THE BELL SYSTEM
TECHNICAL JOURNAL

Volume 47 July–August 1968 Number 6

The Transmission Distortion of a Source as a Function of the Encoding Block Length*

By R. J. PILC

(Manuscript received December 15, 1967)

This paper is concerned with the transmission of a discrete, independent letter information source over a discrete channel. A distortion function is defined between source output letters and decoder output letters and is used to measure the performance of the system for each transmission. The coding block length is introduced as a variable and its influence upon the minimum attainable transmission distortion is investigated.

The lower bound to transmission distortion is found to converge to the distortion level $\mathrm{d_C}$ *(C is the channel capacity) algebraically as* $\mathrm{a/n}$*. The nonnegative coefficient* a *is a function of both the source and channel statistics, which are interrelated in such a way as to suggest the utility of this coefficient as a measure of "mismatch" between source and channel, the larger the mismatch the slower the approach of the lower bound to the asymptote* $\mathrm{d_C}$ *. For noiseless channels* $\mathrm{a} = \infty$ *and for this case the lower bound is shown to converge to* $\mathrm{d_C}$ *as* $\mathrm{a_1(\ln n)/n}$.

For noisy channels the upper bound to transmission distortion is found to converge to the asymptote $\mathrm{d_C}$ *algebraically as* $\mathrm{b[(\ln n)/n]^{\frac{1}{2}}}$*. For noiseless channels, the upper bound converges to* $\mathrm{d_C}$ *as* $\mathrm{a_1(\ln n)/n}$.

[*] The material presented in this paper is based upon the author's thesis, "Coding Theorems for Discrete Source-Channel Pairs," presented to the Massachusetts Institute of Technology in November 1966 in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

I. INTRODUCTION

By now the results originally obtained by Shannon[1] relating reliability and channel capacity are well known. Roughly speaking, they state that perfect transmission can be achieved if, and only if, the capacity of the channel in the transmission link is greater than the information content of the source. For amplitude and time discrete sources the information content is the entropy of the source, but for amplitude continuous sources the entropy and the information content are not the same since the information content is infinite. This, of course, implies that perfect transmission of amplitude continuous sources, or discrete sources with an entropy that is "too large," is impossible with a given finite capacity channel. Yet this is just the situation that is often presented to the communication engineer who must then try to reduce the average distortion to the lowest possible, or practicable, level.

For communication systems in which the capacity of the channel is not sufficient to allow perfect transmission, there are two obvious questions to ask:

(*i*) How small can the average distortion be made if any transmission strategy at all is allowed?
(*ii*) How much does the system complexity, or cost, increase when you are required to get "closer" to this minimum?

To answer the first question, Shannon generalized his results in a later paper[2] in which the channel requirements are found that are necessary and sufficient to allow transmission at a given level of distortion, or a given error rate. It is our purpose here to consider the second question. We use the coding block length to measure the complexity of the system, and study the behavior of the minimum attainable transmission distortion as the block length is increased.

In the work we restrict our attention to sources and channels that are discrete in amplitude and time, and that are constant and memoryless. This means that successive events are independent and are governed by the same probability distributions. The encoder is a block encoder that we describe later in this section. To measure the distortion in the system, we introduce a nonnegative function $d(w,z)$ which gives the distortion in the event letter z is presented to the user at the decoder output when letter w was transmitted. Normally, this function would be specified by the user of the system to reflect how undesirable any particular misinterpretation of the source output is to him. We will assume that the distortion between two sequences of letters is the averaged sum of the composing letter distortions.

Shannon's theory associates with each source and distortion function a rate-distortion curve which expresses the minimum attainable transmission distortion in terms of the maximum allowable mutual information in the system. Associated with each point (d_R,R) on the rate-distortion curve is a particular set of transition probabilities, called the "test channel," which has the significance that among all channels that transmit the given source with distortion d_R or less, it operates at the lowest transmission rate, R. Equivalently, the test channel is that channel which yields the lowest distortion d_R among those that transmit information from the source at a rate R or less. It is in this sense the cheapest channel one could use and meet a distortion criterion. The rate R can also be interpreted as the equivalent information content of the source when a distortion d_R is tolerable.

That the rate-distortion curve gives the channel capacity sufficient to allow a prescribed performance is shown by Shannon through the intermediate step of proving that the rate-distortion curve actually expresses the entropy and resultant distortion in the "best" discrete representation of an output sequence from the original source. This discrete representation can then be transmitted with no further distortion, if its entropy is less than the channel capacity, by the use of suitable channel coding techniques.

Shannon has found the rate-distortion curves for many discrete sources and an explicit expression for this curve for time discrete gaussian sources. These results, together with Shannon's work with vector sources, were used to get rate-distortion curves for gaussian random processes.[3, 4] Bounds to the rate-distortion curve for non-gaussian sources have also been obtained.[5, 6]

However, all of the rate-distortion results derived for both continuous and discrete sources are limiting results, that is, they can be approached in general only when arbitrarily complex operations on very long sequences of source output are allowed before transmitting the "message" through a correspondingly large use of the channel. T. Goblick was the first to study the rate of approach to these limiting results as the source output block length increases, but limited his work to source representation or source encoding, with a deterministic map between the source and its representation.[7] Our work includes a noisy channel, or probabilistic function, between the source and user.

A performance curve $d(n)$ will be introduced for each source-channel pair as the minimum possible average distortion obtainable using a modulator that encodes a string of n successive source outputs into an input signal acceptable by a channel composed of n uses of the original channel. For a source with the rate-distortion curve of Fig. 1 and a channel with capacity C, the performance curve might look like the one shown in Fig. 2.

From Shannon's theory it is known that the performance curve starts at d_0, the zero-rate distortion, and decreases to asymptotically approach d_C, the distortion corresonding to the information rate C on the rate-distortion curve. The curve, of course, has meaning only for integral values of n. Not all modulators and decoders provide a distortion curve that approaches d_C for large n, but this curve obviously must lie above the performance curve which alternately could have been defined as the lower envelope to the set of distortion curves corresponding to all encoder-decoder pairs.

II. THE LOWER BOUND

Upper and lower bounds to the performance curve have been derived.[8] We present the lower bound in the first part of this paper, and the upper bound in Sections XI through XVII. Most of our effort concerning the lower bound was directed toward finding information about the rate of approach of the performance curve to its asymptote. In particular, we tried to relate the source and channel statistics, as well as the method of encoding that is used, to the rate of approach of $d(n)$ to d_C.

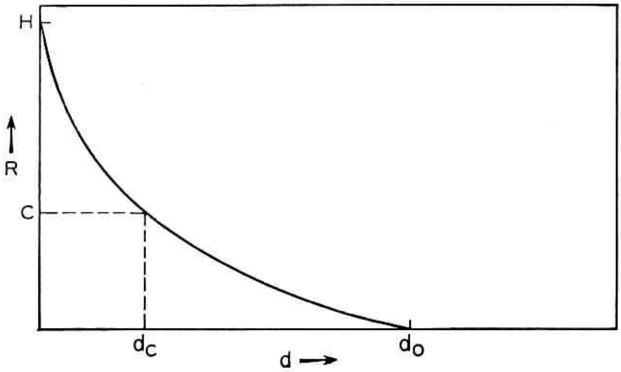


Fig. 1 — The rate distortion curve for $\mathcal{S}$.

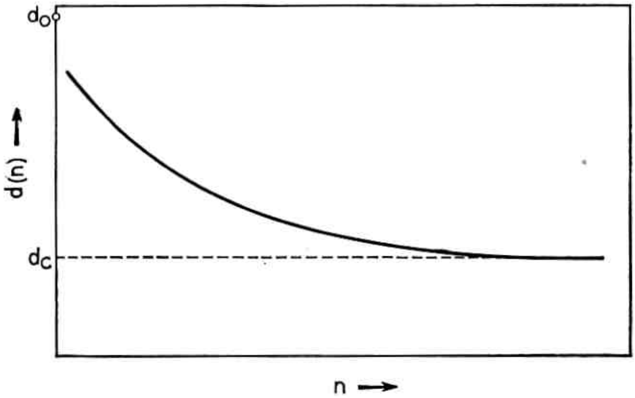


Fig. 2 — The performance curve for $\mathcal{S}$ and $\mathcal{C}$.

Concerning this rate of approach, several interesting situations are known to exist. For one, there are some source-channel pairs for which the minimum attainable transmission distortion is independent of the encoding block length, with the consequence that it is possible to attain the distortion level d_C with a coding block length of one. One example of such a pair is a binary symmetric source (equally likely binary letters with $d(i,j) = 1 - \delta_{ij}$, $i,j = 1,2$) used with a binary symmetric channel, where the optimum encoder is a direct connection. Another example is a gaussian source used with an additive gaussian noise channel, where the optimum encoder is simply an amplifier.[9]

When the source-channel pair is such that the minimum attainable distortion is independent of the coding block length we shall say that the source and channel are "matched." For the more common situation wherein the minimum attainable transmission distortion decreases with increasing encoding block length to asymptotically approach the distortion level d_C, we say that there is a "mismatch" between the source and channel, and suggest as a measure of this mismatch the "slowness" of the approach of the distortion to d_C.

Another interesting situation occurs when there is a choice of using one of several channels of different capacity. Although the channel of highest capacity would be the best choice when one is willing to use infinite block length coding, it might not be the best choice with finite length coding. This could easily happen if the high capacity channel were very much more mismatched to the source than some lower capacity channel.

III. SYSTEM MODEL

Figure 3 is a detailed illustration of the transmission system that we work with. The source $\mathcal{S}$ produces a sequence of letters $\boldsymbol{\omega} = \omega_1, \omega_2, \cdots, \omega_n$, each chosen from the alphabet $W = \{w_1, \cdots, w_H\}$, which is mapped by the encoder into a sequence of channel input letters $\boldsymbol{\xi} = \xi_1, \xi_2, \cdots, \xi_n$, each a member of $X = \{x_1, \cdots, x_K\}$. The channel then transforms the channel input word $\boldsymbol{\xi}$ into a sequence of channel output letters $\mathbf{n} = \eta_1, \eta_2, \cdots, \eta_n$ which are members of $Y = \{y_1, \cdots, y_L\}$, and $\mathbf{n}$ in turn is decoded by the receiver into a sequence $\boldsymbol{\zeta} = \zeta_1, \zeta_2, \cdots, \zeta_n$ of letters from the decoding space $Z = \{z_1, \cdots, z_J\}$.

The source and channel are both assumed to be constant and memoryless; therefore, successive events on each are independent and governed by the same probability distributions. In particular we have

$$p_{\omega}(\mathbf{w}) = \prod_{m=1}^{n} p_{\omega_m}(w^m)$$

$$p_{\mathbf{n}|\xi}(\mathbf{y} \mid \mathbf{x}) = \prod_{m=1}^{n} p_{\eta_m|\xi_m}(y^m \mid x^m),$$

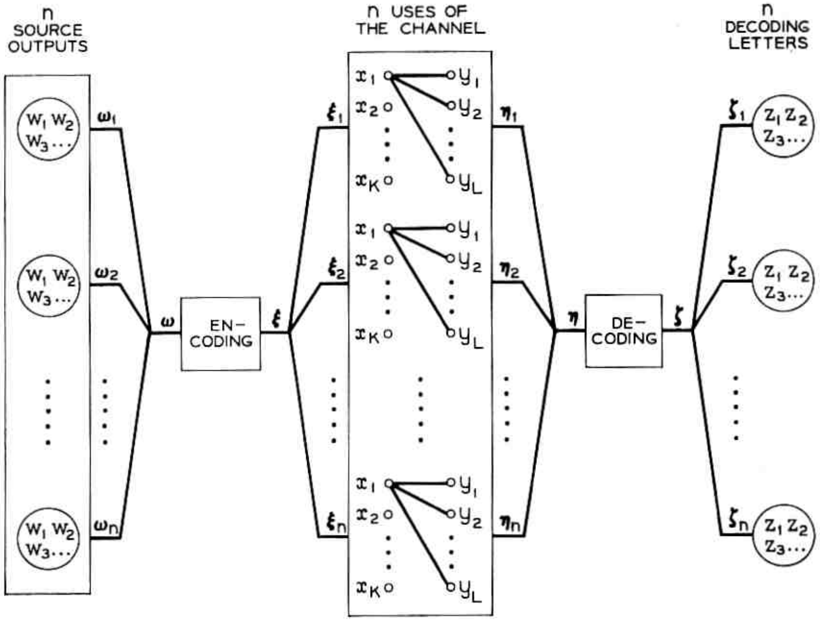


Fig. 3 — Block diagram of the encoding and decoding.

where the superscript on w^m, x^m, y^m is used to denote the m'th letter in the n-letter words **w**, **x**, **y** respectively, and is not to be confused with the particular letters w_m, x_m, and y_m in the alphabets W, X, and Y. The subscripts on the probability distribution are hereafter dropped whenever no confusion will occur.

The distortion in the system when the source word **w** is transmitted but received as **z** is taken to be the normalized sum of the n letter distortions, or

$$d(\mathbf{w}, \mathbf{z}) = \frac{1}{n} \sum_{m=1}^{n} d(w^m, z^m). \tag{1}$$

Finally, although we have set up the problem so that a sequence of n source letters is transmitted as a sequence of n channel letters, different block lengths at the source output and channel input can be allowed by considering a new source and channel that are products of the original ones, with the order of each product adjusted to obtain the desired block length ratio n_s/n_c.

IV. THE SPHERE PACKING ARGUMENT

A generalization of the sphere-packing concept is used to derive the lower bound. We assume the coding block length is n and derive a bound conditioned on the event that a particular source word **w** has occurred at the source output. We further assume that the channel input word **x** is used to transmit **w**, but delay the selection of **x** until the end of the derivation when the result is optimized over all possible choices. The total lower bound to distortion is found by averaging this conditioned lower bound over all source words in W^n. The asymptotic form of this bound is studied in detail and from it a measure of mismatch between the source and channel is defined.

The idea involved can be described with the following simple, but poor, bound which is subsequently improved. Remembering that the source word **w** is assumed transmitted by the channel input word **x**, we list all possible channel output words, **y**, ordered in decreasing conditional probability $p(\mathbf{y} \mid \mathbf{x})$, and pair with each the decoder output word $\mathbf{z}(\mathbf{y})$ to which it is decoded by the optimum decoder. The resulting (conditional) distortion,

$$d(\mathbf{w}) = \sum_{Y^n} p(\mathbf{y} \mid \mathbf{x})\, d[\mathbf{w}, \mathbf{z}(\mathbf{y})], \tag{2}$$

is seen to equal the sum of conditional probability-distortion products on this list. If the set of distortion values that appear on this list is now rearranged (with the list of conditional probabilities fixed) to be ordered according to increasing distortion values, the resulting sum of conditional probability-distortion products must be smaller than, or at most equal to, the sum in equation 2. It therefore provides a lower bound.

The improved lower bound uses the same sort of orderings and rearrangements but includes a probability function, $f(\mathbf{y})$, in the ordering of the channel output words. This function is defined over the set of channel output words, Y^n, and is later chosen to optimize the result. The channel output words are now ordered according to increasing values of the information difference $I(\mathbf{x}, \mathbf{y}) = (1/n) \ln [f(\mathbf{y})/p(\mathbf{y} \mid \mathbf{x})]$ and each is again paired with the decoder output word $\mathbf{z}(\mathbf{y})$ to which it is decoded by the optimum decoder.

The rearrangement of decoder output words is also slightly different. To describe this rearrangement we visualize each channel output word, $\mathbf{y}$, as "occupying" an interval of width $f(\mathbf{y})$ along the line $[0, 1]$. The decoder output word, $\mathbf{z}(\mathbf{y})$, that is paired with a particular channel output word $\mathbf{y}$ is also viewed as occupying the same region along $[0, 1]$ as $\mathbf{y}$, but, because any particular word $\mathbf{z}_o$ might be the decoding result of several channel output words, the region along $[0, 1]$ occupied by $\mathbf{z}_o$ could be a set of separated intervals. The rearrangement of decoder output words is this time a rearrangement of occupancies in $[0, 1]$ toward the desired configuration wherein the decoder words are ordered in increasing distortion along this line, and each occupies the same total width in $[0, 1]$ as it did before the ordering. Thus two monotone nondecreasing functions can be defined along the line $[0, 1]$; one, $I(h)$, giving the information difference $I(\mathbf{x}, \mathbf{y})$ at the point h, $0 \leqq h \leqq 1$, and the other, $d(h)$, giving the distortion $d(\mathbf{w}, \mathbf{z})$ at h. The first theorem presents a lower bound to the single word distortion in terms of these two functions.

Theorem 1: The average transmission distortion, $d(\mathbf{w})$, *conditioned on the occurrence of the source word* $\mathbf{w}$ *and its transmission using the channel input word* $\mathbf{x}$, *satisfies*

$$d(\mathbf{w}) \geqq \int_0^1 d(h) e^{-nI(h)} \, dh. \tag{3}$$

Proof: Figure 4 is used to help prove the inequality. The distortion resulting from optimum decoding is given by equation 2; the conditional probability-distortion products on the previous list *before* rearrangement of the decoder output words. For convenience this is

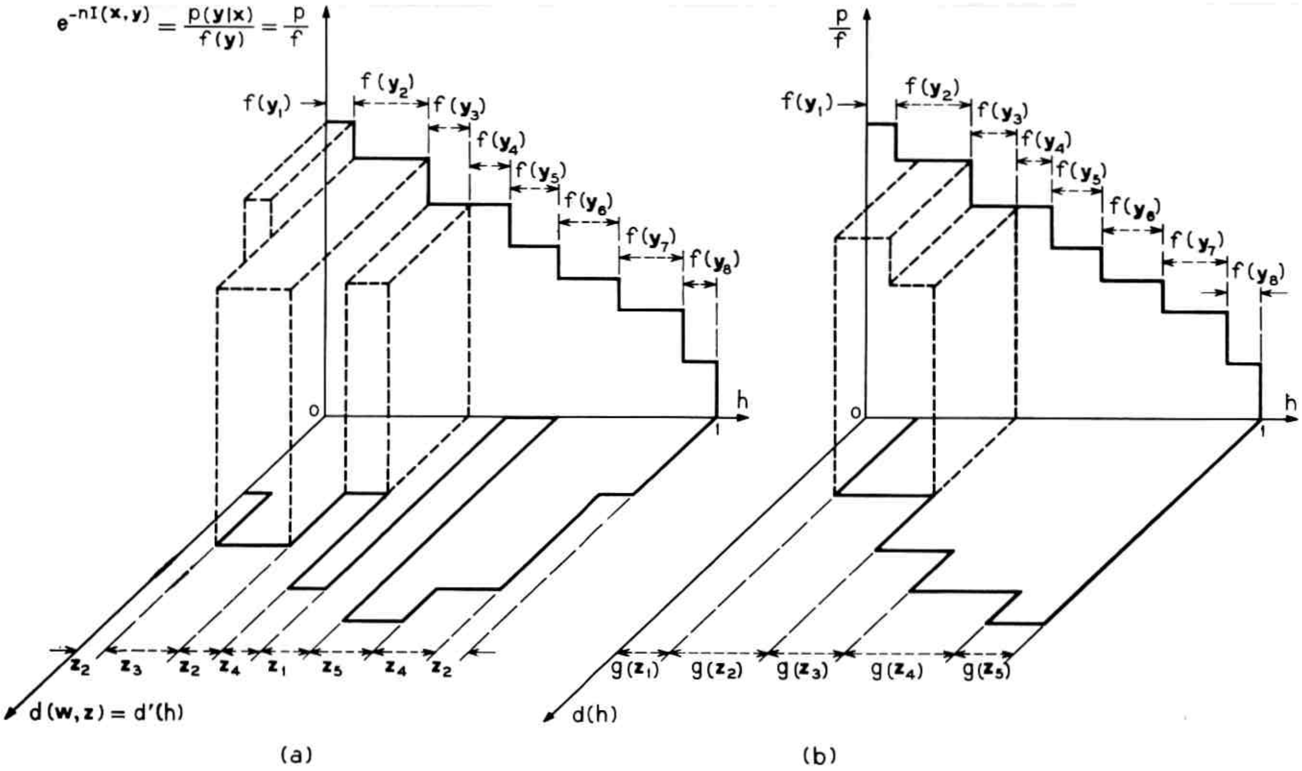


Fig. 4 — The geometry for theorem 1.

rewritten here as

$$d(\mathbf{w}) = \sum_{Y^n} d[\mathbf{w}, \mathbf{z}(\mathbf{y})]\left[\frac{p(\mathbf{y} \mid \mathbf{x})}{f(\mathbf{y})}\right] f(\mathbf{y}) \tag{4}$$

which can be seen equal to the "volume" in Fig. 4a enclosed by the two "amplitude functions" d' and p/f and the "width measure" f.

The rearrangement of the decoder output words to obtain the monotone function $d(h)$ from $d'(h)$ can be accomplished by a sequence of interchanges of the following type. We consider any two points in $0 \leqq h \leqq 1$, say h_1 and h_2 , for which $d'(h_2) \leqq d'(h_1)$ and $p/f(h_2) \leqq p/f(h_1)$. If we consider an interval Δh around each point in which both amplitude functions are single valued and interchange amplitude values of d' in the two intervals, we effect a volume transformation that decreases (or leaves unchanged) the total volume since

$$\begin{aligned}
&\text{initial volume—final volume} \\
&= \left[d'(h_1)\frac{p}{f}(h_1) + d'(h_2)\frac{p}{f}(h_2)\right]\Delta h \\
&\quad - \left[d'(h_2)\frac{p}{f}(h_1) + d'(h_1)\frac{p}{f}(h_2)\right]\Delta h \\
&= [d'(h_1) - d'(h_2)]\left[\frac{p}{f}(h_1) - \frac{p}{f}(h_2)\right]\Delta h \\
&\geqq 0.
\end{aligned}$$

Volume interchanges of this type are repeated until the desired monotonic function $d(h)$ is obtained. The resulting volume configuration is then as shown in Fig. 4b. As each interchange of Δh width volumes decreases the total volume, or leaves it unchanged, the total volume in Fig. 4b is certainly no larger than that in Fig. 4a. We need now only notice that $p/f(h) = \exp -nI(h)$ to recognize that the integral in equation 3 is equal to the volume in Fig. 4b, and, therefore, to establish the inequality claimed in the theorem.

To be sure, the construction in Fig. 4b, and the calculation of the lower bound in equation 2 requires some knowledge of the structure of the optimum decoder. Fortunately, this knowledge is minimal; it is only the total width along [0, 1] occupied by each member, $\mathbf{z}$, of the decoding space Z^n. We refer to this occupancy as the "size" of the decoding set for $\mathbf{z}$ and denote it by $g(\mathbf{z})$.

From the construction of the lower bound volume in Fig. 4b, we see that

$$g(\mathbf{z}) = \sum_{Y(\mathbf{z})} f(\mathbf{y})$$

where $Y(z)$ is the set of channel output words that are decoded into $\mathbf{z}$ by the optimum decoder. Indeed, if we assume unique decoding by the optimum decoder we have

$$\sum_{Z^n} g(\mathbf{z}) = \sum_{Z^n} \sum_{Y(\mathbf{z})} f(\mathbf{y}) = \sum_{Y^n} f(\mathbf{y}) = 1,$$

or that $g(\mathbf{z})$ is also a probability function. Even this function, though, is unknown in the general case or at least is impractical to calculate. The idea of the lower bound development, therefore, is to retain this unknown probability function for the present and subsequently replace it with another such function which minimizes the final lower bound expression. Within this step an approximation involving the form of $g(\mathbf{z})$ is required which is detailed in Section 6.2.

V. FURTHER EVALUATION OF THE LOWER BOUND IN THEOREM 1

The integral in equation 3 can be simplified if we suppress the intermediate variable h and relate the variables d and I directly. The pairings of d and I through a common value of h, $d(h) = I(h)$, does not by itself define a function because several different values of d could be paired with a given value of I, and vice versa. However, we will use the properties that exist among these pairs to define a distortion function $d(I)$ which has the property that for any I, the dependent variable d is at least as small as the smallest $d(h)$ among the pairs that have $I(h) = I$.

To do this, we reinterpret the monotone nondecreasing functions $d(h)$ and $I(h)$. First, we view the distortion $d(\mathbf{w}, \mathbf{z})$ as a random variable on Z^n governed by $g(\mathbf{z})$. Its cumulative distribution function

$$G(d) = \sum_{\substack{\mathbf{z} \ni \\ d(\mathbf{w},\mathbf{z}) \leq d}} g(\mathbf{z}) \tag{5}$$

is then seen to be the "inverse" of $d(h)$. (Strictly speaking, the inverse of a staircase function does not exist, so the term inverse is used here only as an aid in relating $d(h)$ and $G(d)$ pictorially.) In a similar way we also view the information difference $I(\mathbf{x}, \mathbf{y})$ as a random variable on Y^n governed by $f(\mathbf{y})$. Its cumulative distribution function is given by

$$F_1(I) = \sum_{\substack{\mathbf{y} \ni \\ I(\mathbf{x},\mathbf{y}) \leq I}} f(\mathbf{y}), \tag{6}$$

or the "inverse" of $I(h)$. The desired function $d(I)$ can now be defined in terms of $G(d)$ and $F_1(I)$ by relating to any information difference value I the distortion value that satisfies

$$F_1(I^-) = G(d). \tag{7}$$

The following geometric interpretation of $d(I)$ might be helpful. If each size, or "volume," $g(\mathbf{z})$ of the decoding sets is successively placed about the volume $g(\mathbf{z}_1)$ of the decoded word with minimum distortion $d(\mathbf{w}, \mathbf{z}_1)$, and each size, or "volume," $f(\mathbf{y})$ of the channel output words successively placed about the volume $f(\mathbf{y}_1)$ of the channel output word with minimum information difference $I(\mathbf{x}, \mathbf{y}_1)$, the total volume included by a point in the first construction at a distortion "radius" d is $G(d)$ and that included by a point in the second construction at an information difference "radius" I is $F_1(I)$. The function $d(I)$ then gives (except for edge effects) the correspondence between the radii that include the same volume in both geometrical constructions. Figure 5a illustrates the construction of $d(I)$ through the chain $I \rightarrow F_1(I^-) = G(d) \rightarrow d$.

It is convenient at this point to introduce a second random variable of information difference; one which is governed by $p(\mathbf{y} \mid \mathbf{x})$ rather than $f(\mathbf{y})$. Its cumulative distribution function is

$$F_2(I) = \sum_{\substack{\mathbf{y}\mathbf{;} \\ I(\mathbf{x},\mathbf{y}) \leqq I}} p(\mathbf{y} \mid \mathbf{x}). \tag{8}$$

To distinguish the two information difference variables, we will denote by I_1 the variable that has the distribution function in equation 6 and by I_2 the variable that has the distribution function in equation 8.

We are now in a position to rewrite the bound in Theorem 1 in terms of functions that involve only d and I. The distortion function $d(I)$ has been constructed to lower bound all $d(h)$ with $I(h) = I$, thus we can replace $d(h)$ in equation 3 with $d[I(h)]$. As this substitution replaces $d(h)$ with a distortion function that is single valued over subintervals of [0,1] in which I is a constant, we can perform the integration in equation 3 by simply multiplying the integrand in each such constant I interval by the interval width, $dF_1(I)$, and summing. Therefore, we can continue the inequality in equation 3 with

$$d(\mathbf{w}) \geqq \int_{I_{\min}}^{I_{\max}} d(I) \exp(-nI) \, dF_1(I),$$

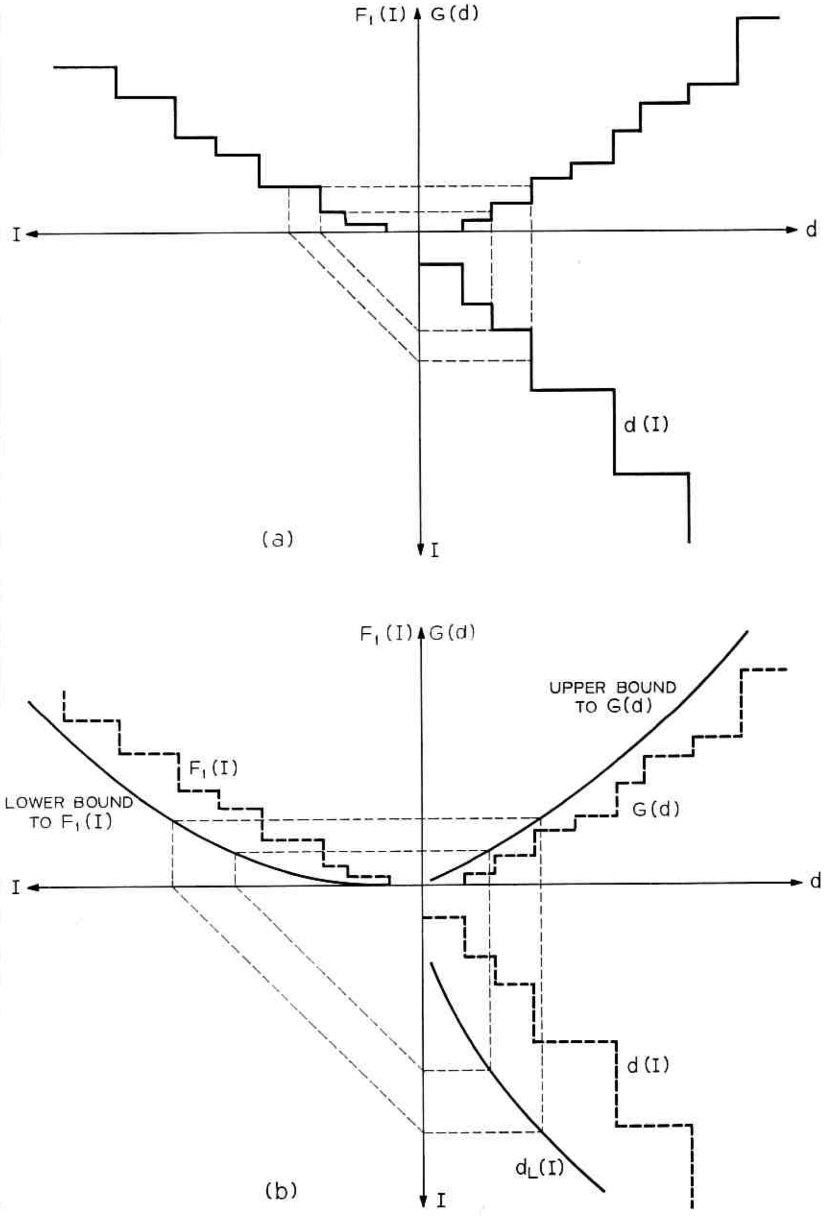


Fig. 5 — The construction of (a) $d(I)$ and (b) $d_L(I)$.

which, upon using $p(\mathbf{y} \mid \mathbf{x}) = \exp(-nI)f(\mathbf{y})$, establishes the lower bound in the next theorem.

Theorem 2: The average transmission distortion, d(**w**), *conditioned on the occurrence of the source word* **w** *and its transmission using the channel input word* **x**, *satisfies*

$$d(\mathbf{w}) \geqq \int_{I_{\min}}^{I_{\max}} d(I)\, dF_2(I). \tag{9}$$

VI. AN ESTIMATE OF THE FUNCTION $d(I)$

6.1 *The Random Variables* I$_1$ *and* I$_2$

To obtain an estimate of $d(I)$ we require an estimate of the two distribution functions, $G(d)$ and $F_1(I)$, from which $d(I)$ was defined. We first focus on $F_1(I)$ and the random variable I_1. Since the lower bounds in Theorems 1 and 2 can be derived for any choice of $f(\mathbf{y})$, we choose a form of $f(\mathbf{y})$ that simplifies the following arguments. We specify that $f(\mathbf{y})$ factors as

$$f(\mathbf{y}) = \prod_{m=1}^{n} f(y^m). \tag{10}$$

One consequence of this assumed form is that the information difference $I(\mathbf{x}, \mathbf{y})$ is given as a sum of n letter information differences:

$$I(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{m=1}^{n} \ln \frac{f(y^m)}{p(y^m \mid x^m)} = \frac{1}{n} \sum_{m=1}^{n} I(x^m, y^m). \tag{11}$$

Among these n letter information differences, however, there are different types, depending on the corresponding transmitted letter x^m in **x**. To separate these, we introduce the vector **c** to denote the letter composition of the channel input word **x**, letting $\mathbf{c} = c_1, c_2, \cdots, c_K$ when there are nc_1 appearances of the letter x_1 in **x**, nc_2 appearances of x_2 in **x**, and so on. Thus we can write the information difference in equation 10 as

$$I(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{r=1}^{nc_k} I_{kr} \tag{12}$$

in which I_{kr} is used to denote the information difference between the r'th appearance of the letter x_k in **x** and the corresponding letter in **y**. The interpretation of the I_{kr} as letter information difference random variables on Y governed by the letter probability function $f(y)$ can now be seen to be consistent with the previous interpretation of I_1

as a word information difference random variable on Y^n governed by $f(\mathbf{y})$. Using the abbreviations

$$f(y_l) = f_l$$
$$p(y_l \mid x_k) = p_{kl} \,,$$

the probability distribution function of I_{kr} can be written as

$$P_{1,I_{kr}}\left[\ln \frac{f_l}{p_{kl}}\right] = f_l \;; \qquad 1 \leqq r \leqq nc_k \;; \qquad 1 \leqq k \leqq K. \tag{13}$$

What this has accomplished is to cast I_1 as the sum of n independent random variables, a step that enables us to use large number laws to estimate $F_1(I)$.[10-13]

In an almost identical way, the random variable I_2 can be cast as a sum of n independent random variables. This can be done if we associate with the variable I_{kr} the probability distribution function

$$P_{2,I_{kr}}\left[\ln \frac{f_l}{p_{kl}}\right] = p_{kl} \;; \qquad 1 \leqq r \leqq nc_k \;; \qquad 1 \leqq k \leqq K \tag{14}$$

instead of that in equation 13. With this distribution the word information difference variable $I(\mathbf{x}, \mathbf{y})$ in equation 12 can be seen to be governed by the probability function $p(\mathbf{y} \mid \mathbf{x})$, therefore, it is equal to the random variable I_2 .

6.2 *The Random Variable* d

In the work so far, the function $g(\mathbf{z})$ is that probability function induced on Z^n by $f(\mathbf{y})$ through the optimum decoder function and cannot, therefore, be freely chosen once $f(\mathbf{y})$ is chosen. On the other hand its precise calculation from the optimum decoder is impractical. The only alternative is to retain the unknown function $g(\mathbf{z})$ in the lower bound expressions and to minimize the final lower bound to distortion over all possible probability functions on Z^n. Since $g(\mathbf{z})$ is one such probability function the inequality in the lower bound is continued. Unfortunately, when this is done it cannot, in general, be shown that the function which minimizes the lower bound factors into n letter probabilities, a form which we were permitted to assume for $f(\mathbf{y})$. However, to proceed beyond the bounds in Theorems 1 and 2, it is necessary to approximate this $g(\mathbf{z})$ by such a product, as in

$$g(\mathbf{z}) = \prod_{m=1}^{n} g(z^m). \tag{15}$$

The necessity for an approximation of this type is, of course, because of the requirement that an estimate be made for the distribution function $G(d)$. The assumed form for $g(\mathbf{z})$ in equation 15, will again allow us to use large number laws to obtain this estimate.

More specifically, the assumed product form for $g(\mathbf{z})$ allows us to cast the word distortion random variable $d(\mathbf{w}, \mathbf{z})$ as a sum of n independent letter variables. This is done in the following way. Among the letter distortions $d(w^m, z^m)$ that sum to the total word distortion there are H different types, corresponding to each of the different letters w_i, $1 \leqq i \leqq H$, that appear in the source word $\mathbf{w}$.

If the composition of this word is $\mathbf{q} = q_1, q_2, \cdots q_H$, that is, if there are nq_1 appearances of w_1 in $\mathbf{w}$, nq_2 appearances of w_2, and so on, the normalized word distortion can be written as

$$d(\mathbf{w}, \mathbf{z}) = \frac{1}{n} \sum_{i=1}^{H} \sum_{r=1}^{nq_i} D_{ir} . \tag{16}$$

In this expression D_{ir} is used to denote the distortion between the r'th appearance of the letter w_i in $\mathbf{w}$ and the corresponding letter in $\mathbf{z}$. Equation 15 now allows the interpretation of the D_{ir} as independent random variables, having the probability distributions

$$\begin{aligned} P_{D_{ir}}(d_{ij}) &= g_j ; \qquad 1 \leqq r \leqq nq_i , \quad 1 \leqq i \leqq H \\ d(w_i , z_j) &= d_{ij} \\ g(z_j) &= g_j , \end{aligned} \tag{17}$$

with the result that $G(d)$ is an n-fold convolution of elementary distribution functions for which there exist many estimating forms.[10–13]

We realize that the approximation in equation 15 is not entirely satisfactory because it eliminates nonproduct probability functions from the minimization of the lower bound and, as far as we know, one of these functions could provide the minimization. However, there is good reason to believe that this approximation does not significantly affect the bound when n is reasonably large. For example, in the next several sections we derive a lower bound to distortion that uses the product from in equation 15. For this bound the required minimization over all probability functions $g(\mathbf{z})$ is reduced to one over all J dimensional vectors $\mathbf{g}$. It can be shown that if in the limit as n becomes large, the product form requirement for $g(\mathbf{z})$ is relaxed, and the minimization of this lower bound is again made over all probability functions $g(\mathbf{z})$, then the optimizing function $g_o(\mathbf{z})$ still has the product form.

Even more significant is the asymptotic form of the lower bound that is derived using equation 15. We later show that it is *only* the final value of the minimizing decoder set size vector $\mathbf{g}_o(n = \infty)$ that affects *both* the asymptote of the lower bound, d_C, and the next lowest order term, which is one proportional to $1/n$. Values of the minimizing vector for finite n, $\mathbf{g}_o(n < \infty)$, affect only terms of $o(1/n)$.

Further, it can be shown that a similar conclusion is reached even if the independence property assumed over letters in equation 15 is generalized to be over blocks of length r, that is if

$$g(\mathbf{z}) = \prod_{m=1}^{n/r} g(\mathbf{z}'^m)$$

$$\mathbf{z}'^m = z_j, z_{j+1}, \cdots, z_{j+r-1}; \qquad j = mr - r + 1.$$

When $g(\mathbf{z})$ is assumed to have this form, the minimization of the lower bound over all decoder set sizes is a minimization over all probability functions $g(\mathbf{z}')$ on Z^r. The conclusion that can be made from the bound derived using this assumption is that it is again *only* the value of the minimizing decoder set size function at $n = \infty$, $g_o(\mathbf{z}', \infty)$, that influences both the asymptote and the term proportional to $1/n$. And, at $n = \infty$, the minimizing decoder set size function on Z^r, $g_o(\mathbf{z}', \infty)$, factors into a product of single letter probability functions on Z. When this solution is substituted in the bound (that uses $r \geqq 1$) the *asymptotic* form is the same for *every* choice of the constant r. Only lower order terms differ for different values of r.

There is one situation in which the assumed product form in equation 15 does not represent an approximation. That is the case of a doubly uniform source, which is a source that has a uniform probability distribution over its letters and has a distortion matrix in which each row and column is the respective permutation of another row and column. For such a source it has been shown[8] that the probability distribution $g(\mathbf{z})$ which minimizes the lower bound in Theorem 1 is uniform for all n, thus has the factorability property in equation 15.

6.3 *A Lower Bound to* d(I)

We now seek an approximation to $d(I)$ that we can substitute in equation 9 and preserve the inequality. A safe approximation to $d(I)$ can be had if, instead of equating $F_1(I^-)$ to $G(d)$ as in equation 7, we equate a lower bound estimate of $G(d)$ to an upper bound estimate of $F_1(I^-)$. Figure 5b illustrates this construction. The result is another distortion function, $d_L(I)$, that satisfies

$$d_L(I) \leqq d(I) \tag{18}$$

which can be used in equation 9 to obtain

$$d(\mathbf{w}) \geqq \int_{I_{\min}}^{I_{\max}} d_L(I)\, dF_2(I). \tag{19}$$

Since the random variable I_2 is a normalized sum of n independent random variables, its variance is proportional to $1/n$. Consequently, when n becomes large the distribution function $F_2(I)$ has almost all of its "rise" around the mean of I_2, which we denote by $\bar{\mathrm{I}}$. In this region, $I \approx \bar{\mathrm{I}}$, $d \approx d(\bar{\mathrm{I}})$, the values of both distribution functions $G(d)$ and $F_1(I)$ are exponentially small. Therefore, the bounds to the tails of distribution functions[10–13] are applicable to the estimation of $G(d)$ and $F_1(I)$ in this region. Indeed, it was with the intended use of these powerful bounds that we formed both the distortion and information difference random variables as sums of n independent letter random variables. All of the bounds, though, are parametric in form and allow only a parametric representation of $d_L(I)$.

We have elsewhere[8] applied strict upper and lower bounds to $G(d)$ and $F_1(I)$, respectively, to obtain the function $d_L(I)$. However, when these bounds are used, the resulting total lower bound to transmission distortion, though applicable for all block lengths n, does not reveal the correct asymptotic behavior inherent to the sphere-packing procedure which has been used. (This happens because the strict bounds to $G(d)$ and $F_1(I)$ themselves do not have the correct asymptotic form to large n.)

In addition, the resulting lower bound to the total distortion is very complex and so does not provide much insight into the factors which affect the rate of approach of the performance curve to its asymptote. For these reasons, we instead use Shannon's[11] and Gallager's[13] asymptotic forms for the tails of distribution functions to bound $G(d)$ and $F_1(I)$. These are:

$$G(d) \leqq \left[\frac{1}{\sqrt{2\pi n s^2 \mu''(s)}} + A_U(n, s) \right] \exp n[\mu(s) - s\mu'(s)] \tag{20a}$$

$$\mu'(s) = d \tag{20b}$$

with

$$0 < d \leqq E(d \mid \mathbf{q}) = \sum_{Z^n} d(\mathbf{w}, \mathbf{z} \mid \text{comp } \mathbf{w} = \mathbf{q}) g(\mathbf{z}),$$

and

$$F_1(I) \geqq \left[\frac{1}{\sqrt{2\pi n t^2 \gamma''(t)}} + A_L(n, t)\right] \exp n[\gamma(t) - t\gamma'(t)] \tag{21a}$$

$$\gamma'(t) = I \tag{21b}$$

with

$$I_{\min} < I \leqq E(I_1 \mid \mathbf{c}) = \sum_{Y^n} I(\mathbf{x}, \mathbf{y} \mid \text{comp } \mathbf{x} = \mathbf{c}) f(\mathbf{y}).$$

In these bounds, $A_U(n, s)$ and $A_L(n, t)$ are sums of rather difficult integrals but each has been shown by Shannon and Gallager to be

$$o\left(\frac{1}{\sqrt{n}}\right).$$

Also within the previous bounds, we have used $\mu(s)$ to denote the semi-invariant moment generating function of the variable d,

$$\begin{aligned} \mu(s) &= \sum_{i=1}^{H} q_i \mu_i(s) \\ &= \sum_{i=1}^{H} q_i \ln \sum_{j=1}^{J} g_j \exp s\, d_{ij}\,, \end{aligned} \tag{22}$$

and $\gamma(t)$ to denote the semi-invariant moment generating function of the variable I,

$$\begin{aligned} \gamma(t) &= \sum_{k=1}^{K} c_k \gamma_k(t) \\ &= \sum_{k=1}^{K} c_k \ln \sum_{l=1}^{L} f_l^{1+t} p_{kl}^{-t}\,. \end{aligned} \tag{23}$$

To guarantee the boundedness of $\gamma(t)$, we restrict the vector $\mathbf{f}$ to have nonzero components. This does not affect the resulting bound. (Actually, these bounds strictly apply only when the variables d and I are nonlattice. For lattice variables the corresponding bounds[11,13] have in their coefficient a quantity Δ which does not change continuously with the argument of the distribution function, and cannot be used within our derivation. One alternative would be to decrease one assigned letter distortion $d(w, z)$ by an arbitrarily small irrational number, and similarly, to change two transition probabilities on the channel in a way consistent with a lower bound to distortion. The new variables d' and I' would then be nonlattice.)

The desired distortion function, $d_L(I)$, can now be defined by equating the two bounds in equations 20 and 21. It can be constructed through the chain: $I^- \to t \to s \to d$ in which the superscript could now be dropped since the bound to $F_1(I)$ is continuous in I. It is important to notice that the region of validity of the previous two bounds allows definition of the function $d_L(I)$ only in a subinterval $[I_a, \mathrm{I}_b]$ of $[I_{\min}, I_{\max}]$ with

$$I_{\min} < I_a < \bar{\mathrm{I}} < I_b \leqq E(I_1 \mid \mathbf{c}), I[E(d \mid \mathbf{q})].$$

Outside the interval $[I_a, I_b]$ we can define $d_L(I)$ equal to zero and write the lower bound in equation 19 as

$$d(\mathbf{w}) \geqq \int_{I_a}^{I_b} d_L(I)\, dF_2(I). \tag{24}$$

We are now faced with the difficult integration of a doubly parametric expression. Rather than integrate directly, we use the following Taylor series expansion for $d_L(I)$ within $[I_a, I_b]$:

$$\begin{aligned} d_L(I) &= d_L(\bar{\mathrm{I}}) + d_L'(\bar{\mathrm{I}})(I - \bar{\mathrm{I}}) + \tfrac{1}{2}\, d_L''(\bar{\mathrm{I}})(I - \bar{\mathrm{I}})^2 + \tfrac{1}{6}\, d_L'''(I')(I - \bar{\mathrm{I}})^3 \\ &\equiv TS(d_L) \end{aligned}$$

with $I_a \leq I' \leq I_b$. (The indicated derivatives can be shown to exist within the restricted interval $[I_a, I_b]$.) Using this form for $d_L(I)$ within equation 24 we see that if the region of integration were $[I_{\min}, I_{\max}]$ instead of $[I_a, I_b]$, the resulting form would be a sum of central moments of I_2 with the Taylor series derivatives as coefficients. To restore this form we rewrite equation 24 as

$$d(\mathbf{w}) \geqq \int_{I_{\min}}^{I_{\max}} \cdots - \int_{I_{\min}}^{I_a} \cdots - \int_{I_b}^{I_{\max}} TS(d_L)\, dF_2(I). \tag{25}$$

In these integrals, the lower limit $I_{\min}$ is finite since f_l is assumed nonzero for all l, and $I_{\max}$ can be taken as the largest finite value of $\ln f_l/p_{kl}$ since this is the largest value of I for which the random variable I_2 has nonzero probability. Therefore the function $TS(d_L)$ is bounded in $[I_{\min}, I_a]$ and $[I_b, I_{\max}]$ with the result that the last two integrals in equation 25 are exponentially small in n. The first integral in this equation has the desired form, involving the central moments of I_2:

$$\begin{aligned} \int_{I_{\min}}^{I_{\max}} TS(d_L)\, dF_2(I) = d_L(\bar{\mathrm{I}}) + d_L'(\bar{\mathrm{I}})E(I - \bar{\mathrm{I}}) + \tfrac{1}{2}\, d_L''(\bar{\mathrm{I}})E[(I - \bar{\mathrm{I}})^2] \\ + \tfrac{1}{6}\, d_L'''(I')E[(I - \bar{\mathrm{I}})^3]. \end{aligned}$$

In the above equation the second term is zero since we have specified that $\bar{\mathrm{I}}$ is the expected value of I_2, and the last term can be shown to be proportional to $(1/n)^2$. This establishes the result in the next theorem.

Theorem 3: The conditional average transmission distortion, $d(\mathbf{w})$, *satisfies*

$$d(\mathbf{w}) \geqq d_L(\bar{\mathrm{I}}) + \tfrac{1}{2}\, d_L''(\bar{\mathrm{I}}) \operatorname{var}(I_2) + o\left(\frac{1}{n}\right). \tag{26}$$

Compared with the last low order term, the variance of I_2 is proportional to $1/n$.

The simplicity in the form of the last result is due to the use of the Taylor series expansion which not only has allowed us to evaluate a difficult integral, but has provided a natural way of separating the important terms in the lower bound to distortion.

6.4 *The Evaluation of* $\mathrm{d}_\mathrm{L}(\bar{\mathrm{I}})$ *and* $\mathrm{d}_L''(\bar{\mathrm{I}})$

We shall denote by s_o and t_o the parameter values consistent with $I = \bar{\mathrm{I}}$ in equations 20 and 21. Since

$$\gamma'(-1) = \sum_{k=1}^{K} \sum_{l=1}^{L} p_{kl} \ln f_l/p_{kl} \,,$$

which is seen equal to $E(I_2) = \bar{\mathrm{I}}$, we can conclude that $t_o = -1$. We also note here for future use that

$$\gamma(-1) = 0.$$

The first of the two significant terms in equation 26 is immediate:

$$d_L(\bar{\mathrm{I}}) = \mu'(s_o).$$

Next, elementary differentiation of the parametric expressions in equations 20 and 21 provides

$$d_L'(\bar{\mathrm{I}}) = \frac{t}{s}\bigg|_{t_o,s_o}$$

$$= -\frac{1}{s_o}$$

and

$$d_L''(\bar{\mathrm{I}}) = \frac{1}{s}\left[\frac{1}{\gamma''(t)} - \frac{t^2}{s^2\mu''(s)}\right]\bigg|_{t_o,s_o}$$

$$= \frac{1}{s_o}\left[\frac{1}{\gamma''(-1)} - \frac{1}{s_o^2\mu''(s_o)}\right].$$

Finally, the variance of I_2 is seen from equation 12 to equal

$$\begin{aligned}\text{Var}\,(I_2) &= \frac{1}{n}\sum_{k=1}^{K} c_k \,\text{Var}\,(I_{kr}) \\ &= \frac{1}{n}\sum_{k=1}^{K} c_k\left[\sum_{l=1}^{L} p_{kl}(\ln f_l/p_{kl})^2 - \left(\sum_{l=1}^{L} p_{kl}\ln f_l/p_{kl}\right)^2\right] \\ &= \gamma''(-1).\end{aligned}$$

With the substitution of these terms in equation 26 we obtain the result in the next theorem.

Theorem 4: The conditional average transmission distortion, $d(\mathbf{w})$, satisfies

$$d(\mathbf{w}) \geqq \mu'(s_o) - \frac{1}{2ns_o}\left[\frac{\gamma''(-1)}{s_o^2\mu''(s_o)} - 1\right] + o\left(\frac{1}{n}\right) \tag{27}$$

in which s_o is given by

$$\mu(s_o) - s_o\mu'(s_o) = \bar{\text{I}} - \frac{1}{2n}\ln\frac{\gamma''(-1)}{s_o^2\mu''(s_o)} + o\left(\frac{1}{n}\right). \tag{28}$$

It remains to average this lower bound over the entire source space W^n.

VII. THE AVERAGE OVER THE SOURCE SPACE

To average the lower bound in Theorem 4 over the source space W^n we assume that channel input words of equal composition are used for all transmissions. It has been shown[8] that this assumption does not affect the asymptotic form of the lower bound to distortion. We first notice that the lower bound in Theorem 4 depends upon the source word $\mathbf{w}$ only through its composition $\mathbf{q}$ which enters in the form of $\mu(s)$. Therefore, we can average $d(\mathbf{w})$ over the set of all compositions for $\mathbf{w}$ rather than over all of W^n. As all composition vectors for $\mathbf{w}$ are probability vectors, they are all located on an $H - 1$ dimensional hyperplane, termed the composition space Q^H, which is in the "first quadrant" of R^H and intersects each axis q_i at one. Not all points in Q^H are possible word compositions for any particular n. For example, with $H = 2$ and $n = 2$ there are only three possible compositions. But as n increases, the points in Q^H that are source word compositions become quite dense.

The probability that any particular composition $\mathbf{q}$ occurs at the source output is

$$P(\mathbf{q}) = N(\mathbf{q}) \prod_{i=1}^{H} p_i^{nq_i} \tag{29}$$

in which $N(\mathbf{q})$ is the number of distinct source sequences with the composition $\mathbf{q}$ and the product is the probability of each. The number $N(\mathbf{q})$ is given by

$$N(\mathbf{q}) = \frac{n\,!}{\prod_{i=1}^{H} (nq_i)\,!}.$$

We now write the total average source distortion, $d(\mathcal{S})$, as

$$d(\mathcal{S}) = \sum_{\substack{\text{all source}\\ \text{compositions}}} d(\mathbf{q})P(\mathbf{q})$$

which we can lower bound by substituting for $d(\mathbf{q})$ the lower bound found in Theorem 4. Rather than write out the entire expression each time we want to use it, we let $d_L(\mathbf{q})$ denote the right side of equation 27, thus have

$$d(\mathcal{S}) \geqq \sum_{\substack{\text{all source}\\ \text{compositions}}} d_L(\mathbf{q})P(\mathbf{q}). \tag{30}$$

Viewed as a function over Q^H, $P(\mathbf{q})$ is a set of impulses. This allows us to consider the distortion function $d_L(\mathbf{q})$ a continuous function over all Q^H, rather than a function defined only at composition points, and to write

$$d(\mathcal{S}) \geqq \int_{Q^H} \cdots \int d_L(\mathbf{q})P(\mathbf{q})\, d\mathbf{q}. \tag{31}$$

Again because the expression for $d_L(\mathbf{q})$ in equations 27 and 28 is parametric, we use a Taylor series expansion of this distortion function to evaluate the integral. The point chosen for the expansion is $\mathbf{p}$, the probability vector characterizing the source. The reason for this choice is that the components of this vector are the means of the coordinates of $\mathbf{q}$ when the latter are considered (dependent) random variables governed by $P(\mathbf{q})$. The Taylor series then contains terms of the type $(q_i - p_i)$, $(q_i - p_i)(q_j - p_j)$, and so on, which, when averaged by $P(\mathbf{q})$, are the central moments of the components of $\mathbf{q}$.

Using the notation $d'_{Li}(\mathbf{p})$ to indicate the partial derivative of $d_L(\mathbf{q})$ with the respect to q_i evaluated at $\mathbf{q} = \mathbf{p}$ (and similarly for higher order derivatives), we have

$$\begin{aligned} d(\mathbb{S}) \geqq \int_{Q^H} \cdots \int \Big[d_L(\mathbf{p}) &+ \sum_{i=1}^{H} d'_{Li}(\mathbf{p})(q_i - p_i) \\ &+ \tfrac{1}{2} \sum_{ij} d''_{Lij}(\mathbf{p})(q_i - p_i)(q_j - p_j) \\ &+ \tfrac{1}{6} \sum_{ijk} d'''_{Lijk}(\varphi)(q_i - p_i)(q_j - p_j)(q_k - p_k) \Big] P(\mathbf{q})\, d\mathbf{q} \end{aligned} \tag{32}$$

with $\varphi \in Q^H$. The central moments of the components of $\mathbf{q}$ can be found to be

$$\begin{aligned} E(q_i - p_i) &= 0, \\ E[(q_i - p_i)(q_j - p_j)] &= \frac{1}{n}(p_i\ \delta_{ij} - p_i p_j) \\ E[(q_i - p_i)(q_j - p_j)(q_k - p_k)] & \\ = \left(\frac{1}{n}\right)^2 [p_i\ \delta_{ijk} - p_i p_j\ \delta_{ki} &- p_j p_k\ \delta_{ij} - p_k p_i\ \delta_{jk} + 2p_i p_j p_k], \end{aligned} \tag{33}$$

which, when substituted in equation 32, yields

$$d(\mathbb{S}) \geqq d_L(\mathbf{p}) + \frac{1}{2n} \Big[\sum_i d''_{Lii}(\mathbf{p}) p_i - \sum_{ij} d''_{Lij}(\mathbf{p}) p_i p_j \Big] + o\left(\frac{1}{n}\right). \tag{34}$$

Referring to equation 27 we see that the required second derivative need only be taken of $\mu'(s_o)$ as the two $1/n$ coefficients allow other terms to be absorbed in those of $o(1/n)$. The differentiation is lengthy, but straightforward, and yields

$$\frac{\partial}{\partial q_i} \mu'(s_o\,, \mathbf{q}) = \frac{\mu_i(s_o)}{s_o}$$

and

$$\frac{\partial^2}{\partial q_i\, \partial q_j} \mu'(s_o\,, \mathbf{q}) = -\frac{\theta_i \theta_j}{s_o^3 \mu''(s_o\,, \mathbf{p})}$$

where

$$\theta_i \equiv \mu_i(s_o) - s_o \mu'_i(s_o).$$

Upon substitution of these derivatives in equation 34 we obtain

$$\begin{aligned} d(\mathbb{S}) \geqq d_L(\mathbf{p}) &- \frac{1}{2ns_o^3 \mu''(s_o)} \Big[\sum_i p_i \theta_i^2 - \sum_{ij} p_i p_j \theta_i \theta_j \Big] + o\left(\frac{1}{n}\right) \\ &= d_L(\mathbf{p}) - \frac{1}{2ns_o^3 \mu''(s_o)} \operatorname{Var}(\theta) + o\left(\frac{1}{n}\right). \end{aligned}$$

With the final substitution of the expression for $d_L(\mathbf{p})$ in equation 27 we have the result in the next theorem.

Theorem 5: The average transmission distortion of the source $\mathcal{S}$, when used with the channel $\mathcal{C}$, is lower bounded by

$$d(\mathcal{S}) \geqq \mu'(s_o\,, \mathbf{p}) - \frac{1}{2ns_o}\left[\frac{\gamma''(-1) + \sigma^2(\theta)}{s_o^2\mu''(s_o\,, \mathbf{p})} - 1\right] + o\left(\frac{1}{n}\right) \tag{35}$$

in which s_o is given by

$$\mu(s_o\,, \mathbf{p}) - s_o\mu'(s_o\,, \mathbf{p}) = \bar{\mathrm{I}} - \frac{1}{2\mathrm{n}} \ln \frac{\gamma''(-1)}{s_o^2\mu''(s_o\,, \mathbf{p})} + o\left(\frac{1}{n}\right). \tag{36}$$

In this bound the vector **g** is, for the reasons previously stated, that which minimizes the bound, the vector **f** is chosen to maximize the bound in order to obtain the tightest bound, and the vector **c** is chosen to minimize the bound, that is to use the best composition for the channel input code words. As formidable as the derivations of these extremum appear, we show in the next section that the work involved in establishing the asymptotic behavior of the bound is actually quite simple.

It should be mentioned that these results do *not* apply when $\gamma''(-1) = 0$, which is a situation that occurs when channel $\mathcal{C}$ is noiseless, for the reason that we have divided by and canceled factors equal to $\gamma''(-1)$. The result for this case is derived separately in Section IX.

VIII. THE ASYMPTOTE AND RATE OF APPROACH

8.1 *The Asymptote*

When n becomes large, the limiting form of the bound in Theorem 5 is:

$$d_\infty(\mathcal{S}) \geqq \mu'(s_o\,, \mathbf{p})$$

in which s_o satisfies

$$\mu(s_o\,, \mathbf{p}) - s_o\mu'(s_o\,, \mathbf{p}) = \bar{\mathrm{I}}$$

with

$$\bar{\mathrm{I}} = \sum_{k=1}^{K} c_k \sum_{l=1}^{L} p_{kl} \ln f_l/p_{kl}\ .$$

The vectors **g**, **f**, and **c** must now be chosen to provide the extremum indicated just after Theorem 5. Since only **f** and **c** enter in the expression for $\bar{\text{I}}$, we can minimize $d_\infty(\mathcal{S})$ with respect to $\mathbf{g}$ for a constant $\bar{\text{I}}$. This minimization provides precisely the expression[7] for the rate-distortion curve for $\mathcal{S}$ at the information rate $\bar{\text{I}}$. It is further shown in the same reference that the value of $\mathbf{g}$ which provides the minimization is the vector that describes the output statistics on the test channel for $\mathcal{S}$ at the point $(d_{\bar{\text{I}}}, \bar{\text{I}})$ on the rate-distortion curve.

The maximization and minimization of $d_\infty(\mathcal{S})$ with $\mathbf{f}$ and $\mathbf{c}$, respectively, can be accomplished by finding the same extremum of $\bar{\text{I}}$. The resulting values for $\mathbf{f}$ and $\mathbf{c}$ are the output and input probabilities, respectively, on channel $\mathcal{C}$ when it is being used to capacity and the value of $\bar{\text{I}}$ at the extremum point is $-C$. Therefore, the resulting expression for the asymptote of the lower bound is

$$d(\mathcal{S}) \geqq \min_{\mathbf{g}} \mu'(s_o, \mathbf{p}) = d_C \tag{37}$$

with s_o satisfying

$$\mu(s_o, \mathbf{p}) - s_o\mu'(s_o, \mathbf{p}) = -C. \tag{38}$$

This agrees with what we know to te the correct asymptote of the performance curve.[2,7]

8.2 *The Rate of Approach to the Asymptote*

Since the lower bound in equations 35 and 36 is parametric in s and includes the vectors $\mathbf{f}$, $\mathbf{c}$, and $\mathbf{g}$, which when optimally chosen are functions of n, the complete asymptotic dependence of this lower bound upon the block length n is not obvious. To establish this dependence, we first find the *full* derivative of the lower bound in Theorem 5 with respect to n and then integrate the result between n and infinity.

We first simplify the procedure slightly by using our freedom to choose $\mathbf{f}$ by setting this vector equal to its value at $n = \infty$; $\mathbf{f}(\infty)$. This does not change the end result. We also drop the terms of $o(1/n)$ in equations 35 and 36, because they clearly do not affect the asymptotic result. Denoting the right side of equation 35 by d_L and using the chain rule several times, we can write the desired derivative as

$$\frac{dd_L}{dn} = \left(\frac{\partial d_L}{\partial n}\right)_{\mathbf{c},\mathbf{g},s} + \left(\frac{\partial d_L}{\partial s}\right)_{\mathbf{c},\mathbf{g},n} \frac{ds}{dn} + \sum_i \left(\frac{\partial d_L}{\partial g_i}\right)_{\substack{g_k \neq i \\ \mathbf{c},n,s}} \frac{dg_i}{dn}$$
$$+ \sum_k \left(\frac{\partial d_L}{\partial c_k}\right)_{\substack{c_l \neq k \\ \mathbf{g},n,s}} \frac{dc_k}{dn}$$

with

$$\frac{ds}{dn} = \left(\frac{\partial s}{\partial n}\right)_{\mathbf{g},\mathbf{c}} + \sum_i \left(\frac{\partial s}{\partial g_i}\right)_{\substack{g_{k \neq i} \\ \mathbf{c},n}} \frac{dg_i}{dn} + \sum_k \left(\frac{\partial s}{\partial c_k}\right)_{\substack{c_{l \neq k} \\ \mathbf{g},n}} \frac{dc_k}{dn}.$$

The notations outside each parentheses indicate the variables which are momentarily held constant. Substitution yields:

$$\begin{aligned} \frac{dd_L}{dn} = \left(\frac{\partial d_L}{\partial n}\right)_{\mathbf{c},\mathbf{g},s} &+ \left(\frac{\partial d_L}{\partial s}\right)_{\mathbf{c},\mathbf{g},n} \left(\frac{\partial s}{\partial n}\right)_{\mathbf{g},\mathbf{c}} \\ &+ \sum_i \left[\left(\frac{\partial d_L}{\partial s}\right)_{\substack{\mathbf{c},\mathbf{g} \\ n}} \left(\frac{\partial s}{\partial g_i}\right)_{\substack{g_{k \neq i} \\ \mathbf{c},n}} + \left(\frac{\partial d_L}{\partial g_i}\right)_{\substack{g_{k \neq i} \\ \mathbf{c},s,n}}\right] \frac{dg_i}{dn} \\ &+ \sum_k \left[\left(\frac{\partial d_L}{\partial s}\right)_{\substack{\mathbf{c},\mathbf{g}, \\ n}} \left(\frac{\partial s}{\partial c_k}\right)_{\substack{c_{l \neq k} \\ \mathbf{g},n}} + \left(\frac{\partial d_L}{\partial c_k}\right)_{\substack{c_{l \neq k} \\ \mathbf{g},n,s}}\right] \frac{dc_k}{dn}. \end{aligned}$$

The bracketed terms represent the respective partial derivatives of d_L with respect to g_i and c_k with s removed from those quantities held constant. Since $\mathbf{g}(n)$ and $\mathbf{c}(n)$ are chosen for each value of n to minimize the lower bound d_L, these partial derivatives must satisfy

$$\left(\frac{\partial d_L}{\partial g_i}\right)_{\substack{g_{k \neq i} \\ \mathbf{c},n}} + \lambda = 0 \qquad 1 \leqq j \leqq J \tag{39}$$

$$\left(\frac{\partial d_L}{\partial c_k}\right)_{\substack{c_{l \neq k} \\ \mathbf{g},n}} + \nu = 0 \qquad 1 \leqq k \leqq K. \tag{40}$$

This presumes that, at least for sufficiently high n, both $\mathbf{g}$ and $\mathbf{c}$ have only nonzero components. This is known to be true for $\mathbf{c}$,[14] which at $n = \infty$ equals the channel input probabilities that use the channel to capacity.

The vector $\mathbf{g}$, though, can at $n = \infty$ have a zero component. For this case, if the approach of $\mathbf{g}(n)$ to $\mathbf{g}(\infty)$ is from within the composition space, that is, if the components of $\mathbf{g}(n < \infty)$ are nonzero, equation 39 is correct as written for all finite n. If, however, the approach of $\mathbf{g}(n)$ to $\mathbf{g}(\infty)$ is along the boundary of the composition space, that is, having one or more components equal to zero for all $n > N$, then equation 39 can be written, not for all $1 \leqq j \leqq J$, but only for the J' nonzero components. Over the region (N, ∞) the other $J - J'$ zero components obviously can be treated as constants and not included in the differentiation process, thus excluded from the previous summations on j. We shall not attempt to deal with the only remaining possibility, which has $\mathbf{g}(n)$ approaching $\mathbf{g}(\infty)$ such that it oscillates between vector values with all nonzero components and values with some zero components, since no example has been found exhibiting this behavior.

We continue the derivation by substituting equations 39 and 40 into the derivative of d_L to obtain

$$\frac{dd_L}{dn} = \left(\frac{\partial d_L}{\partial n}\right)_{\mathbf{c},\mathbf{g},s} + \left(\frac{\partial d_L}{\partial s}\right)_{\mathbf{c},\mathbf{g},n} \left(\frac{\partial s}{\partial n}\right)_{\mathbf{g},\mathbf{c}} - \lambda \sum_i \frac{dg_i}{dn} - \nu \sum_k \frac{dc_k}{dn}. \tag{41}$$

Finally, since both $\mathbf{g}$ and $\mathbf{c}$ are probability vectors, the last two sums are equal to zero (this is true even when the first sum is only over the J' nonzero components of $\mathbf{g}$). It remains only to find the required partial derivatives from equations 35 and 36. These are given by:

$$\left(\frac{\partial d_L}{\partial n}\right)_{\mathbf{c},\mathbf{g},s} = \frac{1}{2n^2 s}\left(\frac{\gamma'' + \sigma^2}{s^2\mu''} - 1\right),$$

$$\left(\frac{\partial d_L}{\partial s}\right)_{\mathbf{c},\mathbf{g},n} = \mu'' + o(1)$$

$$\left(\frac{\partial s}{\partial n}\right)_{\mathbf{g},\mathbf{c}} = \frac{1}{2n^2 s\mu''} \ln \frac{\gamma''}{s^2\mu''}$$

whence substitution in equation 41 provides

$$\frac{dd_L}{dn} = -\frac{1}{n^2}\frac{1}{2\mid s\mid}\left[\left(\frac{\gamma''}{s^2\mu''} - 1\right) - \ln\frac{\gamma''}{s^2\mu''} + \frac{\sigma^2}{s^2\mu''}\right] + o\left(\frac{1}{n^2}\right). \tag{42}$$

At this point, the vectors $\mathbf{g}$, $\mathbf{c}$ and the parameter s are still functions of n chosen to satisfy the prescribed minimizations of Equation 55 and the parametric Equation 35. If, for large n, these functions are written as

$$\mathbf{g}(n) = \mathbf{g}(\infty) + \Delta\mathbf{g}(n)$$

$$\mathbf{c}(n) = \mathbf{c}(\infty) + \Delta\mathbf{c}(n)$$

$$s(n) = s(\infty) + \Delta s(n),$$

the delta terms can be extracted from the first term in Equation 42. Since each has limit zero for large n, they can, together with the $(1/n)^2$ coefficient, be absorbed into the terms of $o(1/n^2)$. Thus, in equation 42, we can use for $\mathbf{g}$, $\mathbf{c}$, and s their *final* values: $\mathbf{g}(\infty)$, $\mathbf{c}(\infty)$, and $s(\infty)$.

Simple integration of equation 42 between n and infinity, and the use of the known final value of $d_L(n)$, $d_L(\infty) = d_C$, provides the final lower bound to distortion. We again point out that the derivation has included the approximation that $g(\mathbf{z})$ factors as in equation 15.

Theorem 6: A lower bound to the minimum attainable transmission distortion in a system that includes the source $\mathcal{S}$ *and the channel* $\mathcal{C}$ *is given by*

$$d(\mathcal{S}) \geqq d_C + \frac{1}{2n\,|\,s\,|}\left[\left(\frac{\gamma''}{s^2\mu''} - 1\right) - \ln\frac{\gamma''}{s^2\mu''} + \frac{\sigma^2}{s^2\mu''}\right] + o\left(\frac{1}{n}\right) \tag{43}$$

in which

C = capacity of $\mathcal{C}$

d_C = the distortion at $R = C$ on the rate-distortion curve for $\mathcal{S}$

$$\mu(s) = \sum_i q_i \ln \sum_j g_j \exp sd_{ij}$$

$$\gamma(t) = \sum_k c_k \ln \sum_l f_l^{1+t} p_{kl}^{-t}$$

$\mathbf{q} = \mathbf{p}$, the source output probabilities
$\mathbf{g}$ = the output probabilities on the test channel for $\mathcal{S}$ at (d_C, C)
$\mathbf{c}, \mathbf{f}$ = the input and output probabilities on $\mathcal{C}$ when it is used to capacity
$t = -1$
s satisfies $\mu - s\mu' = -C$.

The lower bound in equation 43 is seen to approach its limit algebraically as a/n. Since $(w-1)$ is at least as large as $\ln w$ for any w and σ^2 and μ'' are variances, hence nonnegative, the coefficient a cannot be negative. But it can in special cases equal zero. The conditions for this are

$$\gamma'' = s^2\mu''$$

$$\sigma^2 = 0,$$

conditions that are necessarily met when the source and channel are perfectly matched; that is, when $d(\mathcal{S}) = d_C$ for all n.

They do not, however, constitute a sufficient condition for matching since the low order correction terms in equation 43 could still be nonzero. For the more common situations wherein a is nonzero, the form of the lower bound suggests that the larger the value of a, the longer the coding block length must be to obtain a tolerable level of distortion, $d_C + \Delta$. In turn, the more complex the modulator and demodulator must become. These relations all suggest the utility of the coefficient a as a measure of mismatch between the source $\mathcal{S}$ and the channel $\mathcal{C}$; the larger the value of a, the slower the approach of the lower bound to its asymptote and the greater the mismatch between source and channel. Section X gives several numerical examples illustrating different types of mismatch.

IX. THE SPECIAL CASE OF A NOISELESS CHANNEL

As we have stated, Theorem 5 cannot be applied when $\mathcal{C}$ is noiseless because factors equal to $\gamma''(-1)$ have been canceled within its derivation and, for a noiseless channel, $\gamma''(-1)$ equals zero. We return to the lower bound in equation 3 which is still valid. If the vector $\mathbf{f}$ is chosen uniform over Y^n, we see from the definition of a noiseless channel (L^n outputs) and the definition of information difference in Section IV that $I(\mathbf{x}, \mathbf{y})$ is equal to $\ln(1/L)$ for the output $\mathbf{y}_1$ that has $p(\mathbf{y}_1/\mathbf{x}) = 1$, and is infinite for all other outputs. Since $f(\mathbf{y}_1) = L^{-n}$, $e^{-nI(h)}$ is nonzero only in $0 \leqq h \leqq L^{-n}$, where it is equal to L^n. Therefore, equation 3 can be written as

$$d(\mathbf{w}) \geqq L^n \int_0^{L^{-n}} d(h)\, dh. \tag{44}$$

We remember that the distribution function $G(d)$ is the "inverse" function to $d(h)$ and write

$$d(\mathbf{w}) \geqq L^n \int_0^{d(L^{-n})} [L^{-n} - G(d)]\, dd$$

which can be continued, with any $d_2 \leq d(L^{-n})$, by

$$d(\mathbf{w}) \geqq L^n \int_0^{d_2} [L^{-n} - G(d)]\, dd.$$

Upon dividing the region of integration into two parts, $0 \leq d_1 \leq d_2$, and using the monotonicity of $G(d)$, we have

$$d(\mathbf{w}) \geqq d_2 - L^n d_1 G(d_1) - L^n \int_{d_1}^{d_2} G(d)\, dd. \tag{45}$$

A further lower bound results if we use an upper bound to $G(d)$ in each of the last two terms. In particular, we use the asymptotic bound in equation 20 which we denote here by

$$G(d) \leqq H(n, s) \exp n[\mu(s) - s\mu'(s)] \tag{46}$$

$$\mu'(s) = d.$$

We now set d_2 equal to $\mu'(s_o)$ with s_o given by

$$H(n, s_o) \exp n[\mu(s_o) - s_o\mu'(s_o)] = L^{-n} = e^{-nC}. \tag{47}$$

The fact that $G(d_2) \leq L^{-n}$ guarantees the inequality $d_2 \leq d(L^{-n})$ which we have already used. The second term in equation 45 can be shown to be exponentially small in n whenever $d_1 < d_2$; therefore, we also impose this inequality. To bound the last term in the same equation we use the well known Chernov bound inequality:

$$\exp n[\mu(s) - s\mu'(s)] \leqq \exp n[\mu(s_o) - s_o d]$$

$$\mu'(s) = d$$

together with equations 46 and 47 to obtain

$$L^n \int_{d_1}^{d_2} G(d)\, dd \leqq De^{ns_o\mu'(s_o)} \int_{d_1}^{d_2} e^{-ns_o d}\, dd$$

with

$$D = \max_{d_1 \leqq d \leqq d_2} \frac{H(n, s)}{H(n, s_o)}.$$

The resulting bound for $d(\mathbf{w})$, therefore, is

$$d(\mathbf{w}) \geqq \mu'(s_o) + \frac{D}{ns_o}[1 - \exp ns_o(\mu'(s_o) - d_1)] + o\left(\frac{1}{n}\right).$$

If d_1 is chosen in a way to approach $\mu'(s_o)$ with increasing n, this bound becomes:

$$d(\mathbf{w}) \geqq \mu'(s_o) + \frac{1}{ns_o}[1 + o(1)] \tag{48}$$

in which s_o satisfies equation 47, rewritten here as

$$\begin{aligned} \mu(s_o) - s_o\mu'(s_o) &= -C - \frac{1}{n}\ln H(n, s_o) \\ &= -C + \frac{1}{2n}\ln n[1 + o(1)]. \end{aligned} \tag{49}$$

The remaining steps, averaging over the source space and minimizing the resulting bound over all choices of $\mathbf{g}$ (we continue to use the approximation in Equation 15), are identical in procedure to those previously used. We state only the result.

Theorem 7: The minimum attainable transmission distortion of the source $\mathcal{S}$, *when used with a noiseless channel of capacity* C, *satisfies*

$$d(\mathcal{S}) \geqq d_C + \frac{1}{2}\frac{\ln n}{|s_o|\, n}[1 + o(1)] \tag{50}$$

in which s_o *satisfies*

$$\mu(s_o\,,\,\mathbf{p}) - s_o\mu'(s_o\,,\,\mathbf{p}) = -C. \tag{51}$$

We see by comparing equat.ons 43 and 50 that while the lower bound to distortion with a noisy channel approaches its asymptote, d_C, as $1/n$, the lower bound to distortion with a noiseless channel approaches d_C only as $(\ln n)/n$. These bounds are not inconsistent since for a noiseless channel the variance γ'' is zero with the result that the coefficient of $1/n$ in equation 43 is infinite. A similar limiting statement is also true. If a noisy channel is made to approach a noiseless one by reducing the noisy transition probabilities toward zero, at the same time keeping the channel capacity constant by appropriately reducing either the channel input alphabet size or the channel dimensionality, the coefficient of the $1/n$ term increases and is unbounded. These results therefore suggest than when there is a choice between using a noiseless channel or a noisy one *of equal capacity*, the noisy channel is always the better choice. And, inasmuch as we are using the coefficient of the $1/n$ term to measure the source-channel mismatch, the noiseless channel represents the worst possible match to any source.

X. EXAMPLES

In the first three examples, we illustrate different types of source-channel mismatch and calculate the effect of each upon the coefficient a in the lower bound of equation 43. Each of these examples tends to strengthen the suggestion in the lower bound result that this coefficient is a measure of source-channel mismatch since it increases monotonically as the channel is perturbed away from the matching channel.

Because the channel statistics influence only the first two terms of a, we use in these examples a doubly uniform source for which the σ^2 term equals zero. To further isolate the relative matching properties of the source-channel pairs, we keep constant the channel capacity per source output, C, as the channel is varied. Thus the distortion per source component has the same asymptote, d_C, for all source-channel pairs and the only difference in the lower bound curves, at least asymptotically, is in the coefficient a.

Example 1

This example illustrates a dimensionality, or coding block length, mismatch between a source and channel. We take for the source $\mathbb{S}$

the m_s'th product of a binary symmetric source, defined by $\mathbf{p} = (\frac{1}{2}, \frac{1}{2})$ and $d_{11} = d_{22} = 0$, $d_{12} = d_{21} = 1$. For the channel $\mathcal{C}$ we take the m_c'th product of a binary symmetric channel, each component $\mathcal{C}_i$ having a crossover probability p. The channel capacity per source component is m_c/m_s times the capacity of $\mathcal{C}_i$ and is kept constant as m_c/m_s is varied by appropriately changing the crossover probabilities p.

Figure 6 shows the dependence of a upon m_c/m_s. When comparing the two curves in this figure, notice that the ordinate has been normalized by d_C. We know that for $m_c/m_s = 1$ the source and channel are precisely matched and this is indicated in the figure by the value $a = 0$ at that point. Above this point a increases monotonically in m_c/m_s and can be shown to have the asymptotic form $a \sim k(m_c/m_s)^{\frac{1}{2}}$. Below $m_c/m_s = 1$, a also becomes unbounded as m_c/m_s approaches the ratio that requires each component channel $\mathcal{C}_i$ be noiseless. This is not inconsistent with the noiseless channel result (equation 50) which indicated that the rate of approach of the distortion to d_C was not as a/n but as $(\ln n)/n$.

Example 2

Here we do not change the relative dimensionality, only the form of the channel. The source is a binary symmetric source and the channel a binary nonsymmetric channel of varying asymmetry. The crossover probabilities are again changed in a way that does not vary the capacity. We see in Fig. 7 that a is rather insensitive to small perturbations from a binary symmetric channel and in most cases is affected less by this type of mismatch than a dimensionality mis-

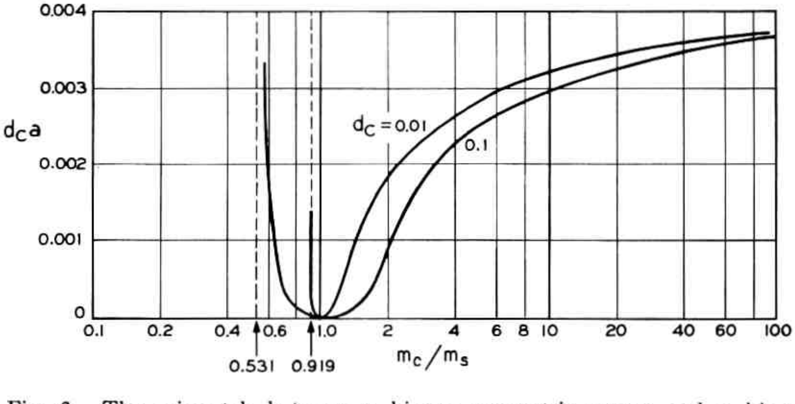


Fig. 6 — The mismatch between a binary symmetric source and a binary symmetric channel of different dimensionality.

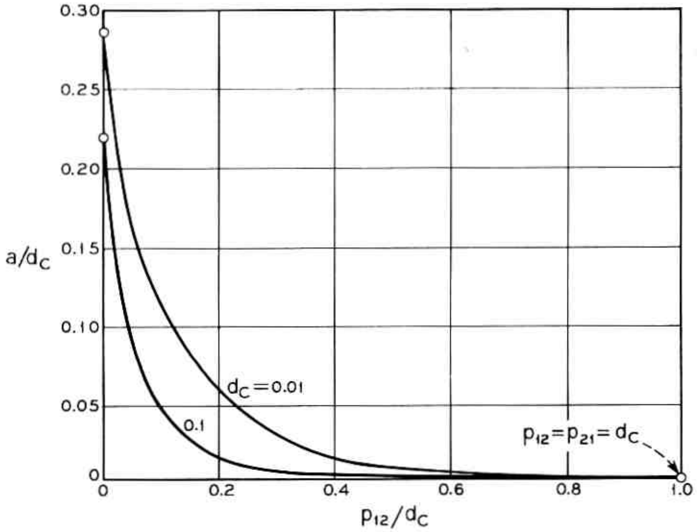


Fig. 7 — The mismatch between a binary symmetric source and a binary nonsymmetric channel.

match. A similar result obtains if the source is also allowed to be nonsymmetric.

Example 3

For this example we use a binary symmetric source and a discrete channel which models the m orthogonal signal modulator used in the next example. The channel has m inputs and m outputs and has from each input one transition of probability $1 - (m-1)p$ and $m - 1$ transitions of probability p. The numbers m and p are varied together in such a way that the capacity of the channel remains constant. We see in Fig. 8 that the mismatch coefficient a is much higher when the binary symmetric source is used with this channel than when it is used with that product binary symmetric channel of Example 1 which has available an input alphabet of equal size. The comparison can be made on Figures 6 and 8 at points for which $m_c/m_s = \log_2 m$.

Example 4

In this, the last example, we include in the system a continuous channel which is to be used by a discrete source with a discrete modulator. Now, as the modulator changes the discrete channel extracted from the actual channel changes and *both* its capacity and its matching characteristics change. It turns out that both properties are not necessarily optimized for the same modulator structure and, therefore, one must strike a compromise (influenced by the block length of interest) between a modulator design that minimizes the asymptote d_C and maximizes the rate of approach to d_C.

To illustrate this we assume the channel to be a band-limited channel with additive white gaussian noise in the allowed bandwidth. During the interval (O,T), the discrete modulator is constrained to transmit one of m orthogonal signals in each of B bauds and altogether an energy no greater than E. To model the bandwidth constraint the mB product is assumed constant, but m and B can otherwise be varied to optimize the system. Thus the equivalent discrete channel is the B'th product of the m input doubly uniform channel of Example 3. The source to be transmitted is a binary symmetric source with an output rate of M_s digits every T seconds.

In Fig. 9 we show the minimum attainable distortion d_C (determined through the channel capacity) and the mismatch coefficient a as a function of m. For the values shown in figure, we see that while d_C is minimized at $m = 15$, the coefficient a is then quite large. And, around $m = 22$, where $a = 0$, the minimum distortion d_C is higher than that which can be realized with a smaller m. The conclusion from this is that the modulator should be designed with $m =$ 15 (to maximize capacity and minimize d_C) only when one is willing to use very long coding block lengths. For shorter block lengths, a larger value of m, and a corresponding smaller value of a, could result in a smaller average distortion even with the larger value of d_C. For

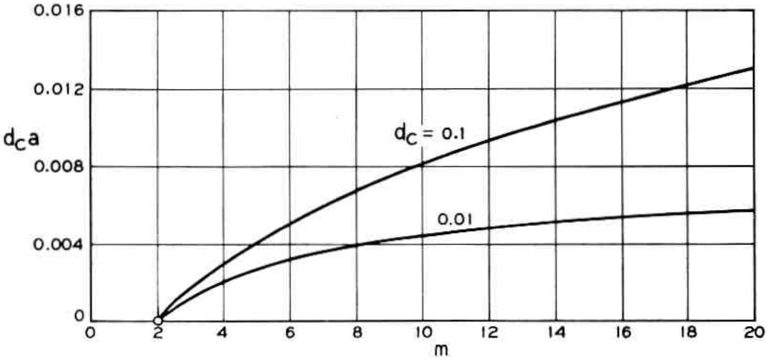


Fig. 8 — The mismatch between a binary symmetric source and the m-orthogonal signal channel.

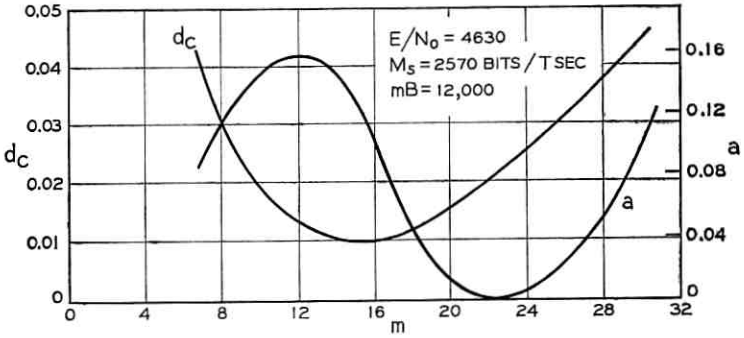


Fig. 9 — The influence of the modulator design in Example 4 on the minimum attainable distortion and the mismatch coefficient.

this example a compromise design with m about 19 would probably be best over a range of intermediate block lengths.

It is interesting to notice in this example that the coefficient a can be zero even when the source and channel are not matched. This is consistent with our previous interpretation of $a = 0$ as a necessary but not sufficient condition for matching. We remember that the coefficient a being zero does not imply that the lower bound in equation 43 is precisely d_C for all n. There are several other terms of $o(1/n)$ in this equation that have not been specified which are not necessarily zero when $a = 0$.

