) and (85d):

$$S_2\left(\frac{d}{h}\right) = \frac{1}{9\pi}\left(\frac{d}{h}\right)^2\left\{2 - \left[2 + 16\left(\frac{h}{d}\right)^2\right]\left(1 + \frac{h^2}{d^2}\right)^{+\frac{1}{2}} E[(1 + h^2/d^2)^{-1}]\right.$$
$$\left. + \left[10\left(\frac{h}{d}\right)^2 + 16\left(\frac{h}{d}\right)^4\right]\left(1 + \frac{h^2}{d^2}\right)^{-\frac{1}{2}} K[(1 + h^2/d^2)^{-1}]\right. \tag{140d}$$

The expansion about $h/d = 0$ is obtained using expressions (140b), (95), and (103):

$$S_2\left(\frac{d}{h}\right) = \frac{1}{\pi}\left\{\left[-\frac{11}{6} - \frac{17}{96}\left(\frac{h}{d}\right)^2 + \frac{53}{288}\left(\frac{h}{d}\right)^4 - \frac{2929}{24576}\left(\frac{h}{d}\right)^6 + \cdots\right]\right.$$
$$\left. + \left[1 + \frac{5}{8}\left(\frac{h}{d}\right)^2 - \frac{35}{192}\left(\frac{h}{d}\right)^4 + \frac{105}{1024}\left(\frac{h}{d}\right)^6 + \cdots\right]\ln\left|\frac{4d}{h}\right|\right\}. \tag{140e}$$

Additional terms may be generated using expressions (140c) and (89) through (92). The expansion about $d/h = 0$ is obtained using expressions (140b), (97b), and (132b):

$$S_2\left(\frac{d}{h}\right) = \frac{2}{9\pi}\left(\frac{d}{h}\right)^2 - \frac{1}{48}\left(\frac{d}{h}\right)^5 + \frac{5}{256}\left(\frac{d}{h}\right)^7 + \cdots . \qquad (140f)$$

Additional terms may be generated using expressions (140b), (97a), and (129b).

B.4 *The General Stability Functions*

Using the definition (66b) and the expression for S_0 [equation (139b)], the S_n are written in terms of the L_n as

$$S_n\left(\frac{d}{h}\right) = -\frac{1}{n^2-1}\frac{1}{2\pi}\left(\frac{d}{h}\right)^2\left[L_n\left(\frac{h^2}{d^2}\right) - L_1\left(\frac{h^2}{d^2}\right) - L_n(0) + L_1(0)\right],$$

$$n \geqq 2. \qquad (141a)$$

The leading term of the expansion about $h/d = 0$ is obtained using expressions (103), (105), (115), (116), (125), and (126):

$$S_n\left(\frac{d}{h}\right) = \frac{1}{\pi}\left[\ln\left|\frac{4d}{h}\right| - \frac{4n^2-1}{2(n^2-1)}\sum_{j=1}^{n}\frac{j}{2j-1} + \frac{2n^2+1}{2(n^2-1)}\right] + O_2\left(\frac{nh}{d}\right),$$

$$n \geqq 2. \qquad (141b)$$

The expansion about $d/h = 0$ is obtained using expressions (115), (116), (129b), and (131a):

$$S_n\left(\frac{d}{h}\right) = \frac{1}{n^2-1}\left\{\frac{2}{\pi}\left(\frac{d}{h}\right)^2\sum_{j=2}^{n}\frac{1}{2j-1} + \frac{1}{2}\sum_{j=1}^{n-1}\frac{j}{j+1}\frac{1}{(-16)^j}\right.$$
$$\left.\cdot\left[\frac{(2j)!}{(j!)^2}\right]^2\left(\frac{d}{h}\right)^{2j+3} + O_{2n+3}\left(\frac{d}{h}\right)\right\}, \qquad n \geqq 2. \qquad (141c)$$

APPENDIX C

Symbol List

Numbers in parentheses are defining equations or figures.

A	exchange constant (71)
a	area
d	mean domain diameter, $2r_0$ (42)
$E(x)$	complete elliptic integral of the second kind ($x = m = k^2$)
E_H	energy due to applied field (9)
E_M	internal magnetostatic energy (10)

E_T total energy (7)
E_W total wall energy (8)
$F(x)$ generalized radial force function (33, 138, Fig. 3)
$\mathbf{H}$ magnetic field vector
H uniform applied field
H_a anisotropy field
H_N nucleation field (74)
$\langle H_z \rangle_{av}$ z-averaged z-component of magnetic field (45)
h plate thickness (Figs. 1, 4)
$K(x)$ complete elliptic integral of the first kind ($x = m = k^2$)
K_u uniaxial anisotropy constant (71)
$L_n(x)$ integrally defined function (34, 98)
l characteristic length, $\sigma_w/4\pi M_s^2$ (67)
l_w wall width (81)
$\mathbf{M}$ magnetization vector
M_s saturation magnetization
n rotational periodicity (1)
O_k terms of order k
q quality factor, $K_u/2\pi M_s^2 = H_a/4\pi M_s$ (72)
r cylindrical coordinate (Figs. 1, 2)
r_f plate radius (Fig. 3)
r_n nth radial Fourier amplitude (1)
r_0 mean domain radius (1)
$S_n(x)$ nth infinite plate stability function (66, 139, 140, 141)
s distance between interacting magnetic charges (10b, 24, 28)
$U(x)$ integrally defined function (84)
$u(x)$ unit step function (5)
V volume
$V(x)$ integrally defined function (85)
z cylindrical coordinate (Figs. 1, 4, 5)
Z operator (23)
z $z - z'$ (21)
$\triangle E$ variation in energy (11, 43)
$\triangle r_n$ variation in r_n (3)
$\triangle \theta_n$ variation in θ_n (3)
$\delta(x)$ dirac delta function
ζ $\theta' - \theta$ (27, Fig. 2)
η polar azimuthal angle (Fig. 5)
η_w polar angle of wall normal (Fig. 5)
θ cylindrical coordinate (Fig. 1)
θ_n nth Fourier phase angle (1)

λ_n wavelength of nth variation (54)
ν azimuthal angle (Fig. 5)
ν_w azimuthal angle of wall normal (Fig. 5)
ξ_0 wall displacement vector (Fig. 5)
ρ coordinate in displaced cylindrical coordinate system (35, Fig. 2)
σ_w wall energy density
φ coordinate in displaced cylindrical coordinate system (35, Fig. 2)
χ_t transverse susceptibility
Ω magnetostatic potential

REFERENCES

1. Michaelis, P. C., "A New Method of Propagating Domains in Thin Ferromagnetic Films," J. Appl. Phys., *39,* No. 2, Part 2 (February 1968), pp. 1224–1226.
2. Bobeck, A. H., "Properties and Device Applications of Magnetic Domains in Orthoferrites," B.S.T.J., *46,* No. 8 (October 1967), pp. 1901–1925.
3. Sherwood, R. C., Remeika, J. P., and Williams, H. J., "Domain Behavior in Some Transparent Magnetic Oxides", J. Appl. Phys., *30,* No. 2 (February 1959), pp. 217–225.
4. Scovil, H. E. D., unpublished work.
5. Kooy, C., and Enz, U., "Experimental and Theoretical Study of the Domain Configuration in Thin Layers of $BaFe_{12}O_{19}$," Philips Research Report, *15,* No. 1 (February 1960), pp. 7–29.
6. Hagedorn, F. B., Conference on Magnetism and Magnetic Materials, Philadelphia, Pa., November 18–21, 1969.
7. Forsyth, A. R., *Calculus of Variations,* London: Cambridge University Press, 1927, pp. 467–468.
8. Bobeck, A. H., unpublished work.
9. Chikazumi, S., *Physics of Magnetism,* New York: John Wiley, 1964, pp. 189–192.
10. Brown, W. F., Jr., *Micromagnetics,* New York: Interscience, 1963.
11. Snow, C., *Magnetic Fields of Cylindrical Coils and Annular Coils,* Nat. Bureau of Standards Appl. Math. Series 38, Washington, D. C., 1953.
12. Alexander, N. B., and Downing, A. C., *Tables for a Semi-Infinite Circular Current Sheet,* Oak Ridge Nat. Laboratory Rep. ORNL-2828, Physics and Mathematics, Oak Ridge, Tennessee.
13. Smith, D. O., "Proposal for Magnetic Domain-Wall Storage and Logic", IRE Trans. Electronic Computers, *10,* No. 4 (December 1961), pp. 709–711.
14. *Handbook of Mathematical Functions,* edited by Abramowitz, M., and Stegun, J. A., Nat. Bureau of Standards Appl. Math. Series 55, Washington, D. C., 1966, pp. 589–626.
15. Jahnke, E., and Emde, F., *Tables of Functions,* 4th Ed., New York: Dover, 1945, p. 73.
16. Groebner, W., and Hofreiter, N., *Integraltafel,* 4th Ed., New York: Springer-Verlag, 1965, pp. 59–72.

Physical and Transmission Characteristics of Customer Loop Plant

By PHILIP A. GRESH

(Manuscript received June 17, 1969)

This report covers the principal physical and transmission characteristics of the Bell System customer loop plant. Items covered include a statistical characterization of physical composition, measured and calculated transmission characteristics, and measured noise and crosstalk performance. A survey conducted in 1964 provided the data base for this report and comparisons of data obtained from a similar survey in 1960 illustrate that, in many respects the composition of loop plant changes only slowly with time. Consequently, the 1964 survey results are believed to be representative of today's plant.

The types of analyses presented in this paper are of increasing interest to certain Bell System customers because of the increasing number and types of services provided over local telephone facilities.

I. INTRODUCTION

This report covers the principal results of the 1964 Bell System customer loop survey. This survey provides a statistical characterization of physical composition, measured and calculated transmission characteristics, and measured noise and crosstalk performance of customer loop plant. Comparisons of data obtained from the 1964 survey and a similar survey made in 1960 are also presented.

Several of the principal transmission characteristics of Bell System customer loop plant as defined by the 1960 loop survey were published in 1962 by R. G. Hinderliter.[1] Additional published data on the transmission characteristics of Bell System toll connections is available in a BSTJ article by I. Nåsell.[2]

The 1964 Bell System survey was comprised of two separate surveys which were merged for analysis and presentation purposes. The basic survey was the general loop survey which consisted of a simple random sample of 1,100 main stations selected from the population of all main

stations (45, 300, 000) as of January 1, 1964. However, since only 3.25 percent of all main stations are served by loops longer than 30 kilofeet, only 35 samples would have been obtained to define the characteristics of the longer loops. Consequently, a long loop survey consisting of a random sample of 955 main stations served by loops longer than 30 kilofeet was obtained. The data obtained from the long loop survey has been used in those instances where characteristics are being expressed as a function of length to permit better resolution of the characteristics for the longer loops. In both of these sub-surveys, official telephones, foreign exchange lines, dial teletypewriter exchange (TWX) lines and special service lines were omitted as it was felt that their design would not be representative of customer loop plant.

II. SUMMARY OF RESULTS

Analyses of data obtained in the 1964 loop surveys lead to six general results.

(*i*) The average customer loop length is 10.6 kilofeet with only 10 percent of the main stations located beyond 21 kilofeet from their serving office. The length distributions show a slight trend toward longer loops between 1960 and 1964, with the average loop length increasing by 300 feet.

(*ii*) The average 1 kHz insertion loss of Bell System loop plant is 3.8 dB and 95 percent of all main stations are served by loops having a 1 kHz loss of less than 8 dB. At 3 kHz, the average loss is 7.8 dB and 95 percent of the main stations have less than 17 dB insertion loss.

(*iii*) The average noise balance of party-line loops is 56 dB, while the balance for individual line loops is 69 dB. Only 5 percent of the individual line loops have a noise balance of less than 50 dB while nearly 20 percent of party-line customers are served by loops with less than 50 dB of balance. The substantially lower balance for party lines is largely due to the inherent circuit unbalancing effect caused by the use of grounded ringers for party-line service.

(*iv*) The average metallic circuit noise (C-message weighted) at a customer's station set is approximately 5.5 dBrnc including the noise contribution of the central office wiring as well the noise contribution of the outside plant facilities. Only 8 percent of the individual lines have noise in excess of the Bell System objective of 20 dBrnc. However, 18 percent of the party-line customers have circuits which have noise in excess of 20 dBrnc because of the generally poorer circuit balance of party-line circuits.

(*v*) Comparison of measured and calculated transmission characteristics of Bell System loop plant has demonstrated that the outside plant cable records are sufficiently accurate to permit characterizing the loop plant transmission performance by theoretical calculations based on the physical composition of the loops as described in the outside plant records.

(*vi*) Main stations served by loops in excess of 30 kilofeet in length were found to be exponentially distributed as a function of working length, with the population of main stations reduced by 50 percent with every 11-kilofoot increase in loop length (see Fig. 6). It is estimated that 1.5 million Bell System customers (3.25 percent of all customers) are presently served by loops in excess of 30 kilofeet in length. Due to the party-line character of longer loops, the 3.25 percent of all Bell System main stations included in the long loop segment of plant used only 1.7 percent of the working Bell System exchange lines.

III. DESIGN OF THE SURVEY

The first steps in the survey were to define the population to be sampled and to obtain a complete list of the sampling units. In these two surveys (that is general loop and long loop), main stations were selected as the sampling units and all Bell System main stations as of January 1, 1964, were taken as the population to be sampled. A simple random sample was chosen as the sampling plan.

The sample size of the general loop survey was selected to provide data of equal precision to that obtained in the 1960 survey. The design parameter chosen was the average distance to the sampled main stations, and the precision was measured in terms of the width of the confidence interval bounding this average distance. The desired confidence interval (at 90 percent confidence level) of ± 450 feet on the average cable distance to the sampled main stations dictated a sample size of 1,100 randomly selected main stations. The actual confidence interval obtained was ± 476 feet.

In the long loop survey, lack of previous knowledge concerning the composition of long loops made it difficult to accurately determine the minimum sample size which would provide sufficient precision. The design parameter selected for the long loop survey was the average noise metallic (C-message weighting) measured at the telephone sets of the sampled main stations to a dialed-up termination. The precision aimed for was a ± 1.0 dB confidence interval (at 90 percent confidence level). A sample size of 955 main stations was collected, and the confidence interval obtained was ± 0.73 dB.

The two surveys had satisfactorily wide geographical dispersion, with every associated company (except Canada) contributing to the survey. Reference to Fig. 1 will illustrate that the large companies and the metropolitan areas contributed heavily to the general loop survey and the rural areas contributed heavily to the long loop survey.

IV. LOOP SURVEY RESULTS—PHYSICAL COMPOSITION

Data obtained in the loop survey included detailed loop schematics indicating the loop composition of each of the loops sampled in the survey. All distributions of physical quantities discussed herein were derived by analysis of these loop schematics. Since similar data were obtained in the 1960 loop survey, comparison of the physical distributions obtained in the two surveys has also been made.

Table I gives a summary of the statistics for the principal physical properties of loop plant. Data are included for both the 1960 and 1964 surveys and significance levels for differences of mean values are presented when meaningful. Cumulative distributions of these factors are shown in Figs. 2 through 5. The distribution of working bridged tap is not given since 82 percent of the sampled main stations were served by loops having zero working bridged tap and consequently the distribution is not particularly enlightening.

As indicated in Table I, the estimated average route distance from serving central office to main station in the Bell System is 10.6 kilofeet with 90 percent confidence that the true mean value lies within ±476 feet of this estimate. Note that although the estimated mean working length in 1964 is over 300 feet longer than that estimated in 1960, it is not statistically possible to claim that the observed increase is indeed

TABLE I — 1964 CUSTOMER LOOP SURVEY SUMMARY OF MAIN STATION CHARACTERISTICS

Main Station Quantity	Mean (ft)		90% Confidence Limits on Mean (± ft)		Sign. Level for Difference of Means in Percent
	1960	1964	1960	1964	
Working length	10,288	10,613	450	476	*
Total bridged tap	2,619	2,478	169	172	*
Working bridged tap	381	228	107	74	95
Airline distance	7,604	7,758	353	386	*
Working length/ airline distance	1.45	1.50	0.02	0.03	98

Drop wire excluded except when individual lengths exceed 400 feet.
* Levels of significance less than 80 percent indicated by asterisk.

an increase. Reference to the cumulative distribution of working length depicted in Fig. 3 will, however, show that shifts in the distribution have occurred since 1960. Note that the percentage of longer loops increased from 1960 to 1964.

Analysis of the long loop survey data has shown that the Bell System main stations served by loops in excess of 30 kilofeet are exponentially distributed as a function of working length as depicted in Fig. 6, with the main station population diminishing by 50 percent with each 11 kilofeet increase in loop length. Survey analysis indicates that about 1.5 million or 3.25 percent (with 90 percent confidence interval of ± 0.2 percent) of all Bell System main stations were located 30 kilofeet or more from their serving central offices as of the end of 1964. Due to the party-line character of the longer loops, the 3.25 percent of all Bell System main stations included in the long loop segment of plant used only 1.7 percent of the 39,300,000 Bell System lines working in 1964.

Analysis of the survey data has also provided valuable insight into the type-of-service distribution of Bell System customers and the physical composition of the plant provided to meet this distribution as shown in Figs. 7 to 10. The type-of-service distribution was derived as a function of length to the sampled main station and took advantage of the pooling of data from the two surveys. To evaluate the physical composition (type of facility, gauge, and pair size) of the loop plant as a function of distance, the sampled loops from the general loop survey were inspected at intervals of 1,000 feet starting at the central office. Both the general loop survey and the long loop survey were similarly inspected to define these distributions beyond 30 kilofeet.

The extent of party-line development as a function of loop length is shown in Fig. 7. Note the rapid increase in eight-party development for loop lengths greater than 30 kilofeet. Examination of the pair size distribution as a function of distance from the central office (Fig. 8) shows rapidly decreasing pair size with distance (at the 50-kilofoot point 50 percent of the sampled loops are contained in cables with fewer than 16 pairs). Similarly, the distribution of gauge shown in Fig. 9 illustrates a rapid transition to coarse gauge with increasing distance from the central office. For example, at 30 kilofeet 60 percent of the sampled loops are composed of gauges coarser than 22 gauge. Note also (Fig. 10) that the longer loops are primarily developed with aerial facilities. For example, 78 percent of all plant is aerial at the 30-kilofoot point from the central office. For the longer loops where small pair sizes are used, the pole line costs become a significant portion of the total loop costs and this factor is one of the reasons for the present trend towards the use

of buried plant. Since the sampled loops were randomly selected from all existing plant, buried plant is not as prominent as it would be in a sample of new construction. Note, however, that beyond 30 kilofeet, buried facilities in 1964 accounted for approximately 20 percent of the loops.

V. 1964 LOOP SURVEY RESULTS—TRANSMISSION PERFORMANCE

Data obtained in the 1964 loop survey have provided considerably more comprehensive knowledge of the transmission performance of customer loop plant than heretofore available. In the 1960 loop survey all transmission performance data were developed by deriving equivalent "T" networks from the information supplied on the loop sketches and analyzing these networks for transmission performance at nine frequencies in the voice band. Similar analysis has been performed for each of the sampled loops in the 1964 loop surveys, and in addition transmission measurements were made. The measurements covered noise, crosstalk, insertion loss at 1, 2 and 3 kHz, and dc resistance. The combination of these two sets of transmission performance data (one calculated and one measured) permits three types of analysis:

(*i*) changes in transmission performance since 1960 by comparison of calculated 1960 data with calculated 1964 data,

(*ii*) comparison of measured versus calculated data for the 1964 survey, and

(*iii*) provision of heretofore unavailable data on the noise and crosstalk performance of customer loop plant.

Since measured insertion loss data was not obtained in the 1960 survey, comparison of 1960 and 1964 data must be based on calculated values. Figure 11 depicts the 1 kHz calculated distributions for both surveys. It can be seen that insertion loss performance has remained virtually unchanged since 1960.

For those transmission characteristics where measured data are available in addition to the analytically derived data, minor differences in performance are exhibited by the two distributions of data (Fig. 12). In this regard it is important to realize that the measured data should provide a more accurate estimate of performance. There are several reasons for greater confidence in the measured data. First, possible inaccuracies in cable records or errors in transferring data from the records to the loop sketches can introduce errors in the calculated data. Second, errors in construction, such as omission or improper connection of loading coils, cannot be detected from the records. Third, use of calculated

data assumes that all cables exhibit nominal characteristics and consequently do not reflect manufacturing tolerances and environmental factors.

The cumulative distributions of insertion loss at 1, 2, and 3 kHz for customer loop plant as derived from both measured and calculated data are presented in Fig. 12. These insertion loss measurements and calculations were made with a 900 ohm source and load as depicted in Fig. 13. Measured loss was found to be consistently higher than calculated loss across the entire voice frequency band. The absolute differences between measured and calculated losses are small however, as indicated by the differences in mean losses. For example, at 1 kHz the measured loss was 3.8 dB and the calculated loss was 3.5 dB. This comparison of measured and calculated insertion losses demonstrates the feasibility of characterizing the loop plant transmission performance by theoretical calculations based on the physical composition of the loops as described by outside plant records. Still referring to Fig. 12, note that approximately 95 percent of all Bell System main stations are served by loops having a 1 kHz insertion loss of less than 8 dB with a mean loss of 3.8 dB. Similarly, at 3 kHz the 95 percent point is 16 dB and the mean loss is 7.8 dB. A scatter diagram of the 1 kHz measured insertion loss as a function of loop length is shown in Fig. 14. This diagram was obtained by merging the data from both the general loop survey and the long loop survey and indicates that the high loss loops are not limited to the long loop category. The high losses observed on some of the short loops generally reflect excessive bridged tap.

An insertion loss measurement of particular interest to designers of data equipment is the slope of loss versus frequency from 1000 to 2750 Hz. Cumulative distributions of the 2750 − 1000 Hz insertion loss (insertion loss measured with 900 ohm source and load) have been provided for all Bell System loop plant and for those loops serving business customers in Figs. 15 and 16 respectively.

Another important transmission characteristic is return loss, significant from echo and singing considerations. Return loss performance was not available from the measured data; consequently, it was calculated. Table II provides 1964 loop survey return loss results for nine frequencies and Fig. 17 presents the cumulative distributions and histograms for the 3 kHz singing return loss and echo return loss (equal weighting of the 500 to 2,500 Hz band). These data are all developed on the basis of looking into the customer loop at the central office end of the loop. The return loss is obtained by matching against a 900 ohm, 2.16 μF termination at the central office, with the customer end of the

TABLE II — CALCULATED RETURN LOSS FROM OFFICE TOWARD STATION*

Frequency (Hz)	1964	
	Mean (dB)	90% Confidence Interval (± dB)
200	8.0	0.11
300	10.2	0.12
500	13.4	0.17
1,000	15.4	0.30
1,500	13.1	0.27
2,000	10.9	0.25
2,500	9.1	0.20
3,000	7.7	0.16
3,200	7.1	0.15
Echo	11.2	0.15

* Station end of loop terminated in impedance of "off-hook" station set.

loop terminated in the impedance of the "off-hook" station set. Similar data on return loss are presented later from the station end of the loop.

The comparison of measured versus calculated loop resistance shown in Fig. 18 indicates that for general loop plant there is no significant difference between measured and calculated data but calculated loop resistance is slightly higher than measured resistance (574 ohms calculated, 567 ohms measured). Measured values may have been influenced by the fact that measurements were made during the winter. Calculated resistances were based on an average temperature of 68°F.

Theoretical calculations cannot be made of all transmission performance characteristics. Two such examples are noise and crosstalk. Since these characteristics are dependent upon external influences (induction from adjacent power lines, cable pair balance and the particular pair assignment), field measurements were made on each of the sampled loops using a Western Electric Company model 3A noise measuring set. The noise and crosstalk measurements were made as depicted in Figs. 19 and 20.

It is convenient to analyze loop noise in terms of the two factors which contribute to the resultant interference. The first of these is the magnitude of open circuit longitudinal voltage induced from power lines and the second is the circuit balance of the cable pairs and central office equipment. The cumulative distribution of the open circuit longitudinal voltage for general loop plant is shown in Fig. 21 for 3 kHz flat weighting. This voltage is induced in a longitudinal mode, and consequently only that portion of it which is converted to the metallic circuit

will create an interference problem. The circuit balance reflects the extent to which the longitudinal voltage is converted to metallic voltage and is, therefore, a measure of the susceptibility of the telephone plant to inductive interference such as power-line hum.*

As seen in Fig. 22 party lines are much more susceptible to power-line hum than individual lines because of the unbalance introduced by the grounded ringers associated with party-line station sets. On the average, individual lines have approximately 12 dB better balance than party lines. Part of this is a result of the shorter length distribution of individual lines which offers less opportunity for cable pair unbalances to accumulate; but it is reasonable to expect a balance improvement of 10 dB with ringers isolated from ground.

The combination of the induced longitudinal voltage and the circuit unbalance produces the metallic noise distribution at customers' station sets (in off-hook state) as shown in Figs. 23 and 24. For comparison purposes, the C-message weighted noise to ground (longitudinal noise) is also shown on these figures. Figure 23 depicts the noise contribution of the loop plant only, while Fig. 24 includes the noise contribution of the central office wiring. In both cases the station end of the loop was terminated in an off-hook 500-type station set with the transmitter and receiver replaced by equivalent resistors. The metallic noise has been measured with C-message weighting to reflect the relative interfering effects of the noise on voice transmission. The important limits to consider are the Bell System long-term noise performance objective of 20 dBrnc and the immediate remedial action limit of 30 dBrnc. As seen in Fig. 24, only 8 percent of the individual lines had total metallic noise in excess of 20 dBrnc. However, 18 percent of the party-line customers have noise in excess of 20 dBrnc.

The near-end crosstalk coupling loss characteristics of customer loop plant as derived from measured data from the general loop survey are shown on Fig. 25. Along with the overall distribution of crosstalk coupling loss is shown the distribution for nonloaded loops only (84 percent of all sampled main stations). The nonloaded loop distribution

---

* Loop circuit noise balance is defined here as

$$20 \log_{10} \frac{\text{open circuit longitudinal voltage}}{\text{metallic voltage}}$$

where both voltages are measured with C-message weighting. The validity of this definition depends on the assumption that the longitudinal voltage induced from adjacent power lines is the only source of metallic noise. This is generally not true when the noise to ground measures less than 20 dBrnc, and consequently loop balance for such loops cannot be computed from the measurements.

can be approximated by a normal distribution with a mean crosstalk coupling loss of 115 dB and a standard deviation of 12 dB. A comparison of these two curves indicates that the poorer crosstalk performance of longer loaded loops dominates the low loss tail of the general loop survey distribution.

The final transmission characteristic to be discussed is the loop input impedance as calculated both at the central office and at the station set. Figure 26 presents the plot of loop input impedance as seen at the central office as a function of frequency (not including central office wiring or equipment). For these calculations the station end of the loop was terminated in an off-hook 500-type station subset with the transmitter and receiver replaced by equivalent resistors. Curves have been provided separately for loaded and nonloaded loops because of the large difference in their characteristic impedance. Also shown is the characteristic impedance of the central office matching network as a function of frequency. The function of this network is to provide high return loss performance across the voice frequency band by matching as close as possible the impedance of the various loops. It is apparent that both the nonloaded loops and loaded loops should have their highest return losses around 1 kilohertz and that the loaded loops should perform more poorly than nonloaded loops at the lower frequencies.

Plots of mean input impedances, such as in Fig. 26, are useful for indicating the general input impedance behavior as a function of frequency. Variations that occur at each frequency, and their effects on return loss, are best shown as scatter diagrams. Figures 27 and 28 present the loop input impedance at 1 kHz for nonloaded and loaded loops. Superimposed on all scatter diagrams are return loss circles referenced to the 900 ohm and 2.16 μF matching network. Any loop having an impedance lying within a particular circle will have a return loss, when measured against the specified matching network impedance, which exceeds the given return loss value. Visual examination of the scatter pattern as it relates to the return loss circles provides a ready means of evaluating the return loss performance of various segments of the loop plant (assuming, of course, that the input impedances of loops in that segment are known).

The range and shape of the input impedance scatter pattern at each frequency are of interest because they point up the difficulty of designing a simple matching network which, at even a single frequency, will provide very high return losses for nearly all loops. Considering the characteristics of the nonloaded loops shown in Figure 27 it is evident

that many of the loops tend to follow a smooth curve, while the others are scattered about this curve. The smooth curve results from the variation in loop length, while the scatter is due to the effects of bridged tap, overgauging, and variations in types of subsets.

Perhaps of particular interest to Bell System customers are the input impedance characteristics of Bell System loop plant as seen from the station end of the customer loop. The input impedance of a customer loop as measured at the station set can vary considerably based on the type of facility connected to the loop at the central office. Various circuit connections may involve use of four-wire trunks, two-wire trunks or intraoffice circuits. In the following analysis a 900 ohm and 2.16 μF central office termination has been used to represent a four-wire trunk termination, and the midsection input impedance of 22 gauge H88 loaded cable has been used to represent a two-wire trunk. For the simulation of intraoffice calls, a Monte Carlo technique was used to select a random sample of 500 pairs of loops from the 1,100 loops in the general loop survey. A loop was randomly selected as the sample loop, and then the input impedance (from the central office) of another randomly selected loop was chosen for the central office termination.

The 1,100 loops of the 1964 general loop survey were segregated into two groups (loaded and nonloaded loops) for all but the simulated intraoffice calls because of the great differences in impedance range of the two populations. Presentation of scatter diagrams of input impedance from the station set has been limited to 1 and 3 kHz. These return loss circles were generated assuming the use of a 500-type station set and it was further assumed that the 500 set was operating on a current equal to the average loop current of 45.5 mA.

Figures 29 through 32 are the input impedance scatter diagrams for loops with a simulated two-wire trunk (22 gauge H88 loaded cable) termination at the central office. The scatter is primarily a result of overgauging, open wire, and bridged tap or varied end section length. Smoothed curves of the mean input impedances of loaded and nonloaded loops with a 22 gauge H88 cable termination are presented in Fig. 33. Scatter diagrams for the loops with a central office termination of 900 ohms and 2.16 μF (simulated four-wire trunk) are presented in Figs. 34 through 37. The general input impedance behavior of these loops as a function of frequency is indicated in Fig. 38 by the plot of the mean input impedances at nine voiceband frequencies.

The scatter diagrams and the mean input impedance curve for the simulated intraoffice calls are shown in Figs. 39 and 40. The effect of connecting together two loops, one of which is terminated by a station

set, and the other whose input impedance is calculated at its station set, is shown by the mean input impedance curve for simulated intra-office connections in Fig. 41. This curve has a shape characteristic of longer nonloaded loops. The mean input impedance curves for non-loaded loops with simulated two- and four-wire trunk terminations at the central office are also shown in Fig. 41. The major differences in the characteristics of these curves are at the low frequencies where the shunt capacitance of the cable masks the termination less than it does at high frequencies.

VI. ACKNOWLEDGMENTS

The author wishes to acknowledge the collaboration of Mr. F. L. Schwartz in the design and execution of this Bell System customer loop survey and his analysis of noise and crosstalk characteristics of loop plant. I also want to thank Mrs. A. F. Rogers for the substantial programming assistance required in the analysis of the data obtained in this study and Mr. D. B. Menist and Mrs. B. J. Hymanson for their contributions on the input impedance characteristics of loop plant.

