THE BELL SYSTEM

Technical Journal

Volume 49 January 1970 Number 1

Analytical Approximations to Approximations in the Chebyshev Sense — **S. Darlington** 1

Multilevel Modulation Techniques for Millimeter Guided Waves — **W. M. Hubbard, G. D. Mandeville and J. E. Goell** 33

Crosstalk in Multiple-Beam Waveguides — **D. Gloge** 55

Asymptotic Analysis of a Nonlinear Autonomous Vibratory System — **J. A. Morrison** 73

The Capacity of Linear Channels with Additive Gaussian Noise — **R. K. Mueller and G. J. Foschini** 81

Theorems on the Analysis of Nonlinear Transistor Networks — **I. W. Sandberg** 95

A Charge Control Relation for Bipolar Transistors — **H. K. Gummel** 115

Rain Attenuation and Radio Path Design — **C. L. Ruthroff** 121

An Improved Thermal Gas Lens for Optical Beam Waveguides — **P. Kaiser** 137

Contributors to This Issue 155

COPYRIGHT © 1970 AMERICAN TELEPHONE AND TELEGRAPH COMPANY

THE BELL SYSTEM TECHNICAL JOURNAL

ADVISORY BOARD

H. G. MEHLHOUSE, *President, Western Electric Company*

J. B. FISK, *President, Bell Telephone Laboratories*

A. S. ALSTON, *Executive Vice President,*
American Telephone and Telegraph Company

EDITORIAL COMMITTEE

W. E. DANIELSON, *Chairman*

F. T. ANDREWS, JR.
E. E. DAVID
W. O. FLECKENSTEIN
W. S. HAYWARD, JR.
C. W. HOOVER, JR.
A. E. JOEL
B. E. STRASSER
M. TANENBAUM
D. G. THOMAS
C. R. WILLIAMSON

EDITORIAL STAFF

G. E. SCHINDLER, JR., *Editor*

W. V. RUCH, *Assistant Editor*

H. M. PURVIANCE, *Production and Illustrations*

F. J. SCHWETJE, *Circulation*

THE BELL SYSTEM TECHNICAL JOURNAL is published ten times a year by the American Telephone and Telegraph Company, B. S. Gilmer, President, C. E. Wampler, Vice President and Secretary, J. J. Scanlon, Vice President and Treasurer. Checks for subscriptions should be made payable to American Telephone and Telegraph Company and should be addressed to the Treasury Department, Room 2312C, 195 Broadway, New York, N. Y. 10007. Subscriptions $7.00 per year; single copies $1.25 each. Foreign postage $1.00 per year; 15 cents per copy. Printed in U.S.A.

THE BELL SYSTEM
TECHNICAL JOURNAL
DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 49 January 1970 Number 1

Copyright © 1970, American Telephone and Telegraph Company

Analytical Approximations to Approximations in the Chebyshev Sense

By SIDNEY DARLINGTON

(Manuscript received February 20, 1969)

This paper concerns approximation in the Chebyshev, or minimax sense such that (i) a minimax approximation implies a maximum number of zero error points separated by equal error extrema, and (ii) the approximating function can be so formulated that the disposable parameters are all the coefficients in a polynomial, which may however be part of a more complicated function the rest of which is prescribed. Weighted minimax polynomial approximations can be included, by multiplying the approximated and approximating functions by the weight factor. Analytic methods are described which yield approximately equal error extrema. They are sufficiently simple so that they may sometimes compete with currently used iterative numerical methods, especially when the degree of the disposable polynomial is large. Their most probable utility concerns explorations of available accuracies over wide ranges of design parameters such as degree of disposable polynomial, interval of approximation, and coefficients in prescribed parts of the approximating function.

I. INTRODUCTION

This paper concerns approximation in the Chebyshev sense, over a prescribed interval $x_a \leqq x \leqq x_b$ of a continuous real variable x. As defined, an approximation in the Chebyshev sense is a minimax approximation—one in which the maximum error is as small as is possible within given constraints on the approximating function. Minimax approximations in which errors are weighted by a prescribed function of the independent variable can also be treated as Chebyshev approximations, by multiplying the approximated and approximating functions by the weight function.

Frequently, but not always, approximation in the Chebyshev sense implies an error of the "equal ripple" sort illustrated in Fig. 1—that is, a sequence of equal positive and negative extrema with monotonic variations in between. General necessary and sufficient conditions for this are not known. However, the following conditions are sufficient: the p disposable parameters of the approximating function are to be such that the approximation error can be made zero at p arbitrary points within the approximating interval. Referring to Fig. 1, the arbitrary error points divide the approximation interval into $p + 1$ segments. There is to be a particular division such that the error function achieves its maximum magnitude $p + 1$ times—at the two edges of the approximation interval and once within each of the $p - 1$ interior segments. There are to be no other local extrema within the approximation interval. Generally, shrinking any one of the $p + 1$ segments (by bringing two zero error points closer together or one closer to an edge of the approximation interval) tends to reduce the corresponding error extremum. Conditions are to be such that all the $p + 1$ equal extrema can be reduced simultaneously only by shrinking all the $p + 1$ segments, which is impossible without shrinking the given approximation interval. These conditions are encountered in many practical problems and are assumed here. Thus we are concerned only with equal ripple approximations like Fig. 1.

Exactly equal ripple approximations have long been known for a very few special cases (which have been useful for example in filter design). Iterative numerical methods have been developed for the solution of various more general problems and are described in text-

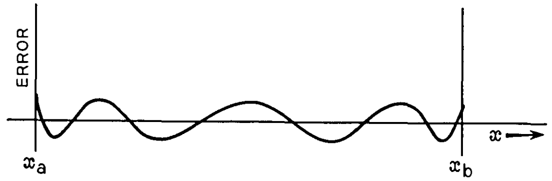


Fig. 1 — An equal ripple error function (p zeros; $p + 1$ segments; $p + 1$ extrema).

books such as Ref. 1. In contrast, this paper describes analytical procedures which yield error extrema of approximately equal amplitude. Their full range of validity has not been determined. However, they are clearly appropriate for a substantial, although poorly defined class of problems. It is characterized further later.

Useful applications are likely to concern equal ripple problems which have not been solved exactly by analysis and which involve so many disposable parameters that iterative numerical solutions are likely to be more costly. The most useful applications probably concern preliminary explorations over primary design parameters (such as intervals of approximation, magnitudes of errors, and degrees of approximating functions) before numerical refinement of specific designs. Accordingly, this paper emphasizes relatively simple means for approximating equal ripples and says little about more complicated higher order approximations.

The procedures apply only to approximating functions characterized as follows. The disposable parameters must be all the coefficients in a polynomial (which may have been obtained, however, by some sort of transformation on the original independent variable and/or the approximating function). This is referred to as the disposable polynomial. On the other hand, the disposable polynomial may be only a part of a more general approximating function the rest of which is prescribed in advance (for example, the numerator of a rational fraction with a prescribed denominator). Weighted as well as unweighted minimax approximations are included. For some problems, closed form formulas are obtained for approximate error size as functions of the degree of the disposable polynomial, usable for degrees of any size. For other problems, the error size is related to an eigenvalue of a certain matrix equation, but the order of the matrix may be small even though the degree of the disposable polynomial is arbitrarily large.

A primary concern here is the distinction between simple truncation of infinite series of Chebyshev polynomials and approximation in the Chebyshev or minimax sense. The functions which we are to approximate can be expanded into infinite series of Chebyshev polynomials. Approximations with polynomials of degree n can be obtained by simply truncating the infinite series after the terms of degree n. However, simple truncation does not usually give an approximation of the minimax sort. A polynomial of degree n which approximates the given function in the minimax manner can be represented as a linear combination of Chebyshev polynomials, but the coefficients are usually different from those in the truncated infinite series.

One way to approach approximation in the Chebyshev sense is to start with the truncated series of Chebyshev polynomials. Then corrections to the coefficients are determined, to obtain equal ripple error functions. Such a procedure has been used before, for example in Refs. 2 through 4, and is used here. Departures from the previous work known to the author include simple approximations to ideal solutions formulated for more general approximating functions and for weighted as well as unweighted minimax approximations, as opposed to more rigorous analyses of more restricted problems.

Sometimes truncation of an infinite series of Chebyshev polynomials yields an approximately equal ripple error function without further adjustment of the coefficients. The procedures for adjusting the coefficients, described herein, sometimes also give an initial insight into whether or not adjustments are needed.

It is interesting to note that some 35 years ago a conference was held in the office of T. C. Fry, at Bell Telephone Laboratories, to consider some filter patents offered for sale by W. Cauer. One of the patents disclosed Cauer's equal ripple image impedance and transfer functions, which soon became famous among circuit theorists, but did not include proofs or derivations. At the conference, S. A. Schelkunoff asserted a very simple principle which enabled him to confirm and interpret Cauer's formulas. However, it did not explain how Cauer might have derived or discovered the formulas. The principle applies also to more general equal ripple approximations. It does not, by itself, solve the approximation problem, but it does furnish a starting point from which to develop procedures which do. We call it Schelkunoff's principle.

Section II describes Schelkunoff's principle. Section III solves two problems for which exactly equal ripple solutions are easily found. Section IV develops general procedures, whereby approximate solutions can be obtained for a large class of problems. Section V further clarifies the general procedures by means of examples.

Various aspects of the procedures described here bear some relation to other work. Section VI notes some of these relationships. Finally, Section VII reviews and summarizes the general conclusions, including a comment on the possibility of generalizations to disposable rational fractions.

II. SCHELKUNOFF'S PRINCIPLE

Consider first a function $T_n(x)$ proportional to a Chebyshev polynomial, defined by

$$\mathrm{T}_n(x) = \frac{K_n}{2^{n-1}} \cos (n \cos^{-1} x). \tag{1}$$

It is illustrated in Fig. 2a, for $n = 4$. It may be regarded as an "equal ripple" approximation to zero, over the interval $-1 \leqq x \leqq +1$, by a polynomial of degree n in which the coefficient of x^n is required to be K_n. Let

$$x = \cos \varphi. \tag{2}$$

Substitution in equation (1) gives

$$\mathrm{T}_n(x) = T_n(\varphi) = \frac{K_n}{2^{n-1}} \cos n\varphi. \tag{3}$$

The new function is illustrated in Fig. 2b, again for $n = 4$. Note that x is periodic in φ with period 2π and $T_n(\varphi)$ is periodic in φ with period $2\pi/n$. Thus there are n periods of $T_n(\varphi)$ in each period of x.

Stated with a little more detail, we have this situation: The original function $\mathrm{T}_n(x)$ has "equal ripples" in the sense of equal extrema. However, the extrema are not uniformly spaced and hence the ripples differ as to width. The periodic transformation from x to φ has two important properties. As φ increases, x sweeps back and forth across the approximation interval, $-1 \leqq x \leqq +1$. In each interval in which x varies monotonically from ± 1 to ∓ 1 the φ scale is a distortion of the x scale such that the ripples of $T_n(\varphi)$ are uniformly spaced and are

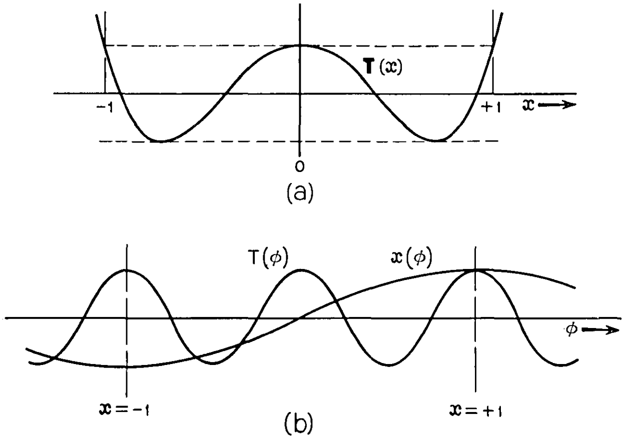


Fig. 2 — Illustrating Schelkunoff's principle.

also otherwise identical. The ripples could not be made identical by any distortion of the x scale alone if $\mathbf{T}_n(x)$ had *unequal* maxima or *unequal* minima. This is a special case of Schelkunoff's principle.

More generally, let $\mathbf{E}(x)$ be a function of x with the following properties over an interval $x_a \leqq x \leqq x_b$: The function $\mathbf{E}(x)$ is real and single valued; there are a number of local maxima, all equal; there are a number of local minima, all equal; the equal extrema for the interval include the end points $\mathbf{E}(x_a)$, $\mathbf{E}(x_b)$ (at which $d\mathbf{E}/dx$ need not $=0$).* Then Schelkunoff's principle asserts the existence of a transformation

$$x = \Gamma(\varphi) \tag{4}$$

with the following properties: The original variable x is periodic in the new variable φ; as φ increases x sweeps back and forth over the given interval $x_a \leqq x \leqq x_b$, monotonically each way once each period; the periodic function

$$E(\varphi) = \mathbf{E}[\Gamma(\varphi)] \tag{5}$$

has a number of periods in each period of x, equal to one less than the number of extrema of $\mathbf{E}(x)$ in the given interval of x (including the end points). In applications to approximation in the Chebyshev sense, $\mathbf{E}(x)$ and $E(\varphi)$ represent the equal ripple error, as functions of x and φ.

The transformation $x = \Gamma(\varphi)$ clearly is not unique, for there are obvious transformations on φ itself which retain the desired character of $E(\varphi)$. For example, φ can be replaced by $\varphi + q(\varphi)$, where $q(\varphi)$ is periodic with the same period as $E(\varphi)$ and is such that $\varphi + q(\varphi)$ is monotonic in φ. When $\mathbf{E}(x)$ is continous (in the given interval of x), a particular $\varphi + q(\varphi)$ will make $E(\varphi)$ sinusoidal.

We do not attempt a very general, rigorous proof of Schelkunoff's principle; we merely use it as a guide to a strategy for solving minimax problems. However, a demonstration of the principle for a specific class of problems will be implicit in what follows, for we shall find transformations which do in fact change our equal ripple errors into sinusoidal errors.

In later sections we will again use the transformation (2), or a generalization for end points other than $x = \pm 1$. Usually, however, it will not be a Schelkunoff transformation. We will use it to transform the disposable polynomial in x into a finite Fourier series in φ. The coefficients of the Fourier series are to be chosen in such a way that the overall approximation is approximately sinusoidal on a *distortion* of the φ scale. Similar strategies have been used before, for example in Refs. 2 through 4.

---

* Problems can be found such that minimax approximations have equal extrema which do not include both end points. Then the number of local extrema for the approximation interval is abnormally large when the end points are counted. Such problems are not considered here.

Means for determining the distortion of the φ scale and the adjustment of the Fourier coefficients are introduced by means of two examples in the next section.

III. TWO GENERALIZATIONS OF CHEBYSHEV POLYNOMIALS

The two problems described below are solved exactly. The form of the solutions suggests approximate solutions to more general problems.

3.1 *A Rational Function Generalization of Chebyshev Polynomials*

Consider the following generalization of Chebyshev polynomials: Let

$$\mathbf{T}_{Dn}(x) = \frac{\mathbf{P}(x)}{\mathbf{D}(x)} \tag{6}$$

in which $\mathbf{P}(x)$ is a polynomial of degree n and $\mathbf{D}(x)$ is a polynomial of degree $\leqq n$. Suppose $\mathbf{D}(x)$ is prescribed in advance and $\mathbf{P}(x)$ is to be chosen in such a way that $\mathbf{T}_{Dn}(x)$ has equal ripples like those of a Chebyshev polynomial in the interval $-1 \leqq x \leqq +1$. More specifically, require that $\mathbf{T}_{Dn}(x) = \pm 1$ at $n - 1$ local extrema within the interval $-1 \leqq x \leqq +1$ and at the end points $x = \pm 1$. Real zeros of $\mathbf{D}(x)$ are to be excluded from the interval $-1 \leqq x \leqq +1$. The conditions on the extrema insure that all n of the zeros of $\mathbf{P}(x)$ will be in the interval.

Let φ be defined again by equation (2) and note that the real axis in the φ plane corresponds to the real interval $-1 \leqq x \leqq +1$ in the x plane. If $\cos \varphi$ is written in terms of exponentials, the polynomial $\mathbf{P}(x)$ can be related to φ by

$$\mathbf{P}(x) = P(e^{i\varphi}) + P(e^{-i\varphi}) \tag{7}$$

in which $P(\cdot)$ is a polynomial of the same degree, n, as $\mathbf{P}(\cdot)$. Given the coefficients of $P(\cdot)$ it is a simple matter to compute the coefficients of $\mathbf{P}(\cdot)$. We shall consider our problem solved when we have found the coefficients of $P(\cdot)$ required for our equal ripple conditions.

It is convenient to relate the prescribed denominator $\mathbf{D}(x)$ to φ in the following slightly different way:

$$\mathbf{D}(x) = D(e^{i\varphi})\, D(e^{-i\varphi})$$

in which $D(\cdot)$ is a polynomial of the same degree, $\leqq n$, as $\mathbf{D}(\cdot)$. If $\mathbf{D}(\cdot)$ and $D(\cdot)$ are written in factored form there is a one to one correspondence between factors. Thus if $(x_\sigma - x)$ is a factor of $\mathbf{D}(x)$ and $(1 - \gamma_\sigma e^{i\varphi})$ a corresponding factor of $D(e^{i\varphi})$,

$$x_\sigma - x = M_\sigma(1 - \gamma_\sigma e^{i\varphi})(1 - \gamma_\sigma e^{-i\varphi}) \tag{9}$$

in which M_σ is a constant scale factor.

By equation (2), $e^{i\varphi} = \pm 1$ at $x = \pm 1$, and hence

$$x_\sigma \pm 1 = M_\sigma(1 \pm \gamma_\sigma)^2. \tag{10}$$

Given x_σ, two solutions for γ_σ are easily obtained, for which $|\gamma_\sigma|$ is respectively < 1 and > 1. (Exclusion of zeros x_σ of $\mathbf{D}(x)$ from the real interval $-1 \leqq x \leqq +1$ removes the possibility of $|\gamma_\sigma| = 1$.) We need the solution for which $|\gamma_\sigma| < 1$, for reasons which will soon be apparent. From equation (11), with the sign of the square root chosen for $|\gamma_\sigma| < 1$,

$$\frac{1 - \gamma_\sigma}{1 + \gamma_\sigma} = \left[\frac{x_\sigma - 1}{x_\sigma + 1}\right]^{\frac{1}{2}}, \qquad \operatorname{Re}\left[\frac{x_\sigma - 1}{x_\sigma + 1}\right]^{\frac{1}{2}} > 0. \tag{11}$$

The scale factor M_σ need not concern us at this time.

The function $\mathbf{T}_{Dn}(x)$ can now be mapped into a function $T_{Dn}(\varphi)$ in terms of equations (7) and (8).

$$\mathbf{T}_{Dn}(x) = T_{Dn}(\varphi) = \frac{P(e^{i\varphi}) + P(e^{-i\varphi})}{D(e^{i\varphi})\,D(e^{-i\varphi})}, \tag{12}$$

By Schelkunoff's principle, our requirements on the extrema of $\mathbf{T}_{Dn}(x)$ imply that $T_{Dn}(\varphi)$ has the following special form

$$\begin{aligned} T_{Dn}(\varphi) &= \frac{P(e^{i\varphi}) + P(e^{-i\varphi})}{D(e^{i\varphi})\,D(e^{-i\varphi})}, \\ &= \tfrac{1}{2}e^{i[n\varphi + f(\varphi)]} + \tfrac{1}{2}e^{-i[n\varphi + f(\varphi)]}. \end{aligned} \tag{13}$$

The variable $\varphi + (1/n)f(\varphi)$ is a distortion of the φ scale for Schelkunoff's principle, for which $f(\varphi)$ is to be periodic in φ with period 2π and $\varphi + (1/n)f(\varphi)$ is to vary monotonically with φ.

Given $f(\varphi)$ one can easily find $P(e^{i\varphi})$ by equation (13). The problem is to find an $f(\varphi)$ for which $P(\cdot)$ is a polynomial of degree n. From equation (13)

$$P(e^{i\varphi}) + P(e^{-i\varphi}) = \left\{ \begin{aligned} &\tfrac{1}{2}e^{i[n\varphi + f(\varphi)]}\,D_n(e^{i\varphi})\,D_n(e^{-i\varphi}) \\ &+ \tfrac{1}{2}e^{-i[n\varphi + f(\varphi)]}\,D_n(e^{i\varphi})\,D_n(e^{-i\varphi}) \end{aligned} \right\}. \tag{14}$$

If $P(e^{i\varphi})$ is to have no terms in $e^{i\sigma\varphi}$ with $\sigma > n$, $e^{if(\varphi)}$ needs to cancel out $D_n(e^{i\varphi})$. This suggests

$$e^{if(\varphi)} = \frac{D(e^{-i\varphi})}{D(e^{i\varphi})} \tag{15}$$

(and note that this does make $f(\varphi)$ real when φ is real). Substitution in equation (14) gives

$$P(e^{i\varphi}) + P(e^{-i\varphi}) = \tfrac{1}{2}e^{in\varphi}\, D^2(e^{-i\varphi}) + \tfrac{1}{2}e^{-in\varphi}\, D^2(e^{i\varphi}). \tag{16}$$

Expanding the right side gives a polynomial in $e^{i\varphi}$. When polynomial $D(\cdot)$ is of degree $\leqq n$ (as assumed) there will be no powers of $e^{i\varphi}$ outside the range $-n$ to $+n$. Collecting positive powers (and half the constant term) gives $P(e^{i\varphi})$.

The function $f(\varphi)$ determined by equation (15) is periodic in φ with period 2π provided the zeros of $D(\lambda)$ lie outside the unit circle in the λ plane, which is assumed by equation (11). It is easily shown that the same condition makes $\varphi + (1/n)f(\varphi)$ monotone in φ. Once $P(\cdot)$ is known it is a simple matter to find $\mathbf{P}(x)$ by means of equations (2) and (7). It is probably simplest to omit the scale factor M_σ of equation (10) in the initial formulation. This does not affect the ratio in equation (15), but only the scale factor of the polynomial $\mathbf{P}(x)$, which can be corrected later on [for example to meet the condition $\mathbf{T}_{Dn}(1) = 1$].

Obvious generalizations of the problem include the following: For extrema $A \pm J$ (instead 0 ± 1) use

$$\mathbf{F}(x) = A \pm J\mathbf{T}_{Dn}(x). \tag{17}$$

In a more general interval of x, say $x_a \leqq x \leqq x_b$, replace equation (2) by

$$x = \frac{x_b + x_a}{2} + \frac{x_b - x_a}{2}\cos\varphi \tag{18}$$

and change equation (10) to

$$x_\sigma - x_a = M_\sigma(1 - \gamma_\sigma)^2, \qquad x_\sigma - x_b = M_\sigma(1 + \gamma_\sigma)^2. \tag{19}$$

The function $\mathbf{F}(x)$ defined by equation (17) has long been used by filter theorists, but previous derivations have been quite different.[5] The form of equation (16) suggests a similar solution to the problem described below.

3.2 *An Irrational Generalization of Chebyshev Polynomials*

Now let

$$\mathbf{T}_{Sn}(x) = \frac{\mathbf{P}(x)}{[\mathbf{S}(x)]^{\frac{1}{2}}} \tag{20}$$

in which $\mathbf{P}(x)$ is again a polynomial of degree n but $\mathrm{S}(x)$ is a polynomial of degree $\leqq 2n$. Suppose $\mathrm{S}(x)$ is prescribed and that $\mathbf{T}_{Sn}(x)$ is to meet the same conditions as to extrema as $\mathbf{T}_{Dn}(x)$ in the previous subsection. In place of equation (8), we can now use

$$\mathrm{S}(x) = S(e^{i\varphi})S(e^{-i\varphi})$$

to determine a polynomial $S(\cdot)$. We can then replace equation (15) by

$$e^{if(\varphi)} = \left[\frac{S(e^{-i\varphi})}{S(e^{i\varphi})}\right]^{\frac{1}{2}} \tag{21}$$

and then equation (16) by

$$P(e^{i\varphi}) + P(e^{-i\varphi}) = \tfrac{1}{2}e^{in\varphi}S(e^{-i\varphi}) + \tfrac{1}{2}e^{-in\varphi}S(e^{i\varphi}). \tag{22}$$

This makes $P(\cdot)$ again a polynomial of degree n. Note that $\mathbf{T}_{Sn}(x)$ cannot be used in place of $\mathbf{T}_{Dn}$ in equation (17), with $A \neq 0$, without changing the polynomial character of the numerator.

IV. GENERAL FORMULATIONS

This section shows how a large class of minimax approximations can be approximated by generalizing the manipulations described above. In Section V, we clarify the general procedures further by providing examples.

4.1 *Unweighted Minimax Approximations*

Let

$$\mathbf{P}(x) = \mathbf{F}(x) + \varepsilon(x), \qquad x_a \leqq x \leqq x_b \tag{23}$$

in which $\mathbf{P}(x)$ is a disposable polynomial of degree n, $\mathbf{F}(x)$ is a given function to be approximated by $\mathbf{P}(x)$ in the interval $x_a \leqq x \leqq x_b$, and $\varepsilon(x)$ is the error in the approximation. For what $\mathbf{P}(x)$ is $\varepsilon(x)$ smallest in the minimax sense? We assume that the minimax $\varepsilon(x)$ has the equal ripple form (Fig. 1) and we seek only approximations to equal ripples. We also restrict the class of applicable functions by certain further assumptions which can best be introduced a little later.

As before, let x and φ be related by equation (18), so that $x_a \leqq x \leqq x_b$ maps into real φ, and replace $\mathbf{P}(x)$ by

$$\mathbf{P}(x) = P(e^{i\varphi}) + P(e^{-i\varphi}). \tag{24}$$

If $P(\cdot)$ is again a polynomial of degree n, it is uniquely determined by $\mathbf{P}(\cdot)$ [and x_a, x_b in equation (18)]. Now, however, we find it expedient to permit $P(z)$ to include negative powers of z, up to z^{-n}. Thus, in equation (24),

$$P(e^{i\varphi}) = \sum_{\sigma=-n}^{n} P_\sigma e^{i\sigma\varphi}. \tag{25}$$

This $P(\cdot)$ is not uniquely determined by $\mathbf{P}(\cdot)$. However, $\mathbf{P}(\cdot)$ *is* uniquely determined by $P(\cdot)$, and it is still a polynomial of degree n. We solve the approximation problem by finding a suitable $P(\cdot)$, from which $\mathbf{P}(\cdot)$ can be easily determined.

We require that the mapping from x to φ maps $\mathbf{F}(x)$ into a function of φ with a convergent Fourier series. This amounts to requiring that $\mathbf{F}(x)$ can be expanded into a convergent series of Chebyshev polynomials (defined to fit the given interval of approximation). Because equation (18) is even in φ, the Fourier series has only cosine terms. Then, replacing cosines by sums of exponentials,

$$\begin{aligned} \mathbf{F}(x) &= F(e^{i\varphi}) + F(e^{-i\varphi}) \\ F(e^{i\varphi}) &= \sum_{\sigma=0}^{\infty} C_\sigma e^{i\sigma\varphi} \end{aligned} \tag{26}$$

in which the series expansion of $F(e^{i\varphi})$ converges when φ is real.

The desired equal ripple error can be written

$$\varepsilon(x) = \epsilon \cos [(n + 1)\varphi + f(\varphi)] \tag{27}$$

in which $f(\varphi)$ is again periodic in φ and represents the distortion of the φ scale per Schelkunoff's principle. In an equivalent exponential form

$$\begin{aligned} \varepsilon(x) &= E(e^{i\varphi}) + E(e^{-i\varphi}) \\ E(e^{i\varphi}) &= \frac{\epsilon}{2} e^{i[(n+1)\varphi + f(\varphi)]}. \end{aligned} \tag{28}$$

The exponent $i(n + 1)\varphi$, instead of $in\varphi$ as in the previous section, reflects the following circumstances: If the Chebyshev polynomial series corresponding to equation (26) is truncated after the polynomial of degree n, the first omitted polynomial is of degree $n + 1$. If all the other omitted polynomials have sufficiently small coefficients, the truncation error will approximate $\mathbf{E}(x)$ of equation (23) with $f(\varphi) = 0$. Note also that a disposable polynomial of degree n has $n + 1$ disposable coefficients. These are an example of the p disposable parameters in the more general description of equal ripple errors in Section I.

Using equations (24), (26) and (28) in equation (23) gives

$$P(e^{i\varphi}) + P(e^{-i\varphi}) = F(e^{i\varphi}) + F(e^{-i\varphi}) + E(e^{i\varphi}) + E(e^{-i\varphi}). \tag{29}$$

We arbitrarily equate the terms in exp $(+i\varphi)$ separately, so that

$$P(e^{i\varphi}) = F(e^{i\varphi}) + E(e^{i\varphi}). \tag{30}$$

If equation (30) is satisfied at all real φ so is the corresponding equation in exp $(-i\varphi)$. Thus a solution of equation (30) is a solution of equation (29). But the converse is not necessarily true. Frequently, an exactly equal ripple approximation corresponds to a solution of equation (29) which is not a solution of equation (30). However, we will find that approximations with approximately equal ripples can frequently be derived from equation (30), and in a much simpler way.

In equation (30), expand $P(\cdot)$, $F(\cdot)$, $E(\cdot)$ per equations (24), (26) and (28). The result can be rearranged as follows:

$$\begin{aligned} \sum_{\lambda=1}^{2n+1} G_\lambda e^{-i\lambda\varphi} &= \frac{\epsilon}{2} e^{if(\varphi)} + \sum_{\lambda=0}^{\infty} C_{n+1+\lambda} e^{i\lambda\varphi}; \\ G_\lambda &= P_{n+1-\lambda} - C_{n+1-\lambda}\,, \qquad \lambda \leqq n+1\ \lambda \leqq n+1; \\ &= P_{n+1-\lambda}\,, \qquad n+1 < \lambda \leqq 2n+1. \end{aligned} \tag{31}$$

In this equation, $C_{n+1+\lambda}$ is fixed by equation (26) but $P_{n+1-\lambda}$ is a disposable parameter in equation (25). Thus we seek an ϵ and exp $[if(\varphi)]$ with the following properties: First, $(\epsilon/2)$ exp $[if(\varphi)]$ is to be expandable in terms of positive and negative powers of exp $(i\varphi)$. Second, the coefficients of positive powers are to cancel the corresponding coefficients $C_{n+1+\lambda}$ in equation (31). Third, the coefficients of negative powers are to be such that, with an appropriate ϵ, $|\exp [\text{if } (\varphi)]| = 1$ when φ is real, so that $f(\varphi)$ is real and the error extrema are equal per equation (27). Sometimes it turns out that there are no negative powers beyond $-(2n + 1)$. Then the left side of equation (31) can be adjusted to match the right side. In many other problems, approximately equal error extrema can be obtained by simply ignoring terms in negative powers beyond $-(2n + 1)$.

Now consider the class of functions $F(\cdot)$ such that, in equation (31)

$$\sum_{\lambda=0}^{\infty} C_{n+1+\lambda} e^{i\lambda\varphi} = \frac{B(e^{i\varphi})}{A(e^{i\varphi})} \tag{32}$$

in which $A(\cdot)$ is a polynomial of degree m and $B(\cdot)$ is a polynomial of degree μ. If the series converges, as assumed, the zeros of $A(z)$ will lie outside the unit circle.

Under conditions which we shall examine further, the appropriate exp $[if(\varphi)]$ is now as follows:

$$e^{if(\varphi)} = \frac{A(e^{-i\varphi})X(e^{i\varphi})}{A(e^{i\varphi})X(e^{-i\varphi})} + \delta(e^{i\varphi}) \tag{33}$$

in which δ is small (at real φ) and $X(\cdot)$ is a polynomial determined by two further conditions. First, the zeros of $X(z)$ must lie outside the unit circle. Second ϵ and $X(\cdot)$ must be such that

$$\frac{\epsilon}{2}\frac{A(e^{-i\varphi})X(e^{i\varphi})}{A(e^{i\varphi})X(e^{-i\varphi})} + \frac{B(e^{i\varphi})}{A(e^{i\varphi})} = \frac{e^{-i\varphi}N(e^{-i\varphi})}{X(e^{-i\varphi})} \tag{34}$$

in which $N(\cdot)$ is a polynomial. Let us examine the implications first and the existence of such an ϵ and $X(\cdot)$ thereafter.

When φ is real exp $(i\varphi)$ and exp $(-i\varphi)$ are conjugates, and so are identical polynomials in these two variables. This makes $f(\varphi)$ real in equation (33), except for small corrections due to δ. When the zeros of $A(z)$ and $X(z)$ lie outside the unit circle, as required, the unit circle in the z plane maps into contours in the polynomial planes which do not enclose 0. This makes $f(\varphi)$ periodic in φ.

The condition on the zeros of $X(z)$ also permits the right side of equation (34) to be expanded:

$$\frac{e^{-i\varphi}N(e^{-i\varphi})}{X(e^{-i\varphi})} = \sum_{\sigma=1}^{\infty} \hat{G}_\sigma e^{-i\sigma\varphi}. \tag{35}$$

Using equations (32), (33), (34) and (35) in equation (31) now gives

$$\delta(e^{-i\varphi}) + \sum_{\sigma=1}^{\infty} \hat{G}_\sigma e^{-i\sigma\varphi} = \sum_{\sigma=1}^{2n+1} G_\sigma e^{-i\sigma\varphi}; \tag{36}$$

$$\begin{aligned} P_\sigma &= \hat{G}_\sigma + C_{n+1-\sigma}, \qquad \sigma \leqq n+1; \\ &= \hat{G}_\sigma, \qquad n+1 < \sigma < 2n+1; \end{aligned}$$

$$\delta(e^{i\varphi}) = -\sum_{2n+2}^{\infty} \hat{G}_\sigma e^{-i\sigma\varphi}.$$

This δ is small provided the zeros z_i of $X(z)$ are such that $z_i^{-(2n+2)}$ is small. When δ is small, the actual error extrema will differ from ϵ, but by no more than $\pm\ |\delta|\ \epsilon$.

Equation (34) requires

$$\frac{\epsilon}{2} A(e^{-i\varphi})X(e^{i\varphi}) + B(e^{i\varphi})X(e^{-i\varphi}) = A(e^{i\varphi})e^{-i\varphi}N(e^{-i\varphi})\ . \tag{37}$$

The appropriate degree η of polynomial $X(\cdot)$ turns out to be one less than the number of poles of $zB(z)/A(z)$ (including any poles at $z = \infty$).

When the degree of $A(\cdot)$ is greater than the degree of $B(\cdot)$ and the zeros of $A(\cdot)$ are distinct, a set of $\eta + 1$ homogeneous equations in the coefficients q_i of $X(\cdot)$ can be derived by evaluating equation (37) at the zeros of $A(\cdot)$. Then

$$\left(M_A + \frac{\epsilon}{2} M_B\right)Q = 0 \tag{38}$$

in which Q is the column matrix of the coefficients q_i of X and M_A and M_B are square matrices of order $\eta + 1$. Under other conditions an equation of the same form can be obtained in other ways.

Equation (38) requires $\epsilon/2$ to be one of $\eta + 1$ eigenvalues for which the matrix coefficient of Q is singular. Each eigenvalue determines a polynomial $X(\cdot)$ [including an arbitrary scale factor which cancels out in equation (33)]. For our purposes, we must choose an ϵ which is real and such that the zeros of $X(z)$ lie outside the unit circle. This raises a question of the existence of a suitable ϵ and $X(\cdot)$.

When degree $\eta = 0$, $X(\cdot)$ is a constant, equation (38) is a real linear equation in ϵ, and there is no zero of $X(\cdot)$. When $\eta = 1$, $X(\cdot)$ is linear, ϵ is a root of a quadratic equation. It is not hard to show that the two roots are real [under our assumptions regarding zeros of $A(\cdot)$] and that one (the larger) yields a zero z_1 of $X(z)$ such that $|z_1| \geq 1$. Equality occurs only in singular cases such that $n + 2$ zero error points are possible [even though there are only $n + 1$ disposable coefficients of $\mathbf{P}(x)$] and can be so placed that there are $n + 3$ equal error extrema, instead of only $n + 2$. This may be seen by assuming that the zero of a linear $X(z)$ is ± 1, and then noting that, in equation (33),

$$\frac{X(e^{i\varphi})}{X(e^{-i\varphi})} = \frac{1 \pm e^{i\varphi}}{1 \pm e^{-i\varphi}} = \pm e^{i\varphi}. \tag{39}$$

Conditions for the existence of a suitable ϵ have not been established for $\eta > 2$. They are probably at least closely related to the (unknown) general conditions under which the minimax approximation has the equal ripple form of Fig. 1.

The procedures described here are appropriate only when a suitable ϵ does in fact exist. However, degrees $\eta = 0$ and 1, for which existence has been established, are sufficient for many practical problems. For approximately equal error extrema, equation (32) itself need be only an approximation, and polynomials $A(\cdot)$ and $B(\cdot)$ for which $\eta = 0$ or 1 are likely to give a good enough approximation. This is particularly true when degree n of $\mathbf{P}(x)$ is sufficiently large so that coefficients

C_σ , $\sigma > n$, in equation (26) approach a simple asymptotic behavior. Percentage variations between error extrema need not have to be very small even though the absolute errors must be very small. For example, a 10 percent variation between very small extrema may be acceptable, compared with large variations obtained by truncation of the infinite Chebyshev polynomial series.

Table I indicates the degree m of $A(\cdot)$ and μ of $B(\cdot)$ for which $\eta = 0$ or 1. The column headed $m + \mu + 1$ indicates the number of disposable parameters in the rational fraction A/B which can be adjusted to approximate the sum in equation (32).

The procedures for $\eta = 0$ and 1 are particularly well suited for rapid explorations of available error magnitudes as functions of initial design parameters, such as degree of the disposable polynomial, extent of the approximation interval, and parameters in the approximated function. When only the error magnitude is needed, it is not necessary to calculate the coefficients of polynomial $P(\cdot)$, which requires the series expansion (35). When $\eta = 0$, the error magnitude is (approximately) the single ϵ determined by equation (38). Then simple closed form formulas can frequently be obtained (and will be included in 4 of the 5 examples in Section V). When $\eta = 1$, ϵ is one of the two roots of the quadratic equation required by equation (38). (To meet the condition on the zero of $X(\cdot)$, the larger ϵ must be chosen.)

When ϵ has been determined, it can be compared with the error ϵ_T obtained by simply truncating the Chebyshev polynomial expansion of $\mathbf{F}(x)$. In terms of equations (26) and (32)

$$\epsilon_T \cong \frac{B(e^{i\varphi})}{A(e^{i\varphi})} e^{i(n+1)\varphi} + \frac{B(e^{-i\varphi})}{A(e^{-i\varphi})} e^{-i(n+1)\varphi}. \tag{40}$$

Comparing the maximum ϵ_T (at real φ) with ϵ indicates the improve-

TABLE I—Value of m and μ for which η is 0 or 1.

Degree m of $A(\cdot)$	Degree μ of $B(\cdot)$	Degree η of $X(\cdot)$	$m + \mu + 1$
0	0	0	1
1	0	0	2
2	0	1	3
0	1	1	2
1	1	1	3
2	1	1	4

ment to be obtained by the minimax refinement of the truncated series. Frequently, max ϵ_T occurs at $\varphi = 0$ or π, and then

$$\max \epsilon_T = 2\,\frac{B(\pm 1)}{A(\pm 1)}. \tag{41}$$

4.2 *Weighted Minimax Approximations*

Let

$$\mathbf{P}(x) = \mathbf{F}(x) + \frac{1}{\mathbf{W}(x)}\,\boldsymbol{\epsilon}(x) \tag{42}$$

in which $\mathbf{P}(x)$ is again a disposable polynomial of degree n, $\mathbf{F}(x)$ is again a given function to be approximated in the interval $x_a \leqq x \leqq x_b$, and the new function $\mathbf{W}(x)$ is a given weight factor. For what $\mathbf{P}(x)$ is $\boldsymbol{\epsilon}(x)$ smallest in the minimax sense? We will again assume that the minimax $\boldsymbol{\epsilon}(x)$ has the equal ripple form and will seek only approximations to equal ripples. We will also assume that $\mathbf{W}(x)$ is bounded and positive definite in the approximation interval. (A point where $\mathbf{W}(x) = 0$ or ∞ would probably spoil the equal ripple character of the minimax approximation.)