XI. THE UPPER BOUND

Now let us present an upper bound to the minimum attainable transmission distortion as a function of the coding block length. As with the lower bound, the upper bound approaches the asymptote d_C, but only as $[(\ln n)/n]^{1/2}$. The reason for the difference, we believe, is that within the upper bound derivation the transmitting signal set was restricted to contain at most $M = e^{nC}$ members, a restriction that was not necessary to impose in the lower bound. We also present an upper bound to the transmission distortion with a noiseless channel. This bound does agree, asymptotically, with the corresponding lower bound.

XII. THE RANDOM CODING ARGUMENT

All of the upper bound derivations in this paper use random coding arguments. That is, we do not explicitly find the encoder and decoder which, when used with $\mathcal{S}$ and $\mathcal{C}$, provide the distortion in the upper bound, but show that one pair does exist. More specifically, we construct a set of encoder-decoder pairs with a probabilistic rule according to which each system is selected to be used. This defines an ensemble of transmission systems, each with its own distortion, corresponding to all possible coding selections. What we calculate is a bound to the average distortion of this ensemble. Clearly, this provides an upper bound to the minimum distortion in the ensemble, hence to the minimum attainable distortion in any system that includes $\mathcal{S}$ and $\mathcal{C}$.

12.1 *The Construction of the Ensemble*

We denote the set of points on the rate distortion curve for $\mathcal{S}$ by (d_R, R) and assume the capacity of $\mathcal{C}$ to be C. We first choose any point (d^*, R^*) on the rate-distortion curve below (d_C, C) and design the code in such a way that the ensemble average distortion approaches d^* with increasing block length. We know this to be possible from Shannon's results.[2] Moreover, we expect, since the situation is somewhat analogous to a channel coding problem with $R^* < C$, that the distortion can be made to approach d^* exponentially fast. The point (d^*, R^*) is subsequently varied to obtain the best result at any particular block length of interest.

For any selection of (d^*, R^*), we then choose the number of signal points, $M = e^{nR}$, used to transmit $\mathcal{S}$. To attain a transmission distortion level d^*, we certainly must have the number of signal points large enough to represent the source to at least within d^*, and this requires that R be greater than R^*. We also require that R be less than C so that in the limit as n becomes large, we are guaranteed correct decoding among the signal points at the receiver. Therefore we have

$$R^* < R < C \tag{52}$$

and, for the corresponding values of distortion on the rate-distortion curve,

$$d_{\max} \geqq d^* > d_R > d_C \,. \tag{53}$$

The value of R can also later be chosen to optimize the result.

An ensemble of codes of length n is constructed for each selection of R and R^*. We use the probability distribution $p(\mathbf{x}, \mathbf{z})$ to generate the ensemble by picking, according to $p(\mathbf{x}, \mathbf{z})$, M independent pairs $(\mathbf{x}, \mathbf{z})$ from X^nZ^n. Thus we have a set of codes containing all possible mappings of the integers 1 through M into pairs of n-letter words $(\mathbf{x}, \mathbf{z})$, or $(JK)^{nM}$ codes in total. (We continue to use here the notation defined in the earlier part of the paper dealing with the lower bound.) Each of these codes has the associated probability

$$\Pr\,(\text{code}) = \prod_{i=1}^{M} p(\mathbf{x}_i\,,\,\mathbf{z}_i).$$

Any probability function $p(\mathbf{x}, \mathbf{z})$ could be used to obtain an upper bound, but we use a distribution that factors into $p(\mathbf{x})g(\mathbf{z})$; therefore, in the ensemble, each set of M decoded words, θ_1, is independent of each set of M channel input words, θ_2. Thus we can write

$$\Pr\,(\text{code}) = p(\theta_1\,,\,\theta_2) = p(\theta_1)p(\theta_2) = \prod_{i=1}^{M} p(\mathbf{x}_i) \prod_{i=1}^{M} g(\mathbf{z}_i).$$

Further, we use for $p(\mathbf{x})$ and $g(\mathbf{z})$ the product forms

$$\prod_{m=1}^{n} p(x^m) \quad \text{and} \quad \prod_{m=1}^{n} g(z^m)$$

in which the letter probability distribution $p(x)$ is that which yields a mutual information C on $\mathcal{C}$ and the letter probability distribution $g(z)$ is that which gives the output statistics on the test channel for $\mathcal{S}$ at the point (d^*, R^*) on the rate-distortion curve.

The encoding and decoding is done as follows: In every ensemble member there is a list θ_1 of allowed decoded words and a list θ_2 of usable channel input words. When a source output $\mathbf{w}$ occurs, the encoder scans θ_1 and chooses any member $\mathbf{z}_o$ in this list for which

$$d(\mathbf{w}, \mathbf{z}_o) \leqq d^*. \tag{54}$$

If there are none, the encoder chooses any member at all on the list θ_1, say $\mathbf{z}_1$. Since the lists are chosen together, there corresponds to $\mathbf{z}_o$ or $\mathbf{z}_1$ a particular $\mathbf{x}$ in θ_2, and this word is used to transmit $\mathbf{w}$. The decoder uses a maximum likelihood decision rule to decode $\mathbf{y}$ into a member of θ_2, which is then associated, through the pairings among the two lists, with a member $\mathbf{z}$ in θ_1. The resulting distortion, by definition, is $d(\mathbf{w}, \mathbf{z})$.

12.2 *The Ensemble Average Distortion*

Each member, θ, of the ensemble is a complete transmission system in itself, and has an average transmission distortion dependent upon the codes, θ_1 and θ_2, that are used. This average distortion, which is an average over all possible source and channel events, is equal to

$$d(\theta) = d(\theta_1\,,\,\theta_2) = \sum_{W^n} p(\mathbf{w}) \sum_{Y^n} p(\mathbf{y} \mid \mathbf{x})\, d(\mathbf{w}, \mathbf{z}).$$

The ensemble average distortion is obtained by averaging $d(\theta_1, \theta_2)$ over all choices of θ_1 and θ_2, hence

$$\langle d(\theta)\rangle_{\text{av}} = \sum_{W^n} p(\mathbf{w}) \sum_{Y^n} [\sum_{\theta_1} \sum_{\theta_2} p(\mathbf{y} \mid \mathbf{x})\, d(\mathbf{w}, \mathbf{z}) p(\theta_1) p(\theta_2)]. \qquad (55)$$

We next separate the events $\mathbf{w}$, θ_1, θ_2, and $\mathbf{y}$ into two sets: (*i*) those quadruples for which *either* there does not exist a $\mathbf{z}$ in θ_1 satisfying equation 54 *or* the received word $\mathbf{y}$ is decoded into a member of θ_2 different from the transmitted word $\mathbf{x}(\mathbf{w})$, and (*ii*) its complement. For quadruples in set one, the distortion $d(\mathbf{w}, \mathbf{z})$ is surely upper bounded by $d_{\max}$, the maximum entry in $\| d(w, z) \|$. For those in the second set, we use equation 54 and the fact that the decoder returns us through $\mathbf{x}(\mathbf{w})$ to $\mathbf{z}_o$ to upper bound the distortion by d^*. Therefore, if the characteristic function Φ is used to indicate the quadruples in set one, we can upper bound the ensemble average with

$$\langle d(\theta)\rangle_{\text{av}} \leqq \sum_{W^n} p(\mathbf{w}) \sum_{Y^n} \sum_{\theta_1} \sum_{\theta_2} p(\mathbf{y} \mid \mathbf{x}) p(\theta_1) p(\theta_2) [d^*(1 - \Phi) + d_{\max}\Phi]$$

$$= d^* + (d_{\max} - d^*) \Pr(\Phi). \qquad (56)$$

Finally, we use the union bound to upper bound $Pr(\Phi)$ and the ensemble average distortion, $\langle d(\theta)\rangle_{\text{av}}$, to upper bound the minimum attainable transmission distortion, $d(\mathcal{S})$, and obtain the result in the next theorem.

Theorem 8: The minimum attainable transmission distortion of the source $\mathcal{S}$*, when used with the channel* $\mathcal{C}$*, satisfies*

$$d(\mathcal{S}) \leqq d^* + (d_{\max} - d^*)[\Pr(\exists' \mathbf{z}_o \text{ in } \theta_1) + \Pr(\text{channel error})] \qquad (57)$$

in which $\exists'$ means "there does not exist," d^* is any distortion greater than d_C, and R (a variable in the bracketed terms) is any rate in the interval $R^* < R < C$. The bound is a function of n through the quantity in the brackets.

The last term in the brackets, the probability of error on the channel, has been approximated by many people, but we will use Gallager's bound[15]

$$\Pr(e) \leqq e^{-nE(R)} \qquad (58)$$

in which $E(R)$ is a positive monotonically increasing function of the difference $C - R$. The next section is devoted to the evaluation of the first term in the brackets, which is the probability that the source word $\mathbf{w}$ and the list θ_1 are such that equation 54 is not satisfied for any $\mathbf{z}$ in θ_1.

XIII. THE PROBABILITY OF FAILURE AT THE ENCODER

We say that failure occurs at the encoder, for the source output **w**, when each of the M allowed decoded words on list θ_1 are at a distortion $d(\mathbf{w}, \mathbf{z})$ from **w** greater than d^*. Because each of the M words in θ_1 is selected independently, we can write the total probability of this failure as

$$\begin{aligned}\Pr(\exists' \mathbf{z}_o \text{ in } \theta_1) &= \sum_{W^n} p(\mathbf{w}) \Pr(\exists' \mathbf{z}_o \text{ in } \theta_1 \mid \mathbf{w}) \\ &= \sum_{W^n} p(\mathbf{w})[1 - \Pr(\mathbf{z} \ni d(\mathbf{w}, \mathbf{z}) \leqq d^* \mid \mathbf{w})]^M.\end{aligned} \tag{59}$$

The last probability is seen equal to the distribution function of the distortion random variable described in Section 6.2 and defined by equations 16 and 17. In these equations $\mathbf{q} = q_1, q_2, \cdots, q_H$ is the composition vector of the source word **w**, and D_{ir} is the letter distortion random variable between the r'th appearance of the letter w_i in **w** and the corresponding letter in **z**.

We again notice that the distribution function of $d(\mathbf{w}, \mathbf{z})$ depends only upon the composition **q** of **w**. Thus we are able to perform the average over W^n in equation 59 as one over all possible compositions of **w**. All possible compositions can be represented as points in the $H - 1$ dimensional hyperplane within the first quadrant of R^H which intersects each axis q_i at one. This hyperplane is called the composition space Q^H. The probability of any composition point is equal to the product of the number of different source words having this composition and the probability of each, therefore, we have

$$\begin{aligned}P(\mathbf{q}) &= N(\mathbf{q}) \prod_{i=1}^{H} p_i^{nq_i} \\ &= \frac{n\,!}{\prod_{i=1}^{H} (nq_i)\,!} \prod_{i=1}^{H} p_i^{nq_i}.\end{aligned}$$

Interpreting $P(\mathbf{q})$ as an impulse function over Q^H we can now write equation 59 as

$$\Pr(\exists' \mathbf{z}_o \text{ in } \theta_1) = \int_{Q^H}\cdots\int P(\mathbf{q})[1 - G(d^* \mid \mathbf{q})]^M \, d\mathbf{q}. \tag{60}$$

To continue the inequality in equation 57, we require a lower bound to $G(d^*)$. For our present purpose, Fano's lower bound[12] is sufficient:

$$\begin{aligned} G(d^* \mid \mathbf{q}) &\geqq K(n, \mathbf{q}) \exp n[\mu(s, \mathbf{q}) - s\mu'(s, \mathbf{q})] \\ &\equiv K(n, \mathbf{q}) \exp - nR(d^*, \mathbf{q}) \end{aligned} \tag{61}$$

in which

$$\mu'(s, \mathbf{q}) = d^* \tag{62}$$

$$0 < d^* \leqq E(d \mid \mathbf{q}) \tag{63}$$

$$\mu(s) = \sum_{i=1}^{H} q_i \ln \sum_{j=1}^{J} g_j \exp sd_{ij}$$

and $K(n, \mathbf{q})$ is a rather complex function of $\mathbf{q}$ and n that goes to zero algebraically in n with increasing n. Its precise form is otherwise unimportant in the following derivation. (The bound in equation 61 can still be used for points $\mathbf{q}$ that violate equation 63 if one uses the value of $s = 0$ rather than that which satisfies equation 62.) We can therefore write

$$\Pr(\exists\, '\mathbf{z}_o \text{ in } \theta_1) \leqq \int \underset{Q^H}{\cdots} \int P(\mathbf{q})[1 - K(n, \mathbf{q}) \exp - nR(d^*, \mathbf{q})]^{\exp nR}\, d\mathbf{q}. \tag{64}$$

The next step is to divide the composition space Q^H into two disjoint subspaces, Q and Q', that are defined by

$$Q = \{\mathbf{q}: R(d^*, \mathbf{q}) < R - \delta\} \tag{65}$$

$$Q' = \{\mathbf{q}: R(d^*, \mathbf{q}) \geqq R - \delta\} \tag{66}$$

with δ any positive number satisfying $R^* < R - \delta$. The idea behind this separation is illustrated in Fig. 10. The bracketed term in the integrand of equation 64 has the form $[1 - \exp(-nA)]^{\exp nB}$ which approaches zero with increasing n when $A < B$, and one when $A > B$. In the first region, which, except for the δ, corresponds to the set Q, we shall use the upper bound

$$[1 - \exp(-nA)]^{\exp nB} \leqq \exp[-\exp n(B - A)] \tag{67}$$

and in the second region, corresponding to Q', the (poorer) bound

$$[1 - \exp(-nA)]^{\exp nB} \leqq 1. \tag{68}$$

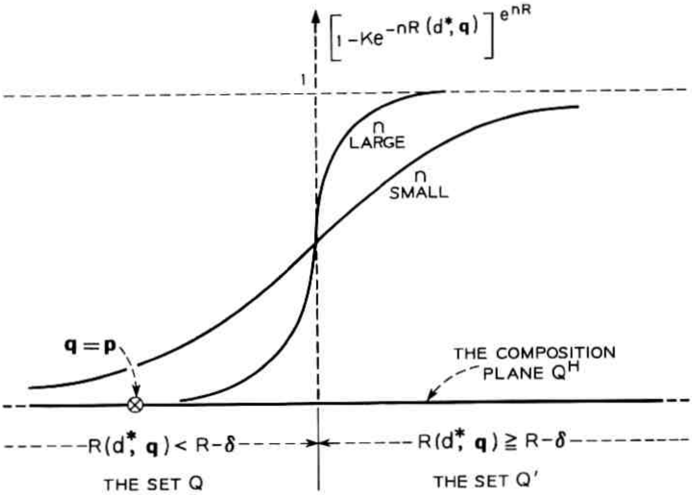


Fig. 10 — The division of the composition plane Q^H into the sets Q and Q'.

The use of these bounds in equation 64 results in

$$
\begin{aligned}
&\Pr(\exists\, '\mathbf{z}_o \text{ in } \theta_1) \\
&\quad \leqq \int_{Q} \cdots \int P(\mathbf{q}) \exp\{-K(n, \mathbf{q}) \exp n[R - R(d^*, \mathbf{q})]\}\, d\mathbf{q} \\
&\quad + \int_{Q'} \cdots \int P(\mathbf{q})(1)\, d\mathbf{q} \\
&\quad \leqq \int_{Q} \cdots \int P(\mathbf{q}) \exp[-K(n, \mathbf{q})e^{n\delta}]\, d\mathbf{q} + \Pr(Q') \\
&\quad \leqq \exp[-K(n)e^{n\delta}] + \Pr(Q')
\end{aligned}
\tag{69}
$$

in which $K(n)$ denotes the minimum of $K(n, \mathbf{q})$ over Q. The first term in this upper bound is a double exponential in n which will turn out to be unimportant. Thus it remains to evaluate $Pr(Q')$.

We shall use what we call the hypercube method to upperbound $Pr(Q')$. Although the resulting bound is not as tight as others that could be derived (see, for example, the maximum probability point method in Ref. 8), it has the advantage of being simpler both to derive and to use and, in addition, does not seriously degrade the final bound to transmission distortion. What is done is to enclose the set Q' by another set Q_1' that has a relatively simple configuration, and to upper bound $Pr(Q')$ by $Pr(Q_1')$.

We construct in R^H a hypercube of dimension $2u$ centered at $\mathbf{q} = \mathbf{p}$,

$$K^H = \{\mathbf{q}: p_i - u \leqq q_i \leqq p_i + u\},$$

and intersect with it the composition space Q^H. The intersection forms a "solid" Q_1

$$Q_1 = Q^H \cap K^H$$

which contains vertices of the form $\mathbf{q}_v = q_{1v}, q_{2v}, \cdots q_{Hv}$, with the components, of course, summing to one. When H is even, q_{iv} equals either $p_i + u$ or $p_i - u$, and when H is odd, q_{iv} has the same values with the addition of one component equal to p_i. The vertices of Q_1 are joined by straight lines.

At this point we use the fact that Q is a convex set,[8] that is, for $0 \leqq \lambda \leqq 1$, $\lambda\mathbf{q}_a + (1 - \lambda)\mathbf{q}_b$ is a member of Q whenever both $\mathbf{q}_a$ and $\mathbf{q}_b$ are. This property ensures us that whenever the vertices of Q_1 are in the set Q, the entire set Q_1 is in Q,

$$Q_1 \subseteq Q,$$

with the consequence that

$$\Pr(Q') \leqq \Pr(Q_1'). \tag{70}$$

The remaining step is to bound the total probability of the set Q_1'. Because this probability equals the probability that *any* of the dependent events $q_i \,\varepsilon'\, [p_i - u, p_i + u]$ occurs, we can use the union bound to upper bound $Pr(Q_1')$ by the sum of the individual probabilities. Thus

$$\Pr(Q_1') \leqq \sum_{i=1}^{H} \Pr[q_i < p_i - u] + \Pr[q_i > p_i + u].$$

These quantities can be further upper bounded by a simple application of Chernov bounds. This has been done for us in Ref. 16, page 102, where the result found is, in our notation,

$$\Pr(Q_1') \leqq \sum_{i=1}^{H} e^{-nX_i} + e^{-nY_i} \tag{71}$$

in which

$$\left.\begin{matrix} X_i \\ Y_i \end{matrix}\right\} = -\ln\left[\left(\frac{p_i}{d_i}\right)^{d_i}\left(\frac{1-p_i}{1-d_i}\right)^{1-d_i}\right]$$

and

$$\begin{aligned} d_i &= p_i - u \quad \text{for} \quad X_i \\ &= p_i + u \quad \text{for} \quad Y_i \, . \end{aligned}$$

In these bounds, the hypercube dimension $2u$ should be maximized, to obtain the tightest bound, subject only to the constraint that all vertices $\mathbf{q}_v$ be in region Q, that is, that they satisfy equation 65.

The bound in equation 71 can be simplified still further by writing

$$\begin{aligned} \Pr(Q_1') &\leqq 2H \exp\left[-n \min\left(X_i \, , \, Y_i\right)\right] \\ &\equiv K_1 \exp - nE_s(R). \end{aligned} \tag{72}$$

Indeed, it can be shown,[8] that there are two, and not $2H$, candidates for the minimizing quantity in the exponent.

XIV. THE SET OF UPPER BOUNDS

Combining equations 57, 58, 69, and 72, we have the following result:

Theorem 9: The minimum attainable transmission distortion of the source $\mathcal{S}$, *when used with the channel* C, *satisfies*

$$\begin{aligned} d(\mathcal{S}) \leqq d^* + (d_{\max} - d^*)\{&\exp\left[-K(n)e^{n\delta}\right] \\ &+ K_1 \exp\left[-nE_s(R)\right] + \exp\left[-nE(R)\right]\} \end{aligned} \tag{73}$$

for any d^* *and* R *that satisfy*

$$d_{\max} \geqq d^* > d_R > d_C \tag{74}$$

$$R^* < R < C. \tag{75}$$

The freedom provided by equations 74 and 75 can be used to generate a set of upper bounds, corresponding to all possible choices of d^* and R, the properties of which depend upon those of the two exponential functions in equation 73. It has been shown elsewhere[8] that $E_s(R)$ is a positive monotone increasing function of the difference $R - R^*$, that $E_s(R^*) = E_s'(R^*) = 0$, and that $E_s''(R^*) \neq 0$. Comparing these with the corresponding properties of the channel reliability function:[15] $E(R)$ a positive monotone increasing function of the difference $C - R$, $E(C) = E'(C) = 0$, $E''(C) \neq 0$; we see that the two functions are quite similar. Typically, their curves would look like those in Fig. 11.

With these curves, we can examine the behavior of the set of bounds in Theorem 9. As shown in Fig. 12, when d^* is chosen much larger than d_C, the nonzero slope of the rate-distortion curve allows

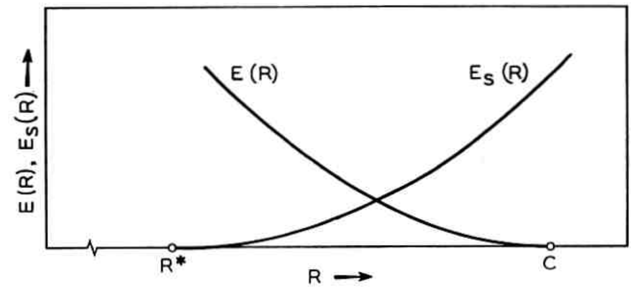


Fig. 11 — Typical behavior of $E_s(R)$ and $E(R)$ near their zero value.

a choice of R that can make both the differences $C - R$ and $R - R^*$ large. In turn, the exponents $E_s(R)$ and $E(R)$ in equation 73 are large and the exponential terms decay very rapidly with n. But for this choice, the asymptote d^* is much greater than the level d_C, which we know can be approached.

On the other hand, if we choose d^* only slightly greater than d_C, we have an upper bound with an asymptote that is nearly d_C, but now the differences $C - R$ and $R - R^*$, and therefore the exponents $E_s(R)$ and $E(R)$, are much smaller and the rate of approach to the asymptote d^* is correspondingly slower. Thus, in the selection of d^* and R there is a trade-off between a small asymptotic value and a fast rate of approach. This is illustrated in Fig. 13 in which we show a set of curves obtained from the upper bound expressions in equation 73. The best compromise for any value of n is given by the

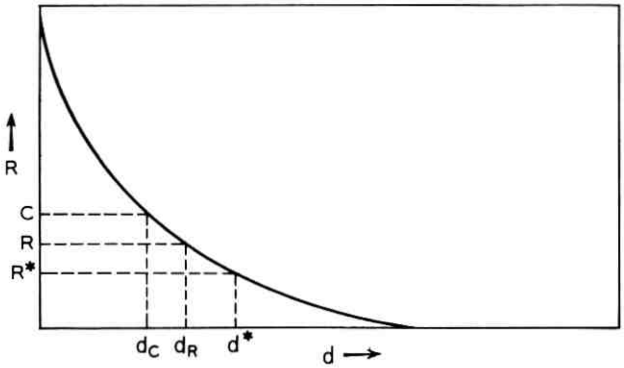


Fig. 12 — The rate-distortion curve for $\mathcal{S}$ illustrating the relations among the parameters in Theorem 9.

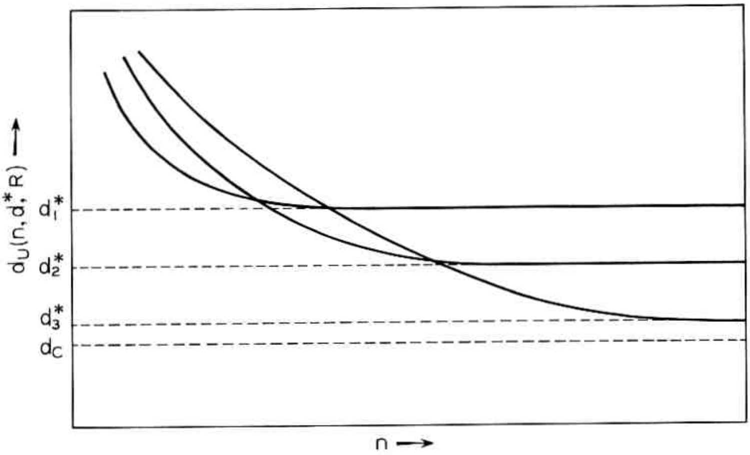


Fig. 13 — The upper bound in Theorem 9 with three different values for d^* and R.

lower envelope to the entire set of bounds in equation 73, therefore we have

Theorem 10: The minimum attainable transmission distortion of the source $\mathcal{S}$*, when used with the channel* $\mathcal{C}$*, satisfies*

$$d(\mathcal{S}) \leqq \min_{d^*,R} d_U(n, d^*, R) \equiv d_U(\mathcal{S}) \tag{76}$$

in which the function $d_U(n, d^*, R)$ is used to denote the right side of equation 73.

In the next section we study the asymptotic behavior of the lower envelope. At this point, though, we wish to include an important conclusion that can be established from the set of upper bounds in equation 73. Each individual bound indicates that, in a system where the distortion level d_C is attainable in the limit, if one would tolerate a distortion $d^* = d_C + \Delta$, this level could be approached exponentially fast as the coding block length is increased.

Actually, a much stronger statement is possible. Since the distortion curve for $d^* = d_C + \frac{1}{2}\Delta$ approaches this level in the limit, it must cross, at some finite n, the level $d_C + \Delta$. Because both curves are for the same source and channel, this proves that the distortion level $d_C + \Delta$ is not only approachable exponentially fast, it is in fact attainable with a finite coding block length. This is true for any $\Delta > 0$, no matter how small.

XV. THE ASYMPTOTIC BEHAVIOR OF THE UPPER BOUND

From the previous discussion it is clear that as n increases, the optimum value of d^* must approach d_C and therefore that the exponents $E_s(R)$ and $E(R)$ must approach zero. For this reason we use the Taylor series representations for these functions at R^* and C in equations 73 and 76, respectively, and obtain

$$d_U(\mathbb{S}) \approx \min_{d^*, R} \{d^* + (d_{\max} - d^*)$$

$$\cdot [K_1 \exp - nb_1(R - R^*)^2 + \exp - nb_2(C - R)^2]\} \qquad (77)$$

with $b_1 = \frac{1}{2}E_s''(R^*)$ and $b_2 = \frac{1}{2}E''(C)$. In using the Taylor series for $E(R)$ and $E_s(R)$ we have dropped the cubic terms since both $E'''(C)$ and $E_s'''(R^*)$ are finite and $C - R$ and $R - R^*$ are $o(1)$. The double exponential term involving δ is also dropped since it can be shown to contribute nothing important in the asymptotic bound.

We next avoid the minimization on R by choosing that value of R which equates the two exponents:

$$b_1(R - R^*)^2 = b_2(C - R)^2. \qquad (78)$$

While this selection of R is nonoptimum for finite n, it can be shown that it asymptotically approaches R_{opt}, and that it does not affect the asymptotic behavior of the upper bound. This particular choice of R allows us to combine the two exponential terms in equation 77. If we start with equation 78 and the obvious equality

$$(C - R) + (R - R^*) = C - R^*,$$

we can establish

$$(C - R) = \frac{\sqrt{b_1}}{\sqrt{b_1} + \sqrt{b_2}}(C - R^*) \qquad (79)$$

$$(R - R^*) = \frac{\sqrt{b_2}}{\sqrt{b_1} + \sqrt{b_2}}(C - R^*), \qquad (80)$$

which further allows us to write the two exponents in terms of the common difference $C - R^*$.

Next, we wish to express the difference $C - R^*$ in terms of the difference $d_C - d^*$. Taylor's formula with remainder is again used:

$$R(d^*) = R(d_C) + R'(d_C)(d^* - d_C) + o(d^* - d_C)$$

or

$$C - R^* = -R'(d_C)(d^* - d_C) - o(d^* - d_C) \tag{81}$$
$$= -s_o(d^* - d_C) - o(d^* - d_C).$$

In the last equation we have used the fact that the slope of the rate distortion curve at the point (d_C, C) is equal to the value of s which satisfies $\mu(s) - s\mu'(s) = -C$.[7, 8]

Finally, we substitute equations 79, 80, and 81 into equation 77, subtract d_C from both sides of this last equation, and change the minimizing variable to $d^* - d_C$ to obtain

$$d(\mathcal{S}) - d_C \leqq \min_x \, [x + (A - x)K_2 \exp - Bnx^2] \tag{82}$$

in which $x = d^* - d_C$, $A = d_{\max} - d_C$, $K_2 = K_1 + 1 = 2H + 1$, and

$$B = b_1 b_2 s_o^2/(\sqrt{b_1} + \sqrt{b_2})^2.$$

We next find the asymptotic behavior of the lower envelope in equation 82.

If x is considered the parameter, each function of n in the set $f(x, n)$ starts at $f(x, 0) = x + (A - x)K_2$ and decreases exponentially to $f(x, \infty) = x$. For any two parameter values, x_1 and x_2, with $x_1 > x_2$ we have

$$f(x_1, 0) - f(x_2, 0) = (1 - K_2)(x_1 - x_2)$$
$$= -2H[f(x_1, \infty) - f(x_2, \infty)].$$

Consequently, any two curves must cross as in Fig. 14.

It follows that the parameter $x_o(n)$, which identifies the minimum of $f(x, n_o)$ at the value $n = n_o$, must change with n. Since this parameter is the solution of

$$f'_x(x, n) = 0,$$

we have

$$\exp(nBx_0^2) - K_2 = 2nK_2Bx_0(A - x_0). \tag{83}$$

Figure 15 shows the required graphical solution which clearly always exists. The substitution of $x_o(n)$ in $f(x, n)$ specifies the single function of n, $f[x_o(n), n]$, which is the desired lower envelope. Unfortunately, an explicit solution is not possible for $x_o(n)$, nor for $f[x_o(n), n]$, but we can obtain bounds to both that are adequate for our purposes.

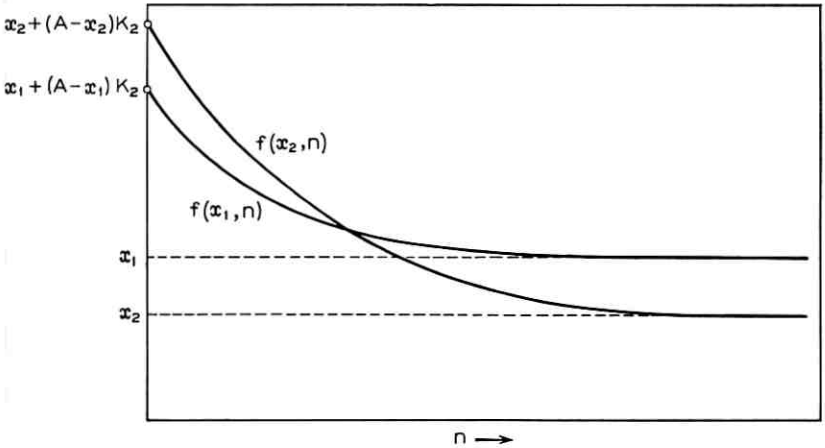


Fig. 14 — Two members of the family of curves: $f(x,n) = x + (A - x)K_2 \exp(-Bnx^2)$.

From the graphical solution in Fig. 15, we see that any conjectured solution, x_o?, must be too large if, in equation 83, the left side exceeds the right and too small if the reverse is true. This criterion could also be used on a trial functional solution $x_o(n)$?. Now, if the left side of equation 83 is functionally stronger in n than the right, we know that our trial solution $x_o(n)$? is too strong in n. Again the reverse is also true.

After several guesses we are led to the trial functional solution $x_o(n) = [a(\ln n)/Bn]^{1/2}$ with which the right side of equation 83 is greater than the left for a $\le \tfrac{1}{2}$, and the reverse is true for $a > \tfrac{1}{2}$.

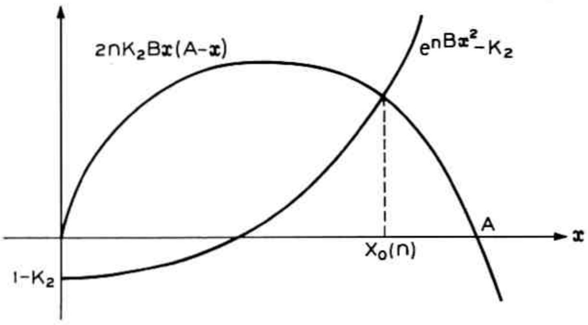


Fig. 15 — The graphical solution of equation 83.

This determines the highest order term of $x_o(n)$ and we can write

$$(\tfrac{1}{2})^{\frac{1}{2}}\left(\frac{\ln n}{Bn}\right)^{\frac{1}{2}}[1 + o(1)] \leqq x_o(n) \leqq (\tfrac{1}{2} + \epsilon)^{\frac{1}{2}}\left(\frac{\ln n}{Bn}\right)^{\frac{1}{2}}[1 + o(1)].$$

It follows that

$$f[x_o(n), n] \geqq \left(\frac{1}{2B}\right)^{\frac{1}{2}}\left(\frac{\ln n}{n}\right)^{\frac{1}{2}}[1 + o(1)]$$

and, since the lower envelope is smaller than any individual $f(x, n)$, that

$$f[x_o(n), n] \leqq f\left[\left(\frac{1}{2B}\right)^{\frac{1}{2}}\left(\frac{\ln n}{n}\right)^{\frac{1}{2}}, n\right] = \left(\frac{1}{2B}\right)^{\frac{1}{2}}\left(\frac{\ln n}{n}\right)^{\frac{1}{2}}[1 + o(1)]. \qquad (84)$$

Although only an upper bound to $f(x, n)$ is required, both upper and lower bounds were found to show that the method used to obtain the desired lower envelope provides asymptotically tight results. Continuing the inequality in equation **82** by that in equation **84** provides our final upper bound to transmission distortion.

Theorem 11: The minimum attainable transmission distortion of the source $\mathcal{S}$, when used with the channel $\mathcal{C}$, is upper bounded by

$$d(\mathcal{S}) \leqq d_C + b\left(\frac{\ln n}{n}\right)^{\frac{1}{2}}[1 + o(1)] \qquad (85)$$

in which

$$b = \left(\frac{1}{2B}\right)^{\frac{1}{2}} = \frac{1}{(2)^{\frac{1}{2}}}\frac{1}{\mid s_o \mid}\left[\frac{1}{(b_1)^{\frac{1}{2}}} + \frac{1}{(b_2)^{\frac{1}{2}}}\right]$$

$$b_1 = \tfrac{1}{2}E_s''(R^* = C)$$

$$b_2 = \tfrac{1}{2}E''(C).$$

For a fixed source $\mathcal{S}$, we see from this theorem that the coefficient b is smallest when $\mathcal{S}$ is used with that channel (among those of equal capacity) for which the constant b_2 is largest. In the same way, the coefficient b is seen to be a decreasing function of b_1 when the channel is fixed. Since the constant b_2 is independent of the source and b_1 independent of the channel, our upper bound does not provide an indicator of matching between the source and channel as we obtained in the lower bound. This was actually expected since here we were forced to separate the source and channel with an interface containing at most e^{nC} points.

The coefficient b_1, though, has an interesting significance. It is equal to one-half the derivative $E_s''(R^* = C)$ which can be thought to indicate how fast the boundary of Q' initially moves away from $\mathbf{p}$ with increasing R. In turn, this indicates, in a reciprocal manner, the necessary rate of change of the rate required to handle source words with compositions just around $\mathbf{p}$, which are just less than typical. Thus, we can think of the coefficient b_1 as a type of "stretch factor"[16] for the source.

When the result in equation 85 is compared with the lower bound to distortion, we see that the $[(\ln n)/n]^{\frac{1}{2}}$ rate of approach to d_C is slower than the $1/n$ rate of approach of the lower bound. Mathematically, at least, the reason for the upper bound decreasing more slowly than $(1/n)^{\frac{1}{2}}$ is that, for small arguments, the lowest order term in the two exponents $E(R)$ and $E_s(R)$ is quadratic. Their form for large n, exp $-n(\Delta R)^2$, shows that values of ΔR larger than $(1/n)^{\frac{1}{2}}$ are required to have these terms go to zero with increasing n. Because the slope of the rate-distortion curve is nonzero, the corresponding values of distortion difference (Δd) must also be larger than $(1/n)^{\frac{1}{2}}$.

There is reason to think that this type of exponential term, and the consequential $[(\ln n)/n]^{\frac{1}{2}}$ rate of approach to d_C, is present in the upper bound because we have used threshold devices in the transmission system. One at the encoder leads to the first exponential term in equation 73 (we again disregard the double exponential term). It uses the rule in equation 54 to choose, for each source word $\mathbf{w}$, any decoder word $\mathbf{z}$ in list θ_1 at a distortion less than d^*. When list θ_1 is lacking such an entry, any $\mathbf{z}$ at all on the list is chosen which, since the members of θ_1 are chosen independently, is then independent of $\mathbf{w}$. The resulting distortion in this circumstance is usually much greater than d^*. In the next section we compare the performance of this encoder with another that does not use such a threshold and show that the source encoding alone need only contribute to a rate of approach to d_C equal to $(\ln n)/n$.

A second threshold operation in our system is at the channel decoder, but it is really dependent upon the coding of the entire system. It leads to the second exponential term in equation 73. To isolate its effect on the system performance, we assume that failure has not occurred at the encoder, that is, there does exist a $\mathbf{z}$ on θ_1 with $d(\mathbf{w}, \mathbf{z}) \leqq d^*$. Now if the channel decoder makes no error, we are assured that the resulting distortion is less than d^*. However, if an error is made, the believed channel input word $\mathbf{x}_1$ is different from the actual word $\mathbf{x}$; therefore the decoded word $\mathbf{z}_1$ is different from $\mathbf{z}_o$. Moreover, since the lists θ_1 and θ_2 are chosen independently, $\mathbf{z}_o$ and $\mathbf{z}_1$ are statistically independent. It follows that $\mathbf{z}_1$ and $\mathbf{w}$ are also statistically independent, and in consequence that the distortion $d(\mathbf{w}, \mathbf{z}_1)$ is usually much greater than d^*.

It is this threshold which, it is believed, cannot be eliminated when the signal space is constrained to contain at most $M = e^{nC}$ points, even if the lists θ_1 and θ_2 are chosen dependently. A heuristic argument in Ref. 8 suggests that with such a constrained signal set, the transmission distortion can approach d_C no more rapidly than as $n^{-\frac{1}{2}}$. This, of course, is a slower rate of approach to d_C than the a/n rate of approach of the corresponding lower bound to distortion that was derived using a signal set not constrained in size.

XVI. AN IMPROVED UPPER BOUND FOR NOISELESS CHANNELS

For the special case of a noiseless channel, the previously derived upper bound can be improved. Since such a channel contains e^C noiseless transitions, or "direct" paths, transmission of the encoder output is trivial and the communication problem is only one of source representation. For this representation we are allowed to choose, from an e^C letter representation alphabet, one representation letter for every source output letter. Just as one is allowed n uses of the channel to transmit an n-letter source output, one is allowed an n-letter representation word to approximate an n-letter source word.

We first state that if the threshold source encoder defined by equation 54 is used in the ensemble of representation codes θ_1 of Section XII, the ensemble average representation error is very similar to the ensemble average transmission error derived in the previous sections. The only difference in the derivation is that the Pr(channel error) term is no longer present in equation 57, nor in any succeeding equation, with the only result being that $b_2 = \infty$ in equation 85.

We note here that this particular result is valid only for sources that are not doubly-uniform, that is, having a uniform probability distribution and a distortion matrix in which all rows are permutations of one row vector and all columns are permutations of one column vector. The reason for this exclusion is that for doubly-uniform sources the exponential term in equation 73 involving $E_s(R)$ also vanishes, and the double exponential term involving δ, previously dropped as insignificant, now remains as the only term. It is instructive to delay further evaluation of the bound in this case until after the following upper bound to representation distortion is derived.

16.1 *Optimum Source Encoder*

We now derive an upper bound to the source representation error when an optimum source encoder is used in place of the threshold encoder of the previous section. The resulting upper bound will be seen to approach the asymptote, d_C, as $(\ln n)/n$. This represents an improvement upon the best previously known upper bound to source representation distortion[7] which approached d_C essentially as $n^{-1/2}$.

The coding ensemble used here is very similar to the set of codes, θ_1, used in Section XII. But now the size of the set, M, is set equal to e^{nC} for all n, rather than have it approach this size with increasing n. And, the probability with which each ensemble member is used,

$$\Pr(\text{code}) = p(\theta_1) = \prod_{i=1}^{M} g(\mathbf{z}_i),$$

is now governed by that probability distribution $g(\mathbf{z})$ equal to the output probability distribution of the test channel at the point (d_C, C) on the rate distortion curve for $\mathcal{S}$. Within each ensemble member the encoder chooses, for any occurring source word $\mathbf{w}$, that member $\mathbf{z}$ on θ_1 for which $d(\mathbf{w}, \mathbf{z})$ is minimum. Therefore, for each ensemble member the average distortion over all possible source events is

$$d(\theta_1) = \sum_{W^n} p(\mathbf{w})\Big[\min_{\substack{1 \le i \le M \\ \mathbf{z}_i \varepsilon \theta_1}} d(\mathbf{w}, \mathbf{z}_i)\Big]. \tag{86}$$

The ensemble average distortion is given by

$$\langle d(\theta_1)\rangle_{\text{av}} = \sum_{W^n} p(\mathbf{w}) \sum_{\theta_1} p(\theta_1)\Big[\min_{\substack{1 \le i \le M \\ \mathbf{z}_i \varepsilon \theta_1}} d(\mathbf{w}, \mathbf{z}_i)\Big]. \tag{87}$$

The set of quantities $d(\mathbf{w}, \mathbf{z}_i)$ in equation 87 could be thought of as a set of M independent and identically distributed random variables, each conditioned on $\mathbf{w}$ and governed by the word probability distribution $g(\mathbf{z})$. The minimum of this set, $d_{\min}(\mathbf{w})$, is then also a random variable, governed by the code probability distribution $p(\theta_1)$. The inner sum in equation 87 is, therefore, the expected value of $d_{\min}(\mathbf{w})$ and we can write

$$\langle d(\theta_1)\rangle_{\text{av}} = \sum_{W^n} p(\mathbf{w}) \int_0^{d_{\max}} d\, dF_{d_{\min}|\mathbf{w}}(d \mid \mathbf{w})$$

which, upon integration by parts, becomes

$$\langle d(\theta_1)\rangle_{\text{av}} = \sum_{W^n} p(\mathbf{w}) \int_0^{d_{\max}} [1 - F_{d_{\min}|\mathbf{w}}(d \mid \mathbf{w})]\, dd. \tag{88}$$

The conditional distortion random variables $d(\mathbf{w}, \mathbf{z}_i)$ are the same distortion variables used in Section XIII. Since they depend only upon the composition of $\mathbf{w}$, we can again perform the summation in equation

88 by integration over the composition space, thus

$$\langle d(\theta_1)\rangle_{\text{av}} = \underset{Q^H}{\int \cdots \int} P(\mathbf{q})\, d\mathbf{q} \int_0^{d_{\max}} [1 - F_{d_{\min}|\mathbf{q}}(d \mid \mathbf{q})]\, dd \tag{89}$$

$$\equiv \underset{Q^H}{\int \cdots \int} P(\mathbf{q})\, d\mathbf{q} \langle d_{\min}(\mathbf{q})\rangle_{\text{av}} . \tag{90}$$

The inner integrand in equation 89 is the probability that all M points on θ_1 have a distortion $d(\mathbf{w}, \mathbf{z})$ from $\mathbf{w}$ greater than d. Using the independence property of the members of θ_1, we can write this probability as

$$1 - F_{d_{\min}\mid\mathbf{q}}(d \mid \mathbf{q}) = [1 - G(d \mid \mathbf{q})]^M. \tag{91}$$

It can be seen from equation 16 that the variance of the variable d is proportional to $1/n$ for every $\mathbf{q}$. Therefore the function $[1 - G(d \mid \mathbf{q})]$, which for every n decreases monotonically from one to zero, approaches, with increasing n, a negative step at the value of distortion $d = E(d \mid \mathbf{q})$.

The same is also true of $[1 - G(d \mid \mathbf{q})]^M$ which approaches a negative step at some lower value of distortion, $d_C(\mathbf{q})$. This can be established using the following asymptotic upper and lower bounds to the distribution function $G(d \mid \mathbf{q})$ which are from Shannon[11] and Gallager[13]:

$$h(n, \mathbf{q}) \exp -nR(d, \mathbf{q}) \leqq G(d \mid \mathbf{q}) \leqq H(n, \mathbf{q}) \exp -nR(d, \mathbf{q}) \tag{92}$$

with

$$R(d, \mathbf{q}) \equiv \mu(s, \mathbf{q}) - s\mu'(s, \mathbf{q}) \tag{93}$$

$$0 < \mu'(s, \mathbf{q}) = d \leqq E(d \mid \mathbf{q})$$

and in which $h(n, \mathbf{q})$ and $H(n, \mathbf{q})$ are algebraically small functions of n. Therefore, within the range $0 < d \leqq E(d \mid \mathbf{q})$, the function in equation 91 can be bounded by

$$[1 - He^{-nR}]^{\exp nC} \leqq [1 - G(d \mid \mathbf{q})]^M \leqq [1 - he^{-nR}]^{\exp nC}; \tag{94}$$

which proves that $[1 - G(d \mid \mathbf{q})]^M$ must approach one when $R(d, \mathbf{q}) > C$ and zero when $R(d, \mathbf{q}) < C$. That the function $R(d, \mathbf{q})$ is monotone decreasing in d within $0 < d \leqq E(d \mid \mathbf{q})$ now establishes the stated limiting step function form of $[1 - G(d \mid \mathbf{q})]^M$ with $d_C(\mathbf{q})$ equal to the distortion value for which

$$R[d_C(\mathbf{q}), \mathbf{q}] = C. \tag{95}$$

The region of integration in equation 89 is thus conveniently divided into two parts: one over $[0, d_C(\mathbf{q}) + \Delta]$ in which the integrand is upper-bounded by unity, and the other $[d_C(\mathbf{q}) + \Delta, d_{\max}]$ in which the integrand is upper-bounded by its value at the lower limit. The result is

$$\langle d_{\min}(\mathbf{q})\rangle_{\mathrm{av}} \leqq d_C(\mathbf{q}) + \Delta + [d_{\max} - d_C(\mathbf{q}) - \Delta][1 - G(d_C(\mathbf{q}) + \Delta \mid \mathbf{q})]^M \tag{96}$$

which, with the use of the lower bound in equation 92, can be continued by

$$\langle d_{\min}(\mathbf{q})\rangle_{\mathrm{av}} \leqq d_C(\mathbf{q}) + \Delta + [d_{\max} - d_C(\mathbf{q}) - \Delta] \cdot\{1 - h \exp[-nR(d_C(\mathbf{q}) + \Delta, \mathbf{q})]\}^{\exp nC}.$$

Equation 67 allows the further continuation of this bound by:

$$\langle d_{\min}(\mathbf{q})\rangle_{\mathrm{av}} \leqq d_C(\mathbf{q}) + \Delta + [d_{\max} - d_C(\mathbf{q}) - \Delta] \cdot\exp(-h \exp\{n[C - R(d_C(\mathbf{q}) + \Delta, \mathbf{q})]\}). \tag{97}$$

Again the monotone decreasing property of $R(d, \mathbf{q})$ in d provides that the quantity $C - R(d_C(\mathbf{q}) + \Delta, \mathbf{q})$ is positive when Δ is positive and, therefore, that the last term in equation (97) is a decreasing double exponential in n.

Equation 97 actually provides, for each $\mathbf{q}$, a set of upper bounds to $\langle d_{\min}(\mathbf{q})\rangle_{\mathrm{av}}$ very similar to the family of curves studied in Section XV. In the choice of the parameter Δ there is once again a trade-off between a small asymptote, $d_C(\mathbf{q}) + \Delta$, and a fast rate of approach. It should, in general, be chosen to optimize the bound at each n. Since we want an upper bound to $\langle d_{\min}(\mathbf{q})\rangle_{\mathrm{av}}$ that approaches $d_C(\mathbf{q})$ with increasing n, the optimizing parameter $\Delta_o(n)$ clearly must approach zero as n increases. But $\Delta_o(n)$ must approach zero in a way that also allows the last term of equation 97 to vanish.

Since an asymptotic bound is our goal, we extract the essential behavior of this term for small Δ by forming a Taylor series of $R(d, \mathbf{q})$ at $d = d_C(\mathbf{q})$:

$$\begin{aligned} C - R(d_C(\mathbf{q}) + \Delta, \mathbf{q}) &= -\Delta R'(d_C(\mathbf{q}), \mathbf{q}) + o(\Delta) \\ &= -s\Delta + o(\Delta). \end{aligned}$$

In this expression s is the parameter value in equation 93 when d equals $d_C(\mathbf{q})$. Thus the lower envelope to the set of bounds in equation 97 can be written, for the purpose of an asymptotic bound, as

$$\langle d_{\min}(\mathbf{q})\rangle_{\mathrm{av}} \leqq \min_{\Delta} \{d_C(\mathbf{q}) + \Delta + [d_{\max} - d_C(\mathbf{q}) - \Delta] \exp(-he^{-sn\Delta})\}.$$

The minimization is found using the same method used in Section XV. In this process, it is important to notice that Shannon's coefficient $h(n, \mathbf{q})$ in equation 92 is proportional to $n^{-\frac{1}{2}}$. The result is that the optimizing parameter satisfies

$$\frac{1}{2}\frac{\ln n}{-sn}[1 + o(1)] \leqq \Delta_o(n) \leqq \left(\frac{1}{2} + \epsilon\right)\frac{\ln n}{-sn}[1 + o(1)]$$

and that $\langle d_{\min}(\mathbf{q})\rangle_{\text{av}}$ satisfies

$$\langle d_{\min}(\mathbf{q})\rangle_{\text{av}} \leqq d_C(\mathbf{q}) + \left(\frac{1}{2} + \epsilon\right)\frac{\ln n}{-sn}[1 + o(1)]. \tag{98}$$

Returning to equation 90, the ensemble average representation error therefore can be upper bounded by

$$\langle d(\theta_1)\rangle_{\text{av}} \leqq \underset{Q^H}{\int \cdots \int} P(\mathbf{q})\left[d_C(\mathbf{q}) + \left(\frac{1}{2} + \epsilon\right)\frac{\ln n}{-sn}\right]d\mathbf{q}. \tag{99}$$

The above integral is evaluated in the same way similar averages were found for the lower bound. The bracketed quantity is expanded in a Taylor series about $\mathbf{q} = \mathbf{p}$ and is truncated after three terms with a Lagrange remainder term. Upon integration of this expansion we find

$$\begin{aligned}\langle d(\theta_1)\rangle_{\text{av}} \leqq{}& d_C(\mathbf{p}) + \left(\frac{1}{2} + \epsilon\right)\frac{\ln n}{-s_on}\\ &+ \sum_i \frac{\partial}{\partial q_i}\left[d_C(\mathbf{q}) + \left(\frac{1}{2} + \epsilon\right)\frac{\ln n}{-sn}\right]_{\mathbf{p}} E(q_i - p_i)\\ &+ \sum_{ij}\frac{\partial^2}{\partial q_i\,\partial q_j}\left[d_C(\mathbf{q}) + \left(\frac{1}{2} + \epsilon\right)\frac{\ln n}{-sn}\right]_{\boldsymbol{\varphi}} E[(q_i - p_i)(q_j - p_j)]\end{aligned} \tag{100}$$

with $s_o \equiv s(\mathbf{p})$ and $\boldsymbol{\varphi} \in Q^H$.

Using the following expected values in equation (100),

$$E(q_i - p_i) = 0$$

$$E[(q_i - p_i)(q_j - p_j)] = \frac{1}{n}(p_i\,\delta_{ij} - p_ip_j),$$

we have the following upper bound to the ensemble average distortion and, therefore, to the minimum attainable representation error.

Theorem 12: The minimum attainable transmission distortion (representation distortion) of the source $\mathcal{S}$, *when used with a noiseless channel* *of capacity C, is upper bounded by*

$$d(\mathcal{S}) \leqq d_C + \left(\frac{1}{2} + \epsilon\right) \frac{\ln n}{-s_o n} [1 + o(1)] \tag{101}$$

in which s_o satisfies

$$\mu(s_o, \mathbf{p}) - s_o\mu'(s_o, \mathbf{p}) = -C.$$

Except for the arbitrarily small positive ϵ, the bound in equation 101 agrees precisely with the asymptotic lower bound that we found earlier in this paper.

We see by comparing equation 85 (with $b_2 = \infty$ for the noiseless channel) and equation 101 that the replacement of the threshold source encoder with an optimum encoder increases the rate of approach to the asymptote from $[(\ln n)/n]^{\frac{1}{2}}$ to $(\ln n)/n$. To obtain some feeling for the reason for this improvement, we might think of the optimum encoder as a threshold encoder, *but* with a threshold that varies depending on the particular source output. Indeed, we used this step within the mathematics when we separated all events (equation 96) into two sets with the separation dependent upon the source word. In particular, for any source output word with composition $\mathbf{q}$, we used a threshold, $d_C(\mathbf{q}) + \Delta$, just large enough so that for large n there is almost surely a representation word in θ_1 that is acceptable. It does not require, as does the fixed threshold encoder, that the set of source words not meeting a fixed distortion level of d^* have a total probability that goes to zero with n. This restriction is really more severe than one would think we need, since some of the source words $\mathbf{w}$ discarded by the fixed threshold encoder are just outside $\mathbf{p}$, having characteristics just less than typical, for which some of the distortions $d(\mathbf{w}, \mathbf{z}_i)$ might be only marginally greater than any fixed d^*.

16.2 *The Special Case of a Double Uniform Source*

There is one situation for which both source encoders provide a representation distortion that approaches the limit d_C as $(\ln n)/n$. This is when the source $\mathcal{S}$ is doubly-uniform. Since $\mu(s, \mathbf{q})$ is independent of $\mathbf{q}$ for such a source, $R(d^*, \mathbf{q})$ in equation 61 is also independent $\mathbf{q}$, with the result that the set Q' in equation 66 is always empty. Therefore, $\Pr(Q') = 0$ in equation 69 and we have for the set of upper bounds to representation distortion, using threshold encoders:

$$d(\mathcal{S}) \leqq d^* + (d_{\max} - d^*) \exp(-he^{n\delta}).$$

In this bound we have used the lower bound in equation 92 rather than that in equation 61. It can now be shown, using precisely the same procedure as before, that this set of bounds approaches the limit d_C as $(\ln n)/n$.

XVII. SUMMARY

We have presented upper and lower bounds to the minimum attainable transmission distortion of a source measured by a specified distortion measure. The bounds, which were derived for both noisy and noiseless channels, have all been shown to converge to the same level of distortion, d_C, algebraically in the block length n. The quantity d_C is that level of distortion shown by Shannon to be the minimum attainable transmission distortion when the channel capacity is C and arbitrarily complex transmission methods are allowed.

For noisy channels, the rate of approach of the lower bound to d_C is as a/n and that of the upper bound as $b[(\ln n/n)]^{\frac{1}{2}}$. The nonnegative coefficients a and b are both functions of the statistics of the source and channel, but have different forms. The lower bound coefficient, a, interrelates these statistics in such a way as to suggest its utility as a measure of "mismatch" between the source and channel, the larger a, the slower the rate of approach of the bound to d_C, and the larger the source-channel mismatch. This coefficient is, of course, necessarily equal to zero whenever the source and channel are perfectly matched, that is, whenever the minimum attainable transmission distortion is equal to d_C for all block lengths, n.

The coefficient b in the upper bound, though, does not present an indicator of source-channel mismatch. It is the sum of two terms which separately contain the source statistics and the channel statistics. The cause of this separation is the interface between the source and channel that results from the use of a transmitting signal set constrained to contain at most e^{nC} members, a constraint which we found necessary to introduce in the development of the bound.

For noiseless channels, both the upper and lower bounds to the transmission distortion (or the source representation distortion) have the same form. They both have been shown to approach the asymptote d_C as $a_1 (\ln n)/n$.

REFERENCES

1. Shannon, C. E., "A Mathematical Theory of Communications," B.S.T.J., 27, Nos. 3 and 4 (July and October 1948), pp. 379–423, 623–656.

2. Shannon, C. E., "Coding Theorems for a Discrete Source with a Fidelity Criterion," IRE National Convention Record, Part 4 (1959), pp. 142–163.
3. Holsinger, J. L., unpublished work.
4. Goblick, T. J., "Theoretical Limitations on the Transmission of Data from Analog Sources," IEEE Trans. Inform. Theory, *IT-11* (October 1965), pp. 558–567.
5. Gerrish, A. M. and Shultheiss, P. M., "Information Rates on Non-Gaussian Processes," IEEE Trans. Inform. Theory *IT-10* (October 1964), pp. 265–271.
6. Pinkston, J. T., "Information Rates of Independent Sample Sources," S. M. Thesis, Department of Electrical Engineering, M.I.T., Cambridge, Massachusetts (1966).
7. Goblick, T. J., "Coding for a Discrete Information Source with a Distortion Measure," Ph.D. Thesis, Department of Electrical Engineering, M.I.T., Cambridge, Massachusetts (1962).
8. Pilc, R. J., "Coding Theorems for Discrete Source-Channel Pairs," Ph.D. Thesis, Department of Electrical Engineering, M.I.T., Cambridge, Massachusetts (1967).
9. Pilc, R. J., unpublished work.
10. Chernov, H., "A Measure of Asymptotic Efficiency for Tests of an Hypothesis Based on a Sum of Observations," Ann. Math. Stat. *23* (1952), pp. 493–507.
11. Shannon, C. E., unpublished work.
12. Fano, R. M., *The Transmission of Information,* New York: Wiley, 1961.
13. Gallager, R. G., "Lower Bounds on the Tails of Probability Distributions," M.I.T. Research Lab. of Electronics, Quart. Progress Report, *77* (April 1965), pp. 277–291.
14. Gallager, R. G., unpublished work.
15. Gallager, R. G., "A Simple Derivation of the Coding Theorem and Some Applications," IEEE Trans. Inform. Theory, *IT-11* (January 1965), pp. 3–18.
16. Wozencraft, J. M. and Jacobs, I. M., *Principles of Communication Engineering,* New York: John Wiley (1965).

Some Considerations of Stability in Lossy Varactor Harmonic Generators

By C. DRAGONE and V. K. PRABHU

(Manuscript received March 4, 1968)

Explicit expressions are derived for the scattering parameters which relate small-signal fluctuations in a lossy varactor harmonic generator of order $N = 2^n$, n an integer. The effect of losses on the stability of the mutiplier is then studied. The very important particular case is then examined in which all the losses occur in the series resistance of the varactor diode, and it is shown that absolute stability is obtained provided the efficiency η_t of the multiplier $\ll N^{-1}$, because of the particular distribution of the losses at various carrier frequencies. Therefore, the conclusion is reached that in most cases of practical interest restrictions have to be placed on the available circuit configurations to prevent instability of the multiplier.

I. INTRODUCTION

A serious limitation to efficient wideband harmonic generation with varactor diodes is that instability in the multiplier might cause the generation of spurious tones.[1] It is the purpose of this paper to study the effect of losses on stability of abrupt-junction varactor frequency multipliers of order $N = 2^n = 2$, 4, and so on, with the minimum number of idlers.

The type of instability considered here is the one discussed in Refs. 2, 3, and 4. It produces undesired low-frequency fluctuations in the amplitude and phase of the output harmonic and is caused by the time-varying elastance of the varactor, which is potentially unstable with respect to phase perturbations.

The stability conditions of lossless abrupt-junction varactor multipliers have already been extensively discussed elsewhere in Refs. 3 and 4. More precisely, these works have shown that, in the absence of any losses in the varactor diode, the frequency characteristics of the input, output, and idler circuits must satisfy certain restrictions in order that the multiplier be stable. The main objective of this paper is to determine the amount of loss that the multiplier must have in order to be absolutely stable, that is, stable for arbitrary linear passive input, output, and idler circuits.

First we show that the over-all multiplier efficiency η_t can be expressed as the product of the efficiencies of the input, output, and idler circuits; that is

$$\eta_t = \eta_1 \times \eta_2 \times \cdots \eta_N ,$$

where η_t represents the ratio of the power P_L delivered to the output at carrier frequency $N\omega_o$ to the power supplied by the input pump at carrier frequency ω_o. The partial efficiency η_r is the efficiency of the circuit at the carrier frequency $r\omega_o$, or $1 - \eta_r$ represents the ratio of the power lost at $r\omega_o$ to the sum of P_L and of the total power lost at the frequencies $r\omega_o$, $2r\omega_o$, $\cdots$, $N\omega_o$.

Next we show that the behavior of the multiplier with respect to small amplitude and phase fluctuations is related in a simple way to the efficiencies η_1, η_2, and so on. For instance, in the case of very slow fluctuations, the *PM* scattering parameters of a doubler are given by the matrix,

$$\begin{bmatrix} 0 & -\eta_1\eta_2 \\ 2 & 1 - 2\eta_2(1 - \eta_1) \end{bmatrix}.$$

In the last two sections we examine the conditions of absolute stability and show that the multiplier may become unstable for some circuit conditions if

$$\eta_t > 1/N.$$

If, on the other hand,

$$\eta_t < 1/N,$$

then the multiplier is absolutely stable if and only if

$$\eta_r < 50\%, \quad \text{for} \quad r = 1, \cdots, N/2.$$

Finally, the important particular case is considered in which all the losses of the multiplier occur in the series resistance of the varactor. It is found that in this case absolute stability is obtained if and only if

$$\frac{\omega_o}{\omega_c} > 0.06, \qquad N = 2,$$

$$\frac{\omega_o}{\omega_c} > 0.1, \qquad N > 2,$$

where ω_c is the cutoff frequency of the varactor. If these conditions are satisfied, then the efficiency of the multiplier is found to be so low that the conclusion is reached that in most cases of practical interest restrictions must be placed on the available circuit configuration in order to obtain stability.

II. SCATTERING RELATIONS

Nominally driven abrupt-junction varactor frequency multipliers of order 2^n come under the general class of pumped nonlinear systems, and the general method presented in Ref. 5 can be used for such systems to obtain the scattering parameters which relate small-signal fluctuations that may be present at various points in the system.* These small-signal fluctuations are assumed to be small and they are at frequencies close to the carriers.

The varactor model that we use is shown in Fig. 1. It is a variable capacitance in series with a resistance R_s. The multiplier has the minimum number of idlers. The linear passive circuits used in the multiplier as input, output, and idler terminations are assumed to produce no amplitude to phase or phase to amplitude conversion.† If input, output, and all idler circuits are tuned,‡ it can be shown[4–6] that the small-signal terminal relations of a harmonic generator can be expressed in the form (see Fig. 2)

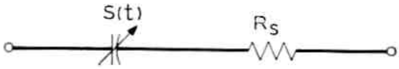


Fig. 1 — Varactor model.

$$\begin{bmatrix} (m_r)_1 \\ (m_r)_{2^n} \\ (\theta_r)_1 \\ (\theta_r)_{2^n} \end{bmatrix} = \left[\begin{array}{c|c} \underline{S}_{aa} & \underline{0} \\ \hline \underline{S}_{pa} & \underline{S}_{pp} \end{array}\right] \begin{bmatrix} (m_i)_1 \\ (m_i)_{2^n} \\ (\theta_i)_1 \\ (\theta_i)_{2^n} \end{bmatrix} \tag{1}$$

* Notation in this paper is identical to that in Refs. 4 and 5. Details of these notations are not given in this paper for the sake of brevity.

† This condition is satisfied by circuits usually used with multipliers.[4]

‡ Tuning of idlers, and input and output circuits usually gives near optimum efficiency for the multipliers. (See Refs. 7, 8, and 9.)

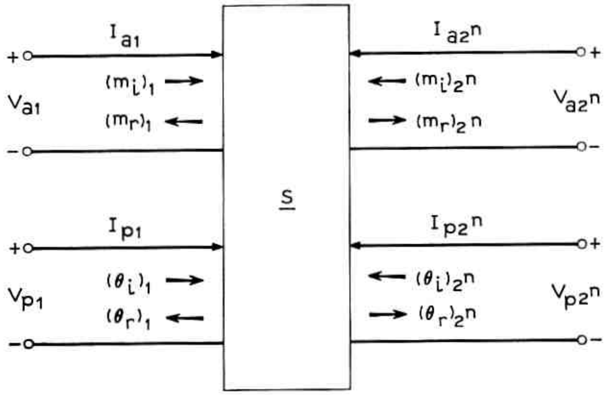


Fig. 2 — Small-signal terminal behavior of a harmonic generator of order 2^n. m is the AM index and θ is the PM index of the multiplier.

or*

$$\mathbf{b} = \underset{\sim}{S}\mathbf{a} \tag{2}$$

where $\underset{\sim}{S}$ is the scattering matrix of the multiplier and $(m_i)_j$ is the incident AM index at carrier frequency $j\omega_o$, $(\theta_r)_k$ is the reflected PM index at carrier frequency $k\omega_o$, and so on. The small-signal fluctuations in the vicinity of carrier frequency $k\omega_o$ are assumed to be at $k\omega_o \pm \omega$, $\omega < \omega_o/2$.

It can also be shown[4] that the AM scattering matrix $\underset{\sim}{S}_{aa}$ and the PM scattering matrix $\underset{\sim}{S}_{pp}$ are independent of the bias source impedance Z_o, and that the stability of the multiplier is completely determined by $\underset{\sim}{S}_{aa}$ and $\underset{\sim}{S}_{pp}$. It can also be shown[4] that a multiplier of order 2^n is stable with respect to its AM fluctuations for all input, output, and idler terminations. In this paper we shall, therefore, obtain an expression for $\underset{\sim}{S}_{pp}$ for a varactor harmonic generator of order 2^n with the minimum number of idlers† and consider its PM stability.

An abrupt-junction varactor multiplier of order 2^n with the least number of idlers can be shown[3,5] to be completely equivalent to a cascade of n lossless doublers‡ as shown in Fig. 3. Z_{2k}, $0 \leqq k \leqq n$, is the termination impedance in the vicinity of carrier frequency $2^k\omega_o$.

* A column matrix is written in the form $\mathbf{a}$, a matrix which is square is written as $\underset{\sim}{A}$, and a unit matrix of order n is written as $\underset{\sim}{1}_n$.

† Methods given in Ref. 5 can, in all cases, be used to obtain $\underset{\sim}{S}$ in equation (2).

‡ The conditions under which a multiplier of order $M_1 \times M_2$ is completely equivalent to a cascade of two multipliers of order M_1 and M_2 are given in Ref. 5.

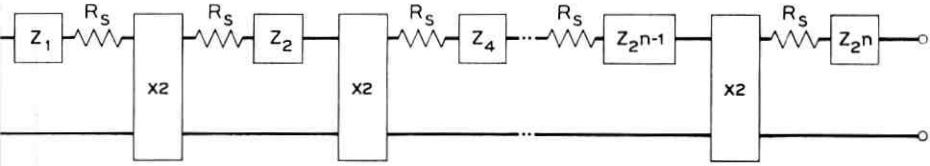


Fig. 3 — Equivalence of an abrupt-junction varactor multiplier of order 2^n to a chain of n doublers. R_s is the series resistance of the varactor diode.

Input and output circuits which are not shown in Fig. 4 can be any arbitrary linear passive circuits. For $\omega/\omega_o \ll 1$, it can be shown that the AM scattering matrix $\underset{\sim}{S}_{aa}$ and PM scattering matrix $\underset{\sim}{S}_{pp}$ of the k^{th} lossless doubler are given[5] by

$$\underset{\sim}{S}_{aa} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ 1 & 0 \end{bmatrix} \tag{3}$$

and

$$\underset{\sim}{S}_{pp} = \begin{bmatrix} 0 & -1 \\ 2 & 1 \end{bmatrix}. \tag{4}$$

Now let us consider the k^{th} lossless doubler. The "input impedance" $(R_{ok})_{\text{in}}$ and the "load impedance" $(R_{ok})_{\text{out}}$ of the lossless doubler are given by[5, 7]

$$(R_{ok})_{\text{in}} = \frac{|S_{2^k}|}{2^{k-1}\omega_o}, \qquad 1 \leqq k \leqq n \tag{5}$$

and

$$(R_{ok})_{\text{out}} = \frac{|S_{2^{k-1}}|^2}{2^{k+1}|S_{2^k}|\omega_o}, \qquad 1 \leqq k \leqq n. \tag{6}$$

Since all impedances are purely resistive, we can define partial efficiencies η_{2^i}'s by the relations

$$\eta_{2^k} = \frac{[R_{o(k+1)}]_{\text{in}}}{Z_{2^k} + R_s + [R_{o(k+1)}]_{\text{in}}}, \qquad 0 \leqq k \leqq n - 1 \tag{7}$$

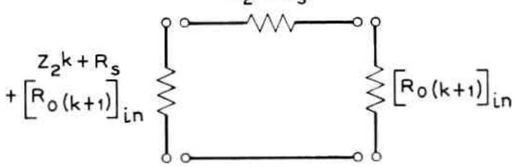


Fig. 4 — An interstage network used with the multiplier.

and

$$\eta_{2^n} = \frac{Z_L}{Z_{2^n} + R_s + Z_L} \tag{8}$$

where Z_L is the load resistance connected to the multiplier. We notice that η_{2^k}, $1 \leqq k \leqq n - 1$ is equal to the ratio of carrier power flowing into the input port of $(k + 1)^{\text{th}}$ doubler to that supplied by the k^{th} doubler, η_1 the ratio of carrier power supplied to the first doubler to the power supplied by the pump, and that η_{2^n} is the ratio of power dissipated in the load resistor Z_L to that supplied by the n^{th} doubler. The over-all efficiency η_t of the multiplier can, therefore, be written as

$$\eta_t = \prod_{r=0}^{n} \eta_{2^r} . \tag{9}$$

Consider Fig. 4. The scattering matrix of the $(k + 1)^{\text{th}}$ interstage network can be shown to be[5, 6, 10]

$$\begin{bmatrix} 0 & \eta_{2^k} \\ 1 & 1 - \eta_{2^k} \end{bmatrix} . \tag{10}$$

If $\omega/\omega_0 \ll 1$, we can then show from equations (4) and (10) that the PM scattering matrix $\underline{S}_{pp}$ for the multiplier shown in Fig. 3 can be written as

$$\underline{S}_{pp} = \begin{bmatrix} 0 & (-1)^n \eta_t \\ 2^n & 1 + \sum_{r=0}^{n-1} (-2)^{r+1} \eta_{2^n} \eta_{2^{n-1}} \cdots \eta_{2^{n-r}} - (-2)^n \eta_t \end{bmatrix} . \tag{11}$$

III. DERIVATION OF THE ABSOLUTE STABILITY CONDITIONS

First, consider the case of a doubler. The scattering matrix of a stage consisting of an ideal doubler with two resistances $R_{s,1}$ and $R_{s,2}$ connected* in series to the input and output ports, respectively, is:†

$$\begin{bmatrix} 0 & -\eta_1 \eta_2 \\ 2 & 1 - 2\eta_2(1 - \eta_1) \end{bmatrix} . \tag{12}$$

By means of standard techniques,[4] one obtains that absolute stability requires that‡

$$2\eta_1\eta_2 + | 1 - 2\eta_2 + 2\eta_1\eta_2 | \leqq 1 \tag{13}$$

* Notice that $R_{s,1} = Z_1 + R_s$ and $R_{s,2} = Z_2 + R_s$.

† Put $n = 1$, $N = 2$, $\eta_t = \eta_1\eta_2$ in equation (11). See also Appendix A for an alternate derivation of absolute stability conditions.

‡ Condition (13) requires that the magnitude of the largest output reflection that can be obtained when the termination of the input port is passive be less than unity.

which is satisfied if and only if

$$\eta_1 < 0.5. \tag{14}$$

It is important to notice that (14) shows that the output circuit losses do not have any effect on the absolute stability conditions of a doubler. This property will be used in the following discussion of the absolute stability of a multiplier of order $N > 2$.

Consider a multplier with $n > 1$. It can be shown that in this case it is necessary and sufficient that

$$\eta_1 < 0.5,\ \eta_2 < 0.5,\ \cdots,\ \eta_{N/2} < 0.5. \tag{15}$$

The fact that (15) guarantees absolute stability follows directly from (14) and the fact that a chain of absolutely stable stages is stable.

In order to show the necessity of (15), consider the k^{th} ideal doubler of Fig. 5, and the impedances presented to its input and output ports by the remaining part of the circuit. The impedance presented to the input port is given by

$$Z_{1,k} = Z_{2^{k-1}} + R_s + Z_{o,(k-1)}\,. \tag{16}$$

Since $Z_{o,(k-1)}$ approaches zero as the magnitude of $Z_{2^{k-1}}$ approaches infinity, $Z_{1,k}$ can have all complex values with nonnegative real part. Furthermore, the impedance $Z_{2,k}$ terminating the output port of the k^{th} ideal doubler has arbitrary imaginary part, because of the presence of Z_{2^k}. Therefore, since (14) shows that the absolute stability of a doubler does not depend on the real part of the output impedance, one concludes that it is necessary that

$$\eta_{2^k} < 0.5, \qquad 0 \leqq k \leqq n - 1, \tag{17}$$

if the chain is to be stable for all allowable values of $Z_{2^{k-1}}$, Z_{2^k}, and $Z_{2^{k+1}}$.

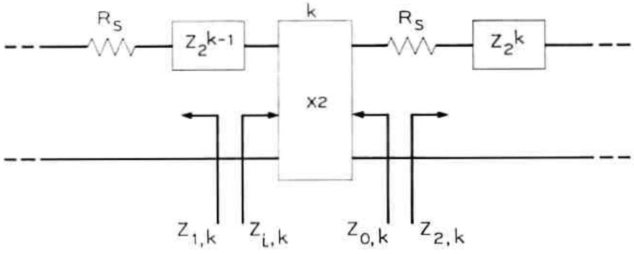


Fig. 5 — Lossless k^{th} doubler.

IV. ABSOLUTE STABILITY CONDITIONS

Equation (15) shows that a multiplier of order $N = 2^n$ of the type considered in this paper will become unstable for some frequency characteristics of the input, output, and idler circuits, if the efficiency η_t is greater than N^{-1}, that is, if

$$\eta_t > 1/N. \tag{18}$$

Therefore, if equation (17) is satisfied then the circuit must satisfy certain conditions such as those derived in Refs. 2, 3, and 4, in order that the multiplier be stable. If, on the other hand,

$$\eta_t < 1/N, \tag{19}$$

then (15) shows that the multiplier will be stable for all circuit conditions if and only if the efficiencies of the input and idler circuits are all less than 50 per cent.

At this point the particular case

$$Z_{2^k} = 0, \qquad 0 \leqq k \leqq n \tag{20}$$

deserves special attention. This represents in fact the important case in which all the losses occur in the series resistance R_s of the varactor. It will be assumed that the output load has the particular value which gives maximum efficiency.[7]

For the absolute PM stability of a doubler, equation (15) requires that

$$\eta_1 < 0.5. \tag{21}$$

We can show* that this condition can only be satisfied if and only if the overall efficiency

$$\eta_t < 36\%. \tag{22}$$

In the case of a quadrupler, the condition of absolute stability requires that

$$\eta_1 < 0.5 \tag{23}$$

and

$$\eta_2 < 0.5. \tag{24}$$

We can show† that equations (23) and (24) can be satisfied if and

* From Ref. 7, p. 331, $\eta_1 < 0.5$ for $\omega_o/\omega_c > 0.06$. For this value of ω_o/ω_c, $\eta_t <$ 36 percent.

† See Ref. 7, pp. 364–365.

only if

$$\eta_t < 0.7\%. \tag{25}$$

It therefore follows that for absolute stability of multipliers of order 2^n, it is necessary that

$$\eta_t \ll 2^{-n}. \tag{26}$$

Thus, if (15) is satisfied then the multiplier is so inefficient that it becomes of little practical interest. Therefore one concludes that, if all the losses occur in the series resistance of the varactor, in most cases of practical interest the question of stability cannot be neglected and the frequency characteristics of the input, output, and idler circuits have to satisfy certain restrictions (such as those given in Refs. 2, 3, and 4) in order to guarantee stability of the multiplier.

V. RESULTS AND CONCLUSIONS

Scattering relations for lossy abrupt-junction varactor harmonic generators are presented in this paper. Explicit expressions have been given for the PM scattering parameters of the multiplier in terms of partial efficiencies defined for the multiplier.

Absolute PM stability of 2^n multipliers is then considered. It is shown that the multiplier is stable if and only if

$$\eta_{2^j} < 0.5, \qquad 0 \leqq j \leqq n - 1. \tag{15}$$

We have also shown that a multiplier of order 2^n and having all the losses occur in the series resistance R_s of the varactor diode is absolutely stable if its efficiency is much lower than 2^{-n}, the inverse of order of multiplication of the multiplier.

The problem of stability is then of major importance in all high efficiency varactor multipliers and proper circuits should always be designed to assure at least the conditional stability of these multipliers.[2–4]

APPENDIX

PM Stability of 2^n Multipliers

Let us investigate by an alternate method absolute PM stability of 2^n multipliers for $n \geq 1$. Let us consider the k^{th} lossless doubler (see Fig. 5) in the equivalent circuit shown in Fig. 3.

If $Z_{i,k}$ and $Z_{o,k}$ are the phase terminating impedances of the k^{th} lossless doubler (see Fig. 5), we can derive from equations (4) through (6) that

$$Z_{o,k} = R_s\left\{1 + \frac{1}{2}\,\frac{\dfrac{m_{2^{k-1}}^2}{2^{2(k-1)}}\left(\dfrac{\omega_c}{\omega_o}\right)^2}{[Z_{o,(k-1)}/R_s] + [Z_{2^{k-1}}/R_s] + 1 - \dfrac{1}{2^{k-1}}\,m_{2^k}\,\dfrac{\omega_c}{\omega_o}}\right\}$$

$$1 \leqq k \leqq n; \qquad (27)$$

and

$$Z_{i,k} = R_s\left\{1 - \frac{1}{2^{k-1}}\,m_{2^k}\,\frac{\omega_c}{\omega_o} + \frac{1}{2}\,\frac{\dfrac{m_{2^{k-1}}^2}{2^{2(k-1)}}\left(\dfrac{\omega_c}{\omega_o}\right)^2}{[Z_{i,(k+1)}/R_s] + [Z_{2^k}/R_s] + 1}\right\},$$

$$1 \leqq k \leqq n \qquad (28)$$

where m_k is the modulation ratio of the varactor at carrier frequency $k\omega_o$.

Since Z_{2^k}'s are all linear passive impedances, it is seen from eqs. (27) and (28) that the multiplier is absolutely stable with respect to PM fluctuations if and only if

$$\frac{m_{2^k}}{2^{k-1}}\left(\frac{\omega_c}{\omega_o}\right) < 1, \qquad 1 \leqq k \leqq n. \qquad (29)$$

If any of these conditions are not satisfied, the multiplier will become unstable for a certain set of Z_{2^k}'s.

REFERENCES

1. Hines, M. F., Bloidsell, A. A., Collins, F., and Priest, W., "Special Problems in Microwave Harmonic Generator Chain," Digest Technical Papers, 1962 International Solid-State Circuits Conference, Philadelphia, Pa.
2. Dragone, C., "AM and PM Scattering Properties of a Lossless Multiplier of Order $N = 2^n$," Proc. IEEE, 54, No. 12 (December 1966), p. 1949.
3. Dragone, C., "Phase and Amplitude Modulation in High Efficiency Varactor Frequency Multipliers of Order $N = 2^n$—Stability and Noise," B.S.T.J., *46,* No. 4 (April 1967), pp. 799–836.
4. Prabhu, V. K., "Stability Considerations in Lossless Varactor Frequency Multipliers," B.S.T.J., *46,* No. 9 (November 1967), pp. 2035–2060.
5. Dragone, C. and Prabhu, V. K., "Scattering Relations in Lossless Varactor Frequency Multipliers," B.S.T.J., *46,* No. 8 (October 1967), pp. 1699–1731.
6. Dragone, C., "Phase and Amplitude Modulation in High Efficiency Varactor Frequency Multipliers—General Scattering Properties," B.S.T.J., *46,* No. 4 (April 1967), pp. 777–798.
7. Penfield, P., Jr. and Rafuse, R. P., *Varactor Applications,* Cambridge, Mass.: M.I.T. Press, 1962.
8. Burckhardt, C. B., "Analysis of Varactor Frequency Multipliers for Arbitrary Capacitance Variation and Drive Level," B.S.T.J., *44,* No. 4 (April 1965), pp. 675–692.
9. Prabhu, V., "Noise Performance of Abrupt-Junction Varactor Frequency Multipliers," Proc. IEEE, *54,* No. 2 (February 1966), pp. 285–287.
10. Carlin, H. J. and Giordano, A. B., *Network Theory,* Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1964.

Computer-Aided Analysis of Cassegrain Antennas

By H. ZUCKER and W. H. IERLEY

(Manuscript received December 8, 1968)

A method of analyzing, in detail, the performance of symmetrical Cassegrain antennas has been developed that uses a digital computer efficiently. For a specified antenna geometry and feed excitation, the program will compute and graphically display the amplitude and phase illumination of the subreflector, main reflector, and far-field pattern. These results may be used to optimize antenna performance by changing parameters and observing the effect.

Analysis of a Cassegrain antenna with a near-field conical horn feed is discussed as an application of the method. Because the radiation characteristics of the horn are determined by the horn flare angle rather than the horn aperture, broadband performance is obtained. It was indeed found that a 50 per cent bandwidth is achieved with a dual mode $TE_{11} - TM_{11}$ *mode feed, provided the proper phase relationship between the modes can be maintained over the band. For dual mode excitation an aperture efficiency of 70% and a noise temperature due to the power loss at the sub and main reflectors of less than 6.5°K was obtained. For a single mode feed* (TE_{11})*, there was a degradation in the E-plane side lobe levels and a corresponding 10°K increase in noise temperature. Excitation in the* TM_{01} *mode was also examined for angle-error sensing purposes. Also, the antenna can be used with reasonable efficiency well below the design frequency in which case it functions as a far-field fed Cassegrain antenna.*

I. INTRODUCTION

The essential radiation characteristics of multiple reflector antennas can be predicted very accurately with existing analytical and computational methods. Previous work on the open Cassegrain antenna showed that good agreement can be achieved between calculated and experimental results.[1] Deviations occurred mainly in the sidelobe regions of the radiation patterns, and these had only a small effect on overall antenna performance.

We are concerned here with a simpler problem, but one of perhaps more general interest: the analysis of symmetrical Cassegrain antennas. We present a computational method in which the amplitude and phase illuminations of the subreflector and main-reflector, as well as the far-field radiation pattern, are determined in detail, given the geometry of the antenna, the dimensions of the feed horn, and the excitation modes of the feed. The analysis includes near-field excitation—an important configuration for broadband operation. Included in the program is a graphic routine which plots all radiation patterns, the intermediate illuminations and the final far-field results. Because of the this feature, and the fact that only seven parameters are required to define the geometry of the antenna, the program is particularly useful for optimizing antenna performance.

The symmetry of the antenna results in improved computational efficiency. For the open Cassegrain antenna, double-integration was required to compute radiation patterns. An approximation recently was obtained[2] which, when applied to symmetrical Cassegrain antennas, eliminates one integration with only a small reduction in accuracy.[2] This makes it possible to compute the radiation characteristics of large Cassegrain antennas a few hundred wavelengths diameter in minutes.

II. NEAR-FIELD SYMMETRIC CASSEGRAIN ANTENNA

The antenna under consideration was intended to be used as the ground station of a satellite communications system. Wide-bandwidth (25 per cent) and low-noise requirements motivated the choice of a near-field conical-horn symmetric Cassegrain configuration. The near-field feed produces relatively low spillover at the subreflector, resulting in a lower noise temperature.[3] Also, because radiation of the feed is virtually confined to the geometrical illumination region of the horn,[4] there is a larger potential bandwidth available.

An additional requirement had to be explored: operation at about ¼ nominal frequency, for target-acquisition, by using both TE_{11} and TM_{01} mode excitation. This leads to the choice of a near-field design at the higher frequency because it would tend to function as a conventional far-field design at the lower frequency. At the lower frequency there would, however, be a shift of the phase center towards the horn aperture, resulting in a phase error in the subreflector illumination and a consequent reduction in efficiency, but perhaps it would be adequate for the intended function. Finally, dual-mode illumination using the TE_{11} and TM_{11} modes was of interest because of the nearly circular symmetric radiation patterns that can be obtained.[5]

III. CASSEGRAIN ANTENNA GEOMETRY

Figure 1 shows the geometry of a Cassegrain antenna. It consists of a conical feed horn, a hyperboloid subreflector and a paraboloid main reflector. One focal point of the hyperboloid coincides with the focal point of the paraboloid and the other focal point with the phase center of the horn. The main reflector illumination angle is equal to the geometrical subreflector illumination angle, θ_m. The feed is located in the geometrical shadow region of the subreflector.

The initial design of a Cassegrain antenna is usually based on geometrical optics, which imposes certain restrictions on the antenna geometry. The constraints are that the feed horn be located in the shadow region of the subreflector and that the subreflector intercept most of the power radiated by the horn.

To relate the radiation properties of the horn to the antenna geometry it is convenient to define a parameter K by:

$$K = \frac{d}{\lambda} \sin \delta \tag{1}$$

where

d = horn aperture diameter
λ = wavelength
δ = the angle subtended by the subreflector with respect to the center of the horn aperture (Fig. 1).

For conventional Cassegrain antennas representative values of K are from 1.2 to 1.6. For these values of K the major portion of the main lobe of a narrow angle horn excited by TE_{11} and TM_{11} modes, is intercepted by the subreflector. The lower value of K is preferable for TE_{11} mode excitation because the major lobe of the horn radiation pattern is narrower in the E plane than in the H plane. Beyond the major lobe region the phase variations are too large for efficient subreflector illumination.

For near field Cassegrains the values for K are much larger, such that the radiation characteristics of the horn are primarily determined by the horn flare angle.