REFERENCES

1. Hinderliter, R. G., "Transmission Characteristics of Bell System Subscriber Loop Plant," AIEE Summer General Meeting, Denver, Colorado, June 17–18, 1962, Paper 62–1198.
2. Nåsell, I., "Some Transmission Characteristics of Bell System Toll Collections," B.S.T.J., *47*, No. 6 (July-August 1968), pp. 1001–1018.

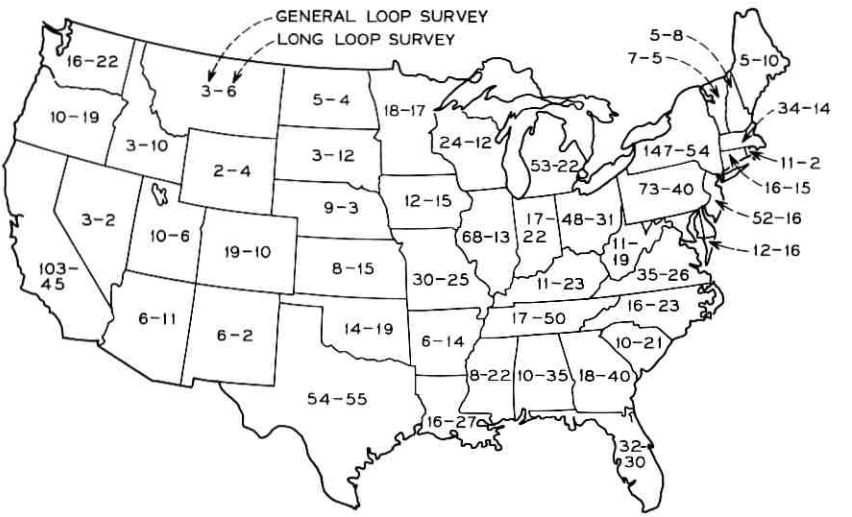

Fig. 1 — Geographic distribution of sampled loops for the 1964 customer loop survey. Eleven hundred loops were sampled for the general loop survey and 955 loops were sampled for the long loop survey.

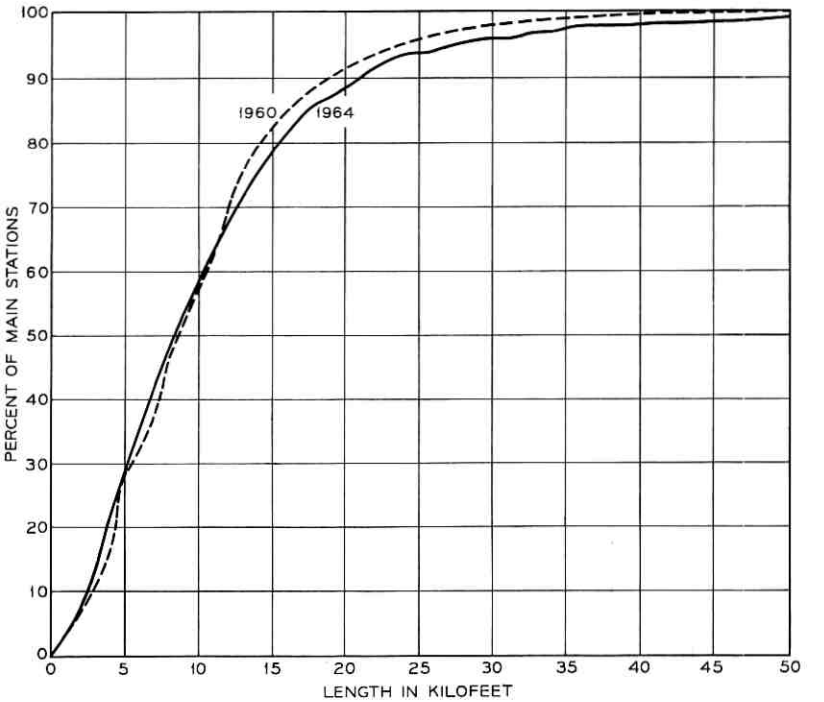

Fig. 2 — Working length to the main station.

	1960	1964
Mean (feet)	10,288	10,613
90 percent confidence limits on mean (feet)	±450	±476

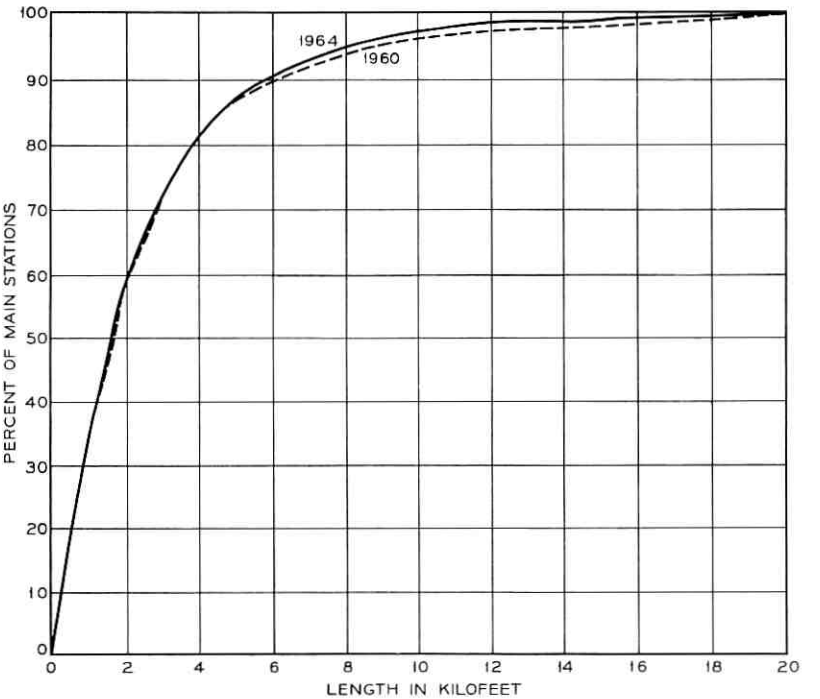

Fig. 3 — Total bridged tap.

	1960	1964
Mean (feet)	2,619	2,478
90 percent confidence limits on mean (feet)	±169	±172

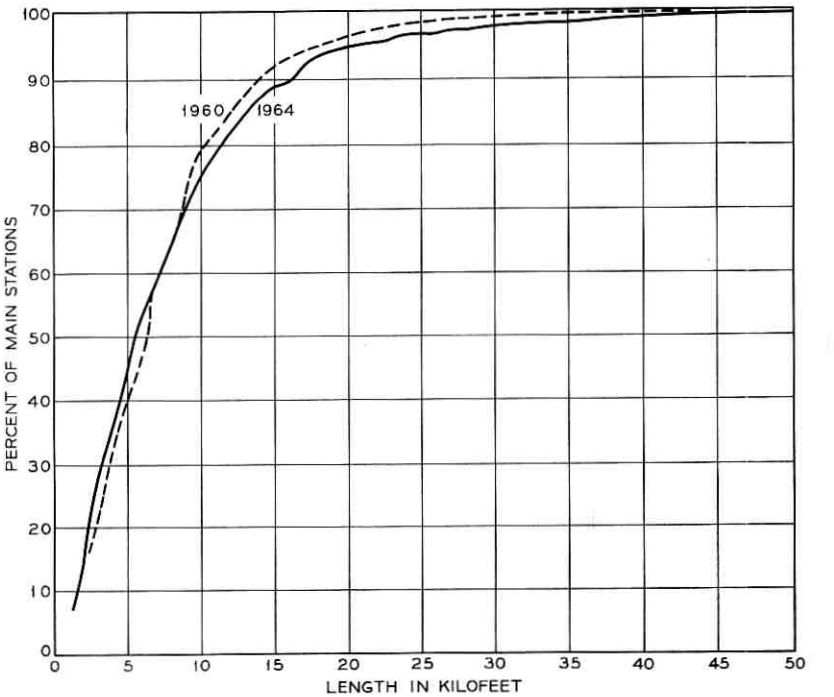

Fig. 4 — Airline distance to main station.

	1960	1964
Mean (feet)	7,604	7,758
90 percent confidence limits on mean (feet)	±353	±386

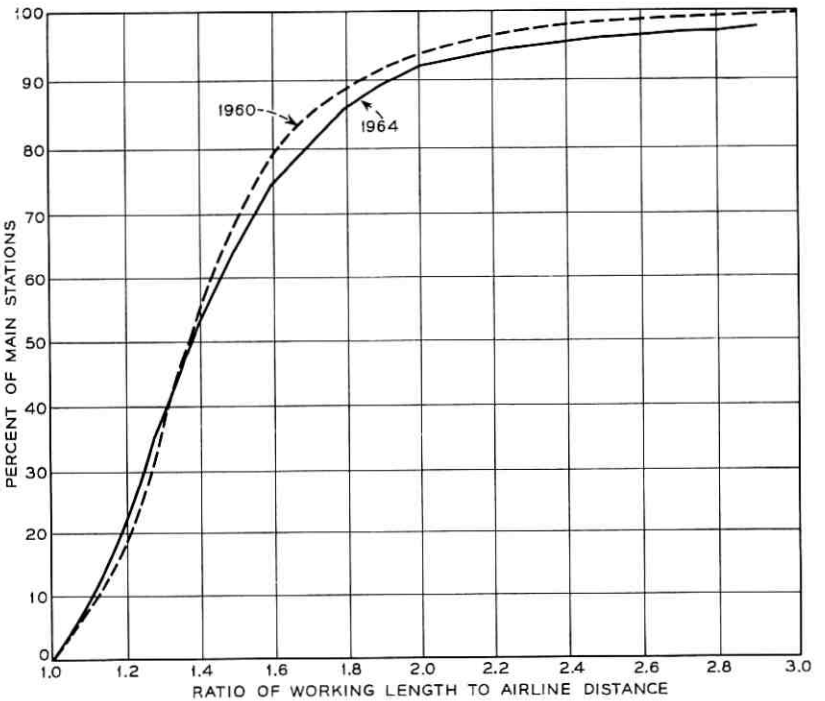

Fig. 5 — Ratio of working length-airline distance to main station.

	1960	1964
Mean	1.45	1.50
90 percent confidence limits on mean	±0.02	±0.03

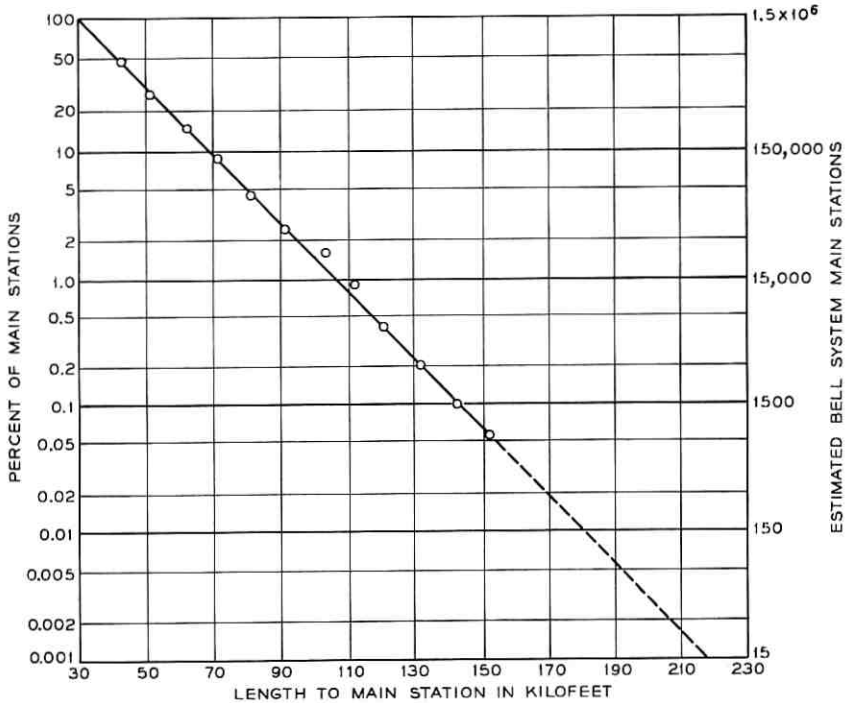

Fig. 6 — Distribution of long loops (1964 long loop survey). The mean was 45,938 feet; 90 percent confidence limits on the mean was ±870 feet.

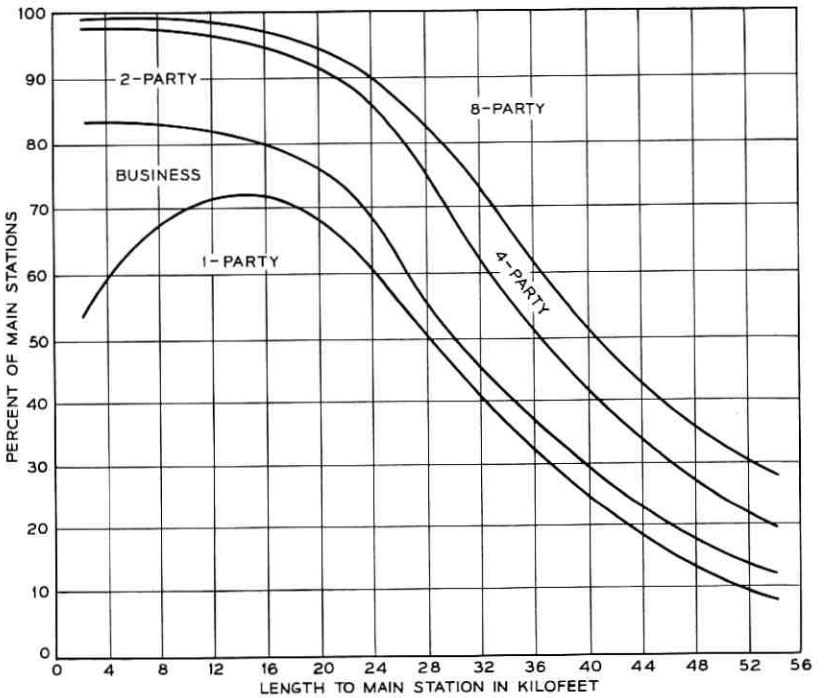

Fig. 7 — Type of service distribution versus loop length (1964 combined loop surveys). One-, two-, four-, and eight-party plots include residence service only; business includes PBX, centrex, and coin.

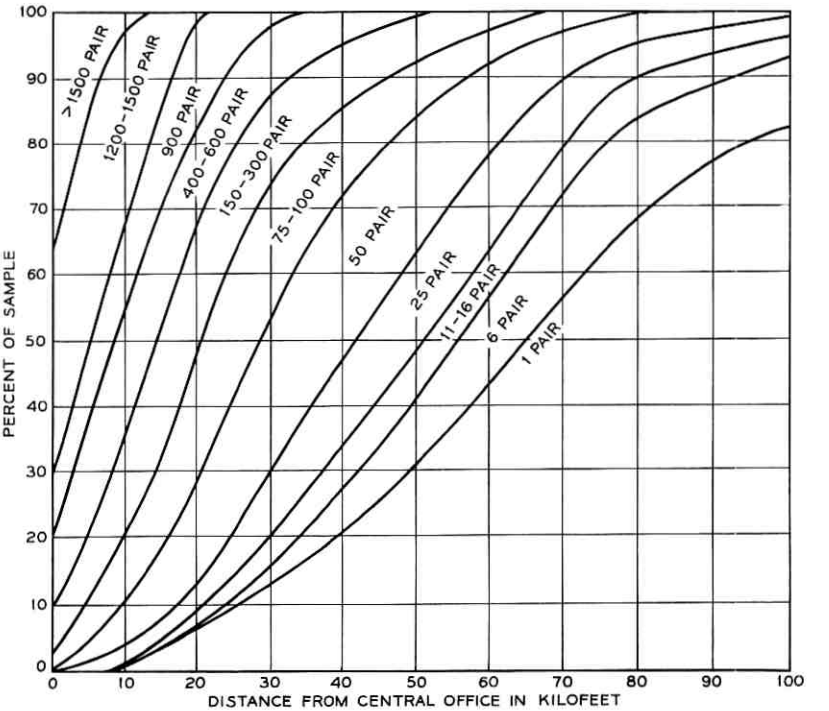

Fig. 8 — Pair size distribution (1964 combined loop surveys—general loop and long loop surveys).

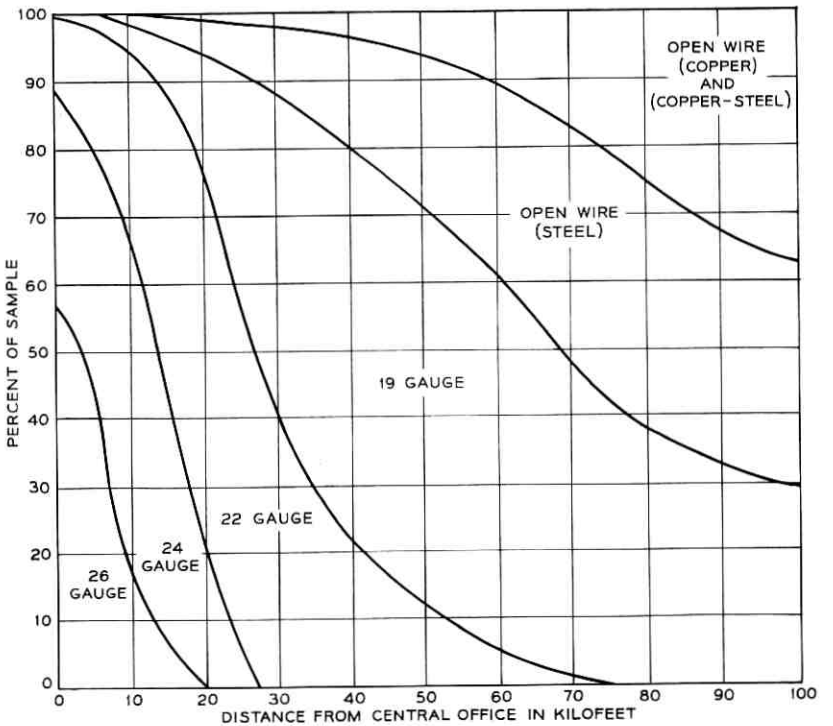

Fig. 9. — Gauge distribution (1964 combined loop surveys—general loop and long loop surveys).

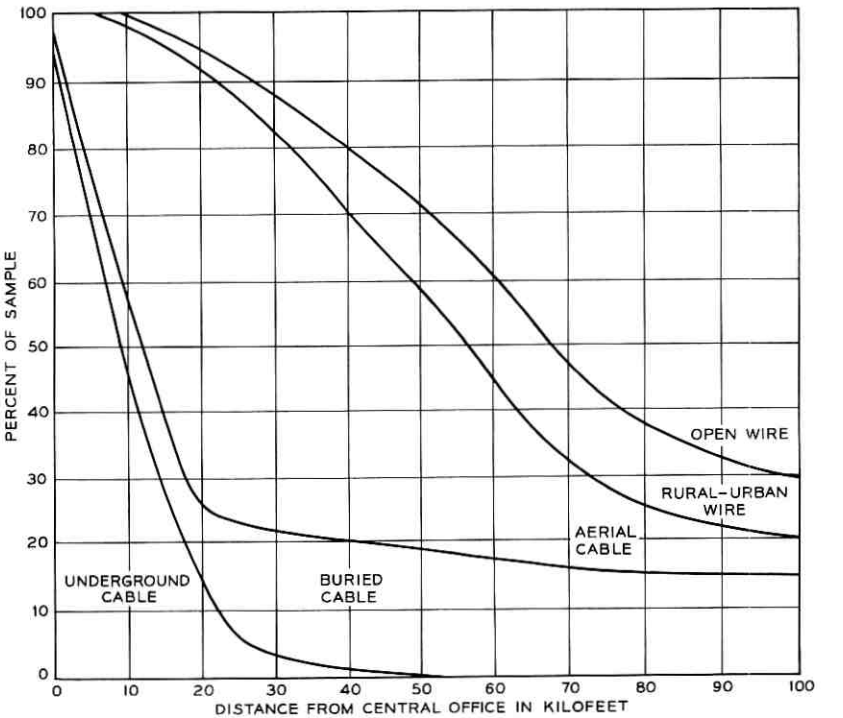

Fig. 10 — Type of construction distribution (1964 combined loop surveys—general loop and long loop surveys).

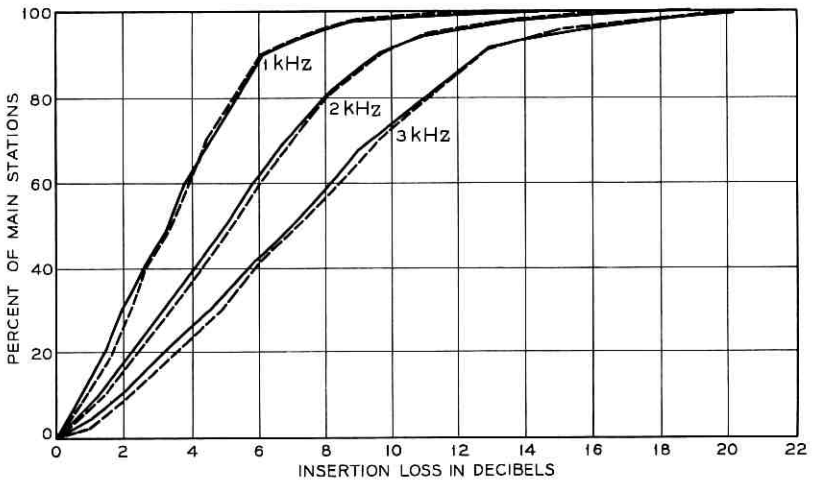

Fig. 11 — Distribution of calculated insertion losses at 1, 2, and 3 kHz (--- 1960; ——— 1964).

	1 kHz		2 kHz		3 kHz	
	1960	1964	1960	1964	1960	1964
Mean (dB)	3.4	3.5	5.4	5.3	7.4	7.3
90 percent confidence limit on mean (dB)	±0.11	±0.10	±0.18	±0.16	±0.24	±0.21

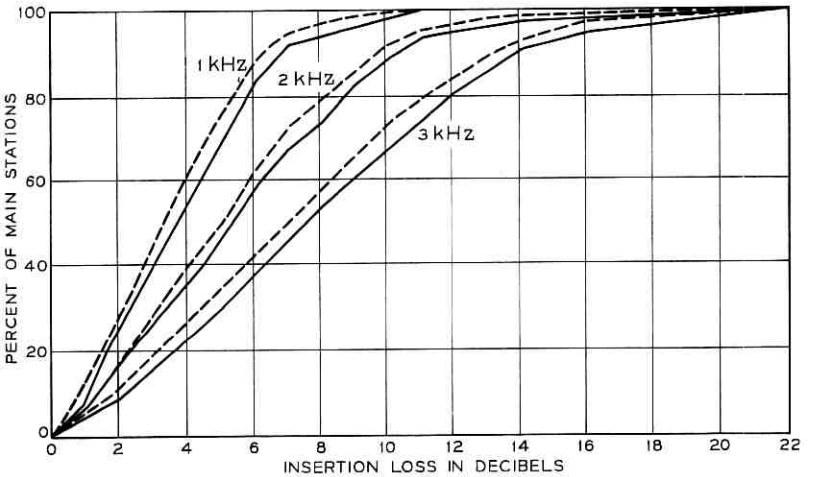

Fig. 12 — Distribution of measured and calculated insertion losses at 1, 2, and 3 kHz (---- calculated; ——— measured).

	1 kHz		2 kHz		3 kHz	
	Measured	Calculated	Measured	Calculated	Measured	Calculated
Mean (dB)	3.8	3.5	5.6	5.3	7.8	7.3
90 percent confidence limit on mean (dB)	±0.12	±0.10	±0.17	±0.16	±0.22	±0.21

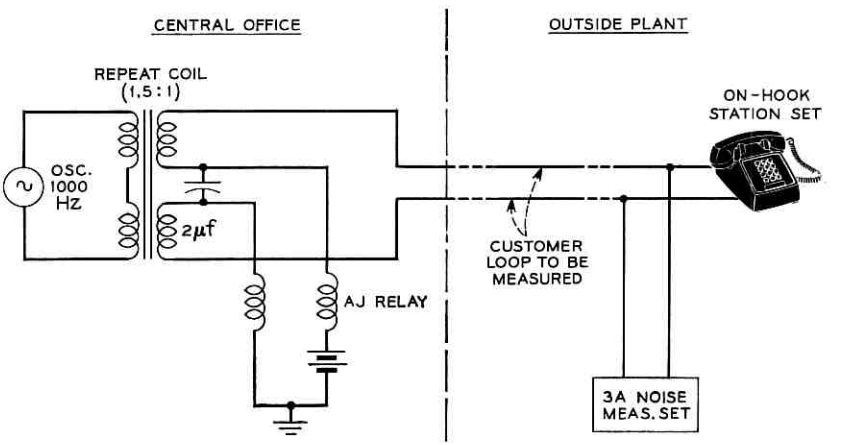

Fig. 13 — Insertion loss measurement technique. The oscillator is set for 600 ohm output termination; the 3 A noise measurement set is equipped with 900 ohm termination.

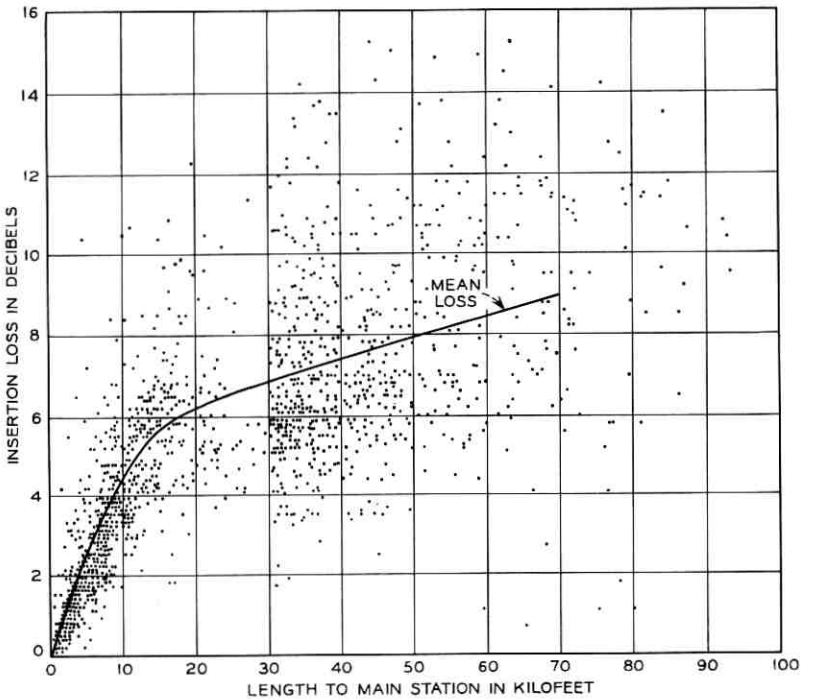

Fig. 14 — Measured 1 kHz insertion loss scatter diagram (1964 combined loop surveys—general loop and long loop surveys).

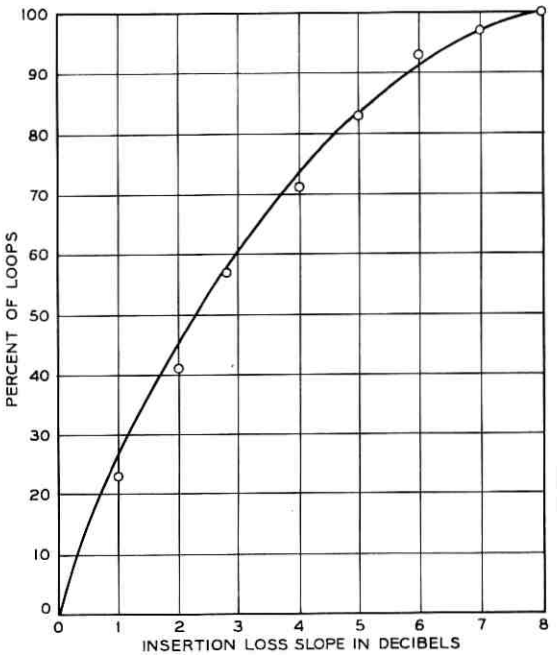

Fig. 15 — Distribution of insertion loss slope between 2750 and 1000 Hz for residential plus business loops.

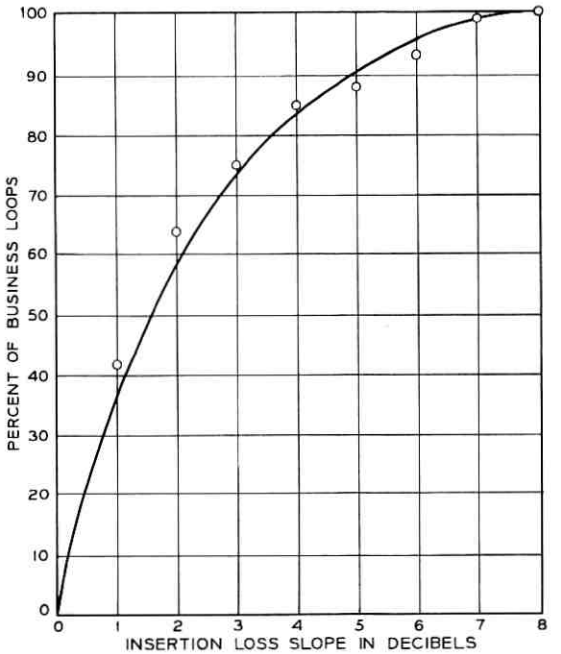

Fig. 16 — Distribution of insertion loss slope between 2750 and 1000 Hz for business loops only.

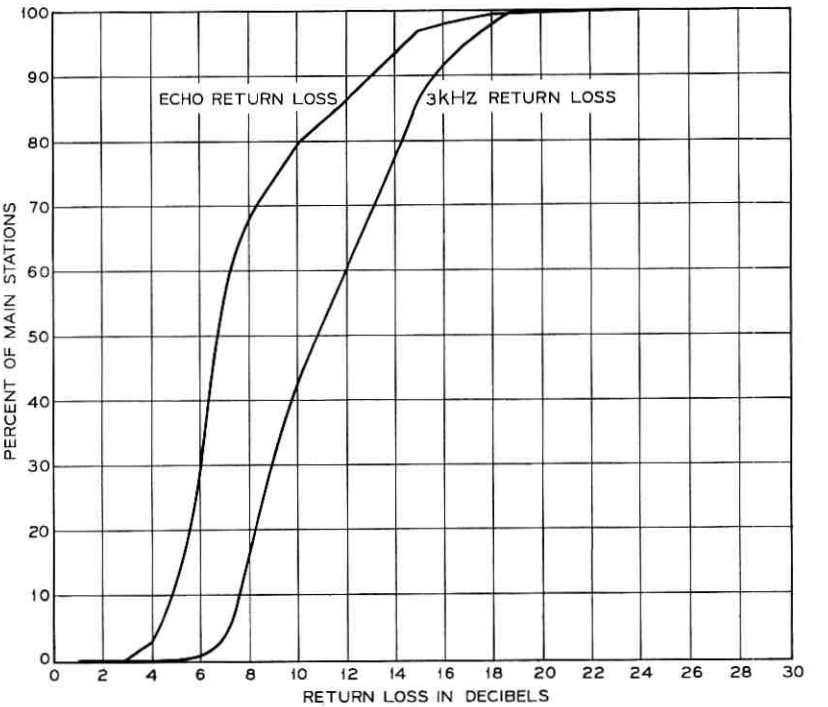

Fig. 17 — Distribution of 3 kHz and echo return losses at central office. Echo return loss distribution assumes flat weighting of 500–2500 Hz band.