Map from x to φ as before and define $P(\cdot)$, $F(\cdot)$ and $E(\cdot)$ again by equations (24), (26), and (28). Express $\mathbf{W}(x)$ also in terms of exponentials, but as a product instead of a sum. More specifically, let

$$\mathbf{H}(x) = -\log \mathbf{W}(x) = H(e^{i\varphi}) + H(e^{-i\varphi}) \tag{43}$$

and assume that $\mathbf{W}(x)$ is sufficiently smooth, as well as bounded and positive definite, so that $H(z)$ is regular at $|z| \leqq 1$. Then

$$\begin{aligned} \frac{1}{\mathbf{W}(x)} &= D(e^{i\varphi})D(e^{-i\varphi}), \\ D(e^{i\varphi}) &= e^{H(e^{i\varphi})} \end{aligned} \tag{44}$$

with $\log D(z)$ regular when $|z| \leqq 1$. This $D(\cdot)$ is a generalization of the $D(\cdot)$ of Subsection 3.1 and of $[S(\cdot)]^{\frac{1}{2}}$ of Subsection 3.2. Frequently it can be found by direct factorization of a function of $e^{i\varphi}$ as in Section III.

Equations like (29) and (30) can now be obtained as before. The only difference is that $E(\cdot)$ must now be multiplied by the product of functions of φ in equation (44). Then equation (30) becomes

$$P(e^{i\varphi}) = F(e^{i\varphi}) + D(e^{i\varphi})D(e^{-i\varphi})E(e^{i\varphi}) \tag{45}$$

and equation (31) becomes

$$\sum_{\lambda=1}^{2n+1} G_\lambda e^{-i\lambda\varphi} = \frac{\epsilon}{2} D(e^{i\varphi})D(e^{-i\varphi})e^{if(\varphi)} + \sum_{\lambda=0}^{\infty} C_{n+1+\lambda}e^{i\lambda\varphi} ; \tag{46}$$

$$\begin{aligned} G_\lambda &= P_{n+1-\lambda} - C_{n+1-\lambda} , \qquad \lambda \leqq n+1; \\ &= P_{n+1-\lambda} , \qquad n+1 < \lambda \leqq 2n+1. \end{aligned}$$

Retain the rational fraction $B(\cdot)/A(\cdot)$ of equation (32), but change equation (33) to

$$e^{if(\varphi)} = \frac{D(e^{-i\varphi})A(e^{-i\varphi})X(e^{i\varphi})}{D(e^{i\varphi})A(e^{i\varphi})X(e^{-i\varphi})} + \hat{\delta}(e^{i\varphi}) \tag{47}$$

so that

$$D(e^{i\varphi})D(e^{-i\varphi})e^{if(\varphi)} = \frac{D^2(e^{-i\varphi})A(e^{-i\varphi})X(e^{i\varphi})}{A(e^{i\varphi})X(e^{-i\varphi})} + \delta(e^{i\varphi}) \tag{48}$$

in which $\hat{\delta}$ and δ are small. Then change equation (34) to

$$\frac{\epsilon}{2}\,\frac{D^2(e^{-i\varphi})A(e^{-i\varphi})X(e^{i\varphi})}{A(e^{i\varphi})X(e^{-i\varphi})} + \frac{B(e^{i\varphi})}{A(e^{i\varphi})} = \frac{e^{-i(\varphi)}N(e^{-i\varphi})}{X(e^{-i\varphi})}. \tag{49}$$

Using equations (32), (47) and (35) in equation (46) now gives equation (36) again. From equation (49), equation (37) must be changed to

$$\frac{\epsilon}{2} D^2(e^{-i\varphi})A(e^{-i\varphi})X(e^{i\varphi}) + B(e^{i\varphi})X(e^{-i\varphi}) = A(e^{i\varphi})e^{-i\varphi}N(e^{-i\varphi}). \tag{50}$$

Equation (50) can be used to find ϵ, and the $X(\cdot)$ and $N(\cdot)$ needed for equations (35) and (36).

4.3 *More General Approximating Functions*

Let

$$\Psi[\mathbf{P}(x), x] = \mathbf{G}(x) + \boldsymbol{\varepsilon}(x) \tag{51}$$

in which $\Psi[\mathbf{P}(x), x]$ is a given function of x and a disposable polynomial $\mathbf{P}(x)$, $\mathbf{G}(x)$ is a given function to be approximated by $\Psi[\mathbf{P}(x), x]$ in the interval $x_a \leqq x \leqq x_b$, and $\boldsymbol{\varepsilon}(x)$ is the error in the approximation. For what $\mathbf{P}(x)$ is $\boldsymbol{\varepsilon}(x)$ smallest in the minimax sense? Under certain further assumptions regarding $\Psi(\cdot\ ,\ \cdot)$ this approximation can be transformed into a weighted minimax polynomial approximation.

Assume an inverse Ψ^{-1} of $\Psi[\mathbf{P}(x), x]$, with respect to $\mathbf{P}(x)$, exists over the approximation interval. Then equation (51) can be replaced by

$$\mathbf{P}(x) = \Psi^{-1}\,\{[\mathbf{G}(x) + \boldsymbol{\varepsilon}(x)], x\}. \tag{52}$$

Assume $\Psi^{-1}(\cdot\ ,\ \cdot)$ is sufficiently smooth and $\varepsilon(x)$ sufficiently small to justify the following approximation (in the interval $x_a \leqq x \leqq x_b$):

$$\mathbf{P}(x) = \Psi^{-1}[\mathbf{G}(x), x] + \frac{\partial \Psi^{-1}[\mathbf{G}(x), x]}{\partial \mathbf{G}(x)} \varepsilon(x). \tag{53}$$

This is in the general form (42) with

$$\begin{aligned} \mathbf{F}(x) &= \Psi^{-1}[\mathbf{G}(x), x] \\ \frac{1}{\mathbf{W}(x)} &= \frac{\partial \Psi^{-1}[\mathbf{G}(x), x]}{\partial \mathbf{G}(x)}. \end{aligned} \tag{54}$$

Thus Subsection 4.2 can now be applied provided the $\mathbf{F}(\cdot)$ and $\mathbf{W}(\cdot)$ determined by equation (54) meet the appropriate conditions. Recall that we required $\mathbf{W}(x)$ to be bounded and positive definite over the interval of approximation. However, reversing the sign of $\mathbf{W}(x)$ merely reverses the sign of $\varepsilon(x)$. Hence, in equation (54), we need only require that the partial derivative must be bounded and either positive definite or negative definite and sufficiently smooth for log $D(z)$ to be regular when $|z| < 1$.

As a first example of the inversion of equation (51), let

$$\mathbf{W}(x)\mathbf{P}(x) = \mathbf{G}(x) + \varepsilon(x) \tag{55}$$

where $\mathbf{W}(x)$ and $\mathbf{G}(x)$ are given functions of x. Then

$$\mathbf{P}(x) = \frac{\mathbf{G}(x) + \varepsilon(x)}{\mathbf{W}(x)} = \frac{\mathbf{G}(x)}{\mathbf{W}(x)} + \frac{1}{\mathbf{W}(x)} \varepsilon(x). \tag{56}$$

As a second example, let

$$[\mathbf{A}(x) + \mathbf{B}(x)\mathbf{P}(x)]^{\frac{1}{2}} = \mathbf{G}(x) + \varepsilon(x) \tag{57}$$

where $\mathbf{A}(x)$, $\mathbf{B}(x)$, and $\mathbf{G}(x)$ are given functions of x, with $\mathbf{B}(x)$ and $\mathbf{G}(x)$ positive definite over the approximation interval. Solving for $\mathbf{P}(x)$ gives

$$\mathbf{P}(x) = \frac{\mathbf{G}^2(x) - \mathbf{A}(x)}{\mathbf{B}(x)} + 2\,\frac{\mathbf{G}(x)}{\mathbf{B}(x)} \varepsilon(x) + \frac{1}{\mathbf{B}(x)} \varepsilon^2(x). \tag{58}$$

If the term in $\varepsilon^2(x)$ is omitted

$$[\mathbf{A}(x) + \mathbf{B}(x)\mathbf{P}(x)]^{\frac{1}{2}} \cong \mathbf{G}(x) + \varepsilon(x) - \frac{\varepsilon^2(x)}{2\mathbf{G}(x)} \tag{59}$$

in which terms in $\varepsilon^{\sigma}(x)$ have been neglected for $\sigma > 2$. An equal ripple $\varepsilon(x)$ in equation (57) yields an approximately equal ripple error if the last term is somewhat smaller than the extrema of $\varepsilon(x)$.

4.4 *Relation to Phase Modulation*

The function of φ defined by equation (28) is similar to functions of time t used in communication theory to describe phase modulated signals. If φ is replaced by t in equation (28),

$$E(e^{it}) = \frac{\epsilon}{2} e^{i[(n+1)t+if(t)]}. \tag{60}$$

This is the exponential representation of a phase modulated signal in which the carrier (radian) frequency is $n + 1$ and the baseband signal $f(t)$ is periodic with one period every $n + 1$ periods of the carrier. The signal may be regarded as the carrier plus sequences of upper and lower sidebands. The upper sidebands are determined by the coefficients $C_{n+1+\lambda}$ in equation (31). The lower sidebands are determined by the requirement of a purely phase modulated signal. Finally, if the sequence of lower sidebands extends as far as the negative carrier frequency $-(n + 1)$ we truncate it at $-n$.

Weighted minimax approximations can be interpreted similarly, in terms of simultaneous phase and amplitude modulation.

4.5 *Alternative Procedures*

It is obvious that the procedures described above can be varied in many different ways. A very few of the possible variations are noted below.

Preliminary manipulations may be needed to obtain a formulation in which the disposable part is a polynomial. Also, the pertinent Fourier series may be sums of sines instead of cosines. Both these situations will be illustrated by Example 5, in Section V.

If $\partial\Psi^{-1}[\mathbf{G}(x), x]/\partial\mathbf{G}(x)$ is expressed as a product of functions of x, $D(e^{i\varphi})$ can be formulated as a product of corresponding factors. A factor of the form $(1 - x/x_\sigma)^\rho$ contributes a factor of the form $[M_\sigma(1 - \gamma_\sigma e^{i\varphi})]^\rho$, as in Section III. More generally, there may be advantages to replacing the $D(\cdot)$ of equation (44) by $\hat{D}(\cdot)$ defined

$$\mathbf{W}^{\pm 2}(x) = \hat{D}(e^{i\varphi})\hat{D}(e^{-i\varphi}). \tag{61}$$

Then the $D^2(\cdot)$ in equations (48), (49) and (50) is replaced by $\hat{D}^{\pm 1}(\cdot)$.

It may sometimes be convenient to express $\mathbf{P}(x)$ and $\mathbf{F}(x)$ as products of factors in exp $(\pm i\varphi)$ instead of sums, say

$$\mathbf{P}(x) = P(e^{i\varphi})P(e^{-i\varphi}), \qquad \mathbf{F}(x) = F(e^{i\varphi})F(e^{-i\varphi}) \tag{62}$$

in which $P(\cdot)$ is a polynomial of degree n with no negative powered terms. If a term in ϵ^2 is neglected, one can now replace equation (29) by

$$P(e^{i\varphi})P(e^{-i\varphi}) = \left[F(e^{i\varphi}) + \frac{E(e^{i\varphi})}{F(e^{-i\varphi})}\right]\left[F(e^{-i\varphi}) + \frac{E(e^{-i\varphi})}{F(e^{i\varphi})}\right]. \tag{63}$$

Equating factors separately replaces equation (30) by

$$P(e^{i\varphi}) = F(e^{i\varphi}) + \frac{E(e^{i\varphi})}{F(e^{-i\varphi})}. \tag{64}$$

Subsequent modifications of our previous procedures are now easily worked out.

It would be possible to replace equation (33) by other functional forms for exp $[if(\varphi)]$. The moduli must approximate unity at real φ and expansions in positive and negative powers of exp $(i\varphi)$ must exist. Disposable parameters are to be adjusted so as to approximate the required coefficients of positive powers. However, except for very special functional forms [such as equation (33)], the adjustment is likely to be a quite complicated task.

V. EXAMPLES

This section further clarifies the general procedures by means of five examples.

5.1 *Example 1*

Let

$$\frac{\mathbf{P}(x)}{(1 - x/x_0)^{\frac{1}{2}}} = 1 + \varepsilon(x), \qquad -1 \leqq x \leqq +1 \tag{65}$$

in which x_0 is a given constant, $|x_0| > 1$, and the degree n of the disposable polynomial $\mathbf{P}(x)$ is large. What is the approximate amplitude $|\epsilon|$ of the equal error extrema of the minimax approximation?

This is a special case of equations (55) and (56), which can be solved as a special case of equation (42) for which

$$\mathbf{F}(x) = \frac{1}{\mathbf{W}(x)} = (1 - x/x_0)^{\frac{1}{2}}. \tag{66}$$

To apply Section IV, define $F(\cdot)$ and $D(\cdot)$ by

$$\begin{aligned} (1 - x/x_0)^{\frac{1}{2}} &= F(e^{i\varphi}) + F(e^{-i\varphi}) = D(e^{i\varphi})D(e^{-i\varphi}), \\ F(e^{i\varphi}) &= \sum_{\sigma=0}^{\infty} C_\sigma e^{i\sigma\varphi}, \end{aligned} \tag{67}$$

Log $D(z)$ regular, $|z| \leqq 1$.

We have already factored a linear function of x in terms of exp $(\pm i\varphi)$, in Section III. A similar factorization now gives

$$D(e^{i\varphi}) = \frac{(1 - \gamma e^{i\varphi})^{\frac{1}{2}}}{(1 + \gamma^2)^{\frac{1}{4}}}, \qquad |\gamma| < 1 \tag{68}$$

and then the coefficients of C_σ correspond to an expansion of

$$\left[\frac{(1 - \gamma e^{i\varphi})(1 - \gamma e^{-i\varphi})}{1 + \gamma^2}\right]^{\frac{1}{2}} = \sum_{\sigma=0}^{\infty} C_\sigma e^{i\sigma\varphi} + \sum_{\sigma=0}^{\infty} C_\sigma e^{-i\sigma\varphi}. \tag{69}$$

To determine ϵ we only need the coefficients C_σ for $\sigma > n$, which we have assumed to be large.

The following expansion of $[1 - \gamma \exp (i\varphi)]$, valid for $|\gamma| \leqq 1$, is well known

$$\begin{aligned} (1 - \gamma e^{i\varphi})^{\frac{1}{2}} &= \sum_{\sigma=0}^{\infty} K_\sigma e^{i\sigma\varphi}; \\ K_0 &= -1, \qquad K_1 = -\gamma/2; \\ K_\sigma &= \frac{-(2\sigma - 3)!}{4^{\sigma-1}(\sigma - 2)!\,\sigma!}\gamma^\sigma, \qquad \sigma \geqq 2. \end{aligned} \tag{70}$$

When n is large and $\lambda \ll n$,

$$\frac{K_{n+\lambda+1}}{K_{n+\lambda}} = \frac{2(n + \lambda) - 1}{2(n + \lambda) + 2}\gamma \cong k\gamma \tag{71}$$

in which

$$k = \frac{2n + 1}{2n + 4} = 1 - \frac{3}{2n + 4} \cong 1. \tag{72}$$

As a result, when n is large

$$(1 - \gamma e^{i\varphi})^{\frac{1}{2}} \cong \sum_{\sigma=0}^{n} K_\sigma e^{i\sigma\varphi} + \frac{K_{n+1}e^{i(n+1)\varphi}}{1 - k\gamma e^{i\varphi}}. \tag{73}$$

Now note that

$$\left[\frac{1 - \gamma e^{-i\varphi}}{1 + \gamma^2}\right]^{\frac{1}{2}}\left[\frac{K_{n+1}}{1 - k\gamma e^{i\varphi}}\right] = \sum_{\sigma=1}^{\infty} L_\sigma e^{-i\varphi} + \frac{(1 - k\gamma^2)^{\frac{1}{2}}K_{n+1}}{(1 + \gamma^2)^{\frac{1}{2}}(1 - k\gamma e^{i\varphi})}. \tag{74}$$

If equation (74) is used to evaluate C_σ in equation (67), only the last term contributes to C_σ when $\sigma > n$. Then

$$\sum_{\lambda=0}^{\infty} C_{n+1+\lambda} e^{i\lambda\varphi} \cong \frac{(1 - k\gamma^2)^{\frac{1}{2}} K_{n+1}}{(1 + \gamma^2)^{\frac{1}{2}}(1 - k\gamma e^{i\varphi})} \tag{75}$$

which is a special case of equation (32) with $A(\cdot)$ a linear polynomial and $B(\cdot)$ a constant. The corresponding $X(\cdot)$ in equation (47) is a constant and cancels out. Then equations (68) and (75) applied to equation (50) give

$$\epsilon = \frac{-2K_{n+1}}{(1 - k^2\gamma^2)(1 - k\gamma^2)^{\frac{1}{2}}}, \tag{76}$$

$$N(e^{-i\varphi}) = G_1 + G_2 e^{-i\varphi}.$$

The constants G_1 and G_2 contribute to the two highest degree terms in the polynomial $P(\cdot)$. They need not be computed unless the specific polynomial is needed as well as the amplitude $|\epsilon|$ of the approximation errors.

The linear $A(\cdot)$ and constant $B(\cdot)$ determined by equation (75) can be used in equation (40) to approximate the truncation error for the polynomial approximation defined by equation (66). The corresponding error in equation (65) can be found by dividing by $(1 - x/x_0)^{1/2}$. This gives

$$r = \frac{\max |\epsilon_T|}{|\epsilon|} = \frac{(1 - k\gamma^2)(1 + k|\gamma|)}{(1 - |\gamma|)}. \tag{77}$$

If $k = 1$, $r = (1 + |\gamma|)^2 < 4$. Actually $k < 1$, but further analysis indicates that r will not be significantly >4 when γ^n is small.

When $|x_0| \to \infty$, $W(x) \to 1$, $C_{n+1+\lambda}/C_{n+1} \to 0$, and ϵ_T is dominated by a single Chebyshev polynomial (which has equal extrema). Consistent with this our $\gamma \to 0$, then G_1, $G_2 \to 0$ in equation (76) and $r \to 1$ in equation (77).

In equation (74), K_{n+1} can be determined by the formula for K_σ in equation (70). However, the following simpler approximate formula may be more useful:

$$K_{n+1} \cong \frac{-\gamma^{n+1}}{2(\pi)^{\frac{1}{2}}(n + 1)(n + 1/4)^{\frac{1}{2}}}. \tag{78}$$

The error amounts to about 0.3 percent at $n = 2$ and about 0.04 percent at $n = 6$. The derivation is related to, but requires more than substitution of Stirling's approximation for the factorials in equation (70).

Fig. 3 illustrates computed errors $\varepsilon(x)$ and $\varepsilon_T(x)$ corresponding re-

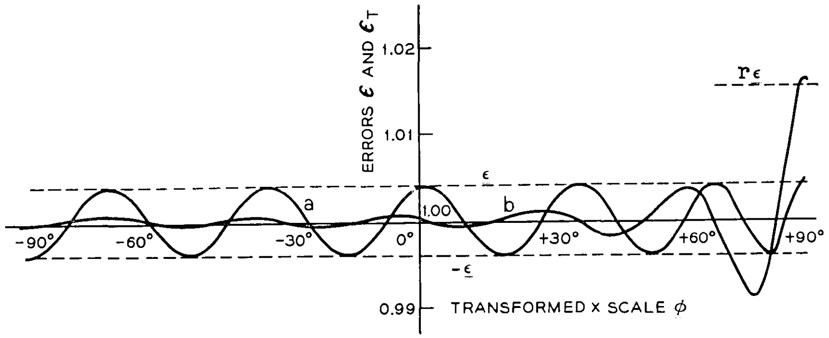


Fig. 3 — Illustrating Example 1: (a) $\varepsilon(x)$ and (b) $\varepsilon_T(x)$.

spectively to our approximately equal ripple solution and truncation of the Chebyshev polynomial series. The constants $\pm\epsilon$ and $r\epsilon$ determined by equations (74) and (75) are included for comparison. The computations started with

$$x_0 = 1.025, \qquad n = 10$$

for which, as computed by equations (76), (77) and (78),

$$\gamma = 0.8, \qquad k = 7/8,$$

$$K_{n+1} \cong 0.0006881, \qquad \epsilon \cong 0.0040678, \qquad r \cong 3.74.$$

5.2 *Example 2*

Let

$$\mathbf{P}(x) = (1 - x/x_0)^{1/2} + \varepsilon(x), \qquad -1 \leqq x \leqq +1 \tag{79}$$

in which the degree n of $\mathbf{P}(x)$ is again large and x_0 is again a given constant, $|x_0| > 1$.

Since the function $\mathbf{F}(x)$ is the same as in Example 1, equation (75) is again valid. Now, however, $\mathbf{W}(x) = 1$ and hence Section 4.1 (on unweighted polynomial approximations) is appropriate. Applying equation (75) to equation (37) gives

$$\epsilon = \frac{(1 - k\gamma^2)^{\frac{1}{2}} K_{n+1}}{(1 + \gamma^2)^{\frac{1}{2}}(1 - k^2\gamma^2)} \tag{80}$$

$$N(e^{-i\varphi}) = G_1$$

in which $N(\cdot)$ contributes only to the highest degree term in $P(\cdot)$. The error ratio r turns out to be

$$r = \frac{\max | \epsilon_T |}{| \epsilon |} = 1 + k \, | \gamma | < 2. \tag{81}$$

5.3 *Example 3**

Let

$$(1 - x/x_0)^{1/2} \mathbf{P}(x) = 1 + \varepsilon(x), \qquad -1 \leqq x \leqq 1 \tag{82}$$

in which degree n of $\mathbf{P}(x)$ is again large and x_0 is given, $|x_0| > 1$.

In the equivalent weighted polynomial approximation

$$\mathbf{F}(x) = \frac{1}{\mathbf{W}(x)} = (1 - x/x_0)^{-\frac{1}{2}}. \tag{83}$$

Proceeding as in Example 1, one now gets

$$\begin{aligned} D(e^{i\varphi}) &= \frac{(1 + \gamma^2)^{\frac{1}{4}}}{(1 - \gamma e^{i\varphi})^{\frac{1}{2}}}, \\ \sum_{\sigma=0}^{\infty} C_{n+1+\lambda} e^{i\lambda\varphi} &\cong \frac{(1 + \gamma^2)^{\frac{1}{2}} K_{n+1}}{(1 - k\gamma^2)^{\frac{1}{2}}(1 - k\gamma e^{i\varphi})}, \\ K_{n+1} &= \frac{(2n + 1)! \, \gamma^{n+1}}{2^{n+1} n! \, (n + 1)!} \cong \frac{\gamma^{n+1}}{[\pi(n + 5/4)]^{\frac{1}{2}}}, \\ k &= \frac{2n + 3}{2n + 4} = 1 - \frac{1}{2n + 4}. \end{aligned} \tag{84}$$

Then equation (49) gives

$$\begin{aligned} \epsilon &= \frac{2K_{n+1}(1 - k\gamma^2)^{\frac{1}{2}}}{1 - k^2\gamma^2}, \\ e^{-i\varphi} N(e^{-i\varphi}) &= \frac{N_1 e^{-i\varphi}}{1 - \gamma e^{-i\varphi}}, \\ N_1 &= \frac{\epsilon}{2} \frac{\gamma(1 - k)(1 + \gamma^2)^{\frac{1}{2}}}{1 - k\gamma^2}. \end{aligned} \tag{85}$$

Equation (35) is now

$$\frac{N_1 e^{-i\varphi}}{1 - \gamma e^{-i\varphi}} = \sum_{\sigma=1}^{\infty} \hat{G}_\sigma e^{-i\sigma\varphi}. \tag{86}$$

The first $2n + 1$ terms in this series contribute to $P(\cdot)$ per equation

---

* The author has encountered this problem in connection with two different circuit theory studies, which will be described in other papers.

(36). The remainder can be summed, to get

$$\delta(e^{i\varphi}) = \frac{N_1\gamma^{2n+1}e^{-i(2n+2)\varphi}}{1 - \gamma e^{-i\varphi}}. \tag{87}$$

Evaluating the error corresponding to simple truncation now gives

$$r = \frac{\max \mid \epsilon_T \mid}{\mid \epsilon \mid} \cong \frac{(1 - \mid \gamma \mid)(1 - k^2\gamma^2)}{(1 - k \mid \gamma \mid)(1 - k\gamma^2)}. \tag{88}$$

When n is large, $k \cong 1$ and r is so close to unity that the minimax refinement of simple truncation is not likely to be justified. However, our analysis has been useful in disclosing this fact, without the detailed computation of any minimax approximations.

5.4 *Example 4*

Previous work, which we shall discuss in Section VI, concerns the following problem: Let

$$\mathbf{P}_n(x) = \mathbf{P}_{n+\nu}(x) + \boldsymbol{\varepsilon}(x), \qquad -1 \leqq x \leqq +1 \tag{89}$$

in which $\mathbf{P}_{n+\nu}(x)$ is a given polynomial of degree $n + \nu$ and $\mathbf{P}_n(x)$ is a disposable polynomial of degree n. For what $\mathbf{P}_n(x)$ does $\boldsymbol{\varepsilon}(x)$ have the equal ripple form?

Equations (32), (33) and (37) now simplify to

$$\begin{gathered} \sum_{\lambda=0}^{\infty} C_{n+1+\lambda}(e^{i\lambda\varphi}) = \sum_{\sigma=0}^{\nu-1} C_{n+1+\lambda}(e^{i\lambda\varphi}) = B(e^{i\lambda\varphi}), \\ e^{if(\varphi)} = \frac{X(e^{i\varphi})}{X(e^{-i\varphi})} + \delta(e^{i\varphi}), \end{gathered} \tag{90}$$

$$\frac{\epsilon}{2} X(e^{i\varphi}) + B(e^{i\varphi})X(e^{-i\varphi}) = e^{-i\varphi}N(e^{-i\varphi})$$

in which $B(\cdot)$ is a polynomial of degree $\nu - 1$, with coefficients $C_{n+1+\lambda}$ and $X(\cdot)$ is a polynomial of degree $\nu - 1$, to be found therefrom. The coefficients of $B(\cdot)$ can be found by expanding the left side of the last equation and equating to zero the coefficients of pgsitive powers of exp $(i\varphi)$. The result can be expressed as the following specialization of equation (38):

$$\left(C + \frac{\epsilon}{2} I\right)Q = 0 \tag{91}$$

in which Q is again a column matrix whose elements are the ν coefficients of $X(\cdot)$ and I is the identity matrix of order ν. The matrix C has the special form (assuming the elements q_σ of Q to be ordered per $X(z) = \sum q_\sigma z^\sigma$)

$$C = \begin{vmatrix} C_{n+1} & C_{n+2} & \cdots & C_{n+\nu-1} & C_{n+\nu} \\ C_{n+2} & C_{n+3} & \cdots & C_{n+\nu} & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ C_{n+\nu} & 0 & \cdots & 0 & 0 \end{vmatrix}, \tag{92}$$

When $\nu = 1$, X is a constant and the solution is elementary. When $\nu = 2$, $X(z)$ is linear. Let z_1 be its zero. Then equations (91) and (92) require

$$\epsilon^2 + 2C_{n+1}\epsilon - 4C_{n+2}^2 = 0,$$

$$z_1 = \frac{\epsilon}{2C_{n+2}}. \tag{93}$$

The roots of the quadratic equation in ϵ are real. When $C_{n+1} \neq 0$, the larger $|\epsilon| > 2|C_{n+2}|$ and $|z_1| > 1$, as required. Then δ in equation (36) turns out to be a power series in $\exp(-i\varphi)$ which can be summed to get

$$\delta(e^{i\varphi}) = \frac{-\gamma^{2n+2}C_{2n+2}e^{-i(2n+2)\varphi}}{1 - \gamma e^{-i\varphi}}, \tag{94}$$

$$\gamma = 1/z_1 .$$

When $C_{n+1} = 0$, $\epsilon = \pm 2C_{n+2}$ and $z_1 = \pm 1$. But then the error due to simple truncation of the Chebyshev polynomial expansion of $\mathbf{P}_{n+2}(x)$ is proportional to a single Chebyshev polynomial of degree $n + 2$, which has equal ripples with $n + 3$ extrema instead of $n + 2$.

5.5 *Example 5**

As a last example consider the following nonalgebraic approximation:

$$\Gamma(\theta) = \sum_{\sigma=1}^{n} \mathrm{A}_\sigma \sin \sigma\theta = \theta + \varepsilon(\theta) \tag{95}$$

$$-\pi < -\theta_c \leqq \theta \leqq \theta_c < \pi.$$

For what coefficients A_σ does the error $\varepsilon(\theta)$ have the equal ripple form and what is the amplitude ϵ of the ripples?

---

* This problem is of interest in, for example, the approximation of differentiation with a tapped delay line. A more detailed treatment is planned for a future paper.

A sequence of transformations changes equation (95) into a weighted polynomial approximation. First, equation (95) is equivalent to

$$\bar{\Gamma}(\theta) = \sin\theta \sum_{\rho=0}^{n=1} \mathbf{B}_\rho \cos(\rho\theta), \tag{96}$$

$$2\mathbf{A}_\rho = \mathbf{B}_{\rho-1} - \mathbf{B}_{\rho+1} \,.$$

Second, relate θ to a new variable φ by

$$\sin\frac{\theta}{2} = q\sin\varphi, \qquad |\,\theta\,| < \pi; \tag{97}$$

$$q = \sin\frac{\theta_c}{2} < 1.$$

Real φ maps into $-\theta_c \leqq \theta \leqq \theta_c$. Also

$$\begin{aligned} \cos\theta &= 1 - q^2 + q^2\cos 2\varphi, \\ \sin\theta &= 2q(1 - q^2\sin^2\varphi)^{\frac{1}{2}}\sin\varphi. \end{aligned} \tag{98}$$

In these terms, equation (96) becomes

$$\bar{\Gamma}(\theta) = 2q(1 - q^2\sin^2\varphi)^{\frac{1}{2}}\sin\varphi \sum_{\rho=0}^{n-1} B_\rho \cos(2\rho\varphi) \tag{99}$$

in which the set of coefficients B_ρ is linearly related to the set $\mathbf{B}_\rho$ of equation (96). In equivalent exponential terms

$$\hat{\Gamma}(\theta) = \frac{q}{i}(1 - q^2\sin^2\varphi)^{\frac{1}{2}}[P(e^{i\varphi}) - P(e^{-i\varphi})] \tag{100}$$

in which $P(z)$ is a polynomial of degree $2n - 1$ in odd powers of z only.

Use equation (100) in equation (95) and solve for the factor in [ ]. The result can be expressed in terms of exponentials:

$$\begin{aligned} P(e^{i\varphi}) - P(e^{-i\varphi}) &= F(e^{i\varphi}) - F(e^{-i\varphi}) \\ &\quad + D(e^{i\varphi})D(e^{-i\varphi})[E(e^{i\varphi}) - E(e^{-i\varphi})] \end{aligned} \tag{101}$$

in which $F(\cdot)$, $D(\cdot)$, and $E(\cdot)$ are related to previous functions by

$$\frac{\theta}{q}(1 - q^2\sin^2\varphi)^{-\frac{1}{2}} = \sum_{\sigma=1}^{\infty} 2C_{2\sigma-1}\sin(2\sigma - 1)\varphi; \tag{102}$$

$$F(e^{i\varphi}) = \sum_{\sigma=1}^{\infty} C_{2\sigma-1}e^{i(2\sigma-1)\varphi};$$

$$\frac{1}{q}(1 - q^2\sin^2\varphi)^{-\frac{1}{2}} = D(e^{i\varphi})D(e^{-i\varphi});$$

$$D(e^{i\varphi}) = \left[\frac{q(1+\gamma e^{i2\varphi})}{1+\gamma}\right]^{\frac{1}{2}}, \qquad |\gamma| < 1;$$

$$\mathbf{E}(\theta) = \frac{1}{i}\,[E(e^{i\varphi}) - E(e^{-i\varphi})];$$

$$E(e^{i\varphi}) = \frac{\epsilon}{2}\,e^{i[(2n+1)\varphi+f(\varphi)]}.$$

It can be shown that the coefficients $C_{2\sigma-1}$ obey a difference equation of order 2. The asymptotic behavior of the difference equation shows that

$$\text{as } \sigma \to \infty, \qquad C_{2\sigma+1} \to -\left(\frac{2\sigma-1}{2\sigma+1}\right)^{\frac{1}{2}} \gamma C_{2\sigma-1}\,. \tag{103}$$

As a result, for a sufficiently large n

$$\sum_{\sigma=n+1}^{\infty} C_{2\sigma-1}e^{i(2\sigma-1)\varphi} \cong \frac{C_{2n+1}e^{i(2n+1)\varphi}}{1+k\gamma e^{i2\varphi}}, \qquad k = \left(\frac{2n+1}{2n+3}\right)^{\frac{1}{2}} \cong \frac{4n+1}{4n+3}. \tag{104}$$

Proceeding almost exactly as in Example 3, using equation (104) and the $D(\cdot)$ of equation (102), one can now obtain an approximation to the minimax error ϵ, to the error $\epsilon_T(\varphi)$ due to simply truncating the expansion of $F(e^{i\varphi})$, and to the ratio r of max $|\epsilon_T|$ and $|\epsilon|$. As in Example 3, it turns out that the minimax approximation is only a little better than the approximation by truncation, at least when n is large.

Figure 4 compares a computed $\epsilon_T(\varphi)$ with the approximate ϵ and ratio r of max $|\epsilon_T|$ to $|\epsilon|$, using

$$\theta_c = 170^\circ \qquad n = 15,$$

for which

$$\gamma = 0.8397 \qquad k = 0.96825,$$

$$\epsilon \cong 5.690^\circ \qquad r \cong 1.0840.$$

VI. COMPARISON WITH OTHER WORK

This section compares the present paper with previous publications in various related fields. It is not intended, however, to be an exhaustive survey of all related publications.

The transformation from x to φ followed by distortion of the φ scale

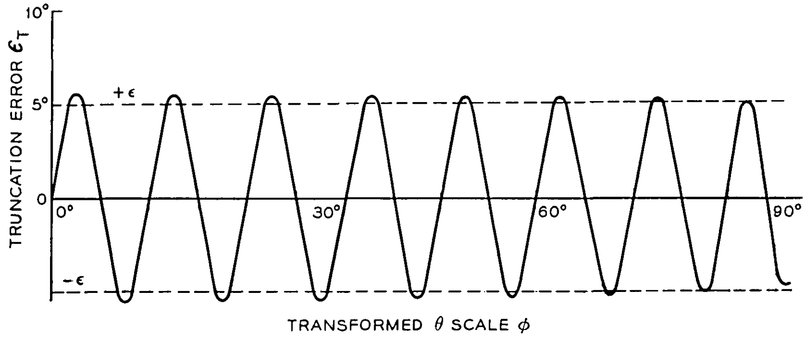


Fig. 4 — Illustrating Example 5.

to obtain an error function of the form $\epsilon \cos [(n + 1)\varphi + f(\varphi)]$ has been used before. Pertinent references are papers by Clenshaw[2] and Stiefel,[3,4] who call our $f(\varphi)$ the "phase function". Of these, Clenshaw's paper is quite close to ours, and in fact our work might be regarded as a generalization of his.

Clenshaw devotes much of his paper to the approximation of a polynomial of degree $n + \nu$ with a polynomial of degree n, which is our Example 4. Clenshaw, (Ref. 2, pp. 30, 31) solves the problem for $\nu = 2$ in a quite similar way, except that he expresses his Fourier series in terms of cosine functions instead of power series in $e^{\pm i\varphi}$. He exhibits an approximate solution which can be shown to be almost, but not quite equivalent to ours. We would have obtained an exact equivalent if we had restricted our polynomial $P(\cdot)$ to positive powers only, corresponding to $\sigma = 0$ to n in equation (25) instead of $-n$ to n. In equation (36), this restricts the disposable G_λ's to $\lambda = 1$ to $n + 1$ and increases the number of terms in $\delta(\cdot)$ to $\lambda = n + 2$ to ∞. The result is a somewhat poorer, but frequently adequate, approximation to an equal ripple error. Clenshaw also notes how his approximation can be improved, but does not fill in the details. It can be shown that the improved approximation would be an exact equivalent of ours. However, we have found that our formulations in terms of $e^{\pm i\varphi}$, instead of $\cos \varphi$, are simpler, and also more revealing concerning, for example, the nature of the approximations.

Clenshaw (Ref. 2, pp. 31–36) also considers $\nu > 2$, and obtains approximate solutions for $\nu = 3, 4$ in terms of roots of cubic and quartic equations. However, he retains the use of $\cos \varphi$, instead of $e^{\pm i\varphi}$. As a result, he does not include a formulation for a general ν in terms of an eigenvalue and eigen vector of a matrix, like our equation (91).

Clenshaw (Ref. 2, p. 29) notes that the equal ripple approximation to $(x_0 - x)^{-1}$ has been solved exactly and cites Hornecker[6] and Rivlin[7]. Any solution to this problem is easily applied to more general problems in which the given function yields the same sort of remainder when the Chebyshev polynomial series is truncated. Examples are our Example 2 and our general formulation for unweighted polynomial approximations with weight factor $W = 1, m = 1, \mu = 0$ in the remainder function (32).

Our procedures are more general in the following ways: First, remainder functions can have the more general form (32). For practical purposes, degrees m and μ should not be large. However, they need not be restricted to the special cases $m = 0, \mu \leqq 4$ and $m = 1, \mu = 0$. Second, minimax *weighted* errors can be obtained [by using suitable weight factors $\mathbf{W}(x)$ in equations (42), (44), and so on]. Third, unweighted minimax approximations can be obtained with approximating functions of which the disposable polynomial is only a part (by solving an equivalent polynomial approximation with a weighted minimax error, as in our Examples 1, 3 and 5). Finally, relatively simple formulations have been obtained by using exponentials instead of cosine functions.

Stiefel's papers[3,4] have much less relevance to our work. They use the error formulation $\epsilon \cos [(n + 1)\varphi + f(\varphi)]$ but obtain solutions by numerical iteration. Reference 3 also includes a general integral equation, which determines the required coefficients implicitly but is not easily solved.

Our use of rational fractions to approximate remainder functions, as in equation (32), and so on, is at least reminscent of the so-called ϵ algorithm. The ϵ algorithm also uses rational function approximations to remainders but for a different purpose—to increase the rate of convergence when functions are evaluated from their power series. It is quite different from the use of rational functions in the formulation of minimax polynomial approximations. References for the ϵ algorithm are Shanks[8] and Wynn.[9]

Our procedures require evaluating certain of the coefficients in the Chebyshev polynomial expansion of a given $\mathbf{F}(x)$ (or in the equivalent Fourier series expansion in terms of φ). Various established numerical methods are available for this.[1] The best choice depends on the form in which $\mathbf{F}(x)$ is specified (for example in closed analytic form, as a power series in x, or numerically at a set of discrete points). When $\mathbf{F}(x)$ satisfies a differential equation with polynomial coefficients, the coefficients in the Chebyshev polynomial series are related by a difference equation of finite order and can be computed recursively. Our Example 5 is a special case. The general relation is described by Clenshaw[10] who also includes numerical tabulations of coefficients for some common functions.

The author's 1952 paper on network synthesis in terms of Chebyshev polynomials is only remotely related to the present work.[11]

VII. CONCLUSIONS

Techniques like those described in Section IV and illustrated in Section V can be applied to many approximation problems in which the disposable part of the approximating function is a polynomial and approximately equal weighted or unweighted error extrema are desired. However, to be useful they must compete with other possible techniques, especially established numerical methods whereby equal-ripple approximations are obtained by iterative improvement of a sequence of unequal-error approximations. This section notes some circumstances under which procedures like those described here may perhaps be preferable.

First, the techniques described here are more likely to be competitive when the degree n of the disposable polynomial is large. When n is large iterative numerical methods are more likely to entail excessive amounts of computing. On the other hand, certain aspects of the more analytic techniques described here are likely to become easier as n becomes larger. These concern particularly the use of a simple rational fraction to approximate a remainder function, as in equation (32).

Second, the techniques described here are particularly suitable for exploring relationships between error amplitude $|\epsilon|$, the limits x_a, x_b of the approximation interval, the degree n of the disposable polynomial, and other parameters in the approximating function (such as x_0 in examples 1, 2, and 3). In explorations of this sort the computation of the actual coefficients of the disposable polynomial $\mathbf{P}(x)$ can usually be omitted. When n is large this can mean omitting most of the computations required for a complete determination of the approximating function. Frequently, computations which end with $|\epsilon|$ remain very simple even though n becomes arbitrarily large.

Third, sometimes, as in our Examples 1, 2, 3 and 5, our techniques give quite simple estimates of the advantage of an equal-ripple approximation over simple truncation of an infinite series of Chebyshev polynomials. Such a comparison may be useful, for example, in deciding what sort of approximation should be computed in detail.

More generally, an attractive combination may be an initial exploration in terms of the techniques described here, followed by the detailed computation of one or more preferred cases by established iterative numerical methods.

We have assumed here that the parameters disposable for purposes of approximation are all the coefficients in a *polynomial.* Preliminary investigation indicates that similar methods may be feasible for disposable rational functions, or ratios of polynomials, provided the polynomials in the denominators are of quite modest degree. This will be the subject of a later paper.