For the feed horn to be located in the geometrical shadow region of the subreflector it is necessary that angle θ_b be not less than angle θ_{bh},

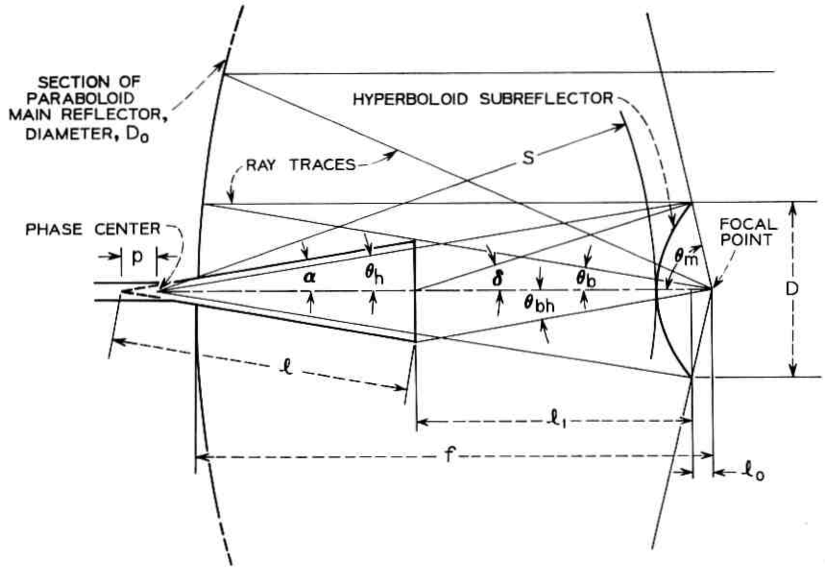


Fig. 1 — Geometry of cassegrain antenna.

that is,

$$\tan \theta_b = B \tan \theta_{bh} \tag{2}$$

where B is a constant with $B \geq 1$. Figure 1 shows the angles θ_b and θ_{bh}. The horn blocking angle θ_{bh} can be expressed in terms of the geometrical parameters and (1) by

$$\tan \theta_{bh} = \frac{K\lambda \sin \theta_m}{D \sin (\delta + \theta_m)}. \tag{3}$$

With all other parameters specified, θ_{bh} is minimum for

$$\delta + \theta_m = \frac{\pi}{2}. \tag{4}$$

Similarly, with θ_{bh} also specified, K has a maximum value when (4) holds.

Using (2) and (3) and expressing θ_b in terms of the geometrical parameters, the following equation is obtained for the subreflector diameter, D.

$$D = \sqrt{\frac{2BKf\lambda \sin \theta_m}{\sin (\theta_m + \delta) + \dfrac{BK\lambda}{16f} \sin \theta_m}} \tag{5}$$

where

$$f = \text{focal length of the paraboloid main reflector.}$$

Equation (5) agrees with the previously given condition for no blocking[6] $D \approx (2Kf\lambda)^{1/2}$. In practical antenna designs $BK\lambda/16f$ is small compared with $\sin(\delta + \theta_m)$, hence (5) may be rewritten in terms of the main reflector diameter, D_o, as:

$$D \geqq \cos\left(\frac{\theta_m}{2}\right)\sqrt{\frac{K\lambda\, D_o}{\sin\,(\delta + \theta_m)}} \tag{6}$$

Equation (6) shows that antennas with large main reflectors also require larger subreflectors, but that the ratio $(D/D_o)^2$ which is a measure of the amount of power blocked by the subreflector is inversely proportional to the main aperture diameter D_o. Hence the condition (6) is of importance primarily in the design of relatively small Cassegrain antennas.

Equation (6) also shows, as expected, that a conventional Cassegrain antenna requires a smaller subreflector than a near-field Cassegrain antenna, since K is smaller for the former. However, this disadvantage of the near-field Cassegrain antenna is offset by other advantageous properties.

Another parameter which influences the antenna design is the total Fresnel number of the horn at the subreflector distance, defined by

$$F_t = \frac{d^2}{4\lambda}\left(\frac{1}{l} + \frac{1}{l_1}\right). \tag{7}$$

For a conventional Cassegrain, F_t can be selected, to a certain extent, independently of the antenna geometry, because the radiation properties of the horn are not directly related to the horn length, l. For such an antenna with combined TE_{11} and TM_{11} mode excitation, a total Fresnel number in the 0.5-0.65 range would provide a nearly uniform subreflector illumination over a wide frequency range (about 30 per cent) with relatively small phase deviations. For TE_{11} mode excitation, a lower Fresnel number is necessary, because the phase of the E-plane horn radiation pattern is more frequency sensitive for larger Fresnel numbers.

For a near field Cassegrain antenna the total Fresnel number is almost directly related to the antenna geometry. Specifically, for an antenna with the horn located in the shadow region of the subreflector and with a subreflector illumination angle equal to the horn flare angle, F_t is given by

$$F_t = \frac{D}{2\lambda} \frac{\tan \theta_b}{1 - \dfrac{l_o}{l_1 + l_o}} \tag{8}$$

with

$$\tan \theta_b = \frac{2 \dfrac{D}{4f}}{1 - \left(\dfrac{D}{4f}\right)^2}. \tag{9}$$

Since l_o is much smaller than l_1 the total Fresnel number, F_t, is mainly determined by the subreflector diameter D.

For a near field Cassegrain it is necessary to have both K and F_t large. Equations (3) and (8) show that these quantities are proportional to D^2.

IV. ANTENNA DIMENSION

For the antenna under consideration the main reflector dimensions were specified. Its diameter, D_o is 224λ (λ = wavelength at the design frequency), its focal length, f, is 72.8λ and the corresponding geometrical illumination angle, θ_m, is 75°.

The initial choice of the other antenna dimensions was based on the following consideration. As shown above, K has a maximum for $\delta + \theta_m = \pi/2$. Using this condition the subreflector diameter, D, has been chosen such that the optimum value of K is unity at the lowest frequency ($\lambda_L = 4.5\lambda$). At the design frequency, K is about 3 times larger than is required for a conventional Cassegrain antenna feed. For this value of K, D is 25λ. With these parameters a horn with a maximum diameter of 17.6λ can be located in the shadow region of the subreflectors. The corresponding horn length, l, is 100λ. However, a feed horn with these dimensions would introduce, at the lowest frequency, appreciable phase variations at the subreflector owing to the shift of the phase center of the horn radiation pattern. For this reason these horn dimensions were not used in the computations.

The horn dimensions used were $d = 14\lambda$ and $l = 42.5\lambda$. With these horn dimensions K is only slightly less than the optimum value. The location of the phase center, which is 5λ in the front of the horn vertex, and the subreflector illumination angle, which is less than the horn flare angle, were chosen on the basis of the computed horn radiation patterns for combined TE_{11} and TM_{11} mode excitations.

Table I summarizes the antenna dimensions.

V. PROGRAM FOR COMPUTING ANTENNA CHARACTERISTICS

Programs have been developed which compute the antenna radiation characteristics and plot the computed radiation patterns. The computational methods are similar to those used in the computation of characteristics of the open Cassegrain antenna.[1] However, only single integrations are used; one integration was eliminated by using the Fresnel region approximation for wide angles and large Fresnel numbers.[2] Appendix A gives the equations used. Appendix B discusses operational aspects of the programs and gives flow diagrams.

The antenna characteristics for combined TE_{11}–TM_{11} and TM_{01} mode excitations are computed in one operation. The computer program consists of three parts which compute (*i*) the horn radiation patterns, (*ii*) the subreflector radiation patterns, and (*iii*) the far field radiation patterns.

The horn radiation patterns are computed at a constant radius, s, corresponding to the subreflector distance. From these computations the power loss at the subreflector is obtained by integration. The horn radiation patterns are also computed at the subreflector surface to obtain the subreflector illumination.

The subreflector radiation patterns are computed at a constant radius, f, and at the main reflector surface. From these computations the power loss at the main reflector and the main reflector illumination is obtained.

From the computed main reflector illumination the aperture gain, aperture efficiency and finally the far field radiation patterns are obtained.

The antenna gain and antenna efficiency are determined from the

TABLE I — ANTENNA DIMENSIONS

Main reflector diameter, D_o	224λ
Focal length, f	72.8λ
Main reflector illumination angle, θ_m	75°
Subreflector diameter, D	25λ
Subreflector illumination angle, θ_h	9.5°
Horn length, l	42.5λ
Horn flare angle, α	9.5°
Phase center location, p	5.0λ

computed aperture gain and efficiency respectively by including the loss at the subreflector and main reflector.

In the computations the phase variations at the sub- and main reflectors are included. Also included is the effect of the main reflector aperture blocking by the subreflector but not the effect of the protruding horn.

An estimate of the antenna noise temperature is obtained by assuming, somewhat arbitrarily, that near the horizon half the power lost at the sub- and main reflectors contributes to noise. At zenith it is assumed that the power lost at the main reflector contributes to noise. A ground temperature of 300°K is used in the computations. The additional noise from possible scattering of the subreflector support and the noise from the wide angle sidelobes of the far field radiation patterns are not included in the computations.

VI. COMPUTED ANTENNA CHARACTERISTICS

The antenna characteristics have been computed with the above computer program for the following feed horn excitations: (*i*) TE_{11} and TM_{11} at the design frequency, f_o, 0.8 f_o and 1.3 f_o, and (*ii*) TE_{11} and TM_{01} modes at f_o and 0.22 f_o. The antenna characteristics for the different modes and frequencies are summarized in Table II. The tabulated power losses are normalized with respect to the total power radiated by the feed horn.

6.1 *TE_{11} and TM_{11} Mode Excitations*

The computations were performed by assuming that the two modes are in phase at the horn aperture. The TM_{11} to TE_{11} power ratio was assumed to be 0.17. This value was used because at the design frequency it minimizes the phase variations of the horn radiation patterns at the subreflector both in the E and H planes.

Computations have been performed at the design frequency, f_o, 0.8 f_o, and 1.3 f_o. This corresponds to about a 50 per cent bandwidth. Except for the expected change in the antenna gain, the antenna radiation characteristics remain virtually the same across this frequency range. This indicates that if a frequency-insensitive conical feed-horn using TE_{11} and TM_{11} mode excitation could be developed, this antenna design is capable of efficient radiation over a 50 per cent bandwidth.

Figures 2 through 6 show the antenna radiation patterns at the design frequency, f_o. Included are: the amplitude and phase of the

Table II — Calculated Antenna Characteristics

Frequency		Design, f_0		$0.8\,f_0$	$1.3\,f_0$	f_0	$0.222\,f_0$	
Mode		TE_{11} and TM_{11}	TM_{01}	TE_{11} and TM_{11}		TE_{11}	TE_{11}	TM_{01}
Per cent of:								
Power loss at subreflector		2.7	20.0	3.4	2.1	9.3	26.0	50.8
Power loss at main reflector		0.7	2.9	0.9	0.5	1.0	4.4	5.6
Power blocked by subreflector		5.0	—	4.8	5.5	3.3	2.9	—
Aperture efficiency		72.7	—	73.4	71.5	75.4	82.2	—
Antenna efficiency		70.4	—	70.3	69.8	67.7	57.2	—
Antenna gain, dB		55.4	49.4 at max	53.4	57.6	55.2	41.4	33.3 at max
Antenna noise temperature, °K	Near horizon	5.1	34.35	6.45	3.9	15.45	45.6	81.6
	At zenith	2.1	8.7	2.7	1.5	3.0	13.2	16.8
First sidelobe, dB	H plane	−22.7	−12.9	−23.9	−21.4	−24.1	−17.2	−13.9
	E plane	−24.5		−23.8	−24.2	−15.1	−21.0	

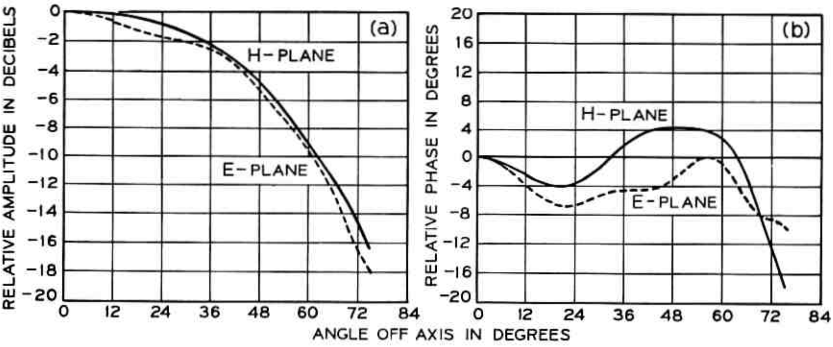


Fig. 2 — Subreflector illumination, TE_{11} and TM_{11} modes, freq. = f_0.

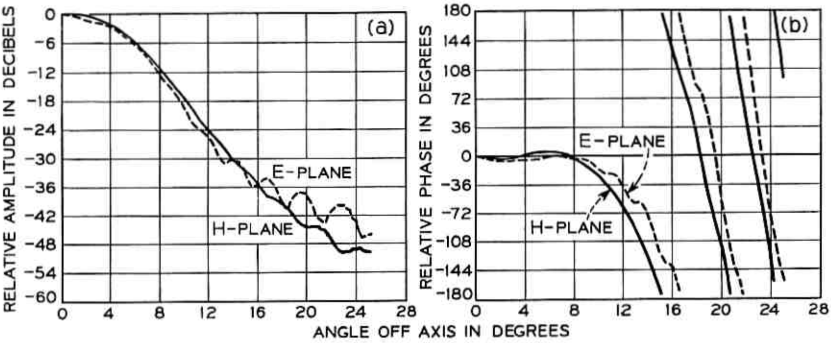


Fig. 3 — Horn radiation pattern at distance S (TE_{11} and TM_{11} modes, freq. $= f_o$)

subreflector illumination, the amplitude and phase of the horn radiation pattern at the subreflector distance, the amplitude and phase of the main reflector illumination, the amplitude and phase of the subreflector radiation pattern at the focal distance, and the far field pattern.

These figures show that the phase variations of the sub- and main reflector illuminations are relatively small at this frequency. This is because the location of the phase center and the ratio of the TM_{11} to TE_{11} modes has been chosen to minimize the phase variations at the subreflector. These figures also show that the far field radiation pattern is virtually the same in the E and H planes. This should result in a nearly circular symmetric far field radiation pattern.

Figures 7 and 8, and Figures 9 and 10 show some of the antenna

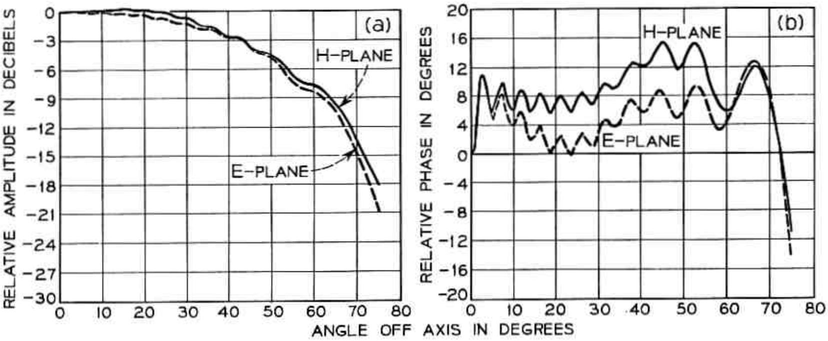


Fig. 4 — Main reflector illumination (TE_{11} and TM_{11} modes, freq. $= f_o$).

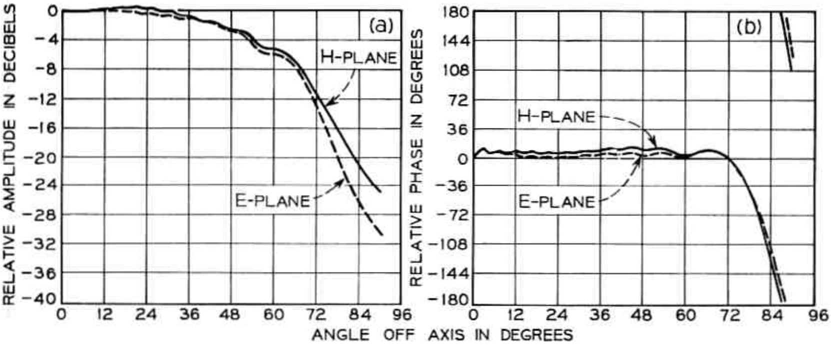


Fig. 5 — Subreflector radiation pattern at focal distance (TE_{11} and TM_{11} modes, freq. = f_o).

radiation characteristics at 0.8 f_o and at 1.3 f_o, respectively. The phase variations of the sub- and main reflector illuminations, though small, are larger than at the design frequency. This primarily results from the shift in the phase center of the horn radiation pattern. It is the shift in the phase center which ultimately limits the upper frequency of operation for this type of antenna.

6.2 *TM_{01} Mode Excitation*

The radiation characteristics for the TM_{01} mode excitation were computed at the design frequency, f_o. Figures 11 through 13 show representative radiation patterns for this mode. Particularly pronounced are the amplitude oscillations of the main and subreflector illuminations. This seems to be characteristic for the TM_{01} radiation

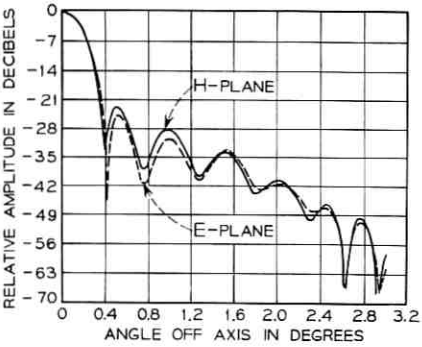


Fig. 6 — Far field radiation pattern (TE_{11} and TM_{11} modes, freq. = f_o).

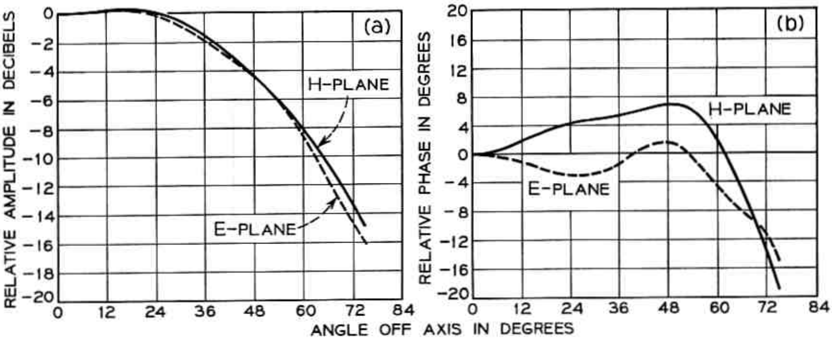


Fig. 7 — Subreflector illumination (TE_{11} and TM_{11} modes, freq. = 0.80 × f_o).

patterns from apertures and reflectors which are large compared with the wavelength and which are illuminated with nearly spherical wave fronts. However, no experimental evidence has been found to confirm these characteristics.

The advantage of this antenna for TM_{01} mode excitation compared with a conventional Cassegrain is less spillover at the subreflector, hence, less antenna noise. However, the sidelobe levels of the far field radiation pattern are perhaps a few dB higher than could be obtained with a conventional Cassegrain.

6.3 *TE_{11} Mode Excitation*

In view of the difficulties in realizing a conical feed horn with TE_{11} and TM_{11} mode excitation which would maintain the proper phase

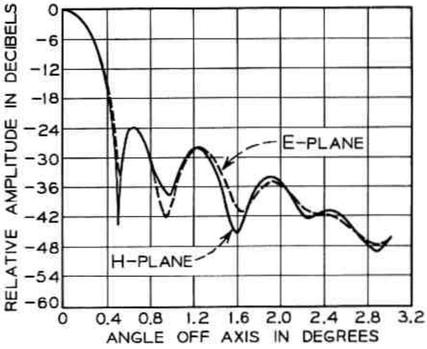


Fig. 8 — Far field radiation pattern (TE_{11} and TM_{11} modes, freq. = 0.80 × f_o).

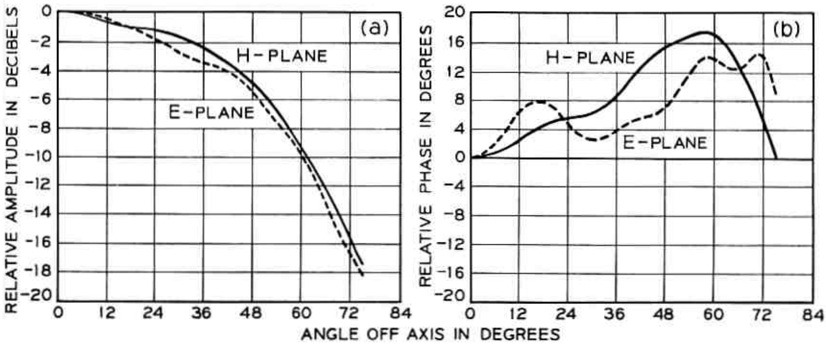


Fig. 9 — Subreflector illumination (TE_{11} and TM_{11} modes, freq. = 1.30 × f_o).

relationship at the horn aperture over a wide frequency range, TE_{11} mode excitation only has been investigated for the same antenna geometry.

Figures 14 and 15 show some of the radiation patterns for this mode at the design frequency, f_o. The computations show that the phase variations of the sub- and main reflector illuminations are considerably larger, particularly in the E plane, compared with those obtained by using combined TE_{11} and TM_{11} mode excitations. Also, the sidelobe levels of the far field radiation pattern in the E plane are considerably higher than in the H plane.

Table II shows that the computed antenna gain is 0.2 dB lower than the computed gain for TE_{11} and TM_{11} modes. However, the most significant difference is the increase in the antenna noise tem-

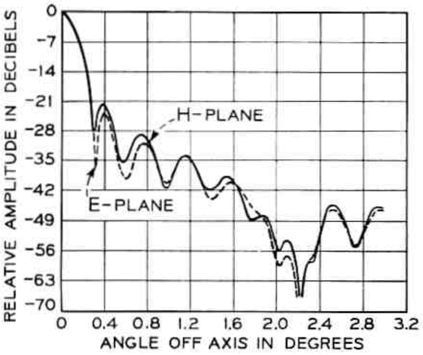


Fig. 10 — Far field radiation pattern (TE_{11} and TM_{11} modes, freq. = 1.30 × f_o).

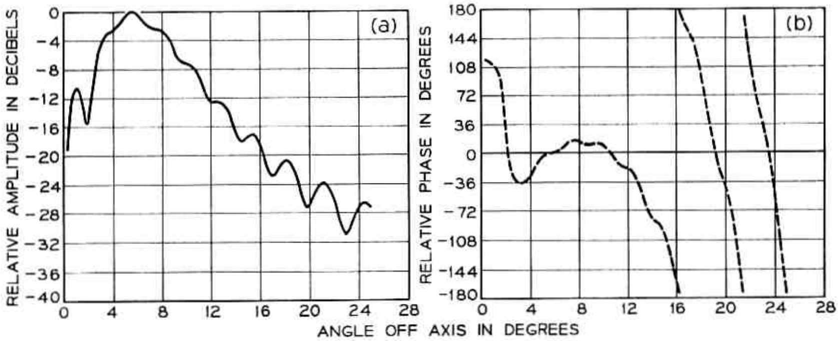


Fig. 11 — Horn radiation pattern at distance S (TM_{01} mode, freq. = f_o).

perature by 10°K near the horizon. This increase is primarily caused by the larger power loss at the subreflector because of the E plane horn radiation pattern characteristics.

A reduction of the antenna noise temperature by a few degrees might be possible by increasing the subreflector illumination angle perhaps even beyond the geometrical illumination angle of the horn. Figure 14 shows that the phase variations in the E-plane radiation pattern are not very large in the vicinity of the presently-used subreflector illumination angle of 9.5 degrees. The computed horn power radiation patterns show that if in the present design the illumination angle were 10.5 degrees the antenna noise temperature would be reduced by 4.8°K. The antenna gain for such a design would be reduced by only a small amount.

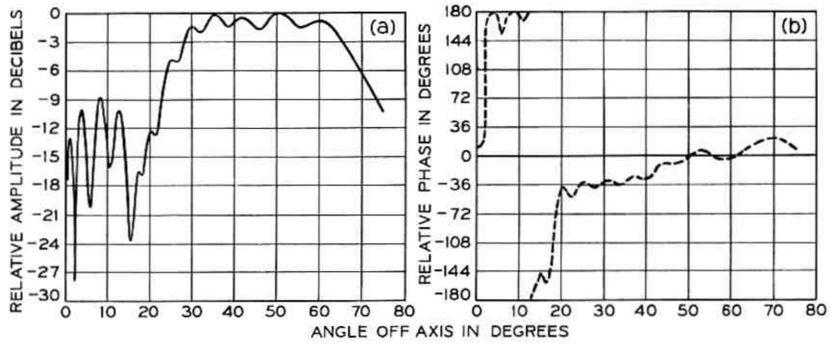


Fig. 12 — Main reflector illumination (TM_{01} mode, freq. = f_o).

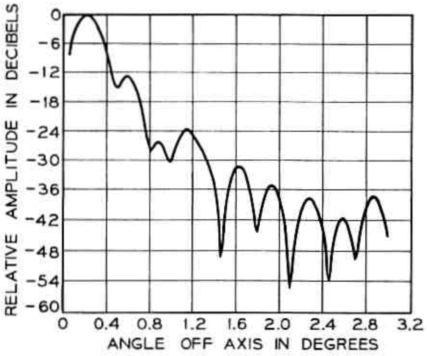


Fig. 13 — Far field radiation pattern (TM_{01} mode, freq. = f_o).

The bandwidth characteristics for this mode in the vicinity of the design frequency should be similar to those of the combined TE_{11} and TM_{11} modes.

6.4 *TE_{11} and TM_{01} Mode Excitation at* 0.22 f$_o$

The horn and far field radiation patterns for these modes are shown in Figs. 16 through 19. The right side of Table II summarizes the computed antenna performance at 0.22 f_o. The antenna efficiency for the TE_{11} mode is relatively high particularly in view of the large phase variations of the subreflector illumination. The far field radiation patterns for both the TE_{11} and TM_{01} modes show good characteristics. The primary disadvantages, however, are the high noise

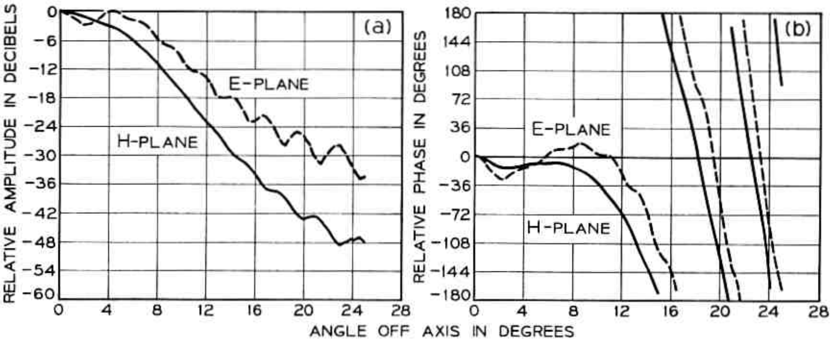


Fig. 14 — Horn radiation pattern at distance S (TE_{11} mode, freq. = f_o).

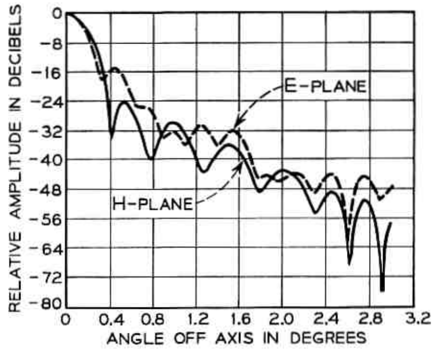


Fig. 15 — Far field radiation pattern (TE_{11} mode, freq. $= f_o$).

temperatures for both modes, owing to the power loss at the subreflector.

VII. SUMMARY AND CONCLUSIONS

Computer programs have been developed for computing the radiation characteristics of Cassegrain antennas and for plotting of the computed radiation patterns. The method is applicable to symmetrical Cassegrain antennas and provides the means of their design for nearly optimum performance.

A Cassegrain antenna with a near field conical feed horn has been investigated for different mode excitations and over a wide frequency range. For large antennas (over 200 wavelength main reflector diam-

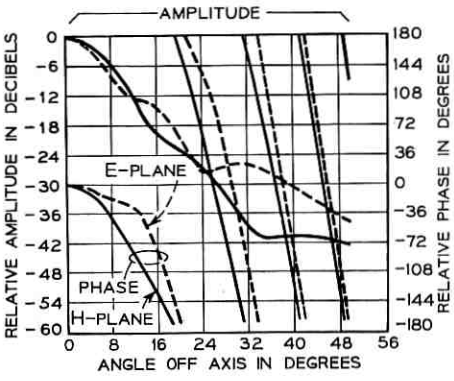


Fig. 16 — Horn radiation pattern at distance S (TE_{11} mode, freq. $= 0.22 \times f_o$).

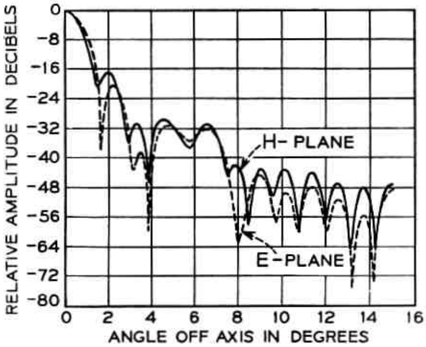


Fig. 17 — Far field radiation pattern (TE_{11} mode, freq. = 0.22 × f_o).

eter) this type of feed can be used over a 50 per cent bandwidth with only small variations in the over-all antenna characteristics, except for the predictable increase in the antenna gain with frequency.

The computed antenna characteristics for the combined TE_{11} and TM_{11} mode excitations show that the advantages of the combined excitation are: (*i*) lower far field E-plane sidelobes, (*ii*) 0.2 dB higher antenna gain, and (*iii*) 10°K lower antenna noise temperature.

At the design frequency for TE_{11} mode excitation, the computed antenna efficiency is 70 per cent and the noise temperature near horizon 15.5°K. With a design modification it should be possible to reduce the noise temperature by a few degrees without affecting the antenna gain.

For the TM_{01} mode, the calculated antenna gain and noise tem-

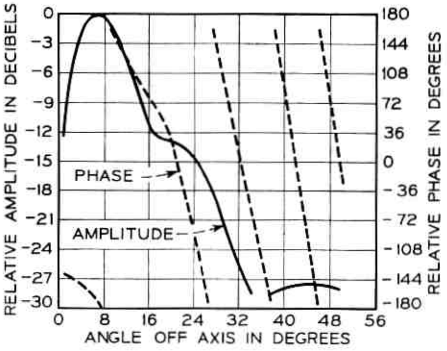


Fig. 18 — Horn radiation pattern at distance S (TM_{01} mode, freq. = 0.22 × f_o).

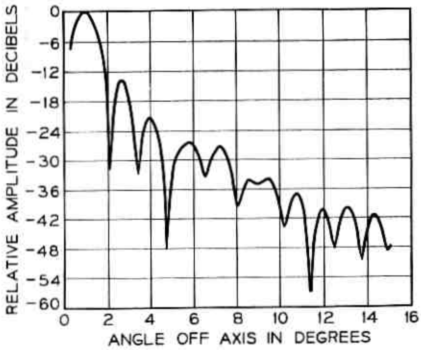


Fig. 19 — Far field radiation pattern (TE$_{11}$ mode, freq. = 0.22 × f_o).

perature at the design frequency are superior to those obtainable by using a conventional feed. The sidelobe levels of the far field radiation pattern are perhaps a few dB higher.

At a frequency 4.5 times below the design frequency the calculated antenna efficiencies for the TE$_{11}$ and TM$_{01}$ modes are relatively high. However, the power loss at the subreflector gives rise to appreciable noise temperatures near the horizon.

APPENDIX A

Formulations Used in Computing the Characteristics of Cassegrain Antennas

A.1 *Horn Radiation Patterns*

The horn radiation patterns have been computed by using the Kirchhoff approximation to the aperture radiation field. With this approximation the electric field $\mathbf{E}_p$ at distances of at least a few wavelengths from the aperture is:[7]

$$\mathbf{E}_p = \frac{jk}{4\pi} \iint_S [\mathbf{E}_a(1 + 1_n \cdot 1_R) - \mathbf{E}_a \cdot 1_R(1_n + 1_R)] \frac{e^{-jkR}}{R} \, ds \tag{10}$$

where

$$S = \text{horn aperture area}$$

$$k = \frac{2\pi}{\lambda}$$

$$\lambda = \text{wavelength}$$

and 1_n and 1_R are unit vectors in the normal, and in the R direction as shown in Fig. 20. $\mathbf{E}_a$ is the electric field in the horn aperture, assumed to be the same as for circular waveguide modes but with spherical wave fronts

Since it has been shown by actual computations[4] that the primary contributions to the horn radiation patterns are due to the first terms of (10), in the computations the term $\mathbf{E} \cdot 1_R(1_n + 1_R)$ has been neglected.

Because of the periodicity of the $e^{-ik}R/R$ with respect to the azimuth coordinates φ' and φ_1, it is sufficient to evaluate (10) in a discrete number of φ_1 planes, the number being equal to number of Fourier components of the aperture field in φ'. In particular it has been shown[1] that for TE_{11} and TM_{11} mode excitations it is sufficient to evaluate one rectilinear x or y component of (10) in the two principal planes $\varphi_1 = 0$ and $\varphi_1 = \pi/2$. Similarly for TM_{01} mode excitation only one rectilinear component of $\mathbf{E}_a$ in one plane needs to be evaluated.

The integrals which are evaluated for the TE_{11}, TM_{11}, and TM_{01} modes are:

$$E_{py} = \frac{kl^2}{4\pi} \int_0^\alpha \int_0^{2\pi} E_{ay} \frac{e^{-jkR}}{R} (1 + 1_n \cdot 1_R) \sin \theta' \, d\theta' \, d\varphi' \qquad (11)$$

with the aperture fields for the different modes given by:

TE_{11} mode

$$(E_{ay})_{\mathrm{TE}_{11}} = J_o\left(k_{\mathrm{TE}_{11}} \frac{\theta'}{\alpha}\right) - J_2\left(k_{\mathrm{TE}_{11}} \frac{\theta'}{\alpha}\right) \cos 2\varphi' \qquad (12)$$

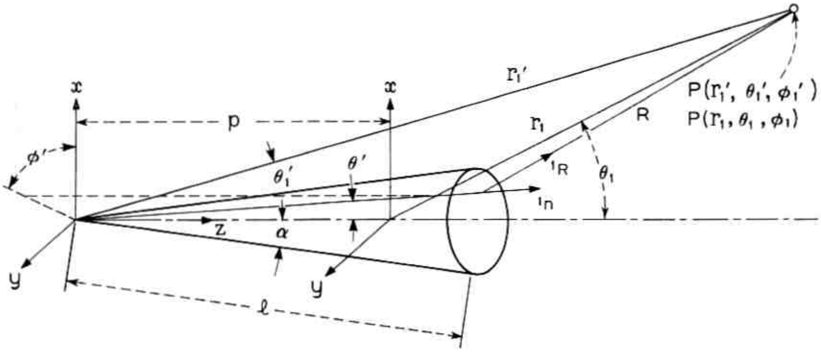


Fig. 20 — Coordinates for horn radiation pattern computation.

with

$$J_1'(k_{\mathrm{TE}_{11}}) = 0. \tag{13}$$

TM_{11} mode

$$(E_{ay})_{\mathrm{TM}_{11}} = J_o\left(k_{\mathrm{TM}_{11}} \frac{\theta'}{\alpha}\right) + J_2\left(k_{\mathrm{TM}_{11}} \frac{\theta'}{\alpha}\right) \cos 2\varphi' \tag{14}$$

with

$$J_1(k_{\mathrm{TM}_{11}}) = 0. \tag{15}$$

TM_{01} mode

$$(E_{ay})_{\mathrm{TM}_{01}} = J_1\left(k_{\mathrm{TM}_{01}} \frac{\theta'}{\alpha}\right) \sin \varphi' \tag{16}$$

with

$$J_o(k_{\mathrm{TM}_{01}}) = 0. \tag{17}$$

J_n = Bessel functions of order n.

α = horn flare angle.

The integration with respect to φ' has been eliminated by approximating the integrals, I_{nh}, given by:

$$I_{nh} = \int_0^{2\pi} \frac{e^{-jkR}}{R} (1 + 1_n \cdot 1_R) \cos n(\varphi_1 - \varphi')\, d(\varphi_1 - \varphi'). \tag{18}$$

The approximations used are modifications of the previously derived[2] approximations, I_n^1, to the integrals (18) with $1_n \cdot 1_R = 0$. The modifications consist of including the values of $1_n \cdot 1_R$ at the stationary phase points since it has been shown that the previously derived approximations, I_n^1, reduce to those obtained by the method of stationary phase, and that I_n^1 can be separated into terms which correspond to the stationary phase terms. It is subsequently shown that R can be expressed as:

$$R = r\sqrt{1 - \frac{2u}{r} \cos(\varphi_1 - \varphi')} \tag{19}$$

where r and u are functions which are independent of φ and φ', on this basis, the first order approximations to (18), I_{nh}^1 are:

$$I_{nh}^{1} = \pi j^{n}\left\{\left[\frac{e^{-jk(R_o+u)}}{R_o}(1 + 1_n\cdot 1_{R_o}) + \frac{e^{-jk(R_1-u)}}{R_1}(1 + 1_n\cdot 1_{R_1})\right]J_n(ku)\right.$$
$$\left. - j\left[\frac{e^{-jk(R_o+u)}}{R_o}(1 + 1_n\cdot 1_{R_o}) - \frac{e^{-jk(R_1-u)}}{R_1}(1 + 1_n\cdot 1_{R_1})\right]J_n'(ku)\right\} \quad (20)$$

with

$$R_o = r\left(1 - \frac{2u}{r}\right)^{\frac{1}{2}} \quad (21)$$

$$R_1 = r\left(1 + \frac{2u}{r}\right)^{\frac{1}{2}} \quad (22)$$

and 1_{R_o} and 1_{R_1} are the unit vectors at $(\varphi_1 - \varphi')$ equal to zero and π respectively.

By using the approximation (20), one integration is eliminated in (11) and the radiation patterns for the different modes are computed from the following integrals:

TE_{11} Mode

$$(E_{py})_{\mathrm{TE}_{11}} = \frac{jkl^2}{4\pi}\int_0^{\alpha}\left[J_o\left(k_{\mathrm{TE}_{11}}\frac{\theta'}{\alpha}\right)I_{oh}^{1} \pm J_2\left(k_{\mathrm{TE}_{11}}\frac{\theta'}{\alpha}\right)I_{2h}^{1}\right]\sin\theta'\,d\theta' \quad (23)$$

where the minus signs give the radiation pattern in the plane $\varphi_1 = 0$ and the plus sign the radiation pattern in the plane $\varphi_1 = \pi/2$.

TM_{11} mode

The integral is analogous to (23).

TM_{01} mode

$$(E_{py})_{\mathrm{TM}_{01}} = \frac{jkl^2}{4\pi}\int_0^{\alpha} J_1\left(k_{\mathrm{TM}_{01}}\frac{\theta'}{\alpha}\right)I_{1h}^{1}\sin\theta'\,d\theta'. \quad (24)$$

Referring to Fig. 20,

$$R = (l^2 + r_1^2 + p^2 + 2r_1p\cos\theta_1 - 2lp\cos\theta' - 2r_1l\cos\gamma_1)^{\frac{1}{2}} \quad (25)$$

with

$$\cos\gamma_1 = \sin\theta_1\sin\theta'\cos(\varphi' - \varphi_1) + \cos\theta_1\cos\theta' \quad (26)$$

and

$$1_n\cdot 1_R = \frac{p\cos\theta' - l + r_1\cos\gamma_1}{R}. \quad (27)$$

Hence a comparison of (25) with (19) shows that

$$r = (l^2 + r_1^2 + p^2 + 2r_1p \cos \theta_1 - 2lp \cos \theta' - 2r_1l \cos \theta_1 \cos \theta')^{\frac{1}{2}} \quad (28)$$

and

$$u = \frac{r_1 l \sin \theta_1 \sin \theta'}{r}. \quad (29)$$

The integrals for the difference modes have been computed at two values of r_1: (*i*) at the subreflector surface to obtain the subreflector illumination, and (*ii*) at a constant distance corresponding to shortest distance, s, from the subreflector to the horn aperture. The latter was performed to obtain the horn radiation pattern in a form which is readily measurable and convenient for subsequent computation of the horn power radiation pattern used for determining the power loss at the subreflector.

For the first computation referring to Fig. 21

$$r_1 = \frac{c}{2} \frac{(1 - \beta^2)}{\cos \theta_1 - \beta} \quad (30)$$

with

$$\beta = \frac{b}{c}.$$

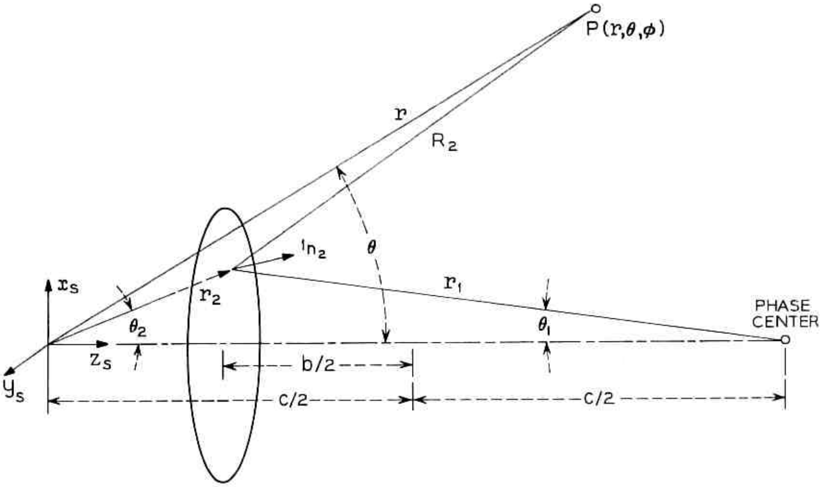


Fig. 21 — Subreflector coordinates.

For determining the subreflector illumination it is preferable to obtain the illumination in terms of subreflector coordinate θ_2. The relationship between the coordinates θ_1 and θ_2 is:

$$\cos \theta_1 = \frac{(1 + \beta^2) \cos \theta_2 + 2\beta}{1 + \beta^2 + 2\beta \cos \theta_2}. \tag{31}$$

The second computation

$$r_1 = \frac{c}{2}(1 + \beta) \tag{32}$$

and the integration is performed as a function of θ_1.

A comparison has been made between some radiation patterns computed by single and double integration. Good agreement was obtained.

The electric field in the spherical θ_1 and φ_1 coordinates can be expressed in terms of the radiation patterns in the principal planes $\varphi_1 = 0, \pi/2$.

For TE_{11} and TM_{11} modes

$$\mathbf{E}_p = 1_{\theta_1} E_{pv}\left(\frac{\pi}{2}\right) \sin \varphi_1 + 1_{\varphi_1} E_{pv}(0) \cos \varphi_1 \tag{33}$$

The TM_{01} mode has only a θ_1 component given by (24).

A.2 *Horn Power Radiation Patterns*

The horn power radiation pattern, P_H, is computed from the following integral:

$$P_H = \frac{r_1^2}{2\eta} \int_0^{2\pi} \int_0^{\theta_1} \mathbf{E}_p \cdot \mathbf{E}_p^* \sin \theta \, d\theta \, d\varphi. \tag{34}$$

η = free space intrinsic impedance.

The total power is obtained by extending the range of θ_1 to the region where $\mathbf{E}_p \cdot \mathbf{E}_p^*$ has a negligible value.

The total radiated power can also be obtained from the assumed fields at the aperture (12) through (17). On this basis the total power for the different modes is approximately

TE_{11} mode

$$P_{H_t} = \frac{2\pi(l\alpha)^2}{2\eta} [k^2_{TE_{11}} - 1]\left[\frac{J_1(k_{TE_{11}})}{k_{TE_{11}}}\right]^2. \tag{35}$$

TM_{11} mode

$$P_{H_t} = \frac{2\pi(l\alpha)^2}{2\eta} J_o^2(k_{TM_{11}}). \tag{36}$$

TM_{01} mode

$$P_{H_t} = \frac{\pi(l\alpha)^2}{2\eta} J_1^2(k_{\mathrm{TM}_{01}})\cdot \tag{37}$$

The computations performed for the different modes and frequencies by using (34) are in agreement with (35) through (37) within 1.8 per cent, with the power computed by using (34) giving larger values for all modes. This is at least partially caused by the higher values that the approximations $I_n{}^1$ give compared with those obtained by precise numerical integration.[2]

A.3 *Subreflector Radiation Patterns*

The subreflector radiation patterns have been computed by using the surface integral relating the radiated fields and the current distribution over a surface.[7] For the distance of at least a few wavelengths from the reflector the radiated electric field with reference to Fig. 21 is:

$$E_s = \frac{j}{\lambda} \iint_{S_s} \mathbf{1}_{R_2} \times (\mathbf{J} \times \mathbf{1}_{R_2}) \frac{e^{-jk(R_2+r_2)}}{R_2}\, ds_2 \tag{38}$$

where

$\mathbf{J}$ = surface current density
S_s = subreflector area.

To evaluate (38) it has been assumed that the reflector is locally plane. With this assumption the current density is directly related to the incident electric field. To simplify the computations $\mathbf{1}_{R_s}$ was replaced by $\mathbf{1}_{r_2}$. (A test computation of the subreflector radiation pattern of the open Cassegrain antenna showed that using $\mathbf{1}_{r_2}$ instead of $\mathbf{1}_{R_2}$ results in a negligible difference.)

For a hyperboloid reflector, the relations between the incident electric field and the current density have been derived.[1] By using the approximation (20) the integration with respect to φ_2 can be eliminated. On this basis the radiated fields of the subreflector resulting from the incident fields of a TE_{11} mode (23), are given by:

In the plane $\varphi_2 = 0$,

$$\begin{aligned}[E_s(0)]_{\mathrm{TE}_{11}} = -\frac{j}{2\lambda} \int_0^{\theta_m} e^{-jkr_2} \Bigg\{ & \left[E_{py}(0) + E_{py}\left(\frac{\pi}{2}\right) \frac{(1+\beta\cos\theta_2)}{(\beta+\cos\theta_2)} \right] I_o^1 \\ & + \left[E_{py}(0) - E_{py}\left(\frac{\pi}{2}\right) \frac{(1+\beta\cos\theta_2)}{\beta+\cos\theta_2} \right] I_2^1 \Bigg\} r_2^2 \sin\theta_2 \, d\theta_2 \,.\end{aligned} \tag{39}$$

In the plane $\varphi_2 = \pi/2$,

$$\left[E_s\left(\frac{\pi}{2}\right)\right]_{\mathrm{TE}_{11}}$$

$$= -\frac{j}{2\lambda}\int_0^{\theta_m} e^{-jkr_2}\left\{\left[E_{py}(0) + E_{py}\left(\frac{\pi}{2}\right)\frac{(1+\beta\cos\theta_2)}{\beta+\cos\theta_2}\right]I_o^1\cos\theta\right.$$

$$-\left[E_{py}(0) - E_{py}\left(\frac{\pi}{2}\right)\frac{(1+\beta\cos\theta_2)}{\beta+\cos\theta_2}\right]I_2^1\cos\theta$$

$$\left.+\,2E_{py}\left(\frac{\pi}{2}\right)\frac{\sin\theta_2}{\beta+\cos\theta_2}\,I_1^1\sin\theta\right\}r_2^2\sin\theta_2\,d\theta_2\,. \tag{40}$$

For the TM_{11} mode, the fields are analogous to (39) and (40) with the corresponding TM_{11} illuminated functions $E_p(0)$ and $E_p(\pi/2)$.

For the TM_{01} mode,

$$[E_s(0)]_{\mathrm{TM}_{01}} = -\frac{j}{2\lambda}\int_0^{\theta_m} E_{py}(0)\,e^{-jkr_2}$$

$$\cdot\left[\cos\theta\,\frac{(1+\beta\cos\theta_2)}{\beta+\cos\theta_2}\,I_1^1 + \frac{\beta\sin\theta\sin\theta_2}{\beta+\cos\theta_2}\,I_o^1\right]r_2^2\sin\theta_2\,d\theta_2 \tag{41}$$

where $E_p(0)$ is given by (24).

In the above r_2 is the equation of the subreflector in the coordinates shown in Fig. 21, and is given by

$$r_2 = \frac{c}{2}\,\frac{(1-\beta^2)}{\beta+\cos\theta_2}. \tag{42}$$

I_n^1 are the first order approximations to the integration with respect to φ_2. They are given by (20) with $1_n \cdot 1_{R_o}$ and $1_n \cdot 1_{R_1}$ set equal to zero. The other parameters are:

$$R_2 = (r^2 + r_2^2 - 2rr_2\cos\gamma_2)^{\frac{1}{2}}. \tag{43}$$

$$\cos\gamma_2 = \mathit{sin}\,\theta\sin\theta_2\cos(\varphi - \varphi_2) + \cos\theta\cos\theta_2\,. \tag{44}$$

$$u_2 = \frac{rr_2\sin\theta\sin\theta_2}{(r^2 + r_2^2 - 2rr_2\cos\theta\cos\theta_2)^{\frac{1}{2}}}. \tag{45}$$

The subreflector radiation patterns have been computed at two distances: at the paraboloid surface

$$r = \frac{2f}{1+\cos\theta} \tag{46}$$

where f = focal length of the main reflector, to obtain the main reflector illumination, and at

$$r = f \tag{47}$$

to determine the power loss at the main reflector.

The φ dependence of the subreflector radiation patterns are obtained in terms of radiation patterns in the principal planes and are:

For TE_{11} or TM_{11} modes

$$[\mathbf{E}_s]_{\mathrm{TE}_{11},\mathrm{TM}_{11}} = 1_\varphi E_s(0) \cos \varphi + 1_\theta E_s\left(\frac{\pi}{2}\right) \sin \varphi. \tag{48}$$

For the TM_{01} mode

$$[\mathbf{E}_s]_{\mathrm{TM}_{01}} = 1_\theta E_s(0). \tag{49}$$

The power radiation pattern of the subreflector is computed by using the integral (34).

A.4 *Aperture Gain and Efficiency*

The aperture gain and efficiency are computed by projecting the incident field on the main reflector aperture. The fields in the rectilinear x, y components are related to θ, φ components by the expressions

$$\begin{aligned} E_{rs} = -1_x(E_{s\theta} \cos \varphi - E_{s\varphi} \cos \varphi) \\ - 1_y(E_{s\theta} \sin \varphi_2 + E_{s\varphi} \cos \varphi_2). \end{aligned} \tag{50}$$

For TE_{11} and TM_{11} mode excitations polarized in the y direction, the gain on axis, G_M, is

$$G_M = \frac{4\pi^2}{\lambda^2} \frac{\left| \int_{\theta_b}^{\theta_m} \left[E_s\left(\frac{\pi}{2}\right) + E_s(0) \right] r^2 \sin \theta \, d\theta \right|^2}{\int_0^{\theta_m} \left[\left| E_s\left(\frac{\pi}{2}\right) \right|^2 + | E_s(0) |^2 \right] r^2 \sin \theta \, d\theta} \tag{51}$$

where r is the equation for the paraboloid (46).

The aperture efficiency, g, is obtained from the relation

$$g = \frac{G}{G_o} \tag{52}$$

with

$$G_o = \left[\frac{4\pi f \sin \theta_m}{\lambda(1 + \cos \theta_m)} \right]^2. \tag{53}$$

The maximum antenna gain for the TM_{01} mode is determined by normalizing the amplitude of the TM_{01} mode electric field at the horn aperture with respect to the amplitudes of the TE_{11} or TE_{11} and TM_{11} modes for the same power input, by using (35 through 37). The gain for the TM_{01} mode is then related to the gain for the TE_{11} mode on axis, referred to the maximum of its pattern.

A.5 *Far Field Radiation Patterns*

The far field patterns are computed from the projected field on the aperture, using the relation

$$\mathbf{E}_f = \frac{j}{\lambda}\frac{e^{-jkr^a}}{r_a}$$

$$\cdot \int_{\theta_b}^{\theta_m}\int_0^{2\pi} \mathbf{E}_{rs} \exp\,[jkr\sin\theta\sin\theta_a\cos(\varphi-\varphi_a)]r^2\sin\theta\,d\theta\,d\varphi \qquad (54)$$

where r_a, θ_a, and φ_a are the coordinates of the far field observation point.

Because of the antenna symmetry the integration with respect to φ can be the readily performed. The resulting integration with respect to θ is for TE_{11} or TM_{11} modes

$$(E_{fy})_{\mathrm{TE}_{11}} = -\frac{j\pi}{\lambda}\frac{e^{-jkr_a}}{r_a}\int_{\theta_b}^{\theta_m}\left\{\left[E_s(0)+E_s\left(\frac{\pi}{2}\right)\right]J_o\left(\frac{2\pi r}{\lambda}\sin\theta\sin\theta_a\right)\right.$$

$$\left.\pm\left[E_s\left(\frac{\pi}{2}\right)-E_s(0)\right]J_2\left(\frac{2\pi r}{\lambda}\sin\theta\sin\theta_a\right)\right\}r^2\sin\theta\,d\theta \qquad (55)$$

where ± signs correspond to the patterns in the $H(\varphi_a=0)$ or $E(\varphi_a=\pi/2)$ planes.

Similarly for the TM_{01} mode

$$(\mathbf{E}_f)_{\mathrm{TM}_{01}} = 1_{\theta_a}\frac{2\pi}{\lambda}\frac{e^{-jkr_a}}{r_a}\int_{\theta_b}^{\theta_m} E_s(0)J_1\left(\frac{2\pi r}{\lambda}\sin\theta\sin\theta_a\right)r^2\sin\theta\,d\theta \qquad (56)$$

where E_s is the main reflector illumination (48) and (49) for the different modes in the planes $\varphi = 0$ and $\varphi = \pi/2$.

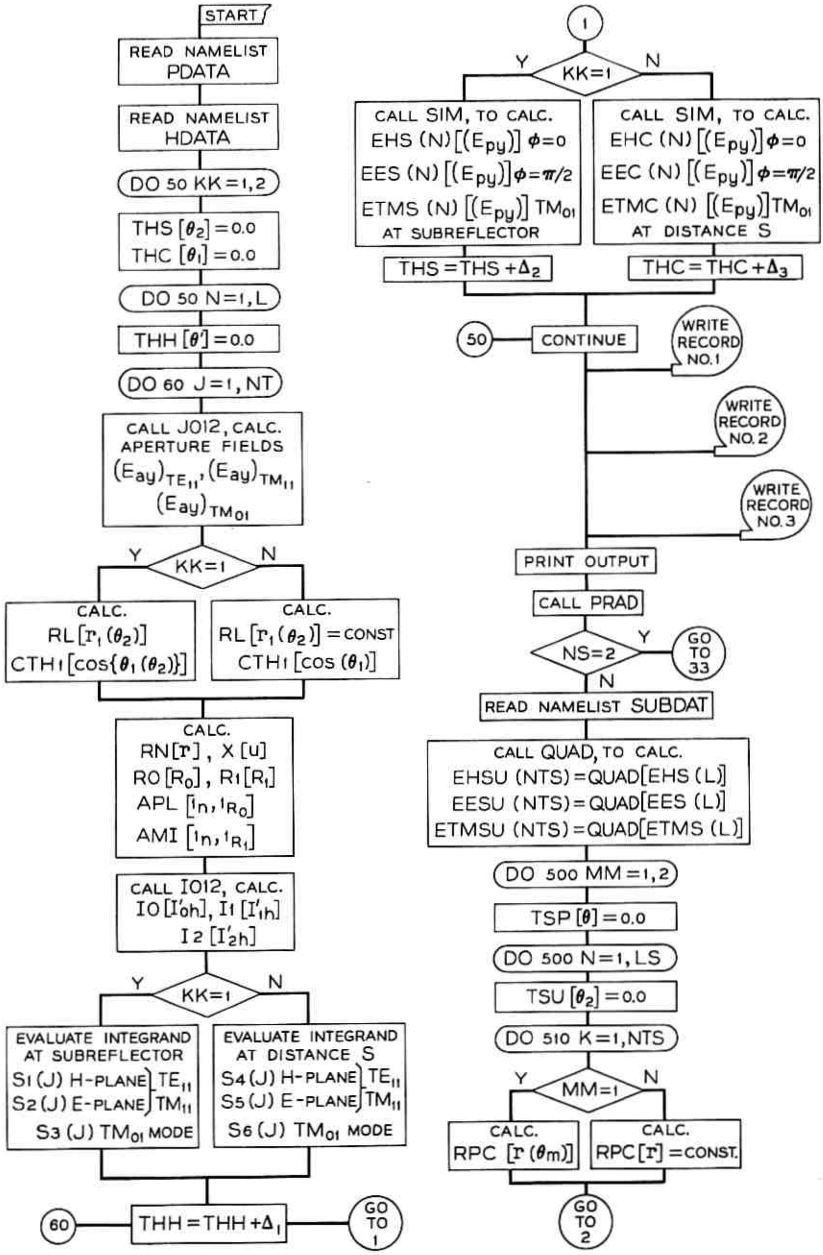
APPENDIX B

Program for Computing the Characteristics of Cassegrain Antennas and for Graphic Display of Radiation Fields

The package consists of two programs: a program to compute the horn radiation patterns, the subreflector radiation patterns, the far field radiation patterns, and other characteristics of Cassegrain antennas described in Appendix A, and a program to scale, label, and plot the radiation fields. The two programs are linked by intermediate storage of computed results and control variables on tape, at the conclusion of Part 1 execution.

B.1 *Computation Program*

Figure 22 is a logic diagram of the program. The following convention is used throughout the logic diagram. Square-bracketed symbols


START
READ NAMELIST PDATA
READ NAMELIST HDATA
DO 50 KK=1,2
THS [θ_2]=0.0
THC [θ_1]=0.0
DO 50 N=1,L
THH [θ']=0.0
DO 60 J=1,NT
CALL J012, CALC. APERTURE FIELDS $(E_{ay})_{TE_{11}}$, $(E_{ay})_{TM_{11}}$ $(E_{ay})_{TM_{01}}$
Y
KK=1
N
CALC. RL [$r_1(\theta_2)$] CTH1 [cos{$\theta_1(\theta_2)$}]
CALC. RL [$r_1(\theta_2)$] = CONST CTH1 [cos (θ_1)]
CALC. RN [r], X [u] R0 [R_0], R1 [R_1] APL [^1n, $^1R_0$] AMI [^1n, $^1R_1$]
CALL I012, CALC. I0 [I'_{0h}], I1 [I'_{1h}] I2 [I'_{2h}]
Y
KK=1
N
EVALUATE INTEGRAND AT SUBREFLECTOR S1 (J) H-PLANE} TE_{11} S2 (J) E-PLANE} TM_{11} S3 (J) TM_{01} MODE
EVALUATE INTEGRAND AT DISTANCE S S4 (J) H-PLANE} TE_{11} S5 (J) E-PLANE} TM_{11} S6 (J) TM_{01} MODE
60
THH = THH + Δ_1
GO TO 1
1
Y
KK=1
N
CALL SIM, TO CALC. EHS (N) [(E_{py})] ϕ=0 EES (N) [(E_{py})] $\phi=\pi/2$ ETMS (N) [(E_{py})] TM_{01} AT SUBREFLECTOR
CALL SIM, TO CALC. EHC (N) [(E_{py})] ϕ=0 EEC (N) [(E_{py})] $\phi=\pi/2$ ETMC (N) [(E_{py})] TM_{01} AT DISTANCE S
THS = THS + Δ_2
THC = THC + Δ_3
50
CONTINUE
WRITE RECORD NO. 1
WRITE RECORD NO. 2
WRITE RECORD NO. 3
PRINT OUTPUT
CALL PRAD
NS=2
Y
GO TO 33
N
READ NAMELIST SUBDAT
CALL QUAD, TO CALC. EHSU (NTS) = QUAD [EHS (L)] EESU (NTS) = QUAD [EES (L)] ETMSU (NTS) = QUAD [ETMS (L)]
DO 500 MM=1,2
TSP [θ]=0.0
DO 500 N=1,LS
TSU [θ_2]=0.0
DO 510 K=1,NTS
Y
MM=1
N
CALC. RPC [r (θ_m)]
CALC. RPC [r] = CONST.
GO TO 2

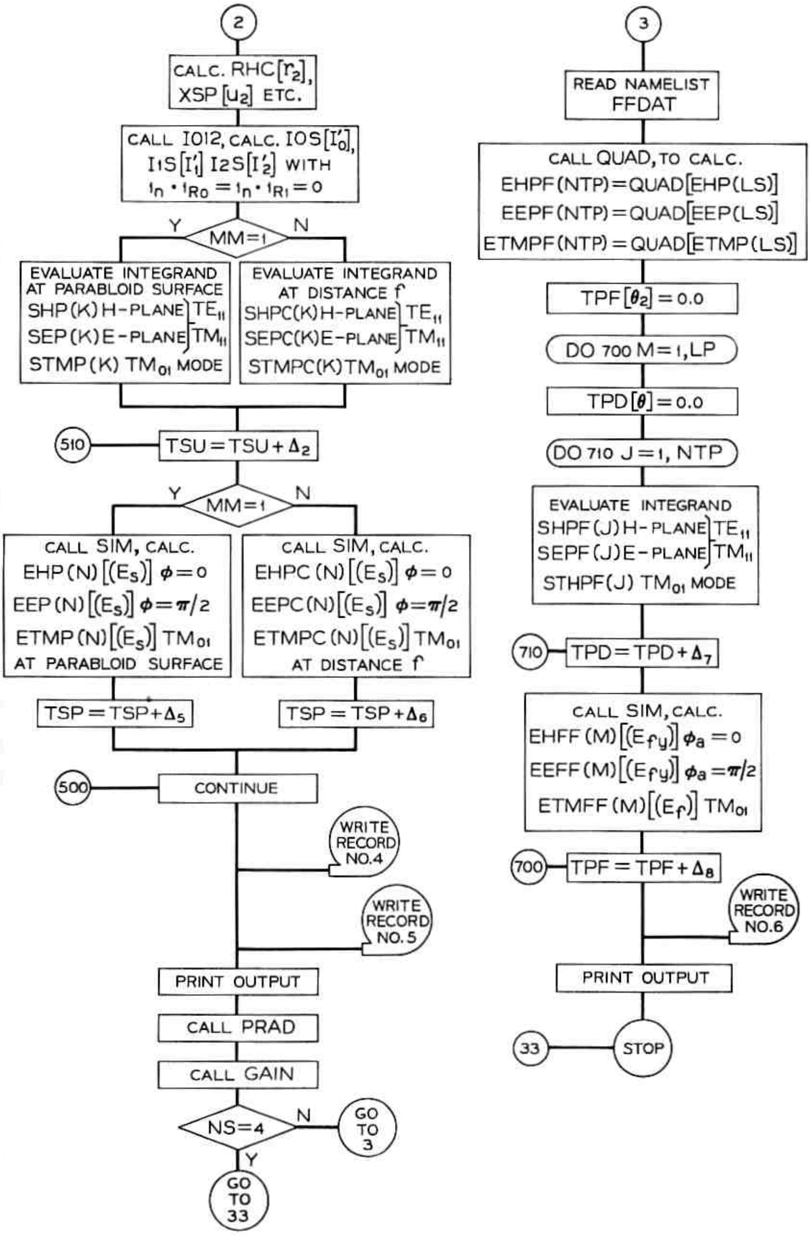

2
CALC. RHC[r_2], XSP [u_2] ETC.
CALL IO12, CALC. IOS[I_0'], I1S[I_1'] I2S[I_2'] WITH $i_n \cdot i_{R0} = i_n \cdot i_{R1} = 0$
Y
MM=1
N
EVALUATE INTEGRAND AT PARABLOID SURFACE
SHP(K) H-PLANE } TE_{11}
SEP(K) E-PLANE } TM_{11}
STMP(K) TM_{01} MODE
EVALUATE INTEGRAND AT DISTANCE f
SHPC(K) H-PLANE } TE_{11}
SEPC(K) E-PLANE } TM_{11}
STMPC(K) TM_{01} MODE
510
TSU = TSU + Δ_2
Y
MM=1
N
CALL SIM, CALC.
EHP (N) [(E_s)] $\phi = 0$
EEP (N) [(E_s)] $\phi = \pi/2$
ETMP (N) [(E_s)] TM_{01}
AT PARABLOID SURFACE
CALL SIM, CALC.
EHPC (N) [(E_s)] $\phi = 0$
EEPC (N) [(E_s)] $\phi = \pi/2$
ETMPC (N) [(E_s)] TM_{01}
AT DISTANCE f
TSP = TSP + Δ_5
TSP = TSP + Δ_6
500
CONTINUE
WRITE RECORD NO.4
WRITE RECORD NO. 5
PRINT OUTPUT
CALL PRAD
CALL GAIN
NS=4
N
GO TO 3
Y
GO TO 33
3
READ NAMELIST FFDAT
CALL QUAD, TO CALC.
EHPF (NTP) = QUAD[EHP(LS)]
EEPF (NTP) = QUAD[EEP(LS)]
ETMPF (NTP) = QUAD[ETMP(LS)]
TPF [θ_2] = 0.0
DO 700 M = 1, LP
TPD [θ] = 0.0
DO 710 J = 1, NTP
EVALUATE INTEGRAND
SHPF (J) H-PLANE } TE_{11}
SEPF (J) E-PLANE } TM_{11}
STHPF (J) TM_{01} MODE
710
TPD = TPD + Δ_7
CALL SIM, CALC.
EHFF (M) [(E_{fy})] $\phi_a = 0$
EEFF (M) [(E_{fy})] $\phi_a = \pi/2$
ETMFF (M) [(E_f)] TM_{01}
700
TPF = TPF + Δ_8
WRITE RECORD NO.6
PRINT OUTPUT
33
STOP

Fig. 22 — Logic diagram for field calculation program.

follow the notation of Appendix A, while the preceding alphanumeric symbols are the Fortran IV source program names.

Four groups of input data are required, designated by the NAMELIST names PDATA, HDATA, SUBDAT, and FFDAT. Although some of the following data is redundant, the formats are designed for convenience and precaution.

PDATA

NS — Number of sets of field patterns to be computed, that is,

NS = 2 Horn radiation patterns only
NS = 4 Horn and subreflector patterns
NS = 5 The above plus far field patterns

NPS — Number of patterns per set, that is,

NPS = 2 E & H planes only
NPS = 3 The above plus TM_{01} mode

ITE — Control bit for plotting program (see Section 2)

HDATA

HL — Horn length
HLAMD — Horn length normalized with respect to design wavelength
FQ — Frequency normalized with respect to design frequency
C — Distance between foci of hyperboloid
BETA — Defined by equation (30)
PL — Location of phase center normalized with respect to horn length
ALPHA — Horn flare angle
L — Number of points at which horn radiation pattern will be evaluated
NT — Number of points at which integrand will be calculated for evaluation of integrals in equations (23) and (24).
DEG — Angular increment (in degrees) for obtaining subreflector illumination
ANG — Angular increment (in degrees) for obtaining horn radiation patterns
TM_{11} — Complex constant which determines TM_{11} mode to TE_{11} mode ratio
I1P — Control bit for plotting program (See Section B.2.)

SUBDAT

F — Focal length of paraboloid
Note: The dimensional unit for F, C, HL must be the same.

FLAMD — Focal length normalized with respect to design wavelength

CLAMD — C normalized with respect to design wavelength

LS — Number of points at which subreflector radiation pattern will be evaluated

NTS — Number of points at which integrand will be calculated for evaluation of integrals in equations (39), (40), and (41)

GAMA — Main reflector illumination angle (degrees)

DEGP — Angular increment (in degrees) for obtaining main reflector illumination

ANGP — Angular increment (in degrees) for obtaining subreflector radiation patterns

INCS — $(N + 1)$ where N is the number of points to be interpolated between previously computed subreflector illumination points

LB — Number of points which are not included in integral (51) because of subreflector blocking

I2P — Control bit for plotting program (See Section B.2.)

FFDAT

NTP — Number of points at which integrand will be calculated for evaluation of integrals in equations (55) and (56)

GAMAB — Angular portion (in degrees) of main reflector blocked by subreflector

NTPB — Number of points which are not included in integrals in equations (55) or (56) owing to subreflector blocking after interpolation of main reflector illumination

LP — Number of points at which far field radiation patterns will be evaluated

INCP — $(N + 1)$ where N is the number of points to be interpolated between previously computed main reflector illumination points

DEGFF — Angular increment (in degrees) at which far field radiation pattern will be evaluated

I3P — Control bit for plotting program (See Section B.2.)

The following subprograms must be included in the deck before execution.

PRAD — Computes the power radiation patterns in accordance with equation (34)

JO12 — A Bessel function subroutine developed by J. Alan Cochran and P. A. Alsberg. The present version includes two subsidiary subroutines:[8]

DPHASE — Uses phase-amplitude method for large values of argument

JLOW — Uses downward recursion technique for small values of argument

IO12 — Calculates the first order approximations I^1_{nh} (20) to the integrals I_{nh} (18)

QUAD — A quadratic interpolation scheme for complex arrays

INTERP — A subroutine called by QUAD

PROC — A subroutine to format and print output data

SIM — A complex Simpson's rule integration routine. Will accept an even or odd length array with negligible variation in accuracy

GAIN — Computes the antenna gain and aperture efficiency as defined by equations (51), (52), and (53)

TR — A special purpose of Simpson's rule integration to evaluate the integrals in equation (51). This function subprogram is called only by the GAIN subroutine

The program requires approximately $(52{,}660)_8$ or $(22{,}000)_{10}$ words of storage. A representative execution time for both modes (TE_{11} and TM_{11} combined, and TM_{01}) at the design frequency is 14 minutes; the same calculations at 0.22 times the design frequency, where a smaller number of integration points is required, takes approximately 5 minutes.

B.2 *Plotting Program*

A logic diagram of the program is presented in Fig. 23.

All input data required by the plotting program has been stored on tape by the previous program.

The plotting control bits, referred to in Section 1, have the following meaning: if field calculations are to be made in a combined TE_{11} and TM_{11} mode—that is, input data $TM_{11} \neq (0.0, 0.0)$—the control bit ITE in NAMELIST PDATA must be set equal to 2. If the field calculations are to be made in the TE_{11} mode alone—that is, $TM_{11} = (0.0, 0.0)$—the control bit should be set equal to 1. Therefore calculations for the three sets of radiation patterns, horn, subreflector, and far field, will generally be made in two modes, a combined TE_{11} and TM_{11} mode, and the TM_{01} mode, where it is understood that the combined mode may be the pure TE_{11} mode, if $TM_{11} = (0.0, 0.0)$ and ITE = 1.

For the combined mode the radiation fields will be evaluated in both E and H planes. E- and H-plane data are plotted together for ease of comparison. However, in some cases (particularly for certain far field patterns) a rapidly varying phase plot superimposed on a rapidly varying amplitude plot may result in an unclear graph. For this reason the control bits I1P, I2P, and I3P are introduced. I1P controls horn radiation pattern plotting, I2P controls subreflector plotting, and I3P, far field plotting. If the control bit for a particular field is set equal to 0, two plots will be generated, that is,

Combined Mode

(*i*) E-plane and H-plane amplitude and phase

TM_{01} Mode

(*ii*) Phase and amplitude

However, if the control bit is set equal to 1, four plots will be generated:

Combined Mode

(*i*) E-plane amplitude and H-plane amplitude
(*ii*) E-plane phase and H-plane phase

TM_{01} Mode

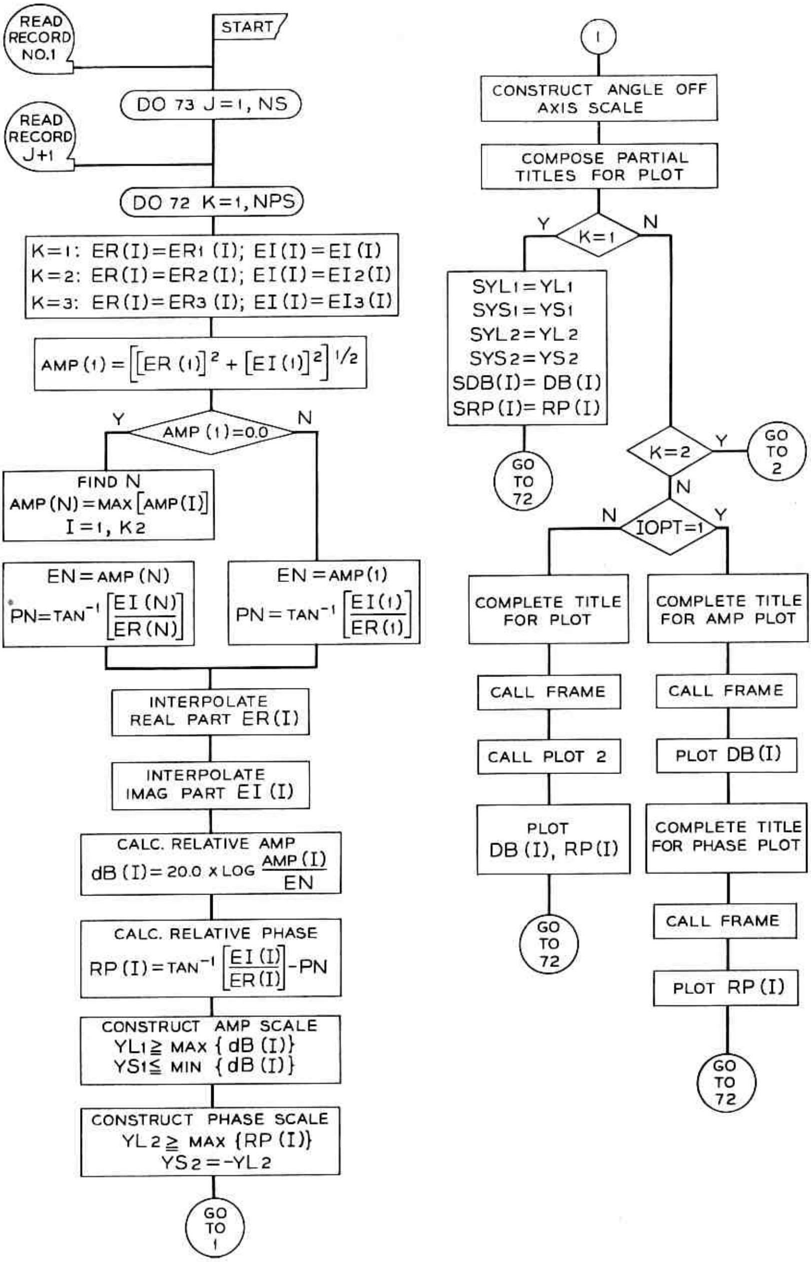
(*iii*) Amplitude
(*iv*) Phase

Vertical scales are restricted to allow only 20 divisions, therefore a preferred set of increments for the various scales has been selected. The allowed increments in dB for the amplitude scale, stored in array ADB(I), are:

0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0;

for the phase scale, in degrees in array APH(I):

1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 8.0, 10.0, 12.0, 15.0, 18.0;


READ RECORD NO.1
START
DO 73 J=1, NS
READ RECORD J+1
DO 72 K=1, NPS
K=1: ER(I)=ER1 (I); EI(I)=EI(I)
K=2: ER(I)=ER2(I); EI(I)=EI2(I)
K=3: ER(I)=ER3 (I); EI(I)=EI3(I)
AMP (1) = [[ER (1)]² + [EI (1)]²] ¹/²
Y
AMP (1)=0.0
N
FIND N
AMP (N)=MAX [AMP(I)]
I=1, K2
EN=AMP (N)
PN=TAN⁻¹ [EI (N) / ER (N)]
EN=AMP(1)
PN=TAN⁻¹ [EI(1) / ER(1)]
INTERPOLATE REAL PART ER(I)
INTERPOLATE IMAG PART EI (I)
CALC. RELATIVE AMP
dB (I)=20.0 x LOG AMP (I) / EN
CALC. RELATIVE PHASE
RP (I)=TAN⁻¹ [EI (I) / ER(I)]-PN
CONSTRUCT AMP SCALE
YL1≧ MAX {dB (I)}
YS1≦ MIN {dB (I)}
CONSTRUCT PHASE SCALE
YL2≧ MAX {RP (I)}
YS2=-YL2
GO TO 1
1
CONSTRUCT ANGLE OFF AXIS SCALE
COMPOSE PARTIAL TITLES FOR PLOT
Y
K=1
N
SYL1=YL1
SYS1=YS1
SYL2=YL2
SYS2=YS2
SDB(I)= DB(I)
SRP(I)= RP(I)
GO TO 72
K=2
Y
GO TO 2
N
N
IOPT=1
Y
COMPLETE TITLE FOR PLOT
CALL FRAME
CALL PLOT 2
PLOT DB (I), RP(I)
GO TO 72
COMPLETE TITLE FOR AMP PLOT
CALL FRAME
PLOT DB (I)
COMPLETE TITLE FOR PHASE PLOT
CALL FRAME
PLOT RP (I)
GO TO 72

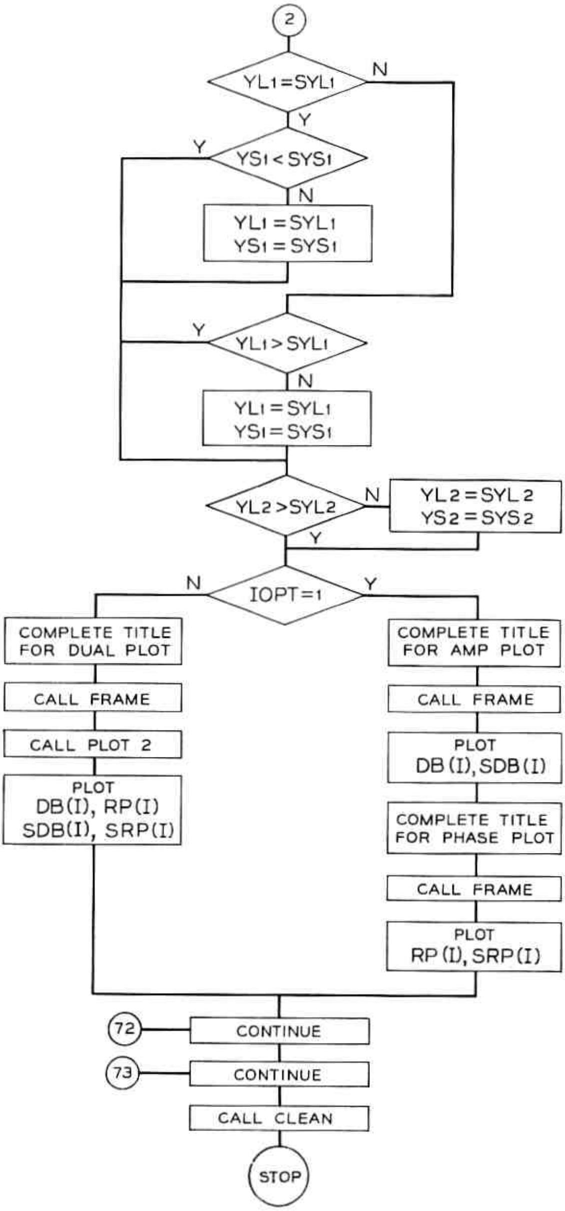


Fig. 23 — Logic diagram for field plotting program.

for the angle-off-axis scale, in degrees, stored in array AS(I):

6.0, 5.0, 4.0, 3.0, 2.5, 2.0, 1.5, 1.0, 0.75, 0.5, 0.4, 0.3, 0.2.

The following subroutines must be included in the deck before execution:

PLOT 2 — A subroutine to generate a grid with two independently labeled ordinates sharing a common abscissa

MINMAX — A subroutine to select the algebraicly largest or smallest entry in an array and specify its index

INTERP — A quadratic interpolation scheme for real arrays

FILTER — Adjusts plotting data for phase variations in the vicinity of ±180°

LABEL 2 — A modified version of the microfilm subroutine LABEL. Called only by PLOT 2.

The program requires approximately $(51{,}536)_8$ or $(22{,}000)_{10}$ words of storage, and about 0.5 minute execution time for all twelve plots.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the comments and suggestions of J. S. Cook, J. N. Hines, R. Lowell, and E. R. Nagelberg.

REFERENCES

1. Cook, J. S., Elam, E. M., and Zucker, H., "The Open Cassegrain Antenna: Part I—Electromagnetic Design and Analysis," B.S.T.J., *44,* No. 7 (September 1965), pp. 1255–1300.
2. Zucker, H., "Fresnel Region Approximation for Wide Angles and Large Fresnel Numbers," IEEE Trans. Antennas and Propagation, *AP-14* (November 1966), pp. 684–688.
3. Hogg, D. C. and Semplak, R. A., "An Experimental Study of Near Field Cassegrain Antennas," B.S.T.J., *43,* No. 6 (November 1964), pp. 2677–2703.
4. Li, T. and R. H., Turrin, "Near Zone Field of a Conical Horn," IEEE Trans. Antennas and Propagation, *AP-12* (November 1964), pp. 800–802.
5. Potter, P. D., "A New Horn Antenna with Suppressed Sidelobes and Equal Beamwidths," Microwave J. *6* (June 1963), pp. 71–78.
6. Hannan, P. W., "Microwave Antennas Derived from a Cassegrain Telescope," IRE Trans. Antennas and Propagation, *AP-9* (March 1961), pp. 140–153.
7. Silver, S., *Microwave Antenna Theory and Design,* New York: McGraw-Hill, 1949.
8. Alsberg, P. A. and Cochran, J. Alan, unpublished work.

Precise 50 to 60 GHz Measurements on a Two-Mile Loop of Helix Waveguide

By D. T. YOUNG and W. D. WARTERS

Precise measurements made in the 50 GHz to 60 GHz band on a two-mile triangular loop of 2 inch diameter helix waveguide are presented. The measuring technique is discussed in some detail regarding accuracy. A brief comparison of the experimental results with theory is made. The average measured attenuation of the waveguide varies smoothly from 2.62 dB per mile at 50 GHz to 2.32 dB per mile at 60 GHz. Fast variations versus frequency were within experimental error. Several short-radius bends of different angles were measured; losses less than 0.8 dB across the band were observed for a 42° bend made of mitered elbows.

I. INTRODUCTION

Low-loss transmission via the TE_{01} mode in circular waveguide has been studied for many years for use as a wideband communication medium.[1] Much work has been done on the design of improved waveguides,[2, 3] the understanding of the effects of spurious modes,[4, 5] and the measurement of sample guides over wide frequency bands.[6, 7]

Interesting waveguide communication system layouts have proposed repeater spacings in the range of 10 to 20 miles. Reasonable design requires that the total loss of such a waveguide section be predictable to within a few dB. However, the longest guides on which measurements have been reported are a few hundred yards, and the variation to be expected between different samples of similar construction is unknown.

This paper describes measurements made in the 50 GHz to 60 GHz band on a two-mile triangular loop of helix waveguide. Extremely precise observations were made on many sections of the loop in order to:

(*i*) Test whether the loss of a long line is indeed the sum of the losses of its component sections as is expected if the sections act independently.

(*ii*) Discover the statistical variations between sections, both in average loss and loss fluctuations with frequency, so that confidence limits can be found for predicting the behavior of very long lines from measurements on shorter lines.

(*iii*) Allow accurate measurements of bends and other components by taking the difference between the losses of sections with and without the test component included.

II. THE TWO-MILE WAVEGUIDE LOOP

The two-mile facility was constructed at Holmdel, New Jersey, by A. C. Beck and C. F. P. Rose. The permanent installation consists of a triangular shaped loop of two parallel 4-inch steel conduits buried below the frost line, with poured concrete ties every 10 feet, along a precisely aligned path. The layout is shown in Fig. 1. The loop begins and ends in a laboratory building, and large waterproof access manholes are provided approximately every 400 feet, as indicated by the letters in Fig. 1. The waveguide was installed in the conduit by adding sections in one manhole and pulling the assembled guide through the conduit to the next manhole with a cable and winch.

The vertical profile of the path is quite smooth, with no radii of curvature less than 4,000 feet. The horizontal plan of the path consists of straight lines, as shown in Fig. 1, except the two sections between manholes U, V, and W. These sections have a constant radius of curvature of 708 feet. The angles at the corners of the loop are 90°, 90°, and 42°.

The waveguide was two-inch inside diameter steel-jacketed helix waveguide. It was constructed at the Holmdel Laboratory by A. C. Beck and C. F. P. Rose and has been described by them.[3] It was made in 15-foot lengths which were connected with precision threaded couplings. The guide rests on its couplings in the steel conduit between manholes, and is thus supported at 15-foot intervals. A short connecting section is provided in each manhole; it is easily removed to allow insertion of measuring gear.

The total added loss at 55 GHz owing to the horizontal and vertical path bends has been calculated by A. C. Beck to be 0.045 dB and 0.002 dB, respectively, for the whole two miles. The former was readily measured, the latter was beyond our measurement accuracy.

To complete the loop, various types of sharp-radius bends were placed in the corner manholes H, O, and U. Section IV describes these bends and experimental measurements on them.