	Echo	3 kHz
Mean (dB)	7.7	11.2
90 percent confidence limits on mean (dB)	±0.16	±0.15

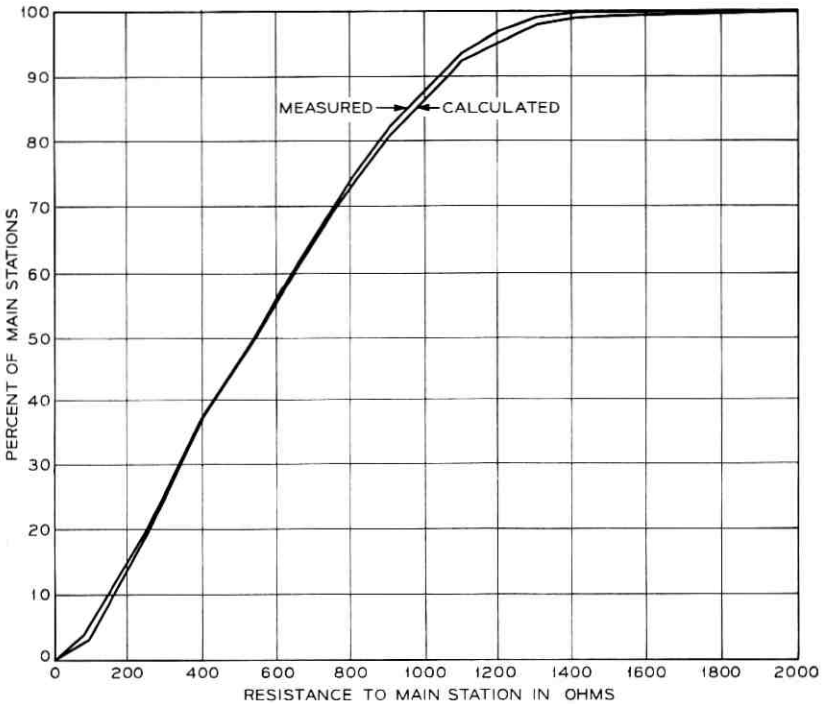

Fig. 18 — Measured and calculated distribution of resistance to main station.

	Measured	Calculated
Mean (ohms)	567	574
90 percent confidence limits on mean	±15.8	±17.0

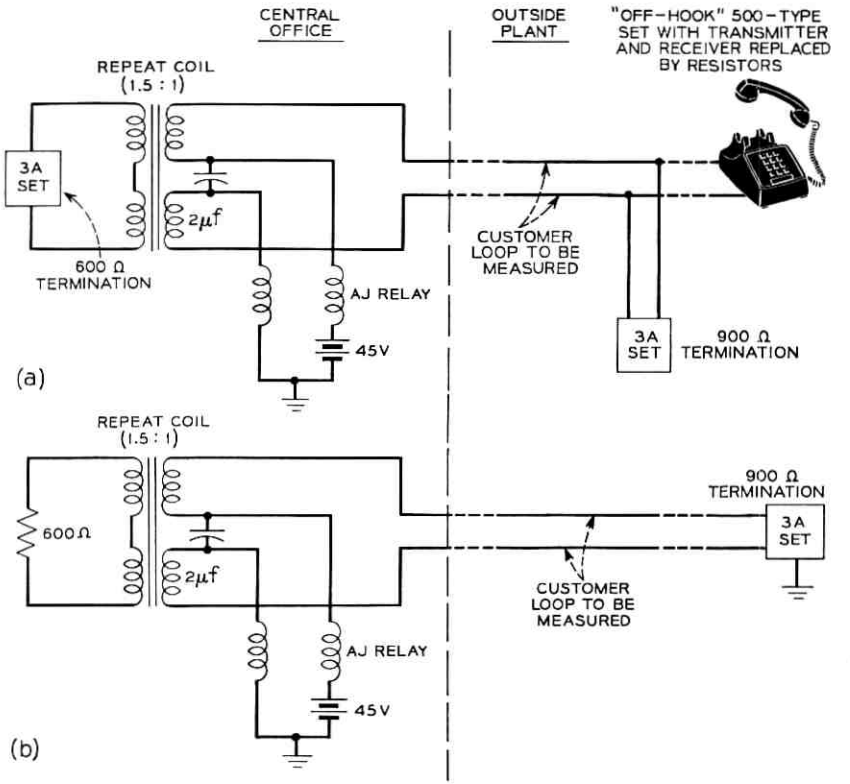

Fig. 19 — Noise measurement technique for (a) noise metallic—loop only and (b) noise longitudinal—loop only.

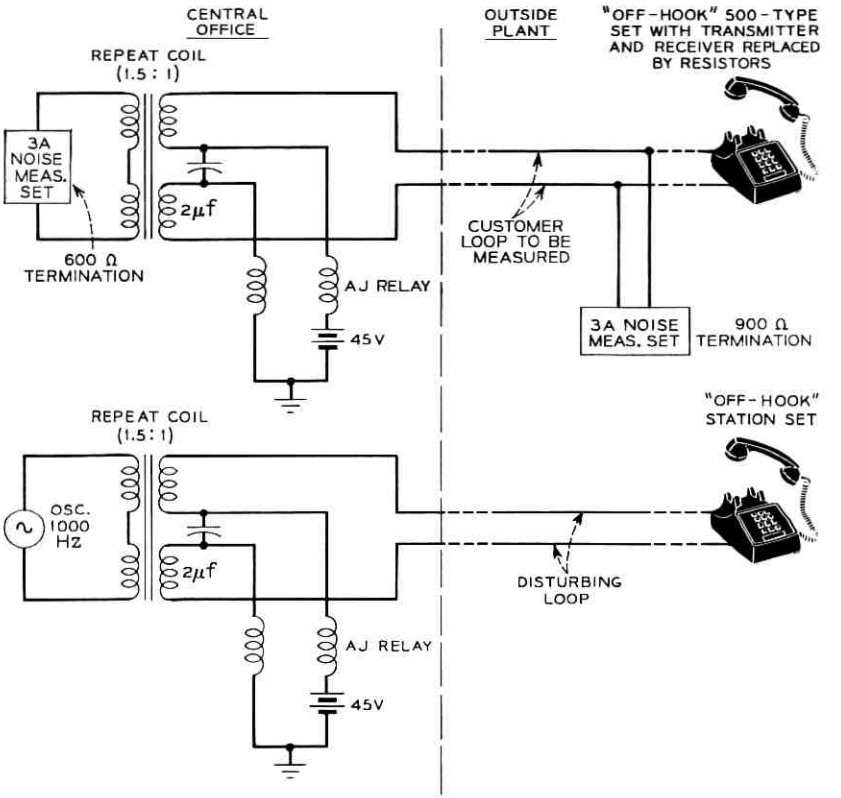

Fig. 20 — Crosstalk measurement technique. Disturbing loop—randomly selected pair in the same 100-pair group as the sample loop on which the measurements are being made. The oscillator on disturbing loop is set for 600-ohm output termination.

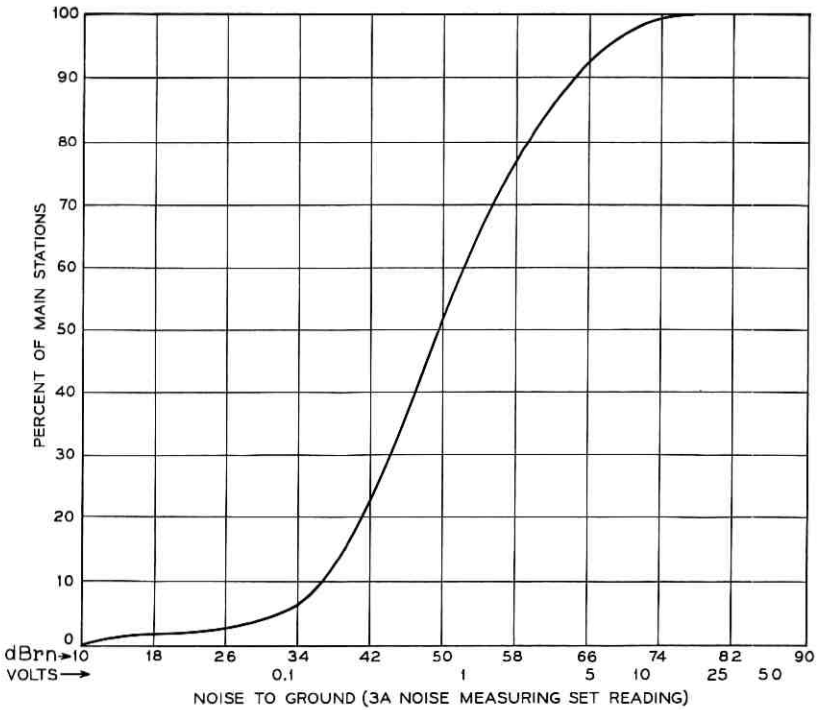

Fig. 21 — Noise to ground at main station for 3 kHz flat weighting. The mean was 49.2 dBrn; the 90 percent confidence limits on the mean was ±0.56 dBrn.

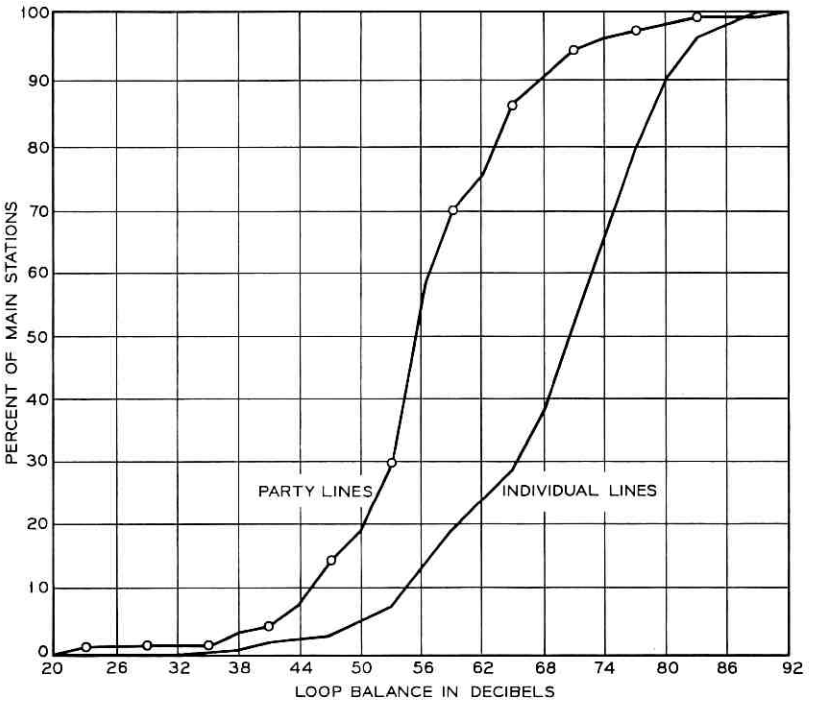

Fig. 22 — Loop circuit noise balance. These distributions are based on the 476 loops where measurements permitted an accurate estimate of loop balance to be made. These are for loops having noise to ground greater than 20 dBrnc.

	Party	Individual
Mean (dB)	56.1	68.7
90 percent confidence limits on mean (dB)	±1.55	±0.87

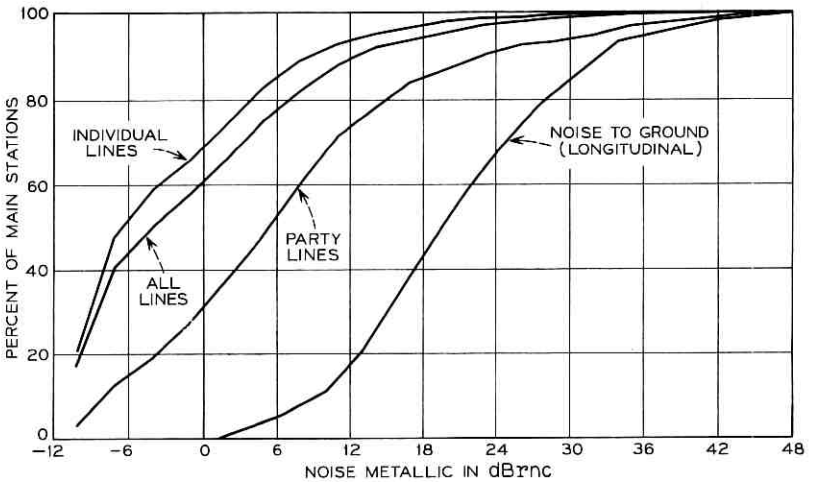

Fig. 23 — Noise metallic at main station with C-message weighting for loop plant only.

	Noise to Ground	Noise Metallic		
		All	Individual	Party
Mean (dBrnc)	19.1	−1.1	−3.1	6.7
90 percent confidence limits on mean (Bdrnc)	±0.5	±0.5	±0.5	±1.3

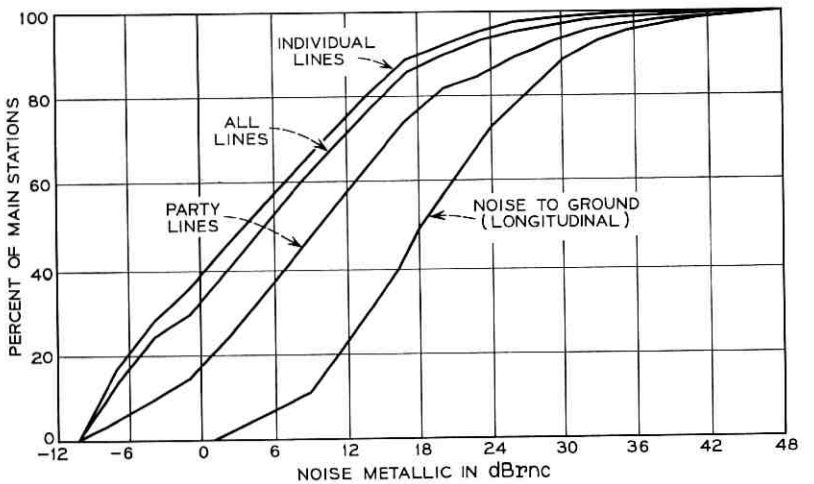

Fig. 24 — Noise metallic at main station with C-message weighting for loop plant plus central office.

	Noise to Ground	Noise Metallic		
		All	Individual	Party
Mean (dBrnc)	19.1	5.6	4.3	10.6
90 percent confidence limits on mean (dBrnc)	±0.5	±0.5	±0.6	±1.2

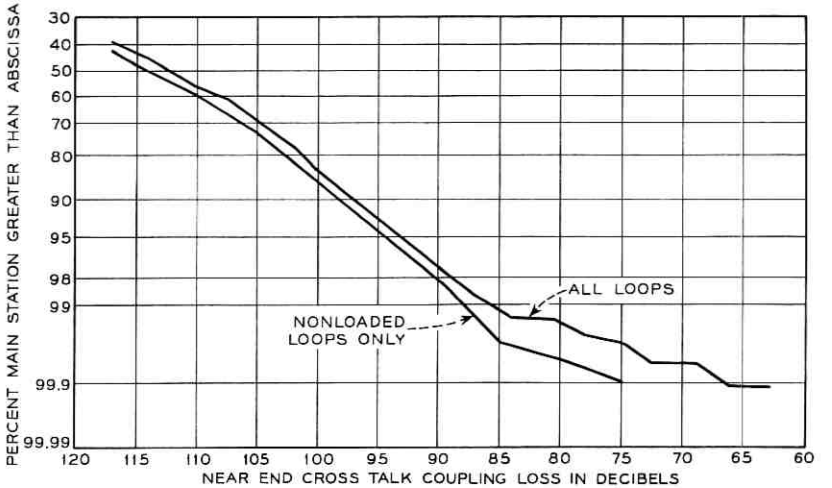

Fig. 25 — 1 kHz near end crosstalk coupling loss at central office. Central office terminated in 900 ohms; customer end terminated in receiver off hook station set.

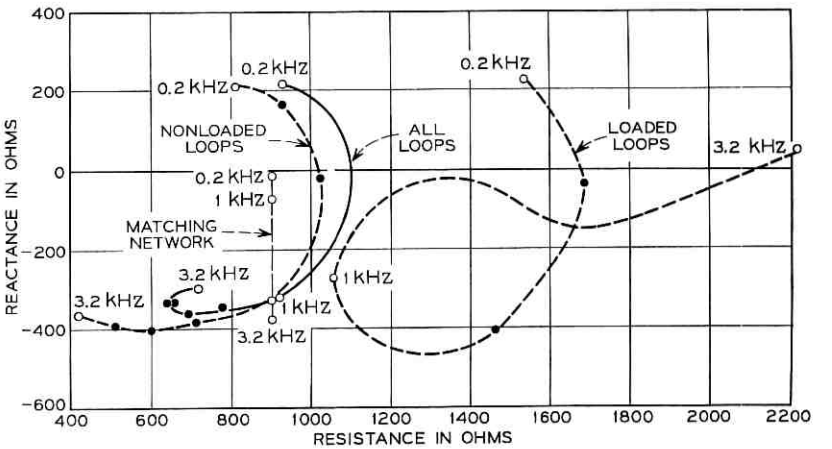

Fig. 26 — Loop impedance at central office.

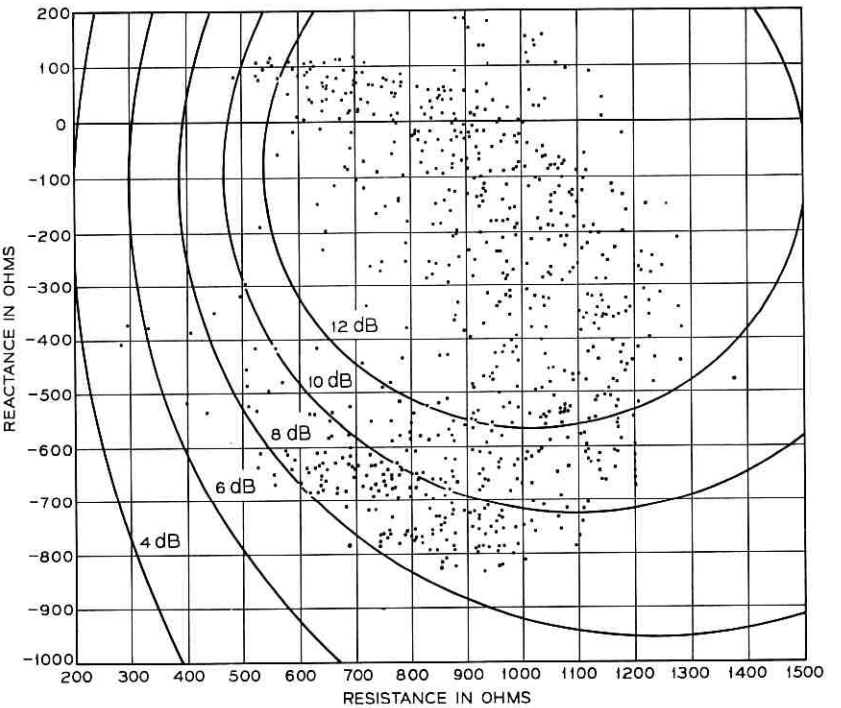

Fig. 27 — Nonloaded loop input impedances at 1 kHz measured from the central office. Return loss circles based on 900 ohm + 2 μF network.

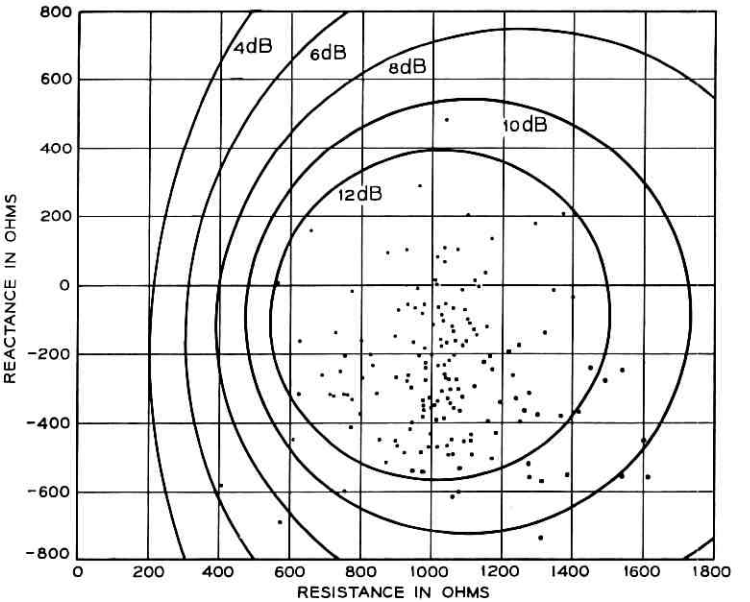

Fig. 28 — Loaded loop input impedances at 1 kHz as measured from the central office. Return loss circles based on 900 ohms + 2 μF network.

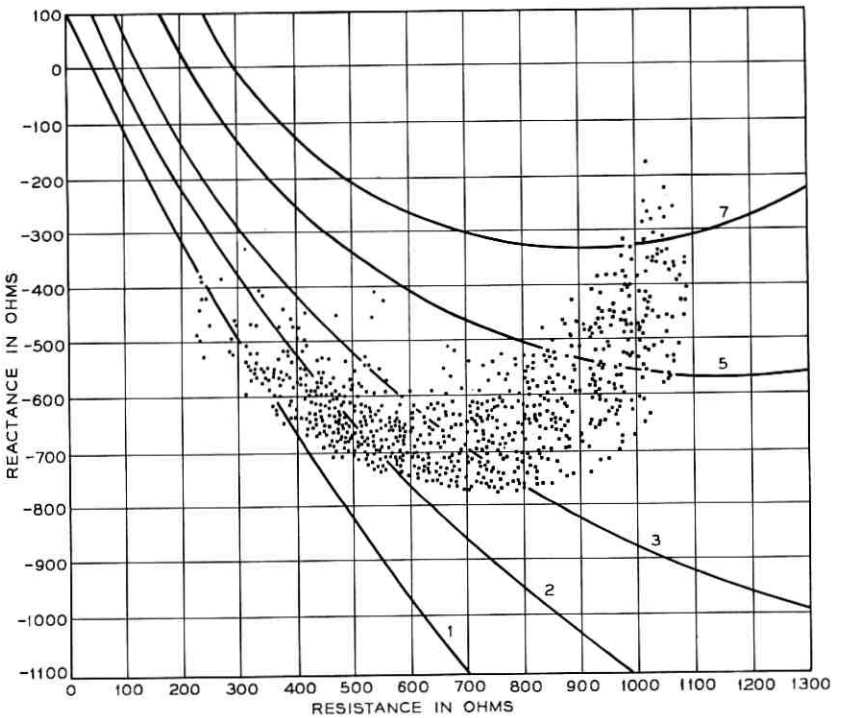

Fig. 29 — Nonloaded loop input impedances at 1 kHz measured from station set with a simulated two-wire trunk termination at the central office. Return loss circles based on 500-type subset impedance.

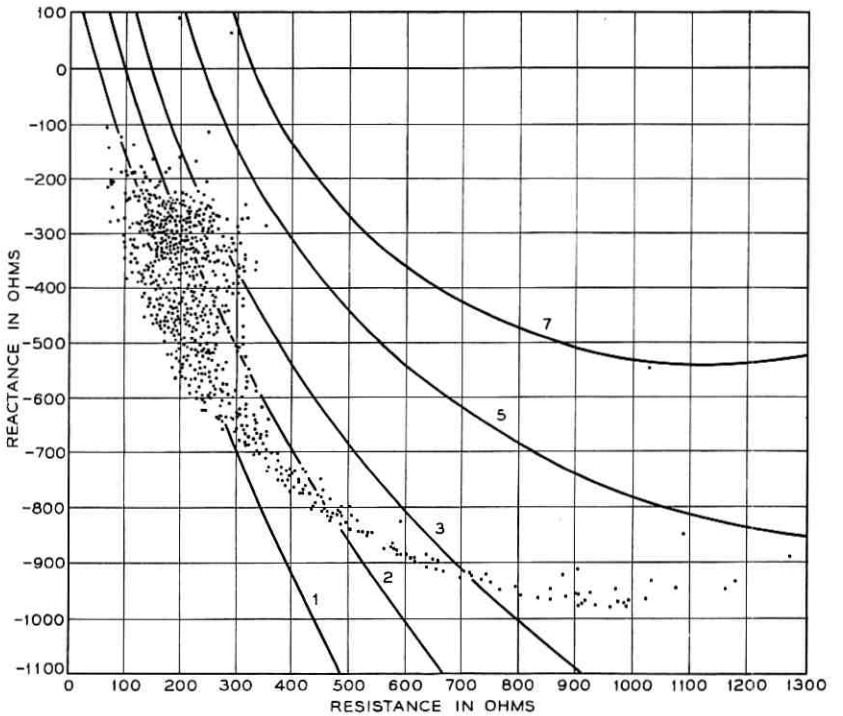

Fig. 30 — Nonloaded loop input impedances at 3 kHz measured from station set with a simulated two-wire trunk termination at the central office. Return loss circles based on 500-type subset impedance.

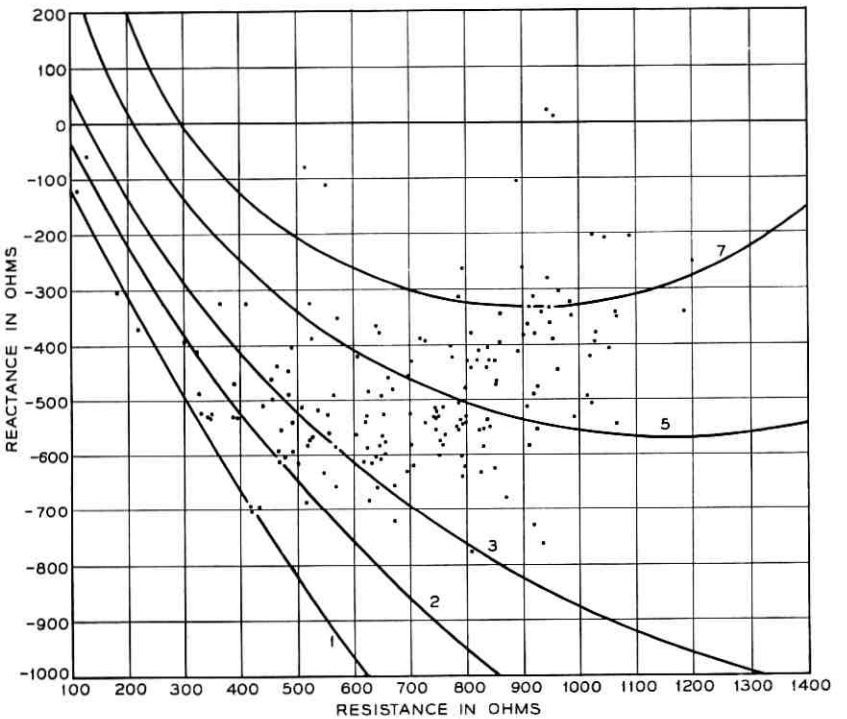

Fig. 31 — Loaded loop input impedances at 1 kHz measured from station set with a simulated two-wire trunk termination at the central office. Return loss circles based on 500-type subset impedance.

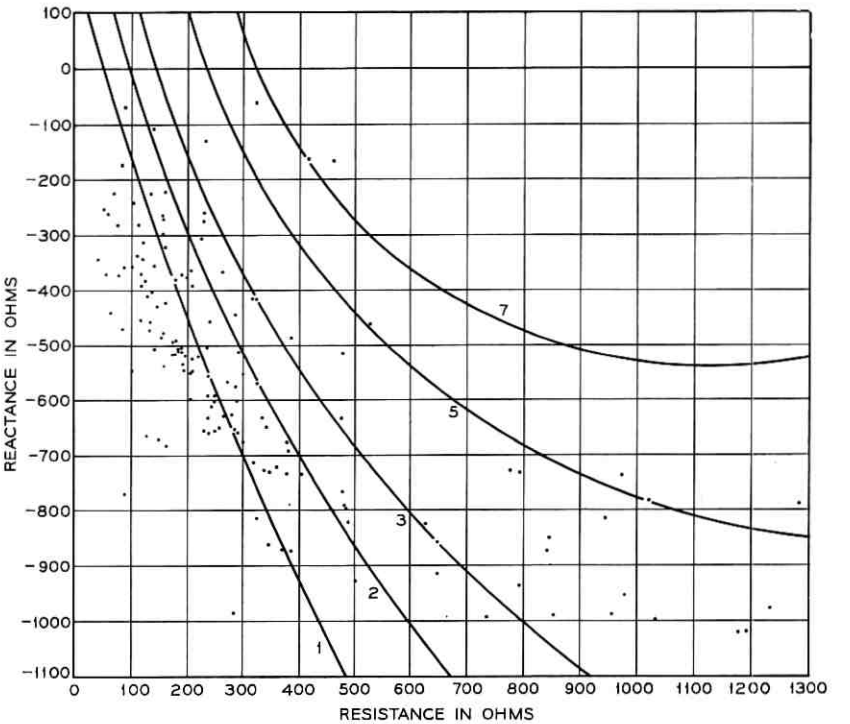

Fig. 32 — Loaded loop input impedance at 3 kHz measured from station set with a simulated two-wire trunk termination at the central office. Return loss circles based on 500-type subset impedance.

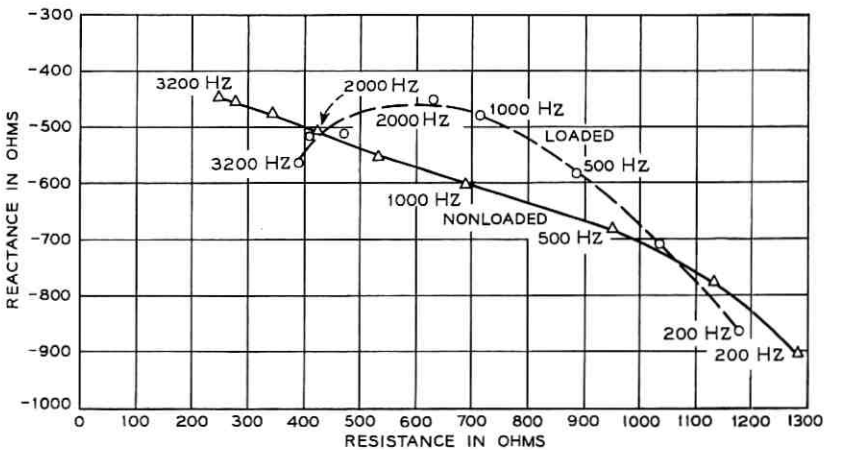

Fig. 33 — Mean value of loop input impedance from station set with simulated two-wire trunk termination at the central office.