VIII. ACKNOWLEDGMENT

The author is indebted to M. J. D. Powell of the Atomic Energy Research Establishment, Harwell, England for the more important references (including references suggested by Powell and references in the suggested references).

REFERENCES

1. Snyder, M. A., *Chebyshev Methods in Numerical Approximation,* Englewood Cliffs, New Jersey: Prentice-Hall, 1966, pp. 114.
2. Clenshaw, C. W., "A Comparison of 'Best' Polynomial Approximations with Truncated Chebyshev Series Expansions," J. SIAM Numerical Analysis Series B, *1* (1964), pp. 26–37.
3. Stiefel, E. L., "Methods—Old and New—for Solving the Tchebycheff Approximation Problem," J. SIAM Numerical Analysis Series B, *1* (1964), pp. 164–176.
4. Stiefel, E. L., "Phase Methods for Polynomial Approximation of Functions," *Proceedings of the Symposium on Approximation of Functions,* General Motors Res. Laboratories, New York: Elsevier, 1965, pp. 68–82.
5. Darlington, S., "Synthesis of Reactance 4-Poles," J. of Math. and Phys., *18,* No. 4 (September 1939), pp. 257–353.
6. Hornecker, G., "Evaluation Approcheé de la Meilleure Approximation Polynomiale d'Ordre n de f(x)sur un Segment Fini (a,b)," Chiffre, *1* (1958), pp. 157–169.
7. Rivlin, T. J., "Čebyšev Expansions and Best Uniform Approximations," I.B.M. Res. Rep. R2-93, 1962.
8. Shanks, D., "Non-linear Transformations of Divergent and Slowly Convergent Series," J. Math. and Phys., *34,* No. 1 (April 1955), pp. 1–42.
9. Wynn, P., "On a Device for Computing the $\epsilon_n(S_n)$ Transformation," Math. Tables and other Aids to Computation, *10,* No. 54 (April 1956), pp. 91–96.
10. Clenshaw, C. W., "Chebyshev Series for Mathematical Functions," Nat. Phys. Laboratory Math. Tables, *5,* London: Her Majesty's Stationery Office, 1962, pp. 1–36.
11. Darlington, S., "Network Synthesis using Tchebycheff Polynomial Series," B.S.T.J., *31,* No. 4 (July 1952), pp. 613–665.

Multilevel Modulation Techniques for Millimeter Guided Waves

By W. M. HUBBARD, G. D. MANDEVILLE, and J. E. GOELL

(Manuscript received July 12, 1969)

This paper describes an investigation of the feasibility of increasing the information transmission capacity of a guided millimeter wave communication system by using quaternary and higher order modulation techniques in place of binary. It first presents a generalization of the binary system to higher orders and then extends the results of previously derived error-rate predictions. An experimental repeater for quaternary modulation which uses components that are similar to those used for binary modulation is then described along with associated equipment used for signal generation and performance evaluation. Finally, performance data on the repeater are given and compared with theory.

I. INTRODUCTION

The quaternary and higher-order modulation techniques which are described are extensions of the binary modulation technique previously discussed by the authors.[1] The earlier system uses binary differentially-coherent phase-shift-keyed (B-FMDCPSK) modulation, a form of modulation in which the frequency of the carrier is increased or decreased once during each time slot in such a manner that the phase (the time integral of the frequency shift) is changed by ±90° relative to the phase of the previous time slot. Experimental repeaters were built which could regenerate, with a 10^{-9} error rate, a signal which had been attenuated by an amount equivalent to 15 miles of waveguide. In this repeater the incoming millimeter-wave signals from circular waveguide were passed through band- and channel-demultiplexing filters, down-converted, amplified, differentially phase-detected, and regenerated (utilizing timing information self-contained in the signal) to obtain a polar baseband representation of the information. This baseband signal was used to drive an IF voltage-tuned oscillator, the output of which was amplified, up-converted, passed through channel- and band-multiplexing filters and launched into circular waveguide.

The modulation scheme of the system can be generalized in such a way that systems with $M = 2^m$ levels (quaternary, octonary, and so on) can be built with components most of which are the same as those used in the binary system. With such a system, the information capacity of the wave-guide can be increased by a factor m by using a 2^m level signal in each of the individual channels. The increase in system capacity is accompanied by a decrease in immunity to noise and system degradation (which is common to all multilevel systems) and a slight increase in system complexity.

A theoretical consideration of multilevel systems of all orders is given in Section II. This consideration amounts to an extension of previous error-rate calculations for binary systems.[2] Section III describes a quaternary (Q-FMDCPSK) system which has been built and operated at 320 megabits/s.

Because the theoretical portion of this paper is a direct extension of a previous paper on a binary system it is assumed that the reader is familiar with the contents of that paper.[2] However, references to the binary paper are made where appropriate as an aid to the reader.

II. THEORETICAL CONSIDERATION OF MULTILEVEL SYSTEMS

2.1 *Description of the Quaternary System*

The signal consists of a constant-amplitude angle-modulated carrier. The modulation is achieved by causing a frequency deviation once in each time slot. For the binary case the frequency deviation $\omega(t)$ satisfies the condition

$$\int_{(n-\frac{1}{2})T}^{(n+\frac{1}{2})T} \omega(t')\,dt' = \alpha_n \tag{1}$$

in the nth time slot, where $\alpha_n = \pm\pi/2$ and contains the binary information. This signal can be written in the form

$$s(t) = \cos\left[\omega_0 t + \int_0^t \omega(t')\,dt'\right] \tag{2}$$

where ω_0 is the center frequency about which the signal is deviated. Equations (1) and (2) hold for the quaternary signal as well. The only difference is that now α_n can take on any of the four values $\pm\pi/4, \pm 3\pi/4$.

The signal space diagrams of these signals are shown in Fig. 1. The states marked "$\times$" represent the phase states which are available to

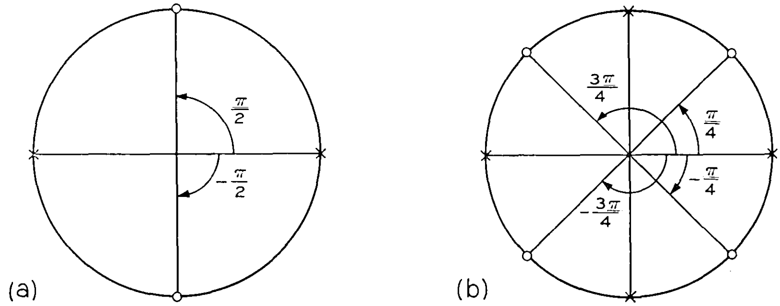

$\frac{\pi}{2}$
$-\frac{\pi}{2}$
(a)
$\frac{3\pi}{4}$
$\frac{\pi}{4}$
$-\frac{\pi}{4}$
$-\frac{3\pi}{4}$
(b)

Fig. 1 — Signal space diagrams for DC phase-shift-keyed signals: (a) binary, (b) quaternary. X represents phase states available in even number time slots and "O" represents the phase states available in odd numbered time slots.

the signal in the even numbered time slots and those marked "O" represent the states which are available in odd numbered time slots. Thus the transition always takes place from a state marked X to a state marked O or vice versa.

Bennett and Davey describe the detection scheme for DCPSK modulation.[3] A particular embodiment of the differential phase detector for the binary signal of Fig. 1a is shown in Fig. 2. Here the relative delay between the two paths is T where T satisfies simultaneously the two constraints

$$\frac{1}{T} = \text{Baud rate}$$

$$\omega_0 T = (n + \tfrac{1}{2})\pi \quad n \text{ an integer.}$$

Figure 3 illustrates how this differential phase detection concept is

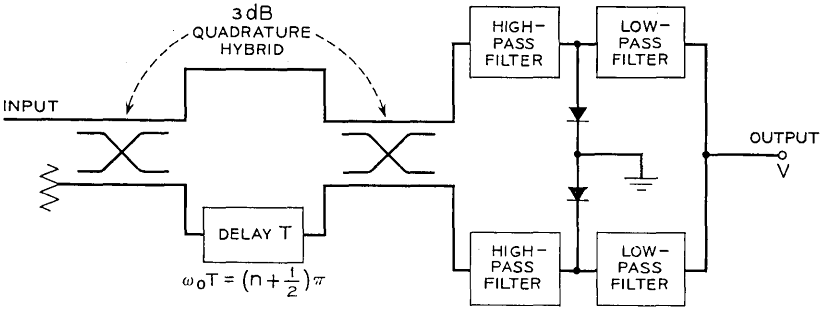

3dB
QUADRATURE
HYBRID
INPUT
HIGH-
PASS
FILTER
LOW-
PASS
FILTER
OUTPUT
V
DELAY T
$\omega_0 T = (n+\frac{1}{2})\pi$
HIGH-
PASS
FILTER
LOW-
PASS
FILTER

Fig. 2 — Binary differential phase detector.

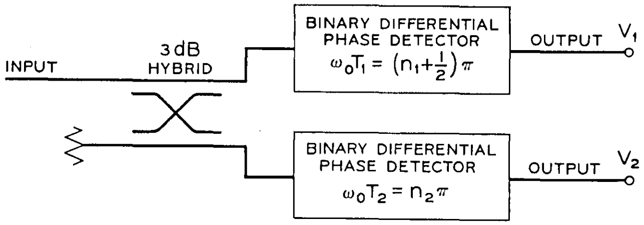

INPUT
3dB HYBRID
BINARY DIFFERENTIAL PHASE DETECTOR
$\omega_0 T_1 = (n_1+\frac{1}{2})\pi$
OUTPUT
V_1
BINARY DIFFERENTIAL PHASE DETECTOR
$\omega_0 T_2 = n_2\pi$
OUTPUT
V_2

Fig. 3 — Quaternary differential phase detector.

extended to the quaternary case. In this case the preceding constraints become

$$\frac{1}{T_1} \approx \frac{1}{T_2} \approx \text{Baud rate}$$

$$\begin{aligned} \omega_0 T_1 &= (n_1 + \tfrac{1}{2})\pi \\ \omega_0 T_2 &= n_2 \pi \end{aligned} \qquad n_1\,, n_2 \quad \text{integers.}$$

If one applies the signal described by equations (1) and (2) to the device illustrated in Fig. 3, he finds that the outputs V_1 and V_2 are given by Table I for the four possible phase changes, α. Thus the binary variable V_1 defines the sign of α while V_2 defines its magnitude. Stated another way, given that the phase state in the $(n - 1)$th time slot was at 0 radians in Fig. 1b, V_1 determines in which half plane (upper or lower) the phase state of the nth time slot lies while V_2 determines in which half plane (left or right) it lies.

Since each branch of the quaternary differential phase detector is identical to the binary device described in Ref. 2 the limiter and the regenerators can also be identical to those described in Sections III and 1.4 of Ref. 2.

For the multilevel system, only one device which is not a direct adaptation of an existing component of the binary system is required. Its function is to translate the regenerated binary signals into a signal

TABLE I — OUTPUTS FOR FOUR POSSIBLE PHASE CHANGES

α	V_1	V_2
$\pi/4$	1	1
$-\pi/4$	-1	1
$3\pi/4$	1	-1
$-3\pi/4$	-1	-1

suitable for driving the FM deviator. Its performance should be virtually error free and need not be considered in the error-rate calculations. A description of this translator is deferred to Section III.

2.2 *Extension of the Error-Rate Calculation*

The bit-error probability of the quaternary system described in Section 2.1 is equal to the probability that an error is made in one of the baseband binary sub-channels. This, however, is just the probability that an error is made in a binary differentially-coherent phase-shift-keyed system in which the expectation value of the second pulse is shifted an amount $\pi/4$ from where it should be for binary operation. (This phase shift from the desired binary-system value is represented by the quantity β_0 in Section III of Ref. 2.) The probability of bit error, Π can be restated for the quaternary case as*

$$\Pi = \tfrac{1}{2}P_0\left(\phi + \frac{\pi}{4} + \delta\right) + \tfrac{1}{2}P_0\left(\phi - \frac{\pi}{4} - \delta\right) \tag{3}$$

where

$$P_0(\Phi) = \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} \exp\left[-\frac{\cos^2 \Phi}{2\sigma^2(1 + \sin \Phi \sin \theta)}\right] d\theta. \tag{4}$$

An approximate solution (in closed form) to this integral is derived in the Appendix. Here the quantities ϕ and δ have the same meaning as in Reference 2, namely, δ is the phase shift due to intersymbol interference and any other phase distortion in the system, and $\phi = \sin^{-1} \epsilon$ where ϵ is given by $S/T = -10 \log \epsilon$† and S/T is the signal-to-threshold ratio of the regenerator in decibels. S/T is defined in Section 1.4 of Ref. 1 as the ratio of the expected value of signal power to the minimum value of signal power which will cause the regenerator to function realiably (in the absence of noise).

Values of P_0 for Φ from 0 to 30° are given in Ref. 2 for signal-to-noise ratios $S/N = 9$ through 15 dB. These results are extended in Fig. 4 to include values suitable for quaternary and higher level systems. Note that $P_0(\Phi)$ is even. Figure 5 shows error rate as a function of S/N for an ideal quaternary system.

The effects of finite S/T and δ are more pronounced for quaternary systems than for binary. The threshold effect noise figure N_T defined in Ref. 4 is shown in Fig. 6 for a quaternary system and for a binary

---

* Equations (3) and (4) follow directly from Equations (20) and (21) of Reference 2 by replacing δ with $\delta + \pi/4$.

† The equation relating S/T and ϵ in Ref. 2 is incorrect. The conclusions of Ref. 2 are not affected by this as the correct form of the relation was used in the calculations.

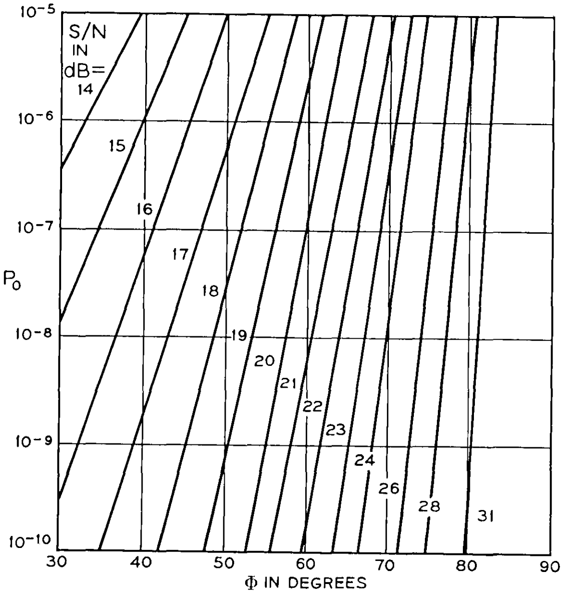


Fig. 4 — Error integral versus Φ for various values of S/N.

system. N_T is the amount by which S/N must be improved in order to offset the effect of a nonideal regenerator.

The effects of δ and S/T are determined directly from Fig. 7 which shows the value of S/N required for 10^{-9}, 10^{-8}, 10^{-6}, and 10^{-4} error-rate as a function of Φ. For comparison we consider the values of S/T and δ which were inferred from the measurements made on the binary repeater described in Ref. 1, namely $S/T = 10$ dB, $\delta = 10°$. For the binary case the combined effect of these degradations is about 0.8 dB whereas for the quaternary case the combined effect is about 3.4 dB. The theoretically predicted values* of S/N for operation with an error probability of 10^{-9} are therefore 13.8 dB and 21.3 dB for binary and quaternary, respectively (with half the bandwidth requirement in a quaternary system with the same bit rate). (The experimentally determined value for the binary system is 13.7 dB.)[1]

2.3 *Extension to Higher Level Systems*

In a system with 2^m levels, the signal described in Section 2.1 could be generalized to have 2^m equally spaced positions around the unit

---

* This value does not include degradation introduced into the quaternary system due to nonlinearity of the FM deviator. Unlike the binary case, this nonlinearity is important in the quaternary and higher order cases.

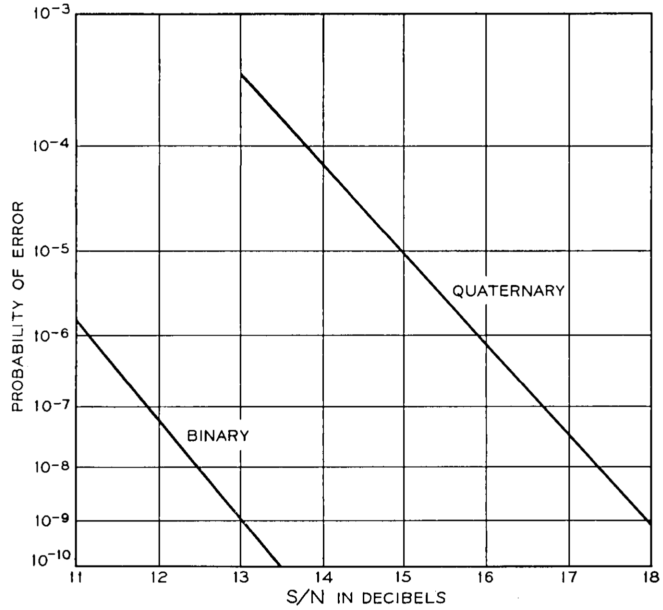


Fig. 5 — Probability of error in ideal binary and quaternary systems.

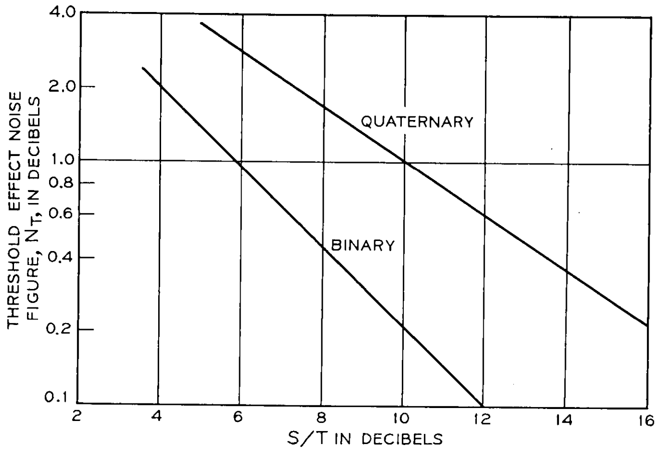


Fig. 6 — Threshold effect noise figures as a function of signal-to-threshold ratio, S/T.

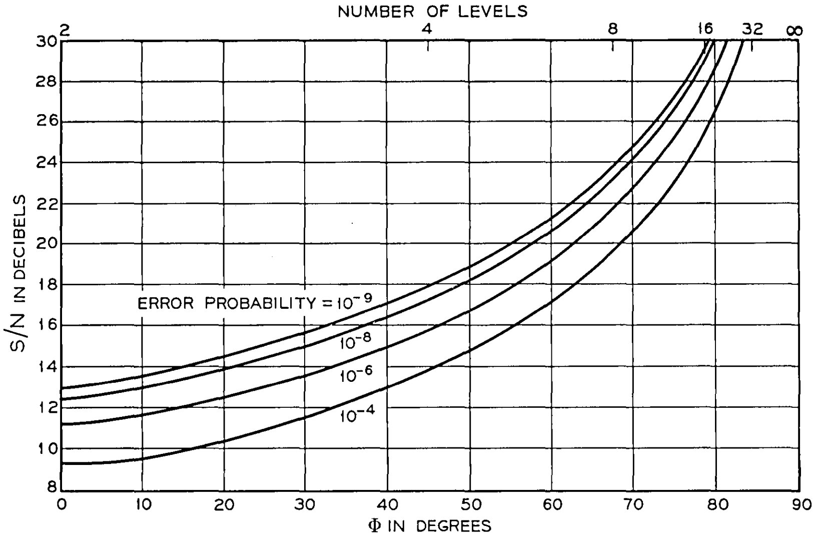


Fig. 7 — Required signal-to-noise ratio for various values of error probability as a function of Φ.

circle in the even numbered time slots. The differential phase detector would then require m of the binary differential phase detectors of Fig. 2, with delay lines T_1, $T_2 \cdots T_m$ such that the values of $\omega_0 T_k$ are chosen to give 2^m equally spaced values around the unit circle beginning at $\pi/2^m$, that is,

$$\omega_0 T_k = \frac{(2k+1)\pi}{2^m}, \qquad k = 0, 1, \cdots, 2^m - 1.$$

In systems with more than four levels this method of detection gives more than one level of pulse height at the regenerator. The following consideration makes the approximation that the worst-case error-rate applies in all cases. This results in a calculated error-rate which is too large by a factor which is less than the ratio $M/(M-4)$ for an M level system ($M > 4$).

Equation (3) then becomes:

$$\Pi = \tfrac{1}{2}P_0\left(\phi + \delta + \frac{\pi}{2} - \frac{\pi}{M}\right) + \tfrac{1}{2}P_0\left(\phi - \delta - \frac{\pi}{2} + \frac{\pi}{M}\right) \qquad (5)$$

and equation (4) is unchanged. Figure 4 gives values of $P_0(\Phi)$ suitable for evaluating Π for values of M up to 16. Figure 7 indicates the values of S/N required for 10^{-9}, 10^{-8}, 10^{-6} and 10^{-4} error-rate for ideal 2^m level systems for $m = 1$ through 5.

2.4 *Results of the Calculations*

In an ideal repeater the signal-to-noise ratio for an error rate of 10^{-9} is 13.0 and 17.9 dB for binary and quaternary signal, respectively. This amounts to a price of 4.9 dB for the doubling of the bit rate (or alternatively the halving of the bandwidth) achieved by quaternary systems. In an actual repeater with intersymbol interference and nonideal regeneration comparable to that in the 306 megabits/s binary repeater of Ref. 1 there is an additional degradation of about 3.4 dB (compared with 0.8 dB for binary signals). Thus a signal-to-noise ratio of 21.3 dB is expected to be necessary for a 10^{-9} error rate in the quaternary repeater—a penalty of 7.5 dB compared with the binary repeater. For a guided-wave system of the sort described in Ref. 1 this requires only a 2.5 mile decrease in repeater spacing (about 10 to 15 percent) which might not be an unattractive price for doubling the channel capacity of the system.

For systems with more than four levels, the degradation in error-rate performance due to S/T and δ is even more severe. For eight levels ($m = 3$) for example, Φ becomes 83.2° (for the worst case) for $S/T = 10$ dB and $\delta = 10°$ and the degradation is (from Fig. 4) intolerable. Clearly an improvement in S/T and a substantial improvement in δ is necessary in order to make systems with more than four levels feasible. Even for ideal systems ($S/T = \infty$, $\delta = 0$), signal-to-noise ratios of about 23.7 dB and 29.7 dB are required for 10^{-9} error-rate for eight and sixteen level systems respectively compared with 13.0 dB and 17.9 dB for two and four level systems.

III. EXPERIMENTAL RESULTS FOR QUATERNARY REPEATERS

3.1 *Description of the Experimental Repeater*

The quaternary experimental system, shown in Fig. 8 is similar to the binary system described in Ref. 1 with the following exceptions.

(*i*) Where one differential phase detector and one regenerator were used in the binary, two are required. In addition, a binary device called a translator is needed.

(*ii*) Conversion into and out of the millimeter medium was omitted.

(*iii*) Modulation of a deviator by the regenerated baseband signal was not attempted.*

---

* The technique for doing this, however, is identical to the technique used to synthesize the signal from the random word generator and should present no additional problems.

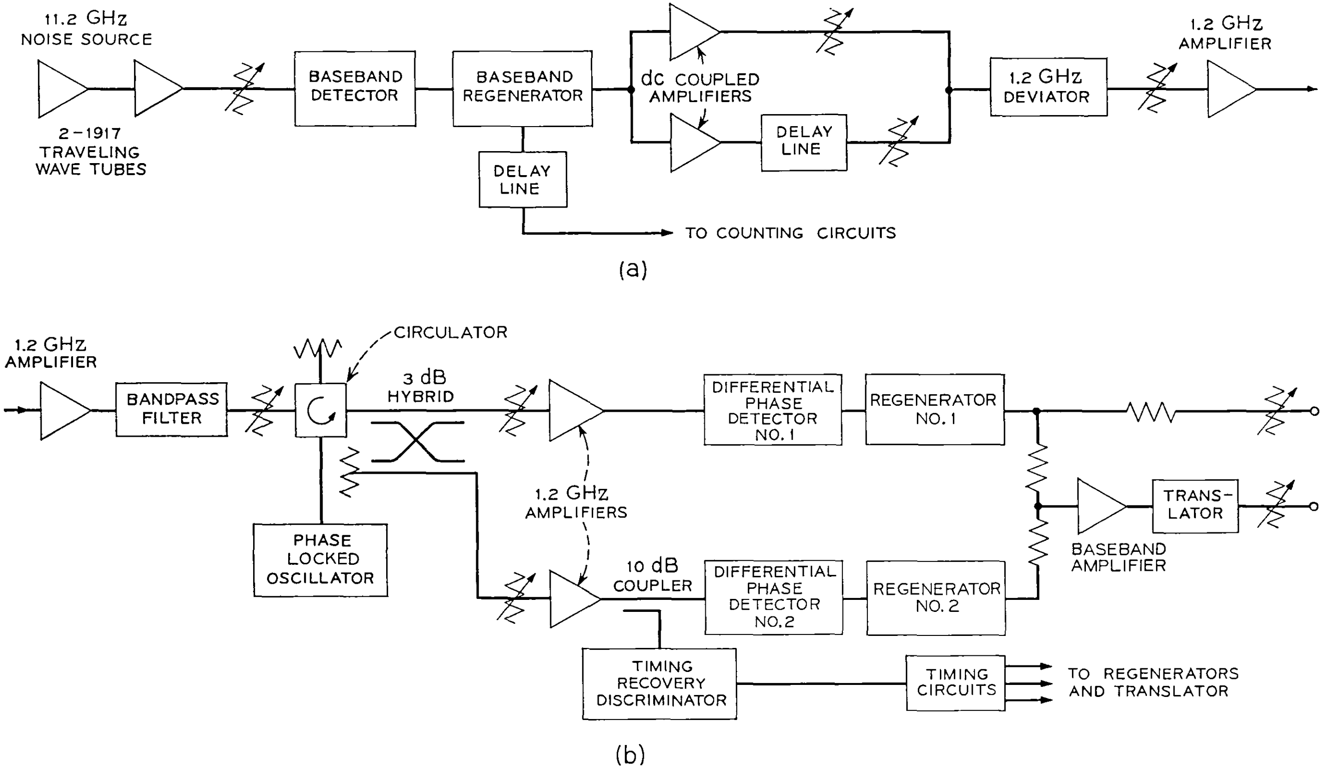

11.2 GHz
NOISE SOURCE
2-1917
TRAVELING
WAVE TUBES
BASEBAND
DETECTOR
BASEBAND
REGENERATOR
DELAY
LINE
TO COUNTING CIRCUITS
dc COUPLED
AMPLIFIERS
DELAY
LINE
1.2 GHz
DEVIATOR
1.2 GHz
AMPLIFIER
(a)
1.2 GHz
AMPLIFIER
BANDPASS
FILTER
CIRCULATOR
3 dB
HYBRID
PHASE
LOCKED
OSCILLATOR
1.2 GHz
AMPLIFIERS
10 dB
COUPLER
DIFFERENTIAL
PHASE
DETECTOR
NO. 1
REGENERATOR
NO. 1
DIFFERENTIAL
PHASE
DETECTOR
NO. 2
REGENERATOR
NO. 2
TRANS-
LATOR
BASEBAND
AMPLIFIER
TIMING
RECOVERY
DISCRIMINATOR
TIMING
CIRCUITS
TO REGENERATORS
AND TRANSLATOR
(b)

Fig. 8 — Block diagram of quaternary experiment. (a) Transmitter (word generator and IF phase modulator) and (b) Receiver (IF amplifier; limiter, phase detector, and regenerator).

(*iv*) A symbol (baud) rate of 160 MHz (320 megabits/s) was used.

In a quaternary FM phase-shift-keyed (Q-FMDCPSK) signal the information is contained in the phase shift between adjacent time slots. The phase shifts used are $\pm 3\pi/4$ and $\pm\pi/4$. The signal is generated by a voltage-controlled oscillator whose frequency is pulsed between sampling instants so that the time integral of the frequency shift is equal to the desired phase shift. An illustration of such a signal is shown in Fig. 9.

A four-level baseband pulse train is derived from a two-level polar source identical to the one used for the binary experiment[1] except for the rate which in the quaternary experiment was chosen to be 160 MHz. The random binary output signal is divided into two signals. One of the signals is delayed a few integral time slots and the other is attenuated 6 dB. They are then recombined. The combination of $+2$, -2 with $+1$, -1 pulses produces a pulse of one of four levels in each time slot (see Fig. 10a). These are $+3$, $+1$, -1, and -3, which produce phase shifts in the deviator $3\pi/4$, $\pi/4$, $-\pi/4$, and $-3\pi/4$, respectively. By delaying one signal a few time intervals, we closely approximate the effect of two independently random binary signals and thus obtain a virtually random four-level signal. Observation of the spectrum displayed by a spectrum analyzer verifies the randomness (see Fig. 10d).

Identifying the larger binary component as signal #1 and the smaller as signal #2 will help clarify the explanation of the regeneration process which follows. For the binary system, the signal was detected by a differential phase detector, the output of which is given by

$$\cos\left[\phi(t) - \phi(t-\tau) + \omega_0\tau\right]$$

where ϕ is the phase angle of the signal, τ is the time delay introduced by a delay line built into the device, and ω_0 is the angular carrier frequency. For the binary case, τ was made equal to the bit interval and $\omega_0\tau$ equal to $\pi/2 + n\pi$. The operation of the differential phase detector is illustrated in Fig. 11. The reference phase is taken as the phase of the signal in the previous time slot. The information given by the device is the projection E_v of the signal S along the vertical axis. The two phase transitions of the binary case, $\pm\pi/2$, are fully determined by the sign of E_v.

For a quaternary signal, it is not possible to distinguish between transition into regions I and II or those into III and IV using only E_v. This problem was solved by splitting the IF signal into two portions, one of which was connected to a differential phase detector with $\tau = T_1$

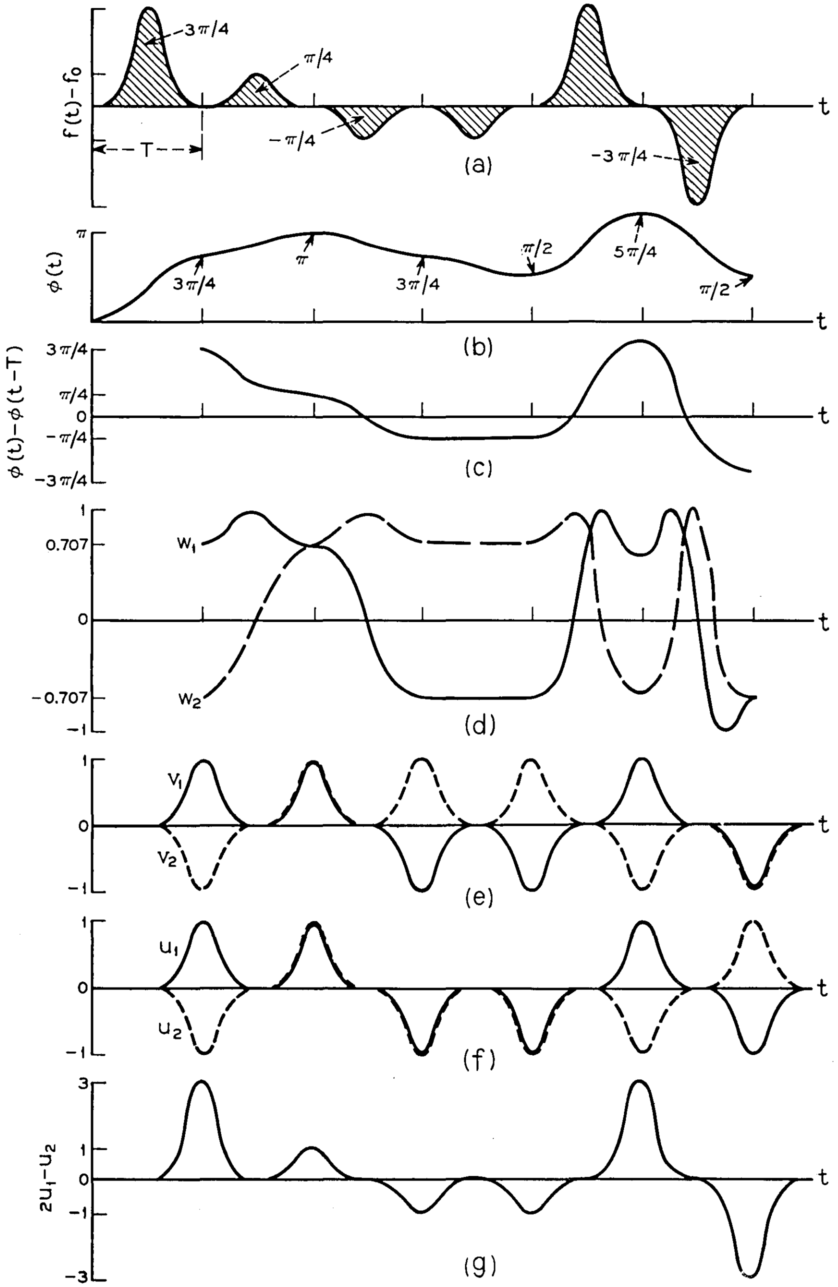


Fig. 9 — Representative quaternary frequency-modulation differentially-coherent phase-shift-keyed signal.

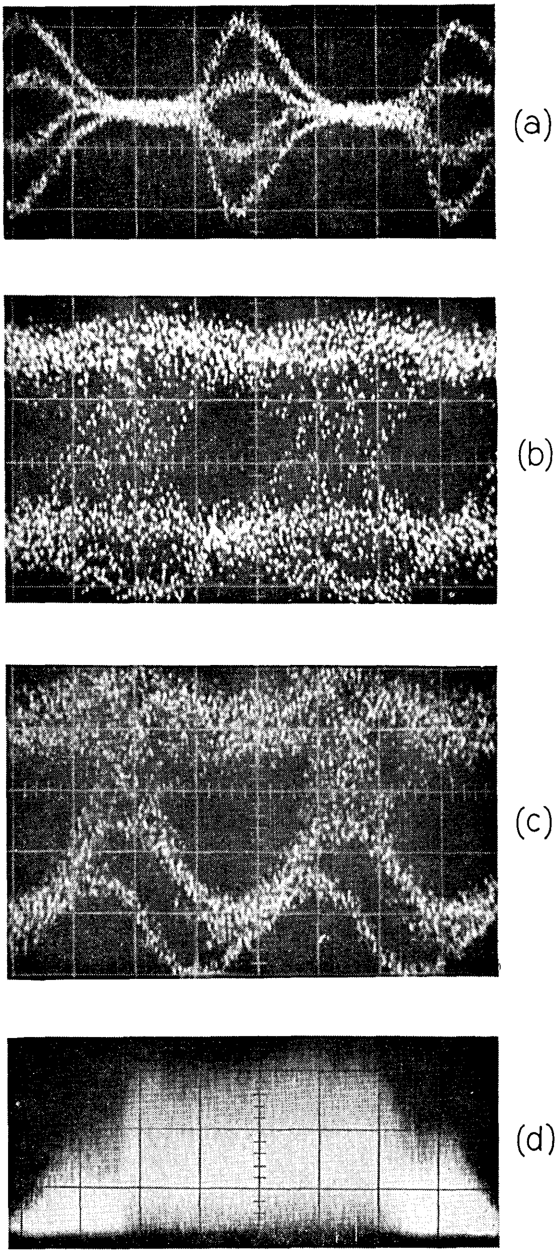


Fig. 10 — (a) Random four-level base-band pulses ($H = 2$ ns/cm); (b) symmetric eye ($H = 2$ ns/cm); (c) asymmetric eye ($H = 2$ ns/cm); and (d) IF spectrum ($H = 30$ MHz/cm).

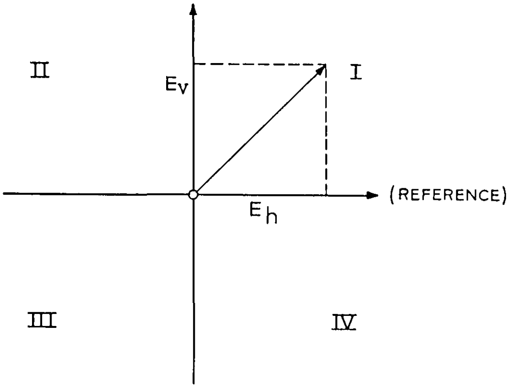


Fig. 11 — Operation of the differential phase detector.

and the other was connected to a differential phase detector with $\tau = T_2$. The quantities ω_0, T_1, and T_2 were chosen so that

$$\omega_0 T_1 = 16.5\pi \quad \text{and} \quad \omega_0 T_2 = 16\pi$$

in order to satisfy the conditions set forth in Section II. Thus the output of the first differential phase detector gives E_v, and the output of the second E_h. From the signs of E_h and E_v the quadrant of the signal can be determined as shown in Table II. Figures 10(b) and (c) are photographs of sampling oscilloscope displays of the two phase-detector outputs of a typical random signal. Because signal state #2 does not correspond to either E_v or E_h a translator is required. If E_h and E_v are regenerated so that they have unit magnitude, then signal #1 is given by E_v and signal #2 is given by the negative of the product, $E_v E_h$. The sign of the product $E_v E_h$ was determined by the translator circuit shown in Fig. 12. This circuit is similar to the balanced-line logic element used in the binary regenerators except for the input circuit; it functions as follows. Without input, once each time slot either diode A or diode B must switch into its high-voltage state while the other diode remains in its low-voltage state. Applying positive bias causes diode B to switch giving positive output pulses while applying negative bias produces the opposite effect. For the translator circuit the bias is set so that diode B always switches into the high-voltage state in the absence of input. The input to the circuit is the sum of the regenerated outputs of the two differential phase detectors E_v and E_h. If the sum of the two signals is zero, diode B switches so the output pulse is positive. If the sum of the two signals is either positive or negative diode A switches and a

TABLE II — DETERMINATION OF THE QUADRANT OF THE SIGNAL

Quadrant	Phase Shift at Transmitter	E_v	E_h	Original Signal #1	Original Signal #2
I	$\pi/4$	+	+	+	−
II	$3\pi/4$	+	−	+	+
III	$-3\pi/4$	−	−	−	−
IV	$-\pi/4$	−	+	−	+

negative output results, since for either polarity, current flowing through the steering diodes, D_1 and D_2 , increases the current through diode B. Thus the translator output is equivalent to the product of the regenerated E_v and E_h signals.

3.2 *Results*

The error rate was measured as in the binary experiment, except that the two binary components of the regenerated quaternary baseband signal were compared separately with their properly delayed counterparts comprising the input baseband signal.

The results of the error-rate measurements are shown in Fig. 13. Curve 1 shows the error-rate performance which was obtained when the equipment was originally built. At high error-rates (above about 5×10^{-6}) the performance was in good agreement with the theoretical

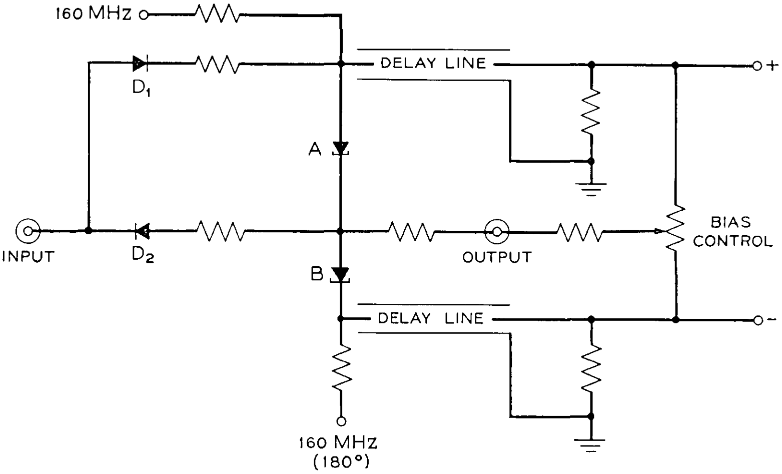


Fig. 12 — Schematic of translator circuit.