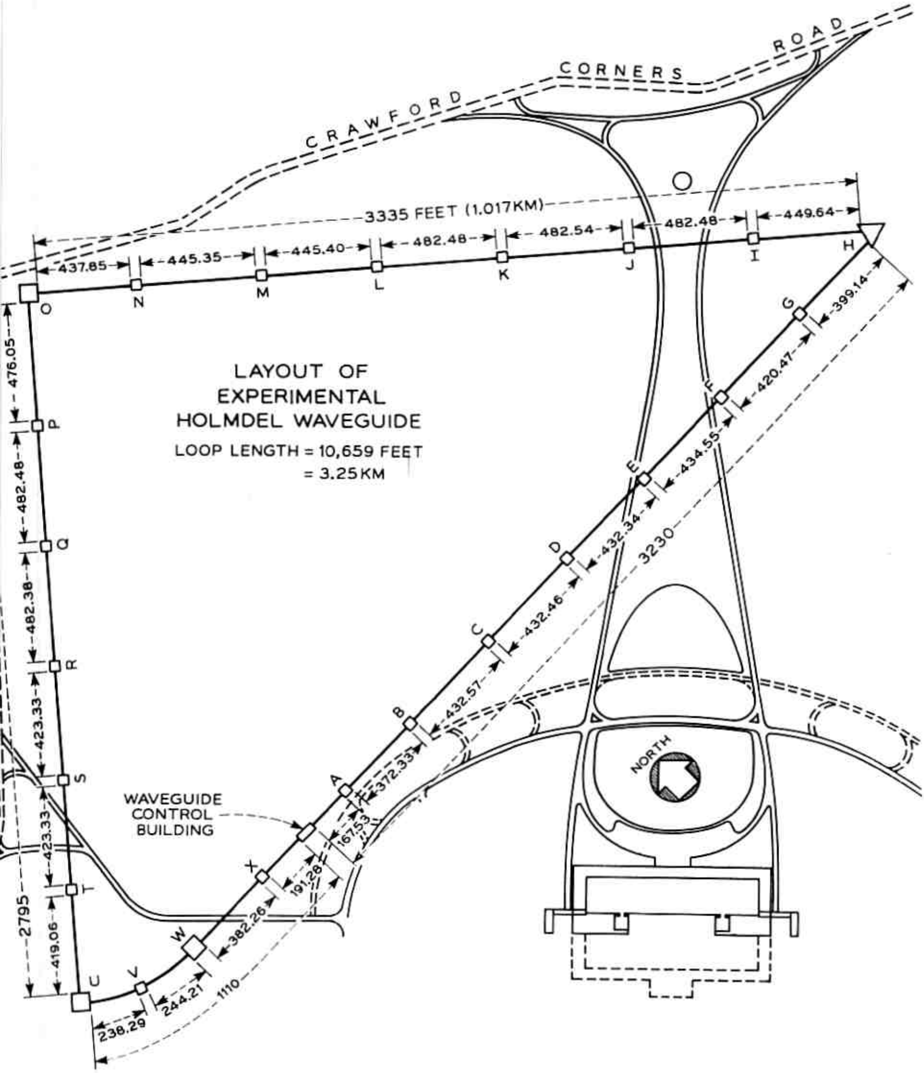


Fig. 1 — Layout of experimental Holmdel waveguide.

The entire waveguide loop, including bends, was made vacuum tight and could be evacuated to a pressure of a few microns of mercury. All measurements were made with the guide filled to slightly over one atmosphere with high purity dry nitrogen. Each time the guide was opened to change experimental conditions it was flushed with nitrogen, pumped, then refilled. These precautions are necessary to eliminate oxygen, which has several strong absorption lines in our band of interest.

III. MEASURING TECHNIQUES AND DATA REDUCTION

The TE_{01} transmission losses of the waveguide sections of interest were measured by the shuttle-pulse method. This method allows highly accurate measurements on low-loss line sections, provided that certain precautions are observed, because it includes observations on many round-trip traversals of the section and because the time resolution of the pulse allows spurious reflections to be avoided.

3.1 *Apparatus*

Figure 2 is a block diagram of the measuring setup. It used a heterodyne receiver system in which the CW beating oscillator signal and the transmitted test pulses are both provided from a single backward wave oscillator by pulsing the beam voltage every 100 μs with a 0.1 μs duration pulse which changes the oscillator frequency by 70 MHz. This scheme was suggested by D. H. Ring and has been described earlier.[6] The test pulses and beating oscillator power driving the converter are reflected from the coupling mesh. A portion of each transmitted signal pulse enters the test section and bounces back and forth between the mesh and piston in the test section many times, thus causing a train of pulses with decreasing amplitudes to be returned to the receiver.

Since the mesh has transmission loss of approximately 17 dB, the level of the signal pulses which have traveled in the test section before returning to the converter is at least 34 dB below the beating oscillator level and good receiver linearity is assured. The transmitter power and receiver noise figure allowed as many as 100 trips to be observed, depending on the length of the test section.

The 70 MHz IF pulse train passes through a precision attenuator adjustable in 0.1 dB steps. The range unit opens a 0.4 μs time gate to select a desired pulse from the train. The selected pulse is peak-detected and read on an expanded-scale levelmeter. The attenuator is set to center the levelmeter, so the entire IF strip operates at constant level. Readings of relative pulse height may readily be made to within 0.05 dB. Measurements are made, after adjusting the BWO and converter for the desired frequency, by selecting a series of pulses (usually 15 to 20) from the train with the range unit and recording

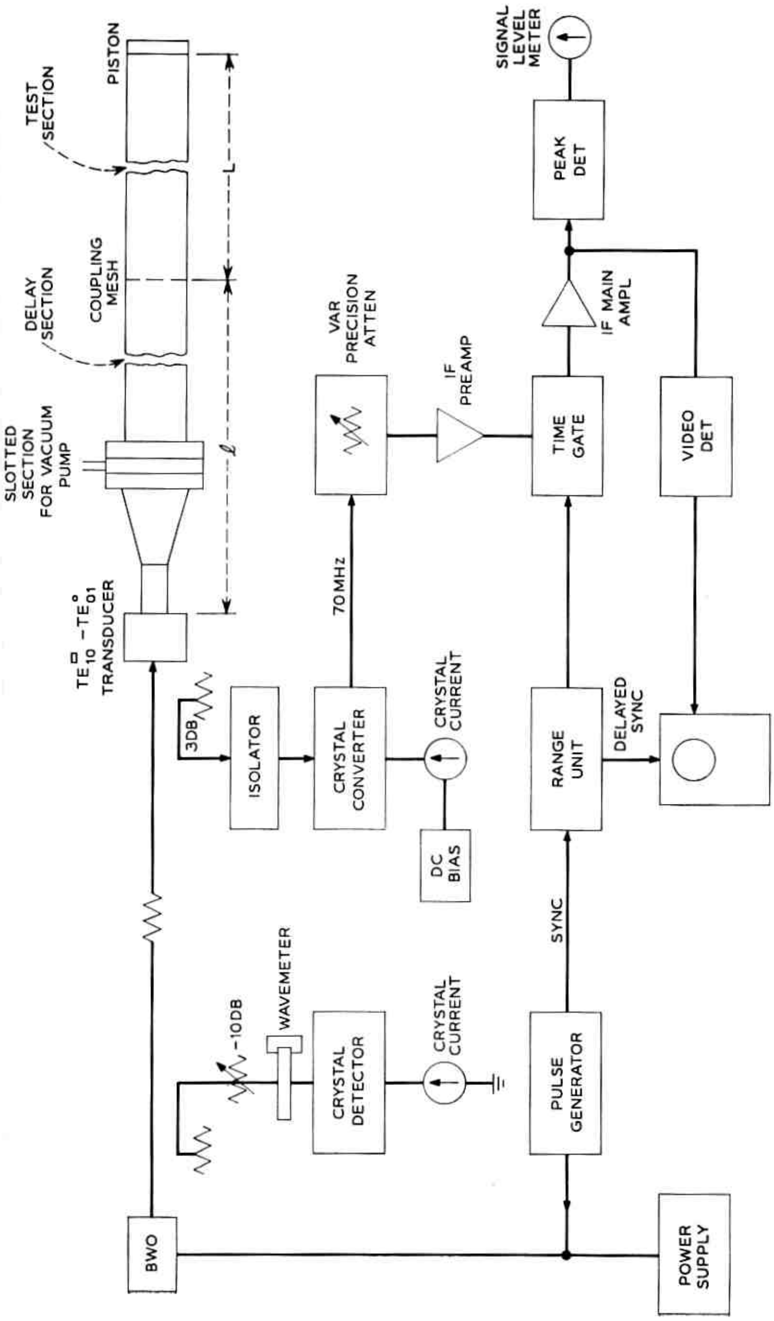


Fig. 2 — Block diagram of measuring setup.

the trip number and relative pulse height of each. This data, together with end corrections for mesh and piston return losses, is reduced by a simple computer program.

The millimeter-wave circuitry is all of precision construction and, with the exception of the $TE_{10}^{\square}$-TE_{01}° transducer, is extremely well matched. The effects of the 15 to 20 dB return loss of the transducer will be discussed later.

The coupling mesh is a flat transverse copper plate 1/32 inch thick with many small uniformly-spaced holes. Care was taken in machining both it and its mounting fixture to insure flatness. The shorting piston was ¼ inch thick solid copper, machined for flatness and polished. The return loss of the mesh was precisely measured by substituting it for the solid piston in a short test line (terminated beyond it) and comparing the relative pulse heights of the 100th trips in the two cases. The mesh return loss varies between 0.054 and 0.114 dB across the 50-60 GHz band. The return loss of the solid copper piston is taken to be the calculated value of 0.005 dB.

The shuttle pulse technique provided a further important advantage for the present experiments, where many sections physically separated by large distances were to be measured, by allowing the test gear to remain in one location. The coupling mesh to the test section was placed in any desired manhole around the loop, and the shorting piston for the far end of the test section was then placed in the appropriate following manhole.

The waveguide between the building where the test gear was located and the manhole where the coupling mesh was located served as a transmission line. Thus measurements could be made on the waveguide section between any two manholes chosen by the experimenter simply by locating the mesh and piston appropriately. The waveguide between the test gear and the coupling mesh also served as a delay line to allow complete time separation between the incident pulse and the pulse reflected from the mismatch at the mode transducer.

Unless these pulses are separated the effective return loss of the coupling mesh will vary considerably. The measured return loss of the coupling mesh is then no longer correct, and this will seriously affect the accuracy of measurement of short waveguides.

3.2 *Data Reduction*

The basic assumption in shuttle pulse measurements is that the loss of each successive trip through the test section is identical, thus the total loss in decibels is a linear function of trip number. The validity of this assumption for our experiment is discussed in Section 3.3.

In order to obtain high precision, many sets of pulse height level vs trip number readings (h_i, n_i) were taken at each test frequency for each test line. To weight the readings equally and to obtain a measure of the experimental precision, a straight line was fitted by the method of least squares. Thus A and B were chosen such that $M(A, B)$ was minimized, where

$$M(A, B) = \sum_{i=1}^{N} (h_i - A + Bn_i)^2.$$

This requires

$$A = \frac{K \sum_{i=1}^{N} n_i h_i - Q \sum_{i=1}^{N} h_i}{K^2 - NQ}$$

$$B = \frac{N \sum_{i=1}^{N} n_i h_i - K \sum_{i=1}^{N} h_i}{K^2 - NQ}$$

where

$$K = \sum_{i=1}^{N} n_i$$

$$Q = \sum_{i=1}^{N} n_i^2$$

and N is the number of data pairs (h_i, n_i). B is therefore the desired experimental loss per trip and A is the intercept at zero trips. A depends upon the transmitter power level and is of interest mainly as an internal check that the data is consistent with other measurements. The attenuation constant α of the test section is then computed from

$$\alpha(f) = \frac{1}{2L} [B(f) - C(f)] \tag{1}$$

where L is the length of the section, B is the measured round trip loss and C is the known end correction.

If we assume that the pulse height measurements h_i are distributed normally about the true line $A_t - B_t n_i$, we can readily calculate the standard deviation of B, our experimental measure of B_t, and thus the accuracy of our experimental value of α.

If we assume the variance $\hat{\Delta}^2$ of the h_i is known so that

$$h_i = A_t - B_t n_i + \Delta_i \tag{2}$$

with $\langle \Delta_i \rangle = \langle \Delta_i \Delta_k \rangle = 0$, except $\langle \Delta_i^2 \rangle = \hat{\Delta}^2$, where $\langle\ \rangle$ is the statistical expected value, then we can readily calculate

$$\langle (B - B_t)^2 \rangle = \frac{N\hat{\Delta}^2}{NQ - K^2}. \tag{3}$$

Thus the accuracy of the experimentally determined loss is related to the individual measurement variance $\hat{\Delta}^2$ by equation (3).

If the variance $\hat{\Delta}^2$ is unknown then one can use the variable

$$t = (B - B_t)\sqrt{\frac{(N-2)(NQ - K^2)}{NM}} \tag{4}$$

where N, M, K and Q are as previously defined. It can be shown[8] that t has Student's t distribution with $N - 2$ degrees of freedom,

$$f(t) = \text{Const}\left(1 + \frac{t^2}{N-2}\right)^{-(N-1)/2}. \tag{5}$$

From (4) we can write

$$\langle (B - B_t)^2 \rangle = \frac{NM}{(N-2)(NQ - K^2)} \langle t^2 \rangle$$

and $\langle t^2 \rangle$ can be evaluated from (5) to give

$$\langle (B - B_t)^2 \rangle = \frac{NM}{(N-4)(NQ - K^2)}, \qquad N > 4. \tag{6}$$

The value of $\langle (B - B_t)^2 \rangle$ was calculated from (6) for each measurement. Comparison with (3) over many measurements gives a value for $\hat{\Delta}$ of about 0.05 dB, which is in agreement with the expected limit of accuracy of our pulse height measurements.

The loss in dB per mile as calculated from (1) and the standard deviation of the measurement as calculated from (6) were plotted versus frequency for each test section by the computer. Some of these results are shown in Figs. 3 to 8 and are discussed in detail in Section 4.

For most test sections, measurements were made at frequencies spaced by 100 MHz from 50 to 51 GHz and from 59 to 60 GHz, and at frequencies spaced by 1 GHz from 51 to 59 GHz. This arrangement allowed a check at the ends of the band on the consistency between the calculated deviations and the actual spread of points, and gave sufficiently fine-grained data across the band to detect any expected

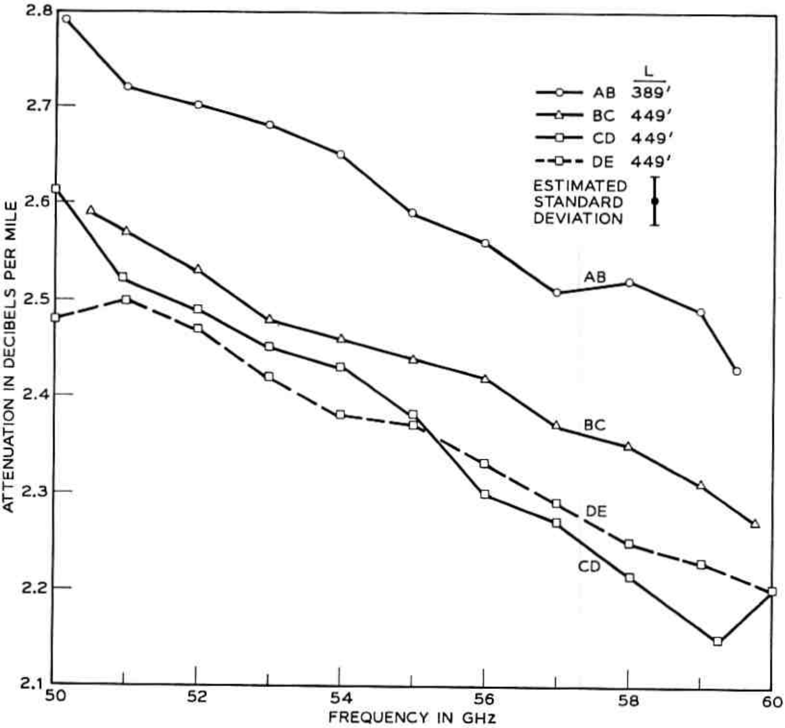


Fig. 3 — Measured attenuation of short length waveguide sections.

variations with frequency. Helix waveguide of the type used in these experiments is not expected to show loss variations vs frequency with periods less than 6 GHz.[9]

3.3 *Experimental Precautions and Limitations*

There are a variety of precautions that must be observed in shuttle pulse measurements in order to avoid anomalies and inaccuracies.

Of prime importance for high precision is that there be no interaction between different traversals of the signal pulse in the test section. Otherwise the observed loss will not be a linear function of the number of trips and the desired single trip loss will be difficult to derive. Interactions can occur in two major ways: (*i*) between successive trips when spurious mode generation is high enough or spurious mode loss is low enough that significant spurious mode power can be built up during one traversal and then be reconverted to the TE_{01} mode in the next

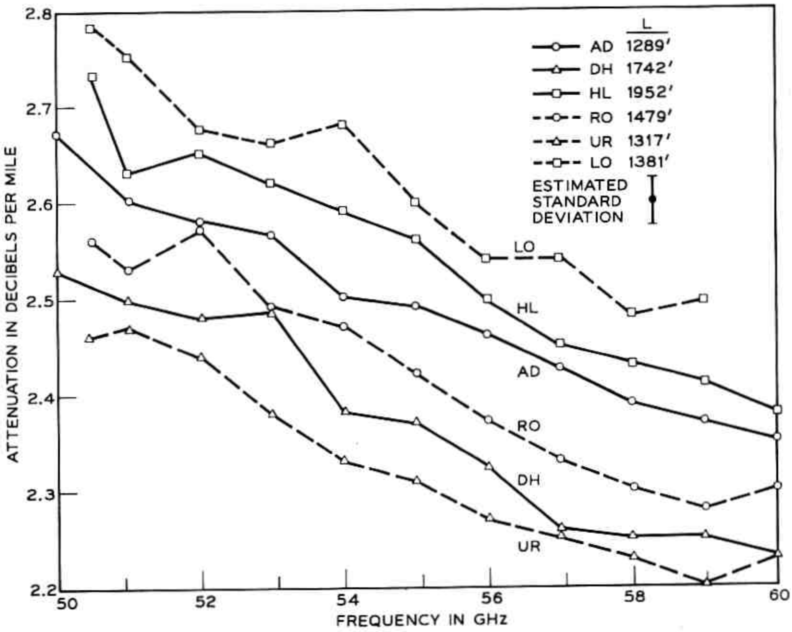


Fig. 4 — Measured attenuation of medium length waveguide sections.

traversal, and (*ii*) between nonsuccessive trips when the test section length L and the delay line length l are related by $mL \approx nl$ so that pulses bouncing in the delay line as a result of reflections or mode conversions at the input transducer can coincide with some of the desired signal pulses bouncing in the test section.

The first type of interaction is readily observed in waveguides with low spurious mode loss and has been discussed[4] in detail. The cure in the low-loss case is to provide mode filters at each end of the test section. For helix waveguide with high spurious mode loss, as in our experiments, it is expected that the spurious mode level is never high enough to cause observable interactions for all except TE_{0n} modes. This expectation was tested in several guide sections at several frequencies by using a movable shorting piston and observing the signal pulse after many round trips. By moving the shorting piston one changes the phases of any reflected spurious modes and thus of the reconverted TE_{01}, causing distinctive variations in the observed signal pulse height.[4] No variations outside experimental uncertainty were observed with the exception of a series of narrow loss peaks at the TE_{02} spacing.

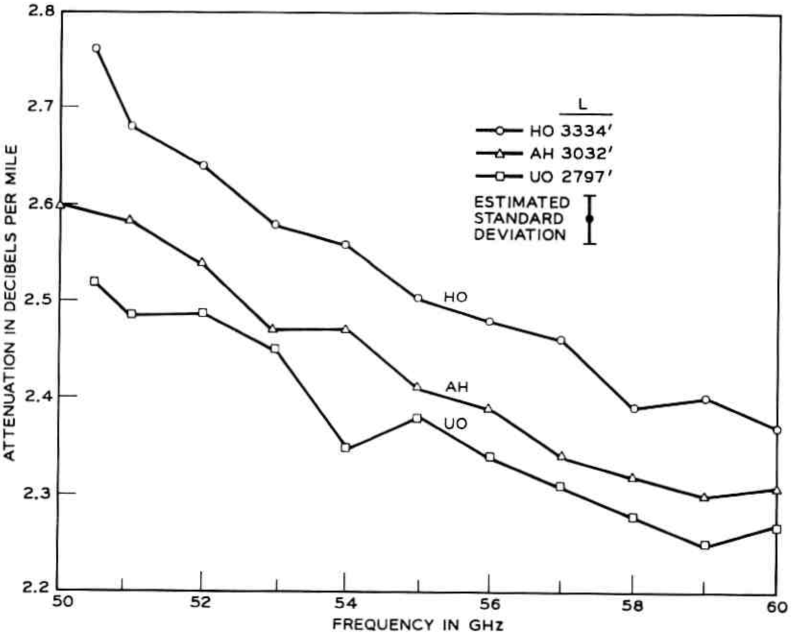


Fig. 5 — Measured attenuation of long length waveguide sections.

The TE_{02} mode is coupled to TE_{01} by imperfections possessing circular symmetry,[10] such as diameter changes or slight dishing of the mesh or end pistons. Its loss in helix waveguide is very low, so it can interact over several trips in short waveguide sections, causing loss peaks when the frequency and section length L are such that $2L$ contains an integral number of TE_{01}–TE_{02} beat-wavelengths, or

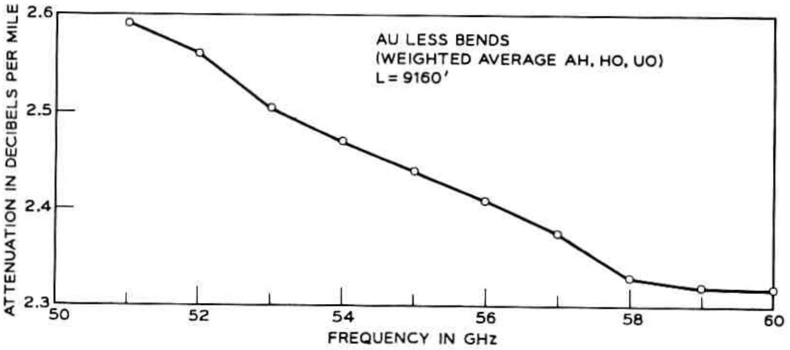


Fig. 6 — Measured attenuation of total straight waveguide.

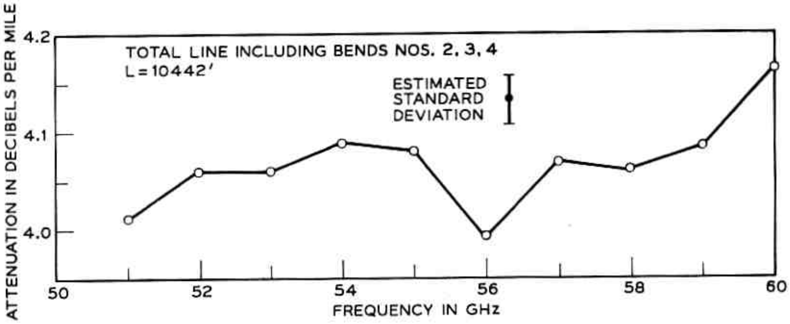


Fig. 7 — Measured attenuation of total line including bends.

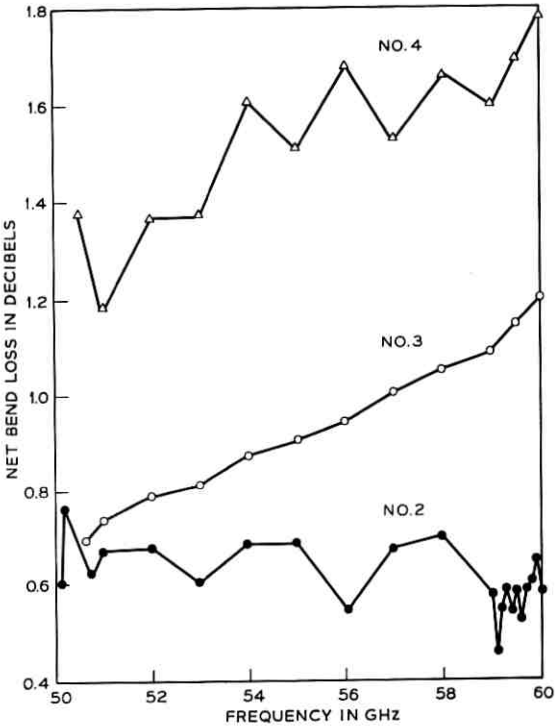


Fig. 8 — Bend losses.

nearly so.[4] The diameter tolerance of the helix waveguide was such that continuous conversion to TE_{02} was not expected to be observable, so conversion at the coupling mesh and end piston was suspected as the cause.

This suspicion was verified by the following experiment. The piston and mesh were fixed, and the test frequency was varied slowly. Loss peaks were observed every 120 MHz although the beat-wavelength condition was satisfied every 60 MHz in the test line. Such an effect should indeed occur if both mesh and piston are converters of roughly similar magnitude. When the coupling mesh was turned around, the loss peaks still occurred every 120 MHz but were shifted 60 MHz to frequencies between those observed originally, thus indicating the expected phase reversal in the coupling at the mesh. Various meshes and end-pistons were tried, with similar results.

It can be shown[5] that conversions 50 dB down at each end of the test section will cause 10 per cent additional loss at the loss-peak frequencies in a 450-foot section; to reduce this number significantly requires flatness beyond that obtainable with simple machining techniques. For our experiments, therefore, we selected the best mesh and piston available, and chose test frequencies which avoided the loss peaks. This precaution was unnecessary in sections 1500 feet or more long, as the extra peak loss was then below experimental uncertainty.

The second type of intertrip interaction was avoided by identifying and observing the spurious pulse trains arising from the delay line. Both TE_{01}, which is reflected for several trips with rapidly decreasing amplitude because of the mismatch at the transducer, and TE_{02}, which is generated at a low level in the taper to the transducer but is then almost totally reflected from taper and mesh on successive trips, are important. Certain test section-delay line combinations with nearly rational length ratios were not measured because these effects were observable. In general they become less important as the delay line length (and therefore its loss) increases.

A third possible cause of nonlinearity between pulse height and trip number is the receiver down-converter noise. This effect was observable only after many trips when the signal pulse was much attenuated and the receiver attenuator was set near zero; it was easily avoided by monitoring the signal-to-noise ratio.

Two other major sources of inaccuracy are oscillator stability and oxygen absorption. At one atmosphere pressure, contamination of the nitrogen filling gas by 0.02 percent of oxygen will increase the measured loss at 60 GHz by approximately 1 per cent. Thus the elaborate flushing procedures mentioned earlier were followed.

The oscillator frequency stability must be sufficient to hold the beating oscillator level at the receiver down-converter constant during a measurement run. The level will vary with frequency because the main return from the mesh at the end of the delay line will phase with the reflection from the transducer mismatch. In addition, the return from the mesh will change when the test line is in the vicinity of resonance for the beating oscillator frequency. These effects become severe as the lengths L and l become large. In the present experiments the BWO beam supply was regulated to a few millivolts, giving frequency stability of a few tens of kHz, but for lengths of either delay or test line of over 1000 feet it was necessary to monitor the converter crystal current very carefully to avoid serious loss of precision, and for lengths over 5,000 feet precise measurements became difficult.

IV. RESULTS AND COMPARISONS

4.1 *Individual Line Sections*

The measured attenuation constants vs frequency for several line sections are plotted in Figs. 3, 4, and 5, grouped roughly by length. The results for sections* AB, BC, CD, and DE, all of which are under 500 feet long, are plotted together in Fig. 3. Sections AD, DH, HL, LO, UR, and RO, from 1289 to 1952 feet long, are shown in Fig. 4. Results for the three long straight runs AH, HO and UO, all around 3000 feet long, are shown in Fig. 5. Notice that in all cases the vertical scales are greatly expanded.

On each figure is indicated the estimated standard deviation of the experimental points, $\langle(\alpha - \alpha_t)^2\rangle^{1/2}$, as calculated from equations (1) and (6). The actual value of this quantity of course varied somewhat from point to point and curve to curve; the indicated amount is a rough average. In general the actual value tended to be a bit larger at the lower frequencies in the band and smaller at the higher frequencies, because of the greater number of trips observable at lower attenuations.

The over-all high quality of the helix waveguide is evidenced by the low observed attenuation constants. The theoretical loss for per-

* The first letter in the section code refers to the manhole in which the coupling mesh was located and the second to the manhole with the piston. Manhole locations are indicated in Fig. 1.

feet solid-copper guide varies from 1.79 dB per mile at 50 GHz to 1.35 dB per mile at 60 GHz. Thus the additional losses from all causes, including finite helix-wire size and pitch, surface roughness, and manufacturing tolerances, total less than 1 dB per mile for most sections.

The rapid variations in loss vs frequency for each section are within the estimated experimental error in most cases. As mentioned earlier, in this waveguide we would expect to see no variations vs frequency with periods less than 6 GHz. None were observed, except the spurious TE_{02} peaks discussed in Section 3.3, and a peak at 54.3 GHz in line LO which is believed to be from a mechanical failure of the steel jacket-lossy lining bond in some of the helix waveguide pieces. Experiments over much wider frequency bands would be necessary to detect the very slow variations which are expected from the random curvature of the waveguide axis.

On the other hand, the difference in measured loss between one line section and another of roughly the same length is much greater than experimental error is most cases, and is therefore quite real. This difference is discussed in detail in Section 4.4, where it is compared with a theoretical estimate. It results from the statistical independence of the loss components between one section and the next; the variations vs frequency for a single section should be as great over frequency differences large enough that the statistical independence again holds.

Figure 6 shows the average attenuation constant for all of the horizontally straight line sections, obtained by adding the measured losses for sections AH, HO and UO and dividing by their total length. Figure 7 shows the average attenuation constant for the entire loop including sharp bends in the corner manholes. The mesh was in manhole A and the shorting piston in the laboratory building at the other end of the waveguide loop; thus everything was included except the short delay-line section between the building and A.

4.2 *Bend Losses*

The losses of several models of sharp bends for use in the corner manholes were measured by taking the difference between the losses of line sections with and without the bends included. The coupling mesh was placed in an appropriate manhole ahead of the corner manhole, and the shorting piston was placed in the corner manhole, first following the bend and then preceding it. In the measurement with bend included, for nonhelix bends, a short section of helix waveguide was placed between the bend and the piston to serve as a mode filter.

The losses of bends 2, 3, and 4 are shown in Fig. 8. These bends were used in manholes H, O, and U, respectively, for the measurement shown in Fig. 7. Bend 2 is made of two 90° mitered elbows back to back, with a rotary joint between them adjusted to give the 42° horizontal angle. The measured loss agrees well with theory.[11] Bends 3 and 4 are 7/8 inch inside diameter helix waveguide with lossy jacket, bent 90° on elastically tapered curves, with effective bend radii of about 3 meters. The loss of bend 3 is in agreement with theory; that of bend 4 is considerably higher.

For all three bends the measurement accuracy is a few hundredths of one dB, thus the plotted variations vs frequency for bends 2 and 4 are real. For bend 2, some phasing between spurious modes generated at the two elbows is to be expected, but for bend 4 the variations further indicate that the helix waveguide was not properly constructed.

Figure 9 shows the measured attenuation of section XU, which contains the 708-foot radius horizontal bend. It also shows the predicted straight loss of XU, obtained by subtracting the calculated

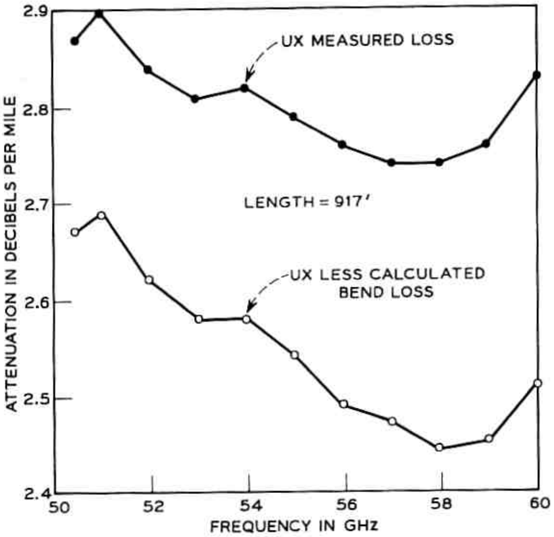


Fig. 9 — Added loss caused by large radius of curvature bend.

bend loss from the measured loss. The agreement with the measured losses for other straight sections as shown in Figs. 3 through 6, indicates that the effect of the horizontal bend is well predicted by theory.

4.3 *Sums of Sections and Residual Errors*

An important purpose of these experiments was to determine whether the sum of the losses of several sections measured individually would be the same as the loss measured for a line made up of the same sections connected. The assumption that this is indeed true is inherent in all predictions of the losses of long lines based on measurements on short lines. It is also inherent in our technique for measuring bends, and it underlies our assumption of the validity of the shuttle-pulse technique in general. Thus, although there were no known reasons to expect the assumption to be false, an experimental verification was considered important.

Figure 10 shows the difference in dB between the measured loss of section AD and the sum of the measured losses of sections AB, BC, and CD, as a function of frequency. The differences are very small indeed. The dashed lines indicate the average across frequency of the estimated standard deviation of the differences about zero as calculated from the sum of the mean square errors of the individual measurements as given by (6). The dashed lines thus indicate only the effect of the scatter of the data points and do not include any effects such as long-term drift of the apparatus between measurements, variations in oxygen contamination between sections, residual tails of the spurious TE_{02} loss peaks, or absolute errors such as in the end correction due to mesh and piston.

The scatter in Fig. 10 of the experimental differences is therefore

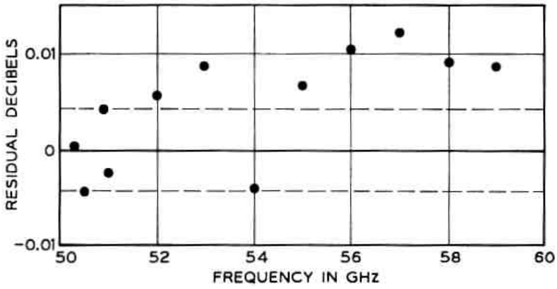


Fig. 10 — Residual dB for difference AD−(AB+BC+CD). The dashed lines are the estimated standard deviations about zero owing to measurement variations only.

quite satisfactory. The measured loss of section AD is about 0.6 dB and of its shorter component sections about 0.2 dB; the largest observed difference is thus just 2 per cent of the loss of AD, and about half of the observed differences are within 1 per cent. All of the differences would be shifted a constant 0.004 dB, or 2/3 per cent of the loss of AD, by a fixed absolute error of 0.002 dB in all measurements. That amount is roughly the limit of accuracy of the measurement of the mesh and piston end correction. In addition, oxygen contamination would cause an error rising from zero at 50 GHz to 1 per cent at 60 GHz in any section with 0.02 per cent oxygen from improper flushing or filling.

The addition of longer sections, where the end correction is unimportant, is shown in Fig. 11. Here the difference is between the loss of section AO and the sum of the losses of sections AH and HO and of bend 2. Bend 2 was itself measured by taking the difference of the losses of section GH with and without the bend included. The dashed lines are again the estimated standard deviation about zero as calculated from data point scatter only. The loss of section AO is about 3.6 dB, so the dashed lines are at slightly over ±1 per cent. The experimental points fall quite satisfactorily within them.

Other additions were checked with similar results. The direct measurements made during these experiments are thus believed to be accurate to the order of about ±1 per cent or ±0.005 dB, whichever is greater, and the sums of losses of individual sections are the same as the loss of the sum of the sections to within that measurement accuracy.

4.4 *Statistical Confidence Limits*

A further purpose of these experiments was to determine experimentally the variation in attenuation between different waveguide sections and to try to discover the length of guide that must be

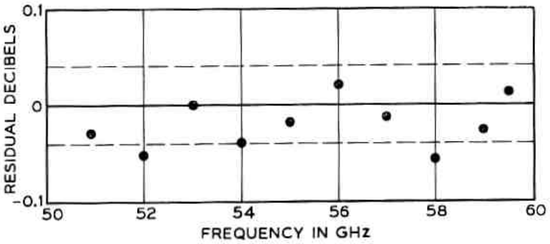


Fig. 11 — Residual dB for difference AO−(AH+HO+bend 2).

measured in order to obtain estimates of a given accuracy on the loss of very long waveguide runs. We assume that the loss variations among sections are caused by variations in the mode conversion in the different sections and are thus determined by a random process whose statistics are related to the statistics of the mechanical tolerances of the guide.[4, 5]

A theoretical solution for the confidence limits on loss as a function of sample length requires knowledge of the probability function for the additional loss caused by mode conversion. An exact solution is difficult when the differential loss between coupled modes is nonzero. An approximate solution for two modes and two polarizations is given in the Appendix; it predicts a normal distribution with mean unity and variance $1/(4 \mid \Delta\alpha \mid L)$ for the quantity $A/\langle A \rangle$. Here A is the additional loss caused by mode conversion and is thus the difference between actual loss and theoretical heat loss. $\langle A \rangle$ is its expected value.

In Fig. 12 the $\pm 2\sigma$ lines for the predicted theoretical distribution are plotted as a function of line length along with experimental values of $A/\langle A \rangle$ for all line sections measured. The experimental value of $\langle A \rangle$ was derived from the curve shown in Fig. 6, so is itself subject to experimental error. The values of $A/\langle A \rangle$ plotted for each line are the means of the maximum and minimum values observed vs frequency.

The fit between theory and experiment would be better if the plotted curves were at $\pm 1\sigma$ instead of $\pm 2\sigma$. However, the approximations of the theory, which includes only one spurious mode, and the experimental accuracies of the points are probably sufficient causes for the poor agreement. In addition the manufacturing variations in our virtually handmade waveguide may be considerable. It should be remembered that

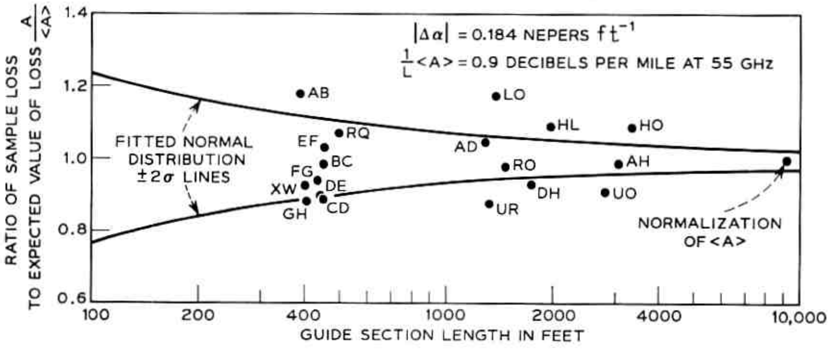


Fig. 12 — Theoretical confidence limits for measurements vs length of waveguide.

the quantity A is the additional loss only, and the value of $\langle A \rangle / L$ is less than one dB per mile. Thus less than 4 per cent variation in the observed total loss will cause a 10 per cent variation in A. The experimental errors are similarly magnified.

Assuming that the theoretically predicted variance of A is correct, one needs to measure 2000 feet of our present waveguide to assure 95 per cent confidence that the measurement is within 5 per cent of the true value of A or thus within 2 per cent of the true value of attenuation constant α. Two per cent gives the loss of a 20-mile section to ±1 dB. If the variance were twice that predicted, one would need to measure four times as much guide, or 8000 feet, for the same confidence.

V. SUMMARY

Precise measurements have been made across the 50 GHz to 60 GHz band on many sections of a two-mile long loop of two-inch inside diameter helix waveguide. The measurement accuracy is approximately ±1 per cent or ±0.005 dB, whichever is greater.

The waveguide is of high quality; the average measured attenuation varies smoothly from 2.62 dB per mile at 50 GHz to 2.32 dB per mile at 60 GHz. Fast variations vs frequency were within experimental error.

The losses of several long sections were compared with the sum of the losses of the smaller sections of which they were composed; the agreement was excellent and within experimental error. Several short-radius bends of different angles were included in the line and measured; losses less than 0.8 dB across the band were observed for a 42° bend made of two mitered elbows.

Differences among line sections in the values of their measured losses were considerably greater than variations vs frequency for any one section, as was expected. It was found that quantities of waveguide from 2000 to 8000 feet must be measured in order to be 95 per cent assured that the measurement is typical of the population to within 2 per cent.

APPENDIX

Approximate Confidence Limits for the Variations Between Guide Sections

For the case of one spurious mode with nonzero differential loss $\Delta\alpha$, Young has shown[9] that the additional loss in a guide of length

L is given by convolving the expression for the additional loss when $\Delta\alpha = 0$ with a particular loss function. Thus

$$A(t) = \int_{-\infty}^{\infty} B(t - s) A_O(s)\, ds, \tag{7}$$

where $A_O(t)$ is the additional loss if $\Delta\alpha = 0$, and where B is the function,

$$B(t) = \frac{2}{\mid \Delta\alpha \mid} \frac{1}{1 + \left(\dfrac{2\pi}{\Delta\alpha} t\right)^2}. \tag{8}$$

The variable t is most conveniently taken as $t = \Delta\beta/2\pi$, where $\Delta\beta$ is the differential phase constant between signal and spurious modes. t is thus roughly proportional to the wavelength λ. The function A_O has been extensively studied by Rowe and Warters,[5] who show that under reasonable restrictions it is a band-limited function and can thus be expressed by its values at its sample points, which are spaced by

$$\Delta t = \frac{1}{2L}. \tag{9}$$

If the convolution function B is much broader than Δt, meaning that $\mid \pi/\Delta\alpha L \mid \ll 1$, we can approximate A_O by constant line segments through its sample-point values, and can estimate the convolution integral (7) as a summation over the sample points. This gives

$$A(t) = \sum_{i=-\infty}^{\infty} B(t - s_i) A_O(s_i)\, \Delta s \tag{10}$$

where

$$s_i = \Delta\beta_i/2\pi = i/2L.$$

Thus

$$A\left(\frac{\Delta\beta}{2\pi}\right) = \frac{1}{\mid \Delta\alpha L \mid} \sum_{i=-\infty}^{\infty} \frac{A_O(s_i)}{1 + \left(\dfrac{\Delta\beta L - i\pi}{\Delta\alpha L}\right)^2}. \tag{11}$$

For convenience we study $A(\Delta\beta/2\pi)$ at the Nth sample-point $(\Delta\beta_N)/2\pi = N/2L$. After substituting $n = N - i$ in the summation, we have

$$A\left(\frac{\Delta\beta_N}{2\pi}\right) = \frac{1}{\mid \Delta\alpha L \mid} \sum_{n=-\infty}^{\infty} \frac{A_O(s_{N-n})}{1 + \left(\dfrac{n}{n_O}\right)^2}, \tag{12}$$

where

$$n_O = \frac{\Delta\alpha L}{\pi}.$$

From our earlier requirement on the width of the loss function B, we require $n_O \gg 1$.

If the coupling coefficient between signal and spurious mode is expressed as a complex Fourier series for the length L, the additional loss A_O is simply expressed in terms of the Fourier coefficients.[5] For two polarizations of the spurious mode,

$$A_O(t) = \frac{1}{2} \sum_{k=1}^{4} I_k^2(t), \tag{13}$$

where

$$I_k(t) = L \sum_{n=-\infty}^{\infty} K_{kn}(-1)^n \frac{\sin \pi(t-n)}{\pi(t-n)}. \tag{14}$$

The index n denotes the nth Fourier coefficient; the index k separates the real and imaginary parts of the coefficients for the two polarizations and thus has four possible values. If the x and y components of the mechanical imperfections are independent random Gaussian variables with white power spectrum, then so are the K_{kn}, at least for large L and over small percentage bandwidths.[5] Under these assumptions one finds that

$$\langle A_O(t_N)\rangle = 2L^2\langle K_N^2\rangle = 2L^2\langle K^2\rangle \tag{15}$$

and

$$\begin{aligned} \langle A_O(t_N)A_O(t_M)\rangle &= 6L^4\langle K^2\rangle^2, \qquad M = N \\ &\approx 4L^4\langle K^2\rangle^2, \qquad M \neq N. \end{aligned} \tag{16}$$

Expressions (15) and (16) are then used to calculate the lower-order statistics of the loss function A from (12), giving

$$\langle A(t_N)\rangle = 2L^2\langle K^2\rangle = \langle A_O(t_N)\rangle \tag{17}$$

$$\begin{aligned} \langle \delta A^2(t_N)\rangle &\equiv \langle (A - \langle A\rangle)^2\rangle \\ &\approx \frac{\langle A_O(t_N)\rangle^2}{4 \mid \Delta\alpha L \mid}. \end{aligned} \tag{18}$$

The requirement $\mid \Delta\alpha L \mid \gg \pi$ has been used to simplify the expressions.

Since A_O is a sum of squares of samples from a Gaussian process, and since A is a weighted sum of values of A_0, it seems reasonable that the distribution function for A should be close to a chi-squared distribution with appropriate normalization. However, for large $\Delta\alpha L$ the approximate chi-squared distribution will have many degrees of freedom, approaching a normal distribution. Therefore, for large $\Delta\alpha$ the variable $A/\langle A\rangle$ becomes normally distributed, with unit mean and variance $1/(4 \mid \Delta\alpha L \mid)$. For this distribution the 95 per cent confidence limits are the $\pm 2\sigma$ lines.

$$\frac{A}{\langle A\rangle}\bigg|_{\pm 2\sigma} = 1 \pm \frac{1}{\mid \Delta\alpha L \mid^{\frac{1}{2}}}. \tag{19}$$

The lines are plotted, together with the experimental observations on various waveguide sections of different lengths, in Fig. 12. The value of $\Delta\alpha$ used is -0.184 neper per foot, which is typical of the differential TE_{12} loss in lossy-jacketed helix waveguide. For sections with L greater than 300 feet, $\mid \Delta\alpha L \mid$ is greater than 55, so the approximation $\mid \Delta\alpha L \mid \gg \pi$ is well satisfied.

REFERENCES

1. Miller, S. E., "Waveguide as a Communication Medium," B.S.T.J., *33,* No. 6 (November 1954), pp. 1209–1266.
2. Unger, H. G., "Helix Waveguide Theory and Applications," B.S.T.J., *37,* No. 6 (November 1958), pp. 1599–1647.
2B. Unger, H. G., "Helix Waveguide Design," Proc. IEE, *106,* Part B, Suppl. 13 (September 1959), pp. 151–155.
3. Beck, A. C. and Rose, C. F. P., "Waveguide for Circular Electric Mode Transmission," Proc. IEE *106,* Part B, Suppl. B (September 1959), pp. 159–162.
4. Rowe, H. E. and Warters, W. D., "Transmission Deviations in Waveguide Due to Mode Conversion: Theory and Experiment," Proc. IEE *106,* Part B, Suppl. 13 (September 1959), pp. 30–36.
5. 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.
6. King, A. P. and Mandeville, G. D., "The Observed 33 to 90 kMc Attenuation of Two-inch Improved Waveguide," B.S.T.J., *40,* No. 5 (September 1961), pp. 1323–1330.
7. Steier, W. H., "The Attenuation of the Holmdel Helix Waveguide in the 100–125 kMc Band," B.S.T.J., *44,* No. 5 (May–June 1965), pp. 899–906.
8. Hoel, P. G., *Introduction to Mathematical Statistics,* New York: Wiley, 1947.
9. Young, D. T., "The Effect of Differential Loss on Approximate Solutions to the Coupled Line Equations," B.S.T.J., *42,* No. 6 (November 1963), pp. 2787–2793.
10. Morgan, S. P., "Mode Conversion Losses in Transmission of Circular Electric Waves Through Slightly Noncylindrical Guides," J. Appl. Phys. *21* (April 1950), pp. 329–338.
11. Marcatili, E. A. J., "Miter Elbow for Circular Electric Mode," Symposium on Quasi-Optics, Polytechnic Institute of Brooklyn (June 1964), pp. 535–542.

A Statistical Theory of Mobile-Radio Reception

By R. H. CLARKE

The statistical characteristics of the fields and signals in the reception of radio frequencies by a moving vehicle are deduced from a scattering propagation model. The model assumes that the field incident on the receiver antenna is composed of randomly phased azimuthal plane waves of arbitrary azimuth angles. Amplitude and phase distributions and spatial correlations of fields and signals are deduced, and a simple direct relationship is established between the signal amplitude spectrum and the product of the incident plane waves' angular distribution and the azimuthal antenna gain.

The coherence of two mobile-radio signals of different frequencies is shown to depend on the statistical distribution of the relative time delays in the arrival of the component waves, and the coherent bandwidth is shown to be the inverse of the spread in time delays.

Wherever possible theoretical predictions are compared with the experimental results. There is sufficient agreement to indicate the validity of the approach. Agreement improves if allowance is made for the nonstationary character of mobile-radio signals.

I. INTRODUCTION

In a typical mobile-radio situation one station is fixed in position while the other is moving, usually in such a way that the direct line between transmitter and receiver is obstructed by buildings. At ultrahigh frequencies and above, therefore, the mode of propagation of the electromagnetic energy from transmitter to receiver will be largely by way of scattering, either by reflection from the flat sides of buildings or by diffraction around such buildings or other man-made or natural obstacles.

1.1 *The Model*

It therefore seems reasonable to suppose that at any point the received field is made up of a number of generally horizontally traveling free-space plane waves whose azimuthal angles of arrival occur at random for different positions of the receiver, and whose phases are completely random such that the phase is rectangularly distributed throughout 0 to 2π. The phase and angle of arrival of each component wave will be assumed to be statistically independent. The probability density function $p(\alpha)$ which gives the probability $p(\alpha)d\alpha$ that a component plane wave will occur in the azimuthal sector from α to $\alpha + d\alpha$ will not be specified, since it will be different for different environments, and is also likely to vary from region to region within one environment; but the assumption that the phase φ has a rectangular probability density function throughout 0 to 2π will be made in all cases.

For simplicity, it will be assumed that at every point there are exactly N component waves and that these N waves have the same amplitude. In addition it will be assumed that the transmitted radiation is vertically polarized, that is, with the electric-field vector directed vertically, and that the polarization is unchanged on scattering so that the received field is also vertically polarized.

The model described so far gives what might be termed the "scattered field," since the energy arrives at the receiver by way of a number of indirect paths. Another term for this scattered field is the "incoherent field," because its phase is completely random. Sometimes a significant fraction of the total received energy arrives by way of the direct line-of-sight path from transmitter to receiver. The phase of the "direct wave" is nonrandom and it may therefore be described as a "coherent wave." It will be seen later that the field in a heavily built-up area such as New York City is entirely of the scattered type, whereas the field in a suburban area with the transmitter not more than a mile or two distant is often a combination of a scattered field with a direct wave.

1.2 *Comparison With Other Proposed Models*

J. F. Ossanna[1] was the first to attempt an explanation of the statistical character of the received mobile-radio signal in terms of a set of interfering waves. He was concerned with measurements taken in a suburban environment, and assumed that reflection occurred at the flat sides of houses and that the incident and reflected waves form an interference pattern through which the receiver moves. He then assumed that all orientations of the sides of houses are equally likely, and hence obtained spectra for the randomly fading signal with the angle between the direction of vehicle motion and the direction to the transmitter as a parameter.

There is quite good agreement between Ossanna's theoretical spectra and those derived from measurements on several suburban streets situated within 2 miles of the transmitter. There is marked disagreement, however, at very low frequencies and at frequencies in the region of the sharp cut-off associated with the maximum Doppler frequency shift. At very low frequencies the spectral energy is always observed to be higher than that predicted by theory, whether Ossanna's or the one we use in this paper. The reason for this is that neither theoretical model takes into account the large-scale variations in total energy which result from the changing topography between transmitter and mobile receiver.

The basic difference between Ossanna's theoretical model and the model used here is that the former is essentially a *reflection* model whereas the latter is essentially a *scattering* model and so includes the former as a special case. An example of the limitations of the reflection model can be seen from the experimental spectra plotted in Ossanna's paper. The spectra are derived from signal-fading records made on several streets whose inclination to the transmitter direction ranged from 15 degrees to 84 degrees, and in each case there is evidence of a shelf which cuts off at twice the maximum Doppler frequency shift. Ignoring the higher harmonics generated in the detection process, the reflection model predicts a spectral cutoff which depends on the direction of the street with respect to the transmitter, ranging from the maximum Doppler frequency shift itself when the street is at right angles to the transmitter direction to twice that value when the street is in line with the transmitter.

With the scattering model, on the other hand, the angular distribution $p(\alpha)$ of scattered waves can be chosen to predict the existence of a spectral shelf out to twice the Doppler frequency shift for any street direction. Another feature of the reflection model which makes it rather inflexible is that for every randomly oriented reflected wave there exists a direct wave incident on the mobile receiver and carrying the same power. Thus the ratio of coherent to incoherent power in the received signal is fixed, whereas in the scattering model this ratio is arbitrary and may be adjusted according to the environment.

In his study of energy reception in mobile radio, E. N. Gilbert[2] examined several models of the scattering type and established a number of important relationships between them. One feature common to all of them, however, was the uniform distribution of waves in angle, although he briefly mentioned the effect of a single strong component arriving directly from the transmitter. The first model Gilbert considered was that of N waves arriving from fixed directions, equally spaced in angle. The phases of the waves were assumed to be independent and uniformly distributed throughout 0 to 2π; their amplitudes were assumed to be Rayleigh distributed and independent, but with the same variance. In a second model the angles of arrival were allowed to occur at random with equal probability for any direction; the phases were again completely random but the amplitudes were assumed to be constant. (This model is the same as the one we use in this paper, with the restriction that $p[\alpha] = [2\pi]^{-1}$.) A third model was an extension of the second to include the case of an arbitrary distribution of the amplitudes. Gilbert showed that the second and third models were equivalent to the first for sufficiently large N.

1.3 *Scope*

This paper shows that the scattering model can be used to predict the statistical characteristics of the signal received at the antenna terminals, hence at the output of a square-law or envelope detector, of the mobile receiving vehicle. These characteristics include the probability distributions of amplitude and phase, spatial correlations, amplitude spectra, and frequency correlations.

A simple relationship is established between the spectrum of the signal input and the product of the azimuthal power gain $g(\alpha)$ of the antenna and the probability distribution function $p(\alpha)$ of the angle of arrival of the component waves. This relationship will be particularly useful in analyzing mobile-radio systems with directional antennas on the mobile unit.

Other topics discussed are the use of space and frequency diversity, coherent bandwidth, and random frequency modulation. Some comments also are made on the nonstationary aspects of mobile-radio fields and on the consequent need for their characterization in terms which will be useful to the mobile-radio system designer. Whenever possible the theory is discussed in the light of available experiments.

II. FIRST-ORDER STATISTICS OF THE FIELD

2.1 *Theory*

Under the assumption that the total field at any receiving point is vertically polarized and is composed of the superposition of N

waves, the n^{th} wave arriving at any angle α_n to the x axis (Fig. 1) with phase φ_n, the field components at point 0 (the zero phase reference point) are

$$E_z = E_0 \sum_{n=1}^{N} \exp \{j\varphi_n\} \tag{1}$$

$$H_x = -\frac{E_0}{\eta} \sum_{n=1}^{N} \sin \alpha_n \exp \quad \{j\varphi_n\} \tag{2}$$

$$H_y = \frac{E_0}{\eta} \sum_{n=1}^{N} \cos \alpha_n \exp \{j\varphi_n\}. \tag{3}$$

In these equations E_0 is the common (real) amplitude of the N waves and η is the intrinsic impedance of free space. The time variation is understood to be of the form $\exp\{j\omega t\}$. Notice that E_z will be proportional to the signal input to the receiver when a vertical dipole antenna is used, and that E_z, H_x, and H_y will be proportional to the three inputs from a Pierce antenna system.[2]

The three field components E_z, H_x, and H_y are complex Gaussian random variables, to a good approximation, provided that N is sufficiently large. This is a consequence of the Central Limit Theorem and the assumption that the phases φ_n are independent of each other and of the angles of arrival α_n. Thus each field component has a real part and an imaginary part which are approximately zero-mean Gaussian random variables of equal variance, the approximation improving for larger N, and provided that the phases φ_n are rectangularly distributed throughout 0 to 2π. Appendix A shows that under the same assumptions the real and imaginary parts of each field component are uncorrelated; they are therefore approximately statistically independent.[3]

An important consequence of this is that the envelope of all three field components (hence of the signals at the terminals of a vertical

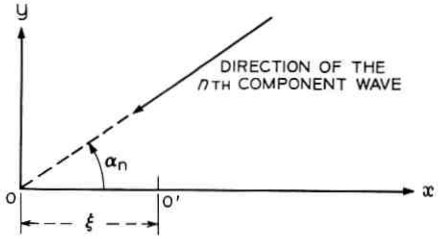


Fig. 1 — A typical component wave and the two field points 0 and 0'.

dipole antenna and of two orthogonal, vertical loops) will be Rayleigh distributed; and their phases will be rectangularly distributed throughout 0 to 2π. (See pp. 160–161 of Ref. 3.)

If, in addition to the N scattered waves, there is a wave of significant magnitude arriving directly from the transmitter, the resulting envelope and phase will no longer be respectively Rayleigh and rectangularly distributed. The relevant distributions will then be those derived by Rice[4] for a sine wave plus random noise. These distributions are, in general, quite complicated (see pp. 165–167 of Ref. 3), but in the limit, when the power in the direct wave is considerably greater than that in the combined scattered waves, both the phase and the envelope are approximately Gaussian distributed; the phase with zero mean and the envelope with a mean value equal to the amplitude of the direct wave.

2.2 *Experiment*

W. R. Young[5] has found that the Rayleigh distribution gives an excellent fit to the observed amplitude fluctuations in mobile-radio reception at 150, 450, 900, and 3700 MHz in New York City, provided that the sample area is less than about 1000 feet square. Trifonov, Budko, and Zutov, in a review of several investigations at 50, 150, and 300 MHz, also found that the Rayleigh distribution fits the data measured in rural suburbs at distances of about 5 and 9 km from the transmitter.[6] The fact that the measured distributions are Rayleigh in the above situations implies that there is no significant directly transmitted component and the fields are wholly of the scattered type, which seems physically reasonable.

Trifonov and his colleagues also found that for short transmission distances in towns (about 1 km), the signal amplitude has a non-zero-mean Gaussian distribution; and that for a transmission distance of 11 km in woodland, the signal has a Rice distribution. In these two cases there is apparently a significant direct component wave, and in the first case, where the transmission distance is only 1 km, the power in the direct component is considerably greater than that in the combined scattered components.

W. C. Jakes and D. O. Reudink have compared the statistical character of the amplitude of the fluctuating signal at the two frequencies of 836 MHz and 11200 MHz on the same street in a suburban environment at about 4 km from the transmitter. They find that the signal amplitudes are Rayleigh distributed at both frequencies, again indicating that the direct wave is not significant.[7] This conclusion is borne out, for reasons discussed in Section 3.2.3, by the shape of the amplitude spectra which were computed from the same data.

The particular section of data which Jakes and Reudink analyzed was chosen with some care. The criterion of choice was that the data should "look" statistically uniform, and although this criterion is both arbitrary and subjective, it is important that it be applied in the absence of any other satisfactory criterion. The point is well illustrated by Fig. 2, which shows a section of signal-amplitude data at 836 MHz, obtained with a vertical dipole on a street adjacent to that used by Jakes and Reudink. The speed of the mobile receiver was 22 feet per second, and each of the five frames lasts about a second (time scale horizontal). The vertical scale is approximately linear in dB, covering a 70 dB spread with about 7 dB to each vertical division.

There is an obvious change in the statistics of the received signal in the fourth frame, compared with the others. (In fact, the fourth frame corresponds to the position of a street intersection, with one

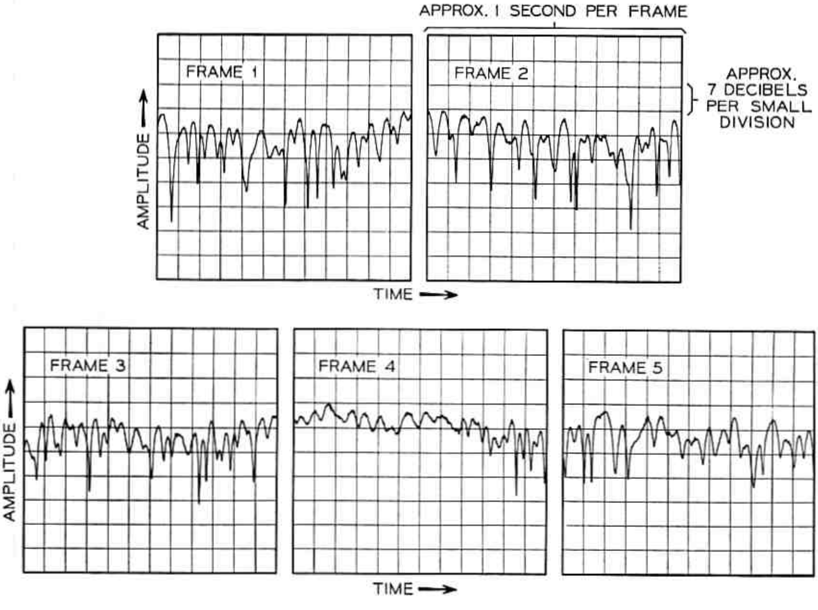


Fig. 2—Section of a mobile radio data run, showing the variation of signal amplitude with time. (One vertical division is approximately 7 dB, and one horizontal frame is approximately 1 second.)

of the intersecting streets pointing in the direction of the transmitter. Then, according to the arguments used above, there will be a strong direct component which will raise the average signal level and change the distribution from a Rayleigh to a Rice or even a Gaussian. The average signal level in the fourth frame does rise, and the distribution does appear to be more symmetrical.) Using all five frames to estimate the probability density function would therefore be misleading in this case since obviously different parts of the data are samples of different distributions.

More subtle differences, as when the distributions underlying the data are all Rayleigh but with different variances over different parts of the run, can be equally misleading. Young found that whereas over fairly small areas of New York City the signal amplitude was accurately described by a Rayleigh distribution, over larger areas—even when the path of the receiver was roughly concentric with the transmitter—the data did not fit a Rayleigh distribution. This is examined in greater detail in Section VI.

III. SPATIAL CORRELATION OF FIELDS

3.1 *Theory*

The field components at some point 0 (see Fig. 1) in the mobile-radio field are given by equations (1), (2), and (3). At another point 0′, a distance ξ away from 0 in the x-direction, the phase of the n^{th} component wave will no longer by φ_n but $\varphi_n + k\xi \cos \alpha_n$, where $k = 2\pi/\lambda$ is the free-space phase constant. In the case of the electric field, the product of the complex conjugate of E_z (the field at 0) with E_z' (the field at 0′) is

$$\begin{aligned} E_z^* E_z' &= E_0^2 \sum_{n=1}^{N} \exp\{-j\varphi_n\} \sum_{n=1}^{N} \exp\{j(\varphi_n + k\xi \cos \alpha_n)\} \\ &= E_0^2 \sum_{n=1}^{N} \sum_{m=1}^{N} \exp\{j(\varphi_m - \varphi_n)\} \exp\{jk\xi \cos \alpha_m\}. \end{aligned} \tag{4}$$

Taking the average (that is, expectation) of both sides of equation (4), the autocovariance function of the electric field is

$$\begin{aligned} R_{E_z}(\xi) &= \langle E_z^* E_z' \rangle_{\text{av}} \\ &= E_0^2 \sum_{n=1}^{N} \sum_{m=1}^{N} \langle \exp\{j(\varphi_m - \varphi_n)\} \rangle_{\text{av}} \langle \exp\{jk\xi \cos \alpha_m\} \rangle_{\text{av}} . \end{aligned} \tag{5}$$

The angular parentheses denote "the average of" the quantity they enclose, and in this case may be thought of as an ensemble average over all the possible situations implied by the assumed statistics of φ and α. The right-hand side is written as the product of two separate averages because of the statistical independence of φ and α. The first of these averages is zero except when $m = n$, so that

$$R_{E_z}(\xi) = E_0^2 \sum_{n=1}^{N} \langle \exp \{ jk\xi \cos \alpha_n \} \rangle_{\text{av}} \tag{6}$$

$$= NE_0^2 \int_{-\pi}^{+\pi} p(\alpha) \exp \{ jk\xi \cos \alpha \} \, d\alpha. \tag{7}$$

In the particular case when the N waves can arrive from any direction with equal probability,

$$p(\alpha) = \frac{1}{2\pi} \qquad -\pi \leqq \alpha \leqq +\pi, \tag{8}$$

the spatial autocovariance function of the electric field becomes

$$R_{E_z}(\xi) = \langle E_z^* E_z' \rangle_{\text{av}} = NE_0^2 J_0(k\xi). \tag{9}$$

The spatial autocovariance functions for the two components H_x and H_y of the magnetic field can similarly be shown to be

$$R_{H_x}(\xi) = \langle H_x^* H_x' \rangle_{\text{av}} = \frac{NE_0^2}{2\eta^2} [J_0(k\xi) + J_2(k\xi)] \tag{10}$$

and

$$R_{H_y}(\xi) = \langle H_y^* H_y' \rangle_{\text{av}} = \frac{NE_0^2}{2\eta^2} [J_0(k\xi) - J_2(k\xi)] \tag{11}$$

for waves arriving from any direction with equal probability. $J_0(\ )$ and $J_2(\ )$ are, respectively, the zero- and second-order Bessel functions of the first kind. The autocovariance functions (9), (10), and (11) are plotted in Fig. 3.

For the same probability density function $p(\alpha)$ of the equation (8) it can be shown in a similar manner that the cross-correlations of the field components are given by the following covariance functions.

$$R_{E_z H_x}(\xi) = \langle E_z^* H_x' \rangle_{\text{av}} = 0 \tag{12}$$

$$R_{E_z H_y}(\xi) = \langle E_z^* H_y' \rangle_{\text{av}} = j \frac{NE_0^2}{2y} J_1(k\xi) = -\langle E_z H_y'^* \rangle_{\text{av}} \tag{13}$$

$$R_{H_x H_y}(\xi) = \langle H_x^* H_y' \rangle_{\text{av}} = 0. \tag{14}$$

These equations show that all three field components are uncorrelated and therefore independent, since the fields are Gaussian at zero spatial

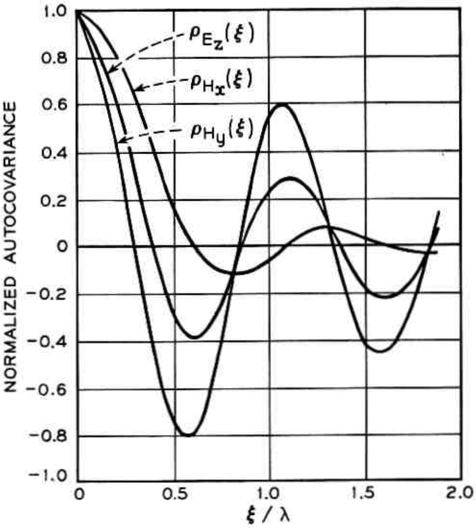


Fig. 3 — The normalized autocovariance functions $\rho_{E_z}(\xi)$, $\rho_{H_x}(\xi)$, and $\rho_{H_y}(\xi)$ from equations (9), (10) and (11).

separation. Further, E_z, H_x and H_x, H_y are uncorrelated and independent for all spatial separations, whereas E_z, H_y are correlated—except at spatial separations corresponding to the zeros of $J_1(\ )$, the first-order Bessel function of the first kind. The normalized covariance function for E_z and H_y is plotted in Fig. 4.

The autocovariance functions (9), (10), and (11) and the covariance functions (12), (13), and (14) are for the particular case of $p(\alpha)$ uniform in the interval $-\pi$ to $+\pi$. The autocovariance and covariance functions for any $p(\alpha)$ can be obtained from equation (7) and similar equations, but those derived here are useful illustrations as well as useful approximations in practice.

In any practical case, however, the complex field components E_z, H_x, and H_y cannot be measured. But their magnitude (that is, envelope or squared magnitude, that is, energy) can. Appendix B shows that the normalized autocovariance function of the departure from their mean of the squared magnitude of complex Gaussian random variables, such as the field components E_z, H_x, and H_y, is equal to the square of the normalized autocovariance function of the complex random variable itself. Taking the electric field E_z as an example, the normalized autocovariance function of the departure $\delta \mid E_z \mid^2$ of the squared modulus from its mean is, from equation (9),

$$\rho_{\delta|E_z|^2}(\xi) = J_0^2(k\xi). \tag{15}$$

Similar normalized autocovariance functions and covariance functions for the squared magnitude of all three field components can be obtained from equations (10) through (14), and they can be shown to agree with the theoretical energy density correlations obtained by Gilbert.[2] This agreement was to be expected since energy density is derived from the squared magnitude of the field components; in addition, Gilbert used a theoretical model which is equivalent to that used here with uniform $p(\alpha)$.

With regard to the envelope of each of the complex field components, Appendix B also shows that the departure of the magnitude of such complex random variables from their mean is described by a normalized autocovariance function which is to a good approximation equal to the square of the normalized autocovariance function of the complex random variable itself. Thus, in the case of the electric field component E_z, again from equation (9),

$$\rho_{\delta|E_z|}(\xi) \cong J_0^2(k\xi). \tag{16}$$

(This quantity is also the normalized correlation coefficient of the signal envelopes at the terminals of two vertical monopole antennas ξ apart on the mobile receiving vehicle which is traveling through an isotropically scattered field.) Similar normalized autocovariance and covariance functions for the magnitudes of all three components can be obtained from equations (10) through (14).

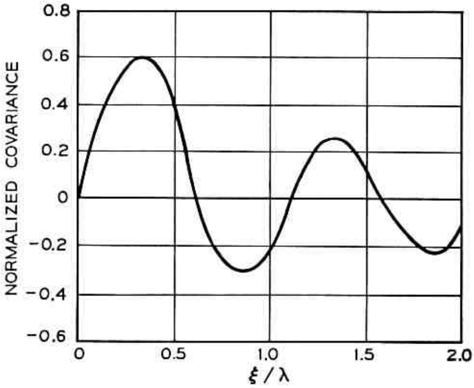


Fig. 4 — The normalized covariance function $\rho_{E_zH_y}(\xi)$, from equation (13).

3.2 *Experiment*

3.2.1 *Spatial Diversity*

Only indirect experimental evidence is available at this time on the spatial correlation of mobile-radio fields. In his measurements of the predetection combining of the signals from several equally spaced vertical monopole antennas, A. J. Rustako found that there was very little difference between the cumulative distributions of the combined amplitudes from four antennas spaced 1/4, 3/4, and 5/4 wavelengths apart.[8] Equation (16) indicates that the correlation coefficients of the signal amplitudes at the antenna terminals at these three separations are about 0.25, 0.06, and 0.03, respectively. Brennan has shown that such correlations produce very little difference in the combined signal from two channels,[9] and so the difference is presumably even less with four channels combined.

3.2.2 *Field Diversity*

Equations (12), (13), and (14) show that all three field components are uncorrelated (and therefore independent, because they are complex Gaussian random variables) at zero separation. The possibility of a "field diversity" system arising from this fact is exploited in the energy density reception scheme from Pierce.[2] (An alternate scheme, proposed by W. C. Jakes, would use predetection combining.[10] This has the advantage that the modulation is not affected.) W. C.-Y. Lee has devised and constructed an energy-density antenna[11] and his analysis of the measurements,[12] based on Gilbert's isotropic scattering model, show sufficient agreement with theory to indirectly confirm equations (12), (13), and (14) at $\xi = 0$.

3.2.3 *Frequency Spectra*

If the mobile receiving vehicle is moving with velocity V in the x direction, the spatial displacement ξ and the corresponding time displacement τ are related by

$$\xi = V\tau. \tag{17}$$

Then all the spatial correlations derived in Section 3.1 can be transformed into time correlations by using equation (17). The Fourier transform of the time autocovariance function then yields the frequency spectrum.

In the case of the signal at the terminals of a vertical monopole antenna in an isotropically scattered field, equations (9) and (17) give the normalized time autocovariance function as

$$\rho_{E_z}(\tau) = J_0(kV\tau). \tag{18}$$

The corresponding input spectrum (see Ref. 3, p. 104) is given by

$$S_{E_z}(f) = \int_{-\infty}^{\infty} \rho_{E_z}(\tau) \exp(-j\omega\tau)\, d\tau \tag{19}$$

$$= \frac{1}{\pi f_m} [1 - f^2/f_m^2]^{-\frac{1}{2}} \qquad |f| \leqq f_m . \tag{20}$$

This spectrum is centered on the carrier frequency and is zero outside the limits $\pm f_m$ on either side of the carrier, where

$$f_m = \frac{V}{\lambda} \tag{21}$$

is the maximum Doppler frequency shift.

Gilbert[2] has shown that the corresponding baseband output spectrum from a perfect square-law detector is given by the complete elliptic integral,

$$S_{\delta|E_z|^2}(f) = \frac{1}{\pi^2 f_m} K\{[1 - (f/2f_m)^2]^{\frac{1}{2}}\}. \tag{22}$$

This output spectrum can be obtained either from the self-convolution of the input spectrum of equation (20) or by taking the Fourier transform of equation (15) expressed as a function of τ by means of equation (17). The spectrum of equation (22) also describes to good approximation the baseband output spectrum from an envelope detector (that is, half-wave linear rectifier). Thus,

$$S_{\delta|E_z|}(f) \cong \frac{1}{\pi^2 f_m} K\{[1 - (f/2f_m)^2]^{\frac{1}{2}}\}. \tag{23}$$

This is a consequence of the approximate equality of the spatial autocovariance functions of equations (15) and (16).

Figure 5 shows input and baseband output spectra for the above case of a vertical monopole antenna in an isotropically scattered field. The sharp cutoff in the baseband spectrum at twice the maximum Doppler shift is observed to some extent in all measured mobile-radio spectra.[1, 8] A small amount of spectral content will occur beyond this cutoff in the case of an envelope detector[13] because of the higher order terms neglected in the analysis, and in all cases because of the finite length of the time series used to compute the spectra. Again,

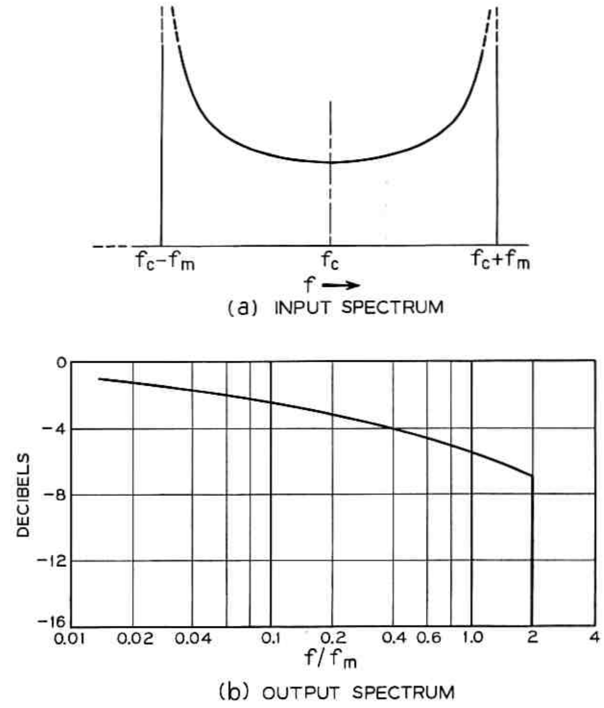


Fig. 5 — Input and baseband output spectra for a vertical monopole antenna in an isotropically scattered field.

in all cases the spectral content at the very low frequency end of the spectrum is much higher than that predicted by theory, owing to the nonstationary character of mobile-radio fields (see Section VI).

But in some cases, such as the spectrum obtained by Rustako,[8] there is reasonably good agreement between the general shape of the spectrum observed and that shown in Fig. 5b. Section IV shows that the theoretical spectra are different, except for the occurrence of the cutoff, if there is a significant directly transmitted component wave in addition to the scattered component waves. Most of the observed spectra seem to be of this latter type.

The above method of deriving spectra, by way of the Fourier transform of the autocovariance function, is not ideal. In all but the simplest cases (for example, when $p(\alpha)$ is uniform), direct integration of equation (7) is often impossibly difficult. As an alternative, the direct method (described in the next section) which depends on associating a Doppler shift with the direction of arrival of each component wave, is much simpler to apply and allows one to retain a clear picture of the underlying physical processes.

IV. SIGNAL SPECTRUM AND ANGULAR PROBABILITY

There is a simple direct relationship between the signal spectrum at the mobile receiver's antenna terminals and the product $g(\alpha)p(\alpha)$. This is the product of the antenna's azimuthal power gain function $g(\alpha)$ and the probability density function $p(\alpha)$, the arrival angles of the plane waves which comprise the field incident on the antenna. Let us look at the use of the relationship for an omnidirectional antenna, the antenna assembly for the Pierce energy density scheme, and an azimuthally directive antenna.

4.1 *The General Relation*

The theoretical model proposed in Section 1.1 describes the field incident on the mobile receiving antenna in terms of a random set of vertically polarized plane waves incident horizontally which occur with probability density $p(\alpha)$, where α is the azimuth angle. Then, because of the vehicle's movement, each angle α (see Fig. 6) will be associated with a Doppler shift f in frequency from the carrier frequency, such that

$$f = f_m \cos \alpha$$

where

$$f_m = \frac{V}{\lambda} \tag{21}$$

is the maximum Doppler shift at the vehicle speed V and carrier wavelength λ.

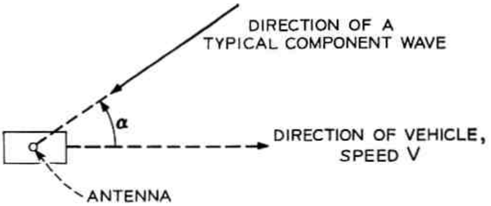


Fig. 6 — Relative directions of the mobile vehicle and a typical component plane wave.

The spectrum of the signal v at the terminals of the receiving antenna on the mobile vehicle will consist of a set of spectral lines which will occur at random in the range $\pm f_m$ about the carrier frequency f_c. The probability that one of these spectral lines will occur in the range from f to $f + df$ is given by the probability density function $p_1(f)$, which may be obtained (see p. 33 of Ref. 3) from the probability density function $p(\alpha)$ by equating the differential probabilities

$$p_1(f)\,|df| = \{p(+\alpha) + p(-\alpha)\}\,|d\alpha| \tag{24}$$

since $+\alpha$ and $-\alpha$ give the same Doppler shift. Then, from equation (23),

$$p_1(f) = \frac{1}{f_m\sqrt{1 - f^2/f_m^2}} \cdot \{p(\alpha)\,|_{\alpha=\cos^{-1}(f/f_m)} + p(\alpha)\,|_{\alpha=-\cos^{-1}(f/f_m)}\}. \tag{25}$$

The signal spectrum $S_v(f)$, the average energy of the signal v in the frequency range f to $f + df$, is given by $p_1(f)$ weighted by the power gain $g(\alpha)$ of the antenna in the corresponding azimuthal direction α. Thus

$$S_v(f) = \frac{1}{f_m\sqrt{1 - f^2/f_m^2}} \cdot \{p(\alpha)g(\alpha)\,|_{\alpha=\cos^{-1}(f/f_m)} + p(\alpha)g(\alpha)\,|_{\alpha=-\cos^{-1}(f/f_m)}\} \tag{26}$$

which is the desired general relation. (See Appendix C for a formal proof.)

4.2 *Application of the General Relation*

4.2.1 *Omnidirectional Antennas*

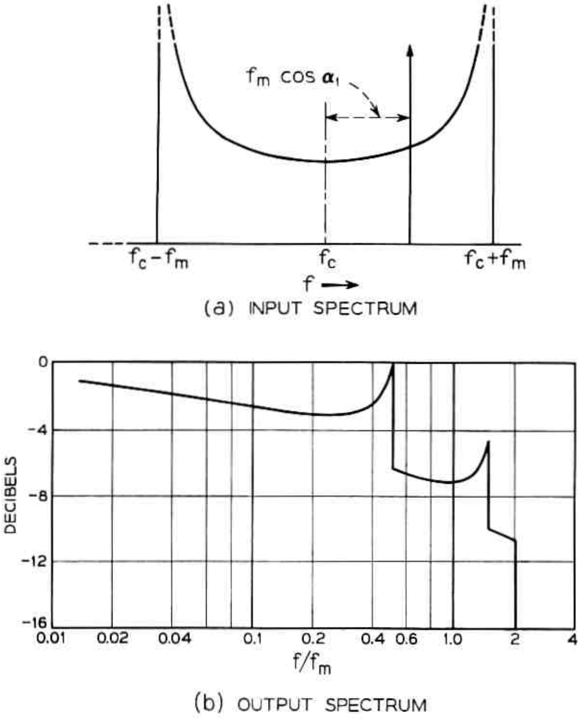
The practical case of most frequent interest is that of a vertical monopole antenna, which has a constant azimuthal gain function, say $g(\alpha) = 1$. Assuming that $p(\alpha)$ is uniform for all angles throughout the range $-\pi$ to $+\pi$, $p(\alpha) = (2\pi)^{-1}$ and the signal spectrum at the antenna terminals would be

$$S_v(f) = \frac{1}{\pi f_m\sqrt{1 - f^2/f_m^2}} \tag{27}$$

for frequency shifts in the range $\pm f_m$ about the carrier frequency f_c, and would be zero outside that range. The spectrum of equation (27) is identical to that of equation (20) which is for the electric field under the same circumstances, an identity that was to have been expected. The spectral shape of equation (27) is therefore that of Fig. 5a. The corresponding receiver baseband output spectrum, assuming square-law detection, would be that of Fig. 5b.

The baseband output spectrum is considerably different if, in addition to the uniformly scattered set of waves, there is a significant wave transmitted directly from the transmitter to the receiver. If the angle of arrival of the direct wave is α_1 the spectrum of the signal at the terminals of an omnidirectional antenna would be that shown in Fig. 7a. This is the basic scattered spectrum of equation (27) together with a spectral line displaced from the carrier frequency by $f_m \cos \alpha_1$.

The corresponding output spectrum from a receiver with a square-law detector (or to good approximation if the detector is half-wave



(a) INPUT SPECTRUM

(b) OUTPUT SPECTRUM

Fig. 7 — Input and baseband output spectra for signals from an omnidirectional antenna, when a uniformly scattered field plus a direct wave are incident.

linear) may be obtained by convolving the above input spectrum with itself. (See p. 255 of Ref. 3.) This yields a baseband output spectrum of the form in Figure 7b, in which α_1 was chosen to be 60 degrees. In general, the high-frequency part of the baseband spectrum ends in a shelf which cuts off at twice the maximum Doppler frequency shift. (In the case of the half-wave linear detector there is a small amount of energy at frequencies beyond the cutoff frequency.)

There are two peaks in the baseband output spectrum which occur at $f = f_m(1 \pm \cos \alpha_1)$. Such peaks, as well as the final shelf, are clearly in evidence in Ossanna's experimental spectra.[1] Figure 8 shows two more experimental spectra, one where the direction to the transmitter was at right angles to the path of the receiving vehicle, and the other where the transmitter was directly ahead. The dashed curves are theoretical spectra with the ratio of power in the direct wave to the total scattered power adjusted arbitrarily. The theory apparently gives the basic form of the experimental spectra, but there are differences in detail.

Of course, complete agreement of theory and experiment is not to be expected. Apart from obvious changes, such as the speed of the vehicle and its inclination to the transmitter direction, the $p(\alpha)$ for the scattered waves and the magnitude of the direct wave will change throughout the entire data run. This means that the time series constituted by the output voltage of the receiver is not a stationary process, whereas the spectra are deduced on the assumption that it is. Methods of approaching this problem of the nonstationarity of mobile-radio data are discussed in Section VI, and methods of making a more valid comparison of theory and experiment are suggested.

4.2.2 *Vertical Loop Antennas*

As a simple example of an azimuthally directional antenna, the vertical loop is interesting because it forms part of the Pierce "total field" antenna system. (See Ref. 2, pp. 14 and 15, where this arrangement of a vertical monopole, together with two orthogonal vertical loops, is discussed in terms of the vertical component of the electric field and the two horizontal components of the magnitude field.)

Assume that the plane of loop 1 (see Fig. 9) lies in the direction of travel and that the plane of loop 2 lies perpendicular to that direction. Then the azimuthal power gain functions for the two orthogonal loops will be of the form

$$g_1(\alpha) = \cos^2 \alpha \tag{28}$$

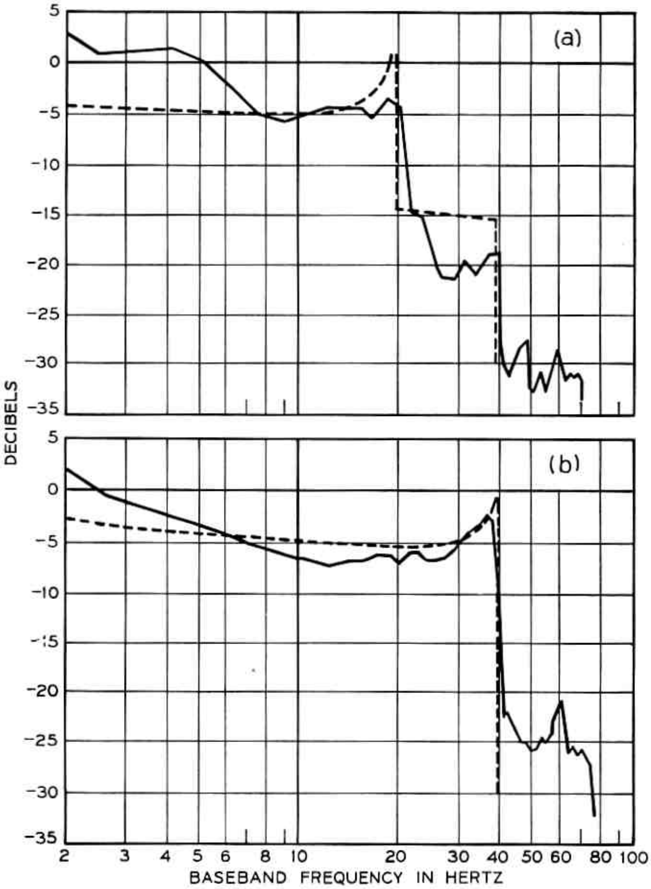


Fig. 8 — Comparison of theoretical (broken line) and experimental baseband output spectra with transmitter (a) at right angles to, and (b) directly ahead of, the vehicle path.

and

$$g_2(\alpha) = \sin^2 \alpha, \tag{29}$$

respectively.