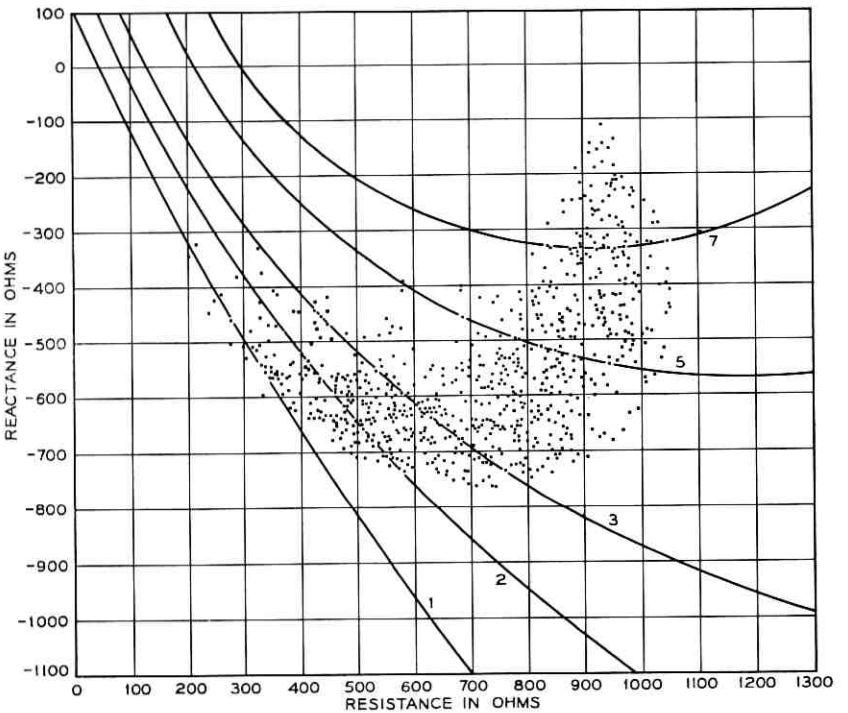

Fig. 34 — Nonloaded loop input impedances at 1 kHz measured from station set with simulated four-wire trunk termination at the central office. Return loss circles based on 500-type subset impedance.

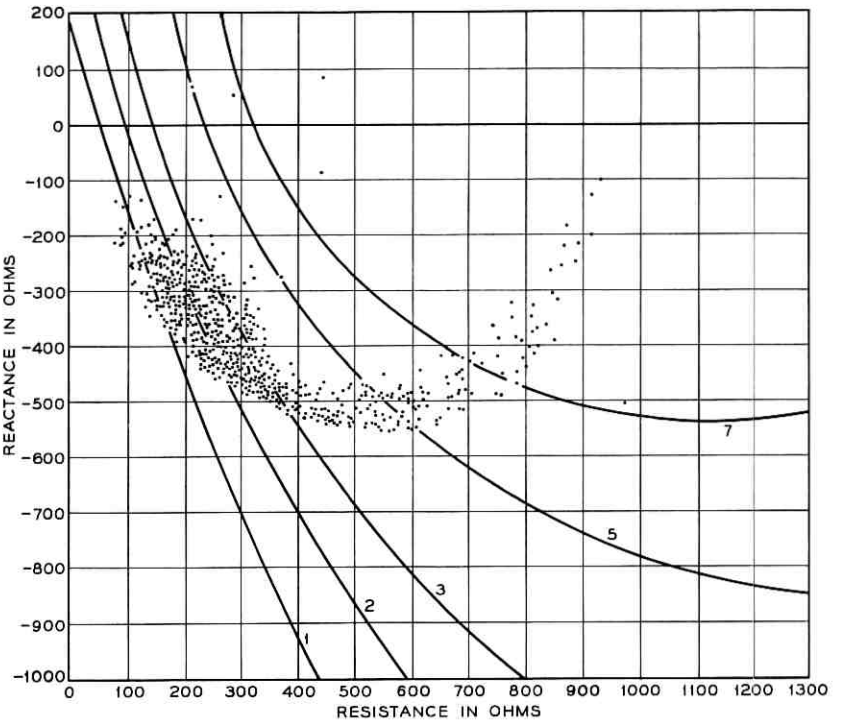

Fig. 35 — Nonloaded loop input impedance at 3 kHz measured from station set with simulated four-wire trunk termination at the central office. Return loss circles based on 500-type subset impedance.

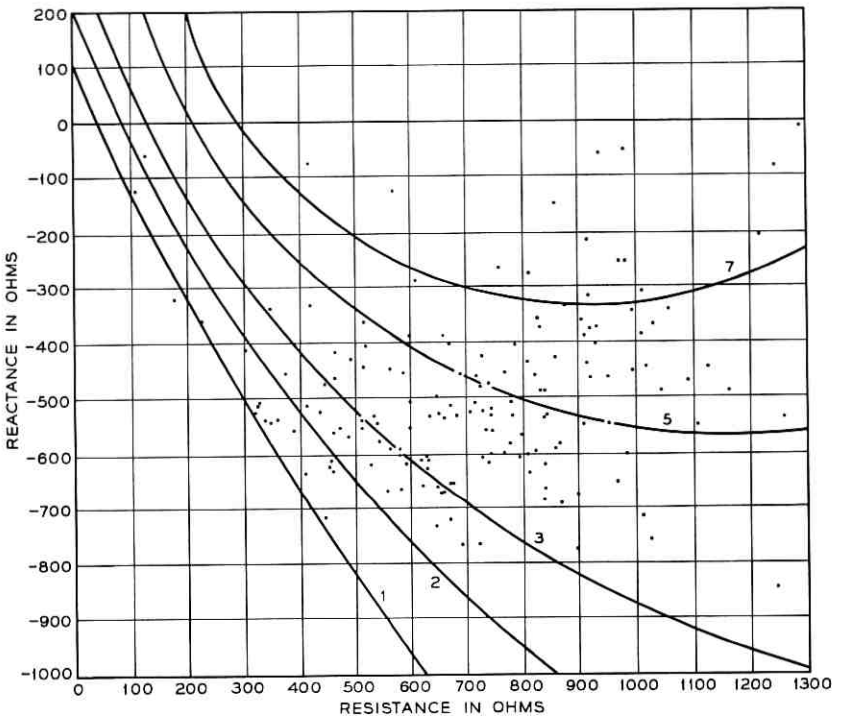

Fig. 36 — Loaded loop input impedances at 1 kHz measured from station set with simulated four-wire trunk termination at the central office. Return loss circles based on 500-type subset impedance.

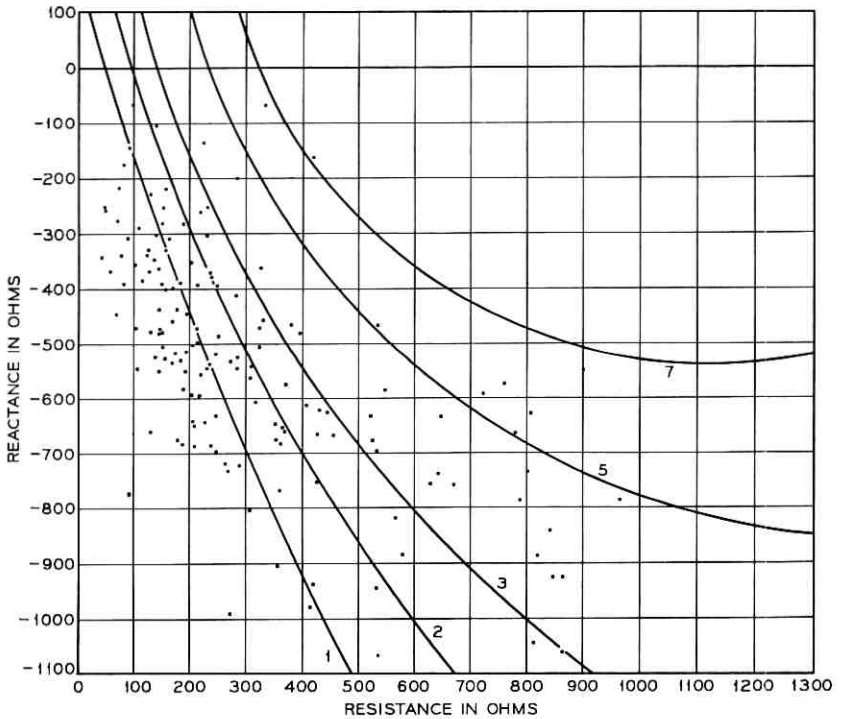

Fig. 37 — Loaded loop input impedances at 3 kHz measured from station set with simulated four-wire trunk termination at the central office. Return loss circles based on 500-type subset impedance.

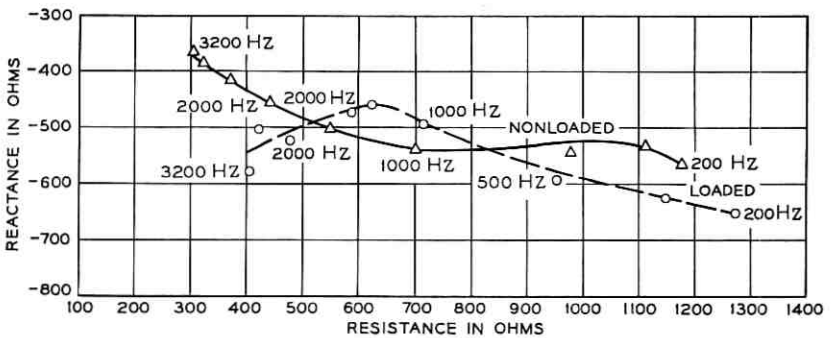

Fig. 38 — Mean value of loop input impedance from station set with simulated four-wire trunk termination at the central office.

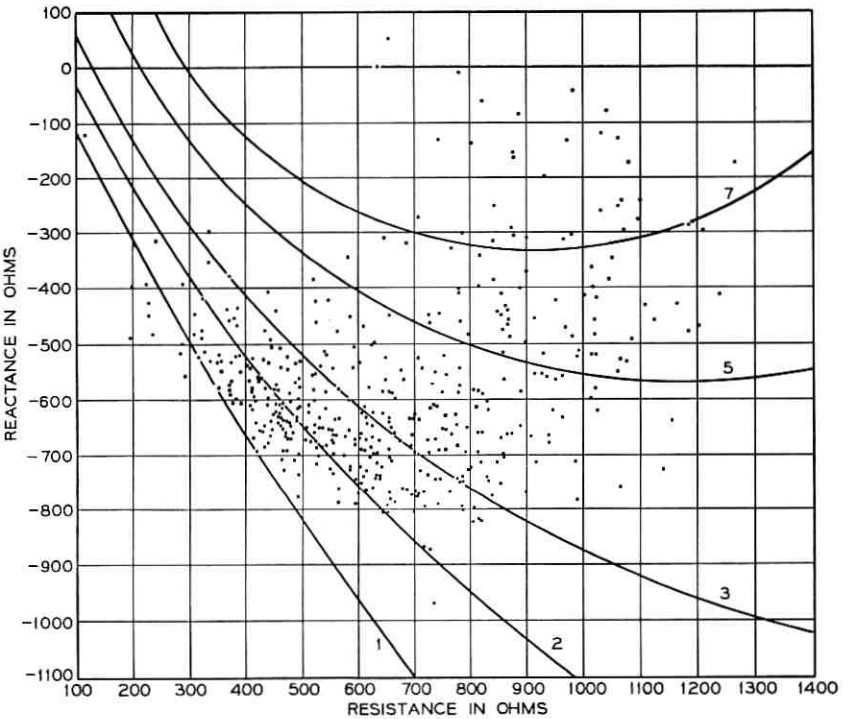

Fig. 39 — Input impedance of all loops at 1 kHz measured from station set with a simulated intraoffice circuit termination at the central office. Return loss circles based on 500-type subset impedance.

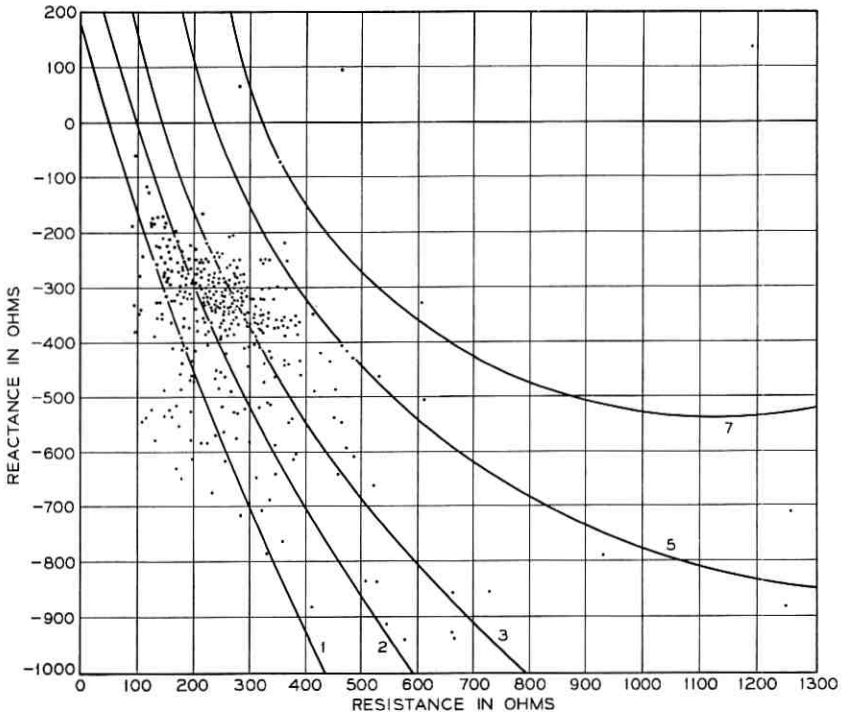

Fig. 40 — Input impedance of all loops at 3 kHz measured from station set with a simulated intraoffice circuit termination at the central office. Return loss circles based on 500-type subset impedance.

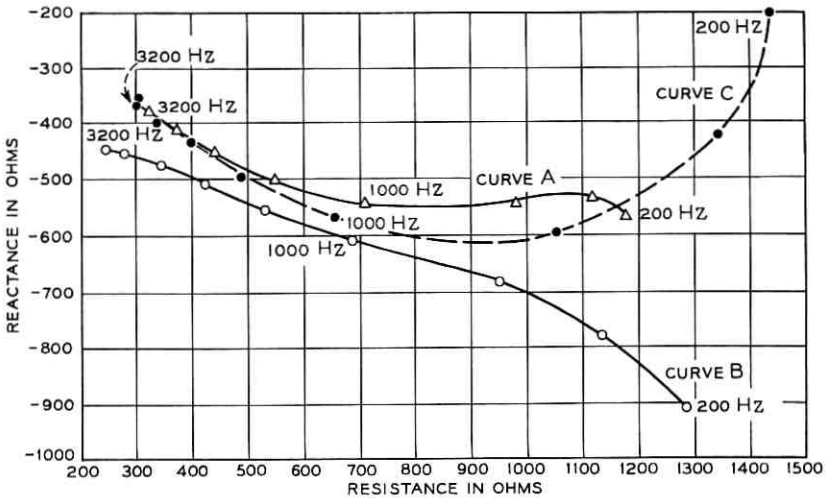

Fig. 41 — Mean input impedance of nonloaded loops measured from station set with: Curve A—simulated four-wire trunk 900 ohms + 2 μF central office termination; Curve B—simulated two-wire trunk 22-gauge H-88 cable central office termination; and Curve C—simulated intraoffice calls condition.

Radio-Relay Antenna Pointing for Controlled Interference with Geostationary Satellites

By C. W. LUNDGREN and A. S. MAY

(Manuscript received July 23, 1969)

We present analytical methods (i) *for calculating microwave radio refraction for negative and positive initial ray angles accounting for station height and* (ii) *for determining refraction-corrected ranges of antenna pointing azimuth within which mutual interference with geostationary satellites in shared frequency bands is likely.*

I. INTRODUCTION

When radio-relay and communication satellite systems share frequency bands, as they do at 4 and 6 GHz, it is necessary to impose restrictions on both systems so that interference is not excessive. The CCIR (International Radio Consultative Committee) recommends that radio-relay antennas maintain a specified angular separation with respect to the geostationary (stationary equatorial) orbit or, where this is not practicable, the application of power limitations to terrestrial radio transmitters involving reception at the satellite. While the above restrictions protect satellites, designers of terrestrial systems should be aware of possible interference into radio-relay systems from satellite radiation arriving at low elevation angles and close to the on-beam directions of receiving antennas. Because the dielectric constant of the earth's atmosphere varies with altitude, the radio-relay beam is not straight, and atmospheric refraction must be considered when computing the directions of radio beams for which the restrictions apply.

1.1 *Simplified Exposure Model*

Figure 1 introduces the geometry of the problem and illustrates significant trends and limits. Radio-relay site P located at North Latitude φ degrees is shown as viewed from above the earth. An arc of the geostationary satellite orbit is also shown. The orbit longitude of point

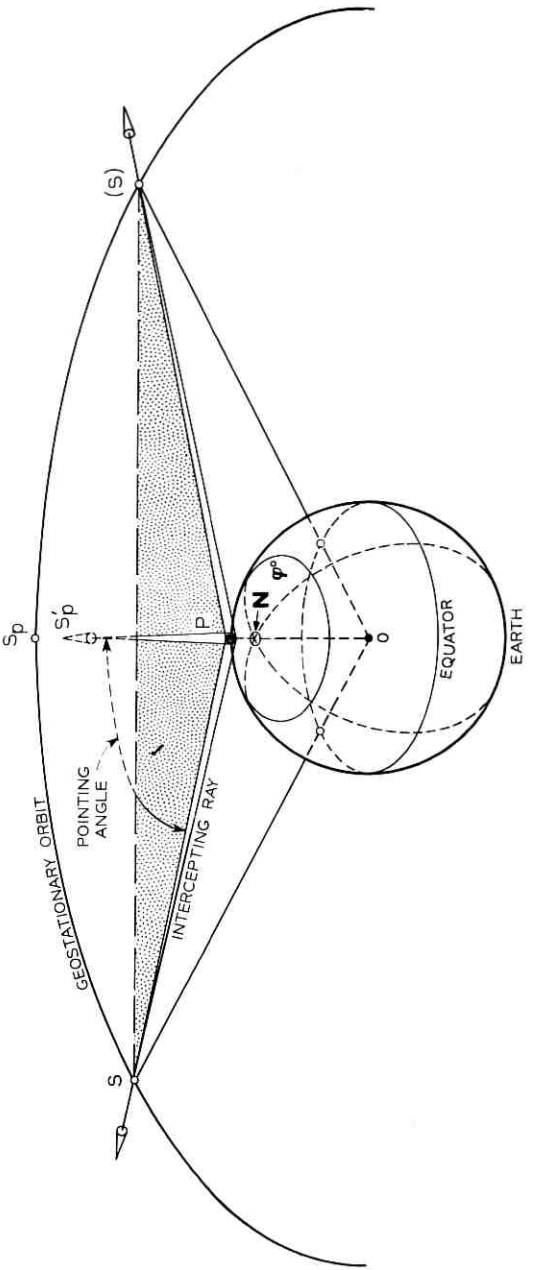

Fig. 1 — Geometry relating unrefracted horizontal radio-relay beams and the geostationary orbit.

S_p and the earth longitude of site P coincide numerically (that is, S_p is in the direction of true south, $\mathbf{PS}'_p$, observed from P).

Two radio rays from P intercept the orbit symmetrically at points S. The controlling geometric relationships are based upon triangles formed by points S, P, and geocenter O. Angle O–P–S is determined by the elevation angle of the radio-relay antenna including ray bending due to atmospheric refraction. A special case is depicted in Fig. 1, wherein unbent intercepting rays **PS** and also ray $\mathbf{PS}'_p$ are assumed to lie in the local horizontal plane at P and remain tangent to the earth sphere at P. Thus, angles S–P–S'_p are always antenna pointing angles to orbit interception referred to "true south."

It is instructive to visualize the relationship between latitude φ at P and the location of orbit intercepts S, for a given fixed triangle OPS. As points S approach S_p, the constraints imposed above require that the station latitude φ approach a maximum latitude "visible" to the orbit. The resulting single pointing direction to orbit intercept is due south (from P to S_p). Conversely, the maximum separation between points S and S_p obtains when φ is zero (for site P located on the equator). Since both intercepting rays are tangent to the equator at P, the limiting pointing directions are due west and due east.

Note that a rotation in azimuth of the antenna (about the local vertical axis at P) between known orbit-intercept directions results in radio rays **PS** which fall below the orbit as viewed from P; rotations beyond these "critical azimuths" result in rays above the orbit.

1.2 *Computations*

Given the latitude and elevation angle of a microwave radio-relay antenna, one can calculate the pointing azimuth for which, neglecting refraction, the main beam axis intercepts the geostationary orbit. This calculation is repeated to produce screening charts like Fig. 2. Such charts are adequate for quickly determining a hazard condition, but often the true critical azimuths must be approached closely while maintaining tolerable interference levels.

A graphical procedure proposed for the convenience of system planners provides pointing angle estimates for most stations when caution is exercised in those steps accounting for atmospheric refraction.[1] An analytical technique adaptable to machine calculation is also required for rapid, accurate screening of large numbers of existing and proposed radio-relay sites for potential interference exposures. Precise evaluations are required for cases of unavoidable exposure.

The method described in following sections can be used by the system

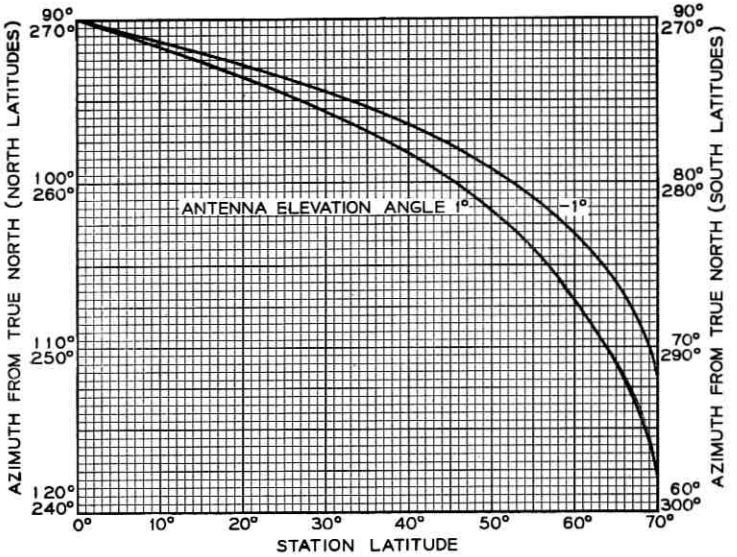

Fig. 2 — Screening chart, neglecting refraction.

planner to determine critical azimuths and the ranges of azimuths to be avoided (accounting for atmospheric refraction) for prescribed minimum angular separations between main beam axes and the geostationary orbit.

The Central Radio Propagation Laboratory (CRPL) Exponential Reference Atmosphere is adopted for the generation of microwave radio refraction curves by accounting for station heights and negative antenna elevation angles, for several representative refractive indexes.[2] Earlier extrapolations are based upon upper limits of the total bending associated with assumed earth-grazing rays.[3,4]

A refraction anomaly arising from a temperature inversion, storm, ducting, or other departure from an assumed representative radial ray-bending model precludes an absolutely confident evaluation of any given exposure at a given time. These phenomena are usually localized. However, the intent of these computations is to protect the geostationary orbit against continuous interference arising simultaneously from a large number of terrestrial systems.

Following a development of the refraction model, we derive the critical pointing azimuth corresponding to orbit intercept. The geocentered longitude displacement between the radio-relay site and the point where the refracted beam intercepts the orbit is next determined by spherical

geometry. Then the apparent slope* of the geometric orbit trace, as if viewed unbent by refraction from the station, is obtained corrected for aspect (also dependent upon antenna elevation angle and concomitant refraction). Using the refraction data converted to geometric elevation angles (as without ray bending or obstruction), the geometric orbit is adjusted to the apparent position and shape that would be observed at the radio site. This apparent orbit is hereafter termed the refracted orbit. Subsequent sections describe the determination of azimuth zones to be avoided for prescribed beam—orbit separations and the permissible transmitted power. Appendices are included to: provide means for estimating initial ray angles from commonly available radio-relay information when actual antenna angles are unknown; solve by manual calculation a representative numerical example, giving the applicable equations for each step; and verify governing equations using a different analytical approach.

II. DETERMINATION OF REFRACTION FOR POSITIVE AND NEGATIVE ANTENNA ELEVATION ANGLES

2.1 *Ray Tracing Equations*

In the following equations θ_o is the initial angle† of a ray as it leaves the earth's surface and τ is the total refractive bending corresponding to θ_o. Figure 3 illustrates this relationship and shows the geometric director with elevation angle ϵ to an intercept with the geostationary orbit at S. Also shown on an arc through S centered at P is the apparent position of the intercept S_a and a horizontal reference, zero-elevation point A. The latter relationships are used in a subsequent section to describe a method for constructing the refracted orbit.

Angle ϵ is approximately, but in general not identical to, $\theta_0 - \tau$‡. However, this approximation is reasonable for rays between terrestrial antennas and geostationary satellites well beyond the earth's atmosphere when relatively small effects of parallax associated with the controlling portion of the atmosphere near the earth's surface are neglected.[5]

Figure 4 depicts a ray entering the earth's atmosphere and the result-

* The first derivative with respect to pointing azimuth (azimuth-elevation plot of the orbit) is more completely defined in Section V.

† θ_o is generally used in ray-tracing equations to denote the initial ray elevation angle with respect to the local horizontal and is synonymous with α_o used in subsequent sections to denote a radio-relay antenna beam elevation angle (namely, α in Figs. 6).

‡ Calculations using equation (5) show that the actual ray angle with respect to the geocenter differs from that resulting from the assumption $\epsilon = \theta_o - \tau$ by approximately 1.5 minutes of arc for limiting conditions used herein.

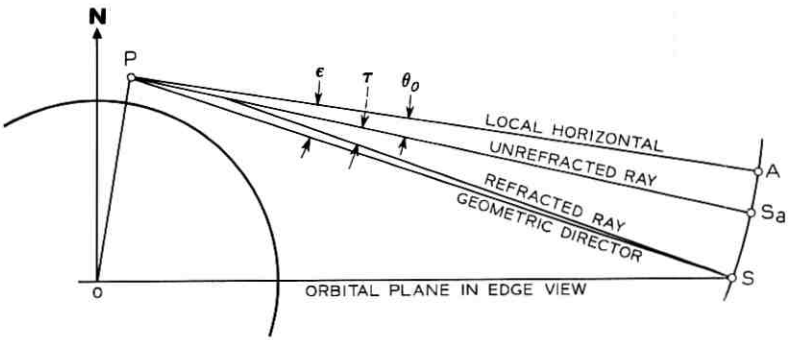

Fig. 3 — Refracted ray.

ing path of the ray due to the refractive gradient. Depending upon the angle of arrival of the ray when it enters the atmosphere, the refraction causes it to intercept the earth, graze the earth, or become tangent to a unique earth-centered sphere at some height above the surface. If the ray does not intercept the earth it continues out into space again, being subjected to approximately the same refraction in exit as it encountered upon entering.

At any point on the ray trace, the angle the ray makes with the local horizontal at that point is denoted by θ. The angle between the tangents to the ray at any two points (a, b) is a measure of the refractive bending between them and is denoted by τ(a, b).

The following presentation of refraction is based largely upon equations given in Refs. 2, 6, and 7 for zero or positive initial ray angles and as interpreted by the authors for application to negative angles. Appropriate uses of the equations and their application to the problem are

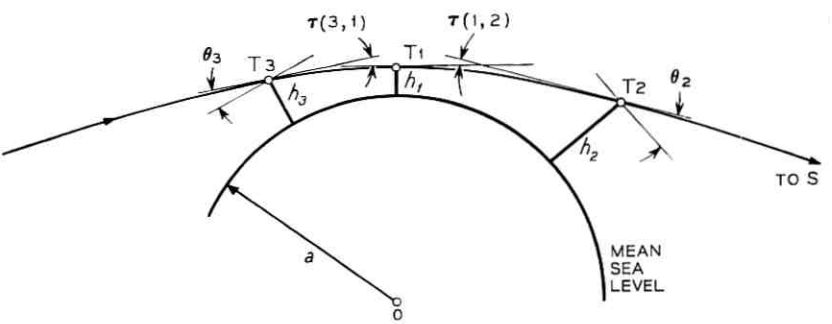

Fig. 4 — Maximally refracted nongrazing ray.

explained, but the reader is directed to the references for complete derivations and limitations including sensitivities to the microwave frequencies involved.

The radio refractive index n varies with atmospheric pressure, relative humidity, saturation vapor pressure and temperature. The lower limit for n is unity—no atmosphere, while the upper limit is determined by local climatic conditions. For the southeastern section of the United States, a typical range of n at mean sea level is 1.00025 to 1.0004. In equations involving refraction it is convenient to express refractivity as N where $N = (n - 1) \times 10^6$ or, for this case, N values are 250 and 400 (N units). The average decay of N is approximately exponential with height and the difference in N between the surface and a height of 1 km above the surface is given by[2]

$$\triangle N = -7.32 \exp (0.005577 N_s). \tag{1}$$

The subscript s denotes N at the surface. The decay constant with height is expressed by[2]

$$C_s = \ln \frac{N_s}{N_s + \triangle N}. \tag{2}$$

N at any height h (kilometers above the radio site surface elevation) is[2]

$$N_h = N_s \exp [-C_s(h - h_s)], \tag{3}$$

where h_s is the surface elevation above mean sea level corresponding to N_s.

N_s is sensitive to local elevations and hence, charts of N_s for mountainous regions are difficult to use because the N_s contours are irregular and closely spaced. Therefore, obtaining an appropriate value of N_s for use in equations (1), (2), and (3) for a particular site is often difficult. However, charts are available giving N_s reduced to mean sea level equivalents N_o which, in effect, reduce the height-dependent N_s values to a common base.[8] Since charts of N_o are more easily interpreted and a single value of N_o usually applies over a large geographical area, they are used in this paper.

N_o and $N_s(h)$ are related[6] by $N_o = N_s \exp (-h/7)$. Conversely, for a given value of N_o, N_s for surface height h_s above mean sea level is determined from the expression

$$N_s = N_o \exp (-h_s/7), \qquad h_s \geqq 0. \tag{4}$$

The N_s value obtained from equation (4) is used in equations (1), (2),

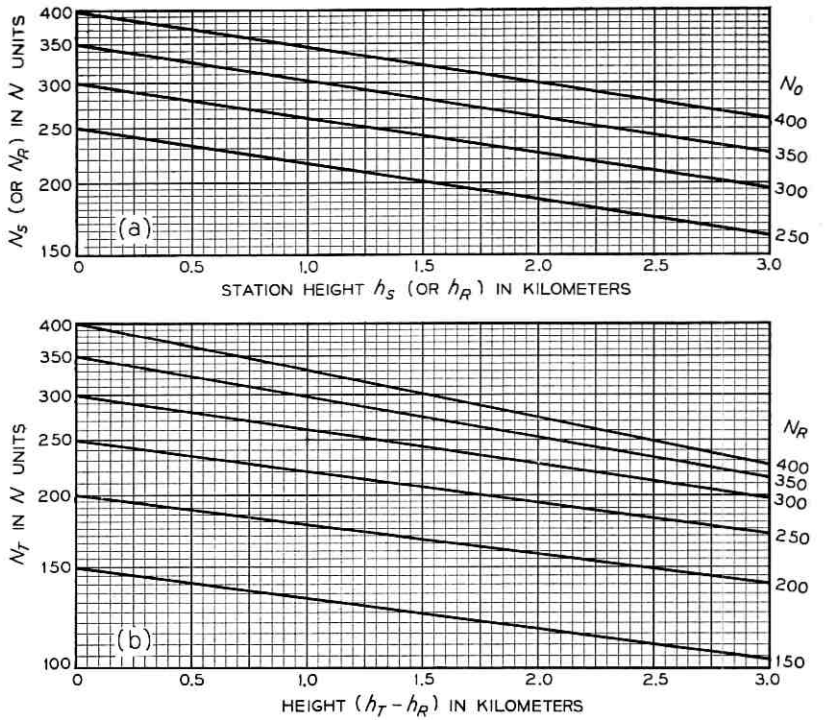

Fig. 5(a) — Radio refractivity conversion — N_0 to N_s (or N_R).
Fig. 5(b) — Radio refractivity conversion — N_R to N_T.

and (3). Figure 5a was generated using equation (4) and, for manual calculations, is convenient for determining N_s.