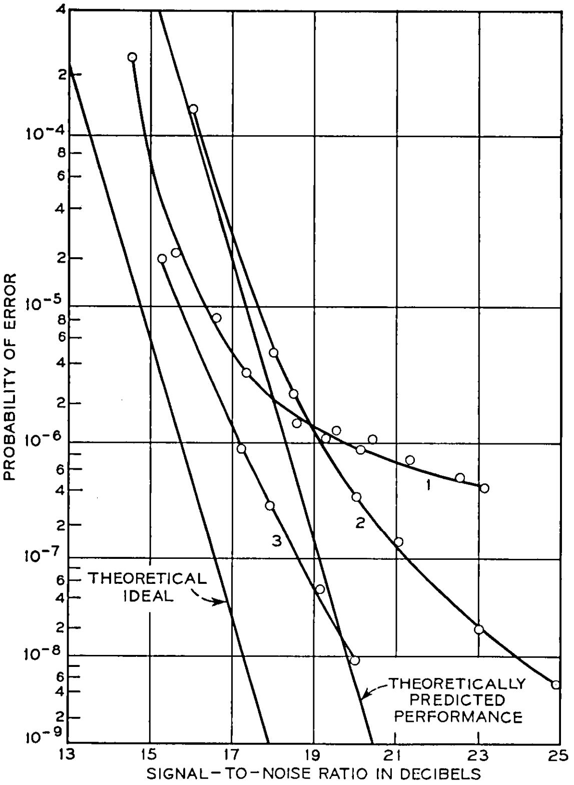


Fig. 13 — Results of error-rate measurements.

predictions. There was, however, a "floor" at an error-rate of about 4×10^{-7} and no increase in the S/N would reduce the error-rate below this value. It should be noted that the degradations considered in Section II shift the error-rate line to larger S/N and decrease the slope slightly (in magnitude) but do not tend to establish a "floor" such as the one shown by curve 1. It might also be mentioned that a "floor" similar to this is characteristic of experimental error-rate curves for the binary repeater of Ref. 1, but there it occurred typically at error-rates below 10^{-10} and was therefore considered insignificant.

Thus the indication was that the "floor" was the result of some degradation which was neglected in the theory and to which the four-level system was far more sensitive than was the two-level system. One possibility was that the impairment was due to slight mismatches among the commercial components. Improved performance was achieved by careful selection of components with particular emphasis on the linearity of the deviator. Curves 2 and 3 of Fig. 13 show the typical and best performance, respectively. The "floor" was at 8×10^{-10} when curve 3 was observed. Thus the agreement between theory and experiment is fairly good at error-rates substantially above the experimentally observed floor.

3.3 *Conclusions*

The quaternary experiment was restricted to the investigation of the modulation and regeneration aspects of a repeater system. However, the data obtained, coupled with the results from the previous binary experiment, suggest that a quaternary (Q-FMDCPSK) repeater system is feasible and might have applications where the conservation of bandwidth is desirable and the cost in terms of noise immunity can be afforded. Systems with eight or more levels do not seem feasible at the present time.

APPENDIX

Evaluation of the Error-Rate Integral

The integral $P_0(\Phi)$ [equation (4)] or its equivalent

$$P_0 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left[-(x - p_x)^2 - (y - p_y)^2\right] \cdot \operatorname{erfc}\left[\frac{xq_x + yq_y}{(x^2 + y^2)^{\frac{1}{2}}}\right] dx\, dy \tag{6}$$

is frequently encountered in error-rate calculations for digital phase and frequency modulated signals.[1,2,4,5] The equivalence of these two forms can be shown by arguments similar to those of Ref. 5. Namely, the integral in the form of equation (6) is written in polar coordinates and the integration over r is performed to give

$$P_0 = \frac{1}{2\pi} \int_0^{2\pi} \exp\left(-p^2 \sin^2 \phi\right) \operatorname{erfc}\left[q \cos\left(\phi + \gamma\right)\right] \cdot \left[\tfrac{1}{2} \exp\left(-p^2 \cos^2 \phi\right) + p \cos \phi \frac{\pi^{\frac{1}{2}}}{2} \operatorname{erfc}\left(-p \cos \phi\right)\right] d\phi$$

where

$$p = (p_x^2 + p_y^2)^{\frac{1}{2}}, \qquad \gamma = \operatorname{atn} \frac{q_y}{q_x} - \operatorname{atn} \frac{p_y}{p_x}, \qquad q = (q_x^2 + q_y^2)^{\frac{1}{2}}.$$

This integral can be greatly simplified by writing the error function compliments in terms of their respective error functions and making use of the odd parity of the error function. When this is done the only nonvanishing terms are (after some simplification)

$$P_0 = \frac{1}{2} - \frac{1}{\pi} \int_{-p}^{p} \int_{0}^{(p/q)[(p^2-x^2)^{\frac{1}{2}} \cos \gamma - x \sin \gamma]} \exp(-x^2 - y^2)\, dx\, dy. \quad (7)$$

The limits of integration in equation (7) form half of an ellipse. Because of the spherical symmetry of the integrand we can pick any half of this ellipse. Thus the integral can be written

$$P_0 = \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} \int_{0}^{R(\phi)} \exp(-r^2) r\, dr\, d\phi$$

where (after some simplification) one finds

$$R(\phi) = \left[\frac{S \cos^2 \Phi}{1 + \sin \Phi \sin 2\phi}\right]^{\frac{1}{2}}$$

and

$$S = \frac{p^2 + q^2}{2} \qquad \cos \Phi = 2\frac{qp \cos \gamma}{p^2 + q^2}.$$

Performing the radial integration gives

$$P_0(\Phi) = \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} \exp\left(-\frac{S \cos^2 \Phi}{1 + \sin \Phi \sin \theta}\right) d\theta$$

which is equation (4) with the substitution

$$S = \frac{1}{2\sigma^2}.$$

As a first step in finding an approximate solution to $P_0(\Phi)$ we can write

$$P(\Phi) = \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} \exp[-S \cos^2 \Phi(1 - \sin \Phi \sin \theta)]\, d\theta.$$

But $P(\Phi)$ can be integrated to give

$$P(\Phi) = \tfrac{1}{2} \exp(-S \cos^2 \Phi) I_0(S \cos^2 \Phi \sin \Phi)$$

where I_0 is the modified Bessel Function of the first kind. Now $P(\Phi)$ must be a good approximation to $P_0(\Phi)$ for sufficiently small Φ. In fact, for $\Phi = 0$ and $\Phi = \pm\pi/2$ we have

$$P(0) = P_0(0) = \tfrac{1}{2} \exp(-S)$$

and

$$P\left(\pm\frac{\pi}{2}\right) = P_0\left(\pm\frac{\pi}{2}\right) = \tfrac{1}{2},$$

respectively. Thus $P(\phi)$ and $P_0(\phi)$ agree at both extremes.

Let $\mathcal{J}(P)$ be the integrand of the integral for P and $\mathcal{J}(P_0)$ the integrand of that for P_0. Their ratio is

$$R(\Phi, \theta) = \frac{\mathcal{J}(P)}{\mathcal{J}(P_0)} = \exp\left\{\frac{S}{4}\cdot\frac{\sin^2(2\Phi)\sin^2\theta}{1+\sin\Phi\sin\theta}\right\}.$$

We observe that $R \geqq 1$ everywhere. In the following argument we assume for clarity that $\Phi > 0$; but it can easily be verified that similar arguments can be constructed for the case $\Phi < 0$ and the same results obtained.

$R(\Phi, \theta)$ has an absolute maximum R_{AM} at $\theta = -\pi/2$ and a relative maximum R_{RM} at $\theta = \pi/2$

$$R_{AM} = \exp\left\{\frac{S}{4}\cdot\frac{\sin^2 2\Phi}{1-|\sin\Phi|}\right\}$$

$$R_{RM} = \exp\left\{\frac{S}{4}\cdot\frac{\sin^2 2\Phi}{1+|\sin\Phi|}\right\}.$$

Thus we have rigorous bounds

$$P(\Phi) \geqq P_0(\Phi) \geqq P(\Phi)/R_{AM}.$$

Unfortunately these bounds are often too loose to be of much value.

Inspection of the integral for P_0 reveals, however, that almost none of the contribution to the integral comes from the region near $\theta = -\pi/2$ and in fact almost all of the contribution comes from the region near $\theta = \pi/2$ where $R(\Phi, \theta) \approx R_{RM}$. This suggests that we consider the expression

$$\begin{aligned} P_1(\Phi) &= P(\Phi)/R_{RM} \\ &= \tfrac{1}{2}\exp\left\{-S\left(1-\sin^3\Phi\right)\right\} I_0(S\cos^2\Phi\sin\Phi) \end{aligned}$$

as an approximation to $P_0(\Phi)$. Note that $P_1(\Phi)$ also possesses the property that

$$P_1(0) = P_0(0), \qquad P_1(\pm\pi/2) = P_0(\pm\pi/2).$$

Finally, one finds empirically that a somewhat better approximation is given by

$$P_2(\Phi) = P_1(\Phi)[1 + \tfrac{1}{2}\,|\sin(2\Phi)|].$$

It has been verified by numerical calculations of P_0 that for values of S and ϕ which give $10^{-11} < P_0 < 10^{-3}$ the accuracy of the approximation is better than 10 percent for $\Phi < 50°$ and remains better than 36 percent up to $\Phi = 80°$. In this range of error rates a variation of 40 percent in P corresponds to a variation of a few tenths of a decible in S/N. This form is most useful for binary systems (Φ small) even though it is valid for a wide range of values of Φ; a simpler form is derived below which is valid for large Φ and is therefore more convenient for quaternary and higher order systems.

It has been pointed out by H. O. Pollak that the integral

$$\int_S^\infty \exp(-x) I_0(x \sin \Phi)\, dx$$

which arises from consideration of a sine wave plus random noise[6] is closely related to the error-rate integral $P_0(\Phi)$. Pollak's proof is sketched below.

Set

$$p = S \cos^2 \Phi, \qquad \alpha = \sin \Phi,$$

$$y = \frac{-1}{1+\alpha} + \frac{1}{1 + \alpha \sin \theta}, \qquad b = \frac{\alpha}{1 - \alpha^2}.$$

Then equation (4) becomes (after some simplification)

$$P_0 = \frac{\exp\left(\dfrac{-p}{1+\alpha}\right)}{2\pi(1-\alpha^2)^{\frac{1}{2}}} \int_0^{2b} \frac{\exp(-py)\, dy}{\left(y + \dfrac{1}{1+\alpha}\right)(2by - y^2)^{\frac{1}{2}}}.$$

But[7]

$$\int_0^{2b} \frac{\exp(-py)}{(2by - y^2)^{\frac{1}{2}}}\, dy = \pi \exp(-bp) I_0(bp)$$

from which it can readily be shown that

$$P_0 = \frac{\cos \Phi}{2} \int_S^\infty \exp(-x) I_0(x \sin \Phi)\, dx.$$

From this form of the error-rate integral an approximate solution can be derived when $S \sin \Phi$ is sufficiently large to expand the Bessel Function as*

$$I_0(z) = \frac{\exp(z)}{(2\pi z)^{\frac{1}{2}}}.$$

---

* For $|z| > 0.15$ this approximation is valid to within 17 percent.

TABLE III—S/N INCREASE FOR VARIOUS NUMBERS OF LEVELS

Number of levels	Approximate increase in S/N in dB from the binary coherent phase-shift-keyed case for same error-rate
2	0.5
4	5.3
8	11.2
16	17.2
32	23.2
and so on	and so on

Making this substitution and performing the integration gives

$$P_0 = \frac{1}{2}\left[\frac{1 + |\csc \Phi|}{2}\right]^{\frac{1}{2}} \operatorname{erfc}\,\{[(1 - |\sin \Phi|)S]^{\frac{1}{2}}\}.$$

When both $|\sin \Phi| > 0.15$ and $|\sin \Phi| > 0.15/S$ hold, this can be written, to an accuracy of a factor of two, as simply

$$P_0 = \tfrac{1}{2} \operatorname{erfc}\,\{[(1 - |\sin \Phi|)S^{\frac{1}{2}}\}.$$

For cases of interest for quaternary and higher order systems these constraints are well satisfied and this approximation is very good.

For the M-level case with $\delta = \phi = 0$, (no phase distortion)

$$\Pi = \tfrac{1}{2} \operatorname{erfc}\left[\left\{\left[1 - \sin\left(\frac{\pi}{2} - \frac{\pi}{M}\right)\right]S\right\}^{\frac{1}{2}}\right]$$

$$= \tfrac{1}{2} \operatorname{erfc}\left\{\left[4\left(\sin^2 \frac{\pi}{2M}\right)S\right]^{\frac{1}{2}}\right\}.$$

For large M this becomes

$$\Pi = \tfrac{1}{2} \operatorname{erfc}\left(\frac{\pi}{M} S^{\frac{1}{2}}\right).$$

Thus, the S/N must be increased by 6 dB each time the number of levels is doubled if the error-rate is to remain constant. This is illustrated in Table III.

REFERENCES

1. Hubbard, W. M., Goell, J. E., Warters, W. D., Standley, R. D., Mandeville, G. D., Lee, T. P., Shaw, R. C., and Clouser, P. L., "A Solid-State Regenerative Repeater for Guided Millimeter-Wave Communication Systems," B.S.T.J., *46*, No. 9 (November 1967), pp. 1977–2018.
2. Hubbard, W. M., "The Effect of Intersymbol Interference on Error-Rate in Binary Differentially-Coherent Phase-Shift-Keyed Systems," B.S.T.J., *46*, No. 6 (July–August 1967), pp. 1149–1172.

3. Bennett, W. R., and Davey, J. R., *Data Transmission,* New York: McGraw-Hill, 1965.
4. Hubbard, W. M., "The Effect of a Finite Width Decision Threshold for Binary Differentially-Coherent PSK System," B.S.T.J., *45,* No. 2 (February 1966), pp. 307–319.
5. Bennett, W. R., and Salz, J., "Binary Data Transmission by FM Over a Real Channel," B.S.T.J., *42,* No. 5 (September 1963), p. 2387,
6. Rice, S. O., "Statistical Properties of a Sine Wave Plus Noise," B.S.T.J., *27,* No. 1 (January 1948), pp. 109–157.
7. Erdelyi, A., editor, *Tables of Integral Transforms,* vol. 1, New York: McGraw-Hill, 1953, p. 138 (14).

Crosstalk in Multiple-Beam Waveguides

By DETLEF GLOGE

(Manuscript received August 6, 1969)

Crosstalk limits the number of communication channels which are spatially resolvable at the end of a beam waveguide. The main sources of crosstalk are scattering and distortion by the focusers. A careful study of high quality front surface mirrors led to the results of this paper. The best choice seems to be a periscopic guide made of dielectric mirrors when used with gaussian beams in a particular mode of multiple transmission. We give a closed description for the expected power profile of a gaussian beam that has passed such a guide and an approximate formula for the mutual crosstalk between several such beams.

I. INTRODUCTION

Arranging many optical channels spatially resolved in the same waveguide is a simple means for high capacity transmission.[1,2] All channels can be modulated in the same frequency band as long as the crosstalk is kept below a certain threshold. One source of crosstalk is the inevitable scattering from the focusing and directing elements.[3] Diffraction from the ideal beams is negligible.[4] Yet we shall see that diffraction must be considered once the beams are distorted by the focusers.

In all likelihood, these focusers would consist of mirrors because lenses of the size needed are apt to have imperfections within the material. The scattering characteristics of high quality front surface mirrors and lenses of the best quality have been measured recently.[5,6] A comparison shows that the lens scattering was about one order of magnitude larger. Directional changes can easily be accomplished by using periscopic mirror arrangements of the kind shown in Fig. 1.[7] Neglecting aberrations, we consider only the first order focusing effect of these periscopes, which is that of thin convex lenses.

Two methods of multiple channel transmission have been suggested.[2] We discuss these two basic methods with respect to their susceptibility

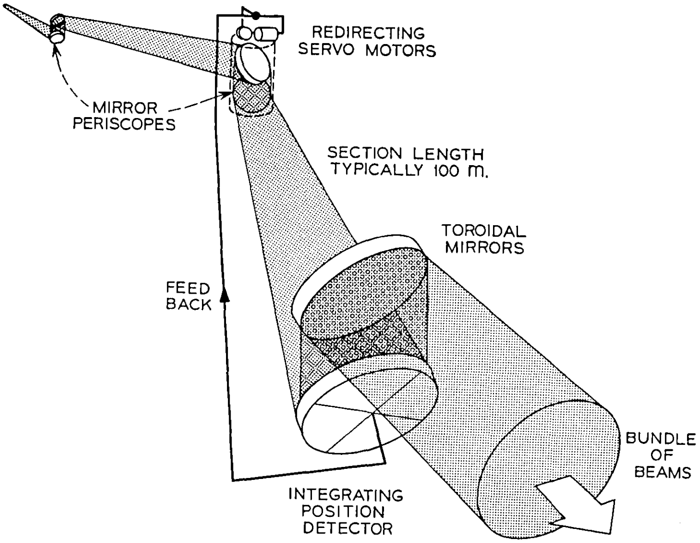

REDIRECTING SERVO MOTORS
MIRROR PERISCOPES
SECTION LENGTH TYPICALLY 100 m.
TOROIDAL MIRRORS
FEED BACK
BUNDLE OF BEAMS
INTEGRATING POSITION DETECTOR

Fig. 1 — Sketch of a periscopic waveguide.

to crosstalk when mirrors of the kind measured in Ref. 5 are used in the waveguide.

II. MULTIPLE CHANNEL TRANSMISSION

The useful cross section of the beam guide is limited by the size of the periscopic mirrors. Without unreasonable effort, mirrors of good optical quality can be made 20 to 30 cm in diameter at the most. The bundle of beams must clear the mirror edges by a wide margin at all times to guarantee safe operation. This implies that diffraction crosstalk caused by the mirror edges is negligible. Tolerances which would allow for controlled diffraction, as suggested by Ref. 4, seem unreasonably tight. Thus we arrive at a radius R of about 10 cm for the useful cross section. The spacing D of the focusers is limited by the terrain and the cost of the straight sections in between. It will most likely be of the order of 100 m. For optimum conditions, the effective focal length of the focusers should be half their spacing, although some deviation can be tolerated in this respect.

Consider the waveguide as a periodic lens system which images an array of transmitters into a similar array of detectors. This is basically what the imaging method (in Fig. 2a) does. Diffraction effects can be minimized if every transmitter radiates a coherent gaussian beam. As these beams propagate in the guide, their sizes change periodically from the fairly small transmitter spot size to a size close to the cross section of the guide.

This periodic change is avoided by the grouping method sketched in Fig. 2b. The beams arrange in groups and open up to the fundamental mode radius

$$w = \left(\frac{D\lambda}{\pi}\right)^{\frac{1}{2}} \tag{1}$$

before they enter the guide. They keep very close to this radius throughout the transmission. Figure 2b shows the grouping method for two groups containing two beams each. Special collector lenses single out the groups at the end and focus the beams well separated on the detector array. For a better understanding of this detection system, consider the focal length f of the collector lenses to be short compared to the distance D between a collector lens and the preceding focuser. In the plane of this focuser, all groups of beams form overlapping patterns. Every collector lens selects the pattern of its group and images it into the detector plane scaled down by a factor f/D. Consequently, the detector array of every group is confined to a circular area with a radius Rf/D.

The density of resolvable beams in the system is determined by beam distortion and scattering rather than the spread of the ideal beams. The distortion of the beam profile determines the receiver size required

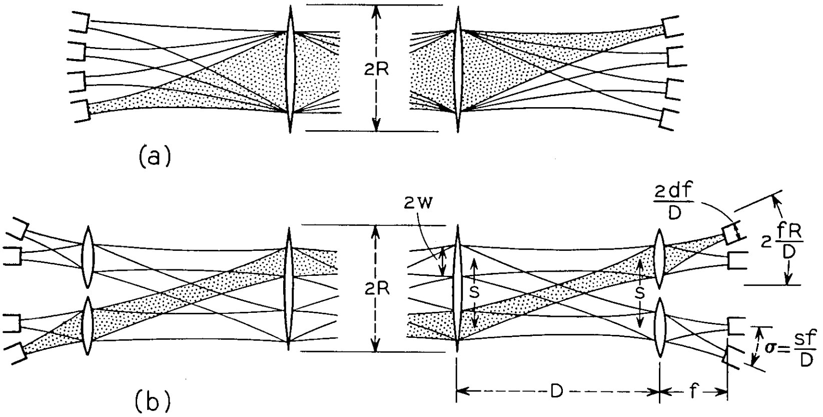


Fig. 2 — Schemes for spatially resolved transmission (a) the imaging method (b) the grouping method.

to secure safe reception. One would like to make the detectors as small as possible in order to minimize the scattering collected from adjacent channels. For further reduction of the crosstalk, one has to increase the spacing between the detectors.

Let us assume that the center-to-center spacing is s for the collector lenses and σ for the detectors in a group. In this case we would have $\pi R^2/s^2$ groups and $\pi R^2 f^2/D^2\sigma^2$ beams per group if the guide were perfectly confocal. If we allow a slight tolerance for the spacing of the focusers, Ref. 2 shows that the groups belonging to off-center collector lenses cannot be completely filled. For this reason the total number of channels is only half the theoretical maximum, namely

$$N = \frac{1}{2}\frac{\pi^2 R^4 f^2}{s^2\sigma^2 D^2}. \tag{2}$$

Rather than considering the detector plane, let us look at the distribution every group has at the focuser preceding a collector lens. This way our results become independent of the focal length f of the collector lenses and a function of the beam waveguide only. In the plane of the last focuser the beam spacing is $D\sigma/f$. As the beams have the same width there as at the collector lenses, it seems reasonable to set

$$s = D\sigma/f. \tag{3}$$

Inserting this into equation (2) yields

$$N = \frac{\pi^2}{2}\left(\frac{R}{s}\right)^4. \tag{4}$$

In the following let us assume that the detectors are simple quantum counters, have a circular area, and have a radius df/D. Transforming this back to the last focuser, we find a circular area of radius d susceptible to crosstalk around every beam. In Section II we evaluate scattering and distortion of a single beam in the plane of the last focuser for the case of a waveguide of n mirror periscopes. Since we use direct detection, we may neglect phase front distortion. The case of heterodyne reception is briefly discussed in the appendix. The results are very similar to those of the direct detectors.

II. DISTORTION AND SCATTERING

Both distortion and scattering are a consequence of irregularities on the mirror surfaces. The distortion originates from smooth imperfections extending over areas comparable to the beam cross section while scattering is caused by a surface roughness correlated over distances much smaller than the beam diameter. Both irregularities are part of a statistical function $\delta(x, y)$ which describes the deviation from the ideal surface. Taking a meaningful average over an ensemble of test surfaces leads to the structure function

$$\Delta(\rho) = \langle[\delta(x, y) - \delta(x - \rho\cos\alpha, y - \rho\sin\alpha)]^2\rangle_{\text{av}} \tag{5}$$

where ρ and α belong to a polar coordinate system which has the point (x, y) as its origin. Writing Δ as a function of ρ only implies the assumption that δ is stationary and isotropic, which seems justified for the statistical properties involved.[5]

A light wave of wavelength λ reflected off the imperfect surface suffers a phase front distortion

$$\varphi(x, y) = \frac{4\pi}{\lambda}\,\delta(x, y). \tag{6}$$

We neglect reflection loss which we assume to be uniform over the surface. For gaussian statistics[4]

$$\langle\exp i[\varphi(x, y) - \varphi(x - \rho\cos\alpha, y - \rho\sin\alpha)]\rangle_{\text{av}} = \exp\left[-\frac{8\pi^2}{\lambda^2}\,\Delta(\rho)\right]. \tag{7}$$

This equality will be used to calculate the power distribution $p_1(r)$ at a distance D from the reflecting surface. Assume that the reflected beam is circular, symmetric, and would have a power profile $p_0(r)$ at a distance D if the reflection were ideal. Then, from Ref. 8, one obtains

$$g_1(\rho) = g_0(\rho)\exp\left[-\frac{8\pi^2}{\lambda^2}\,\Delta(\rho)\right] \tag{8}$$

where $g_1(\rho)$ and $g_0(\rho)$ are the Hankel transforms of $p_1(r)$ and $p_0(r)$, respectively. This Hankel transformation is defined by

$$g_1(\rho) = 2\pi\int_0^\infty p_1(r)J_0(2\pi\rho r/D\lambda)r\,dr \tag{9}$$

or

$$p_1(r) = \frac{2\pi}{D^2\lambda^2}\int_0^\infty g_1(\rho)J_0(2\pi\rho r/D\lambda)\rho\,d\rho \tag{10}$$

where J_0 is the Bessel function of zero order. The quantity $p_1(r)$ has to be understood as an average over an ensemble of equivalent surfaces.

For an accurate confocal spacing of the periscopes, a beam and its distortion in the guide reproduces itself at every second periscope. These periscopes only contribute to the phase front distortion in the detector plane, while all odd ones deteriorate the power profile as well. We have n periscopes with $2n$ surfaces, half of them contributing to the profile distortion. Since the imperfections of all surfaces are uncorrelated, we may write

$$g_n(\rho) = g_0(\rho) \exp\left[-\frac{8\pi^2}{\lambda^2} n\, \Delta(\rho)\right] \tag{11}$$

for the detector plane.

In a guide with thousands of focusers, accurate confocal spacing requires tight tolerances for the focal lengths and spacings. In a practical guide, the focusers will be kept only nearly confocal and, in general, will not be at positions at which previous distortions are reproduced. If all positions are equally probable along the guide, equation (11) can be adapted in the following way[8]

$$g(\rho) = g_0(\rho) \exp\left[-\frac{8\pi^2}{\lambda^2} \frac{2n}{\pi} \int_0^\pi \Delta(\rho \sin \xi)\, d\xi\right]. \tag{12}$$

Notice that for $\xi = 0$ or π, we have $\Delta(0) = 0$ and no change of the power distribution, while for $\xi = \pi/2$ the profile distortion is a maximum.

A fairly reliable functional approximation for Δ in the range $\rho = 0.01$ to 1 mm was derived from scattering measurements around a test beam.[5] The scattering is an effect of the mirror surface roughness averaged over the area covered by the test beam. This average is equivalent to an average over an ensemble of test surfaces. As a consequence, the variance of the scattered power is small, that is, the scattered power actually measured is very close to the average power. The measurements were practically the same for all test surfaces.

This is not true for the processes involved in beam distortion. In this case, the δ-components participating are correlated over areas comparable to the test beam, no averaging is accomplished by the measurement, and the result can be grossly different from one surface to the next. It is this difference between scattering and distortion which makes scattering measurements feasible but distortion measurements tedious and expensive. Distortion is not sufficiently described by an average power profile; instead one needs to know the complete probability distribution of the power at every point in the beam cross section. In this situation some grossly simplifying assumptions are necessary to tackle the distortion problem with the scant experimental evidence available.

We derived the functional approximation

$$\Delta(\rho) = E\rho \quad \text{for} \quad \rho = 0.01 \cdots 1 \text{ mm} \tag{13}$$

with

$$E = 2.4 \times 10^{-9}\mu \tag{14}$$

from measurements.[5] This approximation is plotted in Fig. 3. For larger ρ we have only one reference point: the quality factor of the mirror, given in fractions of the green wavelength, which specifies δ-components correlated over areas comparable to the polishing tool. This will be typically several centimeters. We know Δ decreases for smaller ρ and merges into the linear function (13) at $\rho = 1$ mm. As a convenient approximation, let us assume that Δ is linear everywhere. This function would correspond to about $\lambda_{\text{green}}/50$ at several cm.

Whether equation (13) is a good approximation for $\rho < 0.01$ mm we do not know, but this is of little interest, since components at these small δ generate scattering which does not reach the next focuser, but is absorbed by the guide wall.

A gaussian beam of unit power has the profile

$$p_0(r) = \frac{2}{\pi w^2} \exp(-2r^2/w^2). \tag{15}$$

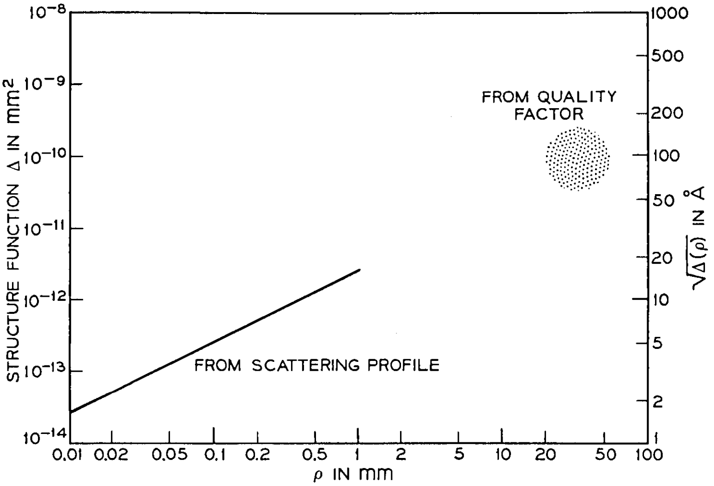


Fig. 3 — The linear structure function $\Delta(\rho)$ calculated from scattering profile and quality tests.

Its Hankel transform is

$$g_0(\rho) = \exp(-\pi^2 w^2 \rho^2 / 2D^2 \lambda^2). \tag{16}$$

Using equations (10), (12), (13), and (16), we find the following average power distribution at the detector

$$p(r) = \frac{2}{\pi w^2} \int_0^\infty \exp\left(-\frac{u^2 + au}{2}\right) J_0\left(\frac{2ur}{w}\right) u \, du \tag{17}$$

where

$$u = \frac{\pi w}{D\lambda} \rho \quad \text{and} \quad a = 64 \frac{wED}{\lambda w}\cdot \tag{18}$$

The average power falling outside a circular area of radius z is

$$P(z) = 2\pi \int_z^\infty p(r) r \, dr. \tag{19}$$

Using the identity

$$J_1(z)z = \int_0^z J_0(v) v \, dv, \tag{20}$$

we arrive at

$$P(z) = 1 - \frac{2z}{w} \int_0^\infty \exp\left(-\frac{u^2 + au}{2}\right) J_1\left(\frac{2uz}{w}\right) du. \tag{21}$$

The result of the machine evaluation of equation (21) is plotted in Fig. 4. For $a = 0$ the beam is undistorted and $P(z)$ is gaussian. Yet $P(z)$ has a tail decreasing with $1/z$ for finite a.

In the course of our calculation we want to know the radius z outside of which a given power P can be found for a certain parameter a. For this purpose the function $z(P, a)$ is plotted in Fig. 5. It can be approximated by the expression

$$z = \left[w^2 \ln \frac{1}{P^{\frac{1}{2}}} + \left(\frac{16EnD}{\lambda}\right)^2 \left(\frac{1}{P^2} - 1\right) \right]^{\frac{1}{2}} \tag{22}$$

where equation (18) was inserted for a.

The first part of equation (22) is an inverse gaussian and depicts the coherent beam, while the second part accounts for the incoherent portion. Equation (22) can be used, for example, to calculate the detector radius required at the end of a periscopic guide. In this case one will probably allow the second term in equation (22) to be about

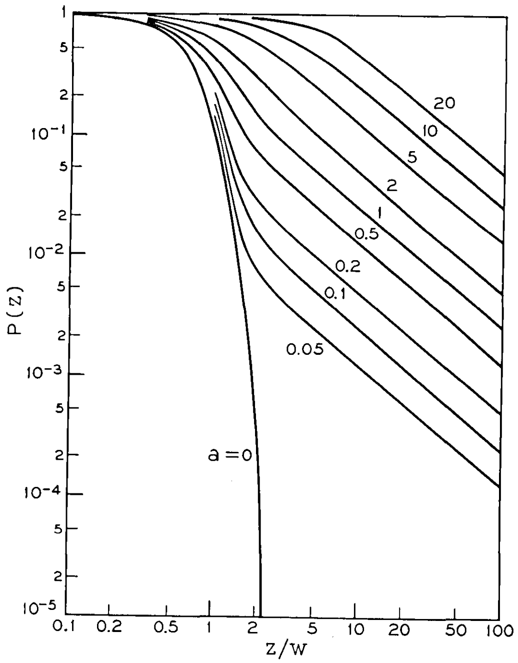


Fig. 4 — The power fraction P expected outside the circular area with the normalized radius z/w for various distortion coefficients a.

equal to the first. This would mean that the beam deterioration becomes just noticeable, but not yet dominating, at the end.

We shall find another application for equation (22) in the course of calculating the scattering crosstalk. In this case we require P to be so small that the second term in equation (22) exceeds the first even for moderate distortions, and equation (22) can be approximated by

$$z \cong \frac{16EnD}{\lambda P}. \tag{23}$$

Notice that the only guide dimension that enters into this formula is the total transmission distance

$$L = nD \tag{24}$$

from one repeater to the next. Equation (22) leads to an estimate for the detector sizes required and with this information and the help of equation (23) we can evaluate the scattering crosstalk.

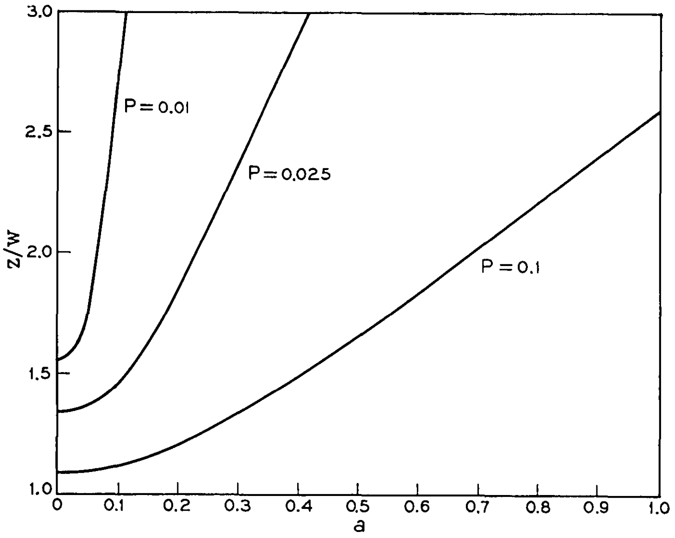


Fig. 5 — The functional relation $z(a)$ with P as a parameter.

IV. BEAM SPREAD AND CROSSTALK

The interchannel crosstalk at the end of the multiple beam guide is a function of the detector size. In order to minimize the amount of light collected from other channels, the detectors should not be larger than is absolutely necessary for signal detection. A few percent of the signal power could even be sacrificed. The signal fraction to be detected will depend on the signal levels available and on the noise sources involved, but it is probably safe to assume that, on the average, 75 percent of the total signal power will be sufficient. Thus we obtain from equations (22) and (24) for the detector radius

$$d = \left[w^2 \ln 2 + 15\left(16 \frac{LE}{\lambda}\right)^2\right]^{\frac{1}{2}}. \tag{25}$$

In a more general sense, we may interpret d as the average radius of a distorted beam at the end of a guide of length L. In equation (25), w is the radius of the ideal gaussian beam which may vary considerably along the guide as, for example, in the case of the imaging method. For the grouping method, w is constant and given by equation (1).

To compare both methods, let us consider a practical example of a waveguide with 100 m section length operating at a wavelength of 1 μm over a distance of 50 km. If we use the grouping method, we find

$w = 5.65$ mm from equations (1) and (25) yields a beam radius of 8.8 mm. In the case of the imaging method, w varies about 5.65 mm from lens to lens, being much smaller than 5.65 mm at the detectors. Yet at the end of a 50 km path, the average radius of the distorted beams will not be smaller than 7.5 mm because of the second part of equation (25). This is only slightly less than the radius of the grouped beams. It is obvious from Fig. 2 that under these circumstances the imaging method loses its advantage. Actually, in this case, the imaging method can only accommodate the beams contained in one group of the grouping method. Therefore, in the following we consider only the grouping method.

For the calculation of the scattering crosstalk we restrict ourselves to paraxial beams. Any two beams of this kind are equivalent in the sense that the amount of light scattered from a beam 1 into another beam 2 is equal to the amount scattered from 2 into 1. In the same way, the scattering from one beam into all others is what the beam receives from all others. This is exactly the crosstalk we want to calculate. Thus, in order to consider the worst situation, let us select the center beam and calculate what it scatters into all extraneous receivers. To do this we have to integrate the scattered power falling into the detector plane, excepting the center detector and the blind area between all detectors. We remember that the detectors have a radius d and are spaced by a distance s center to center. We obtain a reasonable and conservative approximation if we collect all the scattering outside a circle with radius $s/2$, which is $P(s/2)$ from equation (21), and multiply this by a density factor $\pi d^2/s^2$. Consequently, the crosstalk which the center beam inflicts upon, and receives from, all other beams is

$$C = \frac{\pi d^2}{s^2} P(s/2). \tag{26}$$

For all practical cases the tolerable C is so small that we may use the approximation (23) for P. Inserting equations (1), (24), and (25) into equation (26) we obtain

$$C = 32 \frac{LDE}{s^3} \ln 2 + 30\pi \left(\frac{16LE}{\lambda s} \right)^3 . \tag{27}$$

Figure 6 shows the signal to crosstalk ratio $1/C$ plotted in decibel versus the spacing s for the previous example, that is, $D = 100$ m and $\lambda = 1$ μm. Also shown is the guide capacity N to be achieved by the grouping method in a guide of 10 cm radius. The transmission distance L between repeaters is the parameter. For reasons explained previously,

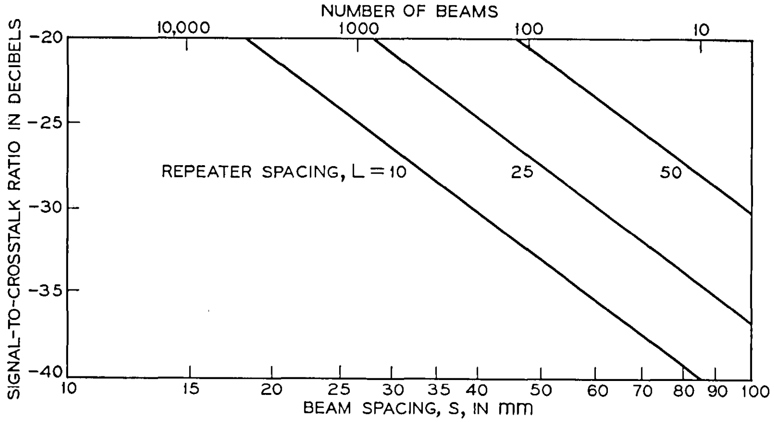


Fig. 6 — Crosstalk versus the beam density and the capacity for various repeater spacings ($\lambda = 1\ \mu$m, $D = 100$ m, and $R = 100$ mm).

the capacity achieved by the imaging method is less, for practical systems only about $N^{\frac{1}{2}}$. Figure 7 shows the useful guide radius required for a certain capacity at various repeater spacings if a crosstalk level of 23 dB is tolerable. Both Figs. 6 and 7 exhibit a system with $N = 100$, $L = 50$ km, $R = 10$ cm, $D = 100$ m, $\lambda = 1\ \mu$m, and $C = 23$ dB as feasible but also (more or less) as a practical limit.

The grouping method uses collector lenses in front of the detectors. Diffraction at these lenses must not cause excessive crosstalk even if the beams are badly distorted. For this reason the apertures have to be fairly large. If the available space is fully used, the lenses touch one another and are arranged as in a fly's eye lens. The lenses should be so large that the power at the lens edges is mainly incoherent and not part of the coherent, though distorted, beam. In this case, diffraction does not substantially increase the total power outside the signal beams. This requirement sets a lower limit to the beam spacing s which is simultaneously the diameter of the collector lenses. How far this limit is approached by the system depicted in Figs. 6 and 7 is a difficult question to answer.

A qualitative approach is tried in Fig. 8 where the previous results are also summarized. Figure 8 shows the power expected outside a given aperture after a beam has passed a length L of periscopic waveguide. The beam is supposed to start with a fundamental mode radius $w = 5.65$ mm in a guide with 100-m section length. Also shown is the power $P(s/2)$ falling outside a collector lens of radius $s/2$ where s is chosen with the help of equation (27) to guarantee a crosstalk level of C dB. Thus, once we have decided on the crosstalk level and the transmission distance, we find the radius of the collector lens and the power falling outside this lens from Fig. 8. For $L = 50$ km and $C = 23$ dB, this power seems to be composed mainly of scattered light so that diffraction should contribute little to the overall crosstalk.

Several discrepancies become apparent when the results of this paper are compared to previous publications. The power P falling outside a circle with a radius z after only one reflection is obtained if we replace equation (12) by equation (11) and set $n = 1$ in the derivation of equation (23). In the case of a linear structure function, equations (11) and (12) differ by a factor $\pi/4$, and therefore

$$P = 4\pi \frac{ED}{\lambda z}. \tag{28}$$

The same physical problem was approached on a different course in Ref. 5 and is expressed in equation (16) there. That result differs from our equation (28) by a factor of four. The reason is a factor of four erroneously introduced in equation (13) of that publication.