If it is further assumed that the scattered waves are uniformly distributed in angle, that is, $p(\alpha) = (2\pi)^{-1}$, and that there is no significant direct wave. Then, using the general relation of equation (26), the spectra of the signals at the terminals of the two loop

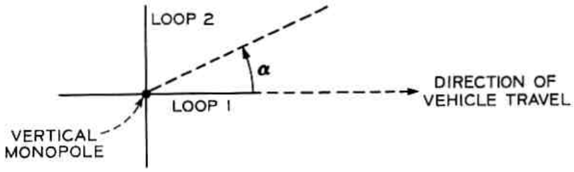


Fig. 9 — Plan view of Pierce antenna system, consisting of a vertical monopole and two orthogonal vertical loops.

antennas will be

$$S_1(f) = \frac{(f/f_m)^2}{\pi f_m \sqrt{1 - f^2/f_m^2}} \tag{30}$$

and

$$S_2(f) = \frac{\sqrt{1 - f^2/f_m^2}}{\pi f_m}. \tag{31}$$

Figure 10 shows these spectra with their corresponding baseband output spectra, assuming square-law detection in the receiver.

The spectra of equations (30) and (31) could also have been obtained from the autocovariance functions of equations (10) and (11) by substituting equation (17) and taking their Fourier transforms. However, the general relation is much simpler to use and indeed is the only reasonable method to use in cases where $p(\alpha)$ and $g(\alpha)$ are other than of the simplest functional form. In addition, the general relation preserves the physical description of the problem. Thus the shapes of the spectra in Fig. 10a have a straightforward explanation in terms of the antenna patterns emphasizing the Doppler shifts resulting from waves arriving from some directions and deemphasizing others—which is precisely the meaning of the general relation of equation (26).

4.2.3 *Beam Antennas*

The general relation of equation (26) gives a simple and direct solution for a beam antenna. The use of such highly directive antennas in mobile radio was suggested by W. C. Jakes[10] with a view to reducing the spectral width, and hence the rate of fading, of the received signal. The general relation shows immediately that such a reduction in spectral width does indeed occur, and gives the precise nature of that reduction.

Consider the idealized beam antenna pattern shown in Fig. 11. The power gain function $g(\alpha)$ in this case can be considered to be unity over the beamwidth β and zero in all other directions. If it is again assumed that the scattered waves are uniformly distributed in angle and that there is no significant direct wave, the effect of the antenna pattern on the spectrum of the signal at the antenna terminals can be thought of in terms of the pattern being a sectoral slice of a fictitious omnidirectional pattern. Hence the spectrum for the beam antenna is a slice taken from the spectrum for an omnidirectional pattern. See equation (27) and Fig. 5a.

When the beam antenna is directed broadside to the direction of

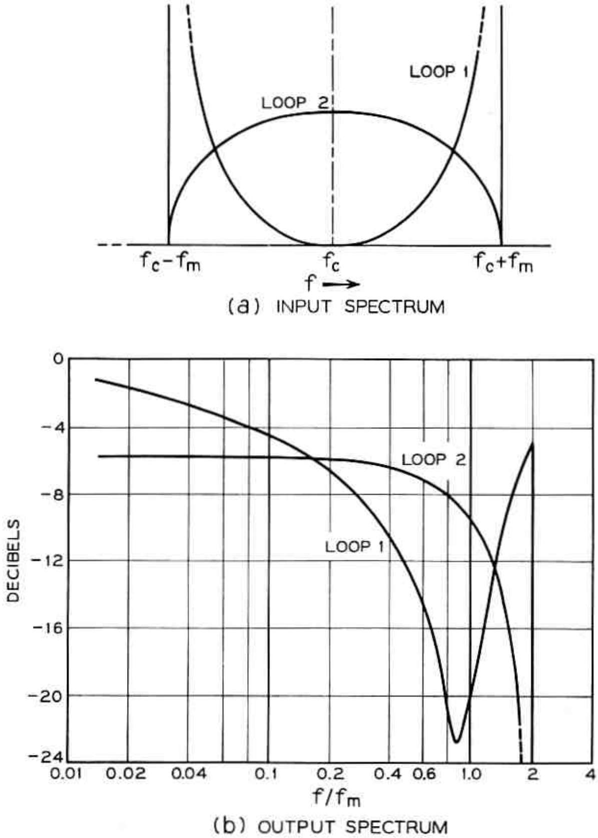


Fig. 10 — Receiver input and baseband output spectra for the two orthogonal loop antennas of Fig. 9.

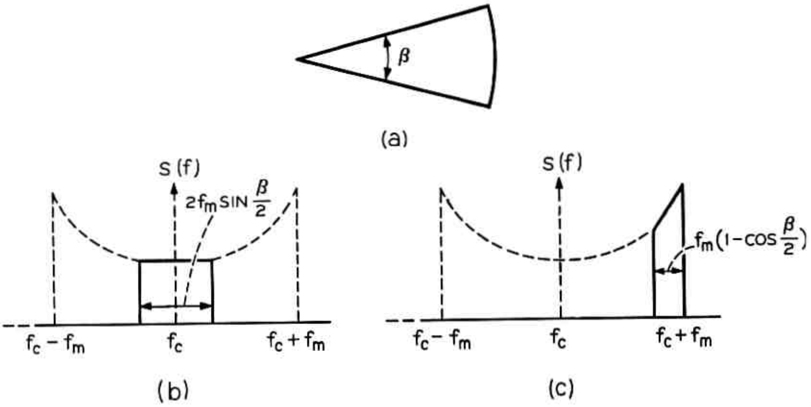


Fig. 11 — Receiver input spectra for an idealized beam antenna used in a uniformly scattered field. (a) Beam antenna pattern. (b) Spectrum for antenna directed broadside. (c) Spectrum for antenna directed straight ahead.

vehicle travel, the spectrum of the signal at the antenna terminals will be that shown in Fig. 11b, where the dashed curve shows the "remainder" of the omnidirectional spectrum. The spectrum is almost flat and is $2f_m \sin(\beta/2)$ wide.

When the beam antenna is pointed straight ahead, along the direction of vehicle travel, the spectrum is that shown in Fig. 11c. Instead of being centered on the carrier frequency, as in the broadside case, the spectrum occurs at the extreme right of the omnidirectional spectrum, and is $f_m[1 - \cos(\beta/2)]$ wide.

Thus it is apparent that the use of highly directive antennas in mobile radio will lead to a reduction in spectral width. W. C.-Y. Lee has confirmed this experimentally, using an array antenna at 836 MHz in a suburban environment.[14] Lee derived from the measured data the rate of crossing of the signal at a certain level and plotted this against antenna beamwidth. Rice has shown that for a narrowband random signal which has a symmetrical spectrum about the carrier frequency, the rate of signal crossing at a certain level is just the probability density at that level multiplied by the square root of the second moment of the spectrum about the carrier frequency.[4] In this way the level crossing rate at a particular level is a measure of the width of the spectrum of the fading signal.

The sectoral beam pattern assumed in the early part of this section never occurs in practice. It is worth emphasizing this rather obvious point in connection with calculating spectral second moments. Because, even though the antenna sidelobe level might be uniformly low, there will be spectral content throughout the entire range of $f_c \pm f_m$. Also, the basic omnidirectional spectrum emphasizes the contributions at the extremes of this range. Hence calculations of the spectral second moments might well be in error if they are based on the assumption that the side-lobe level is zero.

V. CORRELATION BETWEEN SIGNALS OF DIFFERENT FREQUENCIES

The problem of correlating two signals of slightly different frequencies occurs in mobile radio when questions of maximum usable bandwidth, or the use of a pilot signal at a frequency other than the carrier frequency, arise. Let us show that the covariance of two signals as a function of their frequency separation is simply the characteristic function of the probability density function of the time delays suffered by the component plane waves which are assumed to compose the mobile radio field.

5.1 *Theory*

Suppose that the transmitted signal contains two unmodulated signals of frequencies ω_1 and ω_2, whose difference $\Delta\omega = \omega_2 - \omega_1$ is small enough not to violate the following assumptions. Assume that the two signals take exactly the same time to travel from transmitter to mobile receiver along any one of the scattering paths assumed in the model in Section I. This assumption implies that propagation along all paths is by way of freespace type waves (which do not suffer dispersion), and that any phase changes experienced at reflecting or diffracting objects are independent of frequency. Associate a time of travel t_n with the n^{th} component wave, and define a time delay Δt in comparison with the shortest possible time of travel t_o such that

$$\Delta t_n = t_n - t_o . \tag{32}$$

To preserve the assumption made in all previous sections that the phases of the component waves are random and equally probable throughout 0 to 2π it is necessary that the average magnitude of the time delay difference between the n^{th} and m^{th} waves, assumed to be independent, be

$$\langle | t_n - t_m | \rangle_{\text{av}} \gg 1/f_c \tag{33}$$

where f_c is a frequency in the neighborhood of f_1 and f_2.

The electric fields at the two frequencies may be written as

$$E_1 = E_{01} \sum_{n=1}^{N} \exp \{j\omega_1(t - t_n)\}$$

$$E_2 = E_{02} \sum_{n=1}^{N} \exp \{j\omega_2(t - t_n)\}$$

where E_{01} is the amplitude at frequency f_1 of all the waves, and similarly E_{02} is the common amplitude at f_2. Forming the complex product

$$E_1^*E_2 = E_{01}^*E_{02} \exp \{j(\omega_2 - \omega_1)t\} \sum_{n=1}^{N} \sum_{m=1}^{N} \exp \{-j(\omega_2 t_m - \omega_1 t_n)\}$$

and taking the expectation of both sides,

$$\langle E_1^*E_2\rangle_{\text{av}} = E_{01}^*E_{02} \exp \{j(\omega_2 - \omega_1)t\} \sum_{n=1}^{N} \langle \exp \{-j(\omega_2 - \omega_1)t_n\}\rangle_{\text{av}} \quad (34)$$

since it has been assumed that the time delays are independent, and therefore that

$$\langle \exp \{-j(\omega_2 t_m - \omega_1 t_n)\}\rangle_{\text{av}} = 0 \quad \text{for} \quad m \neq n$$

as a consequence of inequality (33). The covariance of the two fields as a function of their frequency separation $\Delta\omega$ is therefore

$$\begin{aligned} R_{12}(\Delta\omega) &= \langle E_1^*E_2\rangle_{\text{av}} \\ &= NE_{01}^*E_{02} \exp \{j\ \Delta\omega t\} \exp \{-j\ \Delta\omega t_o\}\langle \exp \{-j\ \Delta\omega\ \Delta t\}\rangle_{\text{av}} \end{aligned}$$

where the subscript n has been dropped on Δt_n because the average is the same for any n. The normalized magnitude of $R_{12}(\Delta\omega)$ is:

$$|\ \rho_{12}(\Delta\omega)\ | = \langle \exp \{-j\ \Delta\omega\ \Delta t\}\rangle_{\text{av}} \quad (35)$$

is simply the characteristic function, with negative argument, of the probability density function for the time delays Δt. (See Ref. 3, p. 50.)

As an example, suppose that the time delays are exponentially distributed, so that the probability density function of Δt is

$$p(\Delta t) = \frac{1}{T} \exp \left\{-\frac{\Delta t}{T}\right\} \quad \text{for} \quad 0 \leqq \Delta t \leqq +\infty \quad (36)$$

where T is a measure of the spread of the time delays. Then the normalized magnitude of the covariance function in equation (35) becomes

$$|\ \rho_{12}(\Delta\omega)\ | = [1 + (\Delta\omega T)^2]^{-\frac{1}{2}}, \quad (37)$$

which is shown in Fig. 12. It is apparent that the correlation falls off significantly for frequency separations $\Delta\omega > 1/T$, the inverse of the measure of the spread in time delays.

5.2 *Experiment*

Aside from its mathematical convenience, the exponential distribution of time delays seems physically plausible on the grounds that the shorter delays appear more likely to occur than the longer delays. Indeed, the pulse observations made by Young and Lacy at a frequency of 450 MHz in New York City support this contention.[15]

Ossanna has computed the envelope correlations from measurements at 860 MHz in a suburban environment for two-carrier frequency separations of 0.1, 0.5, 1.0, and 2.0 MHz.[16] The corresponding covariances are shown as circles in Fig. 12, where it has been assumed that $T = 1/4$ μsec. A comparison of these experimental points with the theoretical curve indicates that an exponential distribution of time delays is a reasonably good assumption, and that in the suburban environment where the experiments were performed the time-delay spread T is about $1/4$ μsec.

In contrast, Young and Lacy's pulse measurements indicate a time-delay spread about 5 μsec, but with an approximately exponential distribution. The reasons for the difference in time-delay spreads appears to result from the different environments in which the experiments were performed, not to the different frequencies, because their difference is not great. Thus in a suburban environment the component waves are likely to have been redirected by objects within a few hundred feet of the mobile receiver, whereas in New York City

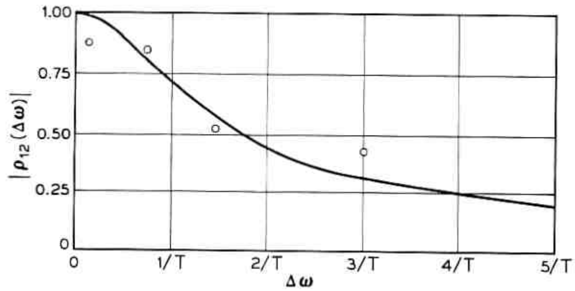


Fig. 12 — Normalized covariance of two signals as a function of their frequency separation, assuming an exponential distribution of time delays with delay spread T. The circles are Ossanna's experimental points.

the range of these objects can reasonably be put at many thousands of feet.

5.3 *Significance of the Random Time Delays*

The immediate benefit of knowing the probability distribution of the time delays of the component waves is that it enables one to deduce the "coherent bandwidth" for that particular system. But the significance of the time delays is much more than this, in that it emerges as a basic characteristic of the system along with the probability distribution of the angles of arrival of the component plane waves.

Indeed, it would appear that *a knowledge of the joint distribution* $\mathrm{p}(\alpha,\Delta \mathrm{t})$ *of the angles of arrival* α *and the delay times* $\Delta \mathrm{t}$ *provides an almost complete description of the mobile radio field; hence, of the mobile radio signals sensed by antennas moving through this field.*

Thus, integration of the joint distribution with respect to α yields the distribution of time delays. Then if the standard deviation of the time delays is large compared with a period of the carrier frequency, the component waves may be said to be completely randomly phased and their phases and angles of arrival to be independent. The results obtained in Sections II, III, and IV would then follow, because they are based solely on the knowledge of $p(\alpha)$ and the assumptions that the phase is completely random and independent of the angle of arrival.

An interesting sidelight is that the cross-covariance of two signals of different frequencies, one shifted in time by τ from the other, depends on the joint distribution $p(\alpha, \Delta t)$. The Fourier transform of this cross-covariance yields the cross-spectrum of the two frequency-separated signals.

It is tempting to assume that α and Δt are independent, thus making the calculation much simpler. But this does not yield answers that accord with experiment; so one must conclude that α and Δt are not independent. This also seems a reasonable conclusion on physical grounds, since it is likely that the shortest time delays will be associated with angles of arrival from the general area of the transmitter, and that the longest delays will be associated with the opposite direction.

VI. THE NONSTATIONARY CHARACTER OF MOBILE RADIO SIGNALS

A perennial difficulty in the analysis of mobile-radio data is its nonstationary character. This makes both the analysis arbitrary and its interpretation uncertain. This section attempts to meet this difficulty directly, rather than trying to find sections of data that "look" stationary or attempting to "doctor" data to that same end before it is analyzed.

The data chosen for analysis were those obtained by Rustako on a single omnidirectional antenna at 836 MHz along Sherwood Drive, a suburban street approximately 2 miles from the transmitter and running at an angle of about 48° to the transmitter direction.[8] The choice of data was made on the grounds that Rustako's computed output spectra most closely resembled the shape of the theoretical output spectrum of Fig. 5b which is for a completely scattered field with no significant directly transmitted component.

Two tests were performed on the data, one to determine the probability distribution of the envelope and the other to determine its time correlation by using Kolmogorov's structure function.

6.1 *The Probability Distribution*

6.1.1 *Theory for a Stationary Process*

According to the theory of Section 2.1, if the field incident on the mobile receiver is of the scattered type, each component wave being independent and randomly phased, then the probability density function (p.d.f.) of the envelope R is Rayleigh, that is,

$$p(R) = \frac{2R}{\sigma^2} \exp\left\{-\frac{R^2}{\sigma^2}\right\} \quad \text{for} \quad 0 \leqq R \leqq +\infty \tag{38}$$

which has the corresponding cumulative distribution function

$$P(R) = \int_0^R p(R)\, dR = 1 - \exp\left\{-\frac{R^2}{\sigma^2}\right\}. \tag{39}$$

This distribution has a root-mean-square value

$$\sqrt{\overline{R^2}} = \sigma \tag{40}$$

a mean value

$$\langle R \rangle_{av} = \frac{\sqrt{\pi}}{2}\,\sigma = 0.886\sigma \tag{41}$$

and a most probable value (or "mode")

$$R\,\big|_{p_{\max}} = \frac{1}{\sqrt{2}}\,\sigma = 0.707\sigma. \tag{42}$$

A convenient method of testing whether or not a given set of statistical data follow an assumed distribution is as follows.[17] First the histogram of the data (that is, relative frequency diagram), which is the practical approximation to the probability density function, is obtained. This is then summed point by point to give the cumulative frequency diagram, which is the practical approximation or estimate $\hat{P}(R)$ of the cumulative distribution function $P(R)$. Then $\hat{P}(R)$ is plotted against $P(R)$. If the two are identical for all R, then the resulting plot will be a straight line from (0, 0) to (1, 1). If not, the departure of the plot from the straight line is a measure of the departure of $\hat{P}(R)$ from $P(R)$.

In analyzing Rustako's data the question to be answered was how closely the data followed a Rayleigh distribution. The appropriate $P(R)$ is then that of equation (39); and the value of σ can be obtained from the maximum of the histogram with the aid of equation (42). The above arguments assume that the data is a stationary process.

6.1.2 *Theory for a Nonstationary Process*

If the theory of Section 2.1 is modified slightly to take account of the undoubted fact that either the number or the magnitude of the component waves will vary as the vehicle moves along its path by normalizing to the local mean, and if the assumption that the field is completely scattered is retained, then the expected distribution of the envelope will again be Rayleigh. However, the root-mean-square value σ will no longer be a constant, but will vary with time in some manner $\sigma(t)$. The envelope can now be classed as a nonstationary Rayleigh process.

It is possible to estimate $\sigma(t)$ from the record by computing the "local" mean $\langle R\rangle_{\text{av}}(t)$; then from equation (41)

$$\langle R\rangle_{\text{av}}(t) = 0.886\sigma(t). \tag{43}$$

Hence, writing the new random variable

$$r = \frac{R}{\sigma(t)} = \frac{0.886R}{\langle R\rangle_{\text{av}}(t)} \tag{44}$$

which in effect has a root-mean-square value of unity. The r process will be a stationary Rayleigh process with a p.d.f.

$$p(r) = 2r \exp \{-r^2\}.$$

Equations (43) and (44), in effect, remove the nonstationary effects from the statistics. The meaning of "local" is explained further in the next section.

6.1.3 *Analysis*

Rustako's data, which had been converted to digital form at 500 samples per second, was taken in sets of 4000 points at a time. Notice that such a length of data contains approximately 200 fading cycles.

Each set was analyzed, first of all, on the assumption that it was stationary, by the method outlined in Section 6.1.1. To obtain the histogram, the amplitude range between the lowest and the highest value was divided into 50 equal slices. The $\hat{P}(R)$ versus $P(R)$ plots for three sets of data are shown on the left side of Fig. 13. Each point corresponding to a partiuclar slice level. The three sets of data were chosen to illustrate where $\hat{P}(R)$ is always greater than $P(R)$, where $\hat{P}(R)$ is always less than $P(R)$, and where they are approximately equal. On the assumption that all three sets of data are stationary it would have to be said that the first two cases are definitely non-Rayleigh while the third case is.

Next, the same sets of data were normalized by the method outlined in Section 6.1.2. The local mean for every point was obtained by averaging the 200 points symmetrically adjacent to that point. The resulting normalized random variable was then treated in exactly the same way as the unnormalized random variable. The right side of Fig. 13 shows plots of $\hat{P}(r)$ versus $P(r)$. It can be seen that in the first two cases the normalized random variable is much more closely Rayleigh distributed than is the unnormalized random variable. The third case is interesting because, although the normalization was not necessary to reduce the data to a stationary Rayleigh process, it demonstrates that the technique of normalization itself does not significantly impair the original process.

In conclusion, it can be said that the technique of normalizing a nonstationary Rayleigh process by way of its running mean can be used to determine whether or not the process is in fact Rayleigh. But it must be emphasized that the technique cannot be applied to processes that are non-Rayleigh. It is certainly possible, however, that different techniques along these same lines might apply to different processes, although it would appear that some knowledge of the expected distribution is essential. The Rayleigh process is one of the simplest to handle because it is determined by a single parameter. In the example used here the Rayleigh process was clearly indicated

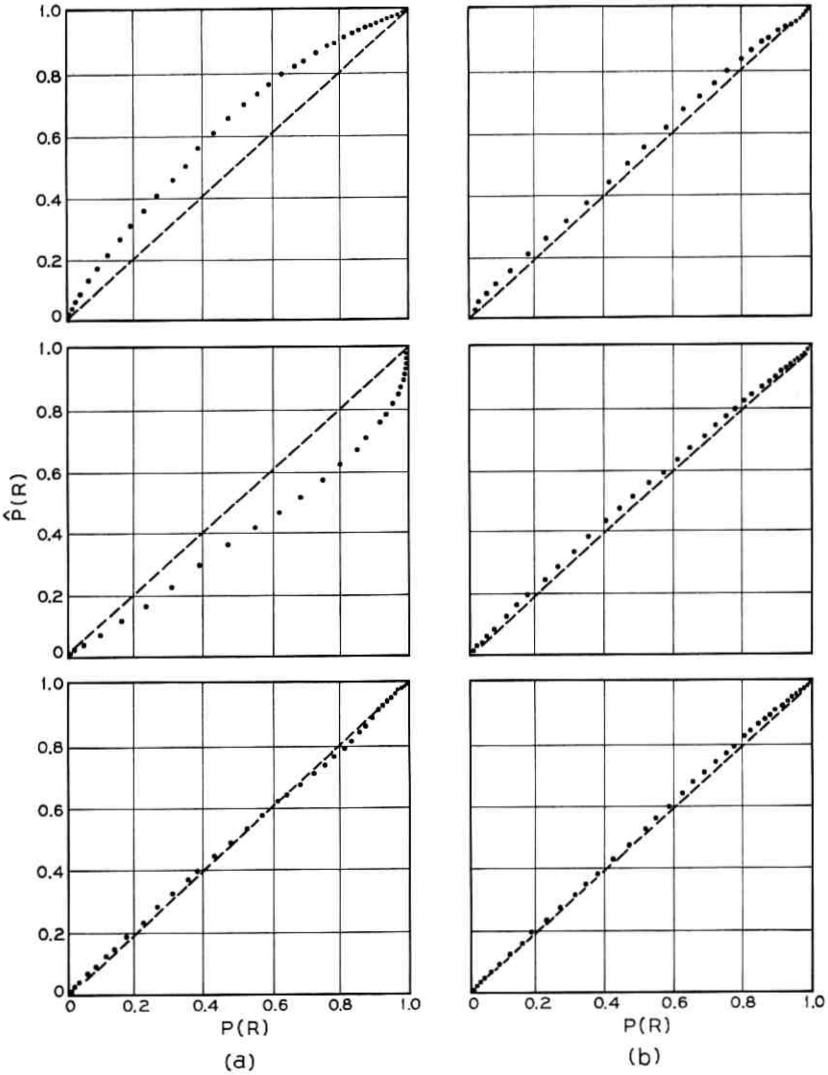


Fig. 13 — Plots of $\hat{P}(R)$ versus $P(R)$. (a) For the raw data. (b) For the same data normalized by its running mean.

by the theory, and the analysis amounts to a positive confirmation of its applicability.

6.2 *Using Kolmogorov's Structure Function*

Tartarski[18] has described the value of using a "structure function" in specifying random variables which are not statistically stationary. (The technique was first used by Kolmogorov to describe meterorological quantities.) The structure function might be of value in analyzing nonstationary mobile radio data.

6.2.1 *Definition and Properties*

The simplest type of structure function, $D_f(\tau)$ of the real random variable $f(t)$, is defined by

$$D_f(\tau) = \langle[f(t+\tau) - f(t)]^2\rangle_{\text{av}} \tag{45}$$

where the angular parentheses denote a time average. This should be compared with the more commonly used autocovariance function, defined for a stationary random variable whose mean is zero by

$$R_f(\tau) = \langle f(t+\tau)f(t)\rangle_{\text{av}}\,. \tag{46}$$

Thus the structure function for a stationary random variable which can be written in terms of the autocovariance function is

$$D_f(\tau) = 2[R_f(0) - R_f(\tau)]. \tag{47}$$

As an example, the structure function for a stationary random variable with a Gaussian autocovariance function, $\exp\{-\tau^2/\tau_0^2\}$ in which τ_0 is constant, is depicted by the solid line in Fig. 14. The equation of this solid line is

$$D_f(\tau) = 2[1 - \exp\{-\tau^2/\tau_0^2\}].$$

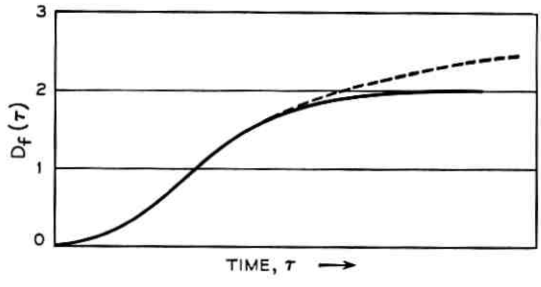


Fig. 14 — Structure functions for stationary (solid line) and nonstationary (dashed line) random variables.

Now, if the random variable is nonstationary in that it has, say, a slowly varying mean value, then the structure function would be modified in some way such as that shown dashed in Fig. 14. This dashed portion would very likely be indeterminate, so that the corresponding autocovariance function would be indeterminate *for all* τ. Hence the value of working, at least initially, with the structure function: if the random variable is stationary, that will immediately be apparent in that $D_f(\tau)$ will approach a horizontal asymptote for large τ; and if it is nonstationary, the portion for small τ can be relied on.

The dashed portion of Fig. 14 can be shown to correspond to an increase in low-frequency spectral energy compared with the stationary case.[18]

6.2.2 *A Structure Function Computed from the Data*

The solid line in Fig. 15 shows the structure function for Rustako's Sherwood Drive data, computed from the definition of equation (45). The data, again consisting of 4000 points, roughly straddled that which gave the first two probability plots of Fig. 13. The structure function is shown out to a time separation τ of 50 data points, or 100 msecs.

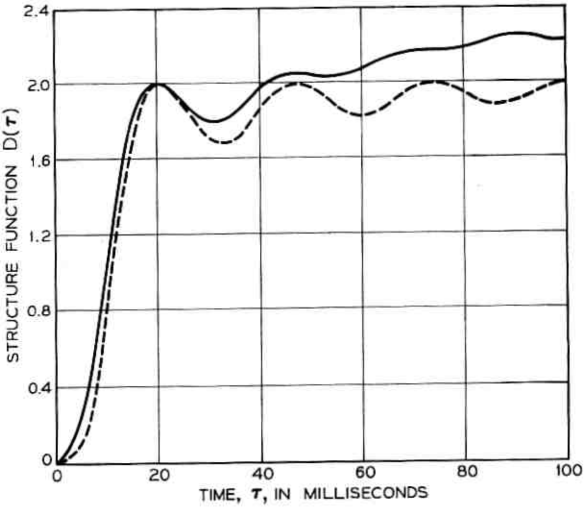


Fig. 15 — Structure function computed for Rustako's data (solid line). The dashed line is the theoretical structure function for a stationary random variable.

The dashed curve is a theoretical structure function for an assumed stationary process with an auto-covariance function of the form $J_0^2(2\pi f_m \tau)$, where f_m is the maximum Doppler shift. This autocovariance function, which is derived from equations (16) and (17), is for the departure of the signal envelope from its mean value for the case of an omnidirectional antenna in a uniformly scattered field. The theoretical and experimental structure functions were arbitrarily made equal at the first maximum.

The experimental structure function, which is typical of many that were obtained, exhibits some of the features that were expected. The initial part of the curve, for small τ, closely follows the theoretical curve, and the quasiperiodic nature of the curve for large τ is also evident. In this region the experimental curve rises systematically above the theoretical curve, as was to be expected for nonstationary data.

This upward trend of the experimental structure function for large τ corresponds to the repeated observation of baseband low-frequency content at a significantly higher level than the theory predicts.

If this large-scale trend in the structure function were removed, then the modified structure function should agree with the theoretical structure function, provided that the basic assumptions of the theory are sound. The curves do differ, both in the amplitude and the period of the quasi-periodic variation. However, this might well result from the wrong choice of $p(\alpha)$, and not to a basic flaw in the theory.

It is evident that the structure function does afford a method of analyzing nonstationary data. The effect of large-scale variations shows up in the structure function and can be removed at that point, rather than by tampering in an arbitrary manner with the original data. Then the modified structure function can be compared with theoretical forms which are appropriate to stationary data.

VII. CONCLUSIONS

The theory presented in this paper attempts to explain the statistical behavior of fields and signals encountered in mobile radio in terms of a set of independent plane waves, redirected by scattering and reflecting obstacles, and incident horizontally on the mobile receiving vehicle. These waves can be described statistically by the joint probability density function $p(\alpha, \Delta t)$ such that the probability of a wave arriving at the azimuthal angle α with a time delay Δt is $p(\alpha, \Delta t) d\alpha d(\Delta t)$.

At ultrahigh frequencies and above, in urban and suburban environments, the spread in the magnitudes of the time delays is sufficiently large, compared with the radio-frequency period for the waves, to be considered randomly phased, in which case the following conclusions apply.

The field components are Gaussian, in the sense that their real and imaginary parts are independent zero-mean Gaussian random variables of equal variance. Thus the envelope of a signal derived from such a field by an antenna will be Rayleigh distributed, unless there is a significant nonscattered wave arriving directly from the transmitter, in which case the envelope will be Rice distributed.

The spatial correlation of the field components may be derived from the probability density function $p(\alpha)$. The spectrum of the signal at the antenna terminals may be derived from the product of $p(\alpha)$ with $g(\alpha)$, the azimuthal gain function of the antenna. The coherence of two radio frequencies, as a function of their frequency separation, may be derived from the probability density function of the time delays $p(\Delta t)$.

A brief examination of available experiments reveals that simple forms of both $p(\alpha)$ and $p(\Delta t)$ give theoretical results which agree broadly with experiment. We do not claim detailed agreement, nor does this seem possible until more complete experimental information is available. It does appear, however, that it is essential to take account of the nonstationary character of the signals obtained in mobile radio when attempting such a comparison.

The theoretical approach we have taken is midway between a purely phenomenological one, based on a complete catalog of the statistical characteristics of mobile-radio signals received under a variety of circumstances, and a purely analytical one in which the transmission environment is specified in detail. The phenomenological approach would be incomplete, in that it would not provide knowledge of why the signals have the character observed. The analytical approach is impossibly difficult to execute. Our approach, which seeks to describe the mobile-radio fields in terms of the compact (though not necessarily simple) quantity $p(\alpha, \Delta t)$, does provide the system designer with information which he can use to advantage in a straightforward way. The following is an example to illustrate this claim.

For example, suppose that experiments in a particular environment have shown that $p(\alpha)$ is roughly uniform and that $p(\Delta t)$ is approximately exponential with parameter T such that T is very large compared with the period of the proposed carrier frequency of the mobile radio system. Then it is known that if an antenna with uniform gain in azimuth is used on the receiving vehicle the received signal will be a Rayleigh distributed fluctuating quantity with a baseband spectrum approximately uniform out to a frequency $2V/\lambda$, where V is the vehicle speed and λ is the carrier wavelength.

This system can be improved in a number of ways. The *depth* of fading, as Rustako has demonstrated,[8] can be reduced by using a number of such antennas separated by a sufficient distance for the signals to be essentially uncorrelated. The signals are then brought to a common phase, at which point they are combined before detection. The resulting signal is therefore the sum of a number of independent, Rayleigh distributed amplitudes, which for a large number will approach a Gaussian distribution with a nonzero mean.

Furthermore, the ratio of the root-mean-square fluctuation to the mean of the combined signal will decrease as the square root of the number of signals combined (by an approximate application of the Central Limit Theorem). Alternatively, the rate of fading, as Lee has demonstrated,[14] can be reduced by using directional antennas, which give a reduced spectral width of the fading* and hence a reduction in its rate.

W. C. Jakes has suggested a system, particularly suited for use at microwave frequencies, which combines the advantages of both a reduced depth and a reduced rate of fading.[10] The system consists of a number of directive antennas mounted on a single mobile unit and pointing in different azimuthal directions. If the signals from the different antennas are brought to a common phase and then combined before detection, the resulting signal will not only be considerably reduced in bandwidth compared with the case if an omnidirectional antenna had been used, but its depth of fading will also be reduced according to the square root of the number of antennas used. The widest coherent bandwidth that can be transmitted in the situation assumed is about T^{-1}.

VIII. ACKNOWLEDGMENTS

I wish to express my gratitude to W. C. Jakes, Jr., for his continued interest and helpful criticism throughout the work reported here, to C. C. Cutler, D. J. Hudson, M. J. Gans, E. N. Gilbert, H. L.

* In this connection, in a strictly literal sense "the medium is the message." If, as has been assumed, an unmodulated carrier is transmitted, then the received signal on a single omnidirectional antenna is both amplitude- and frequency-modulated (see Appendix D) because of the movement of the receiver through the scattering medium.

Schneider, and W. R. Young, Jr., for very useful discussion, to J. F. Ossanna, Jr., and A. J. Rustako, Jr., for generously supplying some of their data and calculations, and to Mrs. C. L. Beattie and Mrs. J. K. Reudink for invaluable assistance with the computations.

APPENDIX A

On the Correlation of the Real and Imaginary Parts of the Field Components

It is important to know the precise conditions under which the six real random variables comprising the real and imaginary parts of the three field components of equations (1), (2), and (3) are uncorrelated. Thus

$$E_z = E_0 \sum_{n=1}^{N} \cos \varphi_n + jE_0 \sum_{n=1}^{N} \sin \varphi_n$$

$$H_x = -\frac{E_0}{\eta} \sum_{n=1}^{N} \sin \alpha_n \cos \varphi_n - j\frac{E_0}{\eta} \sum_{n=1}^{N} \sin \alpha_n \sin \varphi_n$$

$$H_y = \frac{E_0}{\eta} \sum_{n=1}^{N} \cos \alpha_n \cos \varphi_n + j\frac{E_0}{\eta} \sum_{n=1}^{N} \cos \alpha_n \sin \varphi_n \, .$$

Denoting the real and imaginary parts of each field component by the superscripts (r) and (i), the correlation coefficient of the real and imaginary parts of the electric field, is

$$\langle E_z^{(r)} E_z^{(i)} \rangle_{\text{av}} = E_0^2 \sum_{n=1}^{N} \sum_{m=1}^{N} \langle \cos \varphi_n \sin \varphi_m \rangle_{\text{av}} = 0$$

since the φ_n's are independent and rectangularly distributed throughout 0 to 2π.

Similarly,

$$\langle H_x^{(r)} H_x^{(i)} \rangle_{\text{av}} = \frac{E_0^2}{\eta^2} \sum_{n=1}^{N} \sum_{m=1}^{N} \langle \sin \alpha_n \sin \alpha_m \cos \varphi_n \sin \varphi_m \rangle_{\text{av}} = 0$$

and

$$\langle H_y^{(r)} H_y^{(i)} \rangle_{\text{av}} = \frac{E_0^2}{\eta^2} \sum_{n=1}^{N} \sum_{m=1}^{N} \langle \cos \alpha_n \cos \alpha_m \cos \varphi_n \sin \varphi_m \rangle_{\text{av}} = 0$$

with the additional assumption that the φ_n's and α_n's are statistically independent. It can also be shown, based on the foregoing assumptions, that the correlation coefficient for any component real part and any component imaginary part is zero.

Notice that the above correlation coefficients are zero whatever the probability density function $p(\alpha)$ is of the α_n's. Where $p(\alpha)$ is important is in the correlation coefficients for the component real parts with each other and for the component imaginary parts with each other. For example,

$$\langle E_z^{(r)} H_x^{(r)} \rangle_{\text{av}} = -\frac{E_0^2}{\eta} \sum_{n=1}^{N} \sum_{m=1}^{N} \langle \sin \alpha_n \cos \varphi_n \cos \varphi_m \rangle_{\text{av}}$$

is zero if the further assumption is made that $p(\alpha)$ is rectangular throughout $-\pi$ to $+\pi$. Then the correlation coefficient is zero for any pair of component real parts and for any pair of component imaginary parts.

APPENDIX B

Correlation of Fields—Their Magnitudes and Squared Magnitudes

Section 2.1 and Appendix A show that under certain conditions the fields in mobile radio are "Gaussian fields," which means that a typical field component F (either an electric or magnetic component) may be represented by

$$F = x + jy$$

where x and y are real, independent, zero-mean Gaussian random variables of equal variance. Thus

$$\langle x \rangle_{\text{av}} = \langle y \rangle_{\text{av}} = 0$$
$$\langle x^2 \rangle_{\text{av}} = \langle y^2 \rangle_{\text{av}} = \sigma^2$$

and since both x and y are Gaussian distributed, their independence is implied by

$$\langle xy \rangle_{\text{av}} = 0.$$

The theory in the main text is concerned with finding the covariance $\langle F_1^* F_2 \rangle_{\text{av}}$ of two such Gaussian fields, where F_1 and F_2 may be two field components separated in space, in time, in frequency, or in all three. Thus

$$F_1 = x_1 + jy_1$$
$$F_2 = x_2 + jy_2$$

and

$$R_F = \langle F_1^*F_2\rangle_{av} = \langle x_1x_2\rangle_{av} + \langle y_1y_2\rangle_{av} + j(\langle x_1y_2\rangle_{av} - \langle x_2y_1\rangle_{av}).$$

If, as is most often the case, all real parts are uncorrelated with all imaginary parts,

$$\langle x_1y_2\rangle_{av} = \langle x_2y_1\rangle_{av} = 0$$

and

$$R_F = \langle F_1^*F_2\rangle_{av} = \langle F_1F_2^*\rangle_{av} = \langle x_1x_2\rangle_{av} + \langle y_1y_2\rangle_{av} \tag{48}$$

is wholly real.

In practice it is not possible to measure the correlation of the complex fields. But what can be measured is the correlation of their magnitudes (that is, envelopes)

$$A = |F| = \sqrt{x^2 + y^2}$$

and the correlation of their squared magnitudes (that is, energies)

$$A^2 = |F|^2 = FF^* = x^2 + y^2.$$

The relation between the autocovariance functions R_F, R_A, and R_{A^2} is as follows.

Consider first the autocovariance function for squared magnitude

$$\begin{aligned} R_{A^2} &= \langle |F_1|^2 \, |F_2|^2\rangle_{av} = \langle F_1F_1^*F_2F_2^*\rangle_{av} \\ &= \langle x_1^2x_2^2\rangle_{av} + \langle y_1^2y_2^2\rangle_{av} + \langle x_1^2y_2^2\rangle_{av} + \langle x_2^2y_1^2\rangle_{av} . \end{aligned}$$

To evaluate the right-hand side one may use the result that if $x_1, \ldots, x_4$ are real, zero-mean Gaussian random variables (see Ref. 3, p. 168),

$$\langle x_1x_2x_3x_4\rangle_{av} = \langle x_1x_2\rangle_{av}\langle x_3x_4\rangle_{av} + \langle x_1x_3\rangle_{av}\langle x_2x_4\rangle_{av} + \langle x_1x_4\rangle_{av}\langle x_2x_3\rangle_{av} .$$

Then, typically,

$$\langle x_1^2x_2^2\rangle_{av} = \langle x_1x_1x_2x_2\rangle_{av} = \sigma^4 + 2(\langle x_1x_2\rangle_{av})^2$$

and

$$\langle x_1^2y_2^2\rangle_{av} = \langle x_1x_1y_1y_1\rangle_{av} = \sigma^4$$

so that

$$R_{A^2} = 4\sigma^4 + 2[(\langle x_1x_2\rangle_{av})^2 + (\langle y_1y_2\rangle_{av})^2]. \tag{49}$$

Now, in most cases

$$\langle x_1x_2\rangle_{av} = \langle y_1y_2\rangle_{av} . \tag{50}$$

For example (48) and (50) can be shown to follow if F_1 and F_2 are the same field component, but do not hold if F_1 is E_z and F_2 is H_z'. Then equations (48) and (49) combined give

$$R_{A^2} = 4\sigma^4 + R_F^2, \tag{51}$$

or from equations (48) and (50)

$$R_{A^2} = 4\sigma^4(1 + \rho^2), \tag{52}$$

where ρ is the normalized autocovariance function of the x and y random processes.

The corresponding result for the autocovariance function of the magnitudes (see p. 59 of Ref. 13) is

$$\begin{aligned} R_A &= \langle A_1 A_2 \rangle_{\text{av}} = \langle |F_1|\ |F_2| \rangle_{\text{av}} \\ &= \sigma^2[2E(\rho) - (1 - \rho^2)K(\rho)], \end{aligned} \tag{53}$$

where K and E are the complete elliptic integrals of the first and second kind. In series form

$$R_A = \frac{\pi}{2}\sigma^2(1 + \rho^2/4 + \rho^4/64 + \cdots) \tag{54}$$

so that to a good approximation, neglecting powers of ρ higher than the second,

$$R_A \cong \frac{\pi}{2}\sigma^2(1 + \rho^2/4), \tag{55}$$

which has the same form as equation (52).

Finally, in terms of the field autocovariance function,

$$R_A \cong \frac{\pi}{2}\sigma^2\left(1 + \frac{R_F^2}{16\sigma^4}\right). \tag{56}$$

Both autocovariance functions R_{A^2} and R_A take on a much simpler form when normalized in the following way. Define the normalized autocovariance function of the departure δA^2 of the squared magnitude A^2 from its mean as

$$\rho_{\delta A^2} = \frac{\langle (A_1^2 - \overline{A_1^2})(A_2^2 - \overline{A_2^2}) \rangle_{\text{av}}}{\sqrt{\langle (A_1^2 - \overline{A_1^2})^2 \rangle_{\text{av}} \langle (A_2^2 - \overline{A_2^2})^2 \rangle_{\text{av}}}}. \tag{57}$$

Then from equation (52)

$$\rho_{\delta A^2} = \rho^2. \tag{58}$$

Defining the normalized autocovariance function of the departure δA of the magnitude A from its mean in a similar manner, equation (54) gives

$$\rho_{\delta A} = \frac{\pi}{4(4-\pi)} (\rho^2 + \rho^4/16 + \rho^6/64 + \cdots), \tag{59}$$

or to a good approximation

$$\rho_{\delta A} \simeq \rho^2. \tag{60}$$

Equations (48) and (50) show that ρ is the normalized form of the autocovariance function R_F of the complex field component F.

APPENDIX C

Derivation of Equation 26

The complex amplitude of the received signal appearing at the antenna terminals may be written in the form

$$v = E_0 \sum_{n=0}^{N} a(\alpha_n) \exp \{ j\varphi_n \}$$

where E_0 is the common amplitude of the N azimuthal plane waves incident on the mobile receiving antenna. The phase of each wave is φ_n, and $a(\alpha)$ is the voltage response at the antenna terminals owing to a unit-amplitude plane wave arriving at the azimuthal angle α. At another point a distance ξ away (see Fig. 1) the signal at the antenna terminals would be

$$v' = E_0 \sum_{m=1}^{N} a(\alpha_m) \exp \{ j(\varphi_m + k\xi \cos \alpha_m) \}.$$

Forming the complex product v^*v' and taking its expected value to yield the spatial autocovariance function of the two signals, namely

$$\begin{aligned} R_v(\xi) &= \langle v^*v' \rangle_{\text{av}} \\ &= | E_0 |^2 \sum_{n=1}^{N} \sum_{m=1}^{N} \langle a^*(\alpha_n) a(\alpha_m) \exp \{ jk\xi \cos \alpha_m \} \rangle_{\text{av}} \cdot \langle \exp \{ j(\varphi_m - \varphi_n) \} \rangle_{\text{av}} \end{aligned}$$

where it has been assumed that the phases and angles of arrival of the component waves are independent. Making the further assumption that the phases are equiprobable throughout the range 0 to 2π,

$$R_v(\xi) = N \, | E_0 |^2 \int_{-\pi}^{+\pi} p(\alpha) g(\alpha) \exp \{ jk\xi \cos \alpha \} \, d\alpha \tag{61}$$

where $p(\alpha)$ is the probability density function of the component plane waves, and

$$g(\alpha) = a^*(\alpha)a(\alpha) = |a(\alpha)|^2$$

is the azimuthal power gain function of the antenna.

The temporal autocovariance function of v can be derived from equation (61) for a receiver moving with constant velocity V by making the substitution $\xi = V\tau$, where τ is a displacement in time. Then

$$R_v(\tau) = \int_{-\pi}^{+\pi} p(\alpha)g(\alpha) \exp \{j\omega_m\tau \cos \alpha\} \, d\alpha \tag{62}$$

where $\omega_m = 2\pi f_m$ with $f_m = V/\lambda$ the maximum Doppler shift, and $N |E_0|^2$ has been set equal to unity. The spectrum of the signal at the antenna terminals is given by the Fourier transform of the temporal autocovariance function of equation (62) and is

$$\begin{aligned} S_v(f) &= \int_{-\infty}^{\infty} R_v(\tau) \exp \{-j2\pi f\tau\} \, d\tau \\ &= \int_{-\infty}^{\infty} d\tau \int_{-\pi}^{+\pi} d\alpha p(\alpha)g(\alpha) \exp \{j(\omega_m \cos \alpha - 2\pi f)\tau\} \end{aligned} \tag{63}$$

where $f = \omega/2\pi$ is the shift in frequency from the carrier frequency.

Reversing the order of integration in equation (63), the integration w.r.t. τ yields a Dirac δ-function, thus

$$S_v(f) = \int_{-\pi}^{+\pi} p(\alpha)g(\alpha) \; \delta(f_m \cos \alpha - f) \, d\alpha. \tag{64}$$

Now writing

$$h(\alpha) = f_m \cos \alpha - f \tag{65}$$

it may be noticed that the δ-function of a function may be written in the form[19]

$$\delta[h(\alpha)] = \sum_n \frac{\delta(\alpha - \alpha_n)}{| h'(\alpha_n) |} \tag{66}$$

where the α_n are all the values of α for which $h(\alpha) = 0$, and the prime denotes differentiation w.r.t. α. Hence, from equations (64), (65), and (66) the spectrum of the signal at the antenna terminals is

$$S_v(f) = \frac{1}{f_m \sqrt{1 - f^2/f_m^2}} \cdot \{p(\alpha)g(\alpha) \,|_{\alpha = \cos^{-1}(f/f_m)} + p(\alpha)g(\alpha) \,|_{\alpha = -\cos^{-1}(f/f_m)}\}$$

which is equation (26). Notice that since the angle of arrival α must be real, the frequency shift f must lie in the range $\pm f_m$.

APPENDIX D

Random Frequency Modulation of the Carrier

Since frequency modulation is often used in mobile radio systems it is pertinent to inquire what will be the nature of the received audio signal when a single unmodulated frequency is transmitted. The phase of the received signal is changing with time in a random manner; hence its instantaneous frequency is random.

It has been shown,[20] based on the work of Rice,[4] that the p.d.f. of the time-rate of change of phase θ' (the instantaneous frequency) for narrowband Gaussian random noise with an amplitude spectrum which is symmetrical about the carrier frequency, is

$$p(\theta') = \frac{1}{2}\left[\frac{b_2}{b_0}\left(1 + \frac{b_0}{b_2}\,\theta'^2\right)^3\right]^{-\frac{1}{2}} \tag{67}$$

where b_0 and b_2 are the zero$^{\text{th}}$ and second moments, respectively, about the carrier frequency of the amplitude spectrum $S(f)$. Notice that it has been assumed that there is no constant sinusoid present in the noise. It has also been shown[20] that the conditional p.d.f. $p(\theta'|r)$, which is the density of the instantaneous frequency given that the normalized envelope r is a certain value, is

$$p(\theta' \mid r) = \frac{1}{\sqrt{2\pi}\sigma'_\theta} \exp\left\{-\frac{\theta'^2}{2\sigma'^2_\theta}\right\} \tag{68}$$

which is a Gaussian distribution with zero mean and standard deviation

$$\sigma'_\theta = \frac{1}{r}\sqrt{\frac{b_2}{2b_0}}. \tag{69}$$

The above equations can be applied to the case of a mobile radio signal derived from an omnidirectional antenna in a uniformly scattered field.

The appropriate amplitude specturm is that of equation (27) and yields the moments,

$$b_0 = \int_{-\infty}^{\infty} S(f)\, df = 1 \tag{70}$$

and

$$b_2 = (2\pi)^2 \int_{-\infty}^{\infty} f^2 S(f)\, df = (1/2)\omega_m^2 \tag{71}$$

where ω_m is the maximum Doppler frequency shift in radians per second. Equations (67) and (68) then become

$$p(\theta') = \left[2\omega_m^2 \left(1 + \frac{2\theta'^2}{\omega_m^2} \right)^3 \right]^{-\frac{1}{2}} \tag{72}$$

and

$$p(\theta' \mid r) = \frac{1}{\sqrt{2\pi}\sigma_\theta'} \exp \left\{ -\frac{\theta'^2}{2\sigma_\theta'^2} \right\} \tag{73}$$

with

$$\sigma_\theta' = (1/2)\, \frac{\omega_m}{r}. \tag{74}$$

The p.d.f. of equation (72) has a rather sharp maximum at $\theta' = 0$, and falls to about 0.2 of this maximum value at $\theta' = \pm\omega_m$. For large instantaneous frequency deviations the p.d.f. behaves asymptotically as the inverse cube of the frequency. In practical terms this p.d.f. is that of the amplitude of the output of a frequency discriminator in the receiver for a single frequency transmitted.

The conditional p.d.f. of equation (73), which is Gaussian in form, can also be interpreted as the p.d.f. of the amplitude of the discriminator output. But this is the p.d.f. of the frequency deviations measured only when the envelope amplitude is in the neighborhood of a particular level r, which is the envelope normalized by its r.m.s. value. In the particular example chosen the envelope has a Rayleigh distribution.

When $r = 1$ the conditional p.d.f. of the frequency deviations has a spread of the order of the maximum Doppler frequency shift ω_m. The spread will be 10 ω_m when $r = \frac{1}{10}$, the probability that $r \leqq \frac{1}{10}$ being 0.01. Similarly the spread will be 100 ω_m when $r = \frac{1}{100}$, the probability that $r \leqq \frac{1}{100}$ being 0.0001. Thus the wider ranges of random-frequency excursion are associated with only very small fractions of the total time.

REFERENCES

1. Ossanna, J. F., Jr., "A Model for Mobile Radio Fading Due to Building Reflections: Theoretical and Experimental Fading Waveform Power Spectra," B.S.T.J., *43*, No. 6 (November 1964), pp. 2935–2971.
2. Gilbert, E. N., "Energy Reception for Mobile Radio," B.S.T.J., *44*, No. 8 (October 1965), pp. 1779–1803.

3. Davenport, W. B. and Root, W. L., *An Introduction to the Theory of Random Signals and Noise,* New York: McGraw-Hill, 1958, p. 153.
4. Rice, S. O., "Statistical Properties of a Sine Wave Plus Random Noise," B.S.T.J., *27,* No. 1 (January 1948), pp. 109–157.
5. Young, W. R., Jr., "Comparison of Mobile Radio Transmission at 150, 450, 900, and 3700 Mc.," B.S.T.J., *31,* No. 6 (November 1952), pp. 1068–1085.
6. Trifonov, P. M., Budko, V. N., and Zotov, V. S., "Structure of USW Field-Strength Spatial Fluctuations in a City," (English translation from the Russian) Trans. Telecommunications Radio Eng., *9,* Pt. 1 (February 1964), pp. 26–30.
7. Jakes, W. C., Jr., and Reudink, D. O., "Comparison of Mobile Radio Transmission at UHF and X Band," IEEE Trans. Vehicular Technology, *VT-16* (October 1967), pp. 10–14.
8. Rustako, A. J., Jr., "Evaluation of a Mobile Radio Multiple Channel Diversity Receiver Using Pre-Detection Combining," IEEE Trans. Vehicular Technology, *VT-16* (October 1967), pp. 46–57.
9. Brennan, D. G., "Linear Diversity Combining Techniques," Proc. IRE, *47* (June 1959), pp. 1075–1102.
10. Jakes, W. C., Jr., unpublished work.
11. Lee, W. C.-Y., "Theoretical and Experimental Study of the Properties of the Signal from an Energy Density Mobile Radio Antenna," IEEE Trans. Vehicular Technology, *VT-16* (October 1967), pp. 25–32.
12. 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.
13. Lawson, J. L. and Uhlenbeck, G. E., "Threshold Signals," Vol. 24 of MIT Radiation Laboratory Series, New York: McGraw-Hill, 1950, p. 63.
14. Lee, W. C.-Y., "Preliminary Investigation of Mobile Radio Signal Fading Using Directional Antennas on the Mobile Unit," IEEE Trans. Vehicular Comm., *VC-15* (October 1966), pp. 8–15.
15. Young, W. R., Jr. and Lacy, L. Y., "Echoes in Transmission at 450 Megacycles from Land-to-Car Radio Units," Proceedings of IRE, *38,* No. 3 (March 1950), pp. 255–258.
16. Ossanna, J. F., unpublished work.
17. Wilk, M. B., and Ghanadesikan, R., "Probability Plotting Methods for the Analysis of Data," Biometrika, *55,* part 1 (March 1968), pp. 1–19.
18. Tatarski, V. I., *Wave Propagation in a Turbulent Medium,* trans. R. A. Silverman, New York: McGraw-Hill, 1961, Chapter 1.
19. Zadeh, L. A. and Desoer, C. A., *Linear System Theory: The State Space Approach,* New York: McGraw-Hill, 1963, p. 533.
20. R.C.A., Defense Electronics Products, Surface Communications Systems Laboratory, final quarterly and summary report: *UNICOM Long Range Radio Circuits,* prepared under contract to Bell Telephone Laboratories, June 15, 1962.

Some Transmission Characteristics of Bell System Toll Connections

By I. NÅSELL

(Manuscript received January 10, 1968)

A systemwide survey of the transmission performance of built-up toll connections was undertaken in 1966. The sampling plan underlying this survey is discussed briefly. The results are presented in terms of distributions of background noise levels, 1000 Hz loss, phase jitter, time to connect, and airline distance between end offices. The measurement results are broken down by mileage categories. Comparisons are made with the results from the 1962 connection survey. It is found that noise performance has improved since 1962 while loss performance is virtually unchanged.

I. INTRODUCTION

Many systems engineering studies require detailed knowledge about transmission performance and transmission capabilities of the Bell System plant. The need for such information exists both for specific parts or building blocks of the network and for built-up connections between subscribers. A system-wide survey of noise and loss on toll connections was undertaken in 1962.[1] The results of this survey found an important application in the setting of new over-all objectives for background noise.[2]

A similar survey was undertaken in the summer of 1966. It is our purpose to describe this connection survey and to give its results. Present transmission performance of built-up toll connections is given in terms of distributions of noise, loss, and phase jitter. Furthermore, the results include distributions of time to connect, and the distribution of airline distances between end offices of toll calls as presently established by customers.

Connection results discussed in this paper describe the toll plant contribution to the transmission performance on built-up toll connections. In considering complete toll connections from subscriber to subscriber, the influence of the loop plant must also be taken into account. Some of its characteristics have been described by Hinderliter.[3]

II. TARGET POPULATION

The target population is the population about which information is desired. It was defined as the set of all toll calls made in the Bell System during the busy period (9 a.m. to 5 p.m.) of an ordinary business day. A call was considered a toll call if it satisfied the following two conditions: (*i*) the customer received a bill which included a separate charge for the call, and (*ii*) the originating and terminating central offices did not home on the same toll office. The first criterion assures us that the population contains only completed messages rather than call attempts, while the second criterion means that with some minor exceptions the toll calls included in the population require at least one intertoll trunk for their completion.

The main difference between the population defined here and the population defined for the 1962 survey lies in the extension from the busy hour used in 1962 to the busy period. This extension provides for a more satisfactory reflection in the population of the traffic patterns generated by telephone subscribers. For example, cross-continental calls originating on the U. S. east coast were under-represented in the 1962 survey because of the different time zones on east and west coasts. Such under-representation does not exist in the 1966 survey.

III. SAMPLING PLAN AND SAMPLE SIZE

The sampling plan can be described as a two-stage plan with primary stratification and substratification and with the primary units selected with probabilities proportional to measures of size.[4, 5] The primary units were identified with Bell System end-office buildings. Two primary strata were defined, based on the size of the primary units. One of these strata contains those buildings in which at least 400,000 toll messages originate annually; the other contains the remaining smaller buildings.

The first-stage sample contains 40 end-office buildings. Twenty-five of these were selected from the stratum with large offices, and fifteen from the small offices stratum. The sample units in the two strata were selected independently from lists that contained the total of 9052 Bell System end-office buildings that were in service on January 1, 1964.

For each of the selected primary units, information was acquired about the outgoing toll traffic during the busy period of an ordinary business day. This information consisted of lists of terminating endpoints of toll calls originating in the sample office during the indicated time period. Every call in each of these lists was assigned to one of three substrata. The substratification was based on the airline distance between originating and terminating end offices. Toll calls shorter than about 180 miles were assigned to substratum one, while calls longer than about 725 miles were assigned to substratum three.

Independent selections of sample elements were made in each of the substrata for each sampled primary unit. The aim of the substratification was to achieve a sample size that would give acceptable precision in the estimation of transmission performance for toll calls in each of a number of mileage categories. The success of this endeavor is demonstrated by the confidence interval widths listed in the various tables of Section V.

An approximately equal number of toll calls was selected into the sample in each sample office. The resulting sample is not self-weighting. This means that different sample toll calls in general carry different weights in the estimation of population characteristics. The sample contains a total of 1463 calls. Of these, 476 have an airline distance between end offices up to 180 miles, while 554 are between 180 and 725 miles long, and 433 calls are longer than 725 miles.

IV. METHOD OF MEASUREMENT

The measurement procedure in the survey was similar to that used in 1962. Thus, the aim of the measurement phase was to duplicate the calls included in the sample and make transmission measurements in the receive direction on the established connections. In addition, the time required to establish the connection was noted.

All survey connections were established from an ordinary telephone set connected via a test set to a zero loop in the originating central office. The test set consisted of coils and switches and allowed the telephone set to be switched out of the connection and be conveniently replaced by a suitable measurement instrument. This test set and the transmission measuring equipment used in the survey are manufactured by the Western Electric Company for Bell System use only.

Two separate connections were established for each call included in the sample. One of them was made to the balanced (quiet) termination in the distant central office, and the other was made to the far-end milliwatt supply. The first one allowed the measurement of noise on the connection. The 3A noise measuring set[6] was used, and two readings were taken: one with C-message weighting and the other with 3 kHz flat weighting. As in the 1962 connection survey, no information about the physical routing of the call was acquired, and the measured noise levels did not include the subjective penalty due to the possible presence of compandored carrier facilities in the connection.

The second connection was established to record the 1000 Hz loss. The received level was measured with a transmission measuring set and recorded to the nearest tenth of one dB. The peak-to-peak phase jitter of the received signal was measured on the same connection with a voiceband phase jitter meter. The calls to the milliwatt supplies were also used to acquire information about time to connect. This time was measured as the time elapsed after the last digit had been dialed or after the conversation with the operator was finished until the test tone or a ringback signal was heard.

All of the terminating end offices for the sample calls were not equipped with balanced terminations or milliwatt supplies. In order to allow measurements to be made, such sample calls were replaced by calls that terminated in an end office geographically close to the desired one, and equipped with proper test lines. Replacements of this type were made on somewhat less than 10 per cent of the sample calls.

V. SURVEY RESULTS

The survey results presented here have all been evaluated by computer programs based on sample survey evaluation formulas contained in Ref. 4. The transmission results give noise, loss, and phase jitter as measured across a 900Ω termination on a zero length loop.

5.1 *3A Noise with C-Message Weighting*

A scatter diagram showing observed 3A noise levels with C-message weighting as a function of the airline distance between end offices is contained in Fig. 1. The previously observed[1] general trends of increasing mean and decreasing standard deviation as the call distance is increased is visible from this figure. These trends are ex-

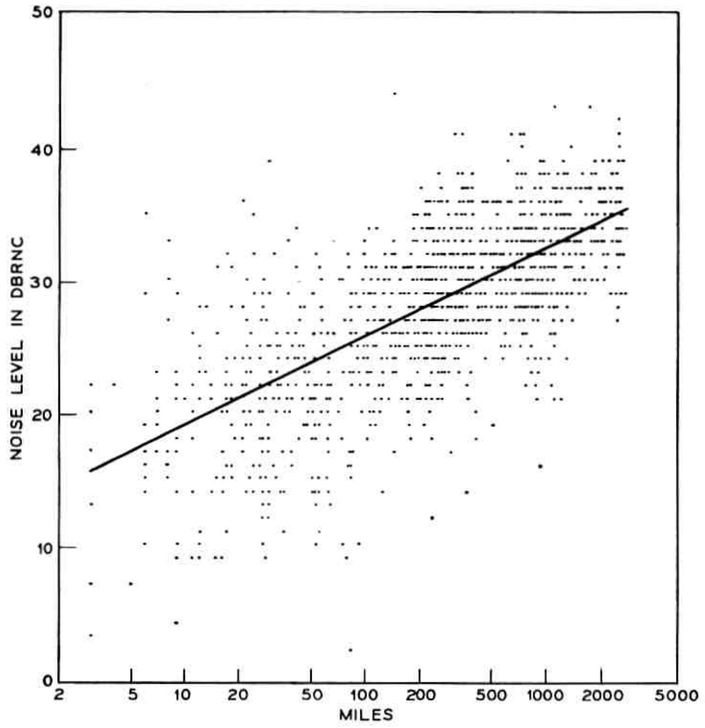


Fig. 1 — Scatter diagram of 3A noise level (C-message weighting) vs airline distance.

plained qualitatively by reference to the theory of power sums of random variables. The noise level on a toll connection can be regarded as the power sum of noise levels from a number of different noise sources, and with the number of noise sources increasing with call distance. Recent results by Marlow[7] and Nåsell[8] show that the mean of a power sum increases with the number of components, while the standard deviation of the power sum decreases as the number of components is increased, in line with the trends observed in Fig. 1.

The regression line in Fig. 1 gives an estimate of the mean noise level under the assumption that the mean noise level is linearly related to the logarithm of the airline distance between end offices. The equation for the regression line is

$$N = 12.6 + 2.0 \log_2 D \tag{1}$$

where D is the airline distance between end offices in miles, and N is the average 3A noise level. This equation shows that the average noise level increases by 2.0 dB for each doubling of the airline distance between end offices. The fact that the variance changes with distance has been accounted for in the regression analysis; weights were applied in inverse proportion to the variance about the regression line.

A summary of the results for 3A noise levels with C-message weighting is contained in Table I. As in most tables in this section, estimates are given of the mean and the standard deviation of the population distribution, and the mean is equipped with its 90 per cent confidence interval. Table I gives such results for each of eight mileage categories. These categories (except the first) are one double distance wide. The first four taken together correspond to the category referred to as "short" (0–180 miles) by D. A. Lewinski,[2] the next two cover the "medium" length and the last two contain the "long" calls (longer than 725 miles). The tendency for the mean to increase, and the standard deviation to decrease with distance is clearly demonstrated in this table.

The noise distributions discussed here are all very close to normal. No significant difference was found between mean noise levels on operator-handled calls and mean noise levels on direct-dialed calls.

A comparison between noise level distributions observed in the 1962 and the 1966 connection surveys is made in Table II. The table indicates improved noise performance of the toll plant in the intervening period; both means and standard deviations show generally lower values in 1966, and the difference between means in the long category is statistically significant. The results given for the 1962

TABLE I—SUMMARY OF RESULTS FOR 3A NOISE LEVELS WITH C-MESSAGE WEIGHTING

Airline distance (miles)	Mean dBrnC	Std. dev. (dB)
0–23	19.8 ± 1.0	6.2
23–45	21.9 ± 1.7	6.5
45–90	22.4 ± 1.6	6.1
90–180	25.3 ± 1.4	5.3
180–360	28.9 ± 1.0	4.3
360–725	31.0 ± 0.8	3.6
725–1450	31.1 ± 1.3	4.2
1450–2900	34.6 ± 0.9	3.1

TABLE II — COMPARISON OF RESULTS FOR 3A NOISE WITH C-MESSAGE WEIGHTING FROM THE 1962 AND 1966 SURVEYS

Airline distance (miles)	1962 Survey		1966 Survey	
	Mean dBrnC	Std. dev. (dB)	Mean dBrnC	Std. dev. (dB)
0–180	23.4 ± 2.6	7.4	21.6 ± 0.8	6.4
180–725	31.0 ± 1.2	5.3	29.6 ± 0.7	4.2
725–2900	35.8 ± 1.5	4.0	32.5 ± 1.0	4.1

survey deviate slightly from those quoted by Lewinski.[2] The reason is that Lewinski's numbers are based on a sub-sample, while the results in Table II are not. The differences are well within the confidence intervals.

Table II also illustrates the improved precision achieved in the 1966 survey compared with the precision of the 1962 survey.

5.2 *3A Noise with 3 kHz Flat Weighting*

A scatter diagram of 3A noise levels with 3 kHz flat weighting as a function of the airline distance between end offices is shown in Fig. 2. It indicates much less of a distance dependence of the observed noise levels than that shown in Fig. 1. This is to be expected since flat weighted noise readings are predominantly caused by low-frequency noise components that fall below the lower cutoff frequency of most carrier facilities used in the toll plant.

A summary of the results for 3A noise with 3 kHz flat weighting is given in Table III. The table reinforces the impression that the distance dependence of both mean and standard deviation is very slight. It does, however, bring out the fact that both means and standard deviations of operator-handled calls are larger than those for direct-dialed calls. This fact is believed to be related to differences in local trunking arrangements. All of the distributions of flat weighted noise levels have a moderate amount of positive skewness.

5.3 *1000 Hz Loss*

The end-office to end-office loss at 1000 Hz is shown as a function of distance in the scatter diagram of Fig. 3. Just as was the case in the 1962 survey, we find the distance dependence of the loss to be only moderate. Table IV summarizes the results for each of the eight mileage categories discussed above. A small trend for both mean and standard deviation to increase with distance is seen to exist.

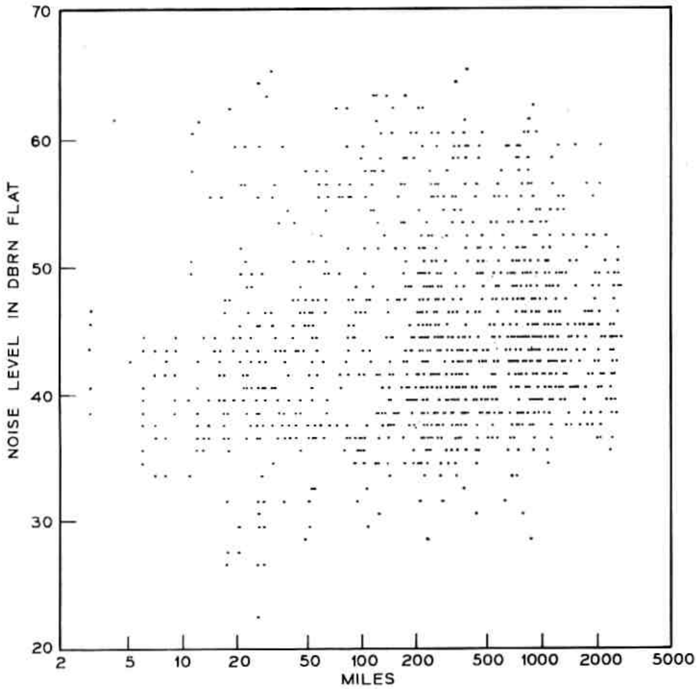


Fig. 2 — Scatter diagram of 3A noise level (3 kHz flat weighting) vs airline distance.

This is related to the higher probability of encountering more than one intertoll trunk in tandem for the longer connections. All the loss distributions deviate somewhat from normality through a moderate amount of positive skewness. Loss values exceeding 20 dB were found both on operator-handled and on direct-dialed calls.

Operator-handled calls will in general require one more trunk for

Table III — Summary of Results for 3A Noise with 3kHz Flat Weighting

Airline distance (miles)	Over-all		Operator		DDD	
	Mean dBrn (3kHz flat)	Std. dev. (dB)	Mean dBrn (3kHz flat)	Std. dev. (dB)	Mean dBrn (3kHz flat)	Std. dev. (dB)
0–180	43.9 ± 1.6	7.4	46.7 ± 3.1	9.1	42.5 ± 1.5	5.8
180–725	45.9 ± 2.4	7.6	47.8 ± 4.0	8.8	43.6 ± 1.4	5.2
725–2900	45.2 ± 1.5	6.0	46.5 ± 2.5	7.0	43.9 ± 1.1	4.2

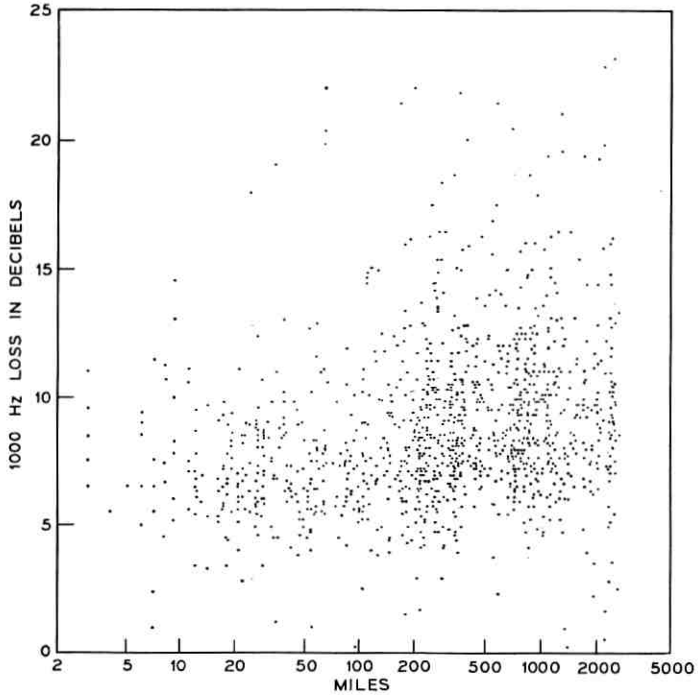


Fig. 3 — Scatter diagram of 1000 Hz loss vs airline distance.

TABLE IV—SUMMARY OF RESULTS FOR END-OFFICE TO END-OFFICE LOSS AT 1000 Hz

Airline distance (miles)	Mean (dB)	Std. dev. (dB)
0–23	6.8 ± 0.6	2.4
23–45	7.7 ± 0.5	2.6
45–90	7.1 ± 0.7	2.6
90–180	7.4 ± 0.6	2.8
180–360	8.7 ± 0.6	2.8
360–725	9.4 ± 1.0	2.9
725–1450	9.5 ± 0.4	2.9
1450–2900	9.7 ± 0.8	3.0

TABLE V — COMPARISON OF LOSS DISTRIBUTIONS FOR OPERATOR-HANDLED AND DIRECT-DIALED CALLS

Airline distance (miles)	Operator		DDD	
	Mean (dB)	Std. dev. (dB)	Mean (dB)	Std. dev. (dB)
0–180	7.5 ± 0.6	3.0	7.0 ± 0.4	2.3
180–725	9.3 ± 0.8	3.1	8.5 ± 0.6	2.5
725–2900	10.2 ± 0.6	2.7	8.9 ± 0.6	3.0

their completion than direct-dialed calls. The total loss on the connection is, therefore, expected to be somewhat higher on operator-handled than on direct-dialed calls. A comparison between the loss distribution parameters on the two types of calls is made in Table V. The table shows a lower mean loss on DDD calls in each of the three mileage categories, and in the third category the difference is significant. The mean loss difference is seen to range from 0.5 dB for short calls to 1.3 dB for long calls. No rationale is known for a distance dependence of this loss difference.

A comparison of means and standard deviations of loss distributions observed in the 1962 and 1966 surveys is made in Table VI. No large changes in the intervening time period are indicated.

5.4 *Phase Jitter*

The phase jitter measurements in the survey reveal the amount of phase modulation that an unmodulated sinusoidal carrier of 1000 Hz is subjected to on a toll connection. These measurements were included since certain types of data transmission are susceptible to phase modulation of transmitted signals. The measurements give the peak-to-peak phase jitter in degrees for jitter components between 10 Hz and 120 Hz on the signal transmitted by the far-end

TABLE VI — COMPARISON OF LOSS DISTRIBUTIONS FROM THE 1962 AND 1966 SURVEYS

	1962 Survey		1966 Survey	
Airline distance (miles)	Mean (dB)	Std. dev. (dB)	Mean (dB)	Std. dev. (dB)
0–180	7.3 ± 0.6	2.8	7.2 ± 0.4	2.6
180–725	8.9 ± 0.7	3.0	8.9 ± 0.7	2.9
725–2900	9.3 ± 1.4	3.8	9.6 ± 0.5	2.9

1000 Hz milliwatt supply. A scatter diagram of observed phase jitter versus connection distance is contained in Fig. 4. The connections for which a phase jitter of 21 degrees is indicated are connections where the phase jitter measurement was larger than or equal to 21 degrees. A trend for the average phase jitter to increase with mileage is indicated by the figure. The phase jitter distributions are definitely not normal with a high amount of positive skewness. Because of this, the summary data in Table VII give 10-, 50-, and 90-percent points of the phase jitter distributions rather than means and standard deviations.

Operator-handled calls that are of short and medium length show a significantly higher median phase jitter than direct-dialed calls of

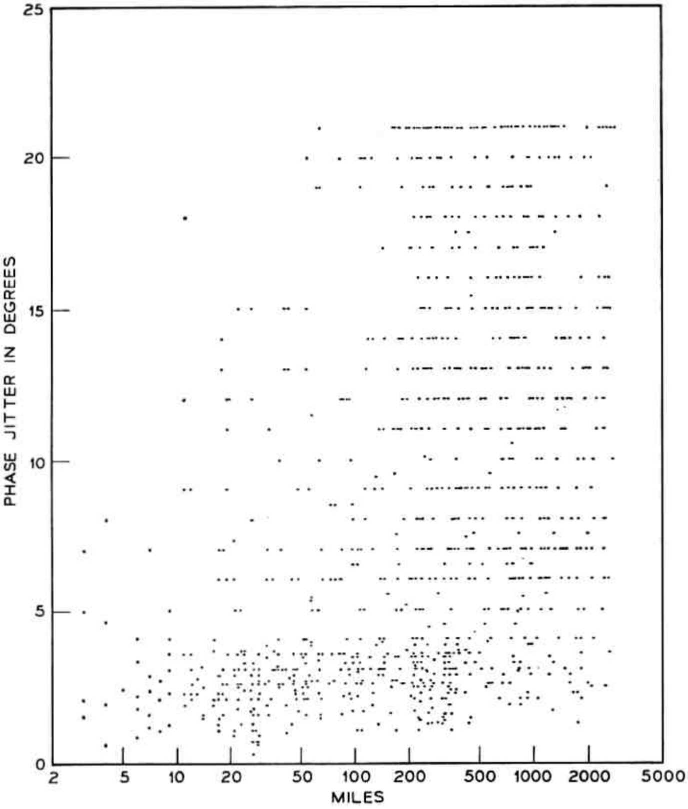


Fig. 4 — Scatter diagram of phase jitter vs airline distance.

TABLE VII — SUMMARY OF RESULTS FOR PEAK-TO-PEAK PHASE JITTER*

Airline distance (miles)	Phase jitter (degrees)		
	10%	50%	90%
0–23	1	3	7
23–45	1	3	7
45–90	2	4	15
90–180	2	7	14
180–360	2	7	20
360–725	2	11	21
725–1450	4	12	20
1450–2900	3	12	21

* The table gives the 10-, 50-, and 90-per-cent points (in degrees) of the phase jitter distributions in each mileage category.

corresponding length, while no apparent difference exists for long calls. A numerical comparison is made in Table VIII.

5.5 *Time to Connect*

The time to connect is shown versus distance in the scatter diagram of Fig. 5. A range up to 100 seconds is used to cover some operator-handled calls that suffered long delays. The scatter diagram shows a tendency for the average time to connect to increase with distance. This is a reflection of the higher average number of intertoll trunks in tandem for the longer connections, which in turn means that a larger number of switching offices is involved in establishing the longer connections.

A separation of operator-handled calls from direct-dialed calls is made in Table IX. It shows that the average time to connect is longer for operator-handled calls than for direct-dialed calls. It also

TABLE VIII — PEAK-TO-PEAK PHASE JITTER FOR OPERATOR-HANDLED AND DIRECT-DIALED CALLS*

Airline distance (miles)	Phase jitter (degrees)					
	Operator			DDD		
	10%	50%	90%	10%	50%	90%
0–180	2	4	11	1	2	9
180–725	3	11	21	2	6	18
725–2900	5	11	20	3	12	21

* The table gives the 10-, 50-, and 90-per-cent points (in degrees) of the phase jitter distributions in each mileage category.

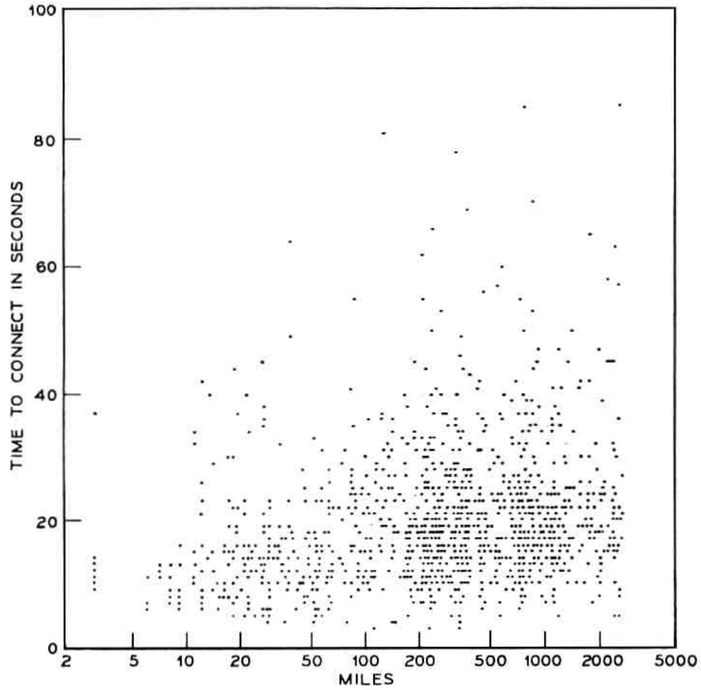


Fig. 5 — Scatter diagram of time to connect vs airline distance.

shows that the average time to connect is virtually independent of distance for operator-handled calls, while a definite trend exists for direct-dialed calls. Finally, we notice that the standard deviations are considerably higher for the operator-handled calls than for those that are direct-dialed. For these reasons, a detailed study of the time to connect for direct-dialed calls is of interest.

Table IX — Comparison of Distributions of Time to Connect for Operator-Handled and DDD Calls

Airline distance (miles)	Time (seconds)			
	Operator		DDD	
	Mean	Std. dev.	Mean	Std. dev.
0–180	24.7 ± 4.2	21.1	11.1 ± 0.9	4.6
180–725	27.0 ± 4.5	20.5	15.6 ± 1.0	5.0
725–2900	24.8 ± 2.4	11.1	17.6 ± 2.1	6.6

A scatter diagram of time to connect versus distance is given for DDD calls in Fig. 6. The regression line shown has the equation

$$T = 7.6 + 0.9 \log_2 D \tag{2}$$

where D is the airline distance between end offices in miles, and T is the average time to connect in seconds. The regression equation shows that the average time to connect increases by 0.9 seconds for each doubling of the airline distance between end-offices.

A summary of the parameters of time to connect distributions for DDD calls is given in Table X. The table indicates that the regression assumption of a linear relation between the mean time to connect and the logarithm of the airline distance may be an oversimplification; the mean time to connect is virtually constant in the first three and in the last two mileage categories; in between it increases by more than 0.9 seconds per double distance.

The distributions of time to connect over all calls have a high positive skewness as indicated by the scatter diagram in Fig. 5. On the other hand, only a small amount of skewness is present in the

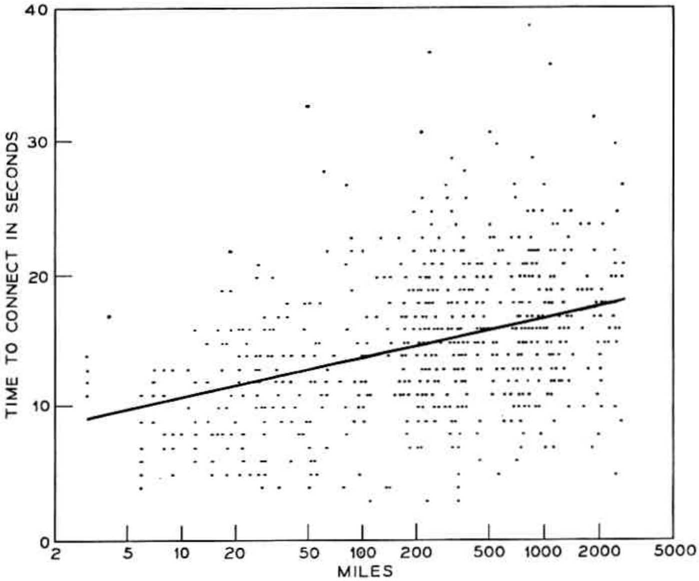


Fig. 6 — Scatter diagram of time to connect on direct-dialed calls vs airline distance.

Table X—Summary of Results for Time to Connect on DDD Calls

Airline distance (miles)	Time (seconds)	
	Mean	Std. dev.
0–23	10.7 ± 1.2	4.6
23–45	11.6 ± 1.2	4.2
45–90	11.2 ± 1.8	4.8
90–180	12.3 ± 3.0	5.2
180–360	15.0 ± 1.0	4.6
360–725	16.8 ± 1.5	5.5
725–1450	17.8 ± 3.1	7.6
1450–2900	17.4 ± 1.1	4.4

distributions for DDD calls, as seen from the scatter diagram in Fig. 6.

5.6 *Distance Distribution*

The distribution of airline distances between end offices of toll calls is given in Fig. 7. The distribution is seen to deviate somewhat from a log-normal distribution, and it is virtually truncated at 2500 miles. Table XI gives estimated percentages of toll calls that fall in each of the eight mileage categories. A comparison with the results from the 1962 survey shows no important changes. The fact that only about four per cent of all toll calls are longer than 725 miles illustrates a problem for the design of the sampling plan. Unstratified sampling would tend to give a sample in which only about four per cent of the sample calls exceed 725 miles in length. In contrast to this, precision requirements dictate approximately equal sample size for short, medium, and long calls. The problem was solved, as mentioned before, by the use of substratification based on the airline distance between end offices of toll calls.

VI. CONCLUDING REMARKS

The 1966 connection survey represents an improvement over the 1962 survey in terms of precision. It also represents a small extension of the measurement program, to include measurements of such entities as phase jitter and time to connect. It does, however, suffer from certain limitations, which it shares with the 1962 survey. Most important is the fact that a number of important transmission parameters, such as frequency response, delay distortion, and impulse noise, were not measured. An additional limitation is that the milliwatt

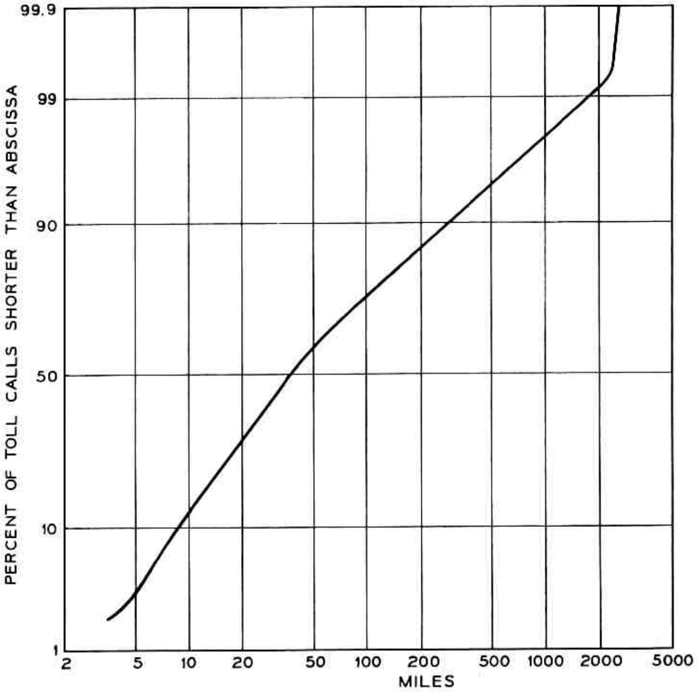


Fig. 7 — Distance distribution of toll calls.

Table XI — Distance Distribution of Toll Calls

Airline distance (miles)	Percent of calls in distance class		
	1966 Survey		1962 Survey
0–23	33.7	83.7	85.0
23–45	20.0		
45–90	18.2		
90–180	11.8		
180–360	8.0	12.4	11.0
360–725	4.4		
725–1450	2.3	3.9	4.0
1450–2900	1.6		

signal source at the far end of each call could not be calibrated.

The use of specially-equipped test teams at both ends of the connections would alleviate both of these limitations. Studies are, therefore, under way to investigate the feasibility of using a 3-stage sampling plan in place of the 2-stage plan that was used in the 1966 survey. The main accomplishment of the 3-stage plan would be to limit the number of far-end end offices involved in the sample connections, thereby reducing the total traveling cost.

A toll connection appraisal program has recently been introduced in the Operating Companies of the Bell System. The procedures of this program are similar to those used in the connection survey described here. However, the main purpose of this appraisal program is to provide data to aid in the location of weak spots and also to aid in managerial decisions affecting the transmission performance of the present plant. In contrast to this, the data collected in the connection survey will find its main application in systems engineering studies conducted at Bell Laboratories and elsewhere in the Bell System.