The conversion between N_0 and N_s above considers only the dry air effects of temperature and pressure related to the difference in elevation between mean sea level and the station surface height, whereas constants of the expression given for $\triangle N$ account for all terms contributing to a change in refractivity of the atmosphere with elevation above the surface height.

A relationship exists between angle θ_0 at the point of origin of the ray and θ at any other point on the ray trace which is expressed by

$$(1 + N_s \times 10^{-6})(a + h_s) \cos \theta_0 = (1 + N_h \times 10^{-6})(a + h) \cos \theta = C, \quad (5)$$

where C is a constant and a is the earth's radius at mean sea level (in kilometers). Thus, angle θ for any point on the ray trace is determined.

The incremental bending of the ray between closely spaced points on the trace is given by

$$\tau(1, 2) = -\int_{n_1}^{n_2} \frac{dn}{n} \cot \theta. \tag{6}$$

As suggested by Schulkin, the term $1/n$ is taken as unity with an error of less than 0.0001 in the computed refraction and, for an iterative solution, equation (6) is expressed as[7]

$$\tau(1, 2) = -\int_{\Delta n_1}^{\Delta n_2} \cot \theta \, d\Delta n. \tag{7}$$

Shulkin also shows[7] that equation (7) is approximated by $(\Delta n1 - \Delta n2)/\theta_m$ where Δn is $n - 1$ and θ_m is $(\theta 1 + \theta 2)/2$. Hence the incremental bending between two closely spaced points on the ray trace is expressed by

$$\tau(1, 2) = \frac{2(N_1 - N_2) \times 10^{-6}}{\theta_1 + \theta_2} \text{ rad.}, \quad 0 \leqq \theta_o \leqq 10°, \tag{8}$$

where N_1 and N_2, θ_1 and θ_2 are the N values and ray angles (in radians), respectively, at the closely spaced points.

2.2 *A Method for Calculating Refraction*

The following paragraphs describe a procedure for calculating refraction curves of the form in Figs. 6 for any values of N_o which is also applicable for the direct calculation of refraction corrections.

2.2.1 *Zero and Positive Initial Angles* $(+\theta_o)$

First assume a value of N_o, a station height h_s, and an initial ray angle θ_0. Equation (4) is used to determine N_s for height h_s and equations (1), (2), and (3) to determine N for a height h, where $h = h_s + \Delta h$*. Equations (5) and (8) then give the ray angle θ at the incremental height and the bending τ in the first increment. For each successive increment of height, the previously solved-for values of N and θ become the initial values for equations (3), (5), and (8). The values of τ for each iteration are accumulated to give the total bending for h_s and θ_o. Repeating the above procedure with other values of θ_o results in data points to be used in plotting the refraction curve for the assumed station height h_s. A complete repetition of the above, beginning with

* Incremental heights must be small in the lower atmosphere where n changes rapidly but may increase at the higher elevations. However, for the generation of Fig. 6, a constant increment of 0.25 km was used to a height of 90 km.

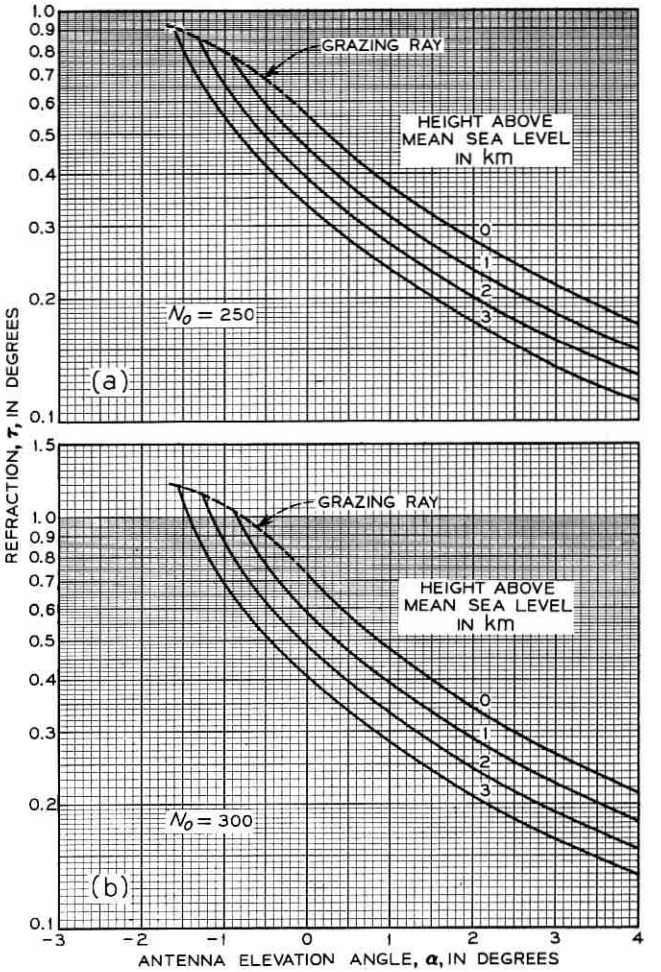

Fig. 6 — Refraction versus antenna elevation angle for (a) $N_s = 250$, and (b) $N_s = 300$.

other station heights, results in a family of curves (namely, h_s) for zero and positive initial ray angles.

2.2.2 *Negative Initial Angles* $(-\theta_o)$

The calculation of refraction for negative initial ray angles requires a modification of the technique used for positive angles. Note in Fig. 4 that the bending of the ray from T3$(h_3, -\theta_3)$ to T1 is the same as that

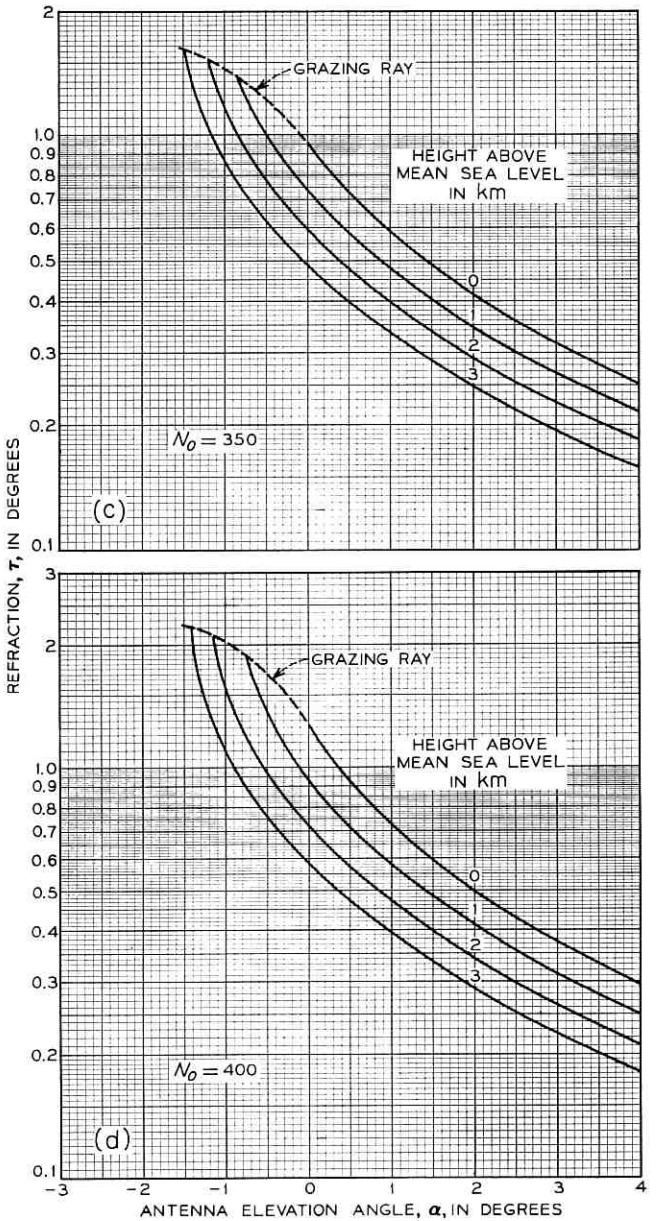

Fig. 6 — Refraction versus antenna elevation angle for (c) $N_o = 350$, and (d) $N_o = 400$.

of the ray from T1 (h_1 , $\theta_o = 0$) to T3. The latter is determined utilizing the equations for positive angles but in a slightly different manner.

First, assume a grazing ray ($h = 0$, $\theta_o = 0$) and, using the iterative method previously described, compute the bending τ and angle θ at each of the specific elevations h_s desired for the chart, that is, in the case of Figs. 6: 1, 2, and 3 km. The value of τ for each height is then added to the maximum τ determined for the same height (when θ_o is zero as found in Section 2.2.1) to obtain the total bending for a ray with the initial angle, $-\theta_o$. In Fig. 4 this is τ(T3, T1) plus τ(T1, S). Repeating the above but beginning with $h = 1$, 2, and so on, provides the data needed to extend the curves for positive initial angles into the negative range.

Note in Figs. 6 that all refraction curves terminate on a dotted extension of the zero-elevation ray representing a maximally refracted grazing ray. For a given height with an initial angle more negative than that represented by the point of termination, the ray is intercepted by the earth.

2.2.3 *Direct Machine Calculation*

For machine calculations it is desirable to compute the refractive bending directly without the use of refraction tables. For zero and positive initial ray angles, the calculation is straightforward as is described in Section 2.2.1. However, for a ray originating at a specific height with a negative initial angle θ_o , it is first necessary to determine the height of the ray where θ is zero (h_1 at T1 in Fig. 4). A variation of equation (5) permits this determination. If the right side of equation (5) represents point T1 in Fig. 4 and the left the initial station, say T3 at height h_3 , then

$$(1 + N_{T3} \times 10^{-6})(a + h_{T3}) \cos \theta_o = (1 + N_{T1} \times 10^{-6})(a + h_{T1}) = C. \quad (9)$$

The variables for the left side of equation (9) are all determined and evaluation yields the constant C. Then, values are assumed for h_{T1} beginning with zero, N_{T1} determined, and the right side evaluated until the result is equal to C. After h_{T1} is determined, τ(T3, T1) and τ(T1, S) are calculated as described in Section 2.2.2 and added numerically to give the total refractive bending. However, when h_{T1} is assumed to be 0 for the initial iteration and the right side is larger numerically than C, the beam intercepts the earth and a solution with angle θ_o is not possible. In such cases it is often desirable to determine the initial grazing-ray angle to the mean sea level horizon, which is accomplished by incrementing θ_o upward until the equation is satisfied.

2.2.4 *Angle from the Transmitting Site to the Radio Horizon*

The radio horizon, accounting for refraction and average local terrain, is determined by a variation of equation (9). Assuming that the height of a receiving station represents the average terrain height for some distance beyond, the initial angle α_H ($= \theta_o$) of a ray which just grazes the terrain represents the angle to the radio horizon as seen from an elevated transmitting site. Letting the left side of equation (9) represent the transmitting location, the right side the receiving location (where θ is zero), and solving for α_H yields

$$\alpha_H = \cos^{-1}\left[\frac{(1 + N_R \times 10^{-6})(a + h_R)}{(1 + N_T \times 10^{-6})(a + h_T)}\right] \text{ rad}, \qquad h_T \geqq h_R\ ,\ \alpha_H \leqq 0, \tag{10}$$

where the subscripts $_R$ and $_T$ refer to the receiving and transmitting stations, respectively. N_R is obtained by equation (4) and N_T by equations (1), (2), and (3), substituting N_R and h_R for N_s and h_s, and h_T for h. (For manual calculations, N_R and N_T may be determined from Figs. 5. First enter Fig. 5(a) with h_R, N_o, and read N_R from the ordinate. Then enter Fig. 5(b) with $(h_T - h_R)$, N_R and read N_T from the ordinate.)

2.2.5 *Refractive Index Limits*

We now illustrate the importance of including the effects of refraction in orbital computations relating to terrestrial radio-relay systems. Table I (reflecting the use of equations and techniques discussed in subsequent sections) demonstrates parametrically the sensitivity to radio refractivity of the orbit-intercept pointing azimuth and the computed terrestrial transmitter power limitation. The 8 dB variation in

TABLE I—INFLUENCE OF REFRACTION

Station Statistics	Assumed Values		
Path azimuth Station latitude Station elevation Antenna elevation angle α_o	103.5 degrees from true north 55.0° north mean sea level 0 degrees		
Parametric results	Computed values		
Radio refractivity N_s, N units	0	250	400
Geometric elevation angle, ϵ_o, degrees Critical azimuth (from north), degrees Maximum transmitter power, dBW	0 102.6 47	−0.555 101.76 50.7	−1.27 100.7 55

the power shown in the table indicates clearly that refraction must be included in calculations and that limits for refractivity should be carefully considered.

World charts of N_o in Ref. 8 indicate appropriate limits of N_o are 250 and 400. At specific locations and for short time intervals, the index may not fall within these limits. However, localized conditions will not affect large numbers of stations at any given time and a wider range of N_o would unnecessarily broaden the restrictive zone for radio-relay systems.

We suggest that the above limits be adopted for standardized calculations. For a specific case where an antenna pointing angle is close to the orbit and transmitter power limitations are restrictive, refractive limits applicable to that locale should be used.

2.2.6 *Adjustment of Computed Geometric Orbit Traces for Atmospheric Refraction*

For many solutions, particularly those involving graphical procedures, it is desirable to "elevate" a computed geometric orbit trace to its apparent (refracted) position and shape. Such an adjustment yields a presentation permitting a given, arbitrarily-shaped radio beam power profile to be related unrefracted and hence undistorted to the easily obtained configuration of the refracted orbit. Figures 7 are charts to enable this manipulation, produced from Figs. 6 by plotting $(\alpha - \tau_\alpha)$ versus τ_α. A method for using these charts is given in Section VI.

III. DETERMINATION OF THE POINTING AZIMUTH TO ORBIT INTERCEPT

Recall that the elevation angle (with respect to local horizontal) of the geometric director shown in Fig. 3 is denoted by ϵ. Hence, station geometric elevation angle ϵ_o may be replaced by $\alpha_o - \tau_{\alpha o}$ where α_o is the initial antenna beam elevation angle and $\tau_{\alpha o}$ is the corresponding refraction correction. (A method for determining α_o is given in Appendix A.)

Note that available information for established radio-relay routes in the United States giving antenna elevations, path distances, and antenna elevation angles is generally expressed in units of feet, statute miles, and degrees, respectively; it is necessary to convert these quantities into kilometers and radians for use in many of the expressions which follow.

Inspection of Fig. 8 shows that the azimuth displacement from the meridian through a station located at P to an intercept with the geo-

stationary orbit is identical to angle A of spherical triangle PES'. From laws for right spherical triangles

$$\cos A = \frac{\tan \varphi}{\tan \beta}, \tag{11}$$

where φ is the latitude at station P, and β is the arc equivalent of angle O of plane triangle OPS. Note that β is numerically equivalent to the maximum visible latitude for assumed radio-relay antenna beam elevation angle α_o and concomitant total ray-bending angle $\tau_{\alpha o}$.

Angle β is determined from triangle OPS using the Law of Sines. This triangle is redrawn in Fig. 9, where

$$\frac{\sin \Omega}{a} = \frac{\sin (\pi/2 + \epsilon_o)}{R},$$

$$\Omega = \sin^{-1} (K^{-1} \cos \epsilon_o), \tag{12}$$

where ϵ_o is the station geometric elevation angle corresponding to α_o, a is the earth radius, and R is the orbit radius, $R/a = K$. From inspection, $\beta = \pi/2 - \Omega - \epsilon_o$. Substituting for Ω yields

$$\beta = \cos^{-1} (K^{-1} \cos \epsilon_o) - \epsilon_o. \tag{13}$$

Substituting equation (13) for β in equation (11) results in

$$A = \cos^{-1} \left[\frac{\tan \varphi}{\tan [\cos^{-1} (K^{-1} \cos \epsilon_o) - \epsilon_o]} \right]. \tag{14}$$

IV. DETERMINATION OF RELATIVE LONGITUDE BETWEEN SITE AND ORBIT INTERCEPT

The earth longitude displacement between radio-relay site P and suborbital intercept S' in Fig. 8 is side λ of right spherical triangle PES'. From the Law of Sines

$$\frac{\sin \lambda}{\sin A} = \sin \beta,$$

from which:

$$\lambda = \sin^{-1} (\sin A \sin \beta), \tag{15}$$

where β and A are found by equations (13) and (14). Note that when $A = \pi/2$, corresponding to $\varphi = 0$, the maximum visible longitude displacement is β.

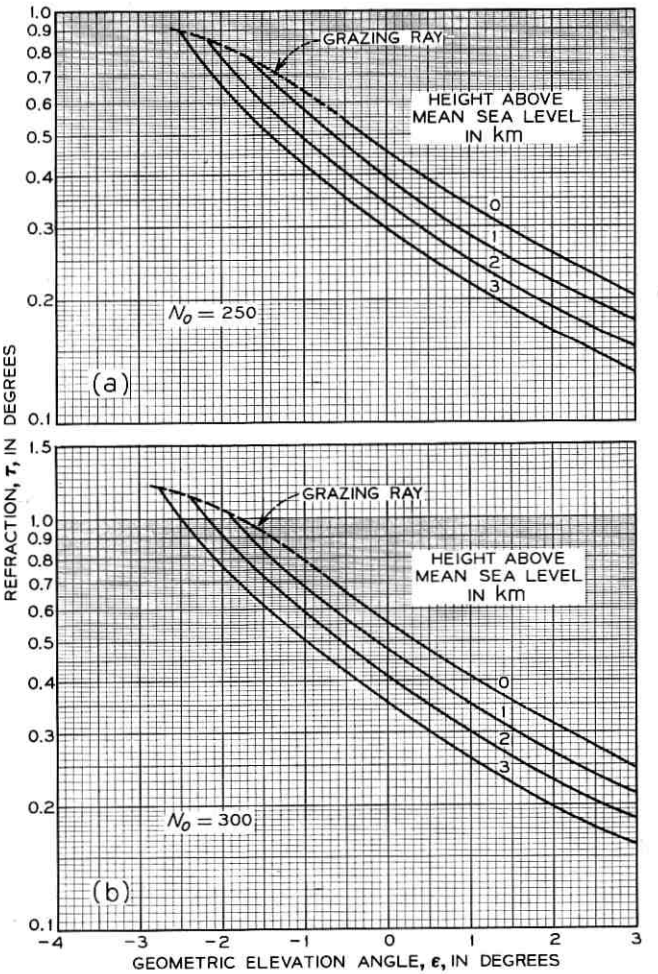

Fig. 7 — Refraction versus geometric elevation angle for (a) $N_0 = 250$, and (b) $N_0 = 300$.

V. GEOMETRIC ORBIT TRACE—CORRECTED FOR ANTENNA ELEVATION AND ATMOSPHERIC REFRACTION

The geostationary orbit and earth's equator are coplanar; hence the orbit near the horizon normally appears to be tilted with respect to the local horizontal plane. Were an equatorial orbit sufficiently distant, as in celestial observations, the angle of tilt would equal the colatitude

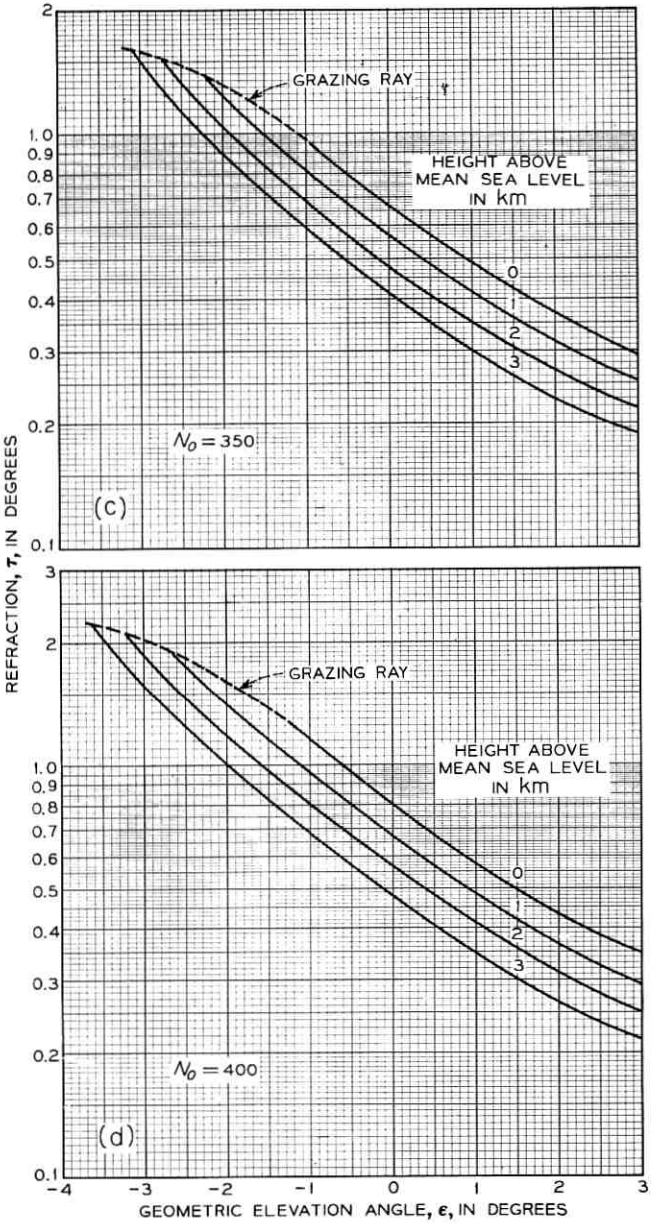

Fig. 7 — Refraction versus geometric elevation angle for (c) $N_0 = 350$, and (d) $N_0 = 400$.

of the observer's position. However, the maximum visibility circle in Fig. 8 is not a great circle, so the angle of tilt is less than the colatitude except where the latter circle crosses the equator. In the following paragraphs, an expression is developed for the tilt of the geostationary orbit as if viewed from a unique location on the visibility circle. This tilt is defined hereafter as the apparent slope of the orbit trace (or first deriva-

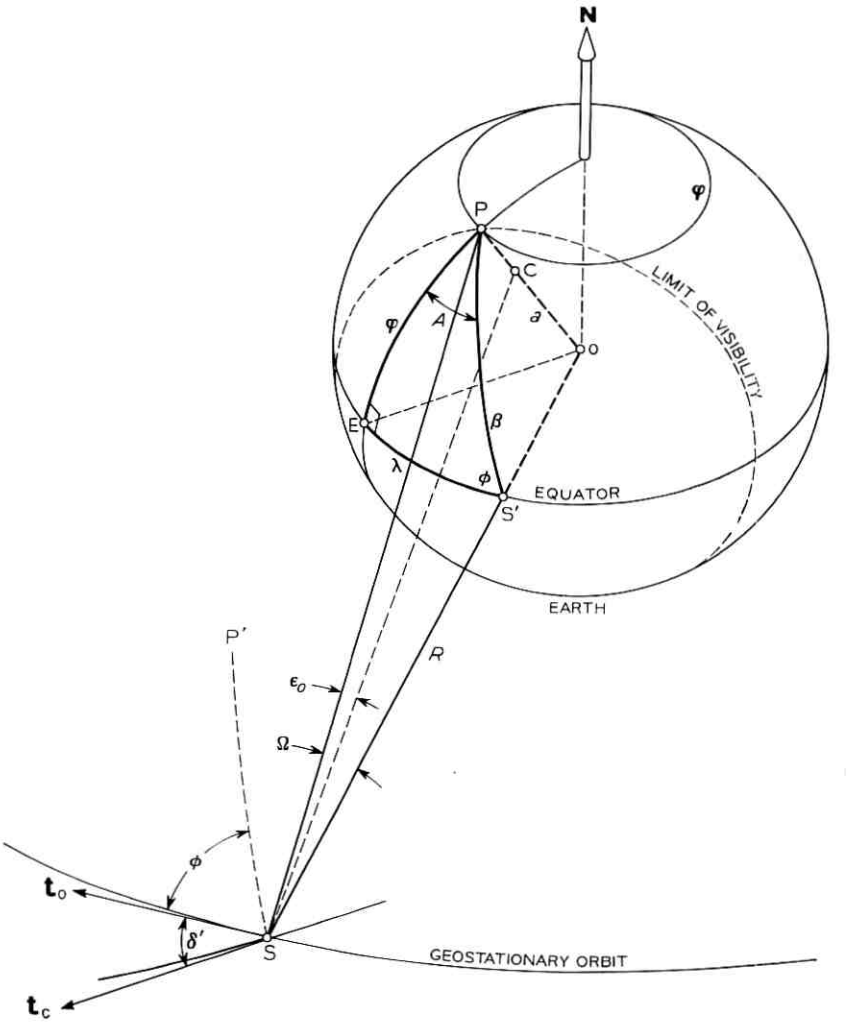

Fig. 8 — Geometry relating site P and geostationary orbit intercept.

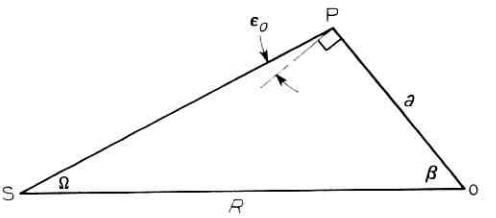

Fig. 9 — Geometry of intercept-determining triangle.

tive of the orbit trace, in elevation, with respect to direction, in azimuth).*

The size of the visibility circle in Fig. 8 depends upon α_o and $\tau_{\alpha o}$; for given site latitude φ, the angular displacement in earth longitude λ between site P and point S where the refracted beam intercepts the geostationary orbit (longitude of suborbital point S′) also depends upon α_o and $\tau_{\alpha o}$. Hence, the slope δ of the geometric orbit trace constructed in Fig. 11 also depends upon α_o and $\tau_{\alpha o}$.

Note in Fig. 8 that the angle between orbit tangent $\mathbf{t}_o$ constructed at S and the local vertical plane at P through S is angle ϕ of spherical triangle PES′. From the Law of Sines

$$\phi = \sin^{-1}\left(\frac{\sin\varphi}{\sin\beta}\right). \tag{16}$$

The complementary angle between orbit tangent $\mathbf{t}_o$ and the plane of a circle generated by radius CS (perpendicular to **OP**) is denoted by $\delta' = \pi/2 - \phi$. Substituting equation (16) for ϕ yields

$$\delta' = \cos^{-1}\left(\frac{\sin\varphi}{\sin\beta}\right). \tag{17}$$

Angle δ' is viewed in true magnitude from O (or S′), but is seen from radio-relay site P as a smaller angle δ when rotated through angle Ω as shown in Fig. 10. Note that $\tan\delta' = y/x$ and $\tan\delta = (y\cos\Omega)/x$, from which

$$\delta = \tan^{-1}(\tan\delta'\cos\Omega). \tag{18}$$

Combining equations (17) and (18) yields

$$\delta = \tan^{-1}\{\tan[\cos^{-1}(\sin\varphi/\sin\beta)]\cos\Omega\}. \tag{19}$$

* The orbit trace is envisioned as the locus of all pointing angles to the orbit, plotted on an azimuth-elevation chart aligned and calibrated according to the location of the observer.

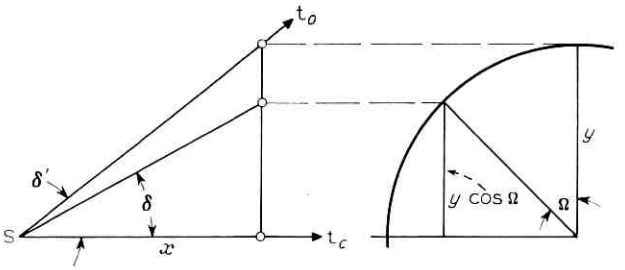

Fig. 10 — Development of geometric orbit slope as observed from site P.

VI. CONSTRUCTION OF THE REFRACTED ORBIT

Figure 11 presents the geometry of the problem viewed from a radio-relay station (similar, but not equivalent to the presentation given in Ref. 1). Reference to Figs. 1 and 3 may assist in interpretation of Fig. 11, wherein points A, S, and S_a correspond to similar points in Fig. 3 and intercept S corresponds to the left (easterly) intercept S in Fig. 1. Origin A is the beam-intercept direction from site P calculated from equation (14), accounting for atmospheric refraction.

Since the orbit is tilted with respect to the local horizontal plane at site P, the elevation angle of the geometric director to the orbit continuously varies as the orbit is scanned in azimuth. The refractive bending of a ray is a function of the geometric angle, so that the position of the apparent, or refracted orbit, is above the geometric orbit and it exhibits a constantly changing slope with respect to the latter.

The bent, refracted orbit is shown through point S_a (apparent position of interception point S with refraction; also, the antenna elevation angle α_o at the azimuth origin). The straight line labeled "geometric orbit" is tangent at S (Fig. 8) to a radial projection of the geostationary orbit on a sphere of radius PS centered at P. The horizontal line shown through S also represents an arc, in edge view of a great circle through S, parallel at S to the local horizontal plane at P, on this same sphere. Hence, the angle δ obtained from equation (19), corrected for refraction while retaining the concept of a constant site latitude φ, is also accurately represented by the apparent slope of the plane-figure geometric orbit trace at the azimuth origin. Figure 11 illustrates the construction of the corresponding refracted orbit trace.

The equation of the linear geometric orbit trace through S is

$$\epsilon' = -\tan(\delta)\triangle A + \alpha_o - \tau_{ao}, \qquad (20)$$

where ϵ' is the elevation angle of the geometric orbit trace corresponding to an arbitrary displacement in azimuth $\triangle A$ from origin A.