In Ref. 3 the crosstalk of one beam into one other beam was measured. This quantity can be calculated on the basis of this paper. The

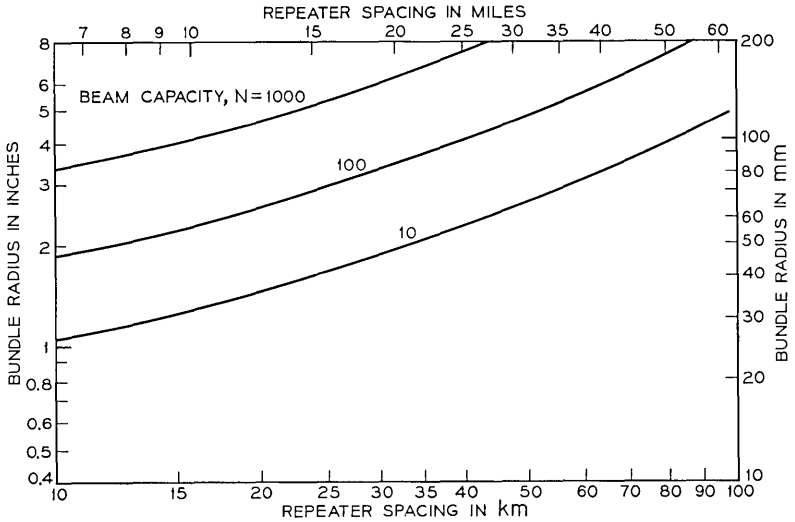


Fig. 7 — The beam bundle radius versus the repeater spacing for a given capacity N (signal/crosstalk $= 23$ dB, $\lambda = 1\ \mu$m, and $D = 100$ m).

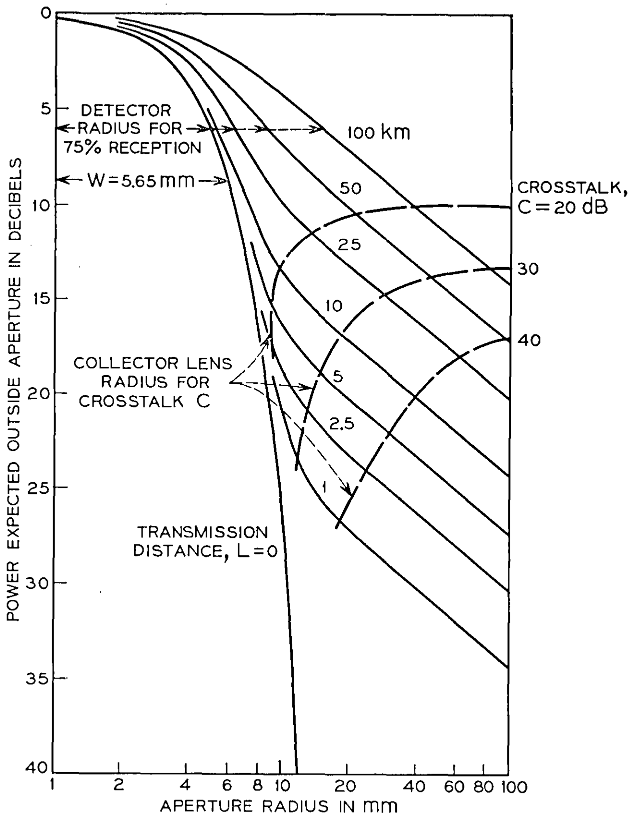


Fig. 8 — The power fraction expected outside the collector lens aperture for various repeater spacings and crosstalk levels (section length = 100 m; wavelength = 1 μm).

power density at a distance z from a scattering beam is

$$q = -\frac{1}{2\pi z}\frac{dP}{dz} \tag{29}$$

with $P(z)$ from equation (23). A small aperture of radius d displaced by a distance s from the beam center collects approximately

$$c = \pi d^2 q(s) \tag{30}$$

with q from equation (29). This is the crosstalk in the second beam. With equations (23), (29), and (30) we obtain

$$c = \frac{8d^2EL}{\lambda s^3}. \tag{31}$$

The specific dimensions in the experimental arrangement of Ref. 3 were $s = 5$ mm, $d = 2.5$ mm, and $\lambda = 0.63$ μm. The transmission distance was equivalent to 8.5 km of a periscopic guide. By inserting these data into equation (31) we obtain a crosstalk level of 19 dB. The measured level was 30 dB. Moreover, equation (31) suggests that the crosstalk decreases with the third power of the beam spacing. In Ref. 3, on the other hand, a decrease with the fourth power of the beam spacing was measured. The reason for the discrepancy is still under investigation. The comparison seems to indicate that the data used here are on the conservative side.

Finally let us compare our results to the diffractional crosstalk which ideal gaussian beams experience along a wave guide. Reference 4 considers this situation giving the following results. A guide of 10,000 focusers, 100 m apart and 5 cm in diameter, could accommodate 16 beams with only 60 dB crosstalk. According to our findings, scattering in this arrangement causes 50 dB crosstalk in one 100 m section. This underlines the severe limit which scattering and distortion set to multibeam transmission. Improving the optical surfaces would be a worthwhile undertaking.

V. CONCLUSIONS

Scattering and distortion of the beams in a beam waveguide are caused by surface irregularities of the focusers. There is experimental evidence that these irregularities can be described, as a first approximation, by a linear structure function. Based on these findings we predict a beam distortion and an incoherent background radiation which both increase with the transmission distance L. The distortion makes it impossible to use a simple transmission method which images an array of transmitters into the detector plane. The method which seems more practical arranges the beams in groups and transmits them with unvariable width.

The incoherent background of scattered light around every beam fades off with the third power of the distance from the beam. This causes a crosstalk inversely proportional to the third power of the beam spacing. The beam density is limited by the crosstalk tolerable after 50 km. If we set this level at 23 dB, allow a beam bundle 20 cm in diameter, and operate at a wavelength of 1 μm, we could accommodate about 100 beams. This is based on the assumption that periscopic focusers are used which are made of high quality front surface mirrors. A critical comparison with previous publications suggests that our results are conservative, if not pessimistic.

APPENDIX

Heterodyne Detection

One might consider heterodyne detection as a way to reduce the scattering received in every channel. Local oscillator beams could be brought in line with the signal beams, utilizing beam splitters. The local oscillator beams would discriminate to a certain extent against the incoherent background of scattered light from other channels. To compare this with the quantum counters discussed in the text, let us assume that the scattered background light is uniform in the vicinity of the local oscillator beam. For this case Siegman has calculated the IF-photocurrent noise of heterodyne reception.[9] He found that the "integrated effective detector area" of the heterodyne detector is equal to λ^2.

We now calculate the "effective detector area" for our quantum detectors. Every detector has an area

$$A_1 = \frac{\pi d^2 f^2}{D^2} \tag{32}$$

and is preceded by an aperture with the area

$$A_2 = \pi s^2. \tag{33}$$

The distance between aperture and detector is f. From Ref. 10 we find the "effective detector area" for this arrangement to be

$$\frac{A_1 A_2}{f^2} = \frac{\pi^2 s^2 d^2}{D^2}. \tag{34}$$

Inserting equation (1) we obtain

$$\frac{A_1 A_2}{f^2} = \frac{s^2 d^2}{w^4} \lambda^2. \tag{35}$$

If we had $sd = w^2$, the quantum counter would discriminate as well against scattering as the heterodyne detectors.

In the case of distorted signal beams, however, both schemes cannot recover the full signal power. What is important in this case is the ratio of signal to background light collected in the respective cases. For this reason, the quantum counter can be equivalent to the heterodyne detector even if s and d are slightly larger than w. However, for reasons explained in Section IV, s will be considerably larger than w to avoid diffractional crosstalk. The heterodyne receiver therefore surpasses the quantum counter. On the other hand, the complexity of the former probably makes up for this advantage.

REFERENCES

1. Rosenthal, J. E., "Modulation of Coherent Light," Bull. Amer. Phys. Soc. II, *6,* No. 1 (February 1, 1961), p. 68.
2. Gloge, D., and Weiner, D., "The Capacity of Multiple Beam Waveguides and Optical Delay Lines," B.S.T.J., *47,* No. 10 (December 1968), pp. 2095–2109.
3. Gloge, D., and Steier, W. H., "Experimental Simulation of a Multiple Beam Optical Waveguide," B.S.T.J., *48,* No. 5 (May 1969), pp. 1445–1457.
4. Goubau, G., and Schwering, F. K., "Diffractional Crosstalk in Beam Waveguides for Multibeam Transmission," Proc. IEEE, *56,* No. 9 (September 1968), pp. 1632–1634.
5. Gloge, D., Chinnock, E. L., and Earl, H. E., "Scattering from Dielectric Mirrors," B.S.T.J., *48,* No. 3 (March 1969), pp. 511–526.
6. Hostetter, G. R., Patz, D. L., Hill, H. A., and Zanoni, C. A., "Measurement of Scattered Light from Mirrors and Lenses," Appl. Opt., *7,* No. 7 (July 1968), pp. 1383–1385.
7. Gloge, D., "Experiments with an Underground Lens Waveguide," B.S.T.J., *46,* No. 4 (April 1967), pp. 721–735.
8. Gloge, D., unpublished work.
9. Siegman, A. E., "The Antenna Properties of Optical Heterodyne Receivers," Proc. IEEE, *54,* No. 10 (October 1966), pp. 1350–1356.
10. Kogelnik, H., and Yariv, A., "Considerations of Noise and Schemes for its Reduction in Laser Amplifiers," Proc. IEEE, *52,* No. 2 (February 1964), pp. 165–172.

Asymptotic Analysis of a Nonlinear Autonomous Vibratory System

By J. A. MORRISON

(Manuscript received June 13, 1969)

A system consisting of a spring, dashpot, and mass upon which is mounted an eccentric driven by a motor with a linear torque-speed characteristic, is analyzed by perturbation procedures based on small reciprocal of rotational inertia. Periodic solutions of the third order system, which arises when the angular position of eccentric mass is taken as the new independent variable, are constructed, and their stability is analyzed. An asymptotic solution is also obtained which is more general than a periodic solution, in that the averaged rotational speed is a slowly varying function, rather than a constant. The results are applicable to the determination of the interaction between the rotational motion of a flexibly mounted motor and the translational vibratory motion of its frame.

I. INTRODUCTION

In a recent paper Senator analyzed a system consisting of a spring, dashpot, and mass upon which is mounted a rotating eccentric weight driven by a motor with a linear torque-speed characteristic.[1] This system has been analyzed by several authors, under different assumptions on the values of the parameters of the system (see Refs. 5 through 9), and Senator discusses their results. The system is a model for the interaction between the rotational motion of a motor driving an eccentric and the translational vibratory motion of the frame, which is caused by this rotation.

In Ref. 1, Senator constructed periodic (rotational) solutions by means of a perturbation technique, based on small reciprocal of rotational inertia. However, he did not analyze the stability of the periodic solutions directly, but proceeded in a somewhat different manner. Thus, he introduced a van der Pol type transformation, but imposed a subsidiary condition on the slowly varying functions of time which differs from the one usually imposed in the method of averaging. He then made assumptions regarding the order of smallness of various derivatives, and dropped all second order terms from the equations for the slowly varying quantities, obtaining what he called averaged equations. The stationary solutions of these averaged equations correspond to periodic solutions of the original system, and he analyzed the stability of the stationary solutions on the basis of the corresponding linearized variational equations.

It is the purpose of this paper to show how the stability condition obtained by Senator may be derived rigorously for sufficiently large values of rotational inertia. This is done by taking the angular position of eccentric mass as the new independent variable, constructing periodic solutions of the resulting third order system, and then analyzing the linearized variational equations corresponding to them. Perturbation procedures, based on the small reciprocal of rotational inertia, are used.

An asymptotic solution is also obtained which is more general than a periodic solution, in that the averaged rotational speed is a slowly varying function, rather than constant. However, this asymptotic solution is not completely general, in that the transients in the translational motion, which decay on a much faster scale, are not included. The asymptotic solution nevertheless provides insight into the manner in which a stable periodic solution is approached, and the analytical results are borne out by some numerical calculations.

II. PERIODIC SOLUTIONS

The equations of motion, in dimensionless form, for the system under consideration are (from Ref. 1),

$$\frac{d^2u}{d\tau^2} + 2\zeta\frac{du}{d\tau} + u = \alpha\left[\left(\frac{d\theta}{d\tau}\right)^2 \sin\theta - \frac{d^2\theta}{d\tau^2}\cos\theta\right] \tag{1}$$

$$\frac{d^2\theta}{d\tau^2} = \epsilon\left(p - b\frac{d\theta}{d\tau} - \alpha\frac{d^2u}{d\tau^2}\cos\theta\right). \tag{2}$$

Here τ is dimensionless time, α, b, p and $\zeta > 0$ are constants, and $\epsilon > 0$, the reciprocal of dimensionless inertia, is a small parameter. Also, u is the dimensionless translational displacement, and θ is the angular position of eccentric mass. Instead of dealing with the fourth order system (1) and (2), as did Senator, it turns out to be more convenient to take θ as the new independent variable. Accordingly, defining

$$\Omega = \frac{d\theta}{d\tau} \tag{3}$$

the third order system

$$\Omega^2 \frac{d^2u}{d\theta^2} + 2\zeta\Omega \frac{du}{d\theta} + u + \Omega \frac{d\Omega}{d\theta}\left(\frac{du}{d\theta} + \alpha \cos \theta\right) = \alpha\Omega^2 \sin \theta \tag{4}$$

$$\Omega \frac{d\Omega}{d\theta}\left(1 + \epsilon\alpha \cos \theta \frac{du}{d\theta}\right) + \epsilon\alpha\Omega^2 \cos \theta \frac{d^2u}{d\theta^2} = \epsilon(p - b\Omega) \tag{5}$$

is obtained.

A periodic solution to (4) and (5) is sought in the form

$$u = \tilde{u}(\theta, \epsilon) \equiv u_0(\theta) + \epsilon u_1(\theta) + \epsilon^2 u_2(\theta) + \cdots \tag{6}$$

$$\Omega = \tilde{\Omega}(\theta, \epsilon) \equiv \omega_0 + \epsilon\Omega_1(\theta) + \epsilon^2\Omega_2(\theta) + \cdots \tag{7}$$

where $u_i(\theta)$ and $\Omega_i(\theta)$ are periodic in θ, with period 2π, and ω_0 is a constant. Substitution of (6) and (7) into (4) and (5), and comparison of the lowest powers of ϵ, leads to

$$\omega_0^2 \frac{d^2u_0}{d\theta^2} + 2\zeta\omega_0 \frac{du_0}{d\theta} + u_0 = \alpha\omega_0^2 \sin \theta \tag{8}$$

$$\omega_0 \frac{d\Omega_1}{d\theta} + \alpha\omega_0^2 \cos \theta \frac{d^2u_0}{d\theta^2} = (p - b\omega_0). \tag{9}$$

The periodic solution of (8) is

$$u_0 = \alpha\omega_0^2\Delta_0[(1 - \omega_0^2) \sin \theta - 2\zeta\omega_0 \cos \theta] \tag{10}$$

where

$$\Delta_0 = [(1 - \omega_0^2)^2 + 4\zeta^2\omega_0^2]^{-1}. \tag{11}$$

In order that $\Omega_1(\theta)$ should be periodic, it is necessary from (9) that

$$p = b\omega_0 + \alpha\omega_0^2\left\langle \cos \theta \frac{d^2u_0}{d\theta^2}\right\rangle_{\text{av}} = b\omega_0 + \alpha^2\zeta\omega_0^5\,\Delta_0 \equiv p^*(\omega_0) \tag{12}$$

using (10). $\langle\ \rangle_{\text{av}}$ denotes average over a period 2π of θ. Equation (12) gives a relationship between ω_0 and p, the dimensionless stall torque, and this relationship is depicted graphically in the figure for $\alpha = 0.707$, $\zeta = 0.2$, $b = 0$. It is noted that ω_0 is a triple valued function of p in part of the range. Senator concluded from his analysis that the middle branch corresponds to unstable periodic solutions, while the outer branches correspond to stable ones, a result verified in this paper.

Now, from (9), (10), and (12) it follows that

$$\Omega_1 = \{\omega_1 - \tfrac{1}{4}\alpha^2\omega_0^3\Delta_0[(1 - \omega_0^2) \cos 2\theta + 2\zeta\omega_0 \sin 2\theta]\} \tag{13}$$

where ω_1 is a constant, which is to be determined from the condition that $\Omega_2(\theta)$ should be periodic. It is clear as to how the higher order terms in the expansions in (6) and (7) may be obtained, but they will not be needed in the subsequent analysis. The periodic solutions in (6) and (7) are equivalent to those derived by Senator as periodic solutions of (1) and (2). It is necessary, of course, to perform a quadrature of equation (3) in order to obtain a relationship between θ and τ.

III. STABILITY ANALYSIS

The variational equations corresponding to the periodic solution $\tilde{u}$, $\tilde{\Omega}$, given by (6) and (7), are formed by substituting

$$u = (\tilde{u} + \xi), \qquad \Omega = (\tilde{\Omega} + \eta) \tag{14}$$

in (4) and (5), and linearizing in ξ and η. Thus,

$$\begin{aligned} &\tilde{\Omega}^2 \frac{d^2\xi}{d\theta^2} + 2\tilde{\Omega} \frac{d^2\tilde{u}}{d\theta^2}\eta + 2\zeta\left(\tilde{\Omega}\frac{d\xi}{d\theta} + \frac{d\tilde{u}}{d\theta}\eta\right) + \xi \\ &\quad + \tilde{\Omega}\frac{d\tilde{\Omega}}{d\theta}\frac{d\xi}{d\theta} + \left(\frac{d\tilde{u}}{d\theta} + \alpha\cos\theta\right)\left(\tilde{\Omega}\frac{d\eta}{d\theta} + \frac{d\tilde{\Omega}}{d\theta}\eta\right) = 2\alpha\tilde{\Omega}\sin\theta\eta, \end{aligned} \tag{15}$$

$$\begin{aligned} &\left(1 + \epsilon\alpha\cos\theta\frac{d\tilde{u}}{d\theta}\right)\left(\tilde{\Omega}\frac{d\eta}{d\theta} + \frac{d\tilde{\Omega}}{d\theta}\eta\right) + \epsilon\alpha\tilde{\Omega}\frac{d\tilde{\Omega}}{d\theta}\cos\theta\frac{d\xi}{d\theta} \\ &\quad + 2\epsilon\alpha\tilde{\Omega}\cos\theta\frac{d^2\tilde{u}}{d\theta^2}\eta + \epsilon\alpha\tilde{\Omega}^2\cos\theta\frac{d^2\xi}{d\theta^2} + \epsilon b\eta = 0. \end{aligned} \tag{16}$$

Equations (15) and (16) are linear equations with periodic coefficients, and the form of solution is known from Floquet theory.[2] Moreover, if all the characteristic exponents of the variational equations have negative real parts, then the periodic solution $\tilde{u}$, $\tilde{\Omega}$ is asymptotically stable. The behavior of the characteristic exponents will be analyzed for $0 < \epsilon \ll 1$.

The limiting case $\epsilon \to 0+$ will first be considered. In this case, from (6) and (7), $\tilde{\Omega} = \omega_0$ and $\tilde{u} = u_0(\theta)$ so that, from (15) and (16), $d\eta/d\theta = 0$ and

$$\omega_0^2 \frac{d^2\xi}{d\theta^2} + 2\zeta\omega_0\frac{d\xi}{d\theta} + \xi = 2\left(\alpha\omega_0\sin\theta - \omega_0\frac{d^2u_0}{d\theta^2} - \zeta\frac{du_0}{d\theta}\right)\eta. \tag{17}$$

Hence one of the characteristic exponents is $\lambda_0 = 0$, and the remaining two characteristic exponents satisfy

$$(\omega_0\lambda_0)^2 + 2\zeta(\omega_0\lambda_0) + 1 = 0 \tag{18}$$

and hence have negative real parts, since $\zeta > 0$ and $\omega_0 > 0$. For sufficiently small ϵ, these real parts will remain negative, so that it suffices to investigate the characteristic exponent which vanishes as $\epsilon \to 0$.

In the light of Floquet theory, a solution of (15) and (16) is sought in the form

$$\xi = e^{\lambda\theta}P(\theta), \qquad \eta = e^{\lambda\theta}Q(\theta) \tag{19}$$

where P and Q are periodic in θ, with period 2π, and

$$\lambda = \epsilon\lambda_1 + \epsilon^2\lambda_2 + \cdots \tag{20}$$

$$P(\theta) = P_0(\theta) + \epsilon P_1(\theta) + \epsilon^2 P_2(\theta) + \cdots \tag{21}$$

$$Q(\theta) = Q_0(\theta) + \epsilon Q_1(\theta) + \epsilon^2 Q_2(\theta) + \cdots . \tag{22}$$

It is a straightforward matter to substitute from (6), (7), and (19)–(22) into (15) and (16), and to compare like powers of ϵ. In particular, it is found from (16) that $dQ_0/d\theta = 0$. Omitting a multiplicative constant and taking $Q_0 = 1$, it is then found that

$$\omega_0^2 \frac{d^2P_0}{d\theta^2} + 2\omega_0 \frac{d^2u_0}{d\theta^2} + 2\zeta\omega_0 \frac{dP_0}{d\theta} + 2\zeta \frac{du_0}{d\theta} + P_0 = 2\alpha\omega_0 \sin\theta \tag{23}$$

and

$$\omega_0\left(\frac{dQ_1}{d\theta} + \lambda_1\right) + \frac{d\Omega_1}{d\theta} + 2\alpha\omega_0 \cos\theta \frac{d^2u_0}{d\theta^2} + \alpha\omega_0^2 \cos\theta \frac{d^2P_0}{d\theta^2} + b = 0. \tag{24}$$

Now, in order for $Q_1(\theta)$ to be periodic, it is necessary from (24) that

$$\omega_0\lambda_1 + b + \alpha\omega_0\left\langle \cos\theta \frac{d^2R_0}{d\theta^2}\right\rangle_{\text{av}} = 0 \tag{25}$$

where

$$R_0 = (\omega_0 P_0 + 2u_0) \tag{26}$$

is periodic in θ, with period 2π. But, from (23),

$$\omega_0^2 \frac{d^2R_0}{d\theta^2} + 2\zeta\omega_0 \frac{dR_0}{d\theta} + R_0 = 2\left(\alpha\omega_0^2 \sin\theta + \zeta\omega_0 \frac{du_0}{d\theta} + u_0\right) \tag{27}$$

and u_0 is given by (10) and (11). Straightforward calculations lead to

$$R_0 = 2\alpha\omega_0^2\, \Delta_0^2\{[(1 - \omega_0^2)^2(2 - \omega_0^2) + 4\zeta^2\omega_0^2(1 - 2\omega_0^2)] \sin\theta - \zeta\omega_0[(1 - \omega_0^2)(5 - \omega_0^2) + 12\zeta^2\omega_0^2] \cos\theta\}. \tag{28}$$

Thus, from (25),

$$\omega_0\lambda_1 + b + \alpha^2\zeta\omega_0^4\,\Delta_0^2[(1-\omega_0^2)(5-\omega_0^2) + 12\zeta^2\omega_0^2] = 0. \tag{29}$$

However, as may be verified from (12), using (11), equation (29) may be written in the form

$$\omega_0\lambda_1 + \frac{dp^*}{d\omega_0} = 0. \tag{30}$$

Since the sign of λ in (20) is determined by the sign of λ_1, for sufficiently small $\epsilon > 0$, it follows that the periodic solution $\tilde{u}$, $\tilde{\Omega}$ is asymptotically stable if $dp^*/d\omega_0 > 0$, and is unstable if $dp^*/d\omega_0 < 0$. That is, the middle branch of the figure corresponds to unstable periodic solutions, while the outer branches correspond to asymptotically stable ones, provided that $\epsilon > 0$ is sufficiently small.

IV. MORE GENERAL SOLUTIONS

In this section a more general solution of (4) and (5) is constructed, for $0 < \epsilon \ll 1$. Thus an asymptotic solution is sought in the form

$$u = v_0(\omega,\theta) + \epsilon v_1(\omega,\theta) + \epsilon^2 v_2(\omega,\theta) + \cdots \tag{31}$$

$$\Omega = \omega + \epsilon w_1(\omega,\theta) + \epsilon^2 w_2(\omega,\theta) + \cdots \tag{32}$$

where $v_i(\omega,\theta)$ and $w_i(\omega,\theta)$ are periodic in θ, with period 2π, and

$$\frac{d\omega}{d\theta} = \epsilon g_1(\omega) + \epsilon^2 g_2(\omega) + \cdots . \tag{33}$$

This procedure may be regarded as a variant of the method of averaging.[3] The above solution is more general than the periodic solutions constructed previously, since the latter correspond to the case in which ω is constant, rather than a slowly varying function of θ. However, this solution is not completely general, since the initial transients in (4) are not taken into account.

Substituting (31) and (32) into (4) and (5), using (33), and comparing the lowest powers of ϵ, it follows that

$$\omega^2\frac{\partial^2 v_0}{\partial\theta^2} + 2\zeta\omega\frac{\partial v_0}{\partial\theta} + v_0 = \alpha\omega^2\sin\theta \tag{34}$$

and

$$\omega\left(\frac{\partial w_1}{\partial\theta} + g_1\right) + \alpha\omega^2\cos\theta\,\frac{\partial^2 v_0}{\partial\theta^2} = (p - b\omega). \tag{35}$$

The periodic solution of (34) is

$$v_0 = \alpha\omega^2\Delta(\omega)[(1 - \omega^2)\sin\theta - 2\zeta\omega\cos\theta] \tag{36}$$

where

$$\Delta(\omega) = [(1 - \omega^2)^2 + 4\zeta^2\omega^2]^{-1}. \tag{37}$$

In order that w_1 should be periodic, it is necessary from (35), using (36), that

$$\omega g_1(\omega) = [p - b\omega - \alpha^2\zeta\omega^5\Delta(\omega)]. \tag{38}$$

Then $w_1(\omega, \theta)$ may be found from (35), to within an arbitrary function of ω. This arbitrariness is usual in averaging procedures, and may be removed by requiring $\langle w_i(\omega, \theta)\rangle_{\text{av}} = 0$, so that, from (32), ω is the averaged value of Ω. Higher order terms in the asymptotic expansion (31) and (32) may be obtained in a systematic manner.

Now, from (33) and (38),

$$\frac{d\omega}{d\theta} = \frac{\epsilon}{\omega}[p - p^*(\omega)] + O(\epsilon^2) \tag{39}$$

where

$$p^*(\omega) = b\omega + \alpha^2\zeta\omega^5\Delta(\omega). \tag{40}$$

As previously remarked, the case in which ω is constant corresponds to the periodic solutions constructed earlier. From (11), (12), (37), and (40), it follows that ω_0 is the lowest order approximation to a stationary solution of (39). If $\omega \neq \omega_0$, equation (39) determines, to lowest order, the slow variation of ω with θ. The direction in which ω changes is determined, to lowest order in ϵ, by the sign of $[p - p^*(\omega)]$, and is illustrated in Fig. 1 for $p = 0.45$, to which there correspond three values of ω_0, denoted by ω_{0l}, ω_{0m}, and ω_{0r}.

Under more general initial conditions similar results should hold, for sufficiently small ϵ, provided that the initial value of Ω is not too close to ω_{0m}. This is because Ω does not change significantly, for sufficiently small ϵ, during the time in which the initial transients in the translational motion die out.

A partial check of these analytical results was made by Senator,[4] who carried out some numerical solutions of (1) and (2). With $\alpha = 0.707$, $\zeta = 0.2$, $b = 0$ and $\epsilon = 0.1$, he chose initial conditions consistent with the unstable periodic solution corresponding to $p = 0.425$, that is, the periodic solution corresponding to $p^*(\omega_{0m}) = 0.425$. He then carried out numerical solutions of (1) and (2) for $p = 0.45$ and $p = 0.4$. He found that for $p = 0.45$ the solution approaches the periodic solution corresponding to ω_{0r} in the figure, that is, to $p^*(\omega_{0r}) = 0.45$, while for

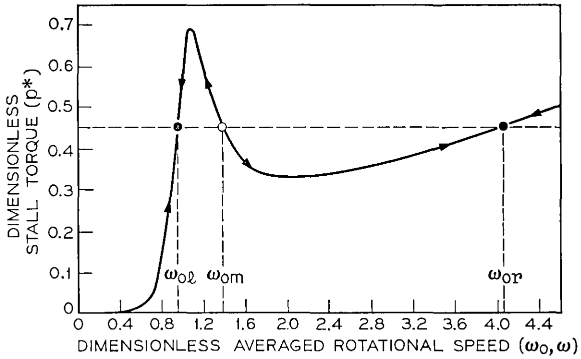


Fig. 1 — Stall torque vs. averaged rotational speed.

$p = 0.4$ the solution approaches the periodic solution corresponding to $p^*(\omega_{0l}) = 0.4$. These results are consistent with our analytical results. Moreover, the number of cycles required before the solution settles down to the appropriate periodic solution was somewhat larger in the case $p = 0.45$ than in the case $p = 0.4$. This is consistent with (39), and the presence of the factor $(1/\omega)$ multiplying $[p - p^*(\omega)]$ therein.

V. ACKNOWLEDGMENT

The author is grateful to M. Senator for discussing the results of this paper, and for carrying out the numerical calculations mentioned above.

REFERENCES

1. Senator, M., "Limit Cycles and Stability of a Nonlinear Two-Degree of Freedom Autonomous Vibratory System," Journal of Engineering for Industry, Trans. ASME, *91*, Series B, No. 4 (November 1969), pp. 959–966.
2. Minorsky, N., *Nonlinear Oscillations,* Princeton: Van Nostrand, 1962, pp. 127–129.
3. Bogoliubov, N. N., and Mitropolsky, Y. A., *Asymptotic Methods in the Theory of Nonlinear Oscillations,* New York: Gordon and Breach, 1962, p. 40.
4. Senator, M., unpublished work.
5. Rocard, Y., *General Dynamics of Vibrations,* New York: Ungar, 1960, trans. 3rd French ed., pp. 362–369.
6. Hoekstra, T. B., "The Response of a Nonlinear Two Degree of Freedom System," Ph.D. dissertation, University of Michigan, Dept. of Eng. Mechanics, 1966, Chapter 5.
7. Kononenko, V. O., "Some Autonomous Problems of the Theory of Nonlinear Oscillations," (in Russian) Trudy Mezhdunarodnovo Simpoziums po Nelineinym Kolebaniyam Izdatel'stvo AN SSSR, *3*, 1963, pp. 151–178.
8. Kononenko, V. O., *Kolebatel'nye Sistemy s Ogranichennym Vozbuzhdeniem,* Moscow, 1964, (L.C. No. QA 871 K7), pp. 51–79.
9. Mazet, R., *Mecanique Vibratoire,* Paris: Dunod, 2nd ed., 1966, pp. 308–318.

The Capacity of Linear Channels with Additive Gaussian Noise

By R. K. MUELLER and G. J. FOSCHINI

(Manuscript received August 8, 1969)

The standard method of computing the mutual information between two stochastic processes with finite energy replaces the processes with their Fourier coefficients. This procedure is mathematically justified here for random signals $w_t(\omega)$ *square-integrable in the product space* $t \times \omega$ *where* $t \in [0, T]$ *and* ω *is an element of a probability space. A natural notion of the sigma field generated by* $w_t(\omega)$ *is presented and it is shown to coincide with the sigma field generated by the random Fourier coefficients of* $w_t(\omega)$ *in any complete orthonormal system in* $L_2[0, T]$. *This justifies the use of Fourier coefficients in mutual information computations.*

Capacity is calculated for finite and infinite-dimensional channels, where the output signal consists of a filter (general Hilbert–Schmidt operator) operating on the input signal with additive Gaussian noise. The finite-dimensional optimal signal is obtained. In the infinite-dimensional case capacity can be approached arbitrarily closely with finite-dimensional inputs. The question of the existence of an infinite-dimensional signal which achieves capacity is considered. There are channels for which no signal achieves capacity. Some results are obtained when the noise coordinates are independent in the eigensystem of the filter.

I. INTRODUCTION

In this paper, we attack a general form of the classical problem of determining the capacity of a linear channel with additive noise. Structurally we have

$$r_t(\omega) = \int_0^T \mathrm{G}(t, \tau) s_\tau(\omega)\, d\tau + n_t(\omega) \tag{1}$$

where the random signals, noise $[n_t(\omega)]$, input $[s_t(\omega)]$, and output $[r_t(\omega)]$ are all defined on $0 \leqq t \leqq T$. All signals as well as the kernel of the channel operator are assumed square integrable in the appropriate product spaces. The noise process, the channel operator, and an average power restriction on $s_t(\omega)$ are assumed to be given. In Section III we begin by defining the capacity of a channel. Our definition is motivated by, but is not a special case of, the generalization of Shannon's notion of capacity that has been indicated by Kolmogorov. The argument for the naturalness of our definition is that any of the above processes can be replaced by their random Fourier coefficients from any expansion using complete orthonormal functions in $L_2[0, T]$. We solve the above problem when $n_t(\omega)$ is Gaussian and independent of $s_t(\omega)$. In Section IV we show that for finite-dimensional inputs there always exists an $s_t(\omega)$ for which capacity is achieved and we find it. The infinite-dimensional case is solved in Section V as a limit of finite-dimensional cases.

II. FUNDAMENTALS

Fundamental to the notion of capacity is the notion of mutual information. We begin with Kolmogorov's definition of the mutual information of two event σ-fields contained in a universal σ-field. Let $\mathcal{A}$ and $\mathcal{B}$ denote two sub σ-fields of a σ-field S_Ω in a probability space (Ω, S_Ω, P). Let α and β denote arbitrary partitions of Ω into a finite number of $\mathcal{A}$ and $\mathcal{B}$ measurable sets A and B. The mutual information $I(\mathcal{A}, \mathcal{B})$ of $\mathcal{A}$ and $\mathcal{B}$ is

$$I(\mathcal{A}, \mathcal{B}) = \sup_{\alpha,\beta} \sum_{A \varepsilon \alpha} \sum_{B \varepsilon \beta} P(A \cap B) \log_e \frac{P(A \cap B)}{P(A)P(B)}. \tag{2}$$

We define $0 \log 0 = 0$. This sum does not decrease as α and β are refined. It can be shown that $I(\mathcal{A}, \mathcal{B}) \geqq 0$ with equality if and only if $\mathcal{A}$ and $\mathcal{B}$ are independent. The nonnegativity and other important properties of I are presented in Ref. 1.

Let $\mathfrak{X}$ be a measurable space with σ-field denoted by $\mathfrak{D}$. A function $\zeta(\omega)$ from (Ω, S_Ω, P) to $\mathfrak{X}$ for which each $D \varepsilon \mathfrak{D}$ has a preimage in S_Ω is called a measurable function.

Let T be an arbitrary index set and let E^1 denote the real line. Endow $\Pi_{t \varepsilon T} E^1$ with the product topology and consider its measurable sets to be the smallest σ-field containing the topology. We are interested in measurable functions from Ω to $\Pi_{t \varepsilon T} E^1$. For our purposes T is either countable or a real compact interval.

Suppose ξ and η are measurable functions from Ω to $\Pi_{t \varepsilon T} E^1$. Then by the mutual information of ξ and η, $I(\xi, \eta)$, we mean the mutual information between the smallest σ-fields with respect to which ξ and η are measurable. We denote these respective σ-fields by $\mathcal{A}_\xi$ and $\mathcal{A}_\eta$.

Let $\zeta(\omega)$ denote any measurable function from Ω to $\Pi_{t\varepsilon T} E^1$. We define the probability distribution P_ζ of $\zeta(\omega)$. The domain of P_ζ is the measurable sets in $\Pi_{t\varepsilon T} E^1$. Let Q be such a measurable set. Then

$$P_\zeta(Q) = P\{\omega : \zeta(\omega) \varepsilon Q\}. \tag{3}$$

If ξ and η are each measurable functions from Ω to $\Pi_{t\varepsilon T_1} E^1$ and $\Pi_{t\varepsilon T_2} E^1$ respectively then (ξ, η) is a measurable function from Ω to $\Pi_{t\varepsilon T_1} E^1 \times \Pi_{t\varepsilon T_2} E^1$ and its distribution function is denoted $P_{\xi,\eta}$. It is called the joint distribution of ξ and η. We can now give an alternate definition of mutual information between ξ and η. Let $\gamma(\delta)$ denote arbitrary partitions of $\Pi_{t\varepsilon T_1}E^1(\Pi_{t\varepsilon T_2}E^1)$ into a finite number of measurable sets $C(D)$. The mutual information $I(\xi, \eta)$ is

$$I(\xi, \eta) = \sup_{\gamma,\delta} \sum_{C\varepsilon\gamma} \sum_{D\varepsilon\delta} P_{\xi\eta}(C \times D) \log \frac{P_{\xi\eta}(C \times D)}{P_\xi(C)P_\eta(D)}. \tag{4}$$

Recall that the inverse image under a measurable function of a σ-field is a σ-field. So it becomes apparent that the two definitions for $I(\xi, \eta)$ are equivalent.

We review without proof some fundamental propositions that will be of use to us later. The following is a result of work by I. M. Gelfand, A. M. Yaglom, and A. Perez.

Theorem: If $P_{\xi\eta}$ is not absolutely continuous with respect to the product measure $P_\xi \times P_\eta$ then $I(\xi, \eta) = \infty$. If $P_{\xi\eta}$ is absolutely continuous with respect to $P_\xi \times P_\eta$, then letting $dP_{\xi\eta}/d(P_\xi \times P_\eta)$ denote the Radon–Nikodym derivative of $P_{\xi\eta}$ with respect to $P_\xi \times P_\eta$ we have

$$I(\xi, \eta) = \int_\Omega \left[\log \frac{dP_{\xi\eta}}{d(P_\xi P_\eta)}\right] dP_{\xi\eta}. \tag{5}$$

Proof: See Ref. 2.

Theorem: Let A be a linear transformation in a k-dimensional vector space and let ξ be a k-dimensional random vector. Then

$$I(\xi, \eta) \geqq I(A\xi, \eta) \tag{6}$$

holds for any random vector η, with equality if the transformation A is nonsingular.

Proof: See Ref. 3.

Theorem: If $I(\xi, \xi) < \infty$, then P_ξ is purely atomic.

Proof: See Ref. 4.

Theorem: *If* $\xi = (\xi_1, \xi_2, \cdots)$, *then*

$$I(\xi, \eta) = \lim_{n \to \infty} I[(\xi_1, \cdots, \xi_n), \eta]. \tag{7}$$

Proof: See Ref. 4.

III. MUTUAL INFORMATION BETWEEN TWO PROCESSES IN $L_2\{(\Omega, S_\Omega, P) \times ([0, T], L, m)\}$

Let $\xi_t(\omega)$ be square integrable on $t \times \omega$. We term $\xi_t(\omega)$ a stochastic process. Notice that it differs from the standard definition of a stochastic process in two ways. First, it is an equivalence class of equal almost everywhere functions in $(t \times \omega)$. Second, not all functions in the equivalence class are stochastic processes in the sense of Ref. 5.; that is, for each t we do not have a random variable but only for almost all t. We assume $E\{\xi_t(\omega)\} = 0$. By Schwarz's inequality and Fubini's theorem it follows that $E\{\xi_{t_1}(\omega)\xi_{t_2}(\omega)\} \varepsilon L_2(t \times t)$. If $\eta_t(\omega)$ and $\zeta_t(\omega)$ are processes of the same type as $\xi_t(\omega)$, $I[\eta_t(\omega), \zeta_t(\omega)]$ is not well defined since $\mathcal{A}_\eta$ and $\mathcal{A}_\zeta$ are not well defined. Because of the central role of these processes in modeling random signals with a finite average power we make $\mathcal{A}_\eta$ and $\mathcal{A}_\zeta$ and hence $I[\eta_t(\omega), \zeta_t(\omega)]$ meaningful here. We need to appeal to the following:

Theorem (F. Riesz): *Let* f_n *converge in measure to* f. *Then there exists a subsequence* f_{n_k} *converging to* f *almost everywhere.*

Proof: See Ref. 6.

Suppose f_n converges in mean square to f. Since convergence in mean square implies convergence in measure, the limit of the subsequence guaranteed by Riesz's theorem is f in the sense that the limit and f agree almost everywhere. This last comment is important since Kolmogorov has given examples of functions g which possess an orthogonal expansion g_n converging in mean square to g, yet pointwise almost everywhere convergence does not occur.

Unless stated otherwise all σ-fields mentioned in the remainder of this section are assumed to be completed. The following new definition is the key to making $\xi_t(\omega)$ meaningful in the information theory sense.

Definition: By the σ-field $\mathcal{A}$ generated by $\xi_t(\omega)$ we mean the smallest σ-field $\mathcal{A}$ satisfying $\xi_t(\omega)$ is $(\Omega, \mathcal{A}) \times ([0, T], L)$ measurable, where L is the sigma field of Lebesgue measurable subsets of $[0, T]$. (This statement is definitive since $\xi_t(\omega)$ is $(\Omega, S_\Omega) \times ([0, T], L)$ measurable and the intersection of σ-fields is a σ-field.)