It might be surprising that a sample of only 1463 calls originating in 40 end offices suffices to estimate the transmission performance of the 15 million toll calls that originate each day in one of more than 9000 end-office buildings. The results presented here show, however, that the achieved precision is indeed acceptable for a number of engineering applications. This fact demonstrates very concretely what can be achieved for data-acquisition purposes by a judicious application of the powerful methods of modern sample survey theory.

VII. ACKNOWLEDGEMENTS

The survey whose results are reported here was executed through the combined efforts of a large number of people. D. T. Osgood of the American Telephone and Telegraph Company served as coordinator between Bell Laboratories and the Operating Companies. Engineering coordinators and plant personnel in each Operating Company represented in the sample contributed to the planning and measuring. The actual measurements were carried out by more than 20 Bell Laboratories employees. Data handling was planned by F. P. Duffy and executed by Mrs. C. A. McFaul. Computer programs for data analysis were written by Miss M. L. Chubb and F. P. Duffy. The inclusion of measurements of time to connect were suggested by C. R. Ellison. I thank them all for their contributions.

REFERENCES

1. Nåsell, I., "The 1962 Survey of Noise and Loss on Toll Connections," B.S.T.J., *43,* No. 2 (March 1964), pp. 697–718.
2. Lewinski, D. A., "A New Objective for Message Circuit Noise," B.S.T.J., *43,* No. 2 (March 1964), pp. 719–740.
3. Hinderliter, R. G., "Transmission Characteristics of Bell System Subscriber Loop Plant," I.E.E.E. Trans. Commun. Elec., *82* (September 1963), pp. 464–470.
4. Hansen, M. H., Hurwitz, W. N., and Madow, W. G., *Sample Survey Methods Theory,* Vols. I and II, New York: John Wiley and Sons, 1953.
5. Cochran, W. G., *Sampling Techniques,* New York: John Wiley and Sons, 1963.
6. Cochran, W. T., and Lewinski, D. A., "A New Measuring Set for Message Circuit Noise," B.S.T.J., *39,* No. 4 (July 1960), pp. 911–931.
7. Marlow, N. A., "A Normal Limit Theorem for Power Sums of Independent Random Variables," B.S.T.J., *46,* No. 9 (November 1967), pp. 2081–2089.
8. Nåsell, I., "Some Properties of Power Sums of Truncated Normal Random Variables," B.S.T.J., *46,* No. 9 (November 1967), pp. 2091–2110.

Negative Impedance Boosting

By L. A. MEACHAM

(Manuscript received February 6, 1968)

Linearized and feedback-stabilized negative impedance circuits having only R, C, and solid state components, powered in series at intervals along a cable pair, offer new possibilities in bilateral transmission. After discussing the basic negative impedance boosting units and the transmission characteristics they impart to a line (computed, with experimental confirmation), this paper describes a field test of two 32-mile telephone lines, largely 22-gauge, each having an insertion loss of only 3 dB at 1,000 Hz. It also shows means for broadening bandwidth and almost eliminating delay distortion over negative impedance boosted lines. Treatment of this sort adapts them to unusual uses. Examples include converting rectangular to raised-cosine pulses in transmission, without pulse-forming circuitry, and the bilateral two-wire transmission of carrier or pulse signals in both directions simultaneously, without frequency separation.

I. INTRODUCTION

The insertion of lumped negative impedances at intervals along each conductor of a cable pair has long been of interest as a means of improving bilateral transmission. In the familiar expressions for propagation constant

$$\gamma = \alpha + j\beta = \sqrt{(R + j\omega L)(G + j\omega C)} \tag{1}$$

and characteristic impedance

$$Z_0 = R_0 + jX_0 = \sqrt{(R + j\omega L)/(G + j\omega C)}, \tag{2}$$

if one lets both G and R go to zero on presumption that the shunt conductance of well-insulated cable is negligible and that the copper resistance can effectively be canceled by active devices, he encounters four challenging approximations:

$$\alpha \approx 0, \qquad \beta \approx \omega\sqrt{LC}, \qquad R_0 \approx \sqrt{L/C} \quad \text{and} \quad X_0 \approx 0. \tag{3}$$

To the extent of their accuracy these describe lossless transmission, free of phase distortion, between matching terminations that are resistive and independent of frequency. Such properties would indeed be of value in either analog or digital transmission.*

In the early 1940s effort toward canceling R was devoted to high speed point-contact thermistors as the requisite "current-controlled" or "open-circuit-stable" negative impedance elements,[2] but lack of stability and uniformity were severe obstacles. Similar handicaps were later encountered with other devices such as avalanche transistors.[3] At least partly for such reasons, development eventually tended to abandon the scheme of distributing bilateral active elements along a pair over which they could also be powered, and instead moved toward combinations of shunt and series type negative impedances (transformer coupled, locally powered, and designed to match the cable in characteristic impedance) that could be installed at convenient points such as in central offices, and there contribute modest amounts of bilateral gain. A well-known outcome was the E-type repeater,[4] of which both vacuum-tube and transistor versions have found extensive use in the exchange plant of the Bell System.

Recently, however, a new look has been taken at negative impedance boosting† (NIB). This paper outlines in chronological order various findings of a small research project that has been in progress for several years at Bell Laboratories.

II. BASIC NIB CIRCUIT

An NIB unit devised early in this study and used as a basic tool appears schematically in Fig. 1. Figure 2 shows its d-c V-I characteristic and equivalent circuit. For convenience the latter represents the total impedance Z_A of a pair of units, one in series with each conductor, at a boosting point.

Accordingly, for small currents (below the first bend of the characteristic) $-R_n = +2R_3$ and $R_p = 4R_2$. At that bend the silicon transistors begin to conduct, while at the second bend they saturate.

* As early as 1887 Oliver Heaviside defined a "distortion constant" $(R/L - G/C) = 2\sigma$ and an "attenuation constant" $(R/L + G/C) = 2\delta$, and showed that distortion could be "annihilated" by increasing G to make $G/C = R/L$[1]. He undoubtedly would have stressed the benefits of making both σ and δ approach zero, had he known of any way to reduce R except the use of more copper.

† "Boosting" is proposed as a better term than "loading," on the grounds that the mass/inductance analogy suggested by the latter is irrelevant.

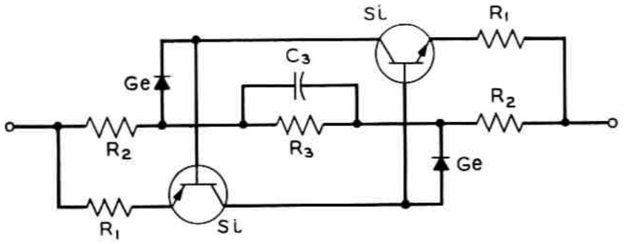


Fig. 1 — Circuit schematic of basic negative impedance booster unit.

In the active region between bends simple circuit analysis shows that

$$R_p = \frac{4R_1R_2}{R_1 + R_2}, \tag{4}$$

$$-R_n = -2R_3\left[\frac{R_2(2\alpha - 1) - R_1}{R_1 + R_2}\right], \tag{5}$$

and

$$-C_n = \frac{R_3C_3}{-R_n}. \tag{6}$$

Here α, the usual ratio of collector to emitter current, is assumed constant and the same for both transistors. Expression (5) tacitly takes into account the nonlinearity of the emitter junctions in Fig. 1; this follows from the fact that the voltage across each emitter junction is compensated, except for an approximately constant voltage difference of about 0.5 volt, by the drop across a germanium junction diode carrying a proportional and almost equal current. The 0.5-volt difference, inherent between silicon and germanium, effectively affords a bias essential to the circuit. The drop across R_2 equals this bias at the first bend, and to a close approximation exceeds the drop across R_1 by the same value of 0.5 volt throughout the active region. The important result of this compensation is a high degree of linearity between the bends, which correspondingly are sharpened almost into cusps.

III. BASIC NIB LINE

Some basic features of an NIB line are illustrated in the telephone customer's loop of Fig. 3. The boosters have a spacing that is (preferably) regular and not much greater than one quarter wavelength at the top of the transmission band. For telephone speech, a suitable

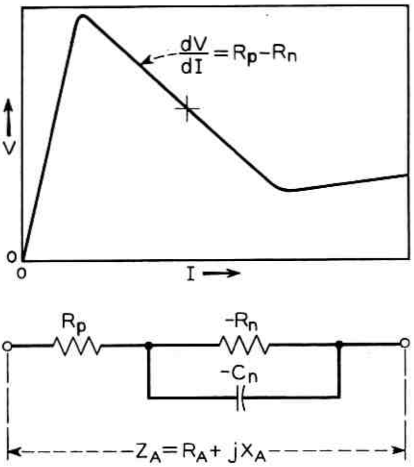


Fig. 2 — DC characteristic and equivalent circuit of basic NIB unit.

spacing would be 12,000 feet. In general, boosting gives the line a characteristic impedance substantially lower than that of ordinary nonloaded or inductively loaded telephone lines. Hence the line circuit at the central office includes an impedance-matching transformer, as well as means for regulating the d-c loop current roughly at the center of the active region of the V-I characteristic. The telephone set can be conventionally powered by this current, and should have a resistive impedance, preferably matching that of the line.

Stability criteria are well known[5] for such arrangements. In practical terms, for regularly spaced NIB units with the equivalent circuit of Fig. 2, the system is found stable (experimentally and by computer) when the net d-c variational resistance ($\Delta V/\Delta I$) of the loop, including its terminations, is positive, provided that the time constant $T_n = R_nC_n$ is greater than a certain critical value. In this study (except where noted) we have consistently made

$$\Sigma R = lR + R_p - R_n = 0,$$

where R is the copper resistance per unit length of cable and l is the NIB spacing. The negative capacitor $-C_n$ bypasses $-R_n$, and with rising frequency gradually reduces the negative real component of terminal impedance of the NIB unit. One way of visualizing the need for such reduction is to notice that the positive copper resistance adjacent to each of the four terminals of the two NIB units at a boosting point is also effectively reduced, being bypassed by the mutual line capacitance. Hence with rising frequency, a point of instability is almost certain to be reached unless the negative resistance diminishes at least as fast as the copper resistance as seen from the NIB terminals.

T_n is therefore an important parameter of the NIB circuit. Increasing it raises the margin of stability, but at the penalty of reducing transmission bandwidth. The midspacing image impedance of the line is also affected by T_n. It is found that when the line conductance per section (lG) is negligible, and when the line resistance per section (lR) is exactly compensated by $R_p - R_n$, the midspacing image impedance Z_H remains essentially constant and resistive as the frequency falls toward zero. As shown in the Appendix, the value it thus approaches is given precisely by

$$Z' = \lim_{\omega \to 0} Z_H = \sqrt{\frac{R_n T_n}{lC} + \frac{L}{C} - \frac{R^2 l^2}{12}} \tag{7}$$

where R, L and C are the usual primary cable constants (per unit length). R_p enters (7) implicitly, being the difference between R_n and lR.

A related effect of T_n is upon the phase velocity $V_H = \omega/\beta_H$, which also approaches an asymptotic value:

$$V' = \lim_{\omega \to 0} V_H = \frac{1}{CZ'}. \tag{8}$$

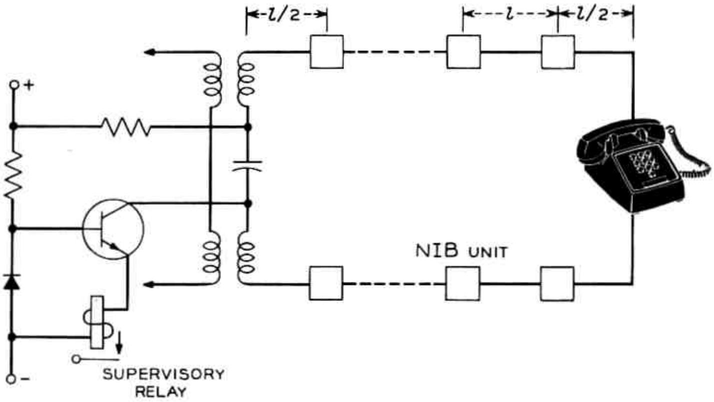


Fig. 3 — Negative impedance boosted subscriber line and central office terminating circuit.

Expressions (7) and (8) are useful, for the values they give hold approximately over a major part of the low-loss frequency band. As an example, take the case of 12,000-foot (2.2727-mile) NIB spacing, along 22-gauge BSA cable that has the primary constants (at low frequency) $R = 173$ ohms per mile, $L = 0.874 \times 10^{-3}$ henry per mile, and $C = 0.825 \times 10^{-6}$ farad per mile. For the NIB parameters (per section) $R_p = 97.3$ ohms, $R_n = 490.5$ ohms, and $T_n = 16 \times 10^{-6}$ second, expressions (7) and (8) tell us

$$Z' = 198 \text{ ohms}$$

and

$$V' = 61{,}200 \text{ miles per second.}$$

For this velocity, the spacing becomes a quarter wavelength at the frequency

$$f_{\lambda/4} = V'/4l = 6{,}730 \text{ Hz.}$$

IV. COMPUTED CHARACTERISTICS

Computer programs have been worked out to give propagation constant and midspacing image impedance as functions of frequency, for any set of cable primary "constants" (which of course actually vary with frequency) and NIB equivalent circuit parameters. Some typical results, plotted in Figs. 4, 5, 6, apply to the set of parameters used in the foregoing example. For comparison, characteristics are included for nonloaded (NL) and loaded (H88) cable, also of 22 gauge. (H88 loading uses 88 mH inductors at 6,000-foot intervals.)

Among varied uses of these programs has been the finding, by successive approximations, of the minimum or "just stable" time constant (jstc) for various gauges and NIB spacings. Figure 7 shows attenuation constant versus frequency for the jstc condition and also for a time constant 10 per cent greater. To illustrate another use, the effect upon attenuation of moderate over- or undercompensation is pictured in Fig. 8. Here the loss per mile between image impedances is shown for errors in compensation of ±20 ohms, or approximately ±5 per cent of the copper resistance. Over most of the useful band, these errors introduce almost flat gain or loss of about 0.2 dB per mile. Their effects upon phase velocity and image impedance (not plotted) are small except at frequencies below 500 Hz.*

* In that region the variation of image impedance is such that if Fig. 8 were a plot of insertion loss between 198-ohm resistive terminations, it would show the gain or loss of 0.2 dB per mile extending almost unchanged all the way down to zero frequency.

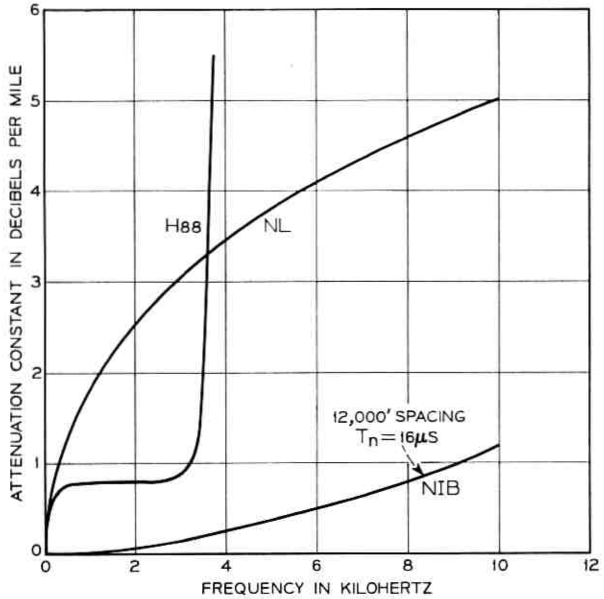


Fig. 4 — Attenuation constant of 22-gauge BSA cable; nonloaded, H-88 loaded, and negative impedance boosted.

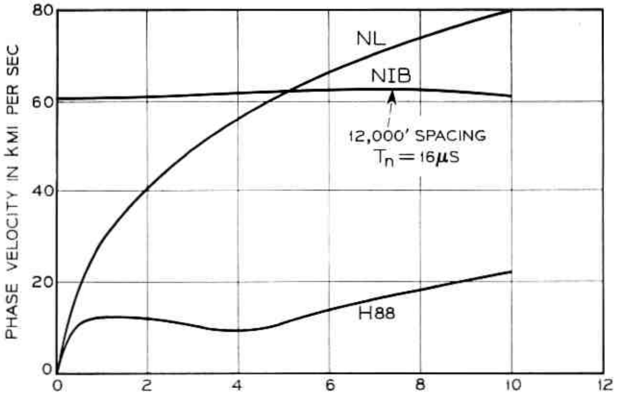


Fig. 5 — Phase velocity of 22-gauge BSA cable; nonloaded, H-88 loaded and negative impedance boosted.

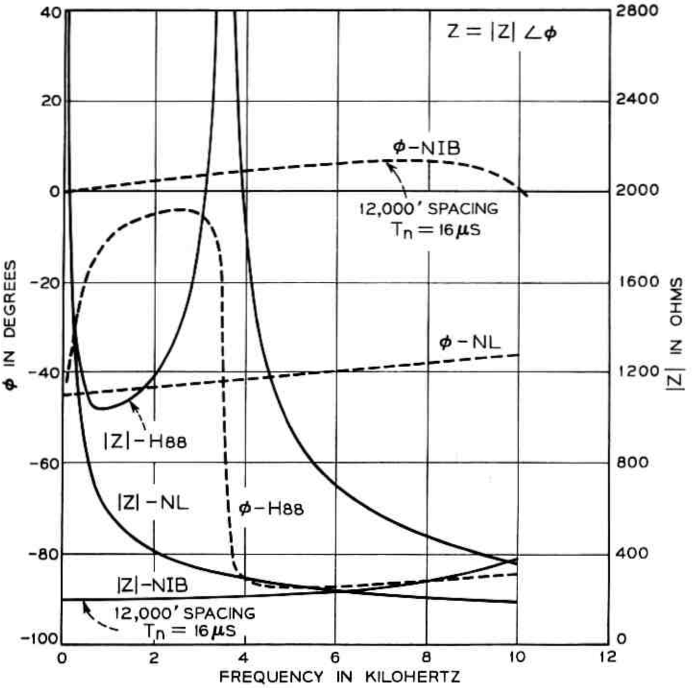


Fig. 6 — Characteristic impedance or midspacing image impedance of 22-gauge BSA cable; nonloaded, H-88 loaded, and negative impedance boosted.

In general, our laboratory tests using either dependably representative artificial lines, or pairs in actual cable on spools, confirmed the computed results very accurately. Conversation over lines several 12,000-foot NIB sections in length was found highly satisfactory—remarkably free of hum, echo, and distortion. But the need was seen for experience with NIB transmission under actual field conditions.

V. "ROUND ROBIN" FIELD TEST

With the cooperation of the New Jersey Bell Telephone Company two NIB lines were set up using pairs in existing interoffice cables over the route shown in Fig. 9. For convenience of measurement, both ends of each line were brought to the same room at the Murray Hill, New Jersey, branch of Bell Laboratories. Experimental applique circuits were provided for coupling to the Murray Hill PBX, permitting each pair to serve as a regular telephone extension when not in use for other tests.

The cable, 32.4 miles long, was all of 22 gauge except for 0.5 per cent of 24 and 4.2 per cent of 26 gauge. Seventy-seven per cent of its length was underground, the rest aerial. All boosting points, one for each of 16 sections ranging from 9,750 to 13,380 feet long, were in manholes. There the NIB units were plugged into jacks within containers that could be conveniently opened and resealed, taking advantage of equipment already installed (for housing regenerative repeaters of the T1 type PCM transmission system).

The NIB circuits were adapted to field conditions in the following ways:

(*i*) By giving R_3 an appropriate positive temperature coefficient, the net coefficient of each NIB unit was matched approximately to that of copper. It was recognized that this compensation would be reasonably accurate for underground cable, but little better than seasonal for aerial.

(*ii*) Taps were provided along R_3 so that any one of four values of

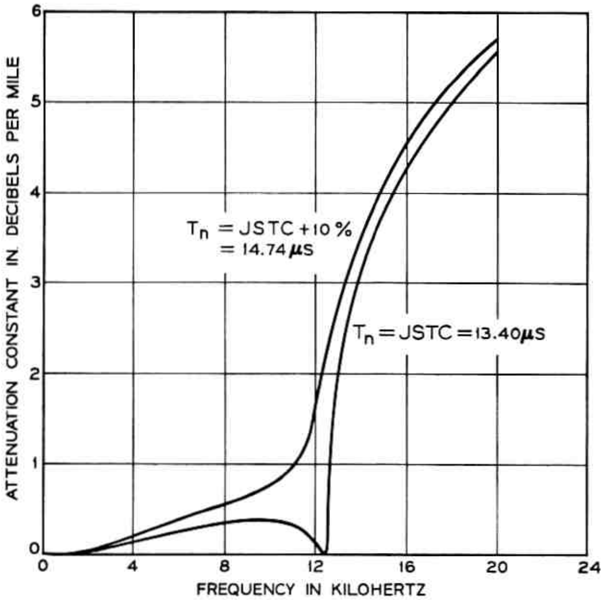


Fig. 7 — Attenuation constant of line with NIB time constant at or near "just stable" value; 22-gauge BSA cable, NIB spacing 12,000 feet.

$-R_n$ could be selected by strapping, as a best fit for the section resistance. No corresponding adjustment of C_3 proved necessary, as the image impedance fortunately turned out to be kept almost constant by the related changes in T_n, $-R_n$ and l.

(*iii*) To increase stability margins in view of the nonuniform NIB spacing, T_n was raised to 20μs for the mean length of 22-gauge section. This gave an image impedance Z' of 225 ohms.

(*iv*) For the two end sections of each line, which happened to include all the 26-gauge cable, T_n was adjusted by changing the capacitor C_3 (Fig. 1) to make the image impedance roughly equal to that of other sections (225 ohms).

Except for these adjustments, the NIB units had the equivalent circuit parameters listed in the discussion of expressions (7) and (8). They were normally powered by 16 mA of loop current, with their linear negative slopes extending from 6 to 26 mA. This range was

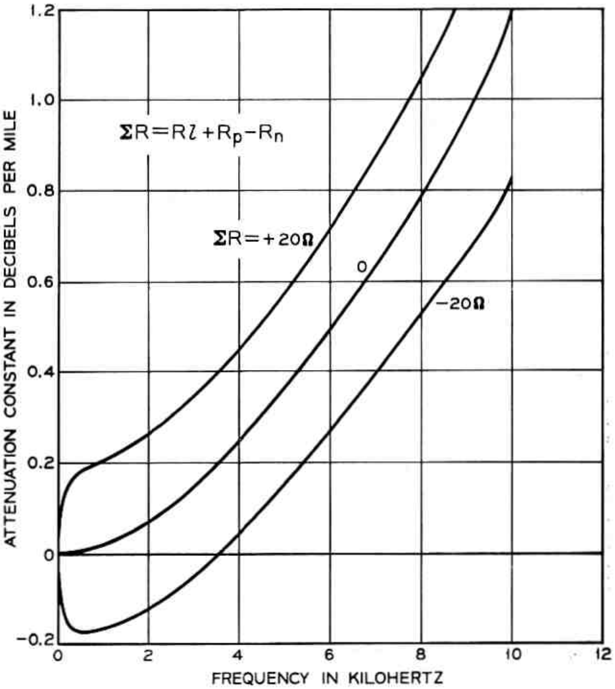


Fig. 8 — Effect of over- or undercompensation of copper resistance; 22-gauge BSA cable, spacing 12,000 feet, 16μs time constant.

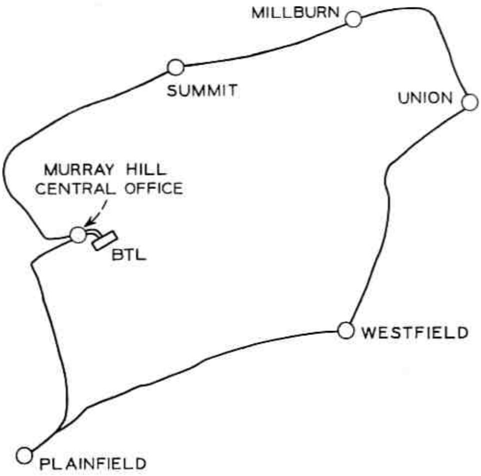


Fig. 9 — Cable route in field test of negative impedance boosting. The line was 32 miles long, mostly buried 22 gauge cable, and it had 16 NIB sections.

twice as great as required for telephone speech; the excess was an allowance for possible hum current. The total IR drop in copper and NIB units of either loop was about 186 volts; hence, with an additional 4-volt drop across a 225-ohm resistive station set, the potentials on tip and ring conductors at the "office end" were approximately ±95 volts from ground.

Touch-Tone® calling was used on one line, rotary dialing on the other. The severe distortion occurring when the rotary dialing pulses were produced by complete interruption of loop current was remedied by having the dial merely insert enough resistance to drop the current from 16 to 6 mA (the regulator going out of range). With the NIB thus left operative, dial pulse distortion became negligible.

Tone ringing[6] was used on both lines, the signal being a 1,000 Hz wave interrupted at 10 Hz. This was applied with a level of about 1 mW at the applique line circuit, under control of the ordinary ringing signal from the PBX.

Supervision was conventional. The current regulator was so designed that when the path was broken by the switchhook the open-circuit voltage on the loop did not greatly exceed the ±95 volt figure. A relay in the applique, responding to the switchhook (and dial pulses) transferred the information to the PBX pair.

The performance of the NIB lines was gratifying. People conversing

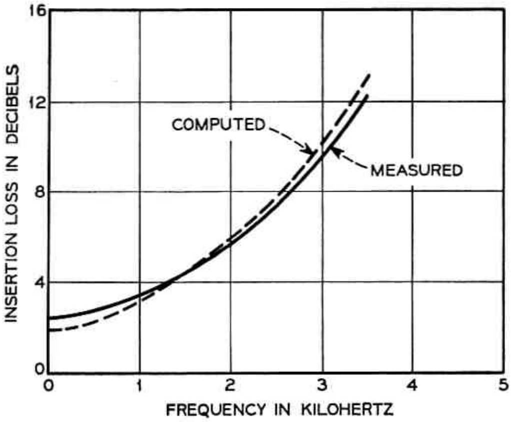


Fig. 10 — Insertion loss of 32-mile field-test NIB line between 225-ohm resistive terminations.

over them were favorably impressed by resemblance of the transmission to that over a short loop, and by freedom from noise, hum, crosstalk and distortion. Fig. 10 shows the insertion loss of one 32.4-mile line measured between 225 ohm resistive terminations. It also shows a computed plot of this loss, using a program that takes account of the individual dimensions of each section and NIB unit.

To help ensure stability in spite of the inherent restrictions on temperature compensation, the total copper resistance (6,000 ohms) was intentionally left undercompensated by about 100 ohms. As a result, the insertion loss had a low-frequency asymptote of roughly 2 dB. Strip chart records of a 1 kHz test tone showed the transmission varying over a typical day and night by about ±0.5 dB. Neither line lost stability at any time during the entire test, which extended over four fall and winter months and encountered large and rapid changes of weather.

Figure 11 shows the input impedance of one line, measured and computed, for a 225-ohm resistive far-end termination. The irregularities of these plots, resulting from nonuniformity of the sections, correspond to echo return losses no smaller than 12 dB, and exceeding 17 dB over most of the band.

Crosstalk loss between the two lines was roughly 88 dB at 1 kHz; there was little difference between near-end and far-end measurements.

In planning the field test, hum was of course recognized as a possible source of trouble. It was known that hum is generally introduced by magnetic induction from power lines, effectively generating equal voltages in series with each conductor. Longitudinal hum currents, impelled by these voltages, could trouble the NIB transmission in two ways: by using up a significant part of the operating range of the NIB units, and by coupling into the metallic circuit as a result of unbalance between the two sides of the line.

Experience and measurements afforded by the test were encouraging, but not extensive enough to be conclusive. In order to minimize hum currents, station grounds were avoided; the only path to ground was via capacitance distributed along the line. At the central office end, the longitudinal termination to ground was roughly matched to the longitudinal impedance of the line, to avoid possible accumulation of multiple reflections. With this arrangement line balance was found adequate to prevent more than a trivial hum level from ever being coupled into the telephones.

Hum voltage to ground (largely 60 Hz) recorded at the station end was found to vary from minute to minute as well as over a daily cycle. The extreme range of these measurements was from 1.8 to 6.5 volts rms, the largest values occurring around 5 to 6 pm. Without knowing the distribution of magnetic induction along the line, one could not determine hum current from such measurements. However,

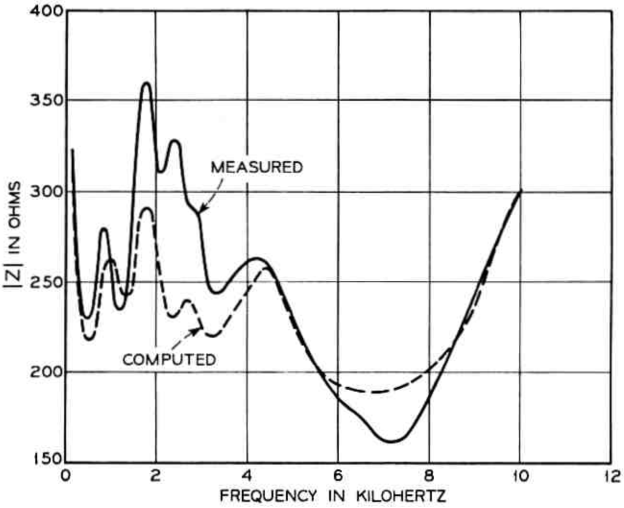


Fig. 11 — Input impedance of 32-mile NIB line with 225-ohm resistive far-end termination.

by varying the d-c loop current away from its usual 16 mA value until current peaks "bumped" an edge of the NIB dynamic range (putting audible 120 Hz pulses into the metallic circuit) one could readily measure the maximum hum current, reached at some boosting point along the line. Typical measurements of this sort gave values around 5 mA peak-to-peak in each conductor, or 25 per cent of the 20 mA dynamic range; under worst conditions at least half the range was undoubtedly filled. Although this amount of hum was found to have no noticeable effect on telephone speech, larger hum currents would probably be encountered at other locations.

Experience with lightning was also encouraging although far from comprehensive. The NIB units were left unprotected except by their own fairly low resistance at large forward currents, and by diodes to bypass reverse currents. No damage was done by thunderstorms, several of which did occur during the field run. Not until after these tests was it recognized that valid protection against large forward currents also could have been provided by merely giving each bypassing diode a zener potential of around 10 volts. Of course, this value is chosen to exceed the drop across the NIB at the "first bend" of its V–I plot. At large forward currents, the emitter and base circuit resistors of Fig. 1 combine to give a terminal resistance of about 48 ohms. With the terminal voltage zener-limited to 10 volts, the current through the NIB could not exceed 0.2 ampere, whereas simulated lightning tests have shown that an unprotected NIB is undamaged by surge currents as great as 5 amperes. Lightning is not expected to present a serious problem.

VI. BAND BROADENING

Shortly after conclusion of this field experiment continuing effort to improve the NIB circuit revealed that by adding to it a resistor and a capacitor, one could flatten and substantially broaden the resulting transmission band, indeed achieving virtually flat lossless transmission almost up to the frequency of quarter-wavelength NIB spacing. The band-broadened equivalent circuit, shown in Fig. 12, is simply that of the basic unit (Fig. 2) shunted by R_s and C_s in series.

The effect of the addition can be seen more readily if one first writes the impedance of the basic unit:

$$Z_A = R_A + jX_A = \frac{-R_n}{1 + (\omega T_n)^2} + R_p + j\omega \frac{T_n R_n}{1 + (\omega T_n)^2}. \qquad (9)$$

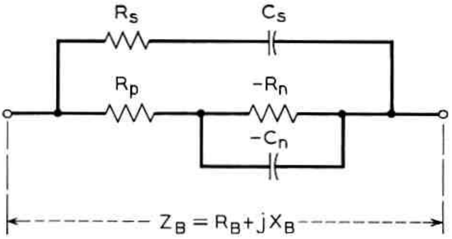


Fig. 12 — Equivalent circuit of NIB unit with band broadening.

When the shunt is applied, the real component R_B (negative) of the resulting terminal impedance Z_B is made larger than the real component R_A (also negative) of Z_A by what amounts to antiresonance between C_s and the positive (inductive) imaginary component of Z_A. Resistor R_s keeps the shunt path from acquiring so low an impedance at any frequency as to bring instability to the "open-circuit-stable" basic unit.

When the straightforward algebraic analysis used to derive expression (7) for the basic unit is repeated for the band-broadened circuit, it shows that the asymptotic low-frequency image impedance (for $G = 0$ and $\Sigma R = 0$) has been slightly modified. With the shunt elements added,

$$Z' = \lim_{\omega \to 0} Z_H = \sqrt{\frac{R_n T_n}{lC} + \frac{L}{C} - \frac{R^2 l^2}{12} - \frac{R^2 l T_s}{R_s C}}. \qquad (10)$$

This expression reverts to (7) when $T_s \to 0$ with $R_s > 0$, or when $R_s \to \infty$ with finite T_s. Computer results confirm the accuracy of (10).

Computed transmission characteristics also support an initial estimate that the time constant $T_s = R_s C_s$ should be made roughly equal to T_n, and show that the revised circuit can be proportioned to sustain its compensation of copper resistance up to higher frequencies, while still letting its negative resistance fall off fast enough above the transmission band to preserve stability.

The effect of band broadening upon the NIB transmission is shown in Figs. 13 and 14 for the case of 22-gauge BSA cable with 12,000-foot NIB spacing, used earlier as an example. Here both time constants (T_s and T_n) are made 16 μs, and curves are shown for three values of R_s. When $R_s = 2{,}000\ \Omega$ the attenuation (Fig. 13) has its widest flat region without appreciable gain over any of the band. For $R_s = \infty$, the circuit reverts to the original or basic NIB. At an inter-

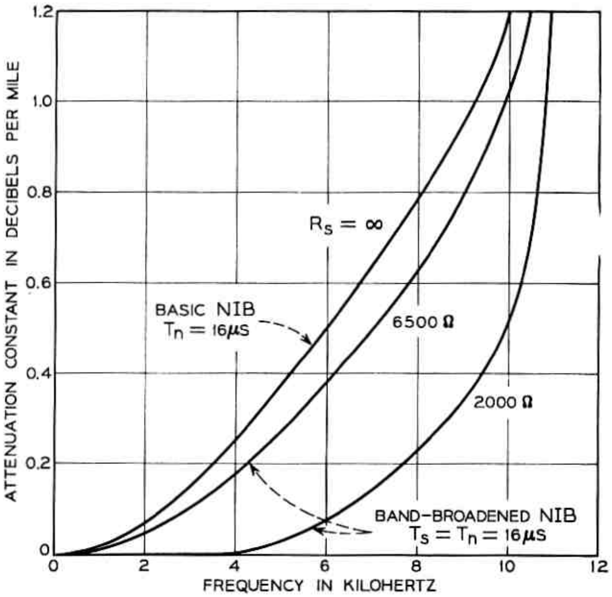


Fig. 13 — Effect of band broadening upon attenuation constant of NIB line; 22-gauge BSA cable, 12,000-foot spacing.

mediate value, $R_s = 6{,}500\ \Omega$, there is less band broadening, but the phase velocity (Fig. 14) becomes remarkably constant from zero frequency up to 6.7 kHz (at which $l = \lambda/4$). A similar change in slope of the phase velocity plot, shifting from positive sign for the basic NIB to negative for the band-broadened version, has consistently been observed over a wide variety of gauges and booster spacings.

VII. PULSE FORMING

The foregoing combination of linear variation of phase with an approximately parabolic variation of loss in dB, both as functions of frequency, clearly offers interesting possibilities in baseband pulse transmission. Under such a condition the line has the properties of a Gaussian filter. If rectangular pulses of a suitable width T and baud rate $f_o = 1/T$ are applied to it, these pulses are shaped in transmission into the raised cosine form. As received, they have the width T at half their peak amplitude and $2T$ along the baseline; they are almost free of tails. For ideal raised-cosine pulse forming, the line or Gaussian filter should have a loss of 1 neper or 8.68 dB at the baud rate f_o. Hence, for the case of $R_s = 6{,}500$ ohms in Figs. 13 and 14, a baud rate of 8 kHz (at which the loss is about 0.635 dB per mile) could be sent over a line 8.68/0.635 = 13.7 miles long. Of course if the line were shorter, or the baud rate slower, the pulses would still be symmetrical and well formed, but would show flatness at their peaks.

Figure 15 shows the output "eye-diagram" formed by a random sequence of 8-level rectangular pulses at a 16.67 kilobaud rate, sent over 10.2 miles of 22-gauge BSA cable, with 6,000-foot NIB spacing. Here the information rate was 3 times the baud rate, or 50 kilobits per second. The NIB parameters were $R_p = 97.3$ ohms, $R_n = 293.9$ ohms, $T_s = T_n = 6.1 \times 10^{-6}$ second and $R_s = 2{,}000$ ohms.

VIII. BIDIRECTIONAL TRANSMISSION

Because of low loss in a broad transmission band, and an image impedance that can be well matched over that band, new possibilities are opened of simultaneous bidirectional carrier transmission; for

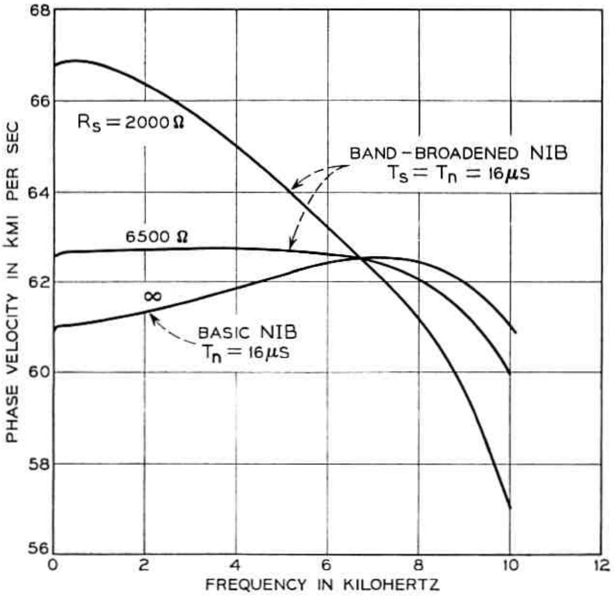


Fig. 14 — Effect of band broadening upon phase velocity of NIB line; 22-gauge BSA cable, 12,000-foot spacing.

example, by double-sideband amplitude modulation of the same carrier frequency at each terminal of the line. Similar possibilities exist for bidirectional baseband pulse transmission. Both of these schemes have been successfully carried out in the laboratory over the same 10.2-mile 22-gauge line with 6,000-foot spacing that was used in obtaining Fig. 15.

In either case, hybrid balance separates the incoming from the outgoing signal. As a result of the low transmission loss, the received signal, if it is a modulated carrier, is left sufficiently free of outgoing carrier (whatever its phase) to be detected without appreciable distortion. Similarly, if the received signal is a pulse train, it is left sufficiently free of interference from the outgoing pulses to be correctly decoded or regenerated.

Figure 16 shows two eye diagrams, received simultaneously at the two ends of the 10.2-mile line while two random 8-level pulse trains were being sent in the respective directions. Some interference may be seen in the interpulse intervals, resulting from imperfection of the hybrid balance presented to the higher frequency components of the rectangular input pulses. For this photograph, the pulse rate was raised slightly (to 16.81 kilobauds), thereby roughly centering the interference in the intervals between eyes of the diagram.

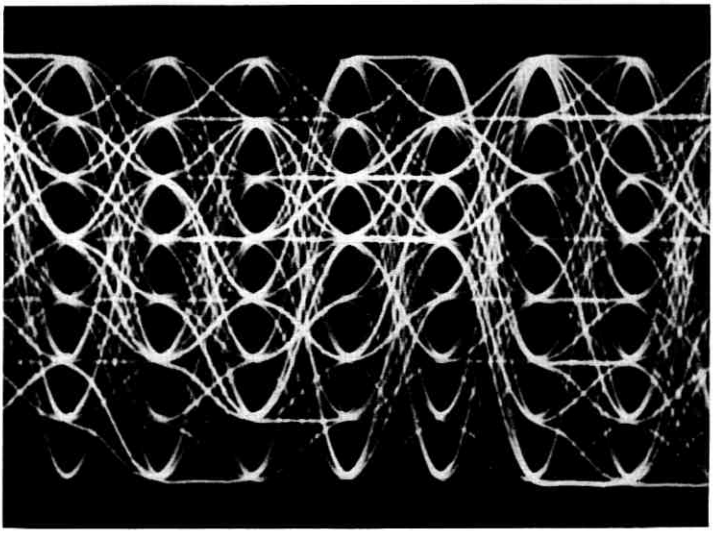


Fig. 15 — Eight-level pulses received over phase-linearized NIB line.

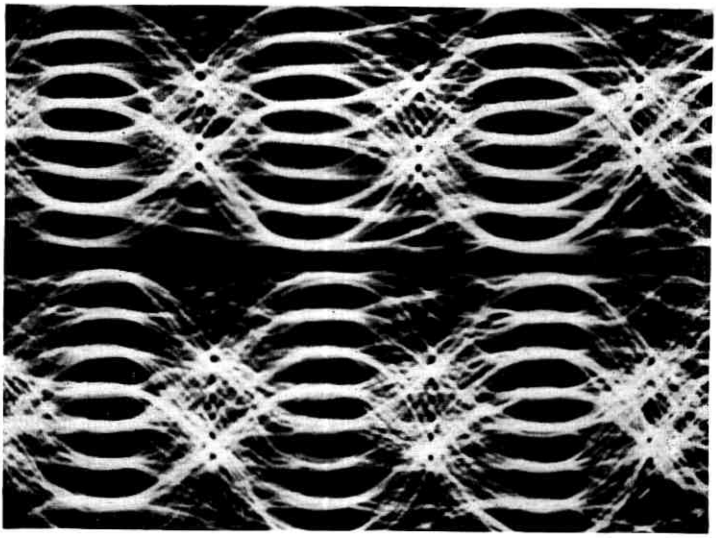


Fig. 16 — Bilateral pulse transmission over phase-linearized NIB line.

APPENDIX

Zero-frequency Asymptotes of Midspacing Image Impedance and Phase Velocity of NIB Lines

In this appendix derivations are given for expressions (7) and (8) of the text. The same method yields (10) when the NIB units include R_s and C_s as in Fig. 12.

Terms

Z_H = Midspacing image impedance of NIB line.

V_H = Phase velocity of NIB line.

Z_A = Total impedance of two NIB units, one on each side of balanced line, serving a single section.

l = Length of NIB section (miles).

Z_O = Characteristic impedance of nonloaded line.

$\gamma = \alpha + j\beta$ = Propagation constant of nonloaded line (per mile).

Z_{OC} = Open-circuit impedance of nonloaded half section (length $l/2$).

Z_{SC} = Short-circuit impedance of nonloaded half section (length $l/2$).

$T_n = R_nC_n$ = Time constant of basic NIB unit.

$lP = l(\alpha_H + j\beta_H)$ = Propagation constant of NIB line (per section).

$Z' = \lim\limits_{\omega \to 0} Z_H$

$V' = \lim\limits_{\omega \to 0} V_H$

Characteristic Impedance

From well-known theory,[4]

$$Z_H = Z_{OC}\sqrt{\frac{Z_A + 2Z_{SC}}{Z_A + 2Z_{OC}}} \tag{11}$$

$$Z_{OC} = \frac{Z_O}{\tanh \frac{\gamma l}{2}} \tag{12}$$

$$Z_{SC} = Z_O \tanh \frac{\gamma l}{2} \tag{13}$$

$$Z_O = \sqrt{\frac{R + j\omega L}{G + j\omega C}} = \frac{a}{b} \tag{14}$$

and

$$\gamma = \sqrt{(R + j\omega L)(G + j\omega C)} = ab \tag{15}$$

where

$$a^2 = R + j\omega L \quad \text{and} \quad b^2 = G + j\omega C. \tag{16}$$

Also

$$\tanh \frac{\gamma l}{2} = \left(\frac{\gamma l}{2}\right) - \frac{1}{3}\left(\frac{\gamma l}{2}\right)^3 + \frac{2}{15}\left(\frac{\gamma l}{2}\right)^5 - \cdots$$

$$= \frac{ablF}{2} \tag{17}$$

where

$$F = \left(1 - \frac{a^2b^2l^2}{12} + \frac{a^4b^4l^4}{120} - \cdots\right). \tag{18}$$

Then

$$Z_{OC} = \frac{2}{b^2lF} \tag{19}$$

$$Z_{SC} = \frac{a^2lF}{2} \tag{20}$$

and from (11),

$$Z_H^2 = \frac{4(Z_A + a^2lF)}{b^2lF(Z_Ab^2lF + 4)}. \tag{21}$$

From Fig. 2

$$Z_A = R_p - R_n \frac{(1 - j\omega T_n)}{1 + \omega^2 T_n^2}$$

$$= \frac{R_p + \omega^2 T_n^2 R_p - R_n + j\omega R_n T_n}{1 + \omega^2 T_n^2}$$

$$= S(R_p - R_n) + T_n S(\omega^2 R_p T_n + j\omega R_n) \tag{22}$$

where

$$S = \frac{1}{1 + \omega^2 T_n^2} = 1 - \omega^2 T_n^2 + \omega^4 T_n^4 - \cdots . \tag{23}$$

Putting (16) and (22) into (21) gives

$$Z_{H^2} = \frac{4[S(R_p - R_n) + T_n S(\omega^2 R_p T_n + j\omega R_n) + (R + j\omega L)lF]}{(G + j\omega C)lF[S(R_p - R_n)(G + j\omega C)lF + T_n S(\omega^2 R_p T_n + j\omega R_n)(G + j\omega C)lF + 4]} . \tag{24}$$

Of present interest is the special case in which $G = 0$ and $R_p - R_n = -lR$. For this condition,

$$Z_H^2 = \frac{4[lR(F - S) + j\omega(lLF + R_n T_n S) + \omega^2 R_p T_n^2 S]}{j\omega ClF[4 - j\omega Cl^2 RFS + j\omega ClFT_n S(\omega^2 R_p T_n + j\omega R_n)]} . \tag{25}$$

Consider now the term $F-S$ in the numerator of (25). From (18) and (23)

$$F - S = -\frac{a^2 b^2 l^2}{12} + \omega^2 T_n^2 + \frac{a^4 b^4 l^4}{120} - \omega^4 T_n^4 \cdots \tag{26}$$

and since $a^2 b^2 = j\omega RC - \omega^2 LC$

$$F - S = -\frac{j\omega l^2 RC}{12} + \omega^2 M\left(\frac{l^2 LC}{12} + T_n^2 - \frac{l^4 R^2 C^2}{120}\right) \tag{27}$$

where $M = 1 +$ terms of positive order in ω. Putting (27) into (25) gives

$$Z_H = \frac{4\left[\frac{-j\omega l^3 R^2 C}{12} + \omega^2 lRM\left(\frac{l^2 LC}{12} + T_n{}^2 - \frac{l^4 R^2 C^2}{120}\right) + j\omega\left(lLF + R_n T_n S\right) + \omega^2 R_p T_n{}^2 S\right]}{j\omega ClF[4 - j\omega Cl^2 RF + j\omega ClFT_n S(\omega^2 R_p T_n + j\omega R_n)]}$$

$$= \frac{4\left[\frac{-l^2 R^2}{12} - j\omega RM\left(\frac{l^2 L}{12} + \frac{T_n{}^2}{C} - \frac{l^4 R^2 C}{120}\right) + \frac{LF}{C} + \frac{R_n T_n S}{lC} - \frac{j\omega R_p T_n{}^2 S}{lC}\right]}{F[4 - j\omega Cl^2 RF + j\omega ClFT_n S(\omega^2 R_p T_n + j\omega R_n)]} . \tag{28}$$

Thus far, although F, S, and M are power series expansions, they are included in their entirety; nothing has been approximated, and (28) is therefore exact.

In passing to the zero-frequency limit we notice that when $G = 0$

$$\lim_{\omega \to 0} F = \lim_{\omega \to 0} S = \lim_{\omega \to 0} M = 1. \tag{29}$$

Accordingly, for the special case considered,

$$(Z')^2 = \lim_{\omega \to 0} Z_H^2 = \frac{4\left[-\frac{l^2R^2}{12} - 0 + \frac{L(1)}{C} + \frac{R_nT_n(1)}{lC} - 0\right]}{(1)[4 - 0 + 0]}$$

$$= \frac{R_nT_n}{lC} + \frac{L}{C} - \frac{l^2R^2}{12}. \tag{30}$$

Phase Velocity

Again from well-known theory,[4]

$$\tanh \frac{lP}{2} = \frac{Z_H}{Z_{OC}}. \tag{31}$$

From (30) the zero-frequency limit of Z_H is finite, real and presumed positive, while from (19) that of Z_{OC} (for $G = 0$) is infinite, imaginary and negative. Hence as $\omega \to 0$,

$$\frac{Z_H}{Z_{OC}} = \tan \frac{lP}{2} \to \frac{lP}{2} \to \frac{j\beta_H l}{2} \to 0. \tag{32}$$

Accordingly,

$$\lim_{\omega \to 0} \beta_H = \lim_{\omega \to 0} \frac{2Z_H}{jlZ_{OC}}, \tag{33}$$

and since $V_H = \omega/\beta_H$,

$$\lim_{\omega \to 0} V_H = \lim_{\omega \to 0} \frac{j\omega l Z_{OC}}{2Z_H}. \tag{34}$$

But from (12), with $G = 0$

$$Z_{OC} = \frac{2}{j\omega C l F},$$

and therefore

$$\lim_{\omega \to 0} V_H = \frac{1}{C} \lim_{\omega \to 0} \frac{1}{Z_H}. \tag{35}$$

REFERENCES

1. Heaviside, Oliver, "Electromagnetic Induction and Its Propagation," Electrician, *40* (June 3, 1887), pp. 79–81.

2. Bullington, R. K., "Negative Resistance Loading," U. S. Patent 2,360,932, April 25, 1942.
3. Miller, S. L. and Ebers, J. J., "Alloyed Junction Avalanche Transistors," B.S.T.J., *34,* No. 5 (September 1955), pp. 883–902.
4. Merrill, J. L., Rose, A. F., and Smethurst, J. O., "Negative Impedance Telephone Repeaters," B.S.T.J., *33,* No. 5 (September 1954), pp. 1055–1092.
5. Gammie, J. and Merrill, J. L., Jr., "Stability of Negative Impedance Elements in Short Transmission Lines," B.S.T.J., *34* No. 2 (March 1955), pp. 333–360.
6. Meacham, L. A., Power, J. R., and West, F., "Tone Ringing and Pushbutton Calling," B.S.T.J., *37,* No. 2 (March 1958), pp. 339–360.

Computation of FM Distortion in Linear Networks for Bandlimited Periodic Signals

By CLYDE L. RUTHROFF

(Manuscript received December 14, 1967)

Computations of the distortion generated in passing large-index, frequency-modulated signals through symmetrical single-pole and three-pole bandpass filters are presented. The computation is for a bandlimited periodic modulation signal; noise modulation is simulated by the use of periodic noise samples in a Monte Carlo procedure.

The convergence of the Monte Carlo procedure is illustrated for the case of the single-pole filter and the results are in good agreement with measurements.

Computations of envelope distortion are also presented. These data give the amplitude-to-phase conversion in the receiver containing the filter to within a constant factor, the constant being the AM/PM conversion coefficient of the limiter.

I. INTRODUCTION

In spite of the efforts of a large number of investigators who have studied the problem over three decades there is no way to compute the distortion caused by filters and other networks for arbitrary angle modulated signals of large index or large baseband bandwidths. However, by use of the Fourier method[1–4] introduced by Roder in 1937, it is possible to compute the exact responses of networks to a frequency-modulated signal for bandlimited periodic modulation signals.

In addition to deterministic signals of this class, noise modulation can also be simulated and the resulting network distortion computed by a Monte Carlo procedure. In an excellent paper, Medhurst and Roberts[4] have described the procedure and given some results for low index FM, pre-emphasized in accordance with CCIR standards.

Their computer program was written in Extended Mercury Autocode. The same method, coded in FORTRAN II, and extended to include the effects of amplitude as well as phase distortion is being used to study large index FM systems.

The results presented are for single sine wave modulation and random noise modulation.

II. ANALYSIS

The modulating signals are restricted to those which are both bandlimited and periodic. This class includes many signals used for test purposes; the notable exception is the signal consisting of bandlimited Gaussian noise. More will be said of noise modulation later.

The analysis and computational procedure follows that of Medhurst and Roberts in Ref. 4 and is outlined briefly here. Specifically, the signals are those which can be written as finite Fourier series.

$$\mu(t) = \sum_{n=1}^{N} (a_n \cos n\omega_a t + b_n \sin n\omega_a t) \text{ radians,} \tag{1}$$

where:

$$\omega_a = 2\pi f_a = 2\pi/T,$$

T is the period of $\mu(t)$,

$$a_n = \frac{2}{T}\int_{-T/2}^{T/2} \mu(t) \cos n\omega_a t \, dt,$$

$$b_n = \frac{2}{T}\int_{-T/2}^{T/2} \mu(t) \sin n\omega_a t \, dt.$$

If $\mu(t)$ is the desired phase modulation, or $\mu'(t) = d\mu(t)/dt$ the frequency modulation, the angle-modulated signal is

$$e = (2)^{\frac{1}{2}} \cos [\omega_c t + \mu(t)] \tag{2}$$

where ω_c is the carrier frequency in radians per second. The FM signal of (2) has a line spectrum with lines at $\omega_c \pm M\omega_a$, $M = 1, 2, 3, \cdots$. The lines always occur at these frequencies, changing only in amplitude and phase as functions of a_n, b_n. It is this feature which makes possible a digital computer solution and, conversely, is the reason for restricting the form of the modulating signal to that of $\mu(t)$ in (1). Beginning with (1) and (2) the major steps in the analysis are:

(*i*) Derive the line spectrum of (2).

(*ii*) Modify the lines in amplitude and phase in accordance with the response of the network being studied.

(*iii*) Derive the envelope and phase of the modified line spectrum, that is, determine $E(t)$ and $\theta(t)$ where the output of the network is written

$$e_o = E(t) \cos [\omega_c t + \theta(t)] \tag{3}$$

(*iv*) Derive the line spectrum of $E(t)$, $\theta(t)$, and $d\theta/dt$.

III. RANDOM MODULATION

An important measuring method in widespread use on FM systems is the noise loading test. The importance of this method arises from the fact that a band of thermal noise is a good approximation to a frequency division multiplex signal which consists of a number of voice channels. In this test a band of thermal noise in the frequency range 0-W Hz is the baseband signal. The noise is removed by band rejection filters in one or more narrow bands or slots ahead of the modulator. At the receiver the power density appearing in the slots is a measure of the intermodulation distortion in the system. The results are usually given in the form of a signal-to-distortion ratio, the signal being the power density at the slot frequency when the band rejection filter is removed, that is, when the signal is present.

Computations of distortion can be made along these lines by following a Monte Carlo procedure with a sequence of random noise samples generated from the periodic form of (1). A set of N sine waves of equal amplitudes and random phases distributed uniformly in the interval $0 - 2\pi$ constitutes the basic signal. Figure 1 is an example of this random noise sample for $N = 10$ and Fig. 2 for $N = 50$. One or more amplitudes are set to zero to form the slots, and the power in the slots as a result of network distortion is computed as outlined in Section II. The process is repeated with a sequence of random noise samples, each sample with a set of N independent random phases. The distortion is averaged for the final result. If N is large enough, if the number of sets is large enough, and if the network transfer function is well-behaved, then the results approach those obtained in a noise loading test.

Rice[5] has shown that such a noise representation has a normal amplitude distribution as $N \to \infty$ and $\omega_a \to 0$. Bennett[6] has computed the amplitude distribution as a function of N. The conclusion is that with respect to amplitude distribution the sets of random signals of the

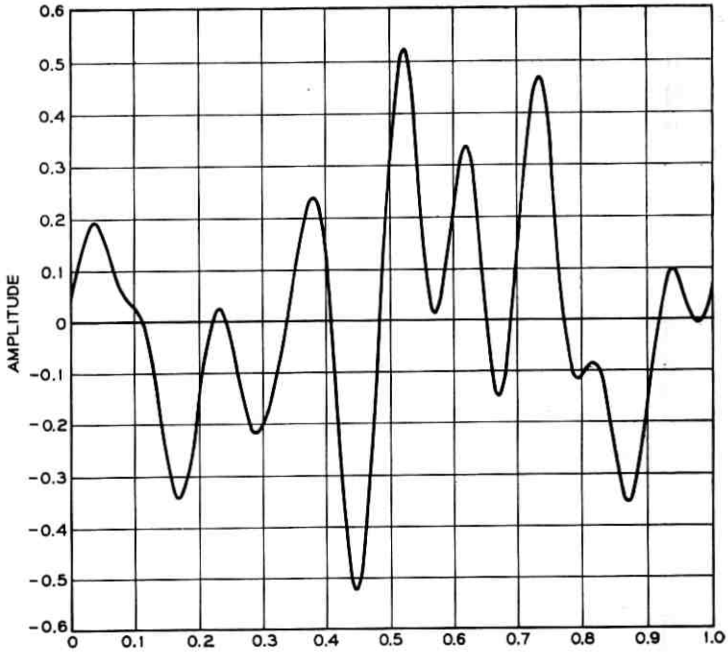


Fig. 1 — A periodic random noise sample for $N = 10$. Peak amplitude/rms amplitude $= | -0.525 |/[1/(2N)^{1/2}] = 2.34$.

form (1) approximate Gaussian noise. With respect to the spectrum the situation is otherwise; the spectrum of noise is continuous whereas the simulation, for finite N, has a line spectrum. This means that the results computed with the simulated noise will approximate the results for real noise only for network responses which are smooth enough. An example of a function which is not smooth enough is a network response of unity at the spectral lines and zero elsewhere. In spite of this limitation it is not expected that smoothness will be a serious problem for most cases of interest.

3.1 *Modulation Index*

The modulating signal $\mu(t)$ can be written as follows:

$$\mu(t) = \sum_{n=1}^{N} A_n \cos (n\omega_a t + \alpha_n) \text{ radians,} \tag{4}$$

where,

$$A_n^2 = a_n^2 + b_n^2$$

$$\alpha_n = -\tan^{-1} \frac{b_n}{a_n}.$$

The baseband is

$$W = N\omega_a . \tag{5}$$

Using (4) to simulate noise in a phase modulation system, the amplitudes A_n are equal and the random phases α_n are uniformly distributed from 0 to 2π. If the rms phase deviation is φ radians,

$$A_n = \varphi(2/N)^{\frac{1}{2}} \quad \text{radians.} \tag{6}$$

For the FM application the amplitude terms of the frequency modulation $\mu'(t)$ are made equal to simulate a flat band of noise, that is, $n\omega_a A_n = \Delta$, the peak frequency deviation per sine wave. The mean

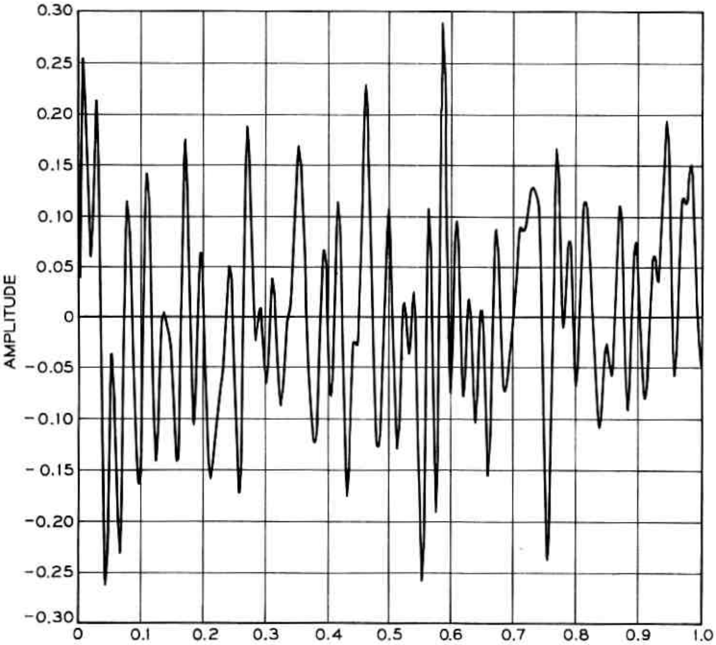


Fig. 2 — A periodic random noise sample for $N = 50$. Peak amplitude/rms amplitude $= |\,0.29\,|/[1/(2N)^{1/2}] = 2.90$.

square frequency deviation is

$$\sigma^2 = \frac{\Delta^2}{2} N = \frac{n^2 \omega_a^2 A_n^2}{2} \times N.$$

Substituting for ω_a from (5) we get

$$A_n = (\sigma/W)[(2N)^{\frac{1}{2}}]/n. \tag{7}$$

The rms phase and frequency deviations can be related to the RF bandwidth by Carson's rule which, for noise modulation, is written

$$B = 2W(1 + 4\sigma/W), \tag{8}$$

where the peak frequency deviation is assumed to be 4σ. Suppose that the line spectrum of (2) contains kN lines in addition to the carrier, then the bandwidth of the computed spectrum is

$$B = kN\omega_a . \tag{9}$$

From (5), (8), and (9) we get the relation between k and σ

$$k = 2(1 + 4\sigma/W). \tag{10}$$

This equation is as accurate as Carson's rule and is useful for estimating k when σ/W is given. If k is chosen too small, significant spectral components are omitted from the spectrum; the effect is to pass the complete spectrum through an ideal filter of bandwidth $kN\omega_a$.

In a similar manner k and φ can be related for the phase modulation case. The rms frequency deviation for the PM case is given by

$$\frac{\sigma}{W} = \varphi \sqrt{\frac{1 + \frac{3}{2N} + \frac{1}{2N^2}}{3}} \tag{11}$$

where N is the number of tones in the baseband. Substitution of (11) into (10) gives the desired result.

3.2 *Limitations on Modulation Index*

It has been shown (9), that the maximum RF spectrum bandwidth is given by $B = kN\omega_a$. From (5) the baseband bandwidth is $W = N\omega_a$. Assuming that only negligible energy falls outside B, then B is the RF bandwidth and the parameter k is a bandwidth expansion factor since

$$k = B/W. \tag{12}$$

Now, k and the rms frequency deviation σ are related by (10). The product kN is limited by the high speed storage capacity of the machine;

this implies a relationship between N and σ/W. Let $M \geqq kN$ be the maximum value of kN which can be accommodated in the machine. Then,

$$\sigma/W \leqq 1/4(M/2N - 1). \tag{13}$$

This expression is dependent upon Carson's rule and has the same unknown precision—but it serves to demonstrate the point that if large σ/W is desired, N must be made small. In the work reported here, $M = 500$ so that for $N = 10$, $\sigma/W \leqq 6$. Conversely for $N = 100$, $\sigma/W \leqq 0.375$.

Because Carson's rule has an unknown precision it is necessary to determine to reasonable accuracy the relationship between k and σ/W. With a perfect rectangular filter of bandwidth $kN\omega_a$, signal-to-distortion ratios have been computed for the case $N = 10$. In these computations, slots 1 and 10 were set to zero separately and the SDR computed for that slot.

The results are shown in Fig. 3 as a function of σ/W with the bandwidth expansion ratio k as a parameter. In all cases slot 1 has the lowest

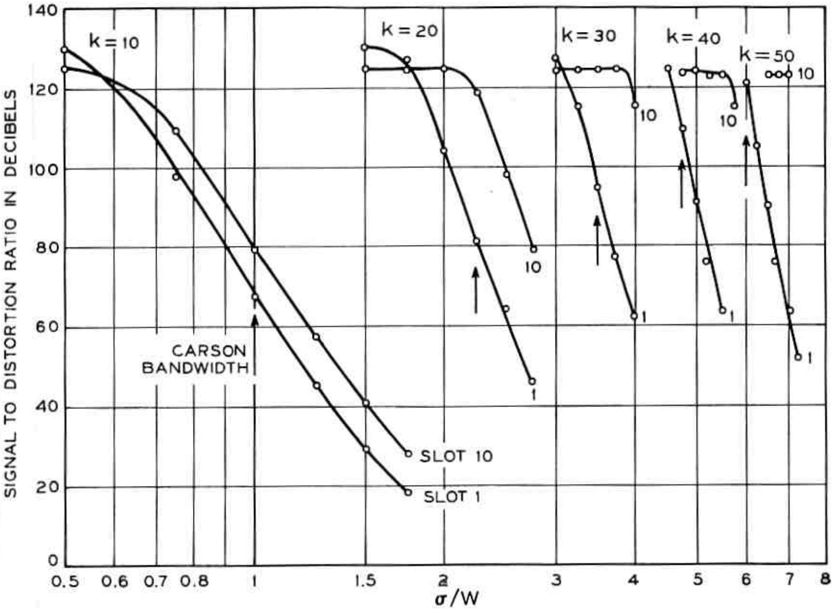


Fig. 3 — FM signal-to-distortion ratios for square filters containing $kN+1$ spectral lines and with $N = 10$.

SDR. The levelling off for SDR near 124 dB is probably caused by the computer round-off error. The negative slopes are the result of the finite filter bandwidth of $kN\omega_a$ and the decreasing accuracy of the method of harmonic interpolation in approximating the spectrum. Increasing k improves the accuracy of the approximation.

Values of σ/W obtained from Carson's rule in the form given in (10) are shown by the arrows in Fig. 3. Fig. 3 can be used to determine the value of k required to compute the SDR for a given σ/W. In all examples reported here, k and N have been chosen so that without a filter an SDR $\geqq$ 100 dB was obtained for the values of σ/W used. The data of Fig. 3 are averages of 20 noise samples.

IV. THE SINGLE POLE FILTER

The single-pole filter is the simplest possible realizable bandpass filter and is important for two reasons.

(*i*) It is widely used. For example, it is nearly optimum for use in the IF section of a frequency feedback receiver.[7]

(*ii*) As simple as it is, no previous method is adequate for the computation of FM distortion for high frequencies and large deviations.

4.1 *Single Sine Wave Modulation*

A number of years ago Bodtmann[8] made extensive measurements on a single-pole filter with both single sine wave and noise modulation.* Let us compare the measured and computed results.

The transfer function of a narrow band single-pole filter is

$$Y = \frac{1}{1 + j\dfrac{f - f_o}{f_c}} \tag{14}$$

where:

f_o is the center frequency and
f_c is the half bandwidth, that is, the frequencies at which the response is down 3 dB are $f_o \pm f_c$.

Bodtmann's filter was centered near 70 MHz with a half bandwidth of 1.223 MHz. The skirts fit the response of (14) to within ±0.1 dB out to the 15 dB loss points. The measured and computed ratios of signal-to-third harmonic distortion power are shown in Fig. 4. Notice

* It was Bodtmann's results which led to the discovery of a simple error in existing theories.[9–11]

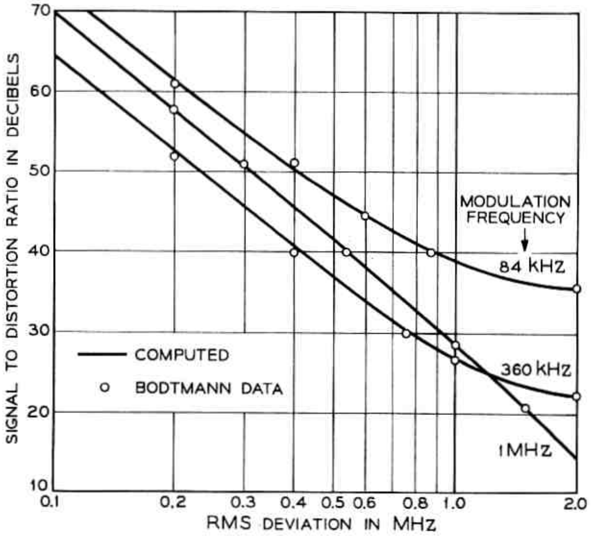


Fig. 4 — Third harmonic distortion in single pole filter.

the peculiarity which occurs at a deviation of 1.2 MHz where the curves for 360 KHz and 1 MHz modulation frequencies cross. Existing theories do not predict this behavior which is verified here by direct computation.

4.2 *Results for Random Modulation*

Computations of SDR have been made for a single pole filter for the random modulation discussed in Section III. The results, for noise samples of 10 and 50 sine waves of equal amplitude and random phase, are shown in Fig. 5 with Bodtmann's measured results. The computations followed the Monte Carlo procedure described previously. The data in Fig. 5 for $N = 50$ is the average over two slots at each frequency for 50 noise samples. The pairs of slots are 4 and 5, 17 and 19, and 49 and 50, corresponding to the slot frequencies 84 KHz, 360 KHz and 1 MHz, respectively. Data for all the slots were computed in the same computer run. In the computations for $N = 10$ one slot at a time was computed, each point being the average of 80 noise samples.

When the noise sample is simulated by 50 sine waves, the agreement with the experimental data is good. The SDR's for the case of 10 sine waves per noise sample are somewhat higher reflecting the

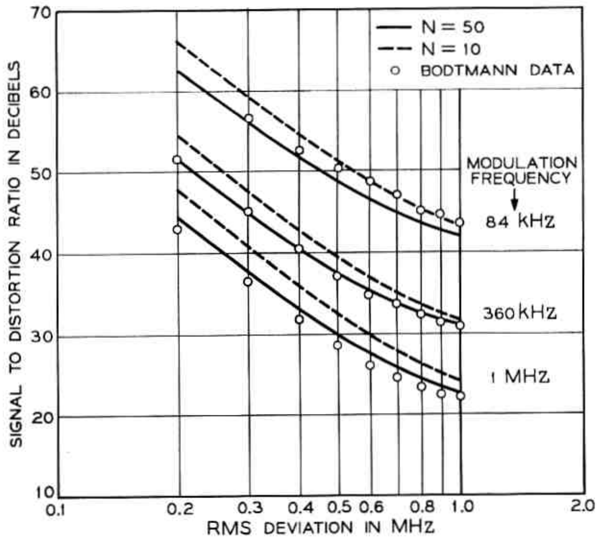


Fig. 5 — Bodtmann's measured results compared with noise samples.

fact that larger modulation peaks are to be found in the sample with the larger number of sine waves.[6]

4.3 *Convergence of the Monte Carlo Process*

The SDR's of 80 individual noise samples for $N = 10$ are shown in Fig. 6 in four sets of 20 each. The average SDR as a function of

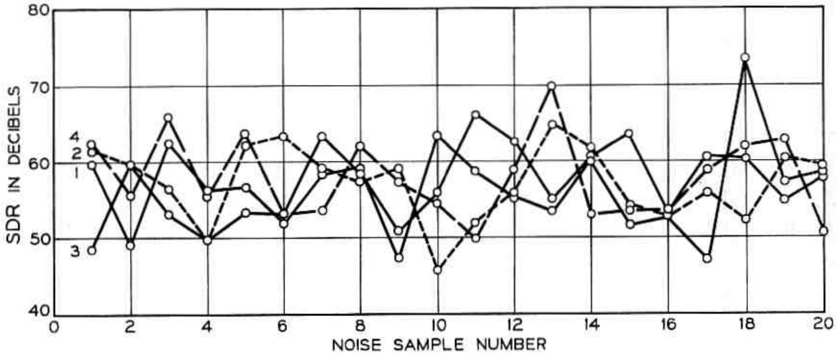


Fig. 6 — FM SDR in a single pole filter. Ten sine waves in baseband; SDR computed in slot 4; bandwidth expansion factor $k = 10$; $\sigma = 0.2$ MHz; $\omega_c/W = 1.223$.

the number of noise samples is shown in Fig. 7; the four sets of Fig. 6 are averaged in sequence. It is interesting to ask how close to the 80-sample average one would get if only 20 samples were used. As a partial answer, the four sets of Fig. 6 were averaged separately and the results are shown in Fig. 8. All four 20-sample averages fall within 1 dB of the 80-sample average.

Similar data for slot 19 is presented for the case $N = 50$ in Figs. 9, 10, and 11. Slot 17 was also computed and the averages for both slots are shown in Figs. 12 and 13. The results for slots 17 + 19 are remarkably similar to those of 19 alone. The 10-sample averages deviate from the 50 sample average by a maximum of 2.7 dB for slot 19 and 2.3 dB for the sum of slots 17 + 19. Interestingly enough, the 10-sample average for $N = 10$ deviates from the 80-sample average by a maximum of 2.2 dB.

The behavior of the SDR of a single noise sample as a function of σ/W is also of interest. Fig. 14 shows this behavior for each of the first six noise samples of set 1, Fig. 6, compared with the 80-sample

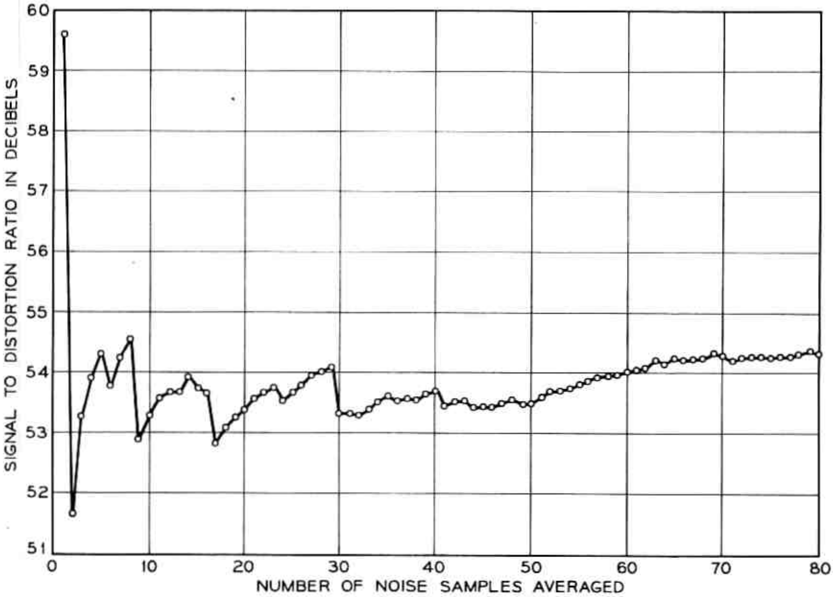


Fig. 7 — Fluctuations in SDR of single pole filter as a function of number of sets of computations. Ten sine waves in baseband; SDR computed in slot 4; bandwidth expansion factor $k = 10$; $\sigma = 0.2$ MHz; $\omega_c/W = 1.223$.

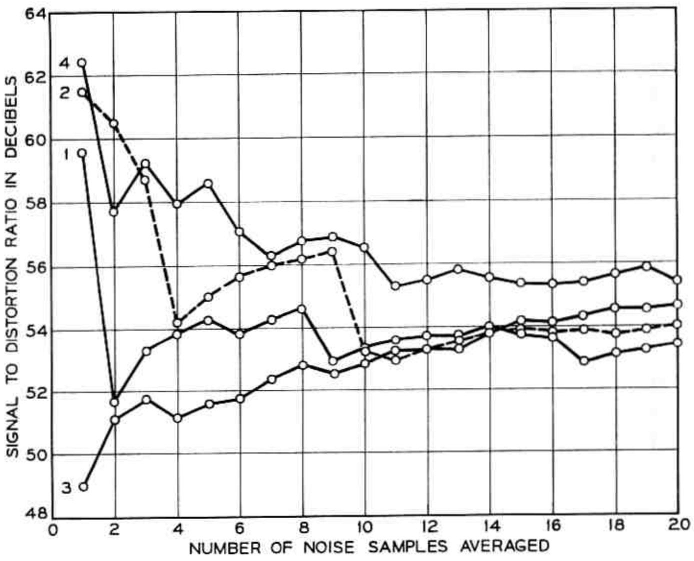


Fig. 8 — Fluctuations in SDR of single pole filter as a function of number of sets of computations. Ten sine waves in baseband; SDR computer in slot 4; bandwidth expansion factor $k = 10$; $\sigma = 0.2$ MHz; $\omega_c/W = 1.223$.

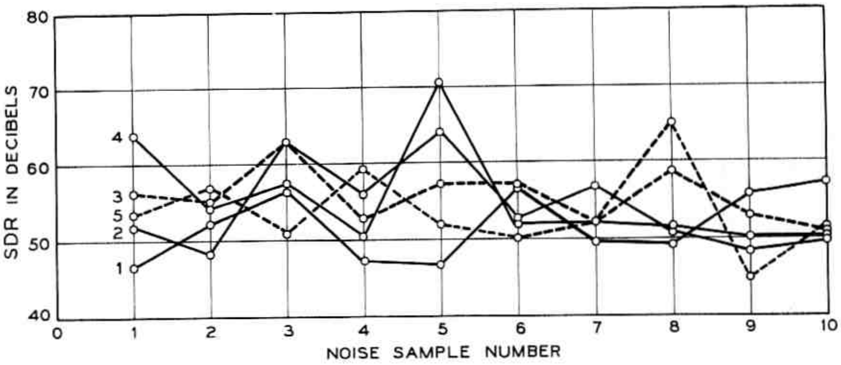


Fig. 9 — FM SDR in a single pole filter. 50 sine waves in baseband; SDR computed in slot 19; bandwidth expansion factor $k = 10$; $\sigma = 0.2$ MHz; $\omega_o/W = 1.223$.

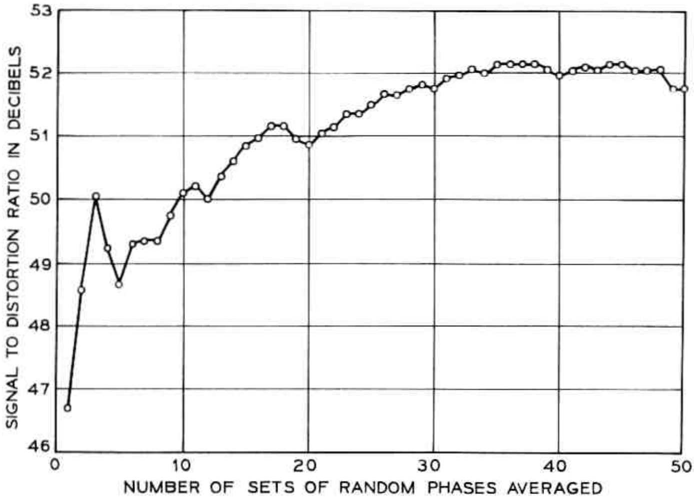


Fig. 10 — Fluctuations in SDR of single pole filter as a function of number of sets of computations. 50 sine waves in baseband; SDR computed in slot 19; bandwidth expansion factor $k = 10$; $\sigma = 0.2$ MHz; $\omega_c/W = 1.223$.

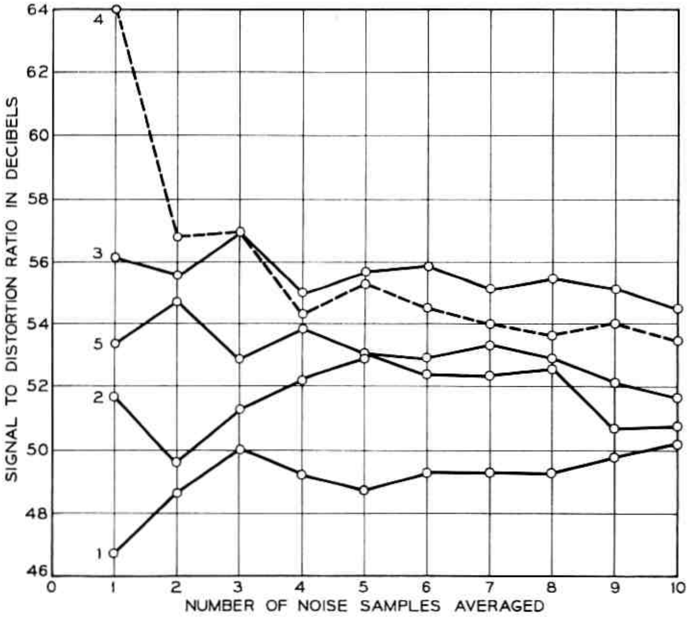


Fig. 11 — Fluctuations in SDR of single pole filter as a function of number of sets of computations. 50 sine waves in baseband; SDR computed in slot 19; bandwidth expansion factor $k = 10$; $\sigma = 0.2$ MHz; $\omega_c/W = 1.223$.

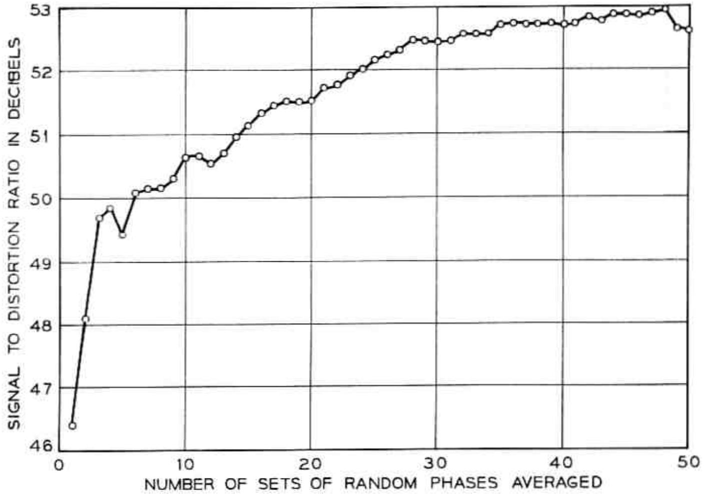


Fig. 12 — Fluctuations in SDR of single pole filter as a function of number of sets of computations. 50 sine waves in baseband; SDR computed in slots 17 + 19; bandwidth expansion factor $k = 10$; $\sigma = 0.2$ MHz; $\omega_c/W = 1.223$.

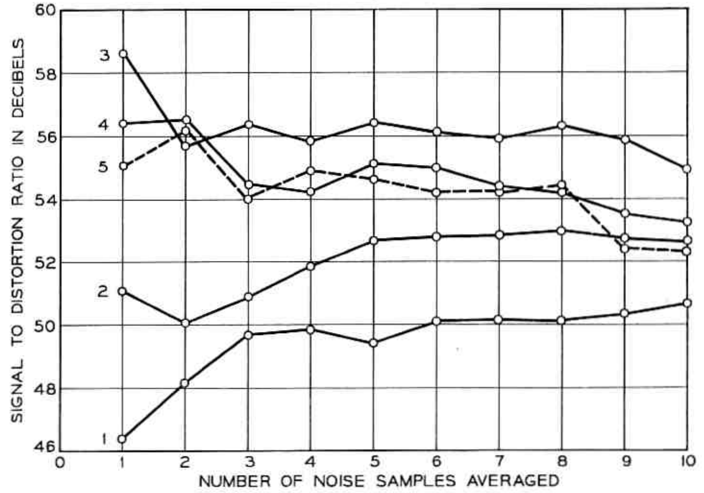


Fig. 13 — Fluctuations in SDR of single pole filter as a function of number of sets of computations. 50 sine waves in baseband; SDR computed in slots 17 + 19; bandwidth expansion factor $k = 10$; $\sigma = 0.2$ MHz; $\omega_c/W = 1.223$.

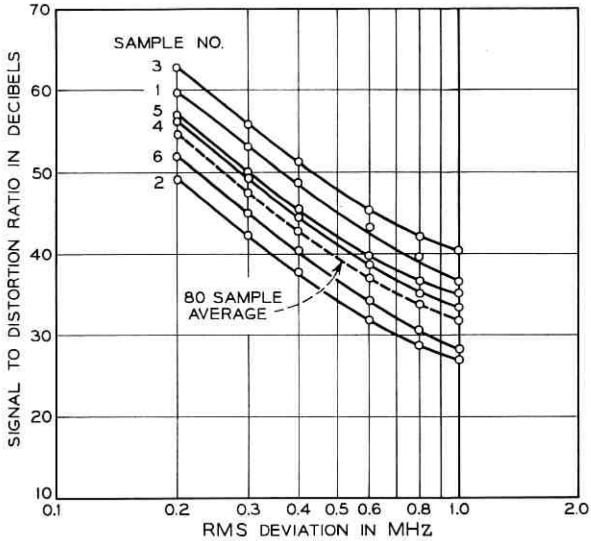


Fig. 14 — Behavior of the SDR in a single noise sample as a function of σ/W. Single pole filter; slot 4; $N = 10$; sample set 1.

average. The same behavior has been observed for other filters. It is clear that almost any noise sample will predict the SDR behavior as a function of σ/W, but the actual SDR computed for the single noise sample depends on the peakiness of the sample.

V. THE THREE-POLE MAXIMALLY FLAT AMPLITUDE FILTER

The maximally flat amplitude filter is used widely in frequency modulation systems; it has the flattest possible amplitude response near the midband frequency and is often used in conjunction with a phase equalizer. The transfer function of a narrow band three-pole bandpass filter is

$$Y = \frac{1}{1 - b_2\left(\dfrac{f - f_o}{f_c}\right)^2 + j\left(\dfrac{f - f_o}{f_c}\right)\left[b_1 - \left(\dfrac{f - f_o}{f_c}\right)^2\right]} \tag{15}$$

where

f_o is the midband frequency and

f_c is the filter half bandwidth; that is, the frequencies at which the response is down 3 dB are $f_o \pm f_c$,

b_1, b_2 are both equal to 2 for an MFA filter.

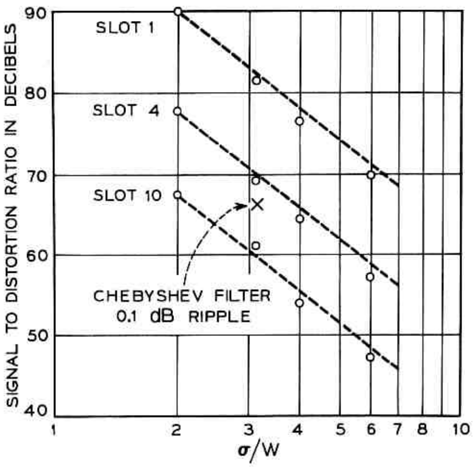


Fig. 15 — FM signal-to-distortion ratios in a three-pole MFA filter. $W = 7$ MHz; 3 dB filter bandwidth = 238 MHz; $N = 10$; $k = 50$; no carrier offset.