Figs. 7 are entered with values of ϵ' derived from equation (20) to obtain total refractive ray-bending angles $\tau_{\epsilon'}$. The refracted orbit trace in Fig. 11 is then constructed by plotting points with coordinates $(A + \triangle A, \alpha')$ and connecting these with a smooth curve, where

$$\alpha' = \epsilon' + \tau_{\epsilon'} . \tag{21}$$

VII. AZIMUTH DISPLACEMENT FROM INTERCEPT TO KEEP THE BEAM CENTER AN ANGULAR SEPARATION v FROM (AND BELOW) THE MINIMALLY REFRACTED ORBIT (N_o MINIMUM)

Figure 11 also illustrates a solution to keeping an angular separation v between the center of a circular beam and the geostationary orbit. The circle centered on S_a and labeled "unrefracted beam" is a cross-section of a conical figure of revolution with apex at antenna site P

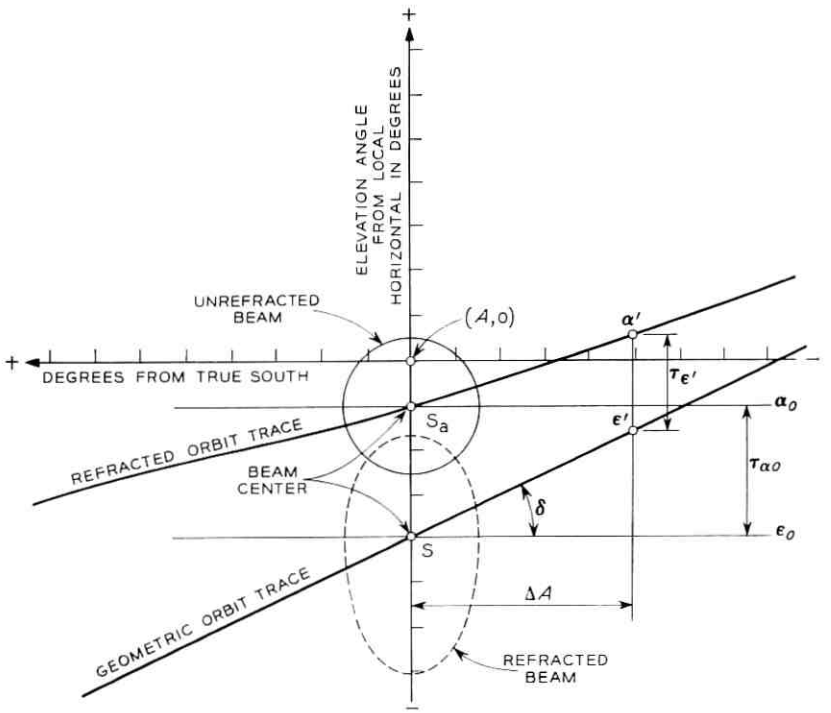

Fig. 11 — Orbit geometry as observed from site P.

and $2v$ included angle (locus of all rays having angle v with respect to the beam center). If elevation angle α_o of the antenna is fixed by radio-relay path parameters, the orbit can be avoided only by an azimuth displacement of the beam. The required angular separation v between the orbit and the beam center results when the latter is moved in azimuth away from intercept until the unrefracted cone is just tangent to the refracted orbit trace. Note that an identical relationship exists between the elongated "refracted beam" and the geometric orbit trace. However, since the former presentation is easier to develop and can readily accommodate unsymmetrical beam cross-sections, it is preferred for following analyses.

Figure 12 represents the analytical solution of an illustrative problem given in Appendix B and is helpful for visualizing subsequent procedures. From inspection

$$\tan \delta = \frac{(\alpha' - \tau_{\alpha'}) - (\alpha_o - \tau_{\alpha o})}{M}, \tag{22}$$

where α' is any point on the refracted orbit corresponding to arbitrary displacement M from an azimuth intercept. Note that in the range of interest M is negative with respect to azimuth $A_{\min}$, obtained from equation (14) using $\epsilon_o = \alpha_o - \tau_{\alpha o}$. Within a few degrees of intercept, the geometric orbit is approximated by a line of constant slope δ. Recall that adjustment of the geometric orbit for refraction results in a refracted orbit trace having a constantly changing slope with azimuth displacement from the intercept. This displacement for a given elevation of the refracted orbit is found by solving equation (22) for M:

$$M = \frac{(\alpha' - \tau_{\alpha'}) - (\alpha_o - \tau_{\alpha o})}{\tan \delta}. \tag{23}$$

Since the refracted orbit slope varies with elevation angle, no direct mathematical solution exists for determining that azimuth displacement providing an angular separation v between the beam center and the refracted orbit, measured in a direction normal to the latter. However, it is closely approximated by determining the slope of an orbit segment including the region of interest. Figure 12 suggests that the refracted orbit slope is everywhere less than the geometric slope. Hence, two judiciously chosen points on the refracted orbit having elevations

$$\alpha 1 = \alpha_o + v \cos \delta, \qquad \alpha 2 = \alpha_o + v,$$

always bracket the appropriate segment. Figures 6 and 12 also show that at these elevation angles the differential refraction is small. This

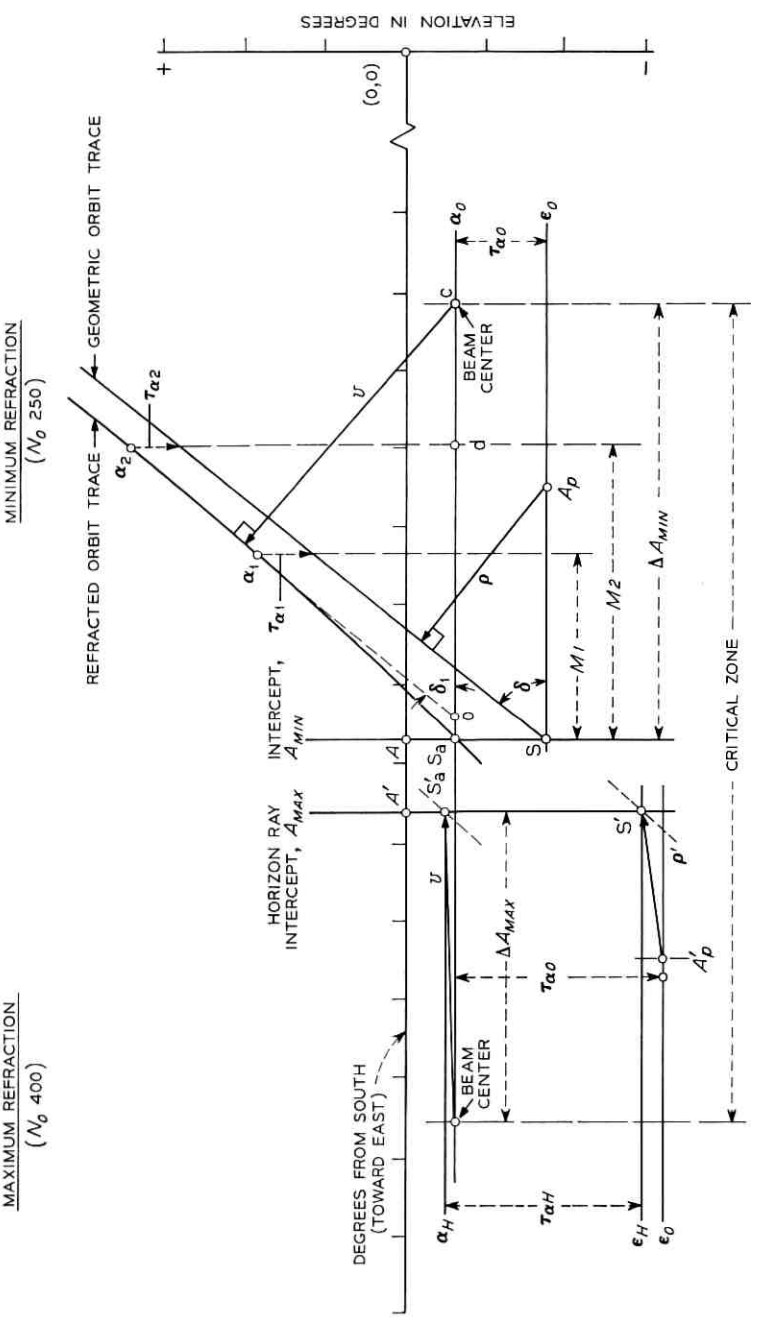

Fig. 12 — Graphical representation of the analytical solution.

supports the assumption that the slope of the small orbit segment in this region is virtually constant.

The refraction corresponding to elevation angles α_o, $\alpha 1$, and $\alpha 2$ are calculated or determined from the charts. Substituting α_o, $\tau_{\alpha o}$, $\alpha 1$, and $\tau_{\alpha 1}$ in equation (23) gives the azimuth displacement $M1$ corresponding to $\alpha 1$. Similarly, $M2$ is determined by substituting α_o, $\tau_{\alpha o}$, $\alpha 2$, and $\tau_{\alpha 2}$. Slope $\delta 1$ between the two chosen points on the refracted orbit is given by

$$\tan \delta 1 = (\alpha 2 - \alpha 1)/(M2 - M1),$$

resulting in

$$\delta 1 = \tan^{-1} [v(1 - \cos \delta)/\triangle M], \tag{24}$$

where $\triangle M = M2 - M1$.

Figure 12 shows that azimuth displacement $\triangle A$ necessary to keep the beam center an angular separation v from the refracted orbit is $S_a c$ (equal to $S_a o$ plus oc).* By inspection, $oc = v/\sin \delta 1$, $od = v/\tan \delta 1$ and $S_a o = M2 - od$. Assembling these into an equation for $\triangle A$ yields

$$\triangle A_{\min} = M2 - v/\tan \delta 1 + v/\sin \delta 1. \tag{25}$$

7.1 *Special Case*

If a ray with initial angle α_o intercepts the earth, a solution is obtained by calculating angle α_H to the radio horizon as is described in Section II. Then, substituting ϵ_H for ϵ_o in equation (14), the orbit intercept for a grazing ray is determined. When solving for the necessary azimuth displacement from this intercept, α_H and $\tau_{\alpha H}$ are substituted for α_o and $\tau_{\alpha o}$ in equation (23). Note, however, that in determining $\alpha 1$ and $\alpha 2$ for substitution in equation (23), the use of ray angle α_o remains valid.

VIII. Azimuth Displacement from Intercept to Keep the Beam Center an Angular Separation v from (and above) the Maximally Refracted Orbit (N_o maximum)

The left side of Fig. 12 shows that the refracted orbit for the case of maximum refraction falls below the radio horizon with azimuth displacement from point $A_{\max}$ obtained from equation (14) using $\epsilon_H = \alpha_H - \tau_{\alpha H}$. Since the earth intercepts all rays below the horizon, they cannot affect the orbit and it is only necessary to displace the beam

---

* Notations such as $S_a c$ represent scalar distances between indicated points in Fig. 12.

center in azimuth sufficiently to maintain angular separation v from the horizon intercept. The required displacement is

$$\triangle A_{\max} = [v^2 - (\alpha_o - \alpha_H)^2]^{\frac{1}{2}}. \quad (26)$$

In Fig. 12, $\alpha_H(-0.25°)$ is the initial angle of a ray which grazes at a receiving station height of 0.4 km and originates at an assumed transmitting height of 0.5 km.

As demonstrated in Appendix B, little increase in the critical zone results if it is assumed that the angle to the radio horizon is equal to the initial ray angle. Therefore, for manual calculations involving small positive values of α_o, these angles are assumed identical so that $\triangle A$ is simply an angular separation v from the orbit intercept determined for α_o. For values of α_o more negative than that for a grazing ray (determined from Figs. 6 or by calculation), it is necessary to determine α_H and the corresponding orbit intercept using equations (10) and (14). Then equation (26) gives the necessary azimuth displacement.

IX. DETERMINATION OF THE CRITICAL ZONES

9.1 *Critical Zones Defined*

The critical zones to be avoided at radio-relay transmitting sites to protect the geostationary orbit are defined:

$$Z_{crit} = A_{\max} + \triangle A_{\max} \quad \text{to} \quad A_{\min} - \triangle A_{\min}$$

(degrees from South), (27)

where values for A and $\triangle A$ for maximum and minimum refraction are obtained as in Sections III, VII, and VIII. These zones are converted to azimuth zones with respect to true north by subtracting the boundaries from 180 degrees for the easterly zone and adding them to 180 degrees for the westerly zone.

Calculations for stations in southern latitudes are identical except that they are referenced to north rather than south. The easterly azimuth zone with respect to true north is obtained directly from equation (27), while the zone boundaries are subtracted from 360 degrees for the westerly zone.

9.2 *Special Case*

At latitudes exceeding $\varphi = \cos^{-1}(K^{-1}\cos\Psi) - \Psi$ it is impossible for the beam's center ray to be below the orbit with angular separation v.* Hence, for such extreme northern latitudes, a single critical zone

* Derivation of this equation is given in Appendix C.

spans south with both easterly and westerly boundaries defined by

$$Z_{crit} = A_{max} + \triangle A_{max} \quad \text{(degrees from south)}. \tag{28}$$

X. DETERMINATION OF MAXIMUM PERMISSIBLE RADIATED POWER

As mentioned in Section I, international agreements exist for maximum radiated powers.[9] For 6-GHz radio-relay transmitters whose antennas point within $v = 2°$ of the geostationary orbit, the power limitation for separations less than 0.5° is 47 dBW relative to the isotropic case (EIRP), increasing 8 dB per degree to a maximum of 55 dBW (occurring at 1.5°). This limitation refers specifically to the center of the major lobe. For this case, only the relationship between the refracted center ray of the beam and the geometric orbit is considered.

If an existing or proposed path direction is between the critical values computed as in Section III for maximum and minimum refraction, the refracted center ray is likely to intercept the orbit for appreciable periods of time. For such cases the maximum power is 47 dBW.

If the path direction for a system in the northern hemisphere is within the critical zone but nearer to south (smaller) than A_{min}, the actual separation ρ illustrated in Fig. 12 is

$$\rho = (A_{min} - A_p) \sin \delta, \tag{29}$$

where A_p is the path direction measured from true south.

Conversely, if A_p exceeds A_{max}, the separation of the beam center from intercept of the geometric orbit and the refracted horizon (Fig. 12) is

$$\rho = [(A_p - A_{max})^2 + | \epsilon_H - \epsilon_o |^2]^{\frac{1}{2}}, \tag{30}$$

where

$$\epsilon_H = \alpha_H - \tau_{\alpha H}, \qquad \epsilon_o = \alpha_o - \tau_{\alpha o}.$$

Note that ϵ_o in equation (30) represents maximum atmospheric refraction.

If a ray having initial angle α_o intercepts the earth, ϵ_o for use in equation (30) is indeterminate. For such special cases, a conservative approximation for the angular separation is $\rho = A_p - A_{max}$.

The maximum permissible effective radiated power P_t dBW (EIRP) is given for separation ρ according to the following criteria:

$$\rho \leqq 0.5°, \qquad P_t = 47;$$

$$0.5° < \rho \leqq 1.5°, \quad P_t = 8(\rho - 0.5) + 47; \tag{31}$$
$$\rho > 1.5°, \quad P_t = 55.$$

XI. CONCLUSIONS

A direct analytical method involving few approximations and assumptions can be used by system planners for calculating refraction-corrected ranges of pointing azimuth for microwave radio-relay antennas within which significant interference with geostationary communication satellites can be expected. Required angular separations between the refracted beam and the geostationary orbit are translated into required azimuth displacements of a radio-relay antenna from that calculated for orbit intercept; conversely, for cases where exposure is unavoidable, means for determining the maximum transmitted powers permitted by international agreement are presented.

Since all analytical expressions including refraction corrections are readily amenable to machine calculation, both speed and improved accuracy in estimating the pointing azimuths are possible. The suggested refractive index limits are believed to be representative for the large majority of exposures and useful for a standardized approach to the problem. For more general applications, where refractive index variations are known to be different, one may use the same principles to generate his own applicable correction curves.

XII. ACKNOWLEDGMENTS

The authors wish to express appreciation to R. C. Harris for motivating this presentation, to G. D. Thayer of the U. S. Department of Commerce, Environmental Science Services Administration, for encouragement in the treatment of refraction for negative angles, and to J. L. Boyette for assistance in mathematical programming.

APPENDIX A

Estimation of Antenna Elevation Angles

The initial beam elevation angle for a radio-relay antenna is determined by the geometry of transmitting and receiving locations, the path length, and refraction. The final alignment based upon transmission measurements, if recorded, is preferred for these calculations. The antenna elevation angle for proposed radio-relay paths can be estimated using the method given below with sufficient accuracy.

Figure 13 depicts radio-relay transmitter T and receiver R at elevations OT and OR above geocenter O, assuming a spherical earth of radius ka. Coefficient k is the ratio of apparent earth radius to true earth radius and accounts for refraction in the lower atmosphere.[10] The path length is represented by arc $\hat{D}$ (great circle length at mean sea level). The transmitting and receiving antenna heights with respect to mean sea level are denoted by h_T and h_R, respectively.

Inspecting Fig. 13, $\phi = \hat{D}/ka$ radians, $C' = 2(ka + h_T) \sin(\phi/2)$, and $\alpha_o = E - \phi/2$ radians. From triangle TT'R and laws for plane triangles

$$\tan E = \frac{(h_R - h_T) \sin(\pi/2 - \phi/2)}{C' + (h_R - h_T) \cos(\pi/2 - \phi/2)}. \tag{32}$$

It can also be shown that

$$\alpha_o = \tan^{-1}\left[\frac{h_R - h_T}{\tan(\hat{D}/2ka) \times (2ka + h_R + h_T)}\right] - \hat{D}/2ka \text{ radians}, \tag{33}$$

where h_R, h_T, $\hat{D}$, and a are expressed in the same units.

Reference 2 provides a formula and a table relating k and N_s. For most calculations, a value of $k = \frac{4}{3}$ results in sufficient accuracy[10] (the

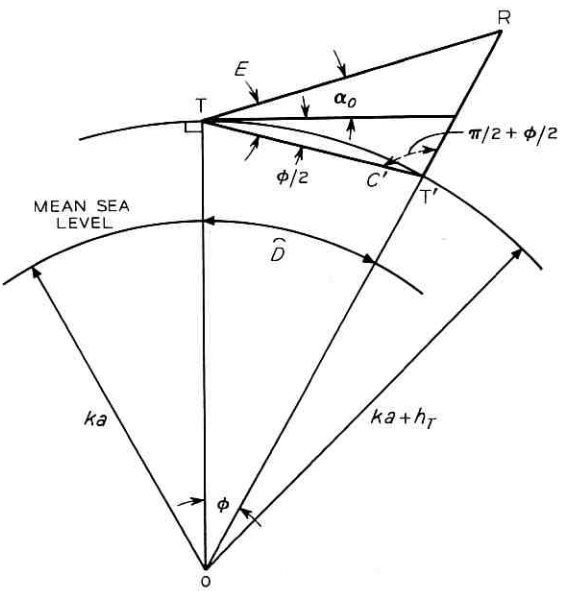

Fig. 13 — Geometry of antenna elevation angle.

geometries of radio-relay systems within limits of normal site elevations require antenna elevation angles which are relatively insensitive to values chosen for k).

APPENDIX B

Illustrative Calculation

Problem Input for a Particular Radio-Relay Site:

Latitude φ	— 38°N,
Transmitter Height h_T	— 0.5 km,
Receiver Height h_R	— 0.4 km,
Path Length $\hat{D}$	— 28 km,
Path Azimuth A_p	— 97.75° with respect to true north, equivalent to 82.25° from south towards east,
N_o Limits	— 250 and 400 N units.

B.1 *Antenna Elevation Angle*

From equation (33) and using $k = \frac{4}{3}$

$$\alpha_o = \tan^{-1}\left[\frac{0.4 - 0.5}{\tan\left[(28/(2.66 \times 6373)](2.66 \times 6373 + 0.4 + 0.5)\right.}\right] - \frac{28}{2.66 \times 6373}$$

$$= -0.00523 \text{ rad, which converts to } -0.3°.$$

B.2 *Geometric Elevation Angle for $N_o = 250$*

From Fig. 6a

$$\epsilon_o = \alpha_o - \tau_{ao} = -0.3 - 0.59 = -0.89°.$$

B.3 *Azimuth Intercept for $N_o = 250$*

From equations (13) and (14)

$$\beta = \cos^{-1}[0.1509 \cos(0.89°)] + 0.89 = 82.2°,$$

$$A_{\min} = \cos^{-1}[\tan(38°)/\tan(82.2°)] = 83.86° \text{ from south.}$$

B.4 *Orbit Slope for $N_o = 250$*

From equations (12) and (19)

$$\Omega = \sin^{-1}[0.1509 \cos(0.89°)] = 8.68°,$$

$$\delta = \tan^{-1}[\tan\{\cos^{-1}[\sin(38°)/\sin(82.2°)]\} \cos(8.68°)] = 51.26°.$$

B.5 *Azimuth Displacement for $v = 2°$ below the Orbit*

From equations (23), (24), (25), and Fig. 6(a)

$$\alpha_o = -0.3°, \qquad \tau_{\alpha o} = 0.59°,$$

$$\alpha 1 = -0.3 + 2\cos(51.26°) = 0.95°, \qquad \tau_{\alpha 1} = 0.35°,$$

$$\alpha 2 = -0.3 + 2 = 1.7°, \qquad \tau_{\alpha 2} = 0.28°,$$

$$M1 = [(0.95 - 0.35) - (-0.3 - 0.59)] \div \tan(51.26°) = 1.19°,$$

$$M2 = [(1.7 - 0.28) - (-0.3 - 0.59)] \div \tan(51.26°) = 1.85°,$$

$$\delta 1 = \tan^{-1}\{2[1 - \cos(51.26°)]/0.66\} = 48.62°,$$

$$\triangle A_{\min} = 1.85 - 2/\tan(48.62°) + 2/\sin(48.62°)$$

$$= 2.76° \text{ toward south from intercept.}$$

B.6 *Horizon Intercept for $N_o = 400$*

From equations (1), (2), (3), (4), (10), (12), (13), (14), and Fig. 6(d)

$$N_R = 400 \exp(-0.4/7) = 377.78 \text{ or from Fig. 5(a)},$$

$$\triangle N = -7.32 \exp(0.005577 \times 377.78) = -60.2,$$

$$C_e = \ln[377.78/(377.78 - 60.2)] = 0.17,$$

$$N_T = 377.78 \exp[-0.17(0.5 - 0.4)] = 371.28 \text{ or from Fig. 5(b)},$$

$$\alpha_H = -\cos^{-1}\left[\frac{1.000377}{1.000371} \times \frac{6373.4}{6373.5}\right] = -0.25°,$$

$$\tau_{\alpha H} = 1.22°,$$

$$\epsilon_H = -0.25 - 1.22 = -1.47°,$$

$$\beta = \cos^{-1}[0.1509\cos(1.47°)] + 1.47 = 82.79°,$$

$$A_{\max} = \cos^{-1}[\tan(38°)/\tan(82.79°)] = 84.33° \text{ from south.}$$

B.7 *Azimuth Displacement for $v = 2°$ from Horizon Intercept*

From equation (26)

$$\triangle A_{\max} = [(2)^2 - (-0.3 + 0.25)^2]^{\frac{1}{2}}$$

$$= 1.99° \text{ toward north from intercept.}$$

B.8 *Critical Zone*

From equation (27)

$$Z_{crit} = 84.33 + 1.99 \quad \text{to} \quad 83.86 - 2.76$$
$$= 86.3° \quad \text{to} \quad 81.1° \text{ from south.}$$

B.9 *Azimuthal Zones*

True-north azimuths:

93.7° to 98.9° easterly, and

266.3° to 261.1° westerly.

The path azimuth of 97.75° falls within the easterly azimuthal zone.

B.10 *Relative Longitude to Suborbit Intercepts*

From equation (15)

$$\lambda_{min} = \sin^{-1} [\sin (83.86°) \sin (82.2°)] = 80.04°,$$

$$\lambda_{max} = \sin^{-1} [\sin (84.33°) \sin (82.79°)] = 80.82°.$$

B.11 *Maximum Permissible Radiated Power*

Comparing the path direction of 82.25° from south with the critical zone found in Section B.8 and with A_{min} and A_{max} found in Sections B.3 and B.6 reveals that equations (29) and (31) are appropriate for calculating the angular separation and permissible power:

$$\rho = (83.86 - 82.25) \sin (51.26°) = 1.23°,$$

$$P_t = 8(1.23 - 0.5) + 47 = 52.8 \text{ dBW} \quad \text{at} \quad 6 \text{ GHz}.$$

B.11.1 *Alternate Power Calculation*

An alternate calculation is indicated when the path direction is further from south than A_{max}. Assume that A_p is, instead, 85.25° from south (left side of Fig. 12). The appropriate equations are now (30) and (31). A value of ϵ_o for maximum refraction is required for equation (30). From Section B.5, Fig. 6(d), equations (30) and (31)

$$\alpha_o = -0.3°,$$

$$\tau_{ao} = 1.3°,$$

$$\epsilon_o = -1.6°,$$

$$\rho = [(85.25 - 84.33)^2 + (-1.47 + 1.6)^2]^{\frac{1}{2}} = 0.93°,$$

$$P_t = 8(0.93 - 0.5) + 47 = 50.4 \text{ dBW at 6 GHz}.$$

APPENDIX C

Alternate Derivations of Basic Equations and Critical Latitudes

All following relationships are expressed in terms of latitude φ of radio-relay site P_v shown in Fig. 14 (a constant), and the maximum latitude for which the refracted geostationary orbit is visible to the antenna (a constant elevation angle).

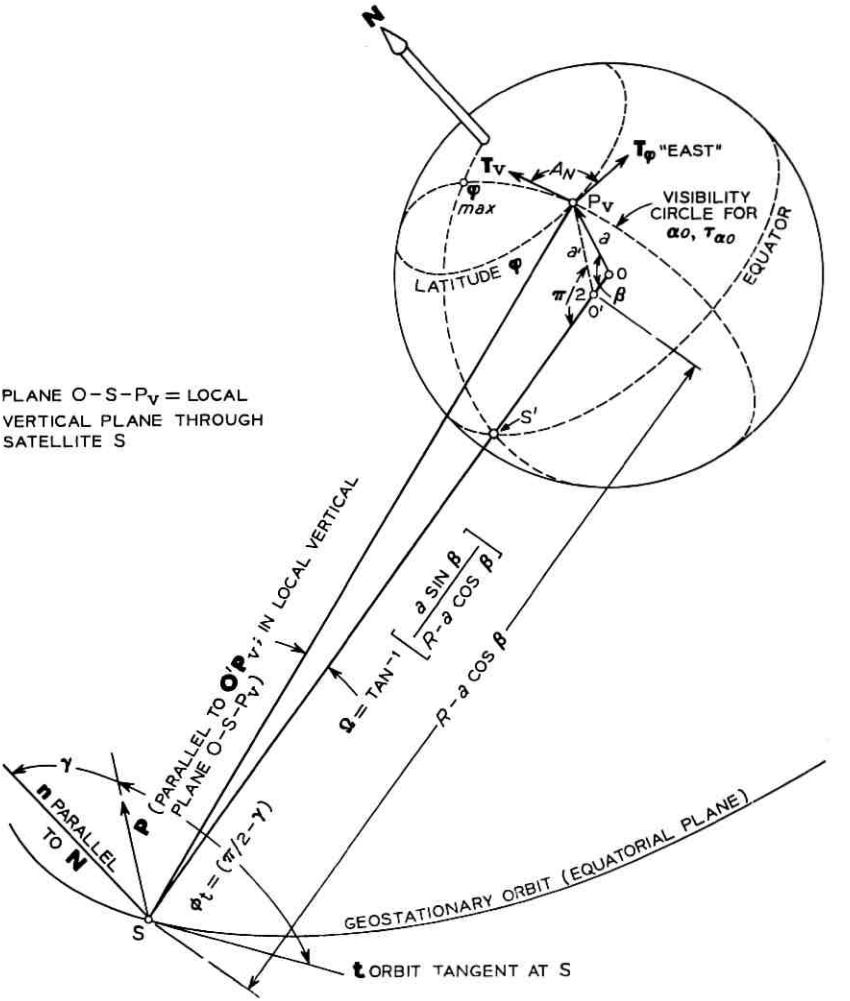

Fig. 14 — Angle ϕ_t between orbit tangent and local vertical plane O-S-P_v.

C.1 *Latitude Limit of Visibility for the Geostationary Orbit*

Figure 14 shows that the maximum latitude visible to the geostationary orbit is

$$\begin{aligned} \varphi_{\max} &= \cos^{-1}\left[K^{-1}\cos\left(\alpha_o - \tau_{\alpha o}\right)\right] - \alpha_o - \tau_{\alpha o} \\ &= \beta, \end{aligned} \tag{34}$$

where

α_o = known elevation angle of the radio-relay antenna,
$\tau_{\alpha o}$ = corresponding total ray-bending angle due to refraction inferred from CRPL Exponential Reference Atmosphere[2] or from Figs. 6 of text, and
K = ratio of orbit radius R to assumed earth radius a.

Hence, equation (34) is a restatement of equation (13).

C.2 *Pointing Azimuth to Orbit Intercept*

The larger of angles formed by intersection of site latitude circle φ and the visibility circle corrected for antenna elevation and refraction shown in Fig. 14 is a direct measure of the pointing azimuth to be avoided for a station located at that intersection (P_v). Hence, angle A_N between tangents $\mathbf{T}_\varphi$ and $\mathbf{T}_v$ is the supplement of pointing angle A given by equation (14) referred to true south. From the geometry of Fig. 14

$$\begin{aligned} A_N = \cos^{-1}\Bigg\{ &-\sin\left[\tan^{-1}\left(\frac{\sin\varphi}{|\,(1-\sin^2\varphi-\cos^2\beta)^{\frac{1}{2}}\,|}\right)\right] \\ &\cdot\cos\left[\tan^{-1}\left(\frac{|\,(1-\sin^2\varphi-\cos^2\beta)^{\frac{1}{2}}\,|}{\cos\beta}\right)\right]\Bigg\}, \end{aligned}$$

$$\pi/2 \leqq |\,A_N\,| < \pi. \tag{35}$$

Equation (35) yields two pointing azimuths which are of interest; one in the second quadrant referred to true north, corresponding to a westerly direction for stations in the northern hemisphere—and one in the third quadrant, or easterly direction. Reversed directions result for stations in the southern latitudes.

Figures 15 illustrate changes of variables which simplify the demonstration of equivalence between equation (35) and equation (14). Since $\cos^2\beta = 1 - \sin^2\beta$ and $\cos^2\varphi = 1 - \sin^2\varphi$,

$$A_N = \cos^{-1}\left\{-\sin\left[\tan^{-1}\left(\frac{\sin\varphi}{|\,(\sin^2\beta-\sin^2\varphi)^{\frac{1}{2}}\,|}\right)\right]\right.$$

$$\cdot \cos\left[\tan^{-1}\left(\frac{|\,(\cos^2\varphi - \cos^2\beta)^{\frac{1}{2}}\,|}{\cos\beta}\right)\right]\Big\}.$$

Substituting μ and ν (Figs. 15) and reducing the result yields

$$A_N = \cos^{-1}(-\sin\mu\cos\nu).$$

Now, substituting for μ and ν provides

$$A_N = \cos^{-1}[-(\sin\varphi/\sin\beta)(\cos\beta/\cos\varphi)]$$

$$= \cos^{-1}\left(-\frac{\tan\varphi}{\tan\beta}\right),$$

which is exactly the supplement of the angle obtained from equation (11) in the text.

C.3 *Longitude Displacement of Orbit Intercept*

The earth-longitude displacement between radio-relay site P_v and suborbital intercept S′ in Fig. 14 is also inferred from equation (35):

$$\lambda = \tan^{-1}\left[\frac{|\,(1 - \sin^2\varphi - \cos^2\beta)^{\frac{1}{2}}\,|}{\cos\beta}\right] \qquad 0 \leqq |\,\lambda\,| < \pi/2,$$

$$= \sin^{-1}(\sin A \sin\beta), \tag{36}$$

from which first and fourth quadrant longitude adjustments are referred to the suborbital longitude. The equivalence with equation (15) is demonstrated using techniques indicated above and identifying angle β uniquely with the maximum latitude for visibility φ_{max} (Section C.1).