Proposition I: Suppose

$$\sum_{i \geqq 1} a_i(\omega)\phi_i(t) = \xi_t(\omega) \tag{8}$$

in the mean square sense in the product space (where $a_i(\omega) = \int \xi_t(\omega)\phi_i(t)\,dt$ *and* $\phi_i(t)$ *are orthonormal on* $[0, T]$*). If* $a(\omega) = [a_1(\omega), a_2(\omega), \cdots]$ *then* $\mathcal{A}_\xi = \mathcal{A}_a$.

Proof: Since the expansion converges to $\xi_t(\omega)$ in mean square in the product space, it converges in measure. By F. Riesz's theorem we can find $n_1 < n_2 < \cdots$ so that

$$\lim_{n_k \to \infty} [a_1(\omega)\phi_1(t) + \cdots + a_{n_k}\phi_{n_k}(t)] = \xi_t(\omega). \tag{9}$$

The sum and product of measurable functions is measurable so that each partial sum is $(\Omega, \mathcal{A}_a) \times ([0, T], L)$ measurable. The limit of measurable functions is measurable so $\xi_t(\omega)$ is $(\Omega, \mathcal{A}_a) \times ([0, T], L)$ measurable. Thus $\mathcal{A}_\xi \subset \mathcal{A}_a$.

Next we project $\xi_t(\omega)$ on $\phi_i(t)$ to get

$$a_i(\omega) = \int \xi_t(\omega)\phi_i(t)\,dt. \tag{10}$$

By Fubini's theorem $a_i(\omega)$ is measurable with respect to every σ-field $\mathcal{B}$ for which $\xi_t(\omega)$ is $(\Omega, \mathcal{B}) \times ([0, T], L)$ measurable. But this is true for each i, so $\mathcal{A}_a \subset \mathcal{A}_\xi$.

Proposition I is of paramount importance. In the sequel it enables us to replace $\xi_t(\omega)$ by $a(\omega)$ when computing mutual information.

It would seem appropriate to express $\mathcal{A}_\xi$ without reference to an expansion. The following proposition accomplishes this. However, our proof does resort to an expansion of $\xi_t(\omega)$. Because the proof is similar to the proof of proposition I, we omit it.

Proposition II: Let $\{\xi_t^\alpha(\omega)\}$ *denote the class of functions in* $\xi_t(\omega)$*. Then* $\mathcal{A}_\xi$ *is the smallest σ-field containing* $\bigcap_\alpha \mathcal{A}_{\xi^\alpha}$ [Here we have the only appearance of possibly noncomplete σ- fields (the $\mathcal{A}_{\xi^\alpha}$)].

We can now define capacity of our noisy linear channel. Let S denote a finite average power restriction on $s_t(\omega)$. Then the capacity of the channel is defined as the supremum of $I[s_t(\omega), r_t(\omega)]$ where the supremum is over all $s_t(\omega)$ satisfying

$$E\left[\frac{1}{T}\int_0^T s_t^2(\omega)\,dt\right] \leqq S. \tag{11}$$

We say $\xi_t(\omega)$ is Gaussian if the linear functionals

$$\int_0^T \xi_t(\omega)\phi(t)\,dt \qquad \{\phi(t) \in L_2[0, T]\}$$

are all Gaussian random variables.

IV. THE FINITE-DIMENSIONAL CASE

For a random variable η possessing density p_η the quantity

$$h(\eta) = -\int p_\eta \log p_\eta \tag{12}$$

arises often in mutual information studies. It is called the differential entropy of η.

The following theorem is proved in (Ref. 7).

Theorem: Let p_u be the density of a k-dimensional random variable **u**. *To maximize*

$$h(\mathbf{u}) = -\int p_u \log p_u$$

subject to the conditions that the mean and dispersion matrix have given values $\boldsymbol{\mu}$ and Γ, choose the normal density

$$Q(\mathbf{u}) = 2\pi^{-k/2} \mid \Gamma \mid^{-\frac{1}{2}} \exp\left[\tfrac{1}{2}(\mathbf{u} - \boldsymbol{\mu})'\Gamma^{-1}(\mathbf{u} - \boldsymbol{\mu})\right],$$

which satisfies the conditions.

We prove a corollary necessary for the sequel.

Corollary: Let p_u be the density of a k-dimensional variable **u**. *We want to choose p_u to maximize*

$$h(\mathbf{u}) = -\int p_u \log p_u$$

subject to the conditions that the mean is $\boldsymbol{\mu}$ and the dispersion matrix satisfies the constraint that its trace is less than or equal to ST. The solution is to choose p_u to be Gaussian with mean $\boldsymbol{\mu}$ and covariance ST/kI, where I is the identity matrix.

Proof: From the preceding theorem we only need to consider Gaussian densities. For a Gaussian density we can write the formula

$$h(u) = \frac{k}{2} \log 2\pi e + \tfrac{1}{2} \log \mid \Gamma \mid. \tag{13}$$

Maximizing $h(u)$ is equivalent to maximizing $\mid \Gamma \mid$. Now by the geometric mean—arithmetic mean inequality

$$|\Gamma| \leqq \left[\frac{(\text{trace } \Gamma)}{k}\right]^k \tag{14}$$

with equality if and only if $\gamma_{11} = \gamma_{22} = \cdots \gamma_{kk}$.

Now we are ready to consider the finite-dimensional version of the problem of finding the optimal power-restricted signal $s_t(\omega)$ which maximizes the mutual information between it and the output $r_t(\omega)$. By what we have shown in the previous section we can replace these processes by their Fourier coefficients when computing mutual information. More specifically let $n_t(\omega)$ be a finite-dimensional Gaussian process of dimension k. Let G denote a nonsingular operator on E^k and let $s_t(\omega)$ be a k-dimensional process that is independent of $n_t(\omega)$. Suppose that the distribution of $n_t(\omega)$ is absolutely continuous with respect to Lebesgue measure in E^k. We want to find $s_t(\omega)$ such that its distribution is absolutely continuous with respect to Lebesgue measure in E^k and $I[s_t(\omega), Gs_t(\omega) + n_t(\omega)]$ is maximized subject to

$$E\left[\int_0^T s_t^2(\omega)\, dt\right] \leqq ST.$$

Now by the theorem concerning linear transformations of random vectors stated earlier, $I[s_t(\omega), Gs_t(\omega) + n_t(\omega)] = I[s_t(\omega), s_t(\omega) + G^{-1}n_t(\omega)]$. Define $\eta_t(\omega)$ as $\eta_t(\omega) = G^{-1}n_t(\omega)$ and let

$$\eta^k = \begin{Bmatrix} \eta_1 \\ \cdot \\ \cdot \\ \cdot \\ \eta_k \end{Bmatrix} \quad \text{and} \quad s^k = \begin{Bmatrix} s_1 \\ \cdot \\ \cdot \\ \cdot \\ s_k \end{Bmatrix}$$

be coordinates of $\eta_t(\omega)$ and $s_t(\omega)$.

Then

$$\begin{aligned} I[s_t(\omega), s_t(\omega) + \eta_t(\omega)] &= \int p_{s^k+\eta^k, s^k} \log \frac{p_{s^k+\eta^k, s^k}}{p_{s^k+\eta^k} p_{s^k}} \\ &= \int p_{s^k+\eta^k, s^k} \log \frac{p_{s^k+\eta^k, s^k}}{p_{s^k}} + h(s^k + \eta^k). \end{aligned}$$

Introducing the transformation

$$\begin{Bmatrix} s^k + \eta^k \\ s^k \end{Bmatrix} \rightarrow \begin{Bmatrix} \eta^k \\ s^k \end{Bmatrix}$$

into the above integral and using the fact that s^k and η^k are independent we have

$$I(s^k, s^k + \eta^k)$$

$$= \int p_{s^k+\eta^k} \log \frac{p_{\eta^k, s^k}}{p_{s^k}} + h(s^k + \eta^k) = -h(\eta^k) + h(s^k + \eta^k). \quad (15)$$

Since $h[\eta_t(\omega)]$ is not a function of $s_t(\omega)$, we have reduced the problem to that of maximizing $h[s_t(\omega) + G^{-1}n_t(\omega)]$ subject to

$$E \int_0^T s_t^2(\omega) \leqq ST.$$

Now $\eta_t(\omega)$ is Gaussian and we know from the corollary stated earlier that $h[s_t(\omega) + G^{-1}n_t(\omega)]$ subject to the above constraint is maximized by a Gaussian process. Thus, without loss of generality, $s_t(\omega)$ can be assumed to be Gaussian. We seek Γ_s the covariance of $s_t(\omega)$ so that $| \Gamma_s + G^{-1}\Gamma_n(G^{-1})' |$ is maximized, since this maximizes $h[s_t(\omega) + G^{-1}n_t(\omega)]$. Let us assume, without loss of generality, that $G^{-1}\Gamma_n G^{-1\prime} = \Gamma_\eta$ is diagonal. Thus the problem is to maximize

$$\left| \Gamma_s + \begin{bmatrix} \eta_1 & & 0 \\ & \ddots & \\ 0 & & \eta_k \end{bmatrix} \right|$$

subject to Γ_s a covariance matrix with trace $(\Gamma_s) \leqq ST$. Since we are maximizing a continuous function over a compact set, we know that the maximum exists.

We use induction to show that the optimal Γ_s is diagonal. For $m = 1$ the statement is a trivial one. For $m > 1$ it shall be convenient to partition Γ_s so that

$$\Gamma_s = \begin{bmatrix} \gamma_{11} & \gamma' \\ \gamma & \hat{\Gamma}_s \end{bmatrix}.$$

Let

$$\Gamma = \hat{\Gamma}_s + \begin{bmatrix} \eta_2 & & 0 \\ & \ddots & \\ 0 & & \eta_k \end{bmatrix}.$$

Now using some standard results on determinants (see Ref. 8, p. 46), it follows that

$$|\Gamma_s + \Gamma_\eta| = \begin{vmatrix} \gamma_{11} + \eta_1 & \gamma' \\ \gamma & \Gamma \end{vmatrix} = (\gamma_{11} + \eta_1) \det \Gamma - (\det \Gamma)\gamma'\Gamma^{-1}\gamma. \tag{16}$$

Note that both Γ and Γ^{-1} are positive definite. It is optimal to choose $\gamma = 0$ to maximize the second term. The first term is also optimal if Γ is diagonal. This follows since any nondiagonal Γ with trace $\sum_{i=1}^{k} \gamma_{ii} = ST - \gamma_{11}$ has determinant less than or equal to some diagonal matrix with trace equal to $ST - \gamma_{11}$ by induction. We now optimally select the diagonal elements of Γ_s. If an $\epsilon > 0$ of ST is to be put on a diagonal element of Γ_s, it is optimal to add it to $\min \{(\gamma_{ii} + \eta_i), i = 1, \cdots, k\}$ so that it will have the largest possible multiplier in the determinant of Γ.

V. CAPACITY FOR THE INFINITE-DIMENSIONAL CASE

We turn to calculating the capacity in the situation where there can be an infinite number of Fourier coefficients of $s_t(\omega)$ and $\eta_t(\omega)$ and where the channel G is an infinite-dimensional Hilbert–Schmidt operator.

Define

$$G = \int_0^T \mathbf{G}(t, \tau)$$

where $\mathbf{G}(t, \tau) \in L_2(t \times \tau)$. Let $\{\phi_i\}$ be a complete set of orthonormal eigenfunctions for $G * G$ and let $\{\lambda_i\}$ be the associated eigenvalues. Define $G\phi_i = \psi_i$; then $(\psi_i, \psi_j) = (G\phi_i, G\phi_j) = (\phi_i, G * G\phi_j) = \lambda_i \delta_{ij}$ and so $\{\psi_i/\lambda_i^{\frac{1}{2}}\}$ is an orthonormal set. We use r and η to denote the infinite vector of Fourier coefficients of $r_t(\omega)$ and $\eta_t(\omega)$ in the system $\{\psi_i/\lambda_i^{\frac{1}{2}}\}$ while $\hat{s}$ denotes the infinite vector of Fourier coefficients of $s_t(\omega)$ in the system $\{\phi_i\}$. Let r^k denote the first k coefficients of r and define $\hat{s}^k$ and η^k similarly. Let D be the doubly infinite diagonal matrix with $\lambda_i^{\frac{1}{2}}$ as the ith diagonal element and define D_k to be the $k \times k$ submatrix of D with indices less than or equal to k. Then $r = D\hat{s} + \eta$ and $r^k = D_k\hat{s}^k + \eta^k$.

We first show that if an optimal input signal exists, then there is an optimal Gaussian input signal. We shall need the following lemma.

Lemma: For any signal $\hat{s}$, $\lim I(r^i, \hat{s}^i) = I(r, \hat{s})$.

Proof: We know that $\lim_i \lim_k I(r^i, \hat{s}^k) = I(r, \hat{s})$. As stated earlier this is proved in Ref. 4. Now

$$I(r^i, \hat{s}^i) = h(r^i) + \int p_{r^i, \hat{s}^i} \log \frac{p_{r^i \hat{s}^i}}{p_{\hat{s}^i}}$$

where $r^i = D_i\hat{s}^i + \eta^i$. If $j \geqq i$,

$$\int p_{r^i\hat{s}^j} \log \frac{p_{r^i\hat{s}^j}}{p_{\hat{s}^j}} = \int p_{\eta^i}p_{\hat{s}^j} \log p_{\eta^i} = \int p_{\eta^i} \log p_{\eta^i} .$$

Thus $I(r^i, \hat{s}^j) = I(r^i, \hat{s}^i)$ for $j > i$; then $\lim_{j\to\infty} I(r^i, \hat{s}^j) = I(r^i, \hat{s}^i) = I(r^i, \hat{s})$. Finally $\lim_{i\to\infty} I(r^i, \hat{s}) = \lim_{i\to\infty} I(r^i, \hat{s}^i) = I(r, \hat{s})$.

Alternately this lemma can be proved by extending some results of Ref. 3 to the infinite-dimensional case.

We now show that if an optimal signal exists for the infinite-dimensional case, then the optimal signal can be assumed, without loss of generality, to be Gaussian.

Proposition III: If $\hat{s}_1$ *is a non-Gaussian optimal signal then* $\hat{s}_2$ *, the Gaussian signal with the same covariance matrix as* $\hat{s}_1$ *, is optimal.*

Proof: Clearly $I(r_1^k, \hat{s}_1^k) \leqq I(r_2^k, \hat{s}_2^k)$ for all k since the Gaussian process is optimal for a fixed covariance matrix in the finite-dimensional case. Thus $I(r_1, \hat{s}_1) = \lim_k I(r_1^k, \hat{s}_1^k) \leqq I(r_2, \hat{s}_2) = \lim_k I(r_2^k, \hat{s}_2^k)$.

Proposition IV: The capacity of the infinite-dimensional channel is the limit of the capacities of the k dimensional truncated approximation of the infinite-dimensional channel.

Proof: Let C_k denote the capacity of the k dimensional channel and let $C = \lim C_k$. We claim the capacity of the infinite-dimensional channel is C. It is evidently at least C. Suppose a signal $s_t(\omega)$ exists satisfying the constraints with mutual information, $I(r, \hat{s})$ greater than C. Then $I(r^k, \hat{s}^k) \leqq C_k$ since $\hat{s}^k$ satisfies the power constraint. Thus $I(r, \hat{s}) \leqq C$, a contradiction.

Corollary 1: There exist finite-dimensional signals whose resulting mutual information is arbitrarily close to the capacity.

Corollary 2: If C_k *is constant for all* k *larger than some integer* l*, then the* $l + 1$*-dimensional optimal signal is optimal for the infinite-dimensional case.*

5.1 *Limiting Covariance Matrices and Optimal Signals When* $\{\eta_i\}$ *Is Independent*

It is not always true that some input signal achieves capacity in the infinite-dimensional case. We first prove this. Then we study the special case when $\{\eta_i\}$ is independent in the $\{\psi_i/\lambda_i^{\frac{1}{2}}\}$ system. This case may be of marginal interest insofar as a model of a realistic system. However it is mathematically tractable and hence serves as a good testing ground for intuition into more general behavior.

We now show that no optimal input signal exists for the case $\lambda_i = 1/i^2$, $E\eta_i^2/\lambda_i = 1$. It is clear that $C_k = \frac{1}{2} \log (1 + ST/k)^k$. Then the capacity is: $C = \frac{1}{2} \lim_{k\to\infty} \log (1 + ST/k)^k = ST/2$. If there exists an optimal signal s, $I(r^i, \hat{s}^i) \to ST/2$. But $I(r^i, \hat{s}^i) = \frac{1}{2} \log | \Gamma_{\hat{s}^i} + I_i |$, where I_i is an $i \times i$ identity matrix. Then $I(r^i, \hat{s}^i) \leqq \frac{1}{2} \sum_{j=1}^{i} \log (1+Es_j^2)$ and $\lim I(r^i, \hat{s}^i) \leqq \frac{1}{2} \sum_{j=1}^{\infty} \log (1 + Es_j^2)$. Recall that $\sum_{j=1}^{\infty} Es_j^2 = ST$ by assumption. We show that $\sum_{j=1}^{\infty} \log (1 + Es_j^2) < ST$. Since $Es_j^2 \geqq \log (1 + Es_j^2)$ with equality if and only if $Es_j^2 = 0$, $\lim I(r^i, \hat{s}^i) \leqq \frac{1}{2} \sum_{j=1}^{\infty} \log (1 + Es_j^2) < \frac{1}{2} \sum_{j=1}^{\infty} Es_j^2 = ST/2$.

Although an optimal signal does not always exist, we can say when it does exist in the special case when the $\{\eta_i\}$ are independent in the system $\{\psi_i/\lambda_i^{\frac{1}{2}}\}$. It will turn out that $\{\hat{\Gamma}_{\hat{s}^k}\}$, the sequence of finite-dimensional optimal covariance matrices for s, converges in some cases to an optimal solution and in other cases the limit is not optimal. The diagonal matrix with $a_i = (1/\lambda_i)E\eta_i^2$ on the ith diagonal element completely determines whether or not an optimal limit is reached.

We define the order of minima of a sequence $\{\xi_i\}_{i=1}^{\infty}$ as follows. The order is 0.5 if no smallest element in $\{\xi_i\}_{i=1}^{\infty}$ exists. If M_1 is defined to be the set of smallest elements in $\{\xi_i\}_{i=1}^{\infty}$ and *Card* $(M_1) = +\infty$, the order of the sequence is 1. If *Card* $(M_1) < +\infty$ but the set $\{\bigcup \xi_i - M_1\}$ has no least element, the order is 1.5. If the set $\{\bigcup \xi_i - M_1\}$ has M_2 smallest elements and *Card* $(M_2) = +\infty$, the order is 2. If *Card* $(M_2) < +\infty$ but the set

$$\left\{\bigcup \xi_i - \bigcup_{j=1}^{2} M_j\right\}$$

has no least element then the order is 2.5, and so on. If the sequence is not assigned a finite order of minima, the order is infinite.

If the order of minima of $\{a_i\}$ is 0.5, $\{\hat{\Gamma}_{\hat{s}^k}\} \to [0]$. To see this we need only consider diagonal elements of $\hat{\Gamma}_{\hat{s}^k}$. Suppose for some j and for some $\epsilon > 0$, $E\hat{s}_j^2 \geqq \epsilon$ in an infinite number of $\hat{\Gamma}_{\hat{s}^k}$. Since no smallest element in $\{a_i\}$ exists, there are an infinite number of a_i, say $\{a_{i'}\}$ smaller than a_j. Then in the optimal covariance matrices where $E\hat{s}_j^2 \geqq \epsilon$ and the i' appear, $Es_{i'}^2 \geqq \epsilon$. But this is not possible with the constraint $\sum E\hat{s}_i^2 \leqq ST$. Thus for each j and $\epsilon > 0$, $E\hat{s}_j^2 < \epsilon$ in all but a finite number of $\hat{\Gamma}_{\hat{s}^k}$.

If the order of the minima of $\{a_i\}$ is 1, $\hat{\Gamma}_{\hat{s}^k} \to [0]$. This follows since it is optimal to put the power on the minima. After some k, only the minima will have positive $E\hat{s}_i^2$. Since there are an infinite number of them and the ST is optimally distributed equally on them, $\hat{\Gamma}_{sk} \to [0]$.

If the order of the minima of $\{a_i\}$ is 1.5, there are two cases. Let $h = \inf \{\bigcup \xi_i - M_1\}$ and $g = \inf \{\bigcup \xi_i\}$. If $(h - g)$ *Card* $(M_1) \geqq ST$, there is an optimal solution as a limit consisting of $E\hat{s}_i^2 = ST/\mathit{Card}\ (M_1)$ for those i corresponding to $a_i \in M_1$. Otherwise the covergence is to a matrix where $Es_i^2 = h - g$ if $a_i \in M_1$ and zero elsewhere, which is clearly not optimal. The analysis for other finite order systems are analogous to the above. Either (*i*) ST is distributed over a finite number of components, in which case the convergence is to an optimal solution, or (*ii*) ST has to be distributed over an infinite number of components, in which case the convergence is not to an optimal solution.

If the order is infinite and we run out of the quantity ST on a finite number of components, the resulting finite-dimensional solution is optimal. Suppose the order is infinite and we do not run out of ST on a finite number of dimensions. Let θ be the smallest accumulation point of $\{a_i\}$. If not all of ST is used in making

$$\frac{Er_i^2}{\lambda_i} = \theta,$$

the limiting covariance is not optimal and no optimal covariance which achieves capacity may be constructed. This follows since a finite amount of ST must be distributed equally to an infinite number of components. If all of the ST is exactly used to make

$$\frac{Er_i^2}{\lambda_i} = \theta,$$

the limiting covariance is optimal. Before proving this we give an example of such a case.

Let

$$\lambda_i = \frac{(i + 1)}{(i + 1)^3 - 1}, \qquad E\eta_i^2 = \frac{1}{(i + 1)^2}$$

and assume the η_i are independent. Then $a_i = 1 - 1/(i + 1)^3$. To bring all components

$$\frac{Er_i^2}{\lambda_i}$$

to 1 we need

$$ST = \sum_{i=1}^{\infty} \frac{1}{(i + 1)^3},$$

and we are then in the case considered above.

We now show that the limiting covariance matrix for the case when there is just enough ST to bring

$$\frac{Er_i^2}{\lambda_i} = \theta$$

is optimal. Let Γ_{s_1} be the limiting matrix with the corresponding Gaussian process $\hat{s}_1$. Let $r_1 = D\hat{s}_1 + \eta$. We show that $I(r_1^k, \hat{s}_1^k) \to C$, $k \to \infty$. Now suppose $\hat{s}^k$ is optimal for k-dimensions. Then

$$I(r^k, \hat{s}^k) - I(r_1^k, \hat{s}_1^k) = h(r^k) - h(r_1^k) = \tfrac{1}{2} \log \left[\theta + \frac{\delta(k)}{k}\right]^k \prod_{i=1}^{k} \lambda_i - \tfrac{1}{2} \log \theta^k \prod_{i=1}^{k} \lambda_i . \tag{17}$$

Here $\delta(k)$ equals that part of ST not used in the matrix Γ_{s_1} in the first k-dimensions. Clearly we are assuming that the smallest elements of a_i appear first. Notice that $\delta(k) \to 0$ as $k \to \infty$. Then

$$I(r^k, \hat{s}^k) - I(r_1^k, \hat{s}_1^k) = \frac{k}{2} \log \left[1 + \frac{\delta(k)}{k\theta}\right] \leqq \frac{\delta(k)}{2\theta} \tag{18}$$

for k sufficiently large. Then

$$\lim I(r^k, \hat{s}^k) = \lim I(r_1^k, \hat{s}_1^k) = I(r_1, \hat{s}_1) = C. \tag{19}$$

VI. SUMMARY

Let us review what we have done. Since we chose to deal with signals $\xi_t(\omega)$ square-integrable on $L_2\{(\Omega, S_\Omega, P) \times ([0, T], L, m)\}$, we define the mutual information between two such signals using Proposition II and equation (2) in such a way that it agrees with the mutual information of their Fourier coefficients defined in equation (10). For the channel defined in equation (1) with input signals constrained by equation (11), we calculate the capacity of the channel. First in Section IV the capacity problem is considered when only a finite number of Fourier coefficients are nonzero. We use the corollary to the theorem in Section IV and equation (15) to show that only Gaussian signals have to be considered. Then equation (16) is used to calculate the finite-dimensional optimal signal by "filling the well." In Section V the case of an infinite number of nonzero Fourier coefficients is considered. We show in Proposition III that optimal signals, if they exist, can be chosen Gaussian. In Proposition IV the capacity of the infinite-dimensional channel is calculated as the limit of finite-dimensional capacities.

Finally in Section 5.1 we deal with the existence of an optimal signal. In general no optimal signal exists. A special case is examined when the noise components are independent in a fixed coordinate system.

APPENDIX

Symbols Used

The following is a list of symbols used throughout the text.

L_2—the set of square-integrable functions
$n_t(\omega)$—a noise process
$\eta_t(\omega)$—a noise process
$s_t(\omega)$—the input signal process
$r_t(\omega)$—the output process
G—the linear channel operator
G^*—the adjoint of G
P_ξ—a probability measure generated by ξ
p_η—the probability density of the random vector η
$\mathcal{A}$—a sigma field
$I(\xi, \eta)$—the mutual information between ξ and η
$h(\eta)$—the differential entropy of η
Γ_s—the covariance of s
E^k—Euclidean k-space
$|\Gamma|$—the determinant of Γ
L—the Lebesgue measurable sets
m—Lebesgue measure
Card—Cardinality
E—expected value

REFERENCES

1. Lloyd, S. P., "On a Measure of Stochastic Dependence," Теория вероятностей и ее применения (Theory of Probability and Its Application), 7 (1962), pp. 312–322
2. Ghurye, S. G., "Information and Sufficient Subfields," Annals of Math. Stat., *39,* No. 6 (December 1968), pp. 2056–2066.
3. Gelfand, I. M., and Yaglom, A. M., "Calculation of the Amount of Information About a Random Function Contained in Another Such Function," Translations of the Amer. Math. Soc., (June 1959), pp. 199–246.
4. Pinsker, M. S., *Information and Information Stability of Random Processes,* San Francisco: Holden Day, 1964.
5. Doob, J. L., *Stochastic Processes,* New York: John Wiley, 1953.
6. Natanson, I. P., *Theory of Functions of a Real Variable,* New York: Ungar, 1955.
7. Rao, C. R., *Linear Statistical Inference and Its Applications,* New York: John Wiley, 1965.
8. Gantmacher, F. R., *The Theory of Matrices,* New York: Chelsea, 1959.

Theorems on the Analysis of Nonlinear Transistor Networks*

By I. W. SANDBERG

(Manuscript received August 19, 1969)

This paper reports on further results concerning nonlinear equations of the form $F(x) + Ax = B$, *in which* $F(\cdot)$ *is a "diagonal nonlinear mapping" of real Euclidean n-space* E^n *into itself,* A *is a real* $n \times n$ *matrix, and* B *is an element of* E^n. *Such equations play a central role in the dc analysis of transistor networks, the computation of the transient response of transistor networks, and the numerical solution of certain nonlinear partial-differential equations.*

Here a nonuniqueness result, which focuses attention on a simple special property of transistor-type nonlinearities, is proved; this result shows that under certain conditions the equation $F(x) + Ax = B$ *has at least two solutions for some* $B \in E^n$. *The result proves that some earlier conditions for the existence of a unique solution cannot be improved by taking into account more information concerning the nonlinearities, and therefore makes more clear that the set of matrices denoted in earlier work by* P_0 *plays a very basic role in the theory of nonlinear transistor networks. In addition, some material concerned with the convergence of algorithms for computing the solution of the equation* $F(x) + Ax = B$ *is presented, and some theorems are proved which provide more of a theoretical basis for the efficient computation of the transient response of transistor networks. In particular, the following proposition is proved.*

If the dc equations of a certain general type of transistor network possess at most one solution for all $B \in E^n$ *for "the original set of* α*'s as well as for an arbitrary set of not-larger* α*'s", then the nonlinear equations encountered at each time step in the use of certain implicit numerical integration algorithms possess a unique solution for all values of the step size, and hence then for all step-size values it is possible to carry out the algorithms.*

---

* The material of this paper was presented at the Advanced Study Institute on Network Theory (sponsored by the N.A.T.O.; Knokke, Belgium; September 1-12, 1969).

I. INTRODUCTION AND DISCUSSION OF RESULTS

References 1 and 2 present some results concerning the equation

$$F(x) + Ax = B, \tag{1}$$

in which, with n an arbitrary positive integer, A is a real $n \times n$ matrix, B is an element of real Euclidean n-space E^n, and $F(\cdot)$ is a mapping of E^n into E^n defined by the condition that* for all $x = (x_1, x_2, \cdots, x_n)^{tr} \in E^n$,

$$F(x) = [f_1(x_1), f_2(x_2), \cdots, f_n(x_n)]^{tr} \tag{2}$$

with each $f_i(\cdot)$ a strictly monotone increasing mapping of E^1 into itself. Equation (1) plays a central role in the dc analysis of transistor networks,** the transient analysis of transistor networks (see Section 1.4), and the numerical solution of certain nonlinear partial differential equations.

In Ref. 1 it is proved that there exists a unique solution x of equation (1) for each strictly monotone increasing mapping $F(\cdot)$ of E^n onto E^n (that is, for each set of strictly monotone increasing mappings $f_i(\cdot)$ of E^1 onto itself) and each $B \in E^n$ if and only if A is a member of the set P_0 of real $n \times n$ matrices with all principal minors nonnegative. It is also proved in Ref. 1 that equation (1) possesses a unique solution x for each continuous monotone nondecreasing mapping $F(\cdot)$ of E^n into E^n (that is, for each set of continuous monotone nondecreasing mappings of E^1 into E^1) and each $B \in E^n$ if A belongs to the set P of all real $n \times n$ matrices with all principal minors positive†. A direct modification of the existence proof given in Ref. 1, as indicated in Ref. 2, shows that equation (1) possesses a unique solution for each strictly monotone increasing mapping $F(\cdot)$ of E^n onto $(\alpha_1, \beta_1) \times (\alpha_2, \beta_2) \times \cdots \times (\alpha_n, \beta_n)$ with each α_i and β_i elements of the extended real line‡ (real line) such that $\alpha_i < \beta_i$ and each $B \in E^n$ if (and only if) $A \in P_0$ and $\det A \neq 0$. Some network theoretic implications of these and related results are discussed in Refs. 1 and 2, where the matter of determining whether or not $A \in P_0$ or $A \in P$ is considered in some detail.

---

* Throughout the paper the superscript *tr* denotes transpose.

** See Ref. 1 for a derivation of the equation within the context of the transistor dc-analysis problem.

† There are some interesting applications of this result in the study of numerical methods for solving certain nonlinear partial-differential equations, in which A has nonpositive off-diagonal terms and is irreducibly diagonally dominant.[3]

‡ The numbers α_i and β_i are members of the *extended real line* if $-\infty \leqq \alpha_i \leqq \infty$ and $-\infty \leqq \beta_i \leqq \infty$.

This paper presents a proof of a nonuniqueness result. The proof focuses attention on a simple special property of transistor-type nonlinearities. The result shows that under certain conditions equation (1) has at least two solutions for some $B \in E^n$. In addition, the paper presents some material concerned with the convergence of algorithms for computing the solution of equation (1) and proves some theorems which provide more of a theoretical basis for the efficient computation of the transient response of transistor networks. The remaining portion of Section I is concerned with a detailed discussion of the results and their significance.

1.1 *An Application of the Nonuniqueness Theorem*

The standard Ebers–Moll transistor model, which is widely used, gives rise to functions $f_i(\cdot)$ which, while continuous and strictly monotone increasing, are mappings of E^1 onto open semi-infinite intervals. For such $f_i(\cdot)$, the results stated above assert that the equation (1) possesses at most one solution x for each $B \in E^n$ if $A \in P_0$; and if $A \in P_0$ and det $A \neq 0$, then equation (1) possesses a solution for each $B \in E^n$. Since, as indicated in Ref. 1, $A = T^{-1}G$ with T a nonsingular matrix which takes into account the forward and reverse transistor α's, and G is the short circuit conductance matrix of the linear portion of the network, the condition that det A not vanish is equivalent to the rather weak assumption that the linear portion of the network possess an open-circuit resistance matrix.

It is natural to ask whether the use of more-detailed information concerning the nonlinearities of the transistor model would enable us to make assertions concerning the existence of a unique solution of equation (1) for all $B \in E^n$ under weaker assumptions on A. In particular, can the condition that A belong to P_0 be relaxed? The first result proved in this paper, Theorem 1 of Section II, shows that if the $f_i(\cdot)$ are exponential nonlinearities of the type associated with the Ebers–Moll model, then the condition that A belong to P_0 cannot be replaced by a weaker condition. More explicitly, in Section II a set $\mathfrak{F}_0^n$ of mappings of E^n into E^n is defined, and $\mathfrak{F}_0^n$ contains all of the mappings $F(\cdot)$ that correspond to Ebers–Moll type $f_i(\cdot)$'s. It is proved there that if $A \notin P_0$, then for *any* $F(\cdot) \in \mathfrak{F}_0^n$, there is a $B \in E^n$ such that equation (1) possesses at least two solutions. In fact, it is proved that if $A \notin P_0$ and if δ is an arbitrary positive number, then for any $F(\cdot) \in \mathfrak{F}_0^n$, there is a $B \in E^n$ such that equation (1) possesses two solutions such that the distance in E^n between the two solutions is δ.

Thus Theorem 1 together with the earlier results mentioned above concerning existence of solutions show that the set of matrices P_0 plays a quite fundamental role in the theory of nonlinear transistor networks.

1.2 *An Algorithm for Computing the Solution of Equation (1)*

Several results which assert that $A \in P_0$ under certain conditions on the transistor α's and the short-circuit conductance matrix of the linear portion of the network are proved in Refs. 1 and 2. In particular, Ref. 1 proves that $A \in P$, and hence that $A \in P_0$, if $A = P^{-1}Q$ with P and Q real $n \times n$ matrices such that for all $j = 1, 2, \cdots, n$

$$p_{jj} > \sum_{i \neq j} | p_{ij} | \quad \text{and} \quad q_{jj} > \sum_{i \neq j} | q_{ij} |.^{*}$$

Theorem 2 of Section II shows that a relatively simple and entirely constructive algorithm can be used to generate a sequence $x^{(0)}, x^{(1)}, \cdots$ of elements of E^n that converges to the unique solution of (1) if $A = P^{-1}Q$ with P and Q as defined above and each $f_i(\cdot)$ is a continuous (but not necessarily differentiable) monotone nondecreasing mapping of E^1 into E^1.[**]

1.3 *Palais' Theorem, Existence of Solutions of Equation (1), and Algorithms for Computing the Solution of Equation (1)*

Reference 1 gives two existence proofs concerning equation (1). One proof, the more basic of the two, is based on first principles and employs an inductive argument in which, with k an arbitrary positive integer less than n, the existence proposition is assumed to be true with n replaced by k and it is proved that then the proposition is true with n replaced by $(k + 1)$. The second proof uses a theorem of R. S. Palais and requires that the $f_i(\cdot)$ be continuously differentiable throughout E^1. More explicitly, Palais' theorem[†] asserts that if $R(\cdot)$ is a continuously differentiable mapping of E^n into itself with values $R(q)$ for $q \in E^n$, then $R(\cdot)$ is a diffeomorphism[‡] of E^n onto itself if and only if

(*i*) $\det J_q \neq 0$ for all $q \in E^n$, in which J_q is the Jacobian matrix of $R(\cdot)$ with respect to q, and

(*ii*) $\| R(q) \| \to \infty$ as $\| q \| \to \infty$.[††]

---

[*] It is proved also that $A \in P_0$ if $A = P^{-1}Q$ with $p_{jj} > \sum_{i \neq j} | p_{ij} |$ and $q_{jj} \geqq \sum_{i \neq j} | q_{ij} |$ for all j.

[**] A related result given in Ref. 4 is not directly applicable here because of assumptions made in Ref. 4 concerning the existence and boundedness of a certain Jacobian matrix.

[†] See Ref. 5 and the appendix of Ref. 6.

[‡] A diffeomorphism of E^n onto itself is a continuously differentiable mapping of E^n into E^n which possesses a continuously differentiable inverse.

[††] Here $\| \cdot \|$ denotes any norm on E^n.

And the second proof of Ref. 1 shows that, with

$$R(q) = F(q) + Aq$$

for all $q \in E^n$, the two conditions (*i*) and (*ii*) are met when $A \in P_0$ and each $f_i(\cdot)$ is a continuously differentiable strictly-monotone-increasing function which maps E^1 onto E^1 and whose slope is positive throughout E^1.*

There are some problems which arise in connection with, for example, the numerical solution of certain nonlinear partial-differential equations** in which one encounters an equation of the form (1) with $A \in P_0$ and $\det A \neq 0$, but with functions $f_i(\cdot)$ which, while continuously differentiable, are monotone nondecreasing (rather than strictly monotone increasing) mappings of E^1 into E^1. We can prove that even in such cases equation (1) possesses a unique solution for each $B \in E^n$ as follows. Here the Jacobian matrix of $F(q)+Aq$ exists and is of the form $D(q)+A$ in which $D(q)$ is a diagonal matrix with nonnegative diagonal elements. Since $A \in P_0$ and $\det A \neq 0$, we have[2] $\det [D(q) + A] \neq 0$ for all $q \in E^n$. An immediate application of Theorem 3 of Section II shows that $\| F(q) + Aq \| \to \infty$ as $\| q \| \to \infty$.† Therefore, by Palais' theorem, $F(x) + Ax = B$ possesses a unique solution for each B.

Theorem 3 is of use not only in connection with the proof given in the preceding paragraph; it also plays a key role in showing that there is an algorithm which generates a sequence of elements of E^n $x^{(0)}, x^{(1)}, \cdots$ that converges to the unique solution of $F(x) + Ax = B$ whenever each $f_i(\cdot)$ is twice continuously differentiable on E^1 and the conditions on A and $F(\cdot)$ of the preceding paragraph are satisfied.‡

More generally, if $R(\cdot)$ is any twice-differentiable mapping of E^n into itself such that conditions (*i*) and (*ii*) of Palais' theorem are satisfied, then, with $R^{-1}(\cdot)$ the continuously-differentiable inverse of $R(\cdot)$, $x = R^{-1}(\theta)$ satisfies $R(x) = \theta$ in which θ is the zero element of E^n, and there are steepest decent as well as Newton-type algorithms each of which generates a sequence in E^n that converges to x. To show this, let[7] $f(y) = R(y) \|^2$ for all $y \in E^n$ in which $\| \cdot \|$ denotes the usual Euclidean norm (that is, the square-root of the sum of squares). Since condition (*i*) of Palais' theorem is satisfied, the gradient ∇f of $f(\cdot)$ satisfies $(\nabla f)(y) \neq \theta$ unless $f(y) = 0$,[*] and since condition (*ii*) of Palais' theorem is satisfied the set $S = \{y \in E^n : f(y) \leqq f(x^{(0)}\}$ is bounded for any $x^{(0)} \in E^n$. Therefore we may appeal to, for example, the theorem of page 43 of Ref. 7 according to which for any $x^{(0)} \in E^n$, for any member of a certain class of mappings $\varphi(\cdot)$ of S into E^n, and for suitably chosen constants γ_0, γ_1, $\cdots$, the sequence $x^{(0)}, x^{(1)}, \cdots$ defined by

---

* The reasons that two proofs were presented in Ref. 1, with the second proof a proof of a somewhat weaker result, are that the arguments needed for the application of Palais' theorem had already been developed in Ref. 1 and used for other purposes there, and it was felt desirable to indicate an alternative approach to essentially the same problem.

** The writer is indebted to J. McKenna and E. Wasserstrom for bringing this fact to his attention.

† More explicitly, Theorem 3 shows that there is a vector $C \in E^n$ such that $\| F(q) + Aq + C \| \to \infty$ as $\| q \| \to \infty$, which is equivalent to the statement concerning $\| F(q) + Aq \|$ made above.