SDR computations for an unequalized filter are presented in Fig. 15 as a function of frequency deviation. The dashed lines are 12 dB per octave slopes placed arbitrarily to coincide with the data at $\sigma/W = 2$. The data points are 20-sample averages. The large cross is the SDR in slot 10 of a three pole 0.1 dB ripple Chebyshev filter with the same skirt selectivity as the MFA filter at a frequency 256

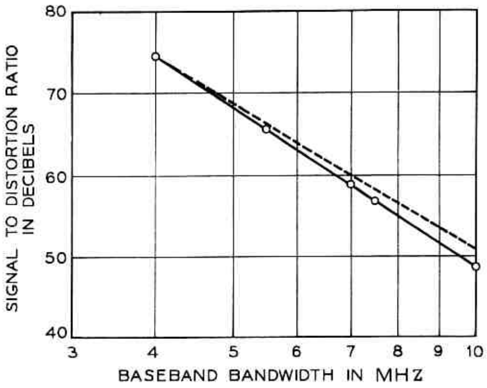


Fig. 16 — FM signal-to-distortion ratios in a three-pole MFA filter. $\sigma/W = 3.12$; 3 dB filter bandwidth = 238 MHz; $N = 10$; $k = 30$; slot 10; no carrier offset.

MHz from the carrier. The Chebyshev filter is clearly superior to the MFA filter in this instance. The SDR is a function of baseband W as shown in Fig. 16 for $\sigma/W = 3.12$ and slot 10. An arbitrary slope of 18 dB per octave is included. As in Fig. 15, the data points are 20-sample averages.

Fig. 17 shows the effect of a carrier frequency offset with respect to the filter midband frequency. In the application for which this filter was chosen, the midband frequency change over the ambient temperature range −40°F to +140°F is about ±6 MHz.

Results for perfect phase equalization are shown in Fig. 18; arbitrary slopes have been added. It is clear that nearly all of the distortion in the unequalized filter is due to nonlinear phase.

VI. AMPLITUDE TO PHASE CONVERSION

In addition to the FM distortion in the filter output there is generally some envelope distortion. Since all known limiters convert envelope modulation to phase modulation this source of distortion must be accounted for in system design. The envelope distortion is computed as described in Section II and it is necessary to relate it to the AM/PM conversion of the limiter.

For good limiters the AM/PM conversion is small and can be assumed linear, that is,

$$\theta = Qm \tag{16}$$

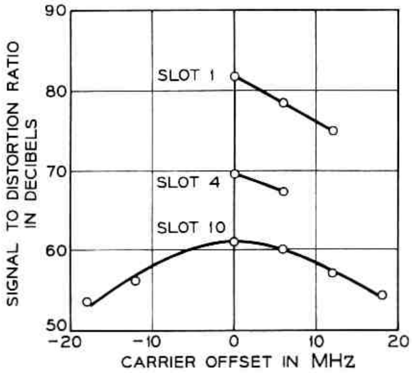


Fig. 17 — FM SDR in three-pole MFA filter as a function of carrier offset. $W = 7$ MHz; 3 dB filter bandwidth $= 238$ MHz; $N = 10$; $k = 50$; $\sigma/W = 3.12$.

where

m is the index of amplitude modulation for the slot of interest,
θ is the phase shift in radians in the same slot caused by m, and
Q is the AM/PM conversion coefficient.

The normal signal in the slot of interest is a sine wave of amplitude A. The signal-to-AM/PM distortion ratio is given by

$$\begin{aligned} \text{SDR (AM)} &= 20 \log A/\theta \\ &= 20 \log A/Qm \\ &= 20 \log A/m - 20 \log Q. \end{aligned} \tag{17}$$

The first term, $20 \log A/m$, can be computed for the network and the AM/PM conversion coefficient can be included separately.

The AM and FM SDR's for transitional Butterworth-Thomson filters[12] are plotted in Fig. 19. For the Chebyshev filter the AM and FM SDR are 72.3 and 66.3 dB, respectively. All filters were adjusted for equal loss 256 MHz from the midband frequency. The trends are

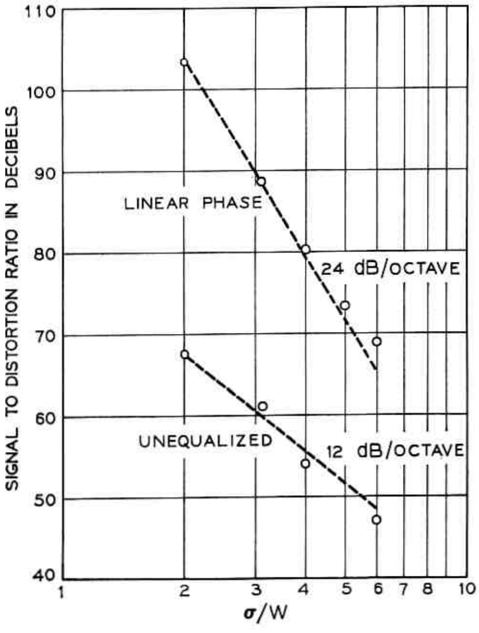


Fig. 18 — FM SDR in a phase-equalized three-pole MFA filter. $W = 7$ MHz; 3 dB bandwidth = 238 MHz; $N = 10$; $k = 50$; slot 10.

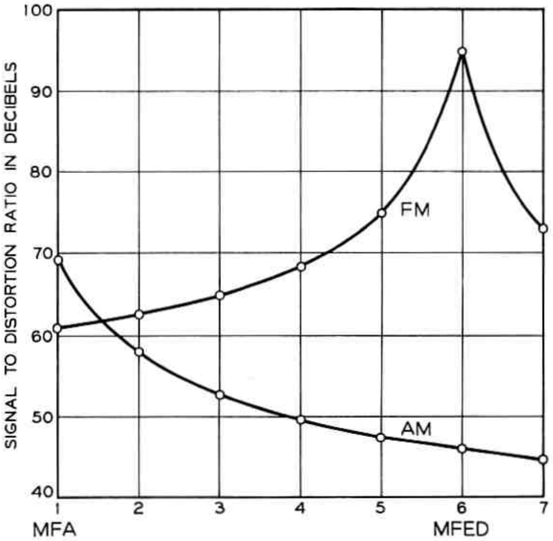


Fig. 19 — FM and AM SDR in three-pole transitional Butterworth-Thomson filters. $W = 7$ MHz; loss 256 MHz from midband = 20 db; $N = 10$; $k = 30$; no carrier offset; $\sigma/W = 3.12$; slot 10.

as expected, as the filter goes from MFA to maximally flat envelope delay (MFED) the FM distortion decreases and the AM/PM distortion increases. The effect of the limiter AM/PM conversion coefficient can be included by adding $-20 \log Q$ to the curve marked AM.

The frequency responses for the filters are given by (15); for the 0.1 dB ripple Chebyshev filter $b_1 = 1.921$, $b_2 = 1.801$. For the transitional Butterworth-Thomson filters the parameters are:

Filter No.	1-MFA	2	3	4	5	6-MFED	7
b_1	2.0	2.103	2.201	2.294	2.383	2.466	2.547
b_2	2.0	2.092	2.182	2.268	2.352	2.433	2.510

VII. DISCUSSION

The Fourier method for the computation of FM distortion in linear networks has been described and some results presented for single sine wave modulation and for random noise modulation simulated by groups of harmonically related sine waves. The method is exact to an accuracy determined by the round-off error in the machine.

Although the computation is exact for any individual input signal, the results for noise modulation are only approximate because the results depend upon averaging over a finite number of periodic noise samples. Much of the work described in this paper has been devoted to describing the behavior of the noise computations and in the determination of the maximum modulation index for which computations can be made with suitable accuracy.

In addition to demonstrating the nature of convergence of the noise averaging method, a detailed comparison of this method with the experimental results of W. F. Bodtmann provides an excellent demonstration of the extent to which a noise sample consisting of as few as 10 sine waves approximates a thermal noise signal. The noise simulation with a 10 sine wave noise sample is sufficient for most applications and accurate computations have been made for modulation indexes of $\sigma \leqq 6W$ where σ is the rms frequency deviation and W is the bandwidth of the modulating signal.

It is notable that a single periodic noise sample is sufficient to determine the shape of the curve describing the signal-to-distortion ratio as a function of the deviation, the baseband bandwidth, or the filter parameters. This result, illustrated in Fig. 14, can be used to conserve computational time when optimizing the parameters of a system.

ACKNOWLEDGMENTS

I am indebted to M. V. Schneider who guided me patiently through the programming maze as far as I have gone, and to T. L. Osborne who insisted from the beginning that the Chebyshev 0.1 dB ripple filter was better than my choice of the MFA filter.

REFERENCES

1. Roder, H., "Effects of Tuned Circuits Upon a Frequency Modulated Signal," Proc. I.R.E., *25,* No. 12 (December 1937), pp. 1617–1647.
2. Stumpers, F. W. L. M., "Distortion of Frequency Modulated Signals in Electrical Networks," Commun. News, *9,* No. 3 (April 1948), pp. 82–92.
3. Panter, P. F., *Modulation, Noise, and Spectral Analysis,* New York: McGraw-Hill, 1965, pp. 273–280.
4. Medhurst, R. G., and Roberts, J. H., "Evaluation of Distortion in FM Trunk Radio Systems by a Monte Carlo Method," Proc. I.E.E., *113,* No. 4 (April 1966), pp. 570–580.
5. Rice, S. O., "Mathematical Analysis of Random Noise," B.S.T.J., *23,* No. 3 (July 1944), pp. 282–332, and *24,* No. 1 (January 1945), pp. 46–156.
6. Bennett, W. R., "Distribution of the Sum of Randomly Phased Components," Quart. Appl. Math., *5,* No. 1 (April 1947), pp. 385–393.
7. Enloe, L. H., "Decreasing the Threshold in FM By Frequency Feedback," Proc. I.R.E., *50* (January 1962), pp. 18–30.

8. Bodtmann, W. F., unpublished work.
9. Enloe, L. H., and Ruthroff, C. L., "A Common Error in FM Distortion Theory," Proc. I.E.E.E., *51,* No. 5 (May 1963), p. 846.
10. Gladwin, A. S., Medhurst, R. G., Enloe, L. H., Ruthroff, C. L., "A Common Error in FM Distortion Theory," Proc. IEEE, *52,* No. 2 (February 1964), pp. 186–189.
11. Magnusson, R. I., Enloe, L. H., Ruthroff, C. L., "A Common Error in FM Distortion Theory," Proc. IEEE, *52,* No. 9 (September 1964) pp. 1082–1084.
12. Peless, Y., Murakami, T., "Analysis and Synthesis of Transitional Butterworth-Thomson Filters and Bandpass Amplifiers," RCA Review, *18,* No. 1 (March 1957), pp. 60–94.

Linear-Real Codes and Coders*

By WILLIAM H. PIERCE

(Manuscript received August 22, 1967)

In linear-real coding, the transmitted signals are (possibly redundant) linear combinations of the data signals. The linear combination of data signals can have a block pattern, resulting in linear-real block coders, or a stationary pattern, resulting in linear-real stationary (shift-register) coders. Stationary coding is shown to be a limiting case of block coding. Both methods appear to be practical for the control of burst and impulse noise. However, stationary coding appears to have some advantages and is the only one we study here. We propose shift register implementations which promise the required precision and dispersion at less cost than tuned RLC circuits.

Error properties of both block and stationary coders are similar, but it is easier to learn concepts by analyzing the block coders. When the receiver is able, by using some of the techniques we discuss, to estimate the noise covariance matrix for each codeblock, the resulting noise power is less than that for receivers not using the statistics for each codeblock.

Nonlinear memoryless filters, such as clippers, are especially effective when used with linear-real coders. We propose a memoryless filter which attenuates the input signal more severely when a second input to the filter indicates the channel is having a noise burst. If the memoryless filter is designed for the worst case noise, then performance will not degrade with decreased noise when the nonlinearity is odd and monotonic.

I. INTRODUCTION

Many communications channels, including telephone channels, contain noise which comes in short bursts, such as noise from impulses. Such noise is particularly deleterious when the channel is used for the transmission of digital data.

* Part of the research for this article was performed at Carnegie Institute of Technology under National Science Foundation grants GP-39 and GK-373. Some of the material contained in this paper is taken from the author's convention article.[9]

At least as early as 1958 it was discovered that it is sometimes possible to reduce digital errors in such channels without reducing the noise power by using a scheme such as Fig. 1 shows. In some formulations[1–7] the transformation A consisted of a continuous all-pass filter whose Fourier transform magnitude was unity at all frequencies but whose phase characteristic varied with frequency; the inverse linear transformation was the continuous all-pass filter with the conjugate phase characteristic. The linear filter was called the smear operation, and its inverse the desmear operation. Later papers considered linear transformations to be real-number matrices operating upon the data in blocks.[8–10]

In all schemes to which Fig. 1 applies, a single impulse of noise into the inverse linear filter will be transformed into an output noise which is dispersed in time. With proper design, this dispersed noise will be small enough at all times to not produce errors at the output of the quantizer.

Our purpose is to investigate coding schemes which fall in the general pattern of Figure 1 to gain conceptual insight and learn practical design. Such study is useful because the practicality of the matrix version has never been studied, and the continuous all-pass filter was limited by cost and filter imprecision. The shift registers we might propose avoid the problems which hindered the application of continuous all-pass filters.

We show that the real-number linearity of the transformations of Fig. 1 will permit the receiver to use any available information about noise correlation or position. All of the proposed means for using this information are simple in concept, and some are simple to implement.

II. DESCRIPTION

Linear-real block coding is a form of coding in which A, an n by k matrix of real numbers, is used to produce an output vector $\mathbf{b}$ from an input vector $\mathbf{r}$ according to the equation

$$\mathbf{b} = A\mathbf{r}. \tag{1}$$

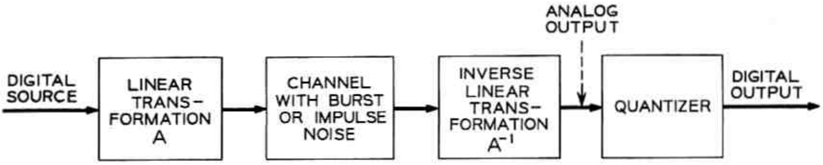

ANALOG OUTPUT
DIGITAL SOURCE
LINEAR TRANS-FORMATION A
CHANNEL WITH BURST OR IMPULSE NOISE
INVERSE LINEAR TRANS-FORMATION A^{-1}
QUANTIZER
DIGITAL OUTPUT

Fig. 1 — A general arrangement for placing linear filters A and A^{-1} to reduce digital errors.

If $n > k$, then $\mathbf{b}$ will be redundant in the sense that not all of its components are independent. The word "real" is used in order to emphasize the fact that the arithmetic in equation 1, and all other equations in this paper, is real number arithmetic. The use of real number arithmetic distinguishes this work from generalized parity-check coders which are linear in finite-field arithmetic.

Stationary (shift register) linear-real coding is a limiting case of linear-real block coding, but is best described as being the convolution summation given by

$$b_i = \sum_{j=-\infty}^{\infty} h_{i-j} r_j \tag{2}$$

where b_i is the i^{th} signal transmitted, r_i is the i^{th} data number, and where h_q can naturally be called the unit pulse response of the encoding filter at time-step q.

The conclusions to be reached on practical applications are that moderate cost encoders and decoders of considerable use for burst and impulse noise channels can be built as soon as low-cost tapped digital delay lines are available. Magnetic domain-wall digital delay lines,[11] for example, might well make these coders practical.

There are two general ways in which noise is controlled by means of linear-real coding. We give the complete details and mathematics later. Briefly, the qualitative aspects are:

The total noise power in the decoded signal is made less than that without coding. We discuss three distinct ways of doing this:

(*i*) When linear-real block coding is used, and when the noise covariance matrix is known (or can be adaptively deduced by the receiver) then this knowledge can be used to reduce the noise power. It can be correlation type knowledge, as accounts for the effectiveness of Wiener filtering. If the noise process is *a posteriori* nonstationary, then a receiver which estimates the noise correlation matrix for each code block may effectively use the available information on the position of burst noises within the block. This is particularly effective in burst noise channels having block coders using rectangular A matrices.

(*ii*) A stationary memoryless nonlinear filter (such as a clipper) can be used to reduce the noise power before the inverse linear transformation is applied. Such a filter would of course reduce noise power in the absence of an inverse filter when it immediately precedes the quantizer, but it would not then reduce errors. When placed before the inverse transformation, the stationary memoryless nonlinear filter reduces both errors and noise power. We refer to equations for analyzing design and performance of the memoryless nonlinear filter. A simulation example in Section VI shows these devices to be surprisingly effective.

(*iii*) A memoryless nonlinear filter can be used which has both the noisy signal and an estimate of the instantaneous noise power for inputs. The output is an optimized estimate of the signal given the estimated instanteous noise power. This filter always reduces noise power, as does the filter in method *ii*, and only reduces errors if there is a filter such as the inverse linear transformation between it and the quantizer. We describe several methods for estimating the instantaneous noise power in Section V. One of these, which appears in Fig. 6, uses the fact that practical pam signals have more bandwidth then the Nyquist bandwidth for their pulse interval.

The remaining noise power is distributed more evenly among all decoded signal components and (in the limit of infinite smearing) made Gaussian. This type of noise control is especially effective in quantized-signal burst and impulse noise channels which have a thermal noise which is small compared with the separation between quantization levels. In this case a burst noise with power which is small compared with the thermal noise would be unable to produce many errors if it were evenly dispersed, although it could when bunched up. Dispersal of the burst noise power is sometimes unfavorable, but if the noise power is reduced enough and the noise dispersed enough, then the effect is very favorable. The decoding operation also tends to make the decoded signal have a Gaussian first-order probability distribution, which reduces the probability of a large peak and thereby reduces errors for quantized signals.

The design equations for the nonlinear memoryless filter (clipper) to which we refer assume a known probability distribution on the noise, as does the simulation reported. In practice, the actual noise can be less noisy than that used for design purposes, and the resulting mean square error will not be larger than that with the design noise, provided the noise probability density is even and the nonlinearity has certain properties. We give precise details in Appendix D.

III. BLOCK CODES AND THEIR NOISE COVARIANCE MATRIX

In general, assuming $\mathbf{r}$ and $\mathbf{c}$ are independent zero-mean column vector random variables, which represent the signal to be encoded and the channel noise, respectively, and assuming **r** and **c** have nonsingular covariance matrixes Q and N, respectively, and assuming that $\mathbf{f} = \mathbf{b} + \mathbf{c}$ is decoded by some linear operator T, where **b** is given in equation 1, then a straightforward evaluation of the covariance matrix of $\mathbf{u} = \mathbf{r} - T\mathbf{f}$ will show that

$$\begin{aligned} M &= E[\mathbf{u}\mathbf{u}'] \\ &= (I_{(k\times k)} - TA)Q(I_{(k\times k)} - TA)' + TNT' \end{aligned} \tag{3}$$

where $(\ )^t$ denotes the transpose of a matrix or column vector. This formula can be used to compare the performance of encoder-decoder pairs with good and bad choices for matrix A, and good and bad choices of matrix T.

Table I shows three possible T matrices. The first was shown to be the least mean square linear estimator in (9), and for Gaussian signal and noise gives the conditional mean of the transmitted vector given the received vector. The second is the first evaluated for infinite signal power in all degrees of freedom (which implies $Q^{-1} = 0$) and produces a decoded error uncorrelated with the signal. The third does not require the use of the N matrix. All assume the columns of A to be linearly independent.

Table II gives further insights into the behavior of the decoded error by presenting a number of special cases of equation (3). The justification of the equations of Table II is given in Appendix A. In one of the special cases in Table II, namely when equation (7) applies, the decoded noise energy is proportional to the arithmetic mean of the received noise energy. In other cases, such as that of equation (12), the eigenvalues of $A^tN^{-1}A$ play a crucial role in formulas for the mean square decoded noise.

Equation (13) of Table II shows that the average of the eigenvalues of $A^tN^{-1}A$ appears in a formula for a lower bound for the mean

TABLE I—THREE DIFFERENT LINEAR OPERATORS FOR DECODING f INTO r.

Name	Formula
Mean estimator (Gives least mean square error)	$T = (Q^{-1} + A^tN^{-1}A)^{-1}A^tN^{-1}$
Unattenuated estimator	$T = (A^tN^{-1}A)^{-1}A^tN^{-1}$
Unadaptive estimator (The generalized inverse of A)	$T = (A^tA)^{-1}A^t$

TABLE II — SOME SPECIAL CASES OF THE ERROR COVARIANCE MATRIX OF EQUATION (3) AND THE RESULTING MEAN SQUARE ERROR

Unadaptive Estimator

$$M = (A^tA)^{-1}A^tNA[(A^tA)^{-1}]^t. \tag{4}$$

$$\text{m.s. error} = 1/k \operatorname{tr} (A^tA)^{-1}A^tNA[(A^tA)^{-1}]^t. \tag{5}$$

When the columns of A are orthogonal and each of length $(n/k)^{1/2}$:

$$M = (k/n)^2A^tNA. \tag{6}$$

When in addition $N = \operatorname{diag}(n_1, n_2, \cdots, n_n)$, and AM is the arithmetic mean of these n_i's, and A is $1/(k)^{1/2}$ times the first k columns of a Hadamard matrix (see Appendix A for a definition):

$$\text{m.s. error} = M_{ii} = (k/n)AM. \tag{7}$$

Unattenuated Estimator

$$M = (A^tN^{-1}A)^{-1}. \tag{8}$$

$$\text{m.s. error} = 1/k \operatorname{tr} (A^tN^{-1}A)^{-1}. \tag{9}$$

Mean Estimator

$$M = (Q^{-1} + A^tN^{-1}A)^{-1}. \tag{10}$$

$$\text{m.s. error} = 1/k \operatorname{tr} (Q^{-1} + A^tN^{-1}A)^{-1}. \tag{11}$$

Mean Estimator ($\Omega = 1$) or Unattenuated Estimator ($\Omega = 0$)

$$\text{m.s. error} = 1/k \sum_{i=1}^{k} \frac{1}{\lambda_i(\Omega Q^{-1} + A^tN^{-1}A)} \tag{12}$$

where $\lambda_i(Z)$ denotes the i^{th} unordered eigenvalue of Z. Special case of above when $Q = sI$, s scalar:

$$\text{m.s. error} = 1/k \sum_{i=1}^{k} \frac{1}{\Omega s^{-1} + \lambda_i(A^tN^{-1}A)} \geq \frac{1}{\Omega s^{-1} + 1/k \sum_{i=1}^{k} \lambda_i(A^tN^{-1}A)}. \tag{13}$$

Special case of equation (12) when $Q = sI$, and A is square, orthogonal, and each column has length $(n/k)^{1/2}$:

$$\text{m.s. error} = 1/k \sum_{i=1}^{k} \frac{1}{s^{-1} + \dfrac{1}{\lambda_i(N)}}. \tag{14}$$

The following assumptions are referred to as equation (15):

$\Omega = 1$: T is the mean estimator.
$\Omega = 0$: T is the unattenuated estimator, and $A^tN^{-1}A$ is positive definite.
$Q = sI$, s scalar.
A is $1/(k)^{1/2}$ times the first k columns of an $n \times n$ Hadamard matrix.
$N = \operatorname{diag}(n_1, n_2, \cdots n_n)$.

The n_i variables are independent, identically distributed random variables such that $E(1/n_i)$ exists, has finite variance σ^2, and the harmonic mean of the n_i variables

$$HM = \left[1/n \sum_{i=1}^{n} 1/n_i \right]^{-1} \tag{15}$$

is finite.
k is large enough for the weak law of large numbers to apply.

Assuming equation (15):

$$\frac{1}{\Omega s^{-1} + \dfrac{n}{k\,HM}} \leq \text{m.s. error} \tag{16}$$

$$\text{m.s. error} \leq \frac{0.5}{s^{-1} + \dfrac{n}{k\,HM} - \sqrt{\dfrac{\sigma 2_n}{\tau}}} + \frac{0.5}{s^{-1} + \dfrac{n}{k\,HM} + \sqrt{\dfrac{\sigma 2_n}{\tau}}} \tag{17}$$

provided that the first denominator is positive, where τ is given by equation 35 of Appendix I.

square error; furthermore, that the mean square error equals this lower bound only when all the eigenvalues are the same. Thus the deviations of the eigenvalues of $A^tN^{-1}A$ detemine the closeness of the lower bound of equation (13), which Appendix A shows is sometimes related to the harmonic mean of the eigenvalues of N, which appears in equations (16) and (17).

A geometric illustration of the eigenvalues of $A^tN^{-1}A$ for rectangular A with orthonormal columns begins with the observation that the eigenvectors of N^{-1} form the semiaxes of an n-dimensional ellipsoid. The projection of this ellipsoid by the transformation A^t forms another ellipsoid, which will be called the k-dimensional shadow of the original n-dimensional ellipsoid.*

The semiaxes of the shadow ellipsoid have the lengths of the eigenvalues of $A^tN^{-1}A$. In order for the equation (13) bound to be close to the actual value, the semiaxes of the shadow ellipsoid have to be generally near their mean length; in other words, the shadow has to be round. A sufficient condition for the shadow to be round is that the ellipsoid is the shadow of a round ellipsoid, but this is not necessary. For some of the possible spacial orientations, for example, a football's shadow is rounder than the football.

IV. THE LIMITING CASE OF STATIONARY (SHIFT REGISTER) CODING

The purpose of this section is to show that—in the limit— all linear-real coding and decoding operations can become time stationary, so that they can be implemented by shift registers with time-invariant impulse responses. The limit is taken in the sense that the transmitted digits are obtained as a single block code whose output is a column

* An ordinary planar shadow of a three-dimensional object will be an orthogonal projection only when the light rays are parallel, and are normal to the plane of the shadow.

vector with components from $-n$ to n, where n approaches infinity.

There are two reasons why a study taking linear-real coding to the limit of being time stationary can be advantageous or useful:

(*i*) Stationary encoders and decoders appear to be more economical to implement than the block type of encoders and decoders.

(*ii*) The mathematical investigations to be made in the passage to the limit will add insights to linear-real coding by showing that a special case of it is Wiener filtering, and will add insights to Wiener filtering by showing that a Wiener filter is related to the least mean square estimator of matrix-encoded noise data vectors.

Toeplitz matrices, defined later, and Z-transforms (Ragazzini and Franklin),[12] are our main mathematical techniques to reach these ends.

4.1 *Stationary Coders*

The transmitted signal b_i is assumed to be obtained from the data stream r_j by the convolution summation of equation (2), which can be put in matrix form by means of the doubly infinite vectors

$$\mathbf{b} = \begin{bmatrix} \vdots \\ b_{-1} \\ b_0 \\ b_1 \\ \vdots \end{bmatrix}, \qquad \mathbf{r} = \begin{bmatrix} \vdots \\ r_{-1} \\ r_0 \\ r_1 \\ \vdots \end{bmatrix}, \qquad \text{etc.,}$$

and the Toeplitz matrix (defined in section 4.2)

$$A_{ij} = a_{i-j} = h_{i-j}$$

so that equation (2) can be expressed in matrix form by

$$\mathbf{b} = A\mathbf{r}.$$

The problem of how to perform the infinite matrix multiplications, either analytically or with hardware, will be shown to be solvable by the use of Z-transforms.

4.2 *Infinite Toeplitz Matrices*

An infinite matrix A, with elements A_{ij} , $i, j = 0, \pm 1, \pm 2, \cdots$, will be called Toeplitz* if some sequence $\ldots , a_{-1} , a_0 , a_1 , \ldots$ exists

* Hermitian matrices of the type of Equation (14) are called Toeplitz forms, and are described by Grenander and Szego.[13] The Hermitian property is not assumed in this paper's definition, since it is not needed for some of the results.

such that

$$A_{ij} = a_{(i-j)} \tag{18}$$

for all i, j. Associated with this Toeplitz matrix will be the two-sided Z-transform

$$a(z) = \sum_{q=-\infty}^{\infty} a_q z^{-q}. \tag{19}$$

The convergence properties of Toeplitz matrices could prove troublesome in some cases, but in this paper most difficulties will be avoided by using only those matrices whose associated Z-transform, according to equations (18) and (19), has all its poles some finite distance from the circle $|z| = 1$, and which is absolutely convergent on $|z| = 1$. (If the matrix is to be inverted, it also must have its zeros some finite distance from $|z| = 1$.)

Any poles outside $|z| = 1$ arise from a_q sequences which are nonzero for $q < 0$. This should not cause alarm, as noncausality of unit pulse reponses for decoders is not a serious practical obstacle, since actual noncausal unit pulse responses can be arbitrarily well approximated by accepting a decoding delay. These restrictions on the poles of the associated Z-transforms require that a_q be bounded by a geometrically decreasing sequence as $q \to \pm\infty$.

Section B.1 of Appendix B presents theorems which are useful in relating Toeplitz matrix operations to Z-transforms, and shows how least mean square matrix operators of the Toeplitz type can be related to Weiner-filter types of sampled data estimators.

4.3 *Error Analysis*

When A, Q, and N are Toeplitz and nonsingular, the expressions for the mean square error equivalent to the equations of Table 2 are

$$T_{\text{MEAN}} = (Q^{-1} + A^t N^{-1} A)^{-1} A^t N^{-1}$$

or

$$t_{\text{MEAN}}(z) = \frac{q(z)a\left(\frac{1}{z}\right)}{n(z) + a\left(\frac{1}{z}\right)q(z)a(z)}$$

gives

$$\begin{matrix}\text{m.s.}\\ \text{error}\end{matrix} = \begin{bmatrix}\text{on diagonal component} \\ \text{of } M_{\text{MEAN}} = (Q^{-1} + A^t N^{-1} A)^{-1}\end{bmatrix} \tag{20}$$

or

$$\begin{matrix}\text{m.s.}\\ \text{error}\end{matrix} = Z^{-1}\left\{\frac{q(z)n(z)}{n(z) + a\left(\frac{1}{z}\right)q(z)a(z)}\right\}\Bigg|_{k=0} . \tag{21}$$

where Z^{-1} is the inverse Z-transform integral operator. When A is either finite or Toeplitz but nonsingular, $T_{\text{UNATTENUATED}}$ and $T_{\text{UNADAPTIVE}}$ give the same decoding matrix, namely A^{-1}, which will be called T_{INVERSE}.

$$T_{\text{INVERSE}} = A^{-1}$$

or for the Toeplitz case

$$t_{\text{INVERSE}}(z) = \frac{1}{a(z)}$$

gives

$$\begin{matrix}\text{m.s.}\\ \text{error}\end{matrix} = \begin{bmatrix}\text{on diagonal component}\\ \text{of } A^{-1}N(A^{-1})^{t}\end{bmatrix} \tag{22}$$

$$\begin{matrix}\text{m.s.}\\ \text{error}\end{matrix} = Z^{-1}\left\{\frac{n(z)}{a(z)a\left(\frac{1}{z}\right)}\right\}\Bigg|_{k=0} . \tag{23}$$

The above error can be evaluated by these three methods:

(*i*) Truncate A and N and then compute an on-diagonal component of $(A^{t}N^{-1}A)^{-1}$ near the center of the matrix.
(*ii*) Use Z-transforms to find $t_{\text{UNATTENUATED}}(z)$. Invert the Z-transform by either
(*a*) Using the inversion integral for Z-transforms, or
(*b*) Using pole-zero expansions and a small table of Z-transforms.

Method (*ii-a*) is the Z-transform analog of using Parseval's theorem to find mean square errors of stationary nonsampled systems.

Lemma 1: *When A is Toeplitz with columns orthogonal and of length* 1, *then*

$$(a) \qquad A^{t}A = I$$

$$(b) \qquad a\left(\frac{1}{z}\right)a(z) = 1.$$

The proof is trivial. Also notice that $(a) \Leftrightarrow (b)$.

Corollary 1: *When* $T = A^{-1}$ *and* A *is orthogonal,*

$$\begin{matrix} m.s. \\ error \end{matrix} = m.s.\ noise.$$

For design purposes it is desirable to make the following definitions; both assume $T = A^{-1}$ which is assumed to exist.

For Toeplitz A *and* N:

$$\begin{matrix}\text{noise} \\ \text{power} \\ \text{amplification}\end{matrix} = \frac{\begin{bmatrix}\text{on diagonal component} \\ \text{of } A^{-1}N(A^{-1})^t\end{bmatrix}\begin{bmatrix}\text{on diagonal component} \\ \text{of } A^tA\end{bmatrix}}{[\text{on diagonal component of } N]} \tag{24}$$

For Block Coders:

$$\begin{matrix}\text{noise} \\ \text{power} \\ \text{amplification}\end{matrix} = \frac{\left[\frac{1}{k}\operatorname{tr} A^{-1}N(A^{-1})^t\right]\left[\frac{1}{k}\operatorname{tr} A^tA\right]}{\left[\frac{1}{k}\operatorname{tr} N\right]}. \tag{25}$$

Physically, this corresponds to the actual amplification of noise in a channel which encodes with a matrix proportional to the A matrix, where the proportionality constant is selected to make the encoder give unity power amplification to a white signal, and where the decoder is T_{INVERSE}. For the stationary coder and channel, the Z-transform version is:

$$\begin{matrix}\text{noise} \\ \text{power} \\ \text{amplification}\end{matrix} = \frac{Z^{-1}\left\{\dfrac{n(z)}{a(z)a\left(\frac{1}{z}\right)}\right\}\Bigg|_{k=0} Z^{-1}\left\{a(z)a\left(\frac{1}{z}\right)\right\}\Bigg|_{k=0}}{n_0}. \tag{26}$$

The block code version of the trace formula can also be used to show that if the impulse response of the stationary encoder is $\ldots a_{-1}, a_0, a_1, \ldots$, and its inverse is $\ldots b_{-1}, b_0, b_1, \ldots$, so that $a_q * b_q = \delta_{q,0}$, then for $N \propto I_\infty$ the noise power amplification can be evaluated from the impulse reponses by:

$$\begin{matrix}\text{noise} \\ \text{power} \\ \text{amplification} \\ \text{(for white noise)}\end{matrix} = \left[\sum_{q=-\infty}^{\infty} a_q^2\right]\left[\sum_{q=-\infty}^{\infty} b_q^2\right]. \tag{27}$$

The Z-transform version for $N \propto I_\infty$ is:

$$\begin{array}{l}\text{noise}\\ \text{power}\\ \text{amplification}\\ \text{(for white noise)}\end{array} = Z^{-1}\left\{\frac{1}{a(z)a\left(\frac{1}{z}\right)}\right\}\Bigg|_{k=0} \quad Z^{-1}\left\{a(z)a\left(\frac{1}{z}\right)\right\}\Bigg|_{k=0} . \tag{28}$$

It can be readily seen from equation (26) that:

Lemma 2: *When A is Toeplitz, the noise power amplification will be unity whenever*

$$a(z)a\left(\frac{1}{z}\right) = \text{constant}$$

whether or not the noise is white, so long as it is Toeplitz.

An equivalent statement is that when A and N are Toeplitz, a sufficient condition for the noise power amplification to be unity is that $A^t A = I_\infty$, which is equivalent to $a(z)a(1/z) = $ constant.

The above lemma will be seen to be especially significant after it is proved that unity noise power amplification is the least which can ever be obtained, and when it is shown that simple $a(z)$ functions, namely all-pass functions, obey the conditions of the lemma. Notice that the noise power amplification definition was based upon a receiver which performed the inverse of the encoding operation, and not upon a receiver which made a least square estimate of the signal given the *a posteriori* noise statistics. Consequently, statements about least possible noise power amplification are not applicable to adapative types of receivers such as those employing T_{MEAN}.

The following theorem is for block codes with $n = k$.

Theorem 1: *When square block coding is used and N is proportional to the identity, then the noise power amplification is always greater than or equal to one, and it is one only when A is proportional to an orthogonal matrix.*

Proof: What is required is a demonstration that:

$$(i) \qquad \frac{1}{k^2}[\text{tr } A^{-1}(A^{-1})'][\text{tr } AA'] \geqq 1 \tag{29}$$

and

(*ii*) Equality occurs if and only if A is proportional to an orthogonal matrix. (30)

These are established in Section 2 of Appendix B.

The following corollary is the Toeplitz matrix limit version of the above.

Corollary 2: When A and N are Toeplitz, and N is proportional to I_∞, a necessary and sufficient condition for unity noise power amplification is that $A^tA = I_\infty$, which is equivalent to a $(1/z)\ a(z) =$ constant. Otherwise the noise power amplification is greater than one.

When stationary (shift-register) linear-real decoding is used, then the decoding filter passes the noise through a Z-transform transfer function. When the noise is statistically stationary, the expected value of the mean square of the output noise is stationary, and depends only upon the amplitude of the transfer function averaged over the values of z. However, for burst noise the variance of the mean square of the decoded noise does depend upon the phase of the transfer function. For burst-noise or impulse-noise channels, this variance is minimized if the impulse response from the noise to the analog output of the decoder consists of many small terms instead of a few big ones.

For quantized signals it is important to minimize the variance of noise power because fluctuations above the mean of the variance increase the error rate far more than fluctuations below the mean of the variance decrease it. In order to make the variance of the noise power small, the impulse response from noise to analog output must be near its peak for many times longer than the periods of fluctuation in the noise process.

Because trace and expected value operators commute, the expected value of the output mean square error can be found by substituting $E(N)$ where N appears, provided the noise process is stationary. This cannot be done for error probabilities after the quantizer, however.

4.4 *All-Pass Z-Transforms*

A Z-transform $a(z)$ is defined to be *all-pass* if $|a(z)| =$ *constant for* $|z| = 1$. These are the Z-transform version of two-sided Laplace (or Fourier) transformed all-pass functions. Figure **2** shows some important properties of all-pass Z-transforms, including the fact that $a(z)a(1/z) =$ *constant* is an alternative definition of an all-pass Z-transform. The proofs of relationships in the figure not proved previously are straightforward. The practical implications of these relationships are that all stationary (shift-register) linear-real coders should have Z-transforms which are all-pass, in order not to increase the noise power amplification.

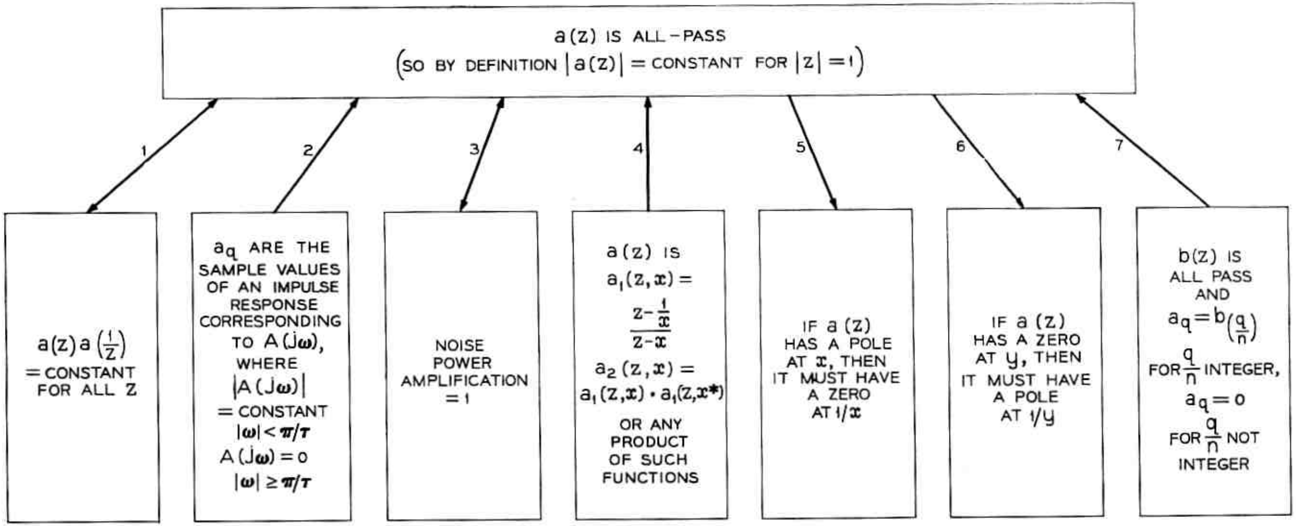

a(z) IS ALL-PASS
(SO BY DEFINITION $|a(z)|$ = CONSTANT FOR $|z| = 1$)
1
2
3
4
5
6
7
$a(z)a\left(\frac{1}{z}\right)$ = CONSTANT FOR ALL Z
a_q ARE THE SAMPLE VALUES OF AN IMPULSE RESPONSE CORRESPONDING TO $A(j\omega)$, WHERE $|A(j\omega)|$ = CONSTANT $|\omega| < \pi/\tau$
$A(j\omega) = 0$ $|\omega| \geq \pi/\tau$
NOISE POWER AMPLIFICATION = 1
a(z) IS
$a_1(z, x) = \frac{z - \frac{1}{x}}{z - x}$
$a_2(z, x) = a_1(z, x) \cdot a_1(z, x^*)$
OR ANY PRODUCT OF SUCH FUNCTIONS
IF a(z) HAS A POLE AT x, THEN IT MUST HAVE A ZERO AT $1/x$
IF a(z) HAS A ZERO AT y, THEN IT MUST HAVE A POLE AT $1/y$
b(z) IS ALL PASS AND
$a_q = b\left(\frac{q}{n}\right)$ FOR $\frac{q}{n}$ INTEGER,
$a_q = 0$ FOR $\frac{q}{n}$ NOT INTEGER

Fig. 2 — Some important relationships for all-pass Z networks.

V. A. Kisel' has made an excellent short study of all-pass Z-transforms, with a view toward using them as phase-correcting networks.[14] He has shown that networks whose Z-transform transfer function are of the form

$$a(z) = \frac{1 + \beta_1 z + \beta_2 z^2 + \beta_3 z^3}{\beta_3 + \beta_2 z + \beta_1 z^2 + z^3}$$

are all-pass, and that Fig. 3 synthesizes such functions. Additional modifications are added to this basic structure and implementations are proposed in the next section.

V. IMPLEMENTATION STUDIES

The decoder for block coding with adaptive mean decoding appears to require a large modern digital computer, and even then it could probably only operate "on line" with a slow channel and a block size not much over one hundred. Further research may lead to A matrices for which $(Q^{-1} + A^t N^{-1} A)$ can be easily inverted for realistic Q and N, or further research may lead to quicker inversion procedures, but with the present techniques, block coding with adaptive mean decoding appears to be decidely less practical than other methods of error control.

The decoder for unadaptive block decoding appears to be generally feasible if certain simplifying techniques are used. The most impor-

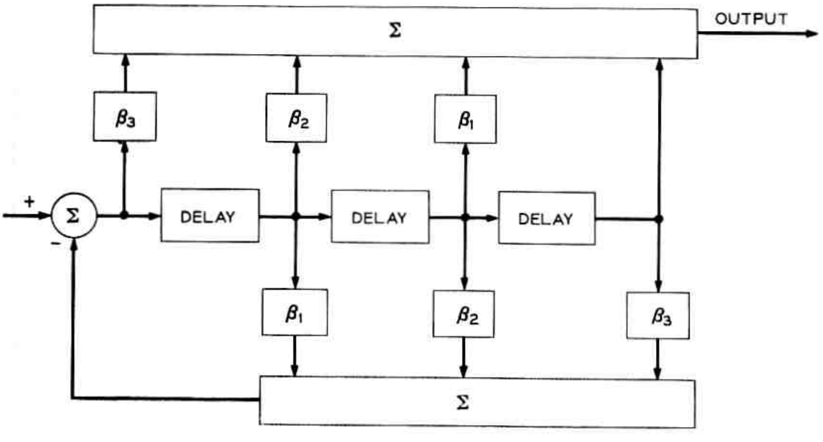


Fig. 3 — A shift register (real-number arithmetic) whose Z-transform transfer function is all-pass. (After V. A. Kisel', with modifications and a correction.)

tant of these is the use of an A matrix which is a permutation matrix* times diag$(A_0, A_0, \ldots \ldots, A_0)$, where A_0 is itself a matrix. A_0 must be large enough to give adaquate smear, whereas A must be large enough to make error burst lengths considerably shorter than the length of a code word. The hardware simplification achieved is that the inverse of the small A_0 can be repeatedly applied in time by the same hardware so as to invert the larger A. The practicality of block coding appears to be slightly overshadowed by stationary (shift register) coding, which offers somewhat simpler circuits and freedom from the problem of block synchronization.

Stationary (shift register) coding appears to be the most practical form of linear-real coding. In effect, such coding is a smear-desmear type of signal processing whenever the encoding and decoding filters are inverses of each other and of the all-pass type. The fundamental reason for the practicality of shift register all-pass filters is that accurately tuned shift registers can be relatively inexpensively synthesized, even when the dispersion times are several seconds. This is partly so because the "absolute" tuning of a shift register is determined by the clock pulses and not the precision of the components used in making the register, and partly because the "relative" tuning in a shift register is controlled by gains which in practice can be resistor values. As will be seen, analog shift registers can be implemented digitally, in which case complexity grows only as the logarithm of accuracy. In RLC filter synthesis, in contrast, cost grows rapidly with accuracy.

Figure 4 is a block diagram for coding of the basic stationary (shift register) type. The decoder, because it must handle the analog signals from the channel instead of the digital input signals, is selected to have the impulse response simplest to implement, namely an all-pass causal $1/a(z)$ obtained by a shift register made from a tapped delay line with a relatively moderate number of taps. The encoder is consequently left with approximating the noncausal $a(z)$, which it does with a delay by means of a tapped delay line.

The decoding shift register of Fig. 4 can be implemented by the arrangement of Fig. 5, which is a particular synthesis of the all-pass shift register shown in Fig. 3. In Fig. 5 all the digital-to-analog conversion is done by resistor summing networks. This is relatively inexpensive, although it does require that the flip-flop registers be designed for relatively precise voltage levels on the "on" and "off" states.

* A permutation matrix is a matrix with a single one in each column and each row; it is always nonsingular.

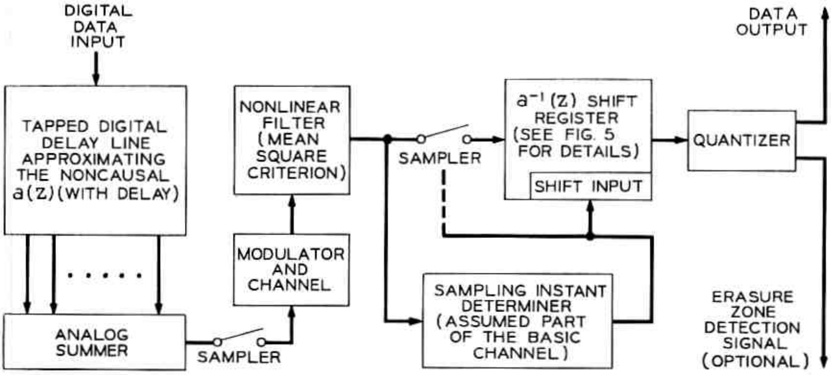

DIGITAL DATA INPUT
TAPPED DIGITAL DELAY LINE APPROXIMATING THE NONCAUSAL a(z) (WITH DELAY)
.....
ANALOG SUMMER
SAMPLER
MODULATOR AND CHANNEL
NONLINEAR FILTER (MEAN SQUARE CRITERION)
SAMPLER
a⁻¹(z) SHIFT REGISTER (SEE FIG. 5 FOR DETAILS)
SHIFT INPUT
SAMPLING INSTANT DETERMINER (ASSUMED PART OF THE BASIC CHANNEL)
QUANTIZER
DATA OUTPUT
ERASURE ZONE DETECTION SIGNAL (OPTIONAL)

Fig. 4 — One possible general arrangement for unadaptive stationary (shift-register) linear-real encoders and decoders. For multilevel signals, a Gray encoder can be used before the analog summer, and the quantizer would incorporate a Gray decoder.

Notice that in Fig. 5 there is only one analog-to-digital converter, because the analog feedback signal is added to the input signal before the conversion which is necessary in order to place the signals in the digital delay line.

The cost of the encoding and decoding shift registers will be roughly proportional to the amount of smear that they introduce. The amount of smear necessary for given performance depends upon the noise power. It follows that a considerable economic saving can be obtained at given performance if circuits, inexpensive compared to the decoder, can be found to reduce the noise during bursts.

A new circuit with this purpose for PAM systems is as shown in Fig. 6. The operation of the circuit requires that the interval between signal pulses be longer than the Nyquist interval for the bandwidth of the pulse shape. A way to find part of the noise component is to sample at the sampling instants, reconstruct the waveform which would be transmitted if these sample values were the data-signal values, and then subtract this signal from the actual received signal. (For proof of this statement, see appendix C.) An estimate of the instantaneous noise power can be made directly from those noise components which can be found. These components, for example, can be used to deduce the presence or absence of a noise burst. The circuit in Fig. 6 can obtain some noise components,* provided that the taps

* Specifically, Fig. 6 obtains the sample values of $\Delta(t)$ of Appendix C at $t = nT/2$, n integer. Notice that by construction, $\Delta(nT/2) = 0$ for n even. By the sampling theorem, just the samples of $\Delta(t)$ will be sufficient to reconstruct $\Delta(t)$ provided that $C(\omega)$ is zero for $|\omega| \geq 2\,\pi/T$.

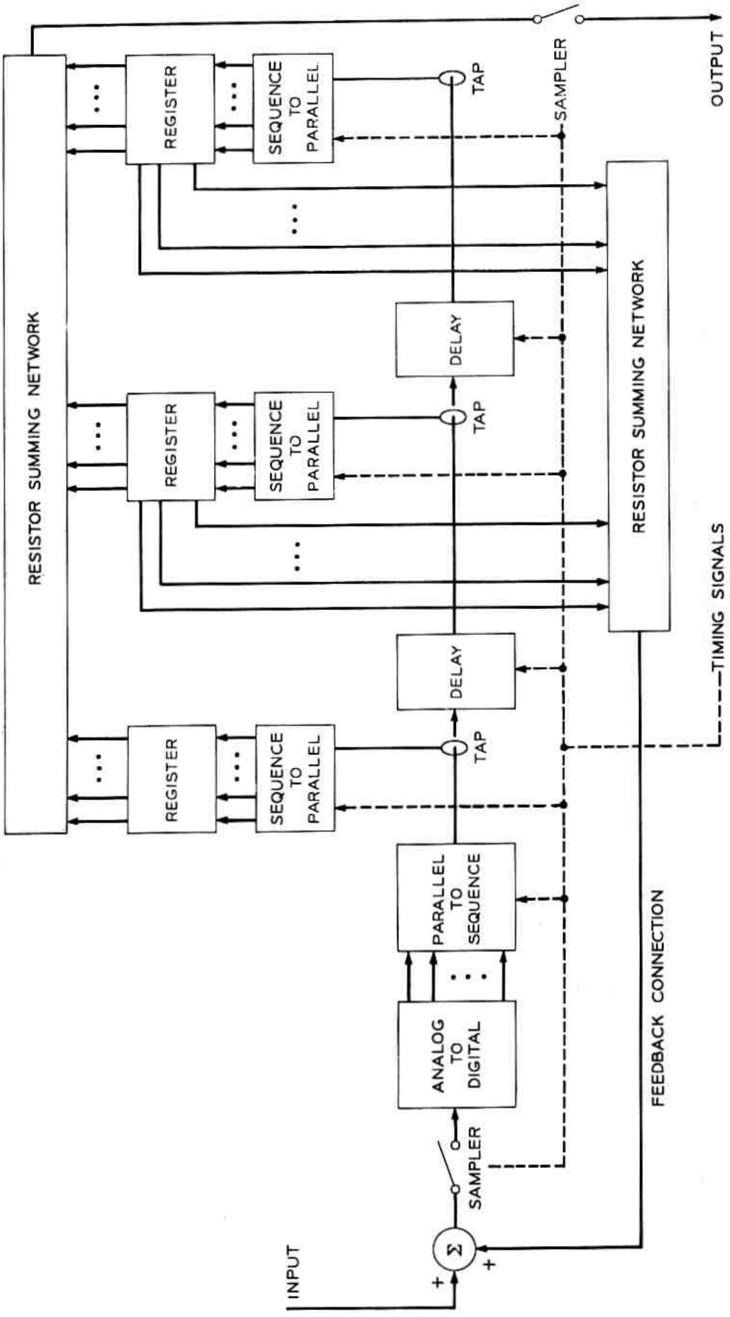


Fig. 5 — A possible arrangement for implementing the decoding shift register.

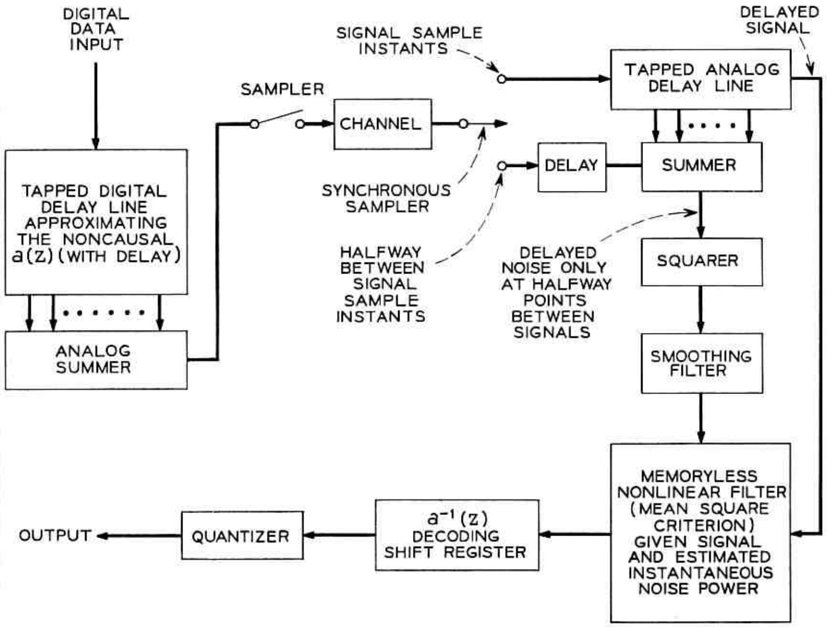

DIGITAL DATA INPUT
SIGNAL SAMPLE INSTANTS
DELAYED SIGNAL
TAPPED ANALOG DELAY LINE
SAMPLER
CHANNEL
DELAY
SUMMER
TAPPED DIGITAL DELAY LINE APPROXIMATING THE NONCAUSAL a(z) (WITH DELAY)
SYNCHRONOUS SAMPLER
HALFWAY BETWEEN SIGNAL SAMPLE INSTANTS
DELAYED NOISE ONLY AT HALFWAY POINTS BETWEEN SIGNALS
SQUARER
ANALOG SUMMER
SMOOTHING FILTER
MEMORYLESS NONLINEAR FILTER (MEAN SQUARE CRITERION) GIVEN SIGNAL AND ESTIMATED INSTANTANEOUS NOISE POWER
OUTPUT
QUANTIZER
$a^{-1}(z)$ DECODING SHIFT REGISTER

Fig. 6 — A stationary (shift register) coder with an adaptive decoder for PAM channels with white burst noise and pulse rates less than the Nyquist rate.

on the delay line represent the PAM pulse value at $t = nT/2$, n odd.

The output noise estimate (specifically $\Delta(nT/2$, n odd, in the language of Appendix C and the previous footnote) is then squared to produce the sample variance of the noise; then the sample variance function is put through a smoothing filter, as shown in Fig. 6. The optimization of this filter is complicated by the absence of an appropriate error criterion, but Wiener filtering principles could be used to optimize a mean square criterion. The problem formulation would specify that the sample variance is the true ensemble variance contaminated by small sample-size noise, and that the cross-correlation between the halfway sample process and the sample process could be found from the autocorrelation function of the channel noise.

Finally, a two-input nonlinear memoryless filter is used, also shown in Fig. 6. It is reasonable to optimize this filter using a mean square criterion because in the limit of infinite smearing only the power of the noise will be significant because of the smearing and Gaussianizing effects of the decoding shift register. Some improvement may be possible by using other criteria, but the details appear to be very difficult and are unsolved.

The general scheme of Fig. 6 appears to be the most economical form of linear-real coding when the channel is used for PAM at less than the Nyquist rate. Telephone lines are used at less than the Nyquist rate because they are used with signals with nonsharp-cutoff frequency characteristics. Radio links can obtain information on nontuned burst noise, such as static, by listening on adjacent frequencies, and could therefore provide the smoothed estimate of instantaneous noise power, needed as an input to the two-input memoryless filter, by other means. Instantaneous carrier-to-noise ratios could be used for carrier systems, for example.

It is also possible to use a different principle of instantaneous noise power estimation which does not require a PAM channel used below the Nyquist rate. The other principle uses the quantized structure of the data stream. It is implemented by a decoder with a "pilot" decoder which decodes, followed by an operator which squares the difference between the signal and the nearest quantization level, which is then smoothed and put into a two-input memoryless filter like that of Fig. 6, following which is the regular decoding shift register and quantizer. This scheme is probably less practical than Figs. 4 and 6, but it does give conceptual insights into some of the signal properties which can be used in decoding, especially for burst channels.

VI. COMMENTS AND SIMULATION RESULTS

Any sample of the decoded noise is a weighted sum of the random channel noises at many other sample instants. When the number of terms in this sum approaches infinity and the relative size of the largest term in the sum approaches zero, the central limit theorem applies. It will probably be true that practical designs will not have the conditions of the central limit theorem fulfilled to the extent that very small digital error probabilities can be computed by using integrals of the tails of the gaussian distribution.

Nevertheless, the fact that the decoded noise at any instant is a sum of the random channel noises at many instants will tend to make the decoded noise have some of the characteristics of a gaussian distribution. One characteristic that the decoded noise will have is the small probability that the decoded noise is larger than three or four standard deviations. This effect of the decoding filter (or matrix) will be called the gaussianizing property.

The use of nonlinear filters in conjunction with linear-real coders is extremely effective, since such filters can considerably reduce both the noise power and the probability that the noise has a large peak. By reducing the probability that the noise has a large peak, the desirable gaussian distribution of the decoded noise occurs with smaller matrices, smaller shift-registers, or simpler all-pass filters. In the limit when the decoded noise is actually gaussian, the noise power is the only significant statistic; the higher-order moments of the noise become insignificant due to the gaussian-distributing property of the decoder. It is therefore quite appropriate to design the nonlinear filter using a mean square error criterion, as is done in Section VIII of Reference 9.

Linear-real coding has features which could greatly improve error detection in channels with burst noise. When erasure zones are used to detect errors, the gaussian-distributing property of the decoder greatly increases the ratio of the probability in the erasure zone to the probability beyond the erasure zone. In addition, the noise spreading gives more opportunities for a signal to land in an erasure zone in the presence of impulses or bursts, because of randomness of the decoded noise, and, with suitable designs, because of deterministic reasons.

If the communications channel is, in order, digital processor to analog transmitter to analog receiver to digital processor, then linear-real block coding permits the energy per transmitted data digit to be altered by reprogramming the digital processors, instead of physically retuning bandwidths of analog equipment. Although this option does not in itself affect error control, it perhaps could greatly simplify the implementation of adaptive communications systems in which the signal energy per digit is adjusted to be appropriate for the transmission conditions, message importance, or message load.

A digital computer simulation was run of an additive-noise channel with a linear-real block-code encoder at the input, and several types of decoders at the output. Table III shows the results of the simulation. The listed results are averages. The A matrix is the Hadamard matrix which is generated recursively according to the procedure described by Golomb and his colleagues (p. 55, first paragraph in proof of Theorem 4.5).[15] The N matrix had zeros in all off-diagonal components, and independent random variables on the diagonals, which were 0.3 with probability 0.7 and 8.3 with probability 0.3. In accordance with Theorem 4 in Appendix D, these can be worst-case values which then give the worst-case decoded mean square error.

Once the N matrix was generated, the channel noises were gen-

TABLE III — SIMULATED PERFORMANCE OF LINEAR-REAL CODERS

	MS ERROR IN DECODED COMPONENTS		COMMENTS
	When receiver uses perfect N matrix	When receiver uses $N = \text{diag}\,(n_1', \cdots, n_n')$ where $n_i' = \max\,(0.3, f_i^2 - 1)$	
Mean estimator	0.516	0.711	The lower bound of equation (13) is somewhat loose; it gives 0.297.
Unattenuated estimator	2.821	2.821	Equation (12) has correctly predicted that the error would be the same as that of the unadaptive estimator because A is square.
Unadaptive estimator	2.821		Equation (7) averaged over the possible N matrices gives m.s. error of 2.70. The randomness of the N matrix accounts for difference.
Clip estimator parameters (1.2, 0.9, 4.0)	1.805		
Clip estimator parameters (1.0, 0.75, 3.0)	1.152		
Clip estimator parameters (0.8, 0.6, 2.0)	0.771		
Clip estimator parameters (0.6, 0.6, 1.5)	0.677		
Clip estimator parameters (0.5, 0.5, 1.3)	0.645		
Clip estimator parameters (0.4, 0.5, 1.0)	0.649		

Channel: Additive noise channel sending +1 and −1 binary numbers and block encoding with an A which is $k^{-\frac{1}{2}}$ times the first k columns of an n by n Hadamard matrix.

$n = 16$.

$k = 16$.

Number of words in simulation: 10. Noise type: Zero-mean white Gaussian noise has variance 0.3 with probability 0.7 and variance 8.3 with probability 0.3.

erated randomly from a Gaussian distribution having the given N for a covariance matrix. The clip estimator used a decoder which first put each received component through a memoryless nonlinearity, and then decoded the resulting components with the unadaptive estimator. The parameters (x, y, z) indicate that the nonlinearity is a continuous odd function having slope 1 for inputs of magnitude less than x, and slope y for inputs of magnitude between x and z, and slope 0 for inputs of magnitude exceeding z. These parameters can be chosen to approximate the least mean square memoryless nonlinear filter referred to earlier, or they can be found by a trial-and-error procedure with either analysis or simulations to evaluate the resulting error.

The following two conclusions can be drawn from the simulation, but it would not be appropriate to generalize them to cases of non-square A matrices:

(*i*) For intermittent additive impulse noise of the type simulated, the simple clip estimator scheme, for appropriate parameters, is almost as good as the mean estimator, even though it is unadaptive and therefore requires only a simple receiver.

(*ii*) The use of rather crude algorithms for generating an estimate of N appeared to be inferior to clip estimator decoding with appropriate parameters.

ACKNOWLEDGMENT

I am grateful to Dr. Sidney Berkowitz for discussions, and the substitution appearing in equation (31). Many astute and helpful suggestions from the referees have been incorporated into the paper.

APPENDIX A

Justification of Table 2 Equations

A.1 *Unadaptive Estimator*

In the case of the unadaptive estimator $TA = I_{(k\times k)}$, so equation (3) reduces to equation (4) shown in Table II. Now in general, when M is the covariance matrix of the decoded noise, the mean square error will be the average of the on-diagonal terms of M, or in other words, $(1/k)\, tr\ M$. In this way (5) follows from (4). Equation (6) follows from (4) because $A^tA = (n/k)I_{(k\times k)}$ in this case.

A Hadamard matrix is a square matrix with $+1$ or -1 elements and orthogonal columns. (Golomb and his associates fully describe Hadamard matrices and their application to binary block codes.[15])

In deriving equation (7), a straightforward evaluation of (6) under the assumption of diagonal N gives the result that

$$M_{ij} = \left(\frac{k}{n}\right)^2 \sum_{l=1}^{n} a_{li} a_{lj} n_l \; .$$

assuming:

T is the unadaptive estimator

$$N = \text{diag}\,(n_1\,,\, n_2\,,\, \cdots\,,\, n_n).$$

The on-diagonal terms of the above can be evaluated by using the Hadamard assumption, which causes $(a_{li})^2$ to equal $1/k$ for all l and i. This gives

$$M_{ii} = \left(\frac{k}{n}\right)\left[\frac{1}{n} \sum_{l=1}^{n} n_l\right]$$

assuming:

T is the unadaptive estimator

$N = \text{diag}\,(n_1\,,\, n_2\,,\, \cdots\,,\, n_n)$

A is $1/(k)^{\frac{1}{2}}$ times the first k columns of any Hadamard matrix.

Notice that the term in brackets is AM, the arithmetic mean of the set $(n_1, n_2, \ldots, n_n)$.

A.2 *Unattenuated Estimator*

Equation (8) comes from (3) by direct substitution for the T matrix.

A.3 *Mean Estimator*

In the case of the mean estimator,

$$\begin{aligned} I_{(k\times k)} - TA &= I_{(k\times k)} - (Q^{-1} + A^t N^{-1} A)^{-1} A^t N^{-1} A \\ &= I_{(k\times k)} - (Q^{-1} + A^t N^{-1} A)^{-1} (A^t N^{-1} A + Q^{-1} - Q^{-1}) \\ &= (Q^{-1} + A^t N^{-1} A)^{-1} Q^{-1}. \end{aligned} \tag{31}$$

Substituting $(Q^{-1} + A^t N^{-1} A)^{-1} Q^{-1}$ for $(I_{(k\times k)} - TA)$ in equation (3) readily shows that

$$M = (Q^{-1} + A^t N^{-1} A)^{-1}$$

assuming T is the mean estimator.

A.4 *Joint Mean and Unattenuated Estimator*

The unattenuated estimator is the special case of the mean estimator when $Q^{-1} \to 0$. It is convenient to handle the two cases together by using the variable $\Omega = 1$ when the mean estimator is used, and $\Omega = 0$ when the unattenuated estimator is used.

The next two equations use an approach from Berkowitz.[10] Equation (9) or (11) can be simplified by using the fact that, for any nonsingular Z,

$$\operatorname{tr} Z^{-1} = \sum_i \frac{1}{\lambda_i(Z)}$$

where $\lambda_i(Z)$ denotes the i^{th} unordered eigenvalue of Z. The result is equation (12). When the signal is white, the relation $\lambda_i(\tau I + Z) = \tau + \lambda_i(Z)$ can be used, giving the equality in equation (13). When $\Omega = 1$ the positive semidefiniteness of $A^tN^{-1}A$ causes its eigenvalues to be real and nonnegative; when $\Omega = 0$ the positive definiteness of $A^tN^{-1}A$ will now need to be assumed. Because $1/(\Omega s^{-1} + \lambda)$ is a convex upward function of λ in the region of possible λ, the inequality part of (13) follows by convexity. This inequality will prove useful later when—under additional assumptions—the term in brackets will be found in closed form.

For square orthonormal A, it follows that $A^{-1} = A^t$, so

$$\lambda_i(A^tN^{-1}A) = \lambda_i(A^{-1}N^{-1}A) = \lambda_i(N^{-1}) = \frac{1}{\lambda_i(N)}.$$

Equation (14) results when the above is substituted into (13). Notice that when Ω is zero and N is diagonal, this will reduce to AM. On the other hand, when Ω is one, this will be less than AM.

When A is rectangular, the next analysis leads to a closed form solution for the average of the eigenvalues of $A^tN^{-1}A$, under the assumptions of equation (15), and it also leads to upper bounds upon the m.s. error. The exact values of the components of A may enter into the formulas for some statistics of the error. However, in the first and second moment statistics to be investigated under the particular assumptions made, it turns out that the only important property of the A matrix is the inner product between the i^{th} and j^{th} columns. This will always be $(n/k)\ \delta_{ij}$, independent of the particular Hadamard matrix upon which A is based. However, since higher-order moments are significant, especially in quantized channels, it is likely that some Hadamard matrices might be more useful for practical purposes than others.

Under the assumptions of equation (15), straightforward calculations will show the following. HM is the harmonic mean of the diagonal components of N; equation (15) includes its formula.

$$(i) \qquad A^tN^{-1}A = \frac{n}{k\,HM}\,I + Y_k$$

where, for large k

$$\left.\begin{aligned} E[(Y_k)_{ij}] &= 0 \\ E[(Y_k)_{ij}^2] &= \frac{\sigma^2 n}{k^2} \end{aligned}\right\} \quad \text{all } i, j \tag{32}$$

$$(ii) \qquad E\left[\frac{1}{k}\,\mathrm{tr}\;A^tN^{-1}A\right] = \frac{n}{k\,HM}. \tag{33}$$

$$(iii) \qquad \mathrm{Var}\left[\frac{1}{k}\,\mathrm{tr}\;A^tN^{-1}A\right] = \frac{n}{k^2}\,\sigma^2. \tag{34}$$

The above equations are especially useful because they show that

$$\textit{the average of the eigenvalues of } A^tN^{-1}A = \frac{n}{k\,HM}.$$

This can be substituted into equation (13) to prove equation (16). Equation (16) becomes an equality when all of the eigenvalues of $A^tN^{-1}A$ are equal; otherwise the mean square error is greater.

Because the m.s. error evaluated according to equation (12) requires the computation of eigenvalues of typically a rather large matrix, or the trace formula of (9) or (11) yields little insight, and because the bound of equation (16) is a simple closed-form equation, the question arises of whether the bound given by (16) is really close enough to be used for design and analysis purposes as an equality. The analysis which follows will derive an upper bound for the m.s. error, which could be used to develop some sufficient conditions for near equality of equation (16)

Let equation (32) be used to define Y_k, let $\lambda'(Y_k)$ denote

$$\max_i |\lambda_i(Y_k)|\,,$$

and let τ be any number such that

$$\tau \leqq \frac{\sum_{i=1}^{k} |\lambda_i(Y_k)|^2}{[\lambda'(Y_k)]^2}. \tag{35}$$

Notice that τ can always be as large as 1 and never exceeds k. The second of the following inequalities is Schur's inequality, which is valid for any square Y_k.[16] The first comes from (35).

$$\tau[\lambda'(Y_k)]^2 \leqq \sum_{i=1}^{k} |\lambda_i(Y_k)|^2 \leqq \sum_{i=1}^{k} \sum_{j=1}^{k} |(Y_k)_{ij}|^2. \tag{36}$$

Assuming that k is large enough for the weak law of large numbers to hold permits (32) to be used to evaluate the above double sum, so that with a few manipulations (36) reduces to

$$\lambda'(Y_k) \leqq \sqrt{\frac{\sigma^2 n}{\tau}}. \tag{37}$$

By using equations (13), (32), (33), and (37), and a relatively obvious property of convex functions,* equation (17) is established.

APPENDIX B

Relating Teoplitz Matrix Operations with Z-Transforms

Theorem 2: If

$$A_{ij} = a_{i-j}$$

and if

$$a(z) = \sum_{q=-\infty}^{\infty} a_q z^{-q}$$

converges on $|z| = 1$ and has no poles or zeros for a finite distance from $|z| = 1$, then A^{-1} exists and

$$a^{-1}(z) = \frac{1}{a(z)}.$$

Proof: Let

$$b(z) = \frac{1}{a(z)} \quad \text{for} \quad |z| = 1.$$

* The property is that if $f(x)$ is convex downward, and

$$\sum_{i=1}^{k} x_i = 0, \qquad \max_i |x_i| \leqq R,$$

then

$$\sum_{i=1}^{k} \frac{1}{k} f(u + x_i) \leqq \tfrac{1}{2} f(u - R) + \tfrac{1}{2} f(u + R).$$

The assumptions on $a(z)$ cause a_q and b_q to have geometrical decay, and therefore the following converge absolutely:

$$(BA)_{ij} = \sum_{q=-\infty}^{\infty} B_{iq}A_{qj} = \sum_{q=-\infty}^{\infty} b_{i-q}a_{q-j} .$$

Also reducing to the above is $(AB)_{ij}$. Letting $q' = q - j$ gives

$$(BA)_{ij} = (AB)_{ij} = \sum_{q'=-\infty}^{\infty} b_{(i-j)-q'}a_{q'} = b_q * a_q \,|_{i-j}$$

where the * denotes the convolution sum in the line above. Because $b(z)a(z) = 1$, it follows that

$$bq * a_q \,|_{i-j} = \delta_{i,j} .$$

So $BA = AB = I_\infty$, thus proving that B is the inverse of A, which completes the proof.

The following have proofs similar to that of the theorem.

Lemma 3: If A and B are Toeplitz, then C = AB is Toeplitz with

$$c(z) = a(z)b(z).$$

Lemma 4: The half-power of a Toeplitz matrix N can be defined by

$$n^{\frac{1}{2}}(z) = \sqrt{n(z)}.$$

The following has a straightforward proof:

Lemma 5: If A is Toeplitz, then A^t is Toeplitz and $a^t(z) = a(1/z)$.

The following relates linear-real coding for Toeplitz matrices with Wiener filtering.

Theorem 3: When A, Q, and N are infinite Toeplitz, then the least mean square estimator

$$T = (Q^{-1} + A^tN^{-1}A)^{-1}A^tN^{-1}$$

is the infinite Toeplitz, and the noncausal Wiener filter, given by

$$t(z) = \frac{q(z)a\left(\frac{1}{z}\right)}{n(z) + a\left(\frac{1}{z}\right)q(z)a(z)}.$$

Proof: By Theorem 2 and Lemmas 3 and 5,

$$t(z) = \frac{1}{\dfrac{1}{q(z)} + \dfrac{a\left(\frac{1}{z}\right)a(z)}{n(z)}} \cdot a\left(\frac{1}{z}\right) \cdot \frac{1}{n(z)}.$$

This equals the stated result, which completes the proof.

Corollary 3: When $A = I_\infty$

$$T = (Q^{-1} + N^{-1})^{-1} N^{-1}$$

and

$$t(z) = \frac{q(z)}{q(z) + n(z)}$$

is the noncausal Wiener filter.

The following proof of equation (29) and statement (30) follows the ideas of J. E. Mazo. For square A,

$$\text{tr}\,[A^{-1}(A^{-1})^t] = \text{tr}\,[(A^{-1})^t A^{-1}],$$

since in general *tr* HC = *tr* CH for square H and C. Now let $B = AA^t$. Notice that $(A^{-1})^t A^{-1}$ is B^{-1}. Equation (29) is then:

$$\frac{1}{k^2} \text{tr}\, B^{-1}\, \text{tr}\, B \geqq 1.$$

But

$$\text{tr}\, B = \sum_{i=1}^{k} \lambda_i(B)$$

$$\text{tr}\, B^{-1} = \sum_{i=1}^{k} \frac{1}{\lambda_i(B)}$$

so

$$\frac{1}{k^2} \text{tr}\, B^{-1}\, \text{tr}\, B = \frac{\dfrac{1}{k} \displaystyle\sum_{i=1}^{k} \lambda_i(B)}{\left[\dfrac{1}{\dfrac{1}{k} \displaystyle\sum_{i=1}^{k} \dfrac{1}{\lambda_i(B)}}\right]}.$$

The numerator and denominator are respectively the arithmetic and harmonic means of the eigenvalues of the B matrix. Hardy, Littlewood, and Polya (p. 26, special case of 2.9.1) show that this ratio always exceeds one, except when the eigenvalues are all the same, in which case it is one.[17] This proves (29).

At equality B has equal eigenvalues, and since it is symmetric the eigenvectors span the space and B is proportional to an orthogonal matrix:

$$\begin{aligned} B = \lambda U &= P^{-1}(\lambda I)P \\ &= P^{t}(\lambda I)P \\ &\qquad \text{because } B \text{ is symmetric} \\ &= \lambda I. \end{aligned}$$

Therefore $AA^t = \lambda I$ and $A^{-1} = \lambda A^t$, so A is proportional to an orthogonal matrix at equality, thereby establishing (30) and completing the proof of (29) and (30), thereby completing the proof of Theorem 1.

APPENDIX C

Finding Noise Component

In the text we discuss the circuit shown in Fig. 6 and state that a way to find part of the noise component is to sample at the sampling instants, reconstruct the waveform which would be transmitted if these sample values were data-signal values, and then subtract this signal from the actual received signal.

The proof of this statement requires the use of the valid converse of the sampling theorem, which states that an arbitrary function with frequency components out to $|\omega| = \pi/T_1$ cannot be reconstructed from samples every T seconds if $T > T_1$. If it is assumed that

(*i*) $h(O) = 1$
(*ii*) $h(nT) = 0$
(*iii*) $H(\omega)$ is nonzero for $|\omega| < \pi/T_1$
(*iv*) $T_1 < T$
(*v*) The additive noise $c(t)$ has components at all frequencies for which $H(\omega)$ has components,

then it follows that

$$\text{actual sample at } t = c(t) + \sum_n r_n h(t - nT)$$

predicted sample at t based upon

samples at nT, $n = 0, \pm 1, \pm 2, \cdots = \sum_n [r_n + c(nT)]h(t - nT)$

$\Delta(t) =$ difference of the above $= c(t) - \sum_n c(nT)h(t - nT)$.

By using a well-known result in sampling theory[12] the Fourier transform of $\Delta(t)$ can be written as either of the following.

$$\mathbf{\Delta}(\omega) = \mathfrak{F}[\Delta(t)] = C(\omega) - H(\omega) \sum_n c(nT)e^{-i\omega nT}$$

$$= C(\omega) - H(\omega) \sum_n C\left(\omega - \frac{n2\pi}{T}\right).$$

By the converse to the sampling theorem, no $H(\omega)$ will make $\mathbf{\Delta}(\omega)$ zero for all ω. Consequently, $\mathbf{\Delta}(\omega)$ contains some components of the additive noise. If $T_1 \geqq T/2$, then the direct sampling theorem shows that samples every $T/2$ are sufficient to reconstruct $\Delta(t)$.

APPENDIX D

The purpose of this appendix is to state and prove the following theorem.

Theorem 4: Assuming

(*i*) *Channel I has additive noise c independent of the signal b*

(*ii*) *Channel II has additive noise g independent of the signal b*

(*iii*) *c and g are zero mean, and each is even about its mean*

(*iv*) $F(\alpha) = p(|\,c\,| \leqq \alpha)$, *($\alpha$ is defined to be nonnegative)*

(*v*) $K(\alpha) = p(|\,g\,| \leqq \alpha)$

(*vi*) *In both channels signal plus noise are passed through the memoryless nonlinearity nl( ) at the receiver*

(*vii*) *$nl(x)$ is odd*

(*viii*) *$nl(x)$ has a slope bounded between* 0 *and* 1 *for all x, and this slope is monotonically decreasing in $|\,x\,|$*

(*ix*) *The mean square errors of channels I and II are MSE_I and MSE_{II}, respectively.*

(*x*) *Channel I is noisier than channel II in the sense that $F(\alpha) \leqq K(\alpha)$ for all α, which means that for every bit of probability density c has at $\pm\beta$, g has an equal amount at a distance which is at least $\pm\beta$,*

then

$$MSE_I \geqq MSE_{II}.$$

(Thus, worst-case noise gives worst-case results with these nonlinearities.)

The next definition and lemma are used in the proof of Theorem 4. Let $MS(\alpha)$ denote the special case of MSE_I of Theorem 4 when

$$p_2(c) = \tfrac{1}{2}\delta(c+\alpha) + \tfrac{1}{2}\delta(c-\alpha) \tag{38}$$

where α is a positive constant, and $\delta(\ )$ denotes the Dirac impulse function.

Lemma 6: Under the conditions of Theorem 4, $(\partial MS(\alpha)/\partial\alpha) \geq 0$.

Proof:

$$MS(\alpha) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[nl(b+c)-b]^2 p_1(b)p_2(c)\,dc\,db. \tag{39}$$

Substituting equation (38) for $p_2(c)$, integrating with respect to c, and then taking partial derivations with respect to α gives

$$\frac{\partial MS(\alpha)}{\partial\alpha} = \int_{-\infty}^{\infty}\left\{\underbrace{[nl(b+\alpha)-b]}_{A}\,\underbrace{\left.\frac{dnl(x)}{dx}\right|_{b+\alpha}}_{B} - \underbrace{[nl(b-\alpha)-b]}_{C}\,\underbrace{\left.\frac{dnl(x)}{dx}\right|_{b-\alpha}}_{D}\right\}p_1(b)\,db. \tag{40}$$

Now

$$[\text{assumptions } 7, 8] \Rightarrow [C \leqq 0 \text{ for } b \geqq 0,\ A \leqq 0 \text{ for } b \leqq 0] \tag{41}$$

$$[\text{assumption } 8] \Rightarrow [D \geqq 0,\ B \geqq 0] \tag{42}$$

$$[\text{assumption } 8] \Rightarrow [B \leqq D \text{ when } b \geqq 0,\ B \geqq D \text{ when } b \leqq 0]. \tag{43}$$

Therefore

$$-CD \geqq -CB \quad \text{when} \quad b \geqq 0 \tag{44}$$

$$AB \geqq AD \quad \text{when} \quad b \leqq 0. \tag{45}$$

Consequently

$$\frac{\partial MS(\alpha)}{\partial\alpha} \geqq \int_{-\infty}^{0-}[(A-C)D]p_1(b)\,db + \int_{0}^{\infty}[(A-C)B]p_1(b)\,db. \tag{46}$$

Now

$$A - C = \int_{b-\alpha}^{b+\alpha}\frac{dnl(x)}{dx}\,dx \tag{47}$$

and by assumption 8 the integrand is nonnegative, so both sides of (47) are nonnegative. This fact and (42), and the nonnegativeness of $p_1(b)$, make the right side of (46) nonnegative, which proves the lemma.

Proof of Theorem 4: The assumed evenness of the noises, and the linearity of the expectation operator, permit the $MS(\alpha)$ function to be used to evaluate the mean square error, as follows

$$MSE_{\text{I}} - MSE_{\text{II}} = \int_0^\infty MS(\alpha)\, dF(\alpha) - \int_0^\infty MS(\alpha)\, dK(\alpha). \tag{48}$$

The above right side can be combined into one integral, such that integrating by parts gives zero for the end conditions plus the resulting integral.

$$MSE_{\text{I}} - MSE_{\text{II}} = \int_0^\infty [K(\alpha) - F(\alpha)]\left\{\frac{dMS(\alpha)}{d\alpha}\right\} d\alpha. \tag{49}$$

Assumption 10 makes the bracketed term nonnegative, whereas Lemma 6 makes the braced term nonnegative, so the right side of (49) when integrated is nonnegative, which proves the theorem.

REFERENCES

1. Knox-Seith, J., unpublished work.
2. Anderson, R. R. and Koll, V. G., unpublished work.
3. Stamboulis, A. P., unpublished work.
4. Gibson, E. D., "A Highly Versatile Corrector of Distortion and Impulse Noise," Proc. Nat. Elec. Conf., 23, 1961.
5. Lerner, R. M. "Design of Signals," Chapter 11 of *Lectures on Communication Theory,* ed. E. J. Baghdady, New York: McGraw-Hill, 1961.
6. Holland-Moritz, E. K., Dute, J. C., and Strember, F. G., "Feasibility of the Swept-Frequency Modulation Technique," Report 4435-16-F, Radar Laboratory, Inst. Sci. and Technology, University of Michigan, August 1962.
7. Wainwright, R., "Overcoming Impulse Noise Interference in Narrowband Data Communication Systems by a Sophisticated Filter Technique," Rixon Eng. Bull. No. 70 (July 1960); also Rome-Utica IRE Conf., October 1960.
8. Helstrom, C. W., "Topics in the Transmission of Continuous Information," Westinghouse Res. Laboratories Report 64-8C3-522-R1, August 27, 1964.
9. Pierce, W. H., "Linear-Real Coding," IEEE Int. Conv. Record, part VII (1966), pp. 44–53.
10. Berkowitz, S., Ph.D. Thesis, Carnegie Inst. Technology, 1966.
11. Smith, D. H., "A Magnetic Shift Register Employing Controlled Domain Wall Motion," IEEE Trans. Magnetics, 1, (December 1965), pp. 281–284.
12. Ragazzini, J. R. and Franklin, G. F., *Sampled-Data Control Systems,* New York: McGraw-Hill, 1958.
13. Grenander, U., and Szego, G., *Toeplitz Forms and Their Applications,* Berkeley, Calif., University of California Press, 1958.
14. Kisel', V. A., "Phase Correcting Circuits Using Delay Lines," *Telecommunications and Radio Engineering* (Elektrosvyaz, Radio Tekhnika), December 1965.
15. Golomb, S. W., Baumert, L. D., Fasterling, M. F., Stiffler, J. J., and Viterbi, A. J., *Digital Communications,* Englewood Cliffs, N.J.: Prentice-Hall, 1964.
16. Schur, I., *Math. Ann.,* 66 (1909), pp. 488–510.
17. Hardy, G. H., Littlewood, J. E., and Polya, G. *Inequalities,* New York: Cambridge University Press, 1959.

Matrix Multiplication and Fast Fourier Transforms

By W. MORVEN GENTLEMAN

(Manuscript received January 29, 1968)

Factoring a matrix and multiplying successively by the factors can sometimes be used to speed up matrix multiplications. This is, in fact, the trick which creates the fantastic gains of the fast Fourier transform.

The same trick which creates the fantastic gains of the fast Fourier transform may be used with other matrices.

As an example, suppose the matrix

$$\begin{bmatrix} 1 & -10 & 4 & 3 & -14 & 12 \\ -5 & 2 & -20 & -7 & 6 & -28 \\ 2 & -20 & 1 & 6 & -28 & 3 \\ -20 & 1 & -10 & -28 & 3 & -14 \\ 4 & -5 & 2 & 12 & -7 & 6 \\ -10 & 4 & -5 & -14 & 12 & -7 \end{bmatrix}$$

is to be multiplied by a large number of different vectors, so that it is worthwhile to try to be as efficient as possible. At first glance, it would appear that (neglecting the possibility that multiplications by one might not actually be performed) multiplying this matrix with a single column vector would require $6^2 = 36$ multiplications and $6(6–1) = 30$ additions. The crafty person, however, might notice that this matrix may be written as the product of two matrices:

$$\begin{bmatrix} 1 & 2 & 4 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -1 & -2 & -4 \\ 2 & 4 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -4 & -1 & -2 \\ 4 & 1 & 2 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -2 & -4 & -1 \end{bmatrix} \begin{bmatrix} 1 & \cdot & \cdot & 3 & \cdot & \cdot \\ \cdot & -5 & \cdot & \cdot & -7 & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & 3 \\ 5 & \cdot & \cdot & 7 & \cdot & \cdot \\ \cdot & -1 & \cdot & \cdot & -3 & \cdot \\ \cdot & \cdot & 5 & \cdot & \cdot & 7 \end{bmatrix}$$

The zero elements in the decomposed form have been written as periods to emphasize that these elements need not really enter into the computation when either of these matrices multiply a vector. In view of this, multiplying sequentially by the two factors would require only $6(2) + 6(3) = 30$ multiplications and $6(1) + 6(2) = 18$ additions.