C.4 *Geometric Orbit Trace*

Since the size of the visibility circle in Fig. 14 depends upon α_o and $\tau_{\alpha o}$, the earth-longitude displacement λ between P_v and S (same longitude as suborbital point S′) also depends upon α_o and $\tau_{\alpha o}$. Because

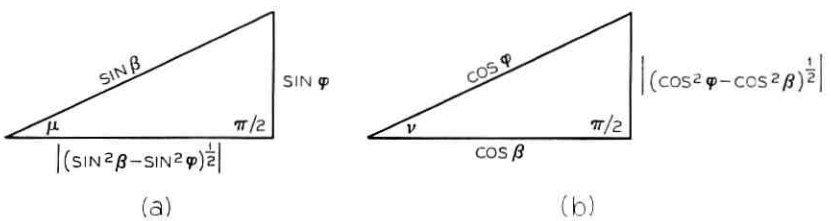

Fig. 15 — Change of variables: (a) angle μ and (b) angle ν.

the resulting viewing aspect depends upon these parameters, the slope of the geometrical orbit trace constructed in Fig. 11 also depends upon α_o and $\tau_{\alpha o}$ as well as φ.

Note that the angle between orbit tangent t constructed at S and the local vertical plane at P_v through S is denoted by ϕ_t. This angle is observed undistorted at point S′, but appears to be a slightly enlarged angle ϕ_t' for an observer at P_v (as if point S were visible without ray bending):

$$\phi_t' = \tan^{-1} (\tan \phi_t / \cos \Omega). \tag{37}$$

Hence, complementary angle δ between orbit tangent **t** and a line through S perpendicular to the geometrical line-of-sight $\mathbf{SP}_v$ and parallel to the horizontal plane at P_v is also the slope of the corrected geometric orbit trace shown in Fig. 11,

$$\begin{aligned} \delta &= \cot^{-1} (\tan \phi_t / \cos \Omega) \\ &= \cot^{-1} \left\{ \frac{\tan [\sin^{-1} (\sin \varphi / \sin \beta)]}{\cos \{\tan^{-1} [\sin \beta / (K - \cos \beta)]\}} \right\} \\ &= \tan^{-1} \{\tan [\cos^{-1} (\sin \varphi / \sin \beta)] \cos \Omega\}, \end{aligned} \tag{38}$$

for which equivalence to equation (19) is shown using the techniques incorporated in Sections C.2 and C.3.

C.5 *Maximum Latitude Permitting Angular Separation v Below the Orbit*

A maximum latitude exists for each antenna elevation angle allowing the beam-center ray to be below the refracted orbit with prescribed angular separation v. Figure 16 illustrates the determination of this critical latitude. S_p is a point on the geostationary orbit having zero relative longitude with respect to station site P_v. A ray emanating from the antenna with initial vertical angle $(\pi/2 + \alpha_o + v)$ must just intercept the orbit. Accounting for refraction, the angle between the geo-

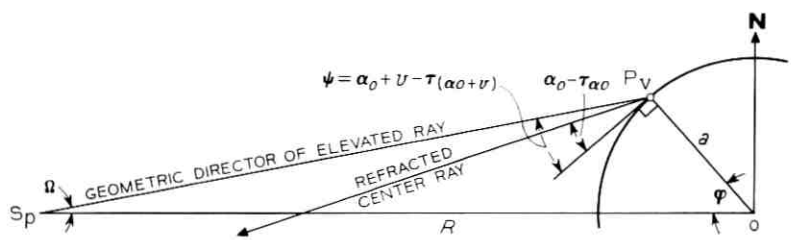

Fig. 16 — Geometry determining the critical latitude.

metric director of the elevated ray and the local horizontal is $\Psi = [\alpha_o + v - \tau_{(\alpha_o+v)}]$. Site latitude φ and Ψ are related by triangle OP_vS_p. From the Law of Sines

$$\frac{\sin(\pi/2 + \Psi)}{R} = \frac{\sin \Omega}{a}.$$

Letting $R/a = K$,

$$\Omega = \sin^{-1}(K^{-1} \cos \Psi).$$

Since $\Omega + \varphi + \pi/2 + \Psi = \pi$,

$$\varphi = \cos^{-1}(K^{-1} \cos \Psi) - \Psi. \tag{39}$$

REFERENCES

1. Gould, R. G., "Protection of the Stationary Satellite Orbit," Telecommunication J., *34*, No. 8 (August 1967), pp. 307–312.
2. Bean, B. R., and Thayer, G. D., "CRPL Exponential Reference Atmosphere," NBS Monograph 4, U. S. Department of Commerce, Washington, D. C., October 29, 1959.
3. Brice, P. J., "Total Atmospheric Refraction of Radio Waves at Small Angles of Declination," unpublished work (WO Branch Memorandum p. 326), Post Office Research Station, General Post Office, London, February, 1967.
4. Lundgren, C. W., unpublished work.
5. Bean, B. R., and Dutton, E. J., "Radio Meteorology," NBS Monograph 92, U. S. Department of Commerce, Washington, D. C., March 1, 1966, p. 345.
6. Bean, B R., "The Radio Refractive Index of Air," Proc. IRE, *50*, No. 3 (March 1962), pp. 260–273.
7. Schulkin, M., "Average Radio-Ray Refraction in the Lower Atmosphere," Proc. IRE, *40*, No. 5 (May 1952), pp. 554–561.
8. Bean, B. R., Horn, J. D., and Ozanich, Jr., A. M. "Climatic Charts and Data of the Radio Refractive Index for the United States and the World," NBS Monograph 22, U S. Department of Commerce, Washington, D. C., November 25, 1960.
9. CCIR Recommendation 406-1, *Documents of the XI Plenary Assembly, Oslo*, 1966, *4*, Part 1.
10. Schelleng, J. C., Burrows, C. R., and Ferrell, E. G., "Ultra-Short-Wave Propagation," Proc. IRE, *21*, No. 3 (March 1933), pp. 427–463.

An Extended Correlation Function of Two Random Variables Applied to Mobile Radio Transmission

By W. C.-Y. LEE

(Manuscript received June 16, 1969)

The definition, properties, and application of an extended correlation function of two random variables involving two common parameters are described and applied to mobile radio systems. The correlation functions of a predetection diversity combined signal (using a scheme of phase equalizing by multiple heterodyning) and of a directional antenna array signal are derived with the help of the extended correlation function.

These correlation functions can be used to determine parameter values giving minimum correlation between two signals desirable for diversity systems. One can also obtain the power spectra by taking the Fourier transform of these correlation functions. Thus extended correlation functions promise to be useful.

I. INTRODUCTION

If two random variables depend on only one common parameter, such as time or distance, conventional correlation formula can be applied to the two variables. However, if both of these variables involve not one but two common variable parameters, then the correlation formula found in the current literature is limited.* Since such cases occur in some of our mobile radio problems, as we discuss later, we need to define an extended correlation function and outline its properties and applications.

II. DERIVATION OF AN EXTENDED CORRELATION FUNCTION OF TWO RANDOM VARIABLES INVOLVING TWO COMMON PARAMETERS

A conventional normalized correlation function of two random variables r_1 and r_2, both of which are functions of one parameter

* Prior to acceptance of this paper for publication, the author was advised that a similar concept was discovered independently by A. Papoulis in his recently published book.[1]

d[that is, $r_1(d_1)$ and $r_2(d_2)$] can be expressed as[2]

$$\rho_{12}(d_1 , d_2) = \frac{R_{12}(d_1 , d_2) - m_1 m_2}{\sigma_1 \sigma_2}$$

$$= \frac{\langle r_1(d_1) r_2(d_2) \rangle_{av} - \langle r_1(d_1) \rangle_{av} \langle r_2(d_2) \rangle_{av}}{[\langle r_1^2(d_1) \rangle_{av} - \langle r_1(d_1) \rangle_{av}^2]^{\frac{1}{2}} \cdot [\langle r_2^2(d_2) \rangle_{av} - \langle r_2(d_2) \rangle_{av}^2]^{\frac{1}{2}}} \quad (1)$$

where m's are the mean values, σ^2's are the covariances, $\rho_{12}(d_1 , d_2)$ is in the range $0 \leqq | \rho_{12}(d_1 , d_2) | \leqq 1$, and $R_{12}(d_1 , d_2) = \langle r_1(d_1) r_2(d_2) \rangle_{av}$ is the correlation function.†

Supposing a random variable $r_1(D_1 ; d_1)$ is a function of two parameters D_1 and d_1, and another variable $r_2(D_2 ; d_2)$ is a function of two parameters D_2 and d_2; the normalized correlation functions of these two variables can be deduced from equation (1):

$$\rho_{12}(D_1 , D_2 ; d_1 , d_2)$$

$$= \frac{R_{12}(D_1 , D_2 ; d_1 , d_2) - m_1 m_2}{\sigma_1 \sigma_2}$$

$$= \frac{\langle r_1(D_1 , d_1) r_2(D_2 , d_2) \rangle_{av} - \langle r_1(D_1 , d_1) \rangle_{av} \langle r_2(D_2 , d_2) \rangle_{av}}{[\langle r_1^2(D_1 , d_1) \rangle_{av} - \langle r_1(D_1 , d_1) \rangle_{av}^2]^{\frac{1}{2}} [\langle r_2^2(D_2 , d_2) \rangle_{av} - \langle r_2(D_2 , d_2) \rangle_{av}^2]^{\frac{1}{2}}}. \quad (2)$$

If the problem we are dealing with is a stationary random process for both of the parameters D and d, then

$$R_{12}(D_1 , D_2 ; d_1 , d_2) = R_{12}(D_1 - D_2 ; d_1 - d_2),$$

$$\langle r_k(D_k , d_k) \rangle_{av} = \langle r_k(0, 0) \rangle_{av} = m_k ,$$

$$\langle r_k^2(D_k , d_k) \rangle_{av} - m_k^2 = \langle r_k^2(0, 0) \rangle_{av} - m_k^2 = \sigma_k^2 ,$$

where $k = 1, 2$. Now m_k and σ_k are constants and we may let $D = D_1 - D_2$ and $d = d_1 - d_2$. Then equation (2) becomes

$$\rho_{12}(D; d) = \frac{R_{12}(D; d) - m_1 m_2}{\sigma_1 \sigma_2}. \quad (3)$$

We call $\rho_{12}(D; d)$ a normalized extended correlation function of the first kind. Also we note that $\rho_{12}(D; d)$ in equation (3) is always smaller than $\rho_{12}(0; 0)$ which is equal to one:

$$\rho_{12}(D; d) \leqq \rho_{12}(0; 0) = 1.$$

† The terms "correlation function $R_{12}(d_1, d_2)$ and normalized correlation function $\rho_{12}(d_1, d_2)$" are adopted from Ref. 3, p. 59.

When the difference D is equal to zero, then

$$\rho_{12}(0; d) = \rho_{12}(d).$$

Now we should illustrate and extend equation (3). We are going to find an extended correlation function of the first kind from a function $\epsilon(D; d_1, d_2, d_3, \cdots, d_m)$ where all the d's are function of D and another parameter α [that is, $d_i(D, \alpha)$ for $i = 1, m$], then

$$R_\epsilon(D; d_1, d_2, d_3, \cdots, d_m) = \langle \epsilon[0; d_1(0, 0), d_2(0, 0), d_3(0, 0), \cdots) \cdot \epsilon(D; d_1(D, \alpha), d_2(D, \alpha), \cdots, d_m(D, \alpha)] \rangle_{av},$$

and the normalized correlation function can be derived from equation (2) as

$$\rho_\epsilon(D; d_1, d_2, \cdots, d_m) = \frac{R_\epsilon(D; d_1, \cdots, d_m) - m_\epsilon^2}{\sigma_\epsilon^2}$$

where

$$m_\epsilon = \langle \epsilon[0; d_1(0, 0), d_2(0, 0), d_3(0, 0), \cdots, d_m(0, 0)] \rangle_{av}$$

$$\sigma_\epsilon^2 = \langle \epsilon^2[0; d_1(0, 0), d_2(0, 0), d_3(0, 0), \cdots, d_m(0, 0)] \rangle_{av} - m_\epsilon^2 .$$

If we consider the case $d_i(D, \alpha)$ is a constant for all D and α itself is a constant, then we may assign a new symbol $R_\epsilon(D \mid d_1, d_2, \cdots, d_m)$ which can be expressed as

$$R_\epsilon(D \mid d_1, d_2, \cdots, d_m) = \langle \epsilon(0; d_1, d_2, \cdots, d_m) \epsilon(D; d_1, d_2, d_3, \cdots, d_m) \rangle_{av} .$$

$R_\epsilon(D \mid d_1, d_2, d_3, \cdots, d_m)$ is a correlation function under a condition that all $d_1, d_2, d_3, \cdots, d_m$ are constants. The normalized correlation is

$$\rho_\epsilon(D \mid d_1, d_2, \cdots) = \frac{R_\epsilon(D \mid d_1, d_2, \cdots, d_m) - m_\epsilon^2}{\sigma_\epsilon^2}, \qquad (4)$$

where m_ϵ and σ_ϵ have been defined previously. We call equation (4) the normalized correlation function of the second kind. As we will show in the Section III, the extended correlation function of first kind $\rho_{12}(D; d)$ and the extended correlation function of second kind $\rho_{12}(D \mid d)$ can be used to obtain the correlation of signals from two diversity scheme receivers easily.

In order to give physical meaning to these functions, let us consider the following two cases. Suppose that two base-station multibranch

diversity receiver arrays are separated by a distance D. The antenna spacing between branch-antenna elements for the first array is d_1 and for the second array is d_2. Both receivers simultaneously receive the signal from a distant mobile radio unit. We would like to determine the values of d_1, d_2, and D to obtain the least cross-correlation desirable for the best diversity reception of these two received signals. The extended correlation of first kind $\rho_{12}(D; d_1, d_2)$ may be used in this case.

The second case assumes that a mobile radio multi-branch diversity receiver array, with given uniform antenna element spacing d, moves along the street with a constant speed V. The autocorrelation of a signal ϵ, received by the mobile receiver, can be obtained from the extended correlation function of the second kind $\rho_\epsilon(D \mid d)$. Alternatively, we can also consider a multielement directive antenna instead of the diversity scheme. In this paper, we only treat the latter case. The former case can be solved following the same technique.

III. APPLICATION TO MOBILE RADIO PROBLEMS

3.1 *Derivation of the Correlation Function of a Signal Received from a Predetection Diversity Combining Receiver*

A multichannel predetection diversity combining system is a scheme for bringing a number of RF carriers to a common phase by means of multiple heterodyning. Then a linear combiner at the IF frequency is used to sum the individual channels.[4,5] A signal received from this system is called a predetection diversity combined signal.

Suppose that a signal consisting of multipath vertically polarized waves is received by an M-branch predetection combining mobile receiver with a M-antenna space diversity array. The M-antennas are spaced by $d_1, d_2, \cdots d_M$ respectively from an arbitrary common point. After the array has moved a distance D, the received signal ϵ, which is the sum of the M individual signal amplitudes received from M individual antennas, can be expressed as[6,7]

$$\begin{aligned}\epsilon(D; d_1, d_2, d_3, d_4, \cdots, d_M) &= r_1(D; d_1) + r_2(D; d_2) + r_3(D; d_3) \\ &\quad + \cdots + r_M(D; d_M) \\ &= \sum_1^M r_m(D; d_m), \end{aligned} \tag{5}$$

where all r_m are functions of distance D and antenna spacing d_m (see Appendix A). For a mobile radio signal,[7,8] or a long range fading signal,[9]

the r_m are usually Rayleigh distributed. Suppose that all d's are constants, then the autocorrelation function of the signal given in equation (5) as a function of the separation distance D is an extended autocorrelation function of second kind which can be expressed as

$$\begin{aligned} R_\epsilon(D \mid d_1, d_2, d_3, d_4, \cdots) &= \langle \epsilon(0; d_1, d_2, d_3, d_4, \cdots)\epsilon(D; d_1, d_2, d_3, d_4)\rangle_{av} \\ &= \left\langle \left[\sum_1^M r_m(0; d_m)\right]\left[\sum_1^M r_m(D; d_m)\right]\right\rangle_{av} \\ &= \left\langle \sum_{m=1}^M \sum_{n=1}^M r_m(0; d_m) r_n(D; d_n)\right\rangle_{av} \\ &= \sum_{m=1}^M \sum_{n=1}^M R_{mn}(D; d_m - d_n). \end{aligned} \tag{6}$$

Using equation (3), this can also be written

$$R_\epsilon(D \mid d_1 \cdots d_M) = \rho_\epsilon(D \mid d_1, \cdots, d_M)(\sigma_\epsilon^2) + m_\epsilon^2, \tag{7}$$

where

$$\sigma_\epsilon^2 = \langle \epsilon^2(0; d_1, \cdots, d_M)\rangle - m_\epsilon^2, \tag{8}$$

$$\begin{aligned} \langle \epsilon^2(0; d_1, \cdots, d_M)\rangle_{av} &= \left\langle \left(\sum_1^M r_m(0; d_m)\right)^2\right\rangle_{av} \\ &= \sum_1^M \sum_1^M \langle r_m(0; d_m) r_n(0; d_n)\rangle_{av} \\ &= \sum_1^M \sum_1^M R_{mn}(0; d_m - d_n). \end{aligned} \tag{9}$$

Substituting equation (8) into equation (7), and combining equations (6) and (7), we obtain

$$\rho_\epsilon(D \mid d_1, \cdots, d_M) = \frac{\sum_1^M \sum_1^M R_{mn}(D; d_m - d_n) - m_\epsilon^2}{\sum_1^M \sum_1^M R_{mn}(0; d_m - d_n) - m_\epsilon^2}. \tag{10}$$

The terms $R_{mn}(D; d_m - d_n)$ can be found from equation (3);

$$R_{mn}(D; d_m - d_n) = \rho_{mn}(D; d_m - d_n)\sigma_m\sigma_n + m_m m_n \tag{11}$$

and

$$m_\epsilon^2 = \langle \epsilon(0; d_1 \cdots d_m) \rangle_{av}^2 = \left\langle \sum_1^M r_m(0; d_m) \right\rangle_{av}^2$$
$$= \left[\sum_1^M \langle r_m(0; d_m) \rangle_{av} \right]^2$$
$$= \sum_1^M \sum_1^M m_m m_n \,. \tag{12}$$

Hence the correlation function of equation (10) becomes, assuming $\sigma_m = \sigma_n$,

$$\rho_\epsilon(D \mid d_1, \cdots, d_M) = \frac{\sum_{m=1}^M \sum_{n=1}^M \rho_{mn}(D; d_m - d_n)}{\sum_{m=1}^M \sum_{n=1}^M \rho_{mn}(0; d_m - d_n)}. \tag{13}$$

If all spacings between two adjacent antennas are equal, then $d_m - d_n = (m - n)d_1$ where d_1 is the distance between two adjacent antennas. We may let $d = d_1$, and simplify the notation of equation (13) to

$$\rho_\epsilon(D \mid d) = \frac{\sum_{m=1}^M \sum_{n=1}^M \rho_{mn}(D; d)}{\sum_{m=1}^M \sum_{n=1}^M \rho_{mn}(0; d)}. \tag{14}$$

Equation (14) shows that a normalized autocorrelation function of an M-branch predetection combined signal is a normalized extended autocorrelation function of second kind in terms of all individual normalized correlation functions between branches. We notice that

$$\rho_\epsilon(D \mid d) \leqq \rho_\epsilon(0 \mid d) = 1, \tag{15}$$

and as stated in Section II

$$\rho_{mn}(0; d) = \rho_{mn}(d). \tag{16}$$

We may also realize that

$$\rho_{12}(D; d) = \rho_{23}(D; d) = \rho_{34}(D; d) = \cdots$$

and

$$\rho_{13}(D; d) = \rho_{24}(D; d) = \rho_{35}(D; d) = \cdots . \tag{17}$$

Hence, equation (14) can be further simplified as

$$\rho_\epsilon(D \mid d)$$
$$= \frac{M\rho_{11} + (M-1)(\rho_{12} + \rho_{21}) + (M-2)(\rho_{13} + \rho_{31}) + \cdots + \rho_{1M} + \rho_{M1}}{M\rho_{11}^0 + (M-1)(\rho_{12}^0 + \rho_{21}^0) + (M-2)(\rho_{13}^0 + \rho_{31}^0) + \cdots + \rho_{1M}^0 + \rho_{M1}^0}, \tag{18}$$

where $\rho_{mn} = \rho_{mn}(D; d)$ and $\rho^0_{mn}(0; d)$ used in equation (18) are for simplicity (ρ_{mn} and ρ^0_{mn} are derived in Appendix A). If we let the antenna spacing $d/\lambda = 0$, then $\rho_\epsilon(D \mid 0)$ from equation (18) represents the correlation function of two single-branch signals

$$\rho_\epsilon(D \mid 0) = J_0^2(\beta D) \tag{19}$$

which agrees with that in Ref. 6.

Several numerical calculations have been carried out for the following example: Two four-branch diversity receivers, each of them with fixed antenna spacing $d/\lambda = 0.5$ or $d/\lambda = 1.0$, are mounted on the roof of the mobile unit, as shown in Fig. 1. These two receivers are separated by a distance D/λ (D/λ varies from 0 to 4) for two cases, $\alpha = 0°$ and $\alpha = 90°$. The calculations of the extended correlation function $\rho_\epsilon(D \mid d)$ of these two signals, obtained from their respective receivers when the mobile unit is moving, are shown in Figs. 2 and 3. Both figures indicate the values of D/λ which give the least correlation between two signals. We also note that the correlations at $\alpha = 0°$ are higher than that at $\alpha = 90°$. Figures 2 and 3 can also represent the auto correlation of a signal received from a single four-branch diversity receiver which has its antenna spacing $d/\lambda = 0.5$ or 1.0 and moves on a street with a constant speed $V(D = Vt)$. The power spectrum of such a signal can be obtained by taking the Fourier transform of its autocorrelation function.

3.2 *Derivation of the Correlation Function of a Signal Envelope Received from a Directional Antenna Array*

Signal reception from a directional antenna array with M antenna elements has been also suggested as a means of overcoming multipath

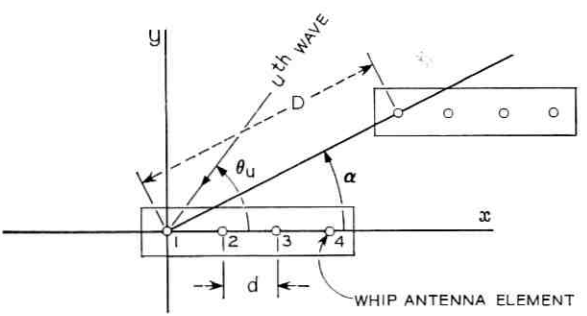

Fig. 1 — Coordinate system of a M-branch diversity mobile radio receiver ($M = 4$ branches).

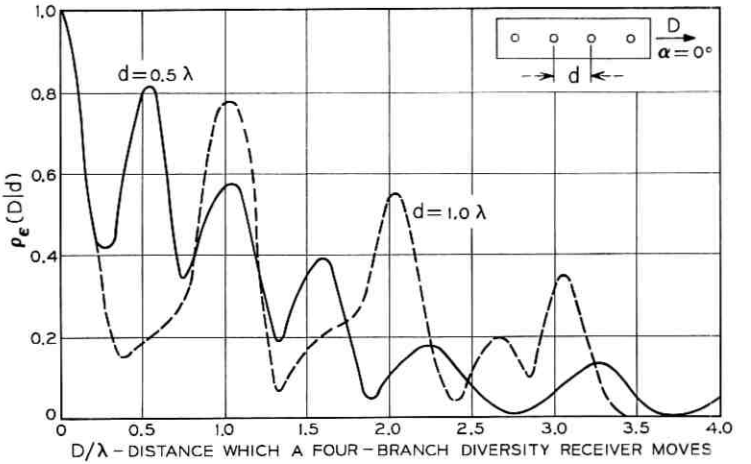

Fig. 2 — Normalized autocorrelation function of a four-branch diversity receiver moving at $\alpha = 0°$.

fading in mobile radio propagation.[10-12] The derivation of the correlation function of this signal envelope is as follows.

Suppose that the same kind of signal which consists of multipath vertical polarized waves as mentioned in Section 3.1 is received by a directional M-antenna array. The M antennas are spaced by d_1, d_2,

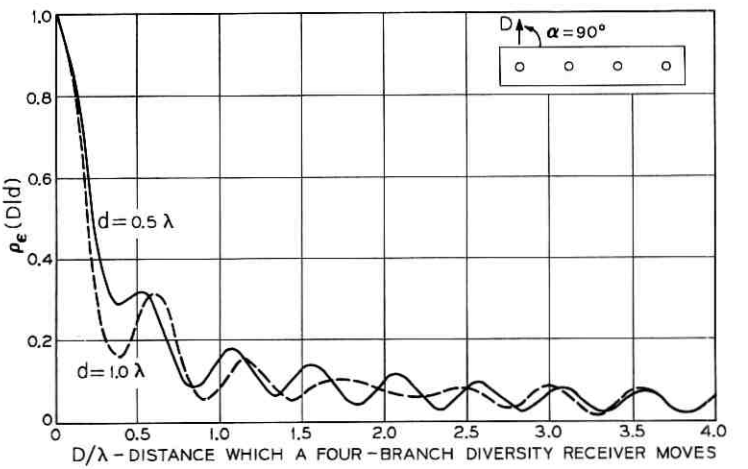

Fig. 3 — Normalized autocorrelation function of a four-branch diversity receiver moving at $\alpha = 90°$.

d_3 , $\cdots$, d_M respectively from an arbitrary common point. After the antenna array is moved by a distance D, (see Fig. 4) the received signal envelope ϵ which is the amplitude of the sum of M individual signals can be expressed as[13]

$$\begin{aligned}\epsilon(D; d_1 , d_2 , d_3 , \cdots , d_M) &= | s_1(D; d_1) + s_2(D; d_2) + \cdots + s_M(D; d_M) | \\ &= \left| \sum_1^M s_m(D; d_m) \right| \\ &= | X(D; d_1 , d_2 , \cdots , d_M) + jY(D; d_1 , d_2 , \cdots , d_M) |, \end{aligned} \tag{20}$$

where s_m is a complex variable which represents the amplitude and the phase of an individual signal. X and Y are the real and imaginary parts of the total signal.

If the spacings between adjacent antennas are equal, then antenna m and antenna n are separate dyb $d_m - d_n = (m - n)d$. Therefore X and Y of equation (20) are functions of D and d only. Suppose that all d's are constants, the autocorrelation function of signal envelope ϵ can be obtained by using the equation:[14]

$$\rho_\epsilon(D \mid d) \doteq \frac{\langle X_1(0; d)X_2(D; d)\rangle_{av}^2 + \langle X_1(0; d)Y_2(D; d)\rangle_{av}^2}{\langle X_1^2(0; d)\rangle_{av}^2} \tag{21}$$

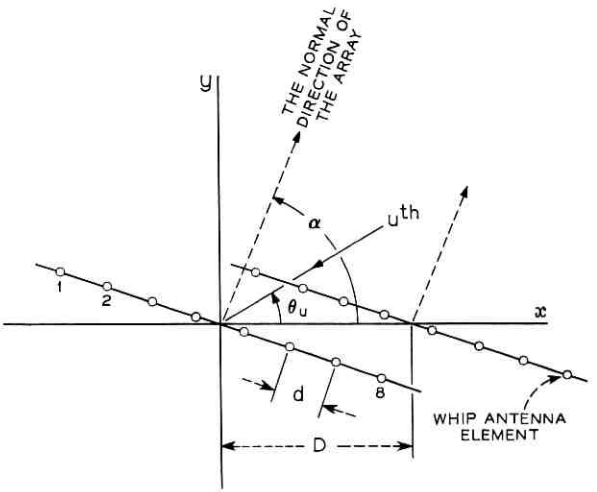

Fig. 4 — Coordinate system of a broadside directional antenna array ($M = 8$ elements).

provided X and Y are gaussian variables, all $\langle X_m \rangle_{av}$ are zeros, and $\langle X_m^2 \rangle_{av}$ and $\langle Y_m^2 \rangle_{av}$ are equal, where $m = 1$ or 2. These facts are shown in Appendix B. If the antenna spacing $d/\lambda = 0$, then $\rho_\epsilon(D \mid 0)$ actually represents the correlation between two single-branch signals, which agrees with equation (19) and Ref. 6.

The normalized correlation function of a signal received from a broadside directional antenna array is

$$\rho_\epsilon(D \mid d) = \frac{1}{4} \frac{\left\{ \sum_{m=1}^{K} \sum_{n=1}^{K} [J_0(A_1) + J_0(B_1) + J_0(A_2) + J_0(B_2)] \right\}^2}{\left[\sum_{m=1}^{K} \sum_{n=1}^{K} [J_0(A_0) + J_0(B_0)] \right]^2} \tag{22}$$

where

$$K = \frac{M}{2} \qquad \text{for} \quad M \text{ is even}$$

$$= \frac{M+1}{2} \quad M \quad \text{for is odd,}$$

and A_1, B_1, A_2, and B_2 are shown in equation (48). A_0 and B_0 are shown in equation (49).

Several numerical calculations have been carried out for the following example: Two eight-element broadside antenna arrays, each of them with fixed antenna spacing $d/\lambda = 0.5$ or $d/\lambda = 1.0$, are mounted on the roof of the mobile unit. These two arrays are separated by a distance D/λ (D/λ varies from 0 to 4) for two cases $\alpha = 0°$ and $\alpha = 90°$. The calculations of the extended correlation function $\rho_\epsilon(D \mid d)$ between two signals received from their respective arrays when the mobile unit is moving are shown in Figs. 5 and 6. Both figures indicate the values of D/λ which have the least correlation between two signals. The extended correlation curve of $d/\lambda = 0.5$ is quite different from that $d/\lambda = 1.0$ in both figures. The curve of $d/\lambda = 0.5$ in Fig. 5 shows that the high correlation and low correlation are about 0.25λ apart; however, this phenomenon does not appear for $d/\lambda = 0.5$, but rather for $d/\lambda = 1.0$ in Fig. 6. It can be explained as follows. For the directional antenna array with spacing $d = \lambda/2$, most of the energy is contained in the two major broadside lobes, while for the directional antenna array with antenna spacing $d = \lambda$, most of the energy is contained in the two major end-fire lobes. As the vehicle moves, strong standing waves may occur when the major antenna lobes lie in line with the motion of the vehicle, such as for the case $\alpha = 0°$ and $d = \lambda/2$; or the case $\alpha = 90°$ and $d = \lambda$.

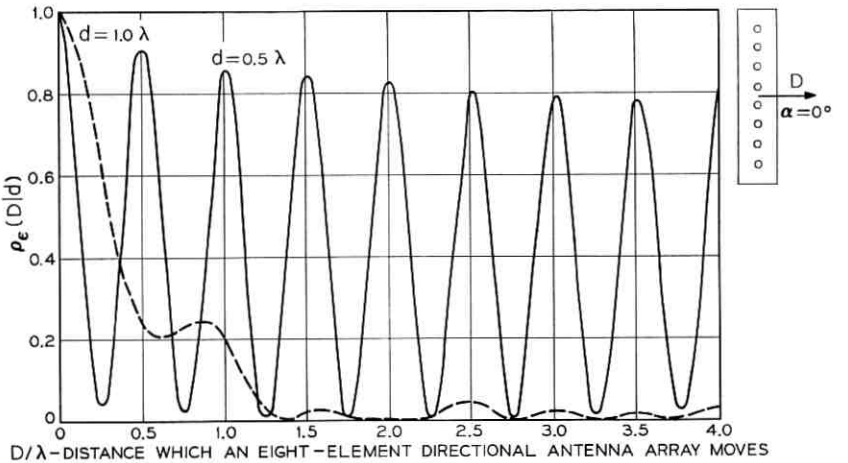

Fig. 5 — Normalized autocorrelation function of an eight-element directional antenna array pointing at $\alpha = 0°$.