‡ The differentiability assumption here is introduced as a matter of convenience, and is certainly satisfied when the $f_i(\cdot)$ are Ebers–Moll exponential-type nonlinearities.

$$x^{(k+1)} = x^{(k)} + \gamma_k \varphi(x^{(k)}) \quad \text{for all} \quad k \geqq 0$$

belongs to S and is such that $\| R(x^{(k)} \| \rightarrow 0$ as $k \rightarrow \infty$. However, since $R^{-1}(\cdot)$ exists and is continuous,[†] it follows from

$$x^{(k)} = R^{-1}[R(x^{(k)})] \quad \text{for all} \quad k \geqq 0$$

and the fact that $R(x^{(k)}) \rightarrow \theta$ as $k \rightarrow \infty$, that $\lim_{k\rightarrow\infty} x^{(k)}$ exists and

$$\lim_{k\rightarrow\infty} x^{(k)} = R^{-1}(\theta),$$

which means that $x = \lim_{k\rightarrow\infty} x^{(k)}$.[‡]

1.4 *Transient Response of Transistor-Diode Networks and Implicit Numerical-Integration Formulas*

At this point we briefly consider some aspects of the manner in which the previous material bears on the important problem of providing more of a theoretical basis for numerically integrating the ordinary differential equations which govern the transient response of nonlinear transistor networks. Although we consider explicitly only networks containing transistors, diodes, and resistors, the material to be presented can be extended to take into account other types of elements as well. In addition, we shall focus attention on the use of linear multipoint integration formulas of closed (that is, of implicit) type, since such formulas are of considerable use in connection with the typically "stiff systems" of differential equations encountered.

---

[*] Here we have used the fact that $(\nabla f)(y) = 2J_y{}^{tr}R(y)$ for all $y \in E^n$.[7]

[†] By Palais' theorem $R(\cdot)$ is a diffeomorphism of E^n onto itself.

[‡] The material of the second part of Section 1.3 was motivated by previous recent work of the writer's colleague A. Gersho who made the observation that the convergence of an algorithm for the solution of equation (1) could be shown by combining results of Ref. 1 with the approaches described by Goldstein.[7] (See the November 1969 B.S.T.J. Brief by A. Gersho.)

A very large class of networks containing resistors, transistors, and diodes modeled in a standard manner is governed by the equation[8]

$$\frac{du}{dt} + TF[C^{-1}(u)] + (I + GR)^{-1}GC^{-1}(u) = B(t), \qquad t \geqq 0 \tag{3}$$

where, assuming that there are q diodes and p transistors,

(*i*) $T = I_q \oplus T_1 \oplus T_2 \oplus \cdots \oplus T_p$, the direct sum of the identity matrix of order q and p 2×2 matrices T_k in which

$$T_k = \left\{\begin{matrix} 1 & -\alpha_r^{(k)} \\ -\alpha_f^{(k)} & 1 \end{matrix}\right\}$$

with $0 < \alpha_r^{(k)} < 1$ and $0 < \alpha_f^{(k)} < 1$ for $k = 1, 2, \cdots, p$.

(*ii*) $R = R_0 \oplus R_1 \oplus R_2 \oplus \cdots \oplus R_p$, the direct sum of a diagonal matrix $R_0 = \text{diag}\,(r_1, r_2, \cdots, r_q)$ with $r_k \geqq 0$ for $k = 1, 2, \cdots, q$ and p 2×2 matrices R_k in which for all $k = 1, 2, \cdots, p$

$$R_k = \left\{\begin{matrix} r_e^{(k)} + r_b^{(k)} & r_b^{(k)} \\ r_b^{(k)} & r_c^{(k)} + r_b^{(k)} \end{matrix}\right\}$$

with $r_e^{(k)} \geqq 0$, $r_b^{(k)} \geqq 0$, and $r_c^{(k)} \geqq 0$. (The matrix R takes into account the presence of bulk resistance in series with the diodes and the emitter, base, and collector leads of the transistors.)

(*iii*) G is the short-circuit conductance matrix associated with the resistors of the network. (It does not take into account the bulk resistances of the semiconductor devices.)

(*iv*) $F(\cdot)$ is a mapping of $E^{(2p+q)}$ into $E^{(2p+q)}$ defined by the condition that

$$F(x) = [f_1(x_1), f_2(x_2), \cdots, f_{2p+q}(x_{2p+q})]^{tr}$$

for all $x \in E^{(2p+q)}$ with each $f_i(\cdot)$ a continuously differentiable strictly-monotone increasing mapping of E^1 into E^1.

(*v*) $C^{-1}(\cdot)$ is the inverse of the mapping $C(\cdot)$, of $E^{(2p+q)}$ into itself, defined by

$$C(x) = \text{diag}\,(c_1, c_2, \cdots, c_{2p+q})x + \text{diag}\,(\tau_1, \tau_2, \cdots, \tau_{2p+q})F(x)$$

for all $x \in E^{(2p+q)}$ with each c_i and each τ_i a positive constant.

(*vi*) $B(t)$ is a $(2p + q)$-vector which takes into account the voltage and current generators present in the network, and (*vii*) u is related to v the vector of junction voltages of the semiconductor devices through $C(v) = u$ for all $v \in E^{(2p+q)}$.

Equation (3) is equivalent to*

$$\dot{u} + f(u, t) = \theta_{(2p+q)}, \qquad t \geqq 0 \tag{4}$$

in which of course

$$f(u, t) = TF[C^{-1}(u)] + (I + GR)^{-1}GC^{-1}(u) - B(t) \tag{5}$$

and $\theta_{(2p+q)}$ is the zero vector of order $(2p + q)$.

It is well known that certain specializations of the general multipoint formula[9,10]

$$y_{n+1} = \sum_{k=0}^{r} a_k y_{n-k} + h \sum_{k=-1}^{r} b_k \tilde{y}_{n-k} \tag{6}$$

in which

$$\tilde{y}_{n-k} = -f[y_{n-k}, (n - k)h] \tag{7}$$

can be used as a basis for computing the solution of equation (4). Here h, a positive number, is the step size, the a_k and the b_k are real numbers, and of course y_n is the approximation to $u(nh)$ for $n \geqq 1$.

In the literature dealing with formulas of the type (6) in connection with systems of equations of the type (4), information concerning the location of the eigenvalues of the Jacobian matrix J_u of $f(u, t)$ with respect to u plays an important role in determining whether or not a given formula will be (in some suitable sense) stable. In particular, an assumption often made is that all of the eigenvalues of J_u lie in the strict right-half plane for all $t \geqq 0$ and all u. For $f(u, t)$ given by equation (5), we have

$$\begin{aligned} J_u = {} & T \operatorname{diag}\left\{\frac{f_j'[g_j(u_j)]}{c_j + \tau_j f_j'[g_j(u_j)]}\right\} \\ & + (I + GR)^{-1}G \operatorname{diag}\left\{\frac{1}{c_j + \tau_j f_j'[g_j(u_j)]}\right\} \end{aligned} \tag{8}$$

in which for $j = 1, 2, \cdots, (2p + q)$ $g_i(u_j)$ is the j^{th} component of $C^{-1}(u)$. Thus here J_u is a matrix of the form

$$TD_1 + (I + GR)^{-1}GD_2 \tag{9}$$

where D_1 and D_2 are diagonal matrices with positive diagonal elements.

---

* Ref. 8 shows that if $B(\cdot)$ is a continuous mapping of $[0, \infty)$ into $E^{(2p+q)}$, then for any initial condition $u^{(0)} \in E^{(2p+q)}$ there exists a unique continuous $(2p + q)$- vector-valued function $u(\cdot)$ such that $u(0) = u^{(0)}$ and (3) is satisfied for all $t > 0$.

A simple result concerning equation (9), Theorem 4 of Section II, asserts that if there exists a diagonal matrix D with positive diagonal elements such that*

(*i*) DT is strongly column-sum dominant, and
(*ii*) $D(I + GR)^{-1}G$ is weakly column-sum dominant,

then for all diagonal matrices D_1 and D_2 with positive diagonal elements, all eigenvalues of (9) lie in the strict right-half plane. This condition on T, G, and R is often satisfied.†

The subclass of numerical integration formulas (6) defined by the condition that $b_{-1} > 0$ are of considerable use[11,12,13] in applications involving the typically "stiff systems" of differential equations encountered in the analysis of nonlinear transistor networks. With $b_{-1} > 0$, y_{n+1} is defined *implicitly* through

$$y_{n+1} + hb_{-1}f(y_{n+1}, (n + 1)h) = \sum_{k=0}^{r} a_k y_{n-k} + h \sum_{k=0}^{r} b_k \tilde{y}_{n-k}$$

in which the right side depends on y_{n-k} only for $k \in \{0, 1, 2, \cdots, r\}$, and for $f(u, t)$ given by equation (5), we have

$$y_{n+1} + hb_{-1}\{TF[C^{-1}(y_{n+1})] + (I + GR)^{-1}GC^{-1}(y_{n+1})\} = q_n \qquad (10)$$

in which

$$q_n = \sum_{k=0}^{r} a_k y_{n-k} + h \sum_{k=0}^{r} b_k \tilde{y}_{n-k} + hb_{-1}B[(n + 1)h].$$

Obviously, the numerical integration formula (10) makes sense only if there exists for each n a $y_{n+1} \in E^{(2p+q)}$ such that equation (10) is satisfied.

Let $x_{n+1} = C^{-1}(y_{n+1})$ for each n. Then equation (10) possesses a unique solution y_{n+1} if and only if there exists a unique $x_{n+1} \in E^{(2p+q)}$ such that

$$C(x_{n+1}) + hb_{-1}[TF(x_{n+1}) + (I + GR)^{-1}Gx_{n+1}] = q_n . \qquad (11)$$

Since $C(x_{n+1}) = cx_{n+1} + \tau F(x_{n+1})$, in which

$$c = \text{diag}\,(c_1, c_2, \cdots, c_{2p+q})$$

and

$$\tau = \text{diag}\,(\tau_1, \tau_2, \cdots, \tau_{2p+q}),$$

---

* The terms "strongly-column-sum dominant" and "weakly-column-sum dominant" are reasonably standard. However they are defined in Section II.

† See Ref. 8 for examples.

equation (11) is equivalent to

$$[\tau + hb_{-1}T]F(x_{n+1}) + [c + hb_{-1}(I + GR)^{-1}G]x_{n+1} = q_n . \qquad (12)$$

The matrices τ and c are both diagonal with positive diagonal elements. Thus it is clear that for all positive h

$$\det [\tau + hb_{-1}T] \neq 0$$

and

$$\det [c + hb_{-1}(I + GR)^{-1}G] \neq 0.^*$$

For all sufficiently small positive h

$$[\tau + hb_{-1}T]^{-1}[c + hb_{-1}(I + GR)^{-1}G] \in P_0 .^\dagger$$

Consequently[††] for all sufficiently small $h > 0$, equation (12) possesses a unique solution for each q_n .[‡] However, our interest in equation (12) is primarily in connection with "large-h" algorithms.

Suppose that det $G \neq 0$ and that $T^{-1}G \in P_0$ for all possible combinations of α_r and α_f ($0 < \alpha_r < 1$, $0 < \alpha_f < 1$) for each transistor (see Ref. 1 for examples). Then, according to Theorem 6 of Section II, for any particular T and R

$$[\tau + hb_{-1}T]^{-1}[c + hb_{-1}(I + GR)^{-1}G] \in P_0$$

for all $h > 0$, and hence equation (10) possesses a unique solution y_{n+1} for all positive values of h.

An important and general proposition concerning (10) is as follows. Suppose that

$$T^{-1}[(I + GR)^{-1}G] \in P_0 \qquad (13)$$

and that condition (13) is satisfied whenever $\alpha_r^{(k)}$ and $\alpha_f^{(k)}$ are replaced with positive constants $\delta_r^{(k)}$ and $\delta_f^{(k)}$, respectively, such that $\delta_r^{(k)} \leqq \alpha_r^{(k)}$ and $\delta_f^{(k)} \leqq \alpha_f^{(k)}$ for $k = 1, 2, \cdots, p$. In other words, assuming that $F(\cdot)$ is as defined in this section and that $F(\cdot) \in \mathfrak{F}_0^{(2p+q)}$ (see Definition 1 of Section 2.1), suppose that the dc equation

$$F(x) + T^{-1}[(I + GR)^{-1}G]x = B$$

possesses at most one solution x for each $B \in E^{(2p+q)}$ for "the original set of α's as well as for an arbitrary set of *not-larger α's*." Then an

---

[*] Here we have used the fact that $(I + GR)^{-1}G$ is positive semidefinite.

[†] See Section 1.2.

[††] See Section 1.3.

[‡] Alternatively, this conclusion could have been obtained by applying the contraction-mapping fixed-point principle to (10), in view of the fact that each of the elements of J_u is bounded on $u \in E^{(p+q)}$ and $t \in [0, \infty)$.

immediate application of Theorem 5 of Section II shows that

$$[\tau + hb_{-1}T]^{-1}[c + hb_{-1}(I + GR)^{-1}G] \in P_0$$

for all $h > 0$, and hence that equation (10) possesses a unique solution y_{n+1} for all $h > 0$ and all $q_n \in E^{(2p+q)}$.

II. THEOREMS, PROOFS, AND SOME DISCUSSION

Throughout this section,

(i) n is an arbitrary positive integer,

(ii) P_0 denotes the set of all real $n \times n$ matrices M such that all principal minors of M are nonnegative,

(iii) real Euclidean n-space is denoted by E^n, and θ is the zero element of E^n,

(iv) v^{tr} denotes the transpose of the row vector $v = (v_1, v_2, \cdots, v_n)$,

(v) $\| v \|$ denotes $(\sum_{i=1}^n v_i^2)^{1/2}$ for all $v \in E^n$,

(vi) if D is a real diagonal matrix, then $D > 0$ $(D \geqq 0)$ means that the diagonal elements of D are positive (nonnegative),

(vii) I_q denotes the identity matrix of order q, and I denotes the identity matrix of order determined by the context in which the symbol is used, and

(viii) we shall say that a real $n \times n$ matrix M is strongly (weakly) column-sum dominant if and only if for $j = 1, 2, \cdots, n$

$$m_{jj} > (\geqq) \sum_{i \neq j} | m_{ij} |.$$

2.1 *Definition 1*

For each positive integer n, let $\mathfrak{F}_0^n$ denote that collection of mappings of E^n into itself defined by: $F \in \mathfrak{F}_0^n$ if and only if there exist for $j = 1, 2, \cdots, n$, continuous functions $f_j(\cdot)$ mapping E^1 into E^1 such that for each $x = (x_1, x_2, \cdots, x_n)^{tr} \in E^n$, $F(x) = [f_1(x_1), f_2(x_2), \cdots, f_n(x_n)]^{tr}$, and

$$(i) \qquad \inf_{\alpha \in (-\infty, \infty)} [f_j(\alpha + \beta) - f_j(\alpha)] = 0$$

$$(ii) \qquad \sup_{\alpha \in (-\infty, \infty)} [f_j(\alpha + \beta) - f_j(\alpha)] = +\infty$$

for all $\beta > 0$ and all $j = 1, 2, \cdots, n$.

2.2 *Theorem 1*

Let $F \in \mathfrak{F}_0^n$, let A be a real $n \times n$ matrix such that $A \notin P_0$, and let δ be a positive constant. Then there exist $B \in E^n$, $x \in E^n$, and $y \in E^n$

such that

(*i*) $$F(x) + Ax = B,$$

(*ii*) $$F(y) + Ay = B,$$

and

(*iii*) $$\| x - y \| = \delta.$$

2.3 *Proof of Theorem 1*

Since $A \notin P_0$, there exists[2] a real diagonal matrix $D > 0$ such that $\det (D + A) = 0$. Thus there exists a $x^* \in E^n$ such that $\| x^* \| = \delta$ and $(D + A)x^* = \theta$.

Since $F \in \mathcal{F}_0^n$, there exists a $x \in E^n$ such that

$$f_j(x_j) - f_j(x_j - x_j^*) = x_j^* \, d_j$$

for all $j = 1, 2, \cdots, n$ in which d_j is the j^{th} diagonal element of D. Let

$$B = F(x) + Ax,$$

and let $y = x - x^*$. Then $A(x - y) = Ax^* = -Dx^*$, and

$$F(x) - F(y) + A(x - y) = \theta. \ \square$$

2.4 *Remarks Concerning Theorem 1*

If, as in the case of standard transistor models,

$$f_j(x_j) = e^{\lambda_j x_j} - 1$$

or

$$f_j(x_j) = 1 - e^{-\lambda_j x_j}$$

with $\lambda_j > 0$, we have, respectively,

$$f_j(\alpha + \beta) - f_j(\alpha) = e^{\lambda_j \alpha}(e^{\lambda_j \beta} - 1)$$

or

$$f_j(\alpha + \beta) - f_j(\alpha) = e^{-\lambda_j \alpha} \, (1 - e^{-\lambda_j \beta})$$

and it is clear that for either type of function conditions (*i*) and (*ii*) of Definition 1 are satisfied.

In Ref. 1 it is proved that if $F(\cdot) \in \mathcal{M}$ the set of all $F(\cdot)$ of the form (2) with each $f_i(\cdot)$ a strictly monotone increasing mapping of E^1 into E^1, and if $A \in P_0$, then equation (1) possesses at most one solution.

Thus, using Theorem 1, we see that for each $F(\cdot) \in \mathfrak{M} \cap \mathfrak{F}_0^n$ there exists at most one solution of $F(x) + Ax = B$ for each $B \in E^n$ if and only if $A \in P_0$. Similarly, with $\mathfrak{M}_0$ the set of all $F(\cdot)$ of the form (2) with each $f_i(\cdot)$ a strictly monotone increasing mapping of E^1 onto (α_i, β_i) with each α_i and β_i such that $-\infty \leqq \alpha_i < \beta_i \leqq \infty$, if $F(\cdot) \in \mathfrak{M}_0 \cap \mathfrak{F}_0^n$ and $\det A \neq 0$, then there exists a unique solution x of $F(x) + Ax = B$ for each $B \in E^n$ if and only if $A \in P_0$. (The "if" part of this statement is proved in Ref. 2.) A parallel development can be carried out for equations of the form $AF(x) + x = B$ with A a real $n \times n$ matrix, $F(\cdot) \in \mathfrak{M}_0 \cap \mathfrak{F}_0^n$, and $B \in E^n$. More explicitly, we can prove that if $F(\cdot) \in \mathfrak{M}_0 \cap \mathfrak{F}_0^n$, then there exists a unique solution x of $AF(x) + x = B$ for each $B \in E^n$ if and only if $A \in P_0$.

There may be a temptation to conjecture that whenever $F(\cdot) \in \mathfrak{M} \cap \mathfrak{F}_0^n$ and $A \notin P_0$ then the equation $F(x) + Ax = B$ does not possess a solution for some $B \in E^n$. The conjecture is false. In fact, with $n = 2$, $f_1(x_1) = e^{x_1}$, $f_2(x_2) = e^{x_2}$, and

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

we have a situation in which (it is easy to show that) there exists a solution for all $B \in E^2$. Of course here for some choices of B the solution is not unique.

2.5 *Theorem 2*

Let P and Q denote real n × n matrices such that

$$p_{jj} > \sum_{i \neq j} |p_{ij}| \quad \text{and} \quad q_{jj} > \sum_{i \neq j} |q_{ij}|$$

for all $j = 1, 2, \cdots, n$. For $j = 1, 2, \cdots, n$ let $f_j(\cdot)$ denote a continuous monotone nondecreasing (but not necessarily differentiable) mapping of E^1 into itself, and let $F(x) = [f_1(x_1), f_2(x_2), \cdots, f_n(x_n)]^{tr}$ for all $x \in E^n$. Then for each $R \in E^n$, there exists a unique $x \in E^n$ such that

$$PF(x) + Qx = R,$$

and, for any $y_0 \in E^n$, x is the limit of the sequence $x^{(0)}, x^{(1)}, \cdots$ defined by

$$y^{(n)} = D_P F(x^{(n)}) + D_Q x^{(n)}$$

$$y^{(n+1)} + (P - D_P)F(x^{(n)}) + (Q - D_Q)x^{(n)} = R$$

for $n \geqq 0$, in which D_P and D_Q are diagonal matrices whose diagonal elements coincide with those of P and Q, respectively.

2.6 *Proof of Theorem 2*

Since the continuous mapping $[D_PF(\cdot) + D_Q]$ of E^n into E^n possesses an inverse $[D_PF(\cdot) + D_Q]^{-1}$, the equation

$$PF(x) + Qx = R$$

possesses a unique solution x if and only if $y = D_PF(x) + D_Qx$ is the unique solution of

$$y + \tilde{P}F[(D_PF(\cdot) + D_Q)^{-1}y] + \tilde{Q}[(D_PF(\cdot) + D_Q)^{-1}y] = R$$

in which $\tilde{P} = (P - D_P)$ and $\tilde{Q} = (Q - D_Q)$.

Therefore, by Banach's contraction-mapping fixed-point theorem, it suffices to show that with the metric $\rho(y, z) = \sum_{i=1}^n | y_i - z_i |$, the operator H defined by

$$H(y) = \tilde{P}F[(D_PF(\cdot) + D_Q)^{-1}y] + \tilde{Q}[(D_PF(\cdot) + D_Q)^{-1}y]$$

for all $y \in E^n$, is a contraction mapping of E^n into itself. We show this as follows. Let $y \in E^n$ and $z \in E^n$. Using the fact that

$$\alpha = d_{Qj}[(d_{Pj}f_j(\cdot) + d_{Qj})^{-1}\alpha] + d_{Pj}f_j[(d_{Pj}f_j(\cdot) + d_{Qj})^{-1}\alpha]$$

for all real α and all $j = 1, 2, \cdots, n$, in which d_{Pj} and d_{Qj} is the j^{th} diagonal element of D_P and D_Q, respectively, it is a simple matter to verify that for all j:

$$f_j[(d_{Pj}f_j(\cdot) + d_{Qj})^{-1}y_j] - f_j[(d_{Pj}f_j(\cdot) + d_{Qj})^{-1}z_j]$$
$$= \frac{r_j}{d_{Qj} + d_{Pj}r_j}(y_j - z_j),$$

and

$$(d_{Pj}f_j(\cdot) + d_{Qj})^{-1}y_j - (d_{Pj}f_j(\cdot) + d_{Qj})^{-1}z_j = \frac{1}{d_{Qj} + d_{Pj}r_j}(y_j - z_j)$$

in which $r_j = 1$ if $y_j = z_j$, and, if $y_j \neq z_j$,

$$r_j = \frac{f_j[(d_{Pj}f_j(\cdot) + d_{Qj})^{-1}y_j] - f_j[(d_{Pj}f_j(\cdot) + d_{Qj})^{-1}z_j]}{(d_{Pj}f_j(\cdot) + d_{Qj})^{-1}y_j - (d_{Pj}f_j(\cdot) + d_{Qj})^{-1}z_j}.$$

Thus

$$H(y) - H(z)$$
$$= \tilde{P} \operatorname{diag} \left\{\frac{r_j}{d_{Qj} + d_{Pj}r_j}\right\}(y - z) + \tilde{Q} \operatorname{diag} \left\{\frac{1}{d_{Qj} + d_{Pj}r_j}\right\}(y - z)$$

in which $r_j \geqq 0$. Therefore

$$\rho(H(y), H(z)) \leqq \max_j \left(\frac{\sigma_{Qj} + \sigma_{Pj} r_j}{d_{Qj} + d_{Pj} r_j}\right) \rho(y, z)$$

in which $\sigma_{Qj} = \sum_{i \neq j} | q_{ij} |$ and $\sigma_{Pj} = \sum_{i \neq j} | p_{ij} |$. Since $\sigma_{Qj} < d_{Qj}$ and $\sigma_{Pj} < d_{Pj}$ for all j, there exists a positive constant $\beta < 1$ such that

$$\max_j \left(\frac{\sigma_{Qj} + \sigma_{Pj} r_j}{d_{Qj} + d_{Pj} r_j}\right) \leqq \beta$$

for all $r_j \geqq 0$. $\square$

2.7 *Theorem 3*

If $A \in P_0$ *and* $\det A \neq 0$, *if for each* $j = 1, 2, \cdots, n$: $f_j(\cdot)$ *is a continuous mapping of* E^1 *into itself such that*

$$f_j(x_j) = 0 \quad \text{for all } x_j$$

or

$$f_j(x_j) > 0 \quad \text{for all } x_j > c$$

and

$$f_j(x_j) < 0 \quad \text{for all } x_j < -c$$

for some $c \geqq 0$, *then, with* $F(x) = [f_1(x_1), f_2(x_2), \cdots, f_n(x_n)]^{\text{tr}}$ *for all* $x \in E^n$,

$$\| F(x) + Ax \| \rightarrow \infty \quad \text{as} \quad \| x \| \rightarrow \infty.$$

2.8 *Proof of Theorem 3*

We note that

$$\| F(x) + Ax \| \rightarrow \infty \quad \text{as} \quad \| x \| \rightarrow \infty$$

if and only if

$$\| A^{-1}F(x) + x \| \rightarrow \infty \quad \text{as} \quad \| x \| \rightarrow \infty.$$

With $M = A^{-1}$, let

$$MF(x) + x = q. \tag{14}$$

Since $A \in P_0$, we have $M \in P_0$.[1] Since $M \in P_0$, we have[1] for any $y \in E^n$ and $y \neq \theta$

$$y_k(My)_k \geqq 0$$

for some index k such that $y_k \neq 0$.

Suppose that $F(x) \neq \theta$. Then there exists an index k_1 such that

$$f_{k_1}(x_{k_1})[MF(x)]_{k_1} \geqq 0$$

with $f_{k_1}(x_{k_1}) \neq 0$. Thus, using (14),

$$f_{k_1}(x_{k_1})[MF(x)]_{k_1} + f_{k_1}(x_{k_1})x_{k_1} = f_{k_1}(x_{k_1})q_{k_1}$$

and

$$f_{k_1}(x_{k_1})x_{k_1} \leqq f_{k_1}(x_{k_1})q_{k_1} .$$

Either $x_{k_1} \in [-c, c]$ or not. If not, then $f_{k_1}(x_{k_1})x_{k_1} > 0$ and $| x_{k_1} | \leqq | q_{k_1} |$. Therefore for some index k_1, $| x_{k_1} | \leqq \delta_1 \triangleq \max (c, | q_{k_1} |)$, whether or not $F(x) = \theta$.

Let $M^{(k_1)}$ denote the matrix obtained from M by deleting the k_1 row and column, and let $M_{(k_1)}$ denote the k_1 column of M with the k_1 entry removed. Similarly, let $x_{(k_1)}$, $q_{(k_1)}$ and $F_{(k_1)}(x_{(k_1)})$ denote the $(n - 1)$-vectors obtained from x, q, and $F(x)$, respectively, by removing the k_1 entry. Then

$$M^{(k_1)}F_{(k_1)}(x_{(k_1)}) + x_{(k_1)} = q_{(k_1)} - M_{(k_1)}f_{k_1}(x_{k_1}).$$

Since $M^{(k_1)} \in P_0$, we can repeat the argument given above. Thus there exists an index k_2, different from k_1, such that

$$| x_{k_2} | \leqq \delta_2 \triangleq \max (c, | q_{(k_1,k_2)} |)$$

in which

$$| q_{(k_1,k_2)} | = \max_{|x_{k_1}| \leqq \delta_1} | [q_{(k_1)} - M^{(k_1)}f_{k_1}(x_{k_1})]_{l_2} |$$

and l_2 is the index of the component of $x_{(k_1)}$ that corresponds to the k_2 component of x. By continuing in this manner we can determine positive constants $\delta_1, \delta_2, \cdots, \delta_n$ depending only on q, F, M, and c such that, with $\delta = \max_j \{\delta_j\}$,

$$| x_j | \leqq \delta \quad \text{for all} \quad j = 1, 2, \cdots, n$$

and each δ_i depends on q such that for any positive constant α, there exists a constant $\beta_i(\alpha)$ with the property that $\delta_i \leqq \beta_i(\alpha)$ provided that $|| q || \leqq \alpha$. Therefore for any $\alpha > 0$ there is a $\beta(\alpha)$ such that $|| x || \leqq \beta(\alpha)$ whenever $|| q || \leqq \alpha$, which implies that $|| q || \to \infty$ as $|| x || \to \infty$. □

2.9 *Theorem 4*

Let P and Q denote real $n \times n$ matrices with P strongly column-sum dominant. Suppose that there exists a real diagonal matrix $D > 0$ such *that DP is strongly column-sum dominant and DQ is weakly column-sum dominant. Then for all real diagonal matrices $D_1 > 0$ and $D_2 > 0$, all eigenvalues of $(PD_1 + QD_2)$ lie in the strict (that is, open) right-half plane.*

2.10 *Proof of Theorem 4*

Since the strict right-half plane contains all of the eigenvalues of P, there exists choices of $D_1 > 0$ and $D_2 > 0$ such that every eigenvalue of $(PD_1 + QD_2)$ lies in the strict right-half plane. Thus it suffices to show that $(PD_1 + QD_2)$ does not possess an eigenvalue on the boundary of the complex plane for all $D_1 > 0$ and all $D_2 > 0$. In other words, it suffices to show that (with $i = (-1)^{\frac{1}{2}}$)

$$PD_1 + QD_2 + i\omega I \tag{15}$$

is nonsingular for all $D_1 > 0$, all $D_2 > 0$, and all real constants ω.

Suppose that (15) is singular for some ω and some $D_1 > 0$ and some $D_2 > 0$. Then $(DPD_1 + DQD_2 + i\omega D)$ is singular. But DPD_1 is strongly column-sum dominant and DQD_2 is weakly column-sum dominant. Thus $M = (DPD_1 + DQD_2)$ is strongly column-sum dominant, and, since

$$| m_{jj} + i\omega d_j | > \sum_{i \neq j} | m_{ij} |$$

for all j, in which d_j is the j^{th} diagonal element of D, it follows that $\det (M + i\omega D) \neq 0$, which is a contradiction. □

2.11 *Definition 2*

With q and p nonnegative integers such that $(p + q) > 0$, let $\mathfrak{J}$ denote the set of all matrices M such that $M = I_q \oplus M_1 \oplus M_2 \oplus \cdots \oplus M_p$ with

$$M_k = \begin{bmatrix} 1 & -\propto_r^{(k)} \\ -\propto_f^{(k)} & 1 \end{bmatrix}$$

and

$$0 < \propto_r^{(k)} < 1$$

$$0 < \propto_f^{(k)} < 1$$

for all $k = 1, 2, \cdots, p$.*

---

* As suggested, if $q = 0$, then $M = M_1 \oplus M_2 \oplus \cdots \oplus M_p$, while if $p = 0$, then $M = I_q$.

2.12 *Definition 8*

With q and p nonnegative integers such that $(p + q) > 0$, let $\mathfrak{I}(\alpha)$ denote the set of all matrices M such that $M = I_q \oplus M_1 \oplus M_2 \oplus \cdots \oplus M_p$ with

$$M_k = \begin{bmatrix} 1 & -\delta_r^{(k)} \\ -\delta_f^{(k)} & 1 \end{bmatrix}$$

and

$$0 < \delta_r^{(k)} \leqq \alpha_r^{(k)}$$
$$0 < \delta_f^{(k)} \leqq \alpha_f^{(k)}$$

for all $k = 1, 2, \cdots, p$.*

2.13 *Theorem 5*

Let $T \in \mathfrak{I}$, let H be a real matrix of order $(2p + q)$, and suppose that $M^{-1}H \in P_0$ for all $M \in \mathfrak{I}(\alpha)$. Then

$$(T + D_1)^{-1}(H + D_2) \in P_0$$

for all diagonal matrices $D_1 \geqq 0$ and $D_2 \geqq 0$.

2.14 *Proof of Theorem 5*

Suppose that for some $D_1 \geqq 0$ and $D_2 \geqq 0$

$$(T + D_1)^{-1}(H + D_2) \notin P_0 .$$

Then there exists[2] a diagonal matrix $D > 0$ such that

$$(T + D_1)^{-1}(H + D_2) + D$$

is singular. It follows that

$$H + \Delta + TD$$

is singular, in which $\Delta = D_2 + D_1 D$. Since

$$\Delta + TD = M(\Delta + D)$$

in which $M \in \mathfrak{I}(\alpha)$, it follows that

$$H + M(\Delta + D)$$

is singular, and therefore that

---

* As suggested, if $q = 0$, then $M = M_1 \oplus M_2 \oplus \cdots \oplus M_p$, while if $p = 0$, then $M = I_q$.

$$M^{-1}H + (\Delta + D)$$

is singular, which is a contradiction since[1] $M^{-1}H \in P_0$ and $(\Delta + D)$ is a diagonal matrix with positive diagonal elements.[4] □

2.15 *Theorem 6*

Let $M^{-1}G \in P_0$ *for all* $M \in \mathfrak{J}$, *and let* $\det G \neq 0$. *Let* R *be as defined in Section 1.4. Then for any* $T \in \mathfrak{J}$

$$(T + D_1)^{-1}[(I + GR)^{-1}G + D_2] \in P_0$$

for all diagonal matrices $D_1 \geqq 0$ *and* $D_2 \geqq 0$.

2.16 *Proof of Theorem 6*

Since $\det G \neq 0$ and $M^{-1}G \in P_0$ for all $M \in \mathfrak{J}$, it follows (see the proof of Theorem 7 of Ref. 2) that

$$M^{-1}(I + GR)^{-1}G \in P_0$$

for all $M \in \mathfrak{J}$.

Suppose that for some $T \in \mathfrak{J}$ and some $D_1 \geqq 0$ and $D_2 \geqq 0$

$$(T + D_1)^{-1}[(I + GR)^{-1}G + D_2] \notin P_0 .$$

Then, following the proof of Theorem 5, we would have

$$\det \{M^{-1}(I + GR)^{-1}G + (\Delta + D)\} = 0$$

for some $M \in \mathfrak{J}$ and some diagonal matrix $(\Delta + D)$ with positive diagonal elements, which is a contradiction. □

III. ACKNOWLEDGMENT

The writer is indebted to A. N. Willson, Jr. for carefully reading the draft.

REFERENCES

1. Sandberg, I. W., and Willson, A. N., Jr., "Some Theorems on Properties of DC Equations of Nonlinear Networks," B.S.T.J., *48*, No. 1 (January 1969), pp. 1–34.
2. Sandberg, I. W., and Willson, A. N., Jr., "Some Network-Theoretic Properties of Non-Linear DC Transistor Networks," B.S.T.J., *48*, No. 5 (May–June 1969), pp. 1293–1312.
3. Varga, R. S., *Matrix Iterative Analysis*, Englewood Cliffs, New Jersey: Prentice-Hall, 1962, p. 23.
4. Stern, T. E., *Theory of Nonlinear Networks and Systems*, Reading, Mass.: Addison-Wesley, 1965, pp. 42–43.

5. Palais, R. S., "Natural Operations on Differential Forms," Trans. Amer. Math. Soc., *92,* No. 1 (1959), pp. 125–141.
6. Holzmann, C. A., and Liu, R., "On the Dynamical Equations of Nonlinear Networks with n-Coupled Elements," Proc. Third Ann. Allerton Conf. on Circuit and System Theory, U. of Illinois, 1965, pp. 536–545.
7. Goldstein, A. A., *Constructive Real Analysis,* New York: Harper & Row, 1967, pp. 41–45.
8. Sandberg, I. W., "Some Theorems on the Dynamic Response of Nonlinear Transistor Networks," B.S.T.J., *48,* No. 1 (January 1969), pp. 35–54.
9. Hamming, R. W., *Numerical Methods for Scientists and Engineers,* New York: McGraw-Hill, 1962.
10. Ralston, A. A., *A First Course in Numerical Analysis,* New York: McGraw-Hill, 1965.
11. Hachtel, G. D., and Rohrer, R. A., "Techniques for the Optimal Design and Synthesis of Switching Circuits," Proc. of the IEEE, *55,* No. 11 (November 1967), pp. 1864–1876.
12. Sandberg, I. W., and Shichman, H., "Numerical Integration of Systems of Stiff Nonlinear Differential Equations," B.S.T.J., *47,* No. 4 (April 1968), pp. 511–527.
13. Calahan, D. A., "Efficient Numerical Analysis of Non-Linear Circuits," Proc. Sixth Ann. Allerton Conf. on Circuit and System Theory, U. of Illinois, 1968, pp. 321–331.

A Charge Control Relation for Bipolar Transistors

By H. K. GUMMEL

(Manuscript received August 1, 1969)

We give a relation which links emitter and collector junction voltages, V_{eb} and V_{cb}, collector current I_c, and the total charge Q_b of carriers that enter through the base terminal (electrons in a pnp *transistor):*

$$I_c = \text{const}\,\frac{e^{(qV_{eb}/kT)} - e^{(qV_{cb}/kT)}}{Q_b}.$$

This relation is valid for high injection conditions, subject only to minor restrictions. The significance for device modeling is discussed.

I. INTRODUCTION

A basic concept of charge control theory is that the controlled current (collector current) equals controlling charge (base charge) divided by a transit time.[1] This paper presents an additional relation which links base charge and collector current with junction voltages. The validity of this relation is subject only to minor restrictions. When these are met the relation holds even under high-level injection conditions.

Section II presents a derivation of the new charge control relation. Equation (15) states the principal result. The discussion in Section III points out the significance of this relation for bipolar transistor models.

II. DERIVATION

Consider a one-dimensional transistor of pnp polarity. The hole current density is given by

$$j_p = q\mu Ep - qDp' \tag{1}$$

where the symbols have their customary meaning. We shall assume that diffusivity D and mobility μ are related through the Einstein relation:

Approximation (a):

$$D = \frac{kT}{q}\mu.$$

Approximation (b_1): It is assumed that electric fields are low enough for avalanche multiplication of carriers to be negligible.

Approximation (b_2): The velocity-field relation is idealized by the field dependent mobility expression:

$$\mu = \frac{\mu_o}{1 + \frac{\mu_o |E|}{v_s}}$$

where $\mu_0 \equiv qD_0/kT$ is the low-field mobility, considered for convenience independent of doping, and where v_s is the scattering limited velocity. Approximation (b_1) places an upper limit on allowable bias. It is known (see for example Ref. 2) that D is underestimated at high fields by approximations (a) and (b_2) and that approximation (b_2) yields too gradual a transition from low field velocities to the high field saturated velocity.[3,4] Nevertheless, approximations (a) and (b) afford significant simplifications in the treatment to follow and are retained for that reason. To the extent that our final result, equation (15), is affected by them, it must be considered approximate. The errors depend on bias and doping profile; they are not expected to exceed a few percent for typical situations. The error due to approximation (b_2) overemphasizes velocity saturation effects and may be alleviated by choice of values of v_s larger than the final saturation value in high field regions. In high-field regions the current is carried predominately as drift current, with a carrier concentration that is nearly constant in such regions, so that errors in D are of minor consequence.

Next we define a quantity $a(x)$ which is the ratio of the hole current density at position x to the current density j_c leaving the collector terminal

$$a(x) = \frac{j_p(x)}{j_c}. \tag{2}$$

For direct current conditions, considered here, a approaches unity at the collector and is $1/\alpha = (1 + \beta)/\beta$ at the emitter. For large common emitter current gain β, a differs negligibly from unity.

We use

$$E = -\frac{kT}{q}\psi' \tag{3}$$

where ψ is the electrostatic potential in units of the Boltzman voltage kT/q, and consider equation (1) as a differential equation for $p(x)$. Its solution, when p is specified at a point x_1, is

$$p(x) = p(x_1)e^{\psi(x_1)-\psi(x)} - \frac{j_c}{qD_o}\int_{x_1}^{x} a(t)e^{\psi(t)-\psi(x)}\,dt - \frac{j_c}{qv_s}\int_{x_1}^{x} a(t)\mid \psi(t)' \mid e^{\psi(t)-\psi(x)}\,dt. \tag{4}$$

Equation (4) is valid for any pair of points x_1 and x. Denote by x_E and x_C the outside edges of emitter and collector transition regions, and use equation (4) with $x_1 = x_E$ and $x = x_C$. Multiplication of equation (4) by $e^{\psi(x)}$ and use* of

$$p(x) = n_i e^{\varphi_p(x)-\psi(x)} \tag{5}$$

$$n(x) = n_i e^{\psi(x)-\varphi_n(x)} \tag{6}$$

where φ_p and φ_n are hole and electron quasi-fermi levels in units of the Boltzman voltage, yield

$$j_c = \frac{qD_o n_i^2(e^{\varphi_p(x_E)} - e^{\varphi_p(x_C)})}{\int_{x_E}^{x_C} a(t)n_i e^{\psi(t)}\,dt + \frac{n_i D_o}{v_s}\int_{x_E}^{x_C} a(t)\mid \psi'(t) \mid e^{\psi(t)}\,dt}. \tag{7}$$

We shall now show that the second term in the denominator is negligible. The integrals in the denominator obtain the largest contribution from the region near x_m where $\psi(x)$ attains its maximum value ψ_m. If in the second integral we replace $a(t)$ by its value a_m at x_m, and if we neglect $e^{\psi(x_E)}$ and $e^{\psi(x_C)}$ in comparison with e^{ψ_m}—all very reasonable assumptions—we obtain for the second integral in the denominator

Approximation (c):

$$\frac{D_o n_i}{v_s}\int_{x_E}^{x_C} a(t)\mid \psi'(t) \mid e^{\psi(t)}\,dt \approx \frac{2a_m D_o n_i}{v_s}e^{\psi_m}.$$

For an assessment of the relative magnitude of the terms in the denominator of equation (7), consider that in a region of width w the potential $\psi(x)$ does not differ markedly from ψ_m; such region is conventionally called the "base" of the transistor. Consider high current gain, that is, $a \approx a_m \approx 1$. Then the value of the first integral is $wn_i e^{\psi_m}$, compared with $(2D_0/v_s)n_i e^{\psi_m}$ for the second. The quantity $2D_0/v_s$ has units of length and is ≈ 200 Å for silicon. This length is small compared to base widths of today's most advanced transistors and hence we will neglect the second term in the rest of this paper. Conceivably, future transistors may have narrow enough bases that the term will have to be kept.