If we are really concerned about efficiency, more can be done by taking into account other special elements. For example, observing that 1 or -1 require only an addition or subtraction would save 3 multiplications in the original form, and 9 multiplications in the decomposed form. Other savings could be made if some of the elements of a column were negatives of other elements in the same column.

In the three years since the fast Fourier transform was first published,[1] there have been numerous accounts of what it is and why it works. The more mathematical of these tend to explain it in terms of the fact that the quotient group of a cyclic subgroup of order MN relative to its cyclic subgroup of order M is itself a cyclic group of order N. Those accounts written by computer people usually consider the binary representation of the time and frequency indices, and observe how each bit enters into the summed products. And accounts written by engineers invariably explain the algorithm in terms of merging the spectra of suitable decimations of the original series to form the spectrum of the original series itself.

These approaches are, of course, all quite valid, but they miss the essence of the fast Fourier transform which is, in fact, contained in the example above. If we wish to multiply a matrix M by a column vector x, it may be possible to find a factorization $M = AB$ such that forming first $y = Bx$ then $z = Ay$ requires less multiplications and additions than would forming $z = Mx$ directly. The factors A and B might themselves be able to be factored further profitably.

The fast Fourier transform is a special case of this, where the matrix of interest is the finite discrete Fourier transform matrix whose elements are $\exp 2\pi i(t\hat{t}/N)$ for $\hat{t}$ and t from 0 to $N - 1$. It is really quite irrelevant that the factors turn out to be (except for a permutation and phase shifts) block diagonal matrices where each block is of the same form as the original matrix—this fact is only used in showing that the factoring can be continued.*

Indeed, the example above has exactly the same structure as a

* In fact, for the fastest programs it is not even quite true. See Bergland.[2] The factors there are not equivalent to each other as the "twiddle factors" have been redistributed to increase the number of coefficients having simple forms.

6 = **3** × **2** point fast Fourier transform, except that the nonzero elements in the factors are different. And it achieves exactly the same savings that the fast Fourier transform does in this case. Even the comments about taking advantage of explicit plus or minus ones or negatives of other elements in the column reflect features currently in the better fast Fourier transform programs.

Having seen that the possibility that matrix factoring will speed things up is not unique to the finite Fourier transform, we might ask when we can expect to take advantage of it. It is immediately evident that it does not improve things all the time. We cannot, for example, reduce the number of operations required to multiply by a diagonal matrix. Can we then identify those matrices for which it is useful? Unfortunately not, except by exhibiting a factorization with the required property.

At this point it is useful to observe that, taking advantage only of zeros and ones, there always exist factorizations which do at least as well as the original matrix. This is trivially true if one of the factors is some permutation matrix, but more interestingly so if we consider factors generated by row (or column) elimination as used in the Gaussian elimination method of solving simultaneous linear equations. In matrix terms this process is based on the observation that

$$\begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1n} \\ m_{21} & m_{22} & \cdots & m_{2n} \end{bmatrix} = \begin{bmatrix} 1 & \cdot \\ r & 1 \end{bmatrix} \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1n} \\ m_{21} - rm_{11} & m_{22} - rm_{12} & \cdots & m_{2n} - rm_{1n} \end{bmatrix}$$

The parameter r is then chosen to make one of the elements in the second row vanish. Since this means that the right factor takes one less multiplication and one less addition than the original matrix did, and since the left factor clearly only requires one multiplication and one addition, the total number of operations for the two factors is exactly the same as for the original matrix.

In other words, row (or column) elimination preserves the number of operations required to form the product of the matrix with an arbitrary vector. This assertion assumes, of course, that in the elimination we do not destroy more special elements (such as zeros or ones) than we create. In fact, if we can create more of these special elements than we had before, we have won: we have achieved a factorization requiring less operations than did the original matrix.

Notice that in the above example we used a nonsquare matrix. In fact, nothing in the whole discussion suggested M be square, and considering nonsquare matrices is no more difficult than considering square ones. An immediate application of this is to the case where a set of only a few Fourier coefficients are required from a large number of very long sequences. Up until now, usually the best that could be done was to compute the complete fast Fourier transforms and discard the unneeded coefficients.

But it is apparent that by carefully factoring the matrix consisting of those rows of the finite Fourier transform matrix which are of interest, a more efficient algorithm can be produced, tailored to the problem. A reasonable factorization to start from might be fast Fourier factorization of the complete matrix. This is illustrated below for the case where three coefficients are wanted from an eight point transform. The four factor matrices are the reordering and the three passes of the Cooley factorization. Only those rows of each matrix which are marked by arrows need actually be computed. ($W = \exp [(2\pi i)/8]$, explicit negatives and ones are represented as such).

$$\begin{array}{c}\to\\ \to\\ \to\\ \\ \\ \\ \\ \\ \end{array}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & W & W^2 & W^3 & -1 & -W & -W^2 & -W^3 \\ 1 & W^2 & -1 & -W^2 & 1 & W^2 & -1 & -W^2 \\ 1 & W^3 & -W^2 & W & -1 & -W^3 & W^2 & -W \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & -W & W^2 & -W^3 & -1 & W & -W^2 & W^3 \\ 1 & -W^2 & -1 & W^2 & 1 & -W^2 & -1 & W^2 \\ 1 & -W^3 & -W^2 & -W & -1 & W^3 & W^2 & W \end{bmatrix}$$

$$= \begin{array}{c}\to\\ \to\\ \to\\ \\ \\ \\ \\ \\ \end{array}\begin{bmatrix} 1 & & & & 1 & & & \\ & 1 & & & & W & & \\ & & 1 & & & & W^2 & \\ & & & 1 & & & & W^3 \\ 1 & & & & -1 & & & \\ & 1 & & & & -W & & \\ & & 1 & & & & -W^2 & \\ & & & 1 & & & & -W^3 \end{bmatrix} \begin{array}{c}\to\\ \to\\ \to\\ \\ \to\\ \to\\ \to\\ \\ \end{array}\begin{bmatrix} 1 & & 1 & & & & & \\ & 1 & & W^2 & & & & \\ 1 & & -1 & & & & & \\ & 1 & & -W^2 & & & & \\ & & & & 1 & & 1 & \\ & & & & & 1 & & W^2 \\ & & & & 1 & & -1 & \\ & & & & & 1 & & -W^2 \end{bmatrix}$$

$$\begin{array}{c}\to\\ \to\\ \to\\ \to\\ \to\\ \to\\ \to\\ \to\end{array}\begin{bmatrix}1 & 1 & & & & & & \\ 1 & -1 & & & & & & \\ & & 1 & 1 & & & & \\ & & 1 & -1 & & & & \\ & & & & 1 & 1 & & \\ & & & & 1 & -1 & & \\ & & & & & & 1 & 1 \\ & & & & & & 1 & -1\end{bmatrix}\begin{array}{c}\to\\ \to\\ \to\\ \to\\ \to\\ \to\\ \to\\ \to\end{array}\begin{bmatrix}1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1\end{bmatrix}$$

We could also have regarded $\pm i$ as special elements.

Our suggestion then is that if one has a matrix which he wants to multiply efficiently into a great number of arbitrary vectors, it might be worthwhile to try to find a factorization of the matrix such that multiplying sequentially by the factors is cheaper than multiplying by the original matrix. Indeed, it is worthwhile to try to find an extremely good, perhaps even the best, such factorization.

Since we cannot identify *a priori* matrices for which this can be done, let alone give an algorithm for finding the best or even just a good factorization, the best we can recommend is to generate trial factorizations and compare them. A useful tool for this is row (or column) elimination: because of the invariance property mentioned earlier, such a factorization cannot lose much, and might gain. As an exercise to the reader, we suggest deriving the factorization of the matrix given at the beginning of this paper, or the eight point fast Fourier transform above. Notice that in the case of the fast Fourier transform it is useful to express the matrix in real arithmetic before reducing it, because then it is more obvious how to go further in the reduction, since in the computer it is usually the number of real operations that counts.

REFERENCES

1. Cooley, J. W., and Tukey, J. W., "An Algorithm for the Machine Calculation of Complex Fourier Series," Mathematics of Computation, *19,* No. 90 (April 1965), pp. 297–301.
2. Bergland, G. D., "A Fast Fourier Transform Algorithm Using Base Eight Iterations," Math. of Computation, *22,* No. 102 (April 1968), pp. 275–279.

On a Class of Configuration and Coincidence Problems

By Z. A. MELZAK*

(Manuscript received August 28, 1967)

Let A and B be sets in E^m where B is convex and symmetric about o. Let n points be taken in A and let B_i be the translate of B centered at the i^{th} one. Let Y be the subset of the Cartesian product A^n, corresponding to the configurations $(B_1, \cdots, B_n)$ such that no more than $p - 1$ sets B_i intersect, or corresponding to any similar configuration condition, expressible in purely Boolean terms. The problem of evaluating various integrals over Y generalizes a number of questions in queuing, telephone traffic, statistical mechanics of hard spheres, and so on. This article gives a complete solution for certain special cases, and discusses numerical (Monte Carlo) techniques.

I. INTRODUCTION

We consider here a number of problems of the following general type. Let A and B be two sets in the m-dimensional Euclidean space $E^m (m \geqq 1)$. B is assumed to have a center of symmetry and for any point x $B(x)$ denotes the translate of B centered at x. An integer $n(n \geqq 2)$ is fixed and the n-fold Cartesian product $A \times A \times \cdots \times A$ is denoted by P. If $u \,\varepsilon\, P$ then $u = (x_1, \cdots, x_n)$ where $x_i \,\varepsilon\, A$ for $i = 1, \cdots, n$; we shall be interested in the sets $B(x_1), \cdots, B(x_n)$. By a configuration condition we shall understand a statement referring to the relative positions of the sets $B(x_1), \cdots, B(x_n)$ and describing their intersection properties in purely Boolean terms.

Examples of admissible configuration conditions are: (*i*) the n sets are pairwise disjoint, (*ii*) their intersection is empty, (*iii*) their union is connected. A configuration condition which generalizes (*i*) and (*ii*) is: an integer p is given $(2 \leqq p \leqq n)$ and no p of the n sets intersect. Any admissible configuration condition C induces a partition of P into two disjoint and complementary sets $Y = Y(C)$ and $N = N(C)$; if $u = (x_1, \cdots, x_n) \,\varepsilon$

[*] University of British Columbia, Vancouver.

P then $u \varepsilon Y$ if and only if the condition C holds for $B(x_1), \cdots, B(x_n)$. Finally, a function $F = F(x_1, \cdots, x_n)$ is defined over P and dV denotes the volume element $dx_1 \cdots dx_n$. Our problem is to evaluate the integral

$$J = \int_Y F\, dV.$$

In all cases to be considered the sets A, B, and Y, as well as the function F, will be sufficiently regular so that the questions of measurability and integrability will not arise. In fact, in most cases of interest B turns out to be a ball, a cube, or an m-dimensional regular octahedron. All these are Minkowski balls for a suitable norm $\rho(\xi) = \rho\ (\xi_1, \ldots, \xi_n)$. We get the Euclidean ball with

$$\rho(\xi) = \left(\sum_1^n \xi_i^2\right)^{\frac{1}{2}},$$

the cube with $\rho(\xi) = \max\limits_i\ (\xi_1, \ldots, \xi_n)$, and the octahedron with

$$\rho(\xi) = \sum_1^n |\ \xi_i\ |\ .$$

It will be therefore assumed throughout that B is a Minkowski ball. This amounts simply to assuming that B is a convex symmetric body. The precise shape of A is of no particular importance, only its content and sufficient regularity are.

The integrand F will be usually of some highly symmetric type such as

$$F = 1, \qquad F = \prod_{i=1}^n f(x_i), \qquad F = \prod_{1 \leq i < j \leq n} f(|\ x_i - x_j\ |),$$

where f is a suitable sufficiently regular function.

In this part of the paper we are concerned with certain special configuration conditions which lead to an explicit expression for J in terms of the so-called cluster-integrals. Later we consider a related expansion of the form

$$J = \sum_{j=0}^{\infty} J_j \lambda^j \tag{1}$$

where the parameter λ measures the ratio of sizes of B to A. We shall take up the questions of the existence of the expansion (1) and the regularity of J as a function of λ.

II. EXAMPLES

Example 1

Let $m = 1$, A is the interval $[0, L]$ and B is the interval $[0, a]$, n is any integer such that $(n - 1)a \leqq L$, the configuration condition is that the sets $B(x_1), \cdots, B(x_n)$ are disjoint, and $F \equiv 1/L$. J is now the probability that with n points at random on the interval $[0, L]$ no two points are closer than a.

Example 2

Let m, A, and B be as above,

$$F(x_1, \cdots, x_n) = \prod_1^n f(x_i)$$

where $f(x)$ is a probability density on A. The configuration condition is: p is an integer $(2 \leqq p \leqq n)$ and some p-tuple of the sets $B(x_1), \cdots, B(x_n)$ is to have a nonempty intersection. Here we have the following interpretation: $[0, L]$ is a basic time interval and n events occur during that time. Each event occurs independently of the others with the probability density $f(x)$. A p-fold coincidence is defined to be the compound event arising when some p events occur closely together—on a time-interval of length a. Now J is the probability that a p-fold coincidence occurs.

The above examples show that problems of our type might be of interest in queuing theory, telephone traffic, the theory of particle counters, and in similar areas. The next example is a scattering problem for a random linear array of n identical isotropic point-scatters, no two of which can be too close together.

Example 3

Let m, A, B, and C be as in example 1. We suppose that the wavelength is 2π and that L is an integral multiple of it. Aside from proportionality factors the signal scattered by the array is the vector (ξ, η) where

$$\xi = \sum_1^n \cos x_i, \qquad \eta = \sum_1^n \sin x_i .$$

We are here interested in the probability $P(u, v)$ that

$$u \leqq \xi \leqq u + du \quad \text{and} \quad v \leqq \eta \leqq v + dv.$$

The Markov method[1] gives

$$P(u, v) = (2\pi)^{-2}J_A^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-i(ru+sv)}B(r, s)\, dr\, ds$$

where

$$B(r, s) = \int_{Y_A} \exp\left[i\left(r \sum_1^n \cos x_i + s \sum_1^n \sin x_i\right)\right] dV.$$

J_A and Y_A are the integral and the region of example 1, respectively. Therefore the spectrum $B(r, s)$ is obtained in the form of our integral J if we take

$$F = \prod_1^n f(x_i), \qquad f(x) = e^{i(r \cos x + s \sin x)}.$$

When $a = 0$ then $P(u, v)$ reduces to the probability density for the isotropic plane random walk of n unit displacements in arbitrary directions.

Example 4

Let $m = 3$, let A be any large and sufficiently regular portion of space, and let B be the ball of radius a. The configuration condition is that no two sets $B(x_i)$ and $B(x_j)$ overlap. There is a suitable given function $\varphi(x)$ and

$$F(x_1 , \cdots , x_n) = \prod_{1 \leq i < j \leq n} e^{-\varphi(|x_i - x_j|)}.$$

Now, aside from some simple normalization factors, J is the so-called partition function for a hard-sphere model of idealized gas with intermolecular potential φ and the hard core radius a.[2]

The knowledge of J is here of considerable importance in statistical mechanics and a great deal of work has been done on the subject of evaluating J in the form (1) which is closely associated with the so-called virial expansion.

III. A SPECIAL CASE

The method to be used involves certain dissections of Cartesian products together with the inclusion-exclusion principle of combinatorics.[3] As an illustration and an introduction to the more complex examples which follow, we consider here at some length example 1 of the previous section. The material is taken from Ref. 4, where some further details can be found. The well-known solution[5, 6] is here

$$J = J(n, a, L) = [1 - (n - 1)a/L]^n \tag{2}$$

and it may be obtained analytically as follows.

Let the coordinates of the n points be $x_1, \cdots, x_n$; these can be ordered in $n!$ ways. Suppose that $0 \leqq x_1 \leqq x_2 \leqq \cdots \leqq x_n \leqq L$; the conditions of the problem are satisfied if and only if

$$0 \leqq x_1 \leqq x_2 - a \leqq x_3 - 2a$$
$$\leqq \cdots \leqq x_n - (n - 1)a \leqq L - (n - 1)a. \tag{3}$$

Let $y_i = x_i - (i-1)a$ $(i = 1, \ldots, n)$, then the probability that (3) holds is L^{-n} times the volume of the region in E^n consisting of the points $y = (y_1, \ldots, y_n)$ for which

$$0 \leqq y_1 \leqq y_2 \leqq \cdots \leqq y_n \leqq L - (n - 1)a.$$

The volume in question is $[L-(n-1)a^n]/n!$; since there are $n!$ equiprobable orderings we get (2) at once.

Consider next an alternative geometrical proof of (2), which is considerably more complicated, but leads to useful generalizations and gives some additional insight.

First, let $n = 2$. The sample space of pairs $(x_1, x_2)(0 \leqq x_1, x_2 \leqq L)$ is the square Q of side-length L, lying in the first quadrant of E^2 and containing the origin as a vertex. Let D be the diagonal of Q through the origin and draw the two lines parallel to D at the distance $2^{-\frac{1}{2}}a$ from it. The hexagonal subset of Q contained between those two lines is the sample space of the forbidden configurations with $|x_1 - x_2| \leqq a$. The remainder of the square Q consists of two congruent triangles which can be moved together so as to form a square Q_1, of side-length $L - a$. By the randomness assumption $J(2, a, L)$ is the ratio of the areas of Q_1 and Q which yields (2) for $n = 2$.

The case of arbitrary n is handled similarly. In E^n we take a Cartesian coordinate system with the n axes $X_1, \ldots, X_n$. The n-dimensional cube

$$H = \{(x_1, \cdots, x_n): 0 \leqq x_i \leqq L, i = 1, \cdots, n\}$$

is then the sample space of all n-tuples of points on the segment $[0, L]$. Let I_i be the interval $[0, L]$ on the X_i- axis. In the two-dimensional square face $Q_{ij} = I_i \times I_j$ of H let D_{ij} be the diagonal through the origin and let H_{ij} be the hexagonal subset of Q_{ij} consisting of all points no further from D_{ij} than $2^{-\frac{1}{2}}a$. Let S_{ij} be the Cartesian product of H_{ij} with all the I_k's for which $k \neq i$ and $k \neq j$.

S_{ij} is now the sample space of the configurations which are forbidden on account of too close approach of the points x_i and x_j for the chosen indices i and j: $|x_i - x_j| \leqq a$. The sample space Y of the allowed configurations is therefore the set

$$H - \bigcup_{1 \leqq i < j \leqq n} S_{ij} \ .$$

When the $\binom{n}{2}$ paradiagonal slabs S_{ij}, based on the paradiagonal sets H_{ij}, are removed from H, the remainder of the cube H consists of $n!$ congruent simplexes which can be reassembled by suitable translations so as to form a smaller cube H_1 of sidelength $L - (n - 1)a$. By the randomness assumption $J(n, a, L)$ is the ratio of the volumes of the cubes H_1 and H, and so (2) is proved again.

The above procedure works on account of a lucky geometrical accident of the fitting of $n!$ simplexes. If A and B were some other, m-dimensional, sets, we could still form the paradiagonal sets and slabs and we could attempt to find the volume of the union $\bigcup S_{ij}$ of all the paradiagonal slabs. This is essentially what is done in the next section by means of the inclusion-exclusion principle[3].

IV. SIMPLE COINCIDENCE WITH SEPARABLE INTEGRAND

In this section we are concerned with a configuration condition corresponding to simple coincidence: $u = (x_1, \cdots, x_n) \varepsilon Y$ if and only if for some i and j $B(x_i)$ and $B(x_j)$ intersect. Subject to general restrictions, A, B, m, and n are arbitrary. We let $N = \binom{n}{2}$ and we form the N paradiagonal sets

$$H_{ij} = \{(x_i, x_j) : B(x_i) \cap B(x_j) \neq \phi\}$$

and the N paradiagonal slabs

$$S_{ij} = \{(x_1, \cdots, x_n) : B(x_i) \cap B(x_j) \neq \phi\}.$$

Let the slabs be enumerated by a single index as $\{S_k\}$, $k = 1, \ldots, N$. Then an application of the inclusion-exclusion principle gives

$$J = \int_Y F\,dV = \sum_{r=1}^{n} (-1)^{r+1} \Bigg[\sum_{1 \leqq k_1 < k_2 < \cdots < k_r \leqq N} \!\!\cdots\! \sum \int_{S_{k_1} \cap \cdots \cap S_{k_r}} F\,dV \Bigg] \qquad (4)$$

$$= \sum_{r=1}^{n} (-1)^{r+1} K_r \ .$$

With the general integrand no further elaboration of (4) is possible. Suppose now that F has the separable form

$$F = \prod_1^n f(x_i). \tag{5}$$

With the double-index enumeration of the S_{ij}'s the first term K_1 can be written as

$$K_1 = \sum_{1 \leqq i_1 < j_1 \leqq n} \int_{S_{i_1 j_1}} \prod_1^n f(x_i)\, dV$$

and since all the N paradiagonal sets are congruent, we have

$$K_1 = N\left(\int_A f(x)\, dx\right)^{n-2} \int_{H_{12}} f(x_1)f(x_2)\, dx_1\, dx_2 \,.$$

For reasons which will be clear shortly we write

$$N_{11} = N, \qquad \int_A f(x)\, dx = J_0\,, \qquad \int_{H_{12}} f(x_1)f(x_2)\, dx_1\, dx_2 = J_{11} \tag{6a}$$

so that

$$K_1 = N_{11} J_0^{n-2} J_{11}\,. \tag{6b}$$

Similarly, the second term K_2 in (4) is

$$K_2 = \sum_{(i_1, j_1)} \sum_{(i_2, j_2)} \int_{S_{i_1 j_1} \cap S_{i_2 j_2}} \prod_1^n f(x_i)\, dV$$

where the summation extends over all distinct pairs (i_1, j_1), (i_2, j_2) such that $1 \leqq i_1 < j_1 \leqq n$, $1 \leqq i_2 < j_2 \leqq n$; no regard is paid to the order of pairs; [(1, 2), (3, 4) is the same as (3, 4), (1, 2)] so that there are exactly

$$\binom{\binom{n}{2}}{2}$$

such pairs of pairs. There are two types of these: N_{21} pairs like (1, 2), (3, 4) with all four indices different, and N_{22} pairs like (1, 2), (1, 3) with one shared index. By a simple calculation

$$N_{21} = n(n-1)(n-2)(n-3)/8, \qquad N_{22} = n(n-1)(n-2)/2,$$
$$N_{21} + N_{22} = \binom{\binom{n}{2}}{2} \tag{7a}$$

and in analogy to (6) we set

$$J_{21} = \int_{H_{12} \cap H_{34}} \prod_{1}^{4} f(x_i)\, dx_1\, dx_2\, dx_3\, dx_4$$

$$= \left[\int_{H_{12}} f(x_1) f(x_2)\, dx_1\, dx_2 \right]^2 = J_{11}^2 \tag{7b}$$

$$J_{22} = \int_{H_{12} \cap H_{13}} \prod_{1}^{3} f(x_i)\, dx_1\, dx_2\, dx_3$$

so that

$$K_2 = N_{21} J_0^{n-4} J_{11}^2 + N_{22} J_0^{n-3} J_{22}\ . \tag{7c}$$

The main purpose of this section is to develop formulae, analogous to (6) and (7), for the general term K_r of (4). The principal difficulty here is that in passing from the single-index formula for K_r

$$K_r = \sum_{1 \leq k_1 < \cdots < k_r \leq N} \cdots \sum \int_{S_{k_1} \cap \cdots \cap S_{k_r}} \prod_{i=1}^{n} f(x_i)\, dV \tag{8a}$$

to the double-index formula

$$K_r = \sum_{(i_1, j_1)} \cdots \sum_{(i_r, j_r)} \int_{S_{i_1 j_1} \cap \cdots \cap S_{i_r j_r}} \prod_{i=1}^{n} f(x_1)\, dV \tag{8b}$$

we need an adequate description of the different types of r-tuples of pairs of indices occurring in (4), together with a hold on the range of summation in (8b). For instance, with $r = 2$ there are two such types, illustrated by (1, 2), (3, 4) and (1, 2), (1, 3). With $r = 3$ there are five types of index-sharing in triples of pairs:

(1, 2), (3, 4), (5, 6); (1, 2), (1, 3), (4, 5); (1, 2), (2, 3), (3, 4);
(1, 2), (1, 3), (1, 4); (1, 2), (1,3) (2, 3); (9)

We may therefore expect that the formula for $r = 3$, analogous to (7c) for $r = 2$, will have five terms rather than two. The number of such types grows very rapidly with r, and as an aid we introduce certain graphs associated with the terms of (8). These graphs reflect completely the intersection properties of the sets $B(x_1), \ldots, B(x_n)$. For $r = 3$ there are five such graphs corresponding to the five types enumerated in (9). These are given in Fig. 1 together with the corresponding B-configurations. (It is, of course, assumed that $n > 2$.)

Each graph is of the following kind:

(*i*) No vertex is isolated.
(*ii*) No pair of vertices is connected by more than one edge.

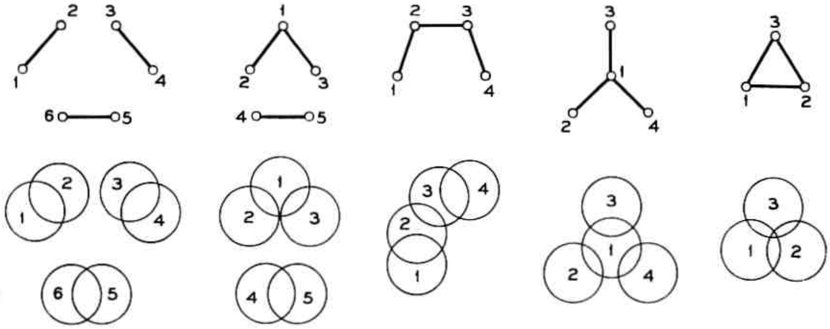


Fig. 1 — Coincidence graphs, $r = 3$.

(*iii*) No edge connects a vertex to itself.
(*iv*) There are exactly r edges.
(*v*) There are exactly v vertices.

One further, and crucial, condition is added:

(*vi*) If the v vertices are enumerated in some order then there exists a configuration of v translates $B_1, \cdots, B_v$ of B, such that B_i and B_j intersect if and only if the i^{th} and the j^{th} vertices are connected by an edge.

For the sake of convenience we make here the following convention: two convex m-dimensional bodies will be said to intersect only if their intersection is itself m-dimensional, otherwise they are to be regarded as disjoint. The reason for this is that we are interested in purely metric properties: the intersections of such sets serve as domains of integration for well-behaved functions in E^m.

A graph satisfying conditions *i* through *vi* will be called a (B, r, v)-graph, one satisfying *i* through *iv* and *vi* a (B, r)-graph, and one satisfying *i* through *iii* and *vi* a B-graph. It must be emphasized that the condition *vi* is not of the usual graph-theoretic kind and it prevents many graphs from being B-graphs. For instance, let $m = 2$ and let B be a circular disk. Since a disk in E^2 cannot intersect six congruent pairwise disjoint disks, the graphs of Fig. 2 are not B-graphs.

The proof of the above assertion for the graph of Fig. 2b is obtained by showing that here the "extreme" configuration is that of Fig. 3.

Similarly, when $m = 2$ and B is a square then B cannot intersect five pairwise disjoint translates of itself (for each translate contains

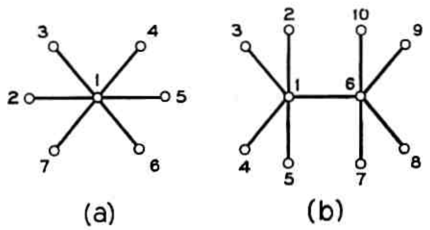


Fig. 2 — Graphs which are not B-graphs. (B is a disk.)

a vertex of B) so that the graph of Fig. 4a is not a B-graph. On the other hand, the graph of Fig. 4b, which corresponds to that of Fig. 2b, is a B-graph as shown by the configuration of Fig. 4c.

Returning to the evaluation of K_r, we start with (8b). Summation there extends over all the

$$\binom{\binom{n}{2}}{2}$$

distinct r-tuples of pairs of indices where for each pair $(i_s, j_s) 1 \leqq i_s < j_s \leqq n$; r-tuples differing only in the order of pairs are not considered distinct. We can now associate the terms of (8b) in a 1 : 1 fashion with the distinct (B, r)-graphs on some n vertices $w_1, \cdots, w_n$. Given a B-graph G let

$$S(G) = \bigcap S_{ij}, \tag{10}$$

where the intersection is taken over all pairs (i, j) for which w_i is connected to w_j by an edge in G. Then (8b) may be written as

$$K_r = \sum_G \int_{S(G)} \prod_{i=1}^{n} f(x_i)\, dV, \tag{11}$$

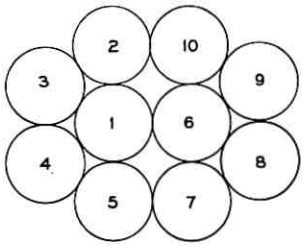


Fig. 3 — An extreme B-configuration.

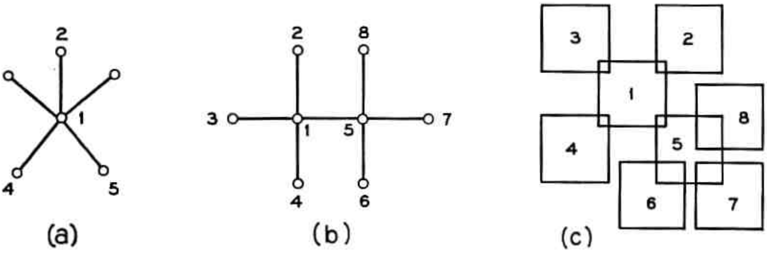


Fig. 4 — Configurations when B is a square.

the summation running over all distinct (B, r)-graphs on n vertices.

Let $v(G)$ denote the number of vertices of G and $C(G)$ a connected component of G. Since the integrand in (11) is completely separable, the integral over $S(G)$ splits into a product of integrals over the connected components and we get

$$K_r = \sum_G J_0^{n-v(G)} \prod_{C(G)} J[C(G)]. \tag{12}$$

Here $J[C(G)]$ is an integral over the connected component and the product is taken over all such components of G. Two examples of integrals $J[C(G)]$ are given in (7b). Owing to the congruence of all the paradiagonal slabs and the form of the integrand, it is not necessary to sum in (12), over all (B, r)-graphs on the vertices $w_1, \ldots,$ w_n, but only over their types.

Suppose that there are exactly $t = t(r)$ types of such graphs and let G_j be any one of the j^{th} type; let also $N_{rj}(n)$ be the number of different (B, r)-graphs on the vertices $w_1, \ldots, w_n$, of the j^{th} type. Then (12) becomes

$$K_r = \sum_{j=1}^{t(r)} N_{rj}(n) J_0^{n-v(G_j)} \prod_{C(G_j)} J[C(G_j)]. \tag{13}$$

Thus the problem of evaluating J has been reduced through (4) and (13) to: the geometrical problem of determining the types of (B, r)-graphs, the combinatorial problem of calculating the coefficients $N_{rj}(n)$, and the analytical problem of evaluating the cluster-integrals over the connected (B, r)-graphs.

V. MULTIPLE COINCIDENCE WITH SEPARABLE INTEGRAND

Formulae analogous to those of the previous section will now be obtained for the case of p-tuple coincidence. Subject to general conditions,

A, B, n, and m are arbitrary and F is of the separable form (5). An integer p is fixed ($2 \leqq p \leqq n$) and the configuration condition is: $u = (x_1, \cdots, x_n) \varepsilon Y$ if and only if there are p indices $i_1, \cdots, i_p (1 \leqq i_1 < \cdots < i_p \leqq n)$ such that

$$\bigcap_{s=1}^{p} B(x_{i_s}) \neq \phi.$$

We observe here our convention that the intersection must be itself m-dimensional. We introduce the analogs of paradiagonal sets and slabs:

$$H_{i_1 \cdots i_p} = \left\{(x_{i_1}, \cdots, x_{i_p}) : \bigcap_{s=1}^{p} B(x_{i_s}) \neq \phi\right\},$$

$$S_{i_1 \cdots i_p} = \left\{(x_1, \cdots, x_n) \quad : \bigcap_{s=1}^{p} B(x_{i_s}) \neq \phi\right\},$$

we let $M = \binom{n}{p}$, and we re-enumerate the M sets $S_{i_1 \cdots i_p}$ with a single index k as $\{S_k\}$, $1 \leqq k \leqq M$. Then we get a formula analogous to (4):

$$J = \int_Y F\,dV = \sum_{r=1}^{n} (-1)^{r+1} \left[\sum_{1 \leqq k_1 < \cdots < k_r \leqq M} \cdots \sum \int_{S_{k_1} \cap \cdots \cap S_{k_r}} F\,dV \right]$$

$$= \sum_{r=1}^{n} (-1)^{r+1} U_r \,. \tag{14}$$

As in (6a) we let

$$M_{11} = M, \quad \int_A f(x)\,dx = J_0\,, \quad \int_{H_{12 \cdots p}} \prod_1^p f(x_i)\,dx_1 \cdots dx_p = J_{11}\,,$$

to get

$$U_1 = M_{11} J_0^{n-p} J_{11}\,.$$

In terms of p-tuple indices the second term U_2 of (14) is

$$U_2 = \sum_{(i_1, \cdots, i_p)} \sum_{(j_1, \cdots, j_p)} \int_{S_{i_1 \cdots i_p} \cap S_{j_1 \cdots j_p}} \prod_1^n f(x_i)\,dV.$$

The summation extends over the $\binom{n}{2}$ distinct pairs of p-tuples. We have now p types of such pairs, depending on the number of shared indices, which may be 0, 1, $\cdots$, or $p - 1$. Let M_{2j} be the number of p-tuple pairs of type j (that is, with $j - 1$ indices shared) and put

$$J_{2j} = \int_{H_{12 \cdots p} \cap H_{p+2-j \cdots 2p-j+1}} \prod_{i=1}^{2p-j+1} f(x_i)\,dx_1 \cdots dx_{2p-j+1}\,,$$

then

$$U_2 = \sum_{j=1}^{p} N_{2j} J_0^{n-2p+j-1} J_{2j} \ .$$

Observe that the integral J_{21} splits into a product: $J_{21} = J_{11}^2$. To get an expression for arbitrary U_r we introduce a higher-dimensional equivalent of B-graphs. Let X be a regular simplex in E^{n-1} on the vertices $w_1, \cdots, w_n$. On account of properties *i* through *iii* listed in Section IV, a (B, r)-graph is simply a set of certain r edges (or one-dimensional faces) of X. A d-dimensional hypergraph G will be just a set of some of the $\binom{n}{d+1}$ d-dimensional faces of X. This takes care of properties *i* through *iii*. When there are r such faces in G we shall speak of an (r)-hypergraph and when these faces comprise between them v vertices of X, G will be called an (r, v)-hypergraph.

An equivalent of the important condition (*vi*) is very naturally obtained: there is a B-configuration of v translates $B_1, \cdots, B_v$ of B, such that any $d + 1$ of them, say, $B_{i_1, \ldots, i_{d+1}}$ intersect if and only if $w_{i_1}, \cdots, w_{i_{d+1}}$ are the vertices of a d-dimensional face of X included in G. Components, types, and so on, for (B, r)-hypergraphs are defined in the same way as before. For instance, a hypergraph G is connected if no plane disjoint from it can strictly separate some of its d-faces from others. All quantities such as $C(G)$ and $v(G)$ have the same meaning as before. Let $t = t\,(r, d)$ be the number of different types of (B, r)-hypergraphs, let G_j be any one hypergraph of the j^{th} type, and let $M_{rj}^{d}(n)$ be the number of different (B, r)-hypergraphs of the j^{th} type on the n vertices. Then, proceeding as before, we get the equivalent of (13):

$$U_r = \sum^{t(r,p-1)} M_{rj}^{p-1}(n) J_0^{[n-v(G_j)]} \prod_{C(G_j)} J[C(G_j)]. \tag{15}$$

VI. SOME COMBINATORIAL PROPERTIES OF B-GRAPHS AND B-HYPERGRAPHS

Let $\varphi(r)$ and $\psi(r)$ be the smallest and the largest number of vertices, respectively, in a (B, r, v)-graph G. From conditions *i* through *iii* we have at once $\psi(r) = 2r$. G is then minimally connected with r components (Fig. 5a). Suppose that r is a triangular number: $r = s(s-1)/2$; there is then a complete graph on s vertices which is clearly a B-graph for any B, so that $s = v$. If r is not a triangular number let $t(t-1) < 2r < t(t+1)$ and put $e = r - t(t-1)/2$.

Let G be the complete graph on t vertices. For the corresponding B-configuration we may assume that the translates $B_1, \cdots, B_t$ of B

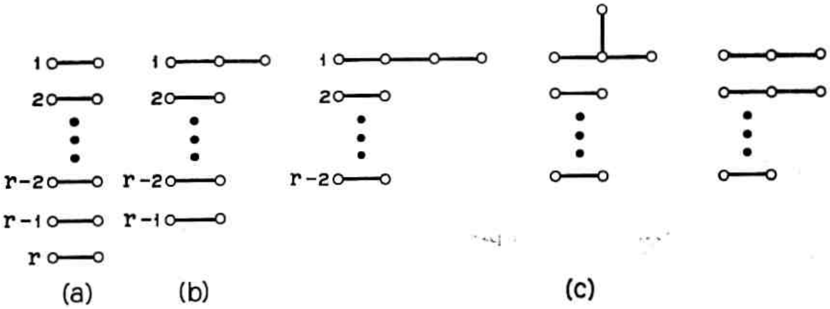


Fig. 5 — (B, r, v)—graphs with high v.

have an interior point in common. We check that $e < t$ and that $B_1, \cdots, B_t$ may be arranged so that a point $z \in \bigcap_1^e B_i$ can be strictly separated from $\bigcup_{e+1}^t B_i$ by a plane P. Let B_{t+1} be a translate of B which contains z and lies strictly on the same side of P as z. Then the resulting B-configuration $B_1, \cdots, B_{t+1}$ corresponds to a (B, r, v)-graph G with $v = t + 1$. This G may be said to be a maximally connected (B, r, v)-graph. We have now

$$\psi(r) = 2r, \qquad \varphi(r) = \min_j \{j : j \geqq [1 + (1 + 8r)^{\frac{1}{2}}]/2\}.$$

Similarly, let $\varphi(r, d)$ and $\psi(r, d)$ be the corresponding minimum and maximum of v for a (B, r, v)-hypergraph. Then clearly $\psi(r, d) = (d + 1)r$. To determine $\varphi(r, d)$ we suppose first that $r = \binom{s}{d+1}$. There is then a complete hypergraph on s vertices, consisting of all the d-dimensional faces of an $(s - 1)$-dimensional simplex. This hypergraph is a B-hypergraph for any B and so $v = s$. If

$$\binom{t}{d+1} < r < \binom{t+1}{d+1}$$

we proceed as before and find that $v = t + 1$. Hence

$$\psi(r, d) = (d + 1)r, \ \varphi(r, d) = \min_j \{j : \\ j \geqq \text{largest pos. root of } x(x - 1) \cdots (x - d) = (d + 1)!r\}.$$

The bounds $\varphi(r)$ and $\psi(r)$ lead us to the possibility of a combinatorial identity

$$\binom{\binom{n}{2}}{r} = \sum_{k=\varphi(r)}^{\psi(r)} A_{rk}\binom{n}{k} \tag{16}$$

and its relation to B-graphs and the numbers $N_{rj}(n)$. For instance, we find for $r = 2$

$$\binom{\binom{n}{2}}{2} = 3\binom{n}{4} + 3\binom{n}{3},$$

$\binom{\binom{n}{2}}{2}$ is the total number $N_{21} + N_{22}$ of graphs of (7a) and

$$3\binom{n}{4} = N_{21}, \qquad 3\binom{n}{3} = N_{22}.$$

To prove the validity of an expansion like (16) for all n observe that the left-hand side is a polynomial in n of degree $2r = \psi(r)$ so that

$$\binom{\binom{n}{2}}{r} = \sum_{k=0}^{2r} A_{rk}\binom{n}{k}.$$

Further, $A_{rk} = 0$ for $k < \varphi(r)$, for we substitute successively $n = 0, 1, \cdots, \varphi(r) - 1$ in (16) and recall that $\binom{p}{r} = 0$ for $p < r$. By expanding both sides of (16) in powers of n and comparing the coefficients we find

$$A_{r\,2r} = (2r)!/2^r r!, \quad A_{r\,2r-1} = (2r-1)!/2^{r-1}(r-2)!,$$
$$A_{r\,2r-2} = (2r-2)!(3r-1)/3.2^{r-1}(r-3)!$$

and so on. Therefore (16) may be written as

$$\begin{aligned}\binom{\binom{n}{2}}{r} &= (n)_{2r}/(2^r r!) + (n)_{2r-1}/[2^{r-1}(r-2)!] \\ &\qquad + (n)_{2r-2}/[3.2^{r-1}(r-3)!\,(3r-1)] + \cdots \qquad (17) \\ &= \sum_{i=0}^{2r-\varphi(r)} (n)_{2r-i}/D_i\,.\end{aligned}$$

$(n)_p$ stands for $n(n-)\ \ldots\ (n-p+1)$.

The denominators D_i have the following interpretation. Consider first the $(B, r, 2r)$-graph of Fig. 5a. The $2r$ vertices can be chosen out of $w_1, \ldots, w_n$ *in* $(n)_{2r}$ ways. We define the symmetry number for a (B, r, v)-graph to be the number of ways in which its vertices can be labelled with integers $1, 2, \ldots, v$, all of which ways are to correspond to the same B-configuration. Here the symmetry number is $2^r r!$, as there are 2^r ways of permuting the labels on the two vertices of a component and $r!$ ways of permuting their components. This leads us to the first term $(n)_{2r}/2^r r!$ in (17) which is precisely the number $N_{r1}(n)$ of (13) provided that we consider G_1 in (13) to be of the type of Fig. 5a.

Similarly, for the $(B,r,2r-1)$-graph of Fig. 5b we find the symmetry number to be $2^{r-1}(r-2)!$. The number of ways to choose the $2r-1$ vertices is $(n)_{2r-1}$ and so we get the second term $(n)_{2r-1}/2^{r-1}(r-2)!$ of (17). The situation gets somewhat more complicated for the $(B,r,2r-2)$-graphs. Here we have three types instead of one, illustrated in Fig. 5c. The $2r-2$ vertices can be selected in $(n)_{2r-2}$ ways, the symmetry numbers for the three types are

$$2^{r-2}(r-3)!, \quad 3.2^{r-2}(r-3)!, \text{ and } 2^{r-1}(r-4)!. \tag{18}$$

Therefore, the corresponding numbers of graphs, say $N_{r3}(n)$, $N_{r4}(n)$, $N_{r5}(n)$ are

$$(n)_{2r-2}/[2^{r-2}(r-3)!], \; (n)_{2r-2}/[3.2^{r-2}(r-3)!], \; (n)_{2r-2}/[2^{r-1}(r-4)!]$$

and their sum is precisely the third term of (17). The corresponding denominator D_2 is therefore three times the harmonic mean of the three symmetry numbers in (18).

Thus the first few terms of (17) give the total numbers

$$\sum_j N_{rj}(n)$$

of (B,r,v)-graphs for $v = 2r, 2r-1$, and so on. However, this pleasing circumstance breaks down as soon as we reach the smallest term i for which one of the types of graphs in question is not a B-graph.

For the case $m = 2$, B a circular disk, this occurs for $i = 7$ and the graph in question is then that of Fig. 2a together with other components containing one edge each. When B is a square the graph of Fig. 4a shows that the breakdown occurs for $i = 6$. On the other hand, the quantity $(n)_{2r-i}/D_i$ from (17) always provides an upper bound for the sum $\sum N_{rj}(n)$, the summation extending over all types j of $(B, r, 2r - i)$-graphs.

The explicit form of (16) is

$$\binom{\binom{n}{2}}{r} = \sum_{k=q}^{2r} A_{rk}\binom{n}{k} \tag{19}$$

where $q = \varphi(r)$ and

$$A_{rk} = \sum_{j=0}^{k-q} (-1)^j \binom{k}{j}\binom{\binom{k-j}{2}}{r}. \tag{20}$$

We prove (2) by induction on k. For $k = q$, (20) holds, suppose it to be proved for $k \leqq q + s - 1$. Let $n = q + s$ in (19), then

$$A_{rq+s} = \binom{\binom{q+s}{2}}{r} - \sum_{i=0}^{s-1} \binom{q+s}{q+i} A_{rq+i}$$

which by the induction hypothesis may be written as

$$A_{rq+s} = \binom{\binom{q+s}{2}}{r} - \sum_{i=0}^{s-i} \sum_{j=0}^{i} (-1)^j \binom{q+s}{q+i} \binom{\binom{q+i-j}{2}}{r}.$$

In the double sum we may sum first over those terms for which the difference $u = i - j$ is constant, then over u. In this way one gets

$$A_{rq+s} = \binom{\binom{q+s}{2}}{r} + \sum_{u=0}^{s-1} (-1)^{s-u} \binom{q+s}{q+u} \binom{\binom{q+s}{2}}{r}$$

which after some simple algebra becomes (20) with $k = q+s$. This completes the induction and the proof of (20).

Some combinatorial identities may be obtained from the above. For example, we know that $A_{r2r} = (2r)!/2^r r!$. Hence, on putting $k = 2r$ in (20), we get

$$\sum_{j=0}^{2r-q} (-1)^j \binom{2r}{j} \binom{\binom{2r-j}{2}}{r} = (2r)!/2^r r!. \tag{21}$$

Similarly, with $k = 2r - 1$ and $k = 2r - 2$ we get

$$\sum_{j=0}^{2r-q-1} (-1)^j \binom{2r-1}{j} \binom{\binom{2r-j-1}{2}}{r} = (2r-1)!/2^{r-1}(r-2)! \tag{22}$$

and

$$\sum_{j=0}^{2r-q-2} (-1)^j \binom{2r-2}{j} \binom{\binom{2r-j-2}{2}}{r}$$
$$= [(2r-2)!\,(r-1/3)]/2^{r-1}(r-3)!. \tag{23}$$

For hypergraphs we have the identity

$$\binom{\binom{n}{d}}{r} = \sum_{k=q}^{dr} A_{rk}(d) \binom{n}{k} \tag{24}$$

where $q = \varphi(r,d)$. The explicit expression for the coefficients $A_{rk}(d)$ can be found in the same way as (20):

$$A_{rk}(d) = \sum_{j=0}^{k-q} (-1)^j \binom{k}{j} \binom{\binom{k-j}{d}}{r}. \tag{25}$$

Some of the higher coefficients $A_{r\,dr}, A_{r\,dr-1}, \ldots$ can be evaluated by comparing the powers of n in (24):

$$A_{r\,dr}(d) = (dr)!/r!(d!)^r, \quad A_{r\,dr-1}(d) = (dr)!d(r-1)/2.r!(d!)^r$$

and so on, so that by putting $k = rd$ and $k = rd - 1$ in (25) we get

$$\sum_{j=0}^{dr-q} (-1)^j \binom{\binom{dr-j}{d}}{r} = (dr)!/r!\,(d!)^r \tag{26}$$

and

$$\sum_{j=0}^{dr-q-1} (-1)^j \binom{dr-1}{j} \binom{\binom{dr-j-1}{d}}{r} = (dr)!\,d(r-1)/2.r!\,(d!)^r. \tag{27}$$

The coefficients $A_{kr}(d)$ have the same interpretation with hypergraphs as the A_{kr} have with ordinary graphs, and they refer to symmetry numbers.

VII. SIMPLE COINCIDENCE IN A CUBE

We consider here the problem of evaluating the probability $P(n,a,L)$ that when n points are taken at random (uniform distribution) in a three-dimensional cube of edge-length L, then no two points are closer than a. The problem occurs in deriving the van der Waals equation from a primitive hard-sphere gas model. See, for instance, Ref. **2**, where the problem is termed "very difficult" and the crude (though sufficient) approximation

$$P(n, a, L) \cong \prod_{j=1}^{n-1} (1 - 4\pi j a^3/3L^3) \cong 1 - 2\pi n^2 (\mathrm{a}/L)^3/3 \tag{28}$$

is used.

From our formulation we find that

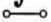
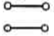
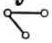
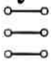
$$L^{3n}[1 - P(n, a, L)]$$

is the J integral for the case $m = 3$, A is a cube of volume L^3, B a ball of radius $a/2$, and the configuration condition is that not all sets $B(x_i)$ be disjoint; in other words, a simple coincidence. Therefore by (4), (13), and an inspection of Fig. 1 we have

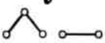
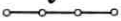
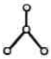
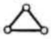
$$\begin{aligned} L^{3n}[1 - P(n, a, L)] = N_{11}L^{3n-6}I_{11} &- (N_{21}L^{3n-12}I_{21} + N_{22}L^{3n-9}I_{22}) \\ &+ (N_{31}L^{3n-18}I_{31} + N_{32}L^{3n-15}I_{32} + N_{33}L^{3n-12}I_{33} \\ &+ N_{34}L^{3n-12}I_{34} + N_{35}L^{3n-9}I_{35}) - \cdots \end{aligned} \tag{29}$$

where the integrals $I_{11}, I_{21}, \ldots$ can be symbolically represented as follows

$$I_{11} = \int , \quad I_{21} = \int = I_{11}^2 , \quad I_{22} = \int , \quad I_{31} = \int = I_{11}^3 , \tag{30}$$

$$I_{32} = \int = I_{11}I_{22} , \quad I_{33} = \int , \quad I_{34} = \int , \quad I_{35} = \int .$$

To obtain an explicit Cartesian expression for an integral, I, we consider its signature graph G which is a (B,r,v)-graph. If the v vertices are enumerated as $1,2, \ldots, v$ in an arbitrary order then I becomes a $3v$-tuple integral

$$I = \int \underset{R_I}{\cdots} \int dr_1 \cdots dr_v \tag{31a}$$

where r_i is the vector (x_i, y_i, z_i), dr_i stands for $dx_i\, dy_i\, dz_i$, and the region of integration R_I is given by $3v + r$ inequalities:

$$0 \leqq x_i \leqq L,\ 0 \leqq y_i \leqq L,\ 0 \leqq z_i \leqq L,\ (i = 1, \cdots, v), \tag{31b}$$

$|r_i - r_j|^2 \leqq a^2$ if the i^{th} and the j^{th} vertices are connected (31c)

in G by an edge.

Further, such an integral occurs in (29) with the multiplier $N_{rj}L^{3n-3v}$ where N_{rj} is the number of distinct graphs on n vertices, which are of the same type as G. Together with each such integral $I = I_{pq}$ we may also consider the corresponding integral K_{pq} given by

$$K_{pq} = \int \underset{Q_I}{\cdots} \int dr_1 \cdots dr_v ,$$

where the region Q_I is given by the $(v^2 + 5v)/2$ inequalities (31b), (31c) and

$|r_i - r_j|^2 \geqq a^2$ if the i^{th} and the j^{th} vertices are not connected in G (31d)

by an edge.

It turns out that the I integrals are expressible in terms of the K integrals, and conversely. For instance, consider the K integral with the signature graph which has four vertices 1, 2, 3, and 4, and edges

12, 13, 23, and 24. We write it in a self-explanatory terminology as

$$(12)(13)(23)(24)[1 - (14)][1 - (34)]$$

and multiply this out to get

$$(12)(13)(23)(24) - (12)(13)(23)(24)(14) - (12)(13)(23)(24)(34) + (12)(13)(23)(24)(14)(24)(34)$$

which yields at once a representation of K as a sum of four I-integrals.

The first integral I_{11} is sixtuple and can be reduced to an iterated integral as follows:

$$I_{11} = \int_{m_6}^{M_6} \cdots \int_{m_1}^{M_1} dx_1\, dx_2\, dy_1\, dy_2\, dz_1\, dz_2 \tag{32}$$

where

$$m_2 = m_4 = m_6 = 0, \quad M_2 = M_4 = M_6 = L$$

and

$$\begin{aligned} m_1 &= \max\{0, x_2 - [a^2 - (y_1 - y_2)^2 - (z_1 - z_2)^2]^{\frac{1}{2}}\}, \\ m_1 &= \min\{L, x_2 + [a^2 - (y_1 - y_2)^2 - (z_1 - z_2)^2]^{\frac{1}{2}}\}, \\ m_3 &= \max\{0, y_2 - [a^2\}(z_1 - z_2)^2]^{\frac{1}{2}}\}, \\ M_3 &= \min\{L, y_2 + [a^2 - (z_1 - z_2)^2]^{\frac{1}{2}}\}, \\ m_5 &= \max\{0, z_2 - a\}, \\ M_5 &= \min\{L, z_2 + a\}. \end{aligned}$$

This arrangement of the limits of integration corresponds to taking two balls of radii $a/2$ and centers (x_1, y_1, z_1) and (x_2, y_2, z_2), and letting the center of the first ball move freely over the cube while the coordinates of the second center vary so that the balls intersect. Accordingly, I_{11} has a simple probabilistic interpretation: $I_{11} = L^6[1 - P(2,a,L)]$, where $P(2,a,L)$ is the probability that two points taken at random in the cube of edge-length L are no nearer than a. Similar probabilistic interpretation holds for any other K integral. If G is its (B,r,v)-graph then K *is* L^{3v} times the probability that when v balls of radius $a/2$ are taken with their centers at random in the cube, then the balls are in the configuration of G (so that two of them intersect if and only if the corresponding vertices of G are connected by an edge).

We evaluate now the integral (32) subject to the condition $a \leqq L$.

Integration with respect to x_1 and x_2 gives

$$L^2 - [\max(0, L - D)]^2 \tag{33}$$

where

$$D^2 = a^2 - (y_1 - y_2)^2 - (z_1 - z_2)^2.$$

Since $a \leqq L$ we have $D \leqq L$ and therefore (33) is $2LD - D^2$. Integrating this with respect to y_1 and y_2 we get first, on putting $y_1 - y_2 = u$,

$$\int_0^L \int_{-m}^{M} [2L(b^2 - u^2)^{\frac{1}{2}} - (b^2 - u^2)]\, du\, dy_2$$

where

$$b^2 = a^2 - (z_1 - z_2)^2,\ m = \min(y_2, b),\ M = \min(L - y_2, b).$$

Again, $a \leqq L$ implies $b \leqq L$ and the double integral is therefore

$$\pi L^2 b^2 - 8Lb^3/3 + b^4/2.$$

Finally, integrating with respect to z_1 and z_2 we get

$$I_{11} = 4\pi a^3 L^3/3 - 3\pi a^4 L^2/2 + 8a^5 L/5 - a^6/6, \quad 0 \leqq a \leqq L. \tag{34}$$

There are two more forms of I_{11}, corresponding to the ranges $L \leqq a \leqq 2^{\frac{1}{2}}L$ and $2^{\frac{1}{2}}L \leqq a \leqq 3^{\frac{1}{2}}L$, but they do not appear to be expressible in terms of elementary or standard transcendental functions. It may be observed that the leading term in (34) is the product of the volumes of the cube and the ball of radius a.

To get a better approximation to $P(n,a,L)$ than (28), we examine (29) and find that for small a every integral I_{ij}, beyond I_{11}, is $O\ (a^6)$. Therefore

$$P(n, a, L)$$
$$= 1 - \binom{n}{2}[4\pi/3(a/L)^3 - 3\pi/2(a/L)^4 + 8/5(a/L)^5] + O[(a/L)^6]. \tag{35}$$

It is possible to find the exact limit of $P(n,a,L)$ as

$$n \to \infty,\ a \to 0,\ (4\pi/3)(n^2/2)(a^3/L^3) \to b.$$

For we have then $P(n,a,L) = P(b)$ and

$$1 - P(b) = N_{11}I_{11}/L^6 - N_{21}I_{21}/L^{12} + N_{31}I_{31}/L^{18} - \cdots$$

and

$$N_{k1} \cong \binom{\binom{n}{2}}{k} \cong n^{2k}/2^k k^2$$

$$I_{k1} = (I_{11})^k.$$

This amounts to neglecting all graphs other than the "principal" one, for each k, that is, the one corresponding to the configuration of Fig. 5a. Hence

$$1 - P(b) = \sum_{j=1}^{\infty} (-1)^{j+1}[(4\pi/3)(n^2/2)(a/L)^3]^j/j! = 1 - e^{-b}$$

so that

$$P(b) = e^{-b}. \tag{36}$$

VIII. NUMERICAL EVALUATION OF THE I-INTEGRALS

Since no I integral beyond I_{11} appears to be explicitly evaluable in terms of standard functions, the possibility was investigated of computing those integrals numerically by the Monte Carlo method. The first set of trial calculations was performed on I_{11} itself, in order to be able to compare the results with the known true value. We assume as before that $a \leq L$ and we put $L = 1$ (homogeneity!) to get

$$I_{11}(a) = 4.1888a^3 - 4.7129a^4 + 1.6000a^5 - 0.1667a^6, \quad 0 \leqq a \leqq 1.$$

We now choose a suitable integer M and set the value of a at $1/M$. Next, two points $p_1(x_1, y_1, z_1)$ and $p_2(x_2, y_2, z_2)$ are taken at random in the unit cube by choosing each coordinate to be a random number from the rectangular distribution on $[0, 1]$. Such pairs of random points are selected N times; suppose that in N_1 of them the distance between the random points does not exceed $1/M$, then the quotient N_1/N is taken as the Monte-Carlo approximation to $I_{11}(1/M)$. Then the whole procedure is repeated with $1/M$ replaced by $2/M$, $3/M$, and so on, until the value $3^{1/2}$ is passed. The whole calculation will be referred to as an N by M Monte Carlo run.

In the first set of trial computations N by M Monte Carlo runs were executed for various values of N and M, and in each case a least-squares fit was done on these data by a polynomial of the form

$$\sum_{j=3}^{6} A_j a^j.$$

The results are shown in Table 1.

TABLE I—FIRST TRIAL COMPUTATIONS

	True value of A_j	4.1888	−4.7129	1.6000	−0.1667
1st	1000 by 20 run	3.1873	−1.0742	−2.5488	1.3448
2nd	1000 by 20 run	3.3296	−1.6727	−1.9088	1.1584
	10000 by 20 run	4.4765	−5.9918	3.4012	−0.9760
	1000 by 200 run	4.3008	−5.2689	2.4437	−0.5641
	10000 by 200 run	4.1974	−4.7337	1.6358	−0.1911
	100000 by 20 run	4.1546	−4.5615	1.4043	−0.0879

It appears from this polynomial that very long and large runs are necessary to determine the coefficients with fair accuracy. However, the values of the integral itself can be computed quite well. To check this we have computed the standard deviations, both for the Monte Carlo data, from

$$\sigma_1^2 = 1/M \sum_{j=1}^{M} [(N_j/N) - I_{11}(j/M)]^2$$

and for the least squares fit from

$$\sigma_2^2 = 1/M \sum_{j=1}^{M} [\bar{I}_{11}(j/M) - I_{11}(j/M)]^2$$

where

$$\bar{I}_{11}(a) = \sum_{j=3}^{6} A_j a^i$$

is the least-squares fit to I_{11}. The results are shown in Table 2.

As a compromise between accuracy and length of the Monte Carlo run, the values $N = 10000$ and $M = 20$ were selected. In this way there were computed the two integrals I_{31} and I_{32} corresponding to the two $(B,r,3)$-graphs, the six integrals $I_{41}, \ldots, I_{46}$ corresponding to the six $(B,r,4)$-graphs, and the 21 integrals $I_{51}, \ldots, I_{521}$ corresponding to the 21 $(B,r,5)$-graphs. The first two series are shown in Figs. 6 and 7. The programming was quite simple and no details need be given. The total time taken up on the CDC 6600 computer was about one hour; this, however, includes a lot of trial runs and tests.

TABLE II—STANDARD DEVIATIONS

N	Monte Carlo σ_1	Least Squares σ_2
1000	0.01154	0.00753
10000	0.00257	0.00183
100000	0.000922	0.000554

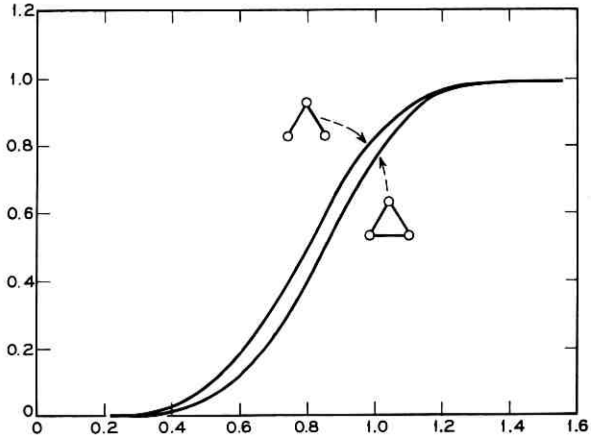


Fig. 6 — Cluster integrals for $v = 3$.

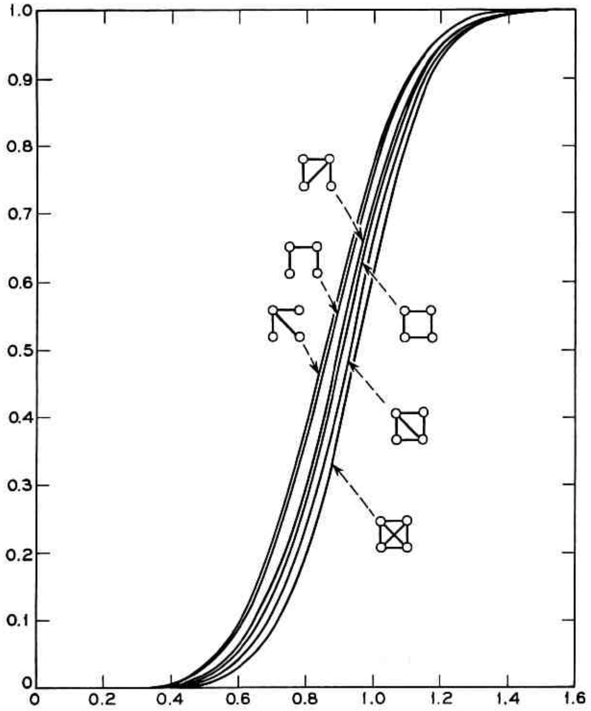


Fig. 7 — Cluster integrals for $v = 4$.

To sum up, it appears that numerical computation of J-type integrals is quite feasible, with the help of an automatic computer, to fairly good approximation. One well known advantage of the Monte Carlo method of evaluating multiple integrals was clearly brought out; namely, its relative independence of the dimension.

IX. ACKNOWLEDGEMENTS

Professors M. Bloom, L. Sobrino, and W. Opechowski, of the Physics Department, University of British Columbia, discussed with the author certain physical problems which led to the general formulation of Section 1. Dr. Max Goldstein, director of the A. E. C. Computation Laboratory at New York University, and his staff, helped with programming and with their CDC 6600 computing machine. Dr. Frank Stillinger of Bell Telephone Laboratories, Murray Hill, pointed out a serious error in the original use of the inclusion-exclusion principle, and suggested the correct version.

The work reported was started while the author was on leave from the University of British Columbia for the year 1966–67 at the Courant Institute of Mathematical Sciences, New York University. During that time he held a Senior Fellowship of the National Research Council. The work was continued while the author was visiting Bell Laboratories in the summer of 1967. I sincerely thank all these persons and institutions.

REFERENCES

1. Chandrasekhar, S., in *Noise and Stochastic Processes,* ed. Nelson Wax, New York: Dover, 1954, Chapter 1, pp. 1–89.
2. Uhlenbeck, G., and Ford, G. W., "Lectures in Statistical Mechanics," A. M. S. publishing, Providence, 1963.
3. Riordan, J., *Combinatorial Analysis,* New York: Wiley, 1958.
4. Melzak, Z. A., "Scattering from Random Arrays," Quart. Appl. Math *20,* No. 2 (1962), pp. 151–159.
5. Gilbert, E. N., and Pollak, H. O., "Coincidences in Poisson Patterns," B.S.T.J., *39,* No. 4 (July 1957), pp. 1005–1033.
6. Molina, E. C., "Theory of Probability," Trans. A.I.E.E., *44* (1929), pp. 294–299.

Data Transmission with FSK Permutation Modulation

By H. L. SCHNEIDER

(Manuscript received November 8, 1967)

Performance characteristics are derived for an FSK data transmission scheme in which M frequencies out of N are used simultaneously. Noncoherent matched filters are applied at the N frequencies, and the filter outputs are compared as in a permutation modulation system.

It is shown that many permutation alphabets provide energy per bit advantage over binary FSK, although the best results are obtained with one-out-of-N alphabets. Considering bits per unit bandwidth, many permutation alphabets perform as well as or better than binary; however, one-out-of-N alphabets carry less information per unit bandwidth when $N > 4$.

I. INTRODUCTION

The technique of N-ary frequency modulation in which energy is transmitted on 1 out of N frequencies to convey $\log_2 N$ bits of information per character has been known for some years.[1,2] David Slepian[3] has recently described a general modulation system, permutation modulation, which is applied here to a multifrequency modulation scheme in which energy is transmitted simultaneously on M frequencies out of N, thus conveying $\log_2 \binom{N}{M}$ bits of information per character. Binary and one out of N FSK modulation are special cases of FSK permutation modulation.

Such a transmission scheme is basically not new; it has been used for many years for transmitting decimal digits, address, and other supervisory information in the telephone plant. This work was motivated by a requirement to compare the information transmission capability of these alphabets. However, the application analyzed here is, in fact, different because we assume a baud synchronous matched filter receiver with a mutually orthogonal set of signals. The channel is assumed to be nonfading, frequency flat, with white gaussian additive noise.

II. GENERAL DESCRIPTION

For convenience, we shall refer to this modulation scheme as PFSK. The PFSK alphabet has $\binom{N}{M}$ characters,

$$\binom{N}{M} = \frac{N!}{M!\,(N-M)!}.$$

The $\binom{2}{1}$ alphabet is the binary FSK modulation with which other $\binom{N}{M}$ alphabets will be compared. The $\binom{N}{1}$ alphabet is commonly referred to as N-ary or MFSK (multiple frequency shift keyed).

PFSK transmission operates in a manner shown for the $\binom{N}{2}$ alphabet in Figure 1. One of the $\binom{N}{M}$ characters is input to the transmitter; the signal out is M simultaneous pulses of energy, one pulse on each of M distinct frequencies, lasting for T seconds. White gaussian noise is added in the channel. Filters, matched to the pulse shape, are tuned to each of the N possible frequencies. The filter outputs are envelope-detected and all N envelope samples are intercompared at the end of the pulse period. The largest M of these outputs determine the transmitted character.*

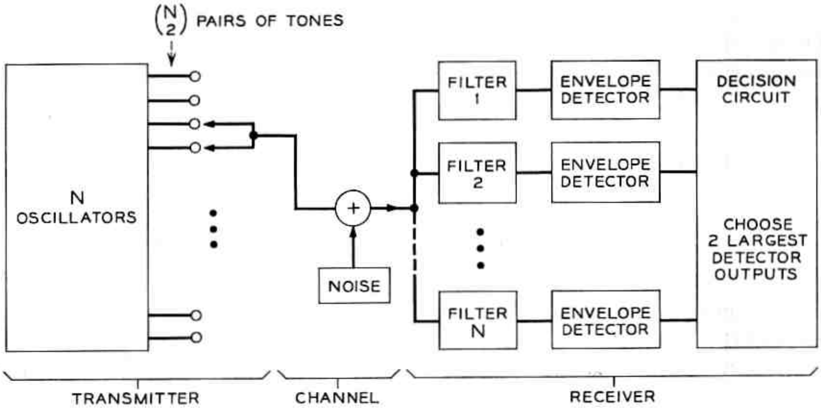


Fig. 1 — Transmission system for $\binom{N}{2}$ alphabets.

* Slepian has shown that this technique of amplitude comparisons minimizes the error probability. See Ref. 3.

III. ANALYSIS

An error is made in this process when any of the noise samples exceeds any of the signal plus noise samples. The error probability is P_e and is one minus the probability of making a correct decision:

$$P_e = 1 - \int_0^\infty p_M(s) P_{N-M}(s) \, ds \tag{1}$$

where

$p_M(s)$ is the p.d.f. of the smallest signal plus noise sample
$P_{N-M}(s)$ is the distribution function of the largest noise sample.

The p.d.f. of the smallest signal plus noise sample is determined as follows. The p.d.f. of the output sample of a matched filter detector can be written as*

$$p(y) = yI_0(y\sqrt{2R}) \exp\left(-\frac{y^2 + 2R}{2}\right) \tag{2}$$

where

y is the output envelope sample amplitude normalized to the rms noise
$I_0(\ )$ is the modified Bessel function
R is $\mathcal{E}_M/\mathfrak{N}_0$
$\mathcal{E}_M$ is the received signal energy in joules at each of the M transmitted frequencies
$\mathfrak{N}_0$ is the noise density, in watts per Hz.

The probability of the smallest of M samples exceeding a value s is the same as the probability that all M samples exceed the value s. This probability is expressed by equation (3), with independence of the M samples following from orthogonality.

$$1 - P_M(s) = \left[\int_s^\infty p(y) \, dy\right]^M = Q^M(\sqrt{2R}, s) \tag{3}$$

where

$P_M(s)$ is the distribution function of the smallest signal plus noise sample
$Q(\cdot, \cdot)$ is the Q function and is tabulated by Marcum[6].

* We view equation 2 as a renormalization of an expression by Helstrom[4] for matched filter detection, although it was originally derived by Rice[5] in a different context.

Thus, the p.d.f. of the smallest output is simply

$$p_M(s) = \frac{d}{ds} P_M(s) = Q^{M-1}(\sqrt{2R}, s)p(s). \tag{4}$$

Similarly, with independent noise samples, we find the distribution function of the largest noise sample.

$$\begin{aligned} P_{N-M}(s) &= \left[\int_0^s y \exp(-y^2/2)\, dy \right]^{N-M} \\ &= \sum_{r=0}^{N-M} (-1)^r \binom{N-M}{r} \exp\left(-\frac{rs^2}{2}\right) \end{aligned} \tag{5}$$

Substitution of equations (4) and (5) into equation (1) yields (after some labor) the character error probability:

$$\begin{aligned} P_e = M \sum_{r=1}^{N-M} (-1)^{r+1} \binom{N-M}{r} \int_0^\infty [Q(\sqrt{2R}, s)]^{M-1} \\ \times sI_0(s\sqrt{2R}) \exp\left[-\frac{(r+1)s^2 + 2R}{2} \right] ds \end{aligned} \tag{6}$$

A closed form expression for the case $M = 1$ was found by Reiger.[2]

$$P_e(M = 1) = \frac{1}{N} \sum_{r=2}^{N} (-1)^r \binom{N}{r} \exp\left[-R\left(1 - \frac{1}{r}\right) \right]. \tag{7}$$

A closed form expression for the case $M = 2$ is obtained from equation (6) using integration forms, having Q function integrands, given by Stein:[7]

$$\begin{aligned} P_e(M = 2) = \frac{2}{N-1} \sum_{r=2}^{N-1} (-1)^r \binom{N-1}{r} \left(\frac{1}{r+1} \right) \\ \times [1 + rQ(\alpha, \beta) - Q(\beta, \alpha)] \exp\left[-R\left(1 - \frac{1}{r}\right) \right] \end{aligned} \tag{8}$$

where

$$\alpha = \left(\frac{2Rr}{r+1} \right)^{\frac{1}{2}}, \qquad \beta = \frac{\alpha}{r}.$$

Closed form expressions for cases of $M > 2$ are not known; however, an asymptotic form for large R is obtained following arguments by Helstrom[8] for approximating the Q function:

$$P_e(M) \approx \frac{M}{N - M + 1} \sum_{r=2}^{N-M+1} (-1)^r \binom{N - M + 1}{r} \cdot \exp\left[-R\left(1 - \frac{1}{r}\right)\right] \tag{9}$$

or taking the predominant first term of equation (9) we have*

$$P_e(M) \approx \frac{M(N - M)}{2} \exp\left(-\frac{R}{2}\right). \tag{10}$$

Equation (10) can also be obtained heuristically. At high signal-to-noise ratios, character errors occur because of a binary decision error; that is, one of the noise samples is mistaken for one of the signal plus noise samples. The probability of a binary decision error is

$$P_e = 1/2 \exp(-R/2).$$

In the multifrequency situation, there are $M(N-M)$ ways for this to happen; the product of these two factors yields equation (10).

IV. COMPARISON OF ALPHABETS

We interrelate the performances of the PFSK alphabets to those of binary FSK using two criteria: energy per bit required for an equivalent error rate, and bits per unit bandwidth.† First, the per character information of these alphabets is defined as k:

$$k \equiv \log_2 \binom{N}{M}.$$

The normalized energy per bit $\mathcal{E}/\mathfrak{N}_0$ is related to the ratio R, defined in equation (2), by

$$R = \left(\frac{k}{M}\right)\left(\frac{\mathcal{E}}{\mathfrak{N}_0}\right). \tag{11}$$

Since the quantity R appears in the exponent of the error rate expression, it is apparent that, for low error rates, the power advantage (over binary FSK) of a PFSK alphabet approaches k/M. We can observe this numerically by comparing error rates on the basis

* It is easy to show that the first term is always an upper bound to P_e.

† The reader can compare the results of the work here with recent work of I. Jacobs,[9] who intercompares coherent modulation systems using virtually the same criteria.

of "equivalent error probability,"[10] which is the binary error probability for which the probability of one or more errors in a binary sequence of k bits is equal to the probability of error in the PFSK case. This equivalent error probability is defined as P_{eq}:

$$P_{eq} \equiv 1 - [1 - P_e(M)]^{1/k} \approx \frac{1}{k} P_e(M). \tag{12}$$

Figure 2 illustrates P_{eq} as a function of $\mathcal{E}/\mathfrak{N}_0$ for several alphabets. At error rates of 10^{-3}, the power advantage is within 0.6 dB of the k/M value for the $\binom{16}{1}$ and $\binom{16}{2}$ alphabets, and closer for the other examples.

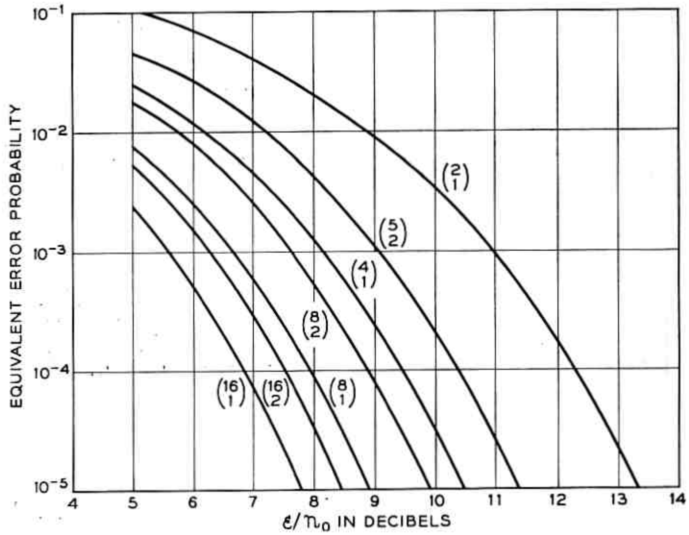


Fig. 2 — PFSK error probabilities.

The number of bits per unit bandwidth for a PFSK alphabet is determined by estimating the bandwidth as N times the frequency separation, which is $1/T$ for noncoherent orthogonal signals with minimum frequency spacing. Since the information rate is k/T, the desired bits per cycle ratio is simply k/N.* Figure 3 shows paired values of 10 $\log_{10}(k/M)$ and k/N for illustrative alphabets.

* It is easy to show that this ratio for PFSK alphabets approaches a maximum value of 1 for large N, with $M = N/2$. At this point $k/M = 2$ for a 3 dB advantage.

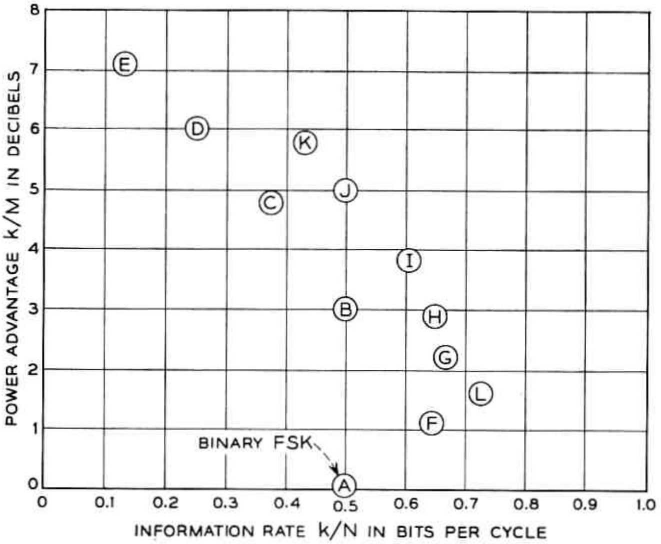


Fig. 3 — Performance comparison of PFSK alphabets.

Symbol	A	B	C	D	E	F	G	H	I	J	K	L
Alphabet	$\binom{2}{1}$	$\binom{4}{1}$	$\binom{8}{1}$	$\binom{16}{1}$	$\binom{32}{1}$	$\binom{4}{2}$	$\binom{5}{2}$	$\binom{6}{2}$	$\binom{8}{2}$	$\binom{12}{2}$	$\binom{16}{2}$	$\binom{6}{3}$

V. SUMMARY AND CONCLUSIONS

It has been shown that the PFSK technique gives significant power advantage over binary FSK. In addition, bandwidth can be controlled by the proper choice of alphabet.

Disadvantages of the technique are practical ones. Implementation of the decision function is relatively complicated. In some applications peak power limitations might make the average power calculations inapplicable.

A generally large number of characters in the alphabet is not suited to all applications, but can be very efficient in some. For example, the $\binom{5}{2}$ alphabet, containing 10 characters, is well suited to decimal digits.

ACKNOWLEDGEMENT

The author would like to thank Ira Jacobs for contributing many comments to this work.

REFERENCES

1. Jordan, D. B., et al., "Multiple Frequency Shift Teletype Systems," Proc. IRE, *43* (November 1955), pp. 1647–1655.
2. Rieger, S., "Error Rates in Data Transmission," Proc. IRE, *46* (May 1958), pp. 919–920.
3. Slepian. D., "Permutation Modulation," Proc. IEEE, *53* (March 1965), pp. 228–236.
4. Helstrom, C. W., *Statistical Theory of Signal Detection,* New York: Pergamon Press, Inc., 1960, p. 152.
5. Rice, S. O., "Mathematical Analysis of Random Noise," B.S.T.J., *23* (July 1944), Section 3.10, pp. 282–332.
6. Marcum, J. I., "Tables of Q Functions," Rand Corporation Report *RM-339,* AD-116551 (January 1950).
7. Stein, S., "Unified Analysis of Certain Coherent and Noncoherent Binary Communication Systems," IEEE Trans. Inform. Theory, *IT-10* (January 1964), pp. 43–51.
8. Helstrom, C. W., "The Resolution of Signals in White Gaussian Noise," Proc. IRE, *43* (September 1955), pp. 1111–1118.
9. Jacobs, I., "Comparison of *M*-ary Modulation Systems," B.S.T.J., *46* (May-June 1967), pp. 843–864.
10. Wolf, J. K., "On Comparing *N*-ary Systems," IRE Trans. Commun. Syst., *CS-10* (June 1962), pp. 216–217.

Contributors to This Issue

Richard H. Clarke, B.Sc. in EE, 1956; Ph.D. in EE, 1960, University College, London; Bell Telephone Laboratories, 1964—. Mr. Clarke has been concerned with radar studies of the troposphere, the statistical properties of microwave and optical propagation through the atmosphere, optical filters, and the analysis of mobile radio propagation. He is now on a leave of absence at Imperial College, London.

Corrado Dragone, Laurea in E.E., 1961, Padua University (Italy); Bell Telephone Laboratories, 1961—. Mr. Dragone has been engaged in experimental and theoretical work on microwave antennas and solid-state power sources. He is involved in solid-state radio system experiments.

W. Morven Gentleman, B.Sc., 1963, McGill University; M.A., 1964, Princeton University; Ph.D. (Mathematics), 1966, Princeton University; Bell Telephone Laboratories, 1965—. Mr. Gentleman has been doing research in numerical analysis and programming languages. Member: ACM, SIAM, IMS, Canadian Mathematical Congress.

William H. Ierley, B.A., 1959, Drew University; M.S., 1967, N.Y.U.; Bell Telephone Laboratories 1966—. As a resident visitor at Bell Laboratories from 1964 to 1966. Mr. Ierley worked on various phases of the Nike-X program. He has since been engaged in applications of computer technology to electromagnetic research, emphasizing computer graphics.

L. A. Meacham, B.S. in E.E., 1929, University of Washington; Certificate of Research, Cambridge University, 1930; Bell Telephone Laboratories 1930—. Mr. Meacham, now Head, Transmission Techniques Research Department, engaged in standard frequency research from 1930 to 1940. Then until 1953 he was concerned with pretransistor electronic switching, radar range units, pulse code modulation, circuitry for the newly-invented transistor, and underwater sound. From 1953 to 1964 he headed a development group to explore possible uses of transistors in telephone sets, and proceeded to develop "low-current sets with tone ringing" used in the Morris trial of electronic switching, *Touch-Tone*® dialing and the *Picturephone*® visual telephone. In 1964 he returned to research in new transmission techniques at low and moderate frequencies. Mr. Meacham, who has 68 patents, received the Eta Kappa Nu Recognition of Outstanding Young Electrical Engineers in 1939. Fellow, IEEE; member, Phi Beta Kappa, Sigma Xi, Eta Kappa Nu, Tau Beta Pi.

Z. A. MELZAK, B.Sc., 1950, and M.Sc., 1951, McGill University; Ph.D., 1956, Massachusetts Institute of Technology, all in mathematics; Bell Telephone Laboratories, 1960–61 and Summer 1967. Mr. Melzak has been engaged in research in applied analysis, combinatorics, and algorithmic complexity. He is a professor in the Department of Mathematics at the University of British Columbia. Member, Amer. Math. Soc. and Canadian Math. Congress.

INGEMAR NÅSELL, Civilingenjör, 1955, Royal Institute of Technology, Stockholm, Sweden; M.E.E., 1962, and M.S. (mathematics), 1965, both from New York University; Research Institute of National Defense, Stockholm, Sweden, 1955–1960; Bell Telephone Laboratories, 1960—. Mr. Nåsell is concerned with characterizing the transmission performance of the Bell System toll network for systems engineering purposes. He is Supervisor of the Transmission Surveys and Models Group. Member, Svenska Teknologföreningen, American Statistical Association, Eta Kappa Nu.

WILLIAM H. PIERCE, A.B., Harvard University, 1955; M.S., 1959, and Ph.D., 1961, both from Stanford University. In the summer of 1966, Mr. Pierce worked at Bell Telephone Laboratories on linear-real coding and pulse amplitude modulation. He is now Professor and Head of the Electrical Engineering Department at the University of Louisville. Member, IEEE.

RANDOLPH J. PILC, B.E.E., 1960, City College of New York; M.E.E., 1962, New York University; Ph.D., 1967, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1960—. Mr. Pilc has been concerned with problems in data transmission and communication theory and is engaged in analysis of data systems. Member, IEEE, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

VASANT K. PRABHU, B.E. (Dist.), 1962, Indian Institute of Science, Bangalore, India; S.M., 1963, Sc.D., 1966, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1966—. Mr. Prabhu has been concerned with various theoretical problems in solid-state microwave devices, noise, and optical communication systems. Member, IEEE, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, AAAS.

CLYDE L. RUTHROFF, B.S.E.E., 1950, and M.A., 1952, University of Nebraska; Bell Telephone Laboratories, 1952—. Mr. Ruthroff has published contributions on the subjects of FM distortion theory, broadband transformers, FM limiters, threshold extension by feedback, and microwave radio systems for satellite and terrestrial use. He is interested in the extension of radio communication into the millimeter and optical wavelengths. Member, A.A.A.S., I.E.E.E., Sigma Xi.

H. L. SCHNEIDER, B.S.E.E., 1949, Purdue University; M.S.E.E., 1955, Pennsylvania State University; Ph.D., 1961, Carnegie Institute of Technology; Bell Telephone Laboratories, 1961—. He had been with the Mobile Radio Research Department and is now in the Military Transmission Systems Department. Senior member, IEEE; member, Sigma Xi.

WILLIAM D. WARTERS, A.B., 1949, Harvard College; M.S., 1950 and Ph.D., 1953, both from California Institute of Technology; Bell Telephone Laboratories, 1953—. Mr. Warters has done research in millimeter waveguide transmission and worked on repeaters for millimeter waveguide systems. He is Director of the Transmission Systems Research Center. Senior member, IEEE; member, American Physical Society, Sigma Xi, Phi Beta Kappa.

D. T. YOUNG, B.S., 1956, MEE, 1960, Ph.D., 1966, University of Oklahoma; Bell Telephone Laboratories 1960–1965, Asst. Prof., Kansas State University, 1966–1967; Bell Telephone Laboratories, 1967—. He initially worked on mode conversion problems in multimode waveguide and a solid state repeater for a waveguide transmission system. At present he is working on optical transmission systems. Member IEEE, Tau Beta Pi, Eta Kappa Nu and Sigma Xi.

H. Zucker, Dipl.-Ing. 1950, Technische Hochschule, Munich, Germany; M.S.E.E., 1954, Ph.D., 1959, Illinois Institute of Technology; Bell Telephone Laboratories, 1964—. He has been engaged in the analysis and design of satellite communication antennas and is now concerned with optical resonators. Member, IEEE, Eta Kappa Nu, Sigma Xi.

T.R
5861-10-2