The autocorrelations obtained from these standing waves, then, become oscillatory in nature, as we would expect.

Figures 5 and 6 can also represent the autocorrelation of a signal received from an eight-element broadside antenna array which has its antenna spacing $d/\lambda = 0.5$ or 1.0 and moves on a street with a constant

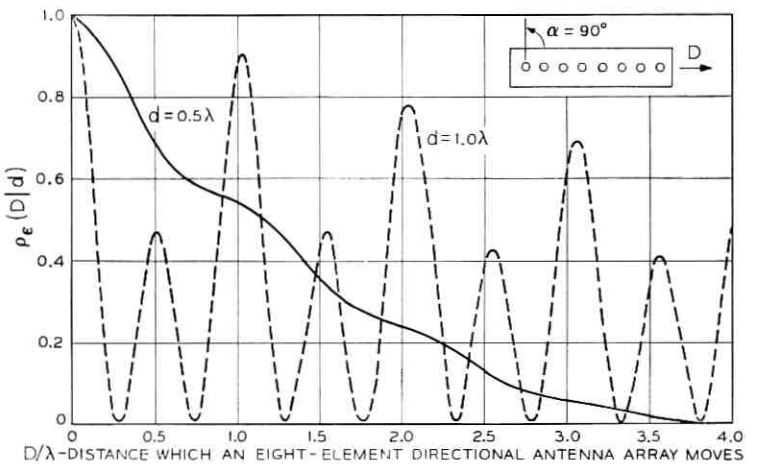

Fig. 6 — Normalized autocorrelation function of an eight-element directional antenna array pointing at $\alpha = 90°$.

speed V ($D = Vt$). The power spectrum of such a signal can be obtained by taking the Fourier transform of its autocorrelation function as we mentioned in Section 3.1.

IV. CONCLUSION

The derivation of a general correlation function of two random variables, each of them involving two parameters, has been obtained. The terms "extended correlation function of first kind" and "extended correlation function of second kind" have been defined. The application of the extended correlation function is demonstrated. The correlation function of a diversity signal and the correlation function of a directional antenna array signal are derived with the help of the extended correlation function in this paper. Several numerical calculations have also been carried out. From these correlation functions we can obtain the least correlations between two signals under certain circumstances. Also, we can obtain the power spectra by taking the Fourier transform of these correlation functions. Thus, it seems likely that these functions will find general application.

V. ACKNOWLEDGMENT

The author is indebted to C. L. Mallows and M. J. Gans with whom previous discussions have been very helpful. He also wishes to thank W. C. Jakes, Jr., for his many suggestions.

APPENDIX A

Finding Normalized Cross Correlation Functions From Individual Branch Signals of a Predetection Diversity Combining Receiver

It is easy to show that the signal from branch m in equation (5) is

$$r_m = \left| \sum_1^N A_u \exp\left[+j\beta D \cos(\theta_u - \alpha) + j(m-1)\beta d \cos\theta_u\right] \right|$$

$$= | X_m + jY_m |, \tag{23}$$

where

$$A_u = R_u + jS_u ,$$

$$X_m = \sum_{u=1}^N R_u \cos\phi_u + S_u \sin\phi_u , \tag{25}$$

$$Y_m = \sum_{u=1}^N S_u \cos\phi_u - R_u \sin\phi_u , \tag{26}$$

$$-\phi_u = \beta D \cos(\theta_u - \alpha) + (m-1)\beta d \cos\theta_u \tag{27}$$

(R_u and S_u are independent gaussian amplitudes with zero mean and unit variance). The diversity receiver is located at the point (D, α) in polar coordinate. The distance D, the angle α, and the arrival of uth wave at angle θ_u are shown in Fig. 1. We assume the N waves are uniformly distributed in angle. Now we can average the product of two components of two branches—branch m and branch n—as

$$\begin{aligned}
&\langle X_m(D_1)X_n(D_1+D)\rangle_{av} \\
&\quad = \langle X_m(0)X_n(D)\rangle_{av} \\
&\quad = NE\{\cos[BD\cos(\theta_u-\alpha) - (m-n)\beta d\cos\theta_u]\} \\
&\quad = N[J_0(a)J_0(b) - 2J_2(a)J_2(b) + 2J_4(a)J_4(b) \\
&\qquad\qquad - 2J_6(a)J_6(b) + \cdots] \\
&\quad = NJ_0(a^2+b^2)^{\frac{1}{2}},
\end{aligned} \tag{28}$$

where[15]

$$a = \beta D\cos\alpha - (m-n)\beta d, \tag{29}$$

$$b = \beta D\sin\alpha, \tag{30}$$

$$a^2 + b^2 = (BD)^2 + (m-n)^2(\beta d)^2 - 2(m-n)\beta^2 Dd\cos\alpha, \tag{31}$$

and

$$\begin{aligned}
&\langle X_m(D_1)Y_n(D_1+D)\rangle_{av} \\
&\quad = \langle X_m(0)Y_n(D)\rangle_{av} \\
&\quad = NE\{\sin[\beta D\cos(\theta_u-\alpha) - (m-n)\beta d\cos\theta_u]\} \\
&\quad = 0.
\end{aligned} \tag{32}$$

Also

$$\langle X_m^2(D_1)\rangle = \langle Y_m^2(D_1)\rangle = N. \tag{33}$$

Substituting equations (28), (32), and (33) into the following equation[6,14]

$$\rho_{mn}(D;d) \doteqdot \frac{\langle X_m(0;d)X_n(D;d)\rangle_{av}^2 + \langle X_m(0;d)Y_n(D;d)\rangle_{av}^2}{\langle X_m^2(0;d)\rangle_{av}^2}. \tag{34}$$

Then we obtain the final result

$$\rho_{mn}(D;d) \doteqdot J_0^2[(\beta D)^2 + (m-n)^2(\beta D)^2 - 2(m-n)^2 Dd\cos\alpha]^{\frac{1}{2}} \tag{35}$$

and

$$\rho_{mn}(d) = \rho_{mn}(0; d) = J_0^2[(m - n)\beta d]. \tag{36}$$

We also can show the relations

$$\rho_{12} = \rho_{23} = \rho_{34}\ ; \qquad \rho_{21} = \rho_{32} = \rho_{43}\ ,$$

$$\rho_{13} = \rho_{24} = \rho_{35}\ ; \qquad \rho_{31} = \rho_{42} = \rho_{53}\ ,$$

$$\rho_{mn} \neq \rho_{nm} \quad \text{for} \quad m \neq n,$$

$$\rho_{mn}(D; d) = \rho_{nm}(-D; d).$$

APPENDIX B

Finding a Normalized Correlation Function From a Real Part and an Imaginary Part of a Signal Received From a Directional Antenna Array

It is easy to show that a signal consisting of N multipath vertical polarized waves received from an equal-spaced directional antenna array at a distance D from a reference position is[16]

$$E_z(D; d) = \sum_{u=1}^{N} A_u\{1 + \exp(j\psi) + \exp(j2\psi) + \exp(j3\psi) + \cdots + \exp[j(M - 1)\psi]\} \cdot \exp(j\beta D \cos\theta_u), \tag{37}$$

where A_u was defined in equation (24),

$\psi = \beta d \sin(\alpha - \theta_u) + \delta$,
d is antenna spacing between two antennas,
M is the number of elements,
α is the normal direction of the array,
δ is the relative phase between antennas,
D is the distance measured from the coordinate origin to the center position of antenna array. (The center position of the antenna array is assumed always on the axis, that is, at the position $(D, 0)$.), and
θ_u is the angle of arrival of the uth wave and is assumed to be uniformly distributed.

The coordinate system of a directional antenna array is shown in Fig. 6. Since the spacings between antennas are equal, we can let the phase refer to the center point of the array. Then equation (37) can be simplified by combining the first term and the Mth term, the second

term and the $(M-1)$th term and so forth.[13] The result becomes

$$E_z(D;d) = \sum_{u=1}^{N} A_u \exp(j\beta D\cos\theta_u) \cdot\left[2\cos\left(\frac{M-1}{2}\psi\right) + 2\cos\left(\frac{M-3}{2}\psi\right) + \cdots + 2Q\right], \tag{38}$$

where

$$Q = 1 \quad \text{if} \quad M = \text{odd}$$
$$= \cos(\tfrac{1}{2}\psi) \quad \text{if} \quad M = \text{even}.$$

Equation (38) can be separated into a real part and an imaginary part as

$$E_z(D;d) = X + jY$$

and

$$\epsilon(D;d) = |E_z(D;d)| = (X^2+Y^2)^{\frac{1}{2}}, \tag{39}$$

where

$$X = 2\sum_{m=1}^{K}\sum_{u=1}^{N}[R_u\cos(\beta D\cos\theta_u) - S_u\sin(\beta D\cos\theta_u)] \cdot\cos\left(\frac{M+1-2m}{2}\psi\right)$$
$$= 2\sum_{m=1}^{K} x_m \tag{40}$$

$$Y = 2\sum_{m=1}^{K}\sum_{u=1}^{N}[R_u\sin(\beta D\cos\theta_u) + S_u\cos(\beta D\cos\theta_u)] \cdot\cos\left(\frac{M+1-2m}{2}\psi\right)$$
$$= 2\sum_{m=1}^{K} y_m\,, \tag{41}$$

where

$$K = \frac{M}{2} \quad \text{if} \quad M \text{ is even}$$
$$= \frac{M+1}{2} \quad \text{if} \quad M \text{ is odd}. \tag{42}$$

Since R_u and S_u are independent gaussian variables, it is easy to realize that all x_m and y_m are independent gaussian variables. Hence X and Y are also gaussian variables. The mean values of X and Y are zeros, and the mean squares of X and Y are the same. Therefore equation (21) can be applied. The following term in equation (21) can be replaced by

$$\langle X_1(0;d)X_2(D;d)\rangle_{av} = 4\left[\sum_{m=1}^{K} x_m(0;d)\right]\left[\sum_{m=1}^{K} x_m(D;d)\right]$$
$$= 4\sum_{m=1}^{K}\sum_{n=1}^{K}\langle x_m(0;d)x_n(D;d)\rangle_{av} . \tag{43}$$

The term $\langle X_1(0;d)Y_2(D;d)\rangle$ also can be obtained, and is equal to equation (43), by replacing x_n by y_n. Then equation (20) becomes

$$\rho_e(D/d) = \frac{\left[\sum_{m=1}^{K}\sum_{n=1}^{K}\langle x_m(0;d)x_n(D;d)\rangle_{av}\right]^2 + \left[\sum_{m=1}^{K}\sum_{n=1}^{K}\langle x_m(0;d)y_n(D;d)\rangle_{av}\right]^2}{\left[\sum_{m=1}^{K}\sum_{n=1}^{K}\langle x_m(0;d)x_n(0;d)\rangle_{av}\right]^2}, \tag{44}$$

where K is shown in equation (42), and

$$\langle x_m(0;d)x_n(D;d)\rangle_{av} = \frac{N}{2}\langle\cos(\beta D\cos\theta_u)\cdot\{\cos[(M+1-m-n)\psi] + \cos[(m-n)\psi]\}\rangle_{av},$$
$$\langle x_m(0;d)y_n(D;d)\rangle_{av} = \frac{N}{2}\langle\sin(\beta D\cos\theta_u)\cdot\{\cos[(M+1-m-n)\psi] + \cos[(m-n)\psi]\}\rangle_{av}, \tag{45}$$

and

$$\psi = \beta d\sin(\alpha-\theta_u) + \delta.$$

Now we may consider only a broadside directional antenna array, that is, $\delta = 0$. Then the following terms can be derived:[15]

$$\langle\cos(a\cos\theta_u)\cdot\cos[b\sin(\alpha-\theta_u)]\rangle_{av}$$
$$= \tfrac{1}{2}\langle\cos[(a+b\sin\alpha)\cos\theta_u - b\cos\alpha\sin\theta_u] + \cos[(a-b\sin\alpha)\cos\theta_u + b\cos\alpha\sin\theta_u]\rangle_{av}$$
$$= \tfrac{1}{2}[J_0(A)+J_0(B)], \tag{46}$$

where

$$A = (a^2 + 2ab \sin \alpha + b^2)^{\frac{1}{2}},$$

$$B = (a^2 - 2ab \sin \alpha + b^2)^{\frac{1}{2}},$$

$$\langle \sin (a \cos \theta_u) \cos [b \sin (\alpha - \theta_u)] \rangle = 0. \tag{47}$$

Inserting the general formulas equation (46) and equation (47) into equation (45), it becomes

$$\langle x_m(0; d) x_n(D; d) \rangle_{av} = \frac{N}{2} [J_0(A_1) + J_0(B_1) + J_0(A_2) + J_0(B_2)]$$

$$\langle x_m(0; d) y_n(D; d) \rangle_{av} = 0 \tag{48}$$

where

$$\left.\begin{matrix} A_1 \\ B_1 \end{matrix}\right\} = \beta[D^2 \pm 2Dd(M + 1 - m - n) \sin \alpha + d^2(M + 1 - m - n)^2]^{\frac{1}{2}}$$

$$\left.\begin{matrix} A_2 \\ B_2 \end{matrix}\right\} = \beta[D^2 \pm 2Dd(m - n) \sin \alpha + d^2(m - n)^2]^{\frac{1}{2}}.$$

From equation (48), we can deduce the results

$$\begin{aligned} &\langle x_m(0; d) x_n(0; d) \rangle_{av} \\ &\quad = N\{J_0[\beta d(M + 1 - m - n)] + J_0[\beta d(m - n)]\} \\ &\quad = N[J_0(A_0) + J_0(B_0)] \end{aligned} \tag{49}$$

and

$$\langle x_m^2(0; d) \rangle = N\{J_0[\beta d(M + 1 - 2m)] + 1\}. \tag{50}$$

Then substituting equations (48) and (49) into equation (44), we complete the derivation of a normalized correlation function of a signal received from a broadside directional antenna array.

REFERENCES

1. Papoulis, A., *System and Transforms with Application in Optics,* New York: McGraw-Hill, 1969, Chapter 8.
2. Schwartz, M., Bennett, W. R., and Stein, S., *Communication System and Techniques,* New York: McGraw-Hill, 1966, p. 8.
3. Davenport, W. B., and Root, W. L., *Random Signals and Noise,* New York: McGraw-Hill, 1958, p. 30, and p. 59.
4. Black, D. M., Kopel, P. S., and Novy, R. J., "An Experimental UHF Dual Diversity Receiver Using a Predetection Combining System," IEEE Trans. on Vehicular Communications, *VC-15* (October 1966), pp. 41–47.

5. Rustako, A. J., Jr., "Evaluation of a Mobile Radio Multiple Channel Diversity Receiver Using Predetection Combining," IEEE Trans. on Vehicular Technology, *VT-16* (October 1967), pp. 46–57.
6. Lee, W. C. Y., "Comparison of An Energy Density Antenna System with Predetection Combining Systems for Mobile Radio," unpublished work.
7. Gans, M. J., "Level Crossing Rates for Arbitrary Directional Antennas and STAR Combining," unpublished work.
8. Lee, W. C.-Y., "Statistical Analysis of the Level Crossings and Duration of Fades of the Signal from an Energy Density Mobile Radio Antenna," B.S.T.J., *46*, No. 2 (February 1967), pp. 417–448.
9. Rice, S. O., "Distribution of the Duration of Fades in Radio Transmission," B.S.T.J., *37*, No. 3 (May 1958), pp. 581–634.
10. Jakes, W. C., Jr., "Mobile Radio: Why not Microwaves?" unpublished work.
11. Lee, W. C. Y., "Preliminary Investigation of Mobile Radio Signal Fading Using Directional Antennas on the Mobile Unit," IEEE Trans. on Vehicular Communications, *VC-15* (October 1966), pp. 8–15
12. Stidham, J. R., "Experimental Study of UHF Mobile Radio Transmission Using a Directive Antenna," IEEE Trans. on Vehicular Communications, *VC-15* (October 1966), pp. 16–24.
13. Kraus, J. D., *Antennas*, New York: McGraw-Hill Book Company, 1950, p. 77.
14. Booker, H. G., Ratcliff, J. A., and Shinn, D. H., "Diffraction from an Irregular Screen with Applications to Ionosphere Problems," Phil. Trans. Royal Soc., London, *242A* (1950), pp. 579–607.
15. Watson, G. N., *Theory of Bessel Functions*, New York: The Macmillan Company, 1948, p. 359.
16. Lee, W. C.-Y., "Theoretical and Experimental Study of the Level Crossings of Signal Fades from Mobile Radio Directional Antennas," unpublished work.

Contributors to This Issue

Richard M. Derosier, A.A.S.E.E., 1967, Hudson Valley Community College; Bell Telephone Laboratories, 1967—. Initially, Mr. Derosier's work concerned the fabrication and development of GaAs injection laser diodes. Currently, he is associated with the studies of mode conversion and radiation losses from various dielectric waveguides.

J. H. Fennick, B.S.E.E., 1959, M.S.E.E., 1961, University of Connecticut; Bell Telephone Laboratories, 1960—. He initially worked on interference problems with emphasis on impulse noise. Currently he is supervisor of the Data Studies Group in the Transmission Systems Division at Holmdel. Member, Tau Beta Pi, Eta Kappa Nu, Phi Sigma Phi, IEEE; associate member, Sigma Xi.

Leonard J. Forys, B.S.E.E., 1963, University of Notre Dame; M.S. and E.E., 1965, Massachusetts Institute of Technology; Ph.D., 1968, University of California at Berkeley; Acting Assistant Professor of Electrical Engineering, University of California at Berkeley, 1967–1968; Bell Telephone Laboratories, 1968—. Mr. Forys is engaged in research and consulting on communication and control theory problems. Member, IEEE.

Philip A. Gresh, B.S.E.E., 1956, Carnegie Institute of Technology; Bell Telephone Laboratories, 1956—. Mr. Gresh supervises a group responsible for the systems planning and economic evaluation of new concepts in the design, layout, and utilization of the telephone exchange outside plant. Member, Tau Beta Pi, Eta Kappa Nu, Phi Kappa Phi.

William C.-Y. Lee, B.Sc. in Engineering, 1954, Chinese Naval Academy; M.Sc. in E.E., 1960, and Ph.D. in E.E., 1963, Ohio State University; Bell Telephone Laboratories, 1964—. He has been concerned with the study of wave propagation in anisotropic medium and antenna theory. His present work has included studies of mobile radio antennas and signal fading problems. Member, Sigma Xi, IEEE.

CARL W. LUNDGREN, E.E., 1957, M.S., 1959, Ph.D., 1961, University of Cincinnati; U. S. Army Electronics Research and Development Laboratory, 1962–1963; Bell Telephone Laboratories, 1961—. Early work in electrodynamics and gyro mechanics resulted in magnetic navigation and spacecraft stabilization techniques. Subsequent interests concerned launch timing for optimum spin-axis orientation and the medium-altitude satellite eclipse environment in support of the *Telstar*® communication satellite experiment. He is studying microwave transmission, interference, and circuit outage problems associated with communication satellite systems. Member, Phi Eta Sigma, Eta Kappa Nu, Tau Beta Pi, Omicron Delta Kappa, IEEE, New York Academy of Sciences.

DIETRICH MARCUSE, Diplom Vorpruefung, 1952, Dipl. Phys., 1954, Berlin Free University; D.E.E., 1962, Technische Hochschule, Karlsruhe, Germany; Siemens and Halske (Germany), 1954–57; Bell Telephone Laboratories, 1957—. At Siemens and Halske, Mr. Marcuse was engaged in transmission research, studying coaxial cable and circular waveguide transmission. At Bell Telephone Laboratories, he has been engaged in studies of circular electric waveguides and work on gaseous masers. He spent one year (1966–1967) on leave of absence from Bell Telephone Laboratories at the University of Utah where he wrote a book on quantum electronics. He is presently working on the transmission aspect of a light communications system. Member, IEEE, Optical Society of America.

A. S. MAY, B.S.E.E., 1939, West Virginia University; Bell Telephone Laboratories, 1939–1962; American Telephone and Telegraph Company, 1962—. At Bell Telephone Laboratories Mr. May was engaged in the design of radar equipment and as a supervisor in the development of microwave radio-relay systems. He is currently engaged in microwave and guided wave planning, and in studies of frequency sharing by terrestrial radio systems and satellites.

A. A. THIELE, B.S. (Physics), 1960, Ph.D. (physics), 1965, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1965—. He was initially engaged in exploratory device development for various optical memory schemes. He is currently working on the development of the theory of circular domains. Member, Sigma Xi.

AARON D. WYNER, B.S., 1960, Queens College; B.S.E.E., 1960, M.S., 1961, and Ph.D., 1963, Columbia University; Bell Telephone Laboratories, 1963—. Mr. Wyner has been doing research in various aspects of information theory. For the year 1969–1970 he is visiting the Dept. of Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. He has also been Adjunct Associate Professor of Electrical Engineering at Columbia University and Chairman of the Metropolitan New York Chapter of the IEEE Information Theory Group. Member, IEEE, Society of Industrial and Applied Mathematics, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.

JACOB ZIV, B.Sc., 1954, Engineering Diploma, 1955, M.Sc., 1957, Technion—Israel Institute of Technology, Haifa, Israel; D.Sc., 1962, Massachusetts Institute of Technology; Technion—Israel Institute of Technology, 1954–1955; Scientific Department, Israel Ministry of Defence, 1955–1959, 1962–1968; Bell Telephone Laboratories, 1968—. (On leave of absence from the Scientific Department, Israel Ministry of Defence.) Mr. Ziv has been engaged in research in information theory and statistical communication theory. Member, IEEE.

B.S.T.J. BRIEFS

Sputtered Glass Waveguide for Integrated Optical Circuits

By J. E. GOELL and R. D. STANDLEY

(Manuscript received September 16, 1969)

A series of papers which appeared in the September 1969 issue of the Bell System Technical Journal treated the theory of dielectric waveguides and stressed the potential use of such media for optical communication circuits.[1-4] Here we report on the realization of low-loss, thin glass films which can be used for circuit fabrication. Methods of preparing planar films and waveguides having rectangular cross section are described along with the techniques used in evaluating their optical characteristics.

The films we used for waveguide fabrication have been prepared by RF Sputtering of suitable glasses. The sputtering system used was oil-diffusion pumped and had five-inch diameter electrodes. Oxygen was used as the sputtering gas. The best films obtained to date were made by sputtering Corning 7059 glass. For convenience, in the early stages of this work, laboratory slides have been used as substrates. Necessary steps were taken to ensure that the substrates were clean.

The index of refraction of the films was measured to be 1.62 by determining Brewster's Angle for the films as described by Abeles.[5] From the color of the film and by interferometer methods the film thickness was found to be about 0.3 μm.

The transmission loss of the films was measured by two methods. Both use prisms to launch a light beam into the film.[6,7] In method 1 it is assumed that the scattering centers in the films are uniformly distributed. A fiber optic probe is then used to measure the intensity of the light scattered at right angles to the film. In method 2, the intensity of the output beam is measured as a function of launcher position along the film. Method 2 appears least accurate due to variations in launching efficiency as a function of prism movement. Method 1 works well to losses of the order of 1 db per cm. Below this level, the variability in the strength of the scattering centers makes reliable measurements difficult. An increase in film length would partially overcome the difficulty of measuring low level scattering from random centers.

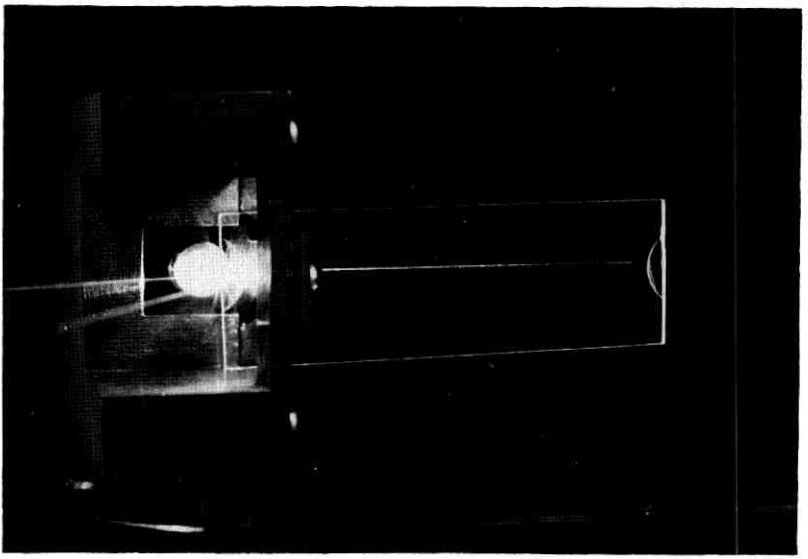

Fig. 1 — Light scattered from a beam propagating in a Corning 7059 glass film.

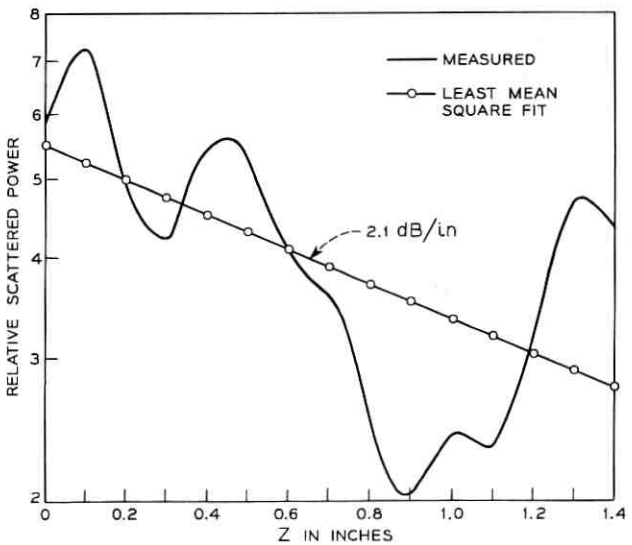

Fig. 2 — Relative scattered power versus length (7059 glass film).

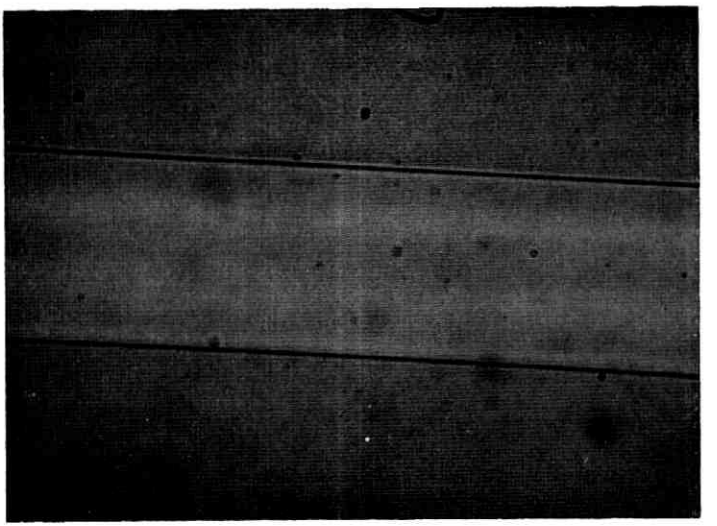

Fig. 3 — Section of a rectangular waveguide (×1000).

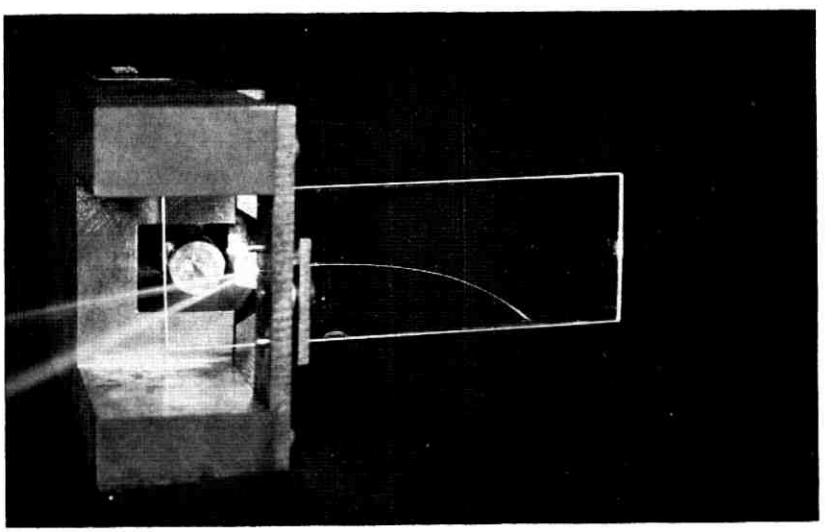

Fig. 4 — Light propagating in a curved section of rectangular waveguide.

Figure 1 is a picture of the light scattered from a beam propagating in the film. The intensity of scattered light as measured by the fiber optic probe is plotted in Figure 2. The average slope is less than -1 dB/cm. This result is in agreement with measurements made by the second method. The lack of uniformity of the scattered light intensity is due, at least in part, to inhomogeneities in the substrate. By using a higher quality substrate this source of scatter can be eliminated.

Curved sections of rectangular waveguides have been constructed from 7059 glass films by back-sputtering using quartz fibers as shadow masks. The waveguides were about 0.3 μm thick, 20 μm wide, and had a radius of curvature of about $\frac{1}{2}$ inch. A photograph of a typical section is shown in Figure 3. Figure 4 shows prism-launched light propagating in such a waveguide. Due to the small size of the waveguide our instrumentation will have to be improved before loss measurements can be made.

Our initial efforts have demonstrated the feasibility of using sputtered glass films and sputter etching in the fabrication of optical waveguides. This approach shows promise as a method of producing low-loss optical integrated circuits.

The authors are indebted to W. R. Sinclair for his valuable comments regarding the sputtering of glass films and the preparation of substrates, and to R. R. Murray who assisted in the preparation of the films and waveguides.

REFERENCES

1. Miller, S. E., "Integrated Optics: An Introduction," B.S.T.J., *48*, No. 7 (September 1969), pp. 2059–2069.
2. Marcatili, E. A. J., "Dielectric Rectangular Waveguide and Directional Coupler for Integrated Optics," B.S.T.J., *48*, No. 7 (September 1969), pp. 2071–2102.
3. Marcatili, E. A. J., "Bends in Optical Dielectric Guides," B.S.T.J., *48*, No. 7 (September 1969), pp. 2103–2132.
4. Goell, J. E., "A Circular-Harmonic Computer Analysis of Rectangular Dielectric Waveguides," B.S.T.J., *48*, No. 7 (September 1969), pp. 2133–2160.
5. Abeles, F., "Determination of the Refractive Index and the Thickness of Transparent Thin Films," J. Phys. Radium, *11*, No. 7 (July 1950), pp. 310–314.
6. Osterberg, H., and Smith, L. W., "Transmission of Optical Energy Along Surfaces: Part II, Inhomogeneous Media," J. Opt. Soc. Amer., *54*, No. 9 (September 1964), pp. 1078–1084.
7. Tien, P. K., Ulrich, R., and Martin, R. J., "Modes of Propagating Light Waves in Thin Deposited Semiconductor Films," Applied Physics Letters, *14*, No. 9 (May 1969), pp. 291–294.