---

* Equation (6) is defined for later reference.

Approximation (d):

$$\frac{n_i D_o}{v_s} \int_{x_E}^{x_C} a(t) \mid \psi'(t) \mid e^{\psi(t)} \, dt \ll \int_{x_E}^{x_C} a(t) n_i e^{\psi(t)} \, dt.$$

If in equation (7) we were to let $v_s \rightarrow \infty$, that is, considered the carrier velocity to be strictly proportional to the electric field, then approximation (d) would be implemented automatically. Note, however, that in making approximation (d) we do not imply $v_s = \infty$, nor do we neglect essential consequences of the finiteness of v_s. A low value of v_s manifests itself in substantial base widening at high currents, that is, in influencing $\psi(t)$ in the remaining (first) term in the denominator of equation (7). In view of the idealized textbook treatments in which the minority carrier concentration at the base side of the collector depletion region is set equal to zero, rather than to a finite value, the following statement of equation (7) may be of interest: For low injection (that is, for currents sufficiently low that the base width is independent of current) the effect of the finiteness of the scattering limited velocity on the dc collector current is equivalent to a base widening of $2D_0/v_s$.

We now make the approximation:

Approximation (e):

The value of the electron quasi-fermi
level in the base is constant.

A gradient in the electron quasi-fermi level in the region where electrons are majority carriers would cause appreciable electron current to flow; for transistors of reasonable current gain, such currents are negligible. Thus, approximation (e) is very reasonable. We denote this value of the electron quasi-fermi level by φ_{nb} and divide numerator and denominator of the right side of equation (7) by exp (φ_{nb}). We define the emitter-base and collector-base junction voltages by

$$V_{eb} = \frac{kT}{q} [\varphi_p(x_E) - \varphi_{nb}] \tag{8}$$

$$V_{cb} = \frac{kT}{q} [\varphi_p(x_C) - \varphi_{nb}]. \tag{9}$$

These voltages differ from terminal voltages by ohmic drops, primarily lateral ohmic drops in the base region. The first integral in the denominator of equation (7), after it is divided by exp (φ_{nb}), contains very nearly the total area density of electrons.

Approximation (f):

$$\int_{x_E}^{x_C} n_i a(t)(e^{\psi(t)-\varphi_{nb}})\,dt = \int_{x_E}^{x_C} a(t)n(t)\,dt.$$

The integrands outside the base region differ, since there the quasi-fermi level is position dependent and does not equal φ_{nb}, but the contribution to the integral outside of the base region is negligible. By defining an average value $\langle a\rangle_{\mathrm{av}}$ of a,

$$\langle a\rangle_{\mathrm{av}} = \frac{\int_{x_E}^{x_C} a(t)n(t)\,dt}{\int_{x_E}^{x_C} n(t)\,dt}. \tag{10}$$

We may write expression (f) as

$$n_i \int_{x_E}^{x_C} a(t)e^{\psi(t)-\varphi_{nb}}\,dt = -\frac{\langle a\rangle_{\mathrm{av}}}{q}\,q_b \tag{11}$$

where q_b is the total charge, per unit area, of those mobile carriers associated with the base terminal, that is, electrons in a pnp transistor. Equation (7) with approximations (d) and (f) may be written

$$j_c = -\frac{(q^2 D_o n_i^2/\langle a\rangle_{\mathrm{av}})[e^{(qV_{eb}/kT)} - e^{(qV_{cb}/kT)}]}{q_b}. \tag{12}$$

We now change from current and charge densities to current and charge. We chose the sign of the collector current according to the convention that an electric current entering the device is positive:

$$I_c = -j_c A \tag{13}$$

$$Q_b = q_b A \tag{14}$$

where A is the device area. Note that the sign of Q_b is such that an electric current flowing into the base tends to increase Q_b. This is the proper sign for charge control theory. Equation (12) can now be written

$$I_c = C\,\frac{e^{(qV_{eb}/kT)} - e^{(qV_{cb}/kT)}}{Q_b} \tag{15}$$

with

$$C = \frac{(qn_iA)^2 D_o}{\langle a \rangle_{\text{av}}}. \tag{16}$$

Equation (15) is the principal result of this paper. Note that Q_b depends on bias, and that the form of the bias dependence is governed by the doping profile. However, the relation among the quantities I_c, V_{eb}, V_{cb}, and Q_b in equation (15) is independent of the details of the doping profile, provided that assumptions (a) through (f) are valid.

III. DISCUSSION

In spite of the simple appearance of equation (15)—indeed because of it—it provides a powerful tool for transistor modeling. It may be written in the form

$$I_c = -a_{21}[e^{(qV_{eb}/kT)} - 1] + a_{22}[e^{(qV_{cb}/kT)} - 1] \tag{17}$$

with

$$a_{22} = -a_{21} = \frac{C}{Q_b} \tag{18}$$

which is the form of one of the Ebers–Moll equations.[5] But whereas in the Ebers–Moll equation the coefficients a_{21} and a_{11} are constant, they depend according to equation (18) on bias through the base charge Q_b (C depends on bias only through $\langle a \rangle_{\text{av}}$ which for $\beta \gg 1$ is nearly unity and varies little with bias). It is this bias dependence of Q_b which contains high-injection effects. Thus, use of equation (15) holds promise for transistor modeling of improved accuracy. The major bias dependence of the collector current is through the exponentials in the numerator of equation (15). These are "ideal" exponentials (unity emission coefficients) and involve no approximations. The actual modeling is now done on Q_b in the denominator. A bipolar transistor model using this approach will be presented in a later paper.

REFERENCES

1. Beaufoy, R., and Sparkes, J. J., "The Junction Transistor as a Charge-Controlled Device," ATE J., *13*, No. 4 (October 1957), pp. 310–324.
2. Persky, G., and Bartelink, D. J., "High Field Energy Distribution and Diffusion Coefficient for Heavy Holes in p-Germanium," Physics Letters, *28A*, No. 11 (March 10, 1969), pp. 749–750.
3. Prior, A. C., "Field-Dependents of Carrier Mobility In Silicon and Germanium," J. Phys. Chem. Solids, *12*, No. 2 (January 1960), p. 175–183.
4. Seidel, T. E. and Scharfetter, D. L., "Dependence of Hole Velocity Upon Electric Field and Hole Density for p-Type Silicon," J. Phys. Chem. Solids, *28*, No. 12 (December 1967), pp. 2563–2574.
5. Ebers, J. J. and Moll, J. L., "Large Signal Behavior of Junction Transistors," Proc. IRE, *42*, No. 12 (December 1954), pp. 1761–1772.

Rain Attenuation and Radio Path Design

By C. L. RUTHROFF

(Manuscript received August 12, 1969)

This paper describes the application of the rain attenuation theory of Ryde and Ryde to the design of Radio Systems. It shows that an upper bound on the outage time due to rain attenuation can be computed from a measured point rain rate distribution. The paper also describes a suitable rain gauge.

I. INTRODUCTION

Heavy rainfall on a radio path absorbs and scatters power transmitted at frequencies above 10 GHz and causes large fading of received signals. At 20 GHz, for example, the attenuation due to a uniform rain rate of 100 mm/hr is about 10 dB/km. Rain attenuation is so severe at these frequencies that for some applications transmission paths must be restricted to a few kilometers or less rather than the tens of kilometers common at lower frequencies. Since the cost of a radio system increases with the number of repeaters it is important to use the longest path allowed by the transmission objectives. This path length can be determined accurately only if the fading outage due to rain attenuation can be predicted.

Bussey estimated fading statistics on a microwave path from point rain rate data.[1] He used the rain attenuation theory of Ryde and Ryde[2-4] to convert rain statistics to fading statistics and since 1950 his results have been used in the design of radio systems.[5] However, as operating frequencies increase and path lengths get shorter, increased precision of fading estimates is required for optimum radio system design.

Over the years a number of experiments have been performed in which attenuation was measured on a path at specific times and compared with values computed from rain rates measured by rain gauges spaced along the path near ground level. Here too, the theory of Ryde and Ryde was used. In general there is wide disagreement between computed and measured values. Also, Medhurst questions the validity of the application of the theory to a practical rainfall situation.[6] Since the theory was derived for uniform rain the question is a good one—rain on a microwave path is seldom uniform.

In this paper the theory of attenuation by uniform rain is applied to the practical rainfall situation. The radio path is defined as the volume of the first Fresnel zone and the rain attenuation is assumed proportional to the number of raindrops in this path volume. Of course, this expression reduces to that of Ryde and Ryde when the rain is uniform.

Rain rate is a vector which can be written as the product of a rain density and the velocity of raindrops. Rain density is proportional to the number of raindrops per unit volume.

Since rain rate is a vector the expression for attenuation in terms of rain density in the path volume can be transformed by the divergence theorem into an expression for attenuation as a function of the rain rate on the surface of the path volume. From this formulation the following results emerge:

(*i*) A natural definition of rain rate which is appropriate to the radio situation.

(*ii*) A time interval, T_o, exists during which no significant fade can occur. T_o is determined by the path length, the frequency of operation and the speed of raindrops.

(*iii*) A rain gauge is described which is suitable for measuring rain rate in accordance with the definition mentioned in (*i*).

By applying these results, an upper bound on the outage time due to rain attenuation is derived. The bound can be computed from a measured point rain rate distribution using the results of uniform rain theory. The bound can be made tight by the proper choice of rain rate integration time interval. A method of estimating this interval from measured path loss distributions is given.

II. GENERAL CONSIDERATIONS

2.1 *The Radio Path*

The radio link consists of two narrow-beam antennas pointing directly at each other over a distance of a few hundred to a few thousand meters. The space, or volume, of the path is taken to be the first Fresnel zone.[7] This means that only the energy confined to that volume contributes significantly to the total energy collected by the receiving antenna.

The first Fresnel zone is a long, thin, prolate ellipsoid of revolution. For a path of length L at wavelength λ, it has a major axis L and equal minor axes $(\lambda L)^{\frac{1}{2}}$ and is terminated at the ends by the antennas. The radio path is defined as the volume enclosed by the first Fresnel zone and the two antennas. Figure 1 is a sketch of the path. When we speak of rain falling on the path we mean rain falling through this volume.

2.2 *Rain Rate as a Vector*

A theory of rain attenuation has been formulated by Ryde and Ryde,[2-4] and others, and a good account of it is given by Medhurst.[6] The attenuation in a radio path depends upon the number and size of the raindrops and not explicitly upon their speed or direction. But the quantity usually measured is rain rate and it does depend on the speed and direction of the raindrops. Since rain rate is the product of a density and a velocity, it can be interpreted as a vector.

Let there be a uniform distribution of N_D drops of water per cubic centimeter in the space between two antennas. The drops are spherical with diameter D and velocity v_D. The fraction of volume occupied by water is defined as the rain density

$$\rho_D \equiv \frac{\pi}{6} N_D D^3. \tag{1}$$

Rain density is a dimensionless, real, nonnegative quantity. The rain rate for drops with diameter D and velocity v_D is

$$R_D \equiv \rho_D v_D .$$

The direction of the rain rate is the direction of travel of the drops.

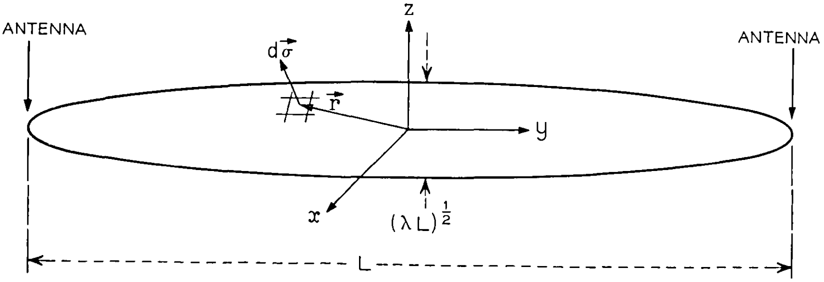


Fig. 1 — Radio path.

The vector expression for rain rate is, therefore,

$$\mathbf{R}_D = \rho_D \mathbf{v}_D , \tag{2}$$

where a boldface letter denotes a vector quantity.

In general a rain storm has drops of many diameters and the total rain rate is a summation over the drop diameters present.

$$\mathbf{R} = \sum_D \rho_D \mathbf{v}_D . \tag{3}$$

2.3 *Attenuation and Rain Density*

For a rain consisting of uniformly distributed drops with diameter D the attenuation of radio waves with wavelength λ is[6]

$$\text{Attenuation} = 4.343 \frac{N_D \lambda^2}{2\pi} A_D \times 10^5 \text{ dB/km} \tag{4}$$

where A_D is a function of the drop diameter, the wavelength, and the dielectric constant of water. Substituting from (1) the attenuation is

$$\alpha_D = k(\lambda, D) L \rho_D \text{ dB}, \tag{5}$$

where

$$k(\lambda, D) = 3 \times 4.343 \frac{\lambda^2 A_D}{\pi^2 D^3} \times 10^5 .$$

The result in (5) is extended to the case of nonuniform spatial distribution of raindrops by replacing the uniform rain density in (5) with the average rain density in the radio path. The path attenuation is therefore assumed to be

$$\alpha_D(t) = k(\lambda, D) \frac{L}{V} \iiint_V \rho_D(x, y, z; t)\, dV \text{ dB}. \tag{6}$$

This expression reduces to (5) for uniform rain density.

There are neither sinks nor sources of rain in the radio path so, for constant drop diameter D, the hydrodynamic equation of continuity which applies is[8]

$$\nabla \cdot (\rho_D \mathbf{v}_D) + \frac{\partial \rho_D}{\partial t} = o,$$

where the time and space variables have been omitted for convenience. Substitution into (6) gives

$$\frac{d\alpha_D(t)}{dt} = k(\lambda, D)\,\frac{L}{V}\iiint_V [-\nabla\cdot(\rho_D\mathbf{v}_D)]\,dV$$

and transforming to a surface integral by the divergence theorem[9] we get

$$\frac{d\alpha_D(t)}{dt} = k(\lambda, D)\,\frac{L}{V}\int_S [-\rho_D\mathbf{v}_D\cdot d\boldsymbol{\sigma}], \tag{7}$$

where the integral is over the surface enclosing the volume V.

Expression (7) relates point rain rate to path attenuation. The vector differential $d\boldsymbol{\sigma}$ has a magnitude equal to the differential area of the surface S, is normal to it and directed outward. Consider a small area on the surface S. The rain rate at this point is $\rho_D\mathbf{v}_D$; the component of this rain rate which is normal to the surface and directed into the path volume is $-\rho_D\mathbf{v}_D\cdot d\boldsymbol{\sigma}/|\,d\boldsymbol{\sigma}\,|$. The integral over the surface S is the increase in water density on the path in unit time which causes an increase in attenuation.

Let the surface S enclosing the radio path volume V be described by orthogonal parametric curves on the surface.[10] A point $P(x, y, z)$ on the surface can be written $P(u, v)$ where,

$$x = \frac{(\lambda L)^{\frac{1}{2}}}{2}\sin u\cos v$$

$$y = \frac{L}{2}\cos u$$

$$z = \frac{(\lambda L)^{\frac{1}{2}}}{2}\sin u\sin v.$$

With this transformation, and substituting R_D from (2), (7) becomes

$$\frac{d\alpha_D(t)}{dt} = k(\lambda, D)\,\frac{L}{V}\iint_S [-\mathbf{R}_D(u, v;\, t)]\cdot d\mathbf{s}, \tag{7a}$$

where $d\mathbf{s}$ is the transformed vector surface differential. Integrating both sides over time T and dividing by T results in the following equation.

$$\frac{\alpha_D(t+T) - \alpha_D(t)}{T} = k(\lambda, D)\,\frac{L}{V}\iint_S\left\{-\left[\frac{1}{T}\int_t^{t+T} R_D(u, v;\, t)\mathbf{r}\,dt\right]\right\}\cdot d\mathbf{s}, \tag{8}$$

where the unit vector $\mathbf{r}$ is defined by $\mathbf{R}_D(u, v;\, t) \equiv R_D(u, v;\, t)\mathbf{r}$.

The left side of (8) is an approximation to the rate of change of attenuation since, by definition,

$$\frac{d\alpha_D(t)}{dt} \equiv \lim_{T \to o} \frac{[\alpha_D(t + T) - \alpha_D(t)]}{T}.$$

Whether attenuation or rain rate is measured, the measuring instrument has some time constant, or integration time, T. Since, in a microwave system, an outage of a millisecond may be significant, an important question arises—can we be sure that the measurement will give us all the important data? In other words does an integration time T_o exist such that for $T \leqq T_o$, (8) is a good approximation to (7a)? The existence of T_o is justified by the following physical argument.

For $t \ll t_o$ let there be no rain on the path. At $t = t_o$ let the rain rate at every point on the surface S be $R_o = \rho_0 v_0$ and be directed inward in the direction of the shortest line from the surface to the path axis. In order that a substantial fade occur, the rain will have to travel a substantial fraction of the shortest distance from the path surface to the path axis. The average of this distance is $(\lambda L)^{\frac{1}{2}}/3$ and the average time required for the rain to reach the axis is $(\lambda L)^{\frac{1}{2}}/3v_o$. An integration time $T_o \ll (\lambda L)^{\frac{1}{2}}/3v_o$ is therefore sufficient since substantial fades will not occur in times less than this. If T is chosen small enough, say $T = T_o$, (8) is a good approximation to (7a), and can be written

$$\frac{d\alpha_D(t)}{dt} \approx k(\lambda, D) \frac{L}{V} \iint_S \left\{ - \left[\frac{1}{T_o} \int_t^{t+T_o} R_D(u, v; t)\mathbf{r}\, dt \right] \right\} \cdot d\mathbf{s}. \tag{9}$$

The quantity in brackets is the rain rate at the point (u, v) averaged over time T_o and defines a suitable rain rate measurement. A rain gauge which measures rain rate in accordance with this definition is described in the Appendix. If the rain rate is known at every point on the surface of the radio path the time derivative of attenuation can be computed from (9).

Since rain rate cannot be measured at all points on the surface of the path we consider what can be done with a more reasonable experiment—one or more rain gauges near the ground in the vicinity of the path. No satisfactory theoretical expression has been derived for computing attenuation from rain rate measurements made near the ground in the vicinity of the path. And the wide disagreement between computed and measured attenuations in experiments of this type described in the literature support the conclusion that the empirical expressions used are also unsatisfactory.[6] Also, the requirements of the sampling theorem must be met if the attenuation is to be computed from sampled rain rates.[11] Visual observation of rainfall reveals a spatial structure so fine that an unreasonably close spacing of rain gauges would be required to meet these requirements.

III. POINT RAIN RATE AND PATH ATTENUATION DISTRIBUTIONS

3.1 *Important Assumptions*

Fortunately, the attenuation as a function of time is not required for the design of radio links; for radio link design the fraction of time that the path attenuation exceeds the fading margin is the important parameter. Thus, we need only a suitable statistic of path attenuation which can be related to a similar statistic of rain rate at ground level in the vicinity of the path. In this section a point rain rate distribution function is defined and related to the path attenuation distribution function.

Let the rain rate be $\mathbf{R}(u, v; t)$ on the surface S of the path. At some point near the path—on the ground beneath the path, for example—a rain gauge measures a sequence of rain rates given by

$$R_n(T, \mathbf{k}, x, y, z) = \frac{1}{T} \int_{t_o+nT}^{t_o+(n+1)T} [-\mathbf{R}(x, y, z; t) \cdot \mathbf{k}] \, dt \tag{10}$$

where $\mathbf{k}$ is a unit vector normal to the collecting surface of the rain gauge and x, y, z are the space coordinates of the rain gauge. Suppose that rain rate measurements have been made for a very long time. The data available is a large number of rain rates $R_n(T, \mathbf{k}, x, y, z)$, one for each interval T. The data is organized by choosing a rain rate R_o and computing the fraction of intervals T for which the rain gauge recorded rates less than R_o. This fraction is denoted by

$$P[R_n(T, \mathbf{k}, x, y, z) \leqq R_o], \tag{11}$$

and is a point rain rate distribution function; it is a function of R_o, the integration time T, the location of the rain gauge, and the pointing direction $\mathbf{k}$. The optimum direction for $\mathbf{k}$ is expected to be vertically upward in many regions; in any case, the rain gauge must be pointed in the direction $\mathbf{k}$ such that for the high rain rates of interest, that is, for $R_o > R_m$,

$$P[R_n(T, \mathbf{k}, x, y, z) \leqq R_0] \leqq P[R_n(T, \mathbf{l}, x, y, z) \leqq R_o],$$

where $\mathbf{l}$ is any unit vector.

For the radio systems of interest, the rain rate, R_m, may be chosen at least one order of magnitude greater than the mean rain rate. In this country the mean rain rate is on the order of 0.1 mm/hr whereas a reasonable value of R_m may be 10 mm/hr.

Let the radio path be divided into a large number, n, of volume elements such that the rain rate is uniform in each element. The average rain rate on the path is

$$R_{ave}(t) \equiv \frac{1}{n} \sum_{i=1}^{n} R_i(x_i , y_i , z_i ; t). \tag{12}$$

In integral form this is written

$$R_{ave}(t) \equiv \frac{1}{V} \iiint_V R(x, y, z; t)\, dV. \tag{12a}$$

Two assumptions are made:

(*i*) In a region containing the radio path the point rain rate distribution function is independent of position and (11) can be written

$$P[R_n(T, \mathrm{k}, x, y, z) \leqq R_o] = P[R_n(T) \leqq R_o]. \tag{13}$$

(*ii*) For the rain rates of interest, that is, for $R_o > R_m$, and for integration time T, the distribution function of the average rain rate on the path is greater than, or equal to, the point rain rate distribution function.

$$P[R_{ave}(t, T) \leqq R_o] \geqq P[R_n(T) \leqq R_o], \tag{14}$$

where,

$$R_{ave}(t, T) \equiv \frac{1}{T} \int_t^{t+T} R_{ave}(t)\, dt.$$

The first assumption is that the point rain rate distribution function is the same whether it is measured below the path, on the path, or near the path. It does not mean that the rain rate at any time t is the same everywhere in the region—an essential distinction. The assumption does not mean that rain rate is a stationary random process in either the wide sense or the strict sense. In statistical language it means that the rain rate can be considered a first order stationary process over a small interval.[12] The second assumption reflects Bussey's observation that high rain rates extend over smaller areas than do lower rain rates.

3.2 *A Bound on the Path Attenuation Distribution*

If the speed of the raindrops does not change while in the radio path, the attenuation can be written in terms of rain rate. From (6),

$$\alpha_D(t) = \frac{k(\lambda, D)}{v_D} \frac{L}{V} \iiint_V R_D(x, y, z; t)\, dV. \tag{15}$$

Let the rain have the Laws and Parsons distribution of drop diameters. Then

$$R_D(x, y, z; t) = R(x, y, z; t)p_D , \tag{16}$$

where p_D is the fraction of water in the rain consisting of drops of diameter D. Substituting into (15) and using the definition (12a) the expression for attenuation is

$$\alpha_D(t) = \frac{k(\lambda, D)}{v_D} Lp_D R_{ave}(t). \tag{17}$$

The total attenuation is

$$\alpha(t) = LR_{ave}(t) \sum_D \frac{k(\lambda, D)}{v_D} p_D . \tag{18}$$

The quantity represented by the summation has been computed by Ryde and Ryde and by Medhurst for the Laws and Parsons drop diameter distribution and for the terminal velocities of water drops in still air.[6]

In Section 2.3 we showed that negligible changes occur in $\alpha(t)$ in a time $T \leqq T_o$. Thus, (18) can be written

$$\alpha(t) \approx \alpha(t, T_o) = LR_{ave}(t, T_o) \sum_D \frac{k(\lambda, D)}{v_D} p_D , \tag{19}$$

where

$$\alpha(t, T) \equiv \frac{1}{T} \int_t^{t+T} \alpha(t)\, dt.$$

The path attenuation for a uniform rain rate R_o is, from (18),

$$\alpha_o = LR_o \sum_D \frac{k(\lambda, D)}{v_D} p_D . \tag{20}$$

The desired bound can be found by substituting from (19) and (20) into (14).

$$P[\alpha(t) \leqq \alpha_o)] \geqq P[R_n(T_o) \leqq R_o]. \tag{21}$$

This bound says that if the measured point rain rate distribution is converted to an attenuation distribution by (20), the outage time predicted is greater than, or equal to, the outage time that occurs on the path. The application of this bound is illustrated in Section V.

IV. EXPERIMENTAL DETERMINATION OF INTEGRATION TIME

It was shown in Section 2.3 that if the rain rate integration time T is short enough no significant fades will be missed; specifically if $T =$

$T_o \ll (\lambda L)^{\frac{1}{2}}/3v_o$, it is small enough. This determination of T_o is based on considerations as to what could happen on a path. Fades may occur slowly compared with T_o so it is worthwhile to determine whether a larger integration time may be practical.

Suppose that path attenuation and point rain rates have been measured with an integration time $T \leqq T_o$. Then the distributions computed from the measurements will satisfy the inequality (21) which can be written

$$P[\alpha(t, T) \leqq \alpha_o] \geqq P[R_n(T) \leqq R_o], \qquad T \leqq T_o. \tag{22}$$

Now, from (14), this inequality holds for any T and since $T \leqq T_o$ all fading of significance is included.

Let the attenuation distribution be computed from the path loss measurements for integration times of $T, 2T, 3T, \cdots$, as long as the distribution remains substantially unchanged. The point rain rate distributions computed for the same integration times are such that the inequality holds and

$$P[\alpha(t, mT) \leqq \alpha_o] \geqq P[R_n(mT) \leqq R_o], \tag{23}$$

for all $m = 1, 2, 3, \cdots$. Now suppose that $P[\alpha(t, mT) \leqq \alpha_o]$ remains substantially unchanged for all integers m up to M. Then all of the corresponding rain rate distributions result in valid upper bounds on outage time as shown by (23); one of these will be the least upper bound.

From experiments of this kind, practical values of integration time can be determined. It is expected (but not proven) that the integration time which results in the least upper bound will be the largest value for which the attenuation distribution remains unchanged, that is, MT.

V. DISCUSSION

Experimental rain rate distributions which meet the requirements of this theory are not available. For this reason a careful experimental verification cannot be made now. There is reason for optimism, however. The upper bound on outage time computed from Bussey's[1] one-minute point rain rate distribution is remarkably close to the one-minute attenuation distributions reported by Semplak and Turrin.[13,14] These distributions are shown in Fig. 2. Semplak and Turrin also report that the attenuation distribution remains unchanged for shorter integration times.[14]

The simplest application of this theory to the design of a radio path requires only a point rain rate distribution measured as described in

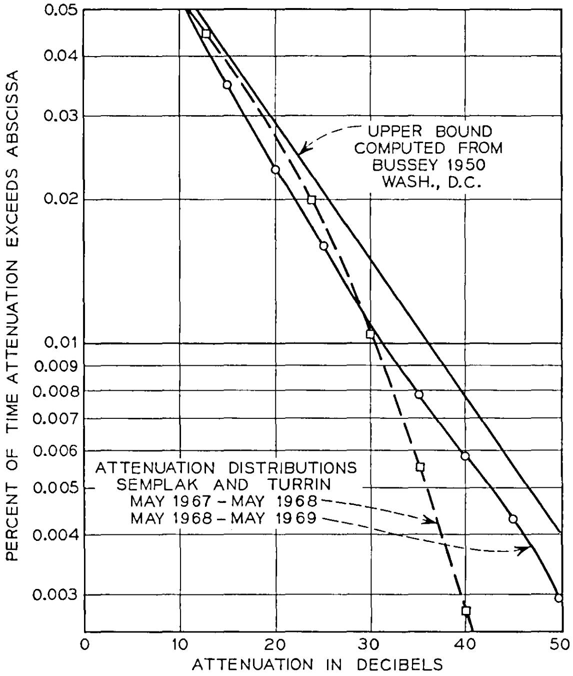


Fig. 2 — Comparison of computed and measured attenuation distributions for one minute integration time.

Section III. In this case the rain rate integration time is computed from the frequency, the path length, and the maximum raindrop velocity. This measured point rain rate distribution can be converted to an upper bound on attenuation outage time by use of expressions (20) and (21). The upper bound obtained this way is not necessarily the least upper bound whereas the least upper bound is required for optimum radio system design. The least upper bound can be computed from the point rain rate distribution if the corresponding optimum rain rate integration time is known; the optimum rain rate integration time can be determined from a path attenuation experiment as described in Section IV.

It is important, then, to determine the optimum integration times in those regions of the country where radio systems above 10 GHz may be used. The optimum integration time is a function of wavelength, path length, and the climate in which the path is located. A few path loss measurements, located in different climatic regions, would yield valuable results. In these experiments attenuation distributions would be measured as functions of path length and wavelength, and optimum integration times computed from this data and the measured point rain rate distributions. It may be, for instance, that the optimum integration times are about the same everywhere; if so, the path loss experiments would show it and, thereafter, only point rain rate measurements would be required.

The accuracy with which the least upper bound predicts the outage time on a radio path cannot be stated precisely until further experimental data is available. It may be anticipated, however, that the accuracy will decrease as the path length increases. For example, if the path is long compared with the dimensions of thunderstorms, the least upper bound prediction will probably be pessimistic, especially for large fades. On the other hand, for paths shorter than the dimensions of thunderstorms the least upper bound prediction may be accurate.

A rain gauge, from which the rain rate in each fixed integration interval can be determined, is suitable for the determination of the point rain rate distribution function defined in Section III. Morgan has built and described such a rain gauge recently, and another rain gauge proposed for this purpose is described in the Appendix.[15]

No attempt has been made to include the effects, on attenuation, of raindrop distortion, temperature, radio wave polarization, and so on, in this theory.

VI. ACKNOWLEDGMENT

I wish to express my sincere appreciation to my colleagues C. Dragone, T. L. Osborne, and V. K. Prabhu for their help and guidance in the preparation of this paper.

APPENDIX

An Instrument for Measuring Point Rain Rates

The instrument described measures rain rates in accordance with the definition of (9). Figure 3 illustrates the basic element in the rain rate instrument—a depth gauge consisting of a funnel, a cylindrical capacitor, and a shutter for draining the capacitor. The funnel is exposed to the rain for a specified time T and then covered. The rain which passed through the collecting area of the funnel drains into the cylindrical capacitor and forms a column of water as shown. The dielectric constant of water increases the capacitance—the greater the height of

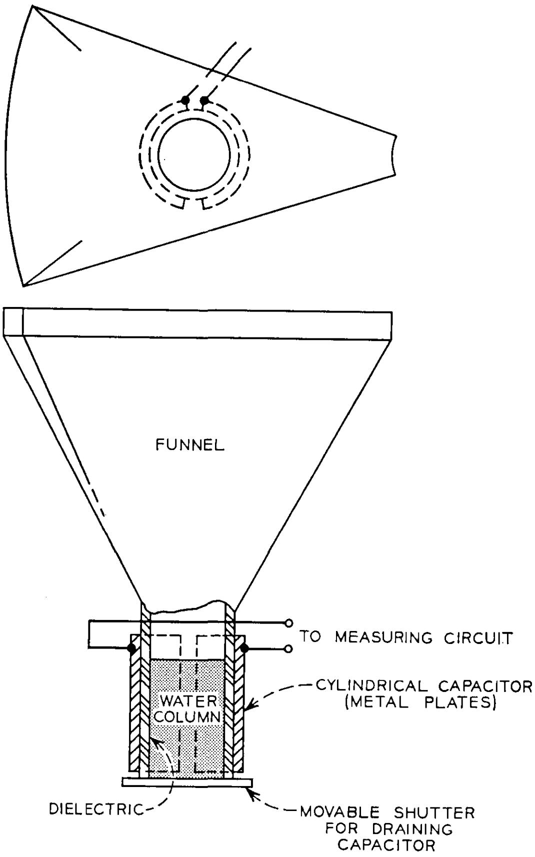

FUNNEL
TO MEASURING CIRCUIT
WATER COLUMN
CYLINDRICAL CAPACITOR (METAL PLATES)
DIELECTRIC
MOVABLE SHUTTER FOR DRAINING CAPACITOR

Fig. 3 — Capacitor depth gauge.

the water column, the greater the capacitance. When the funnel has completely drained, the capacitance (and hence the volume of water) is measured.

There are two features of importance about this depth gauge.

(*i*) The interval of exposure, or integration time, can be determined precisely by careful operation of the shield over the funnel, and is not dependent on the rain rate.

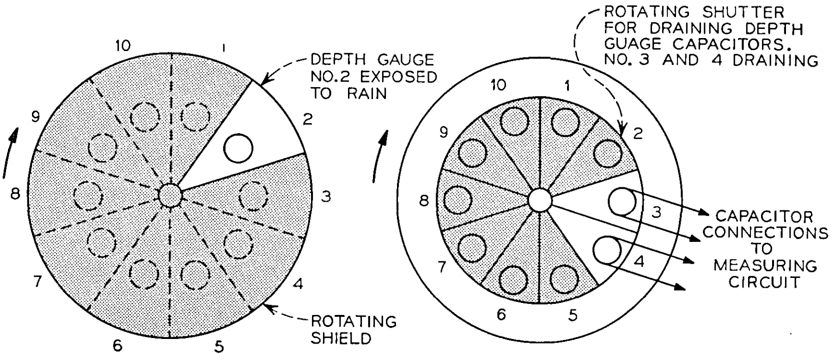
(*ii*) The time allowed to drain the funnel and measure the capacitance is independent of the exposure interval T. Sufficient time can be allowed to drain the funnel and stabilize the water column prior to measuring the capacitor. This eliminates fluctuations due to the random behavior of water flow on the surface of the funnel.

When the shield is over the depth gauge no rain is collected. The complete rain gauge consists of a number of depth gauges arranged as shown in Fig. 4. In this illustration, ten depth gauges are shown, with number 2 in position to collect rain. When the interval ends, the shield rotates, covering depth gauge number 2 and exposing number 3.

A rotating shutter for draining the depth gauge capacitors is shown in Fig. 4B. The shutter is fixed to a common shaft with the funnel shield of Fig. 4A. The operating sequence is as follows. There are ten time intervals of length T in a single rotation of the shaft. With the aid of Fig. 3 the following sequence can be seen to occur.

Time Interval Number	Status of Depth Gauge	
	#2	#3
1	Collecting Water	Draining Capacitor
2	Draining Funnel	Collecting Water
3		Draining Funnel
4		
5		
6		
7		
8	Measure Capacitance	
9	Draining Capacitor	Measure Capacitance
10	Draining Capacitor	Draining Capacitor

The rain shield steps rapidly from gauge to gauge. There is always one gauge collecting rain; one measurement is made in each time interval T. If the measurement starts at time t_o and the rain gauge points


10
1
2
3
4
5
6
7
8
9
DEPTH GAUGE NO.2 EXPOSED TO RAIN
ROTATING SHIELD
ROTATING SHUTTER FOR DRAINING DEPTH GUAGE CAPACITORS. NO. 3 AND 4 DRAINING
10
1
2
3
4
5
6
7
8
9
CAPACITOR CONNECTIONS TO MEASURING CIRCUIT

(a) (b)

Fig. 4(a) Top view of rainrate gauge. (b) Section of rainrate gauge.

vertically upward, the output of the instrument is a sequence of measurements

$$R_n(t, x, y, z) = \frac{1}{T} \int_{t_0+nT}^{t_0+(n+1)T} R(t)\, dt.$$

REFERENCES

1. Bussey, H. E., "Microwave Attenuation Statistics Estimated from Rainfall and Water Vapor Statistics," Proc. I.R.E., *38,* No. 7 (July 1950), pp. 781–785.
2. Ryde, J. W., "Echo Intensity and Attenuation due to Clouds, Rain, Hail, Sand and Dust Storms at Centimeter Wavelengths," Rep. 7831, General Electric Company Research Laboratories, Wembley, England, October 1941.
3. Ryde, J. W., and Ryde, D., "Attenuation of Centimeter Waves by Rain, Hail and Clouds," Rep. 8516, General Electric Company Research Laboratories, Wembley, England, August 1944.
4. Ryde, J. W., and Ryde, D., "Attenuation of Centimeter and Millimeter Waves by Rain, Hail, Fogs and Clouds," Rep. 8670, General Electric Company Research Laboratories, Wembley, England, May 1945.
5. Hathaway, S. D., and Evans, H. W., "Radio Attenuation at 11 KMC," B.S.T.J., *38,* No. 1 (January 1959), pp. 73–97.
6. Medhurst, R. G., "Rainfall Attenuation of Centimeter Waves: Comparison of Theory and Measurement," IEEE Trans. Antennas and Propagation, *AP-13,* No. 4 (July 1965), pp. 550–564.
7. Slater, J. C., and Frank, N. H., *Introduction to Theoretical Physics,* New York: McGraw-Hill, 1933, pp. 307–314.
8. Petterssen, S., *Weather Analysis and Forecasting,* New York: McGraw-Hill, 1956, pp. 7–8.
9. Lass, H., *Vector and Tensor Analysis,* New York: McGraw-Hill, 1950, p. 114.
10. Struik, D. J., *Differential Geometry,* Cambridge, Massachusetts: Addison-Wesley Press, 1950, p. 64.
11. Ruthroff, C. L., "Microwave Attenuation and Rain Gauge Measurements," Proc. IEEE, *57,* No. 6 (June 1969), pp. 1235–1236.
12. Papoulis, A., *Probability, Random Variables, and Stochastic Processes,* New York: McGraw-Hill, 1965.
13. Semplak, R. A., and Turrin, R. H., "Some Measurements of Attenuation by Rainfall at 18.5 GHz," B.S.T.J., *48,* No. 6 (July–August 1969), pp. 1745–1756.
14. Semplak, R. A., unpublished work.
15. Morgan, B. E., "A Rainfall Rate Sensor," Technical Memorandum ERLTM-NSSL 42, Experimental Science Services Administration Research Laboratories, National Severe Storms Laboratory, Norman, Oklahoma, November 1968.

An Improved Thermal Gas Lens for Optical Beam Waveguides

By P. KAISER

(Manuscript received September 3, 1969)

The quality of thermal gas lenses can be significantly improved if the heated gas is exhausted radially. This type of exhaust is typically created in a counter-flow lens where opposing gas flows pass through two closely spaced lens sections and exhaust radially through a small gap between these sections. We found that the stability of the focusing behavior depends on a laminar flow transition region of sufficient length, a properly chosen width and shape of the exhaust gap, and on mechanical symmetry and smoothness.

We describe the guiding properties of a single lens as well as a beam waveguide consisting of up to eleven counter-flow lenses. Off-axis injections of the light beam with displacements amounting up to 45 *percent of the lens radius resulted only in changes of the beam width (less than* 23 *percent for* 11 *lenses), whereas the gaussian profile was essentially maintained.*

I. INTRODUCTION

Gravitational and spherical aberrations in thermal gas lenses are known to cause distortions of light beams with gaussian intensity distribution.[1–5] The eventual use of this lens in long distance optical communication links depends in part on our ability to improve its quality and to develop effective beam control devices which periodically free the distorted beam from higher order modes.[6]

By analyzing the severe aberrations of the thermal gas lens reported in an earlier paper in this journal,[4] we found that the axial exhaust of the heated gas was primarily responsible for the low quality of that lens. We demonstrate that a significant improvement of the focusing properties of the thermal gas lens can be achieved if the heated gas is exhausted in the radial direction.

II. CRITICAL APPRAISAL OF AN EARLIER THERMAL GAS LENS

The thermal gas lens used in a recent experiment[4] was highly aberrated and caused severe beam deformations already for small off-axis injections of a light beam into the lens guide. Attempts to decrease the predominantly gravitational aberrations by increasing the flow velocity resulted in instabilities of the focusing action. This was the apparent consequence of residual turbulent motion of a gas flow with nonuniform temperature distribution. We found that the occurrence of flow instabilities was attributable to its construction (see Fig. 1). Exponential inlet and exhaust regions connected the lens elements with a larger diameter spacing tube which served as heat sink for cooling purposes. At higher flow rates, the laminar flow transition tubes preceding the heated lens elements were too short (stable focusing action requires the gas flow to be fully developed and free from residual random motion when it enters the heated lens section). For that purpose the gas has to pass through a tubular transition region with the same inner diameter as the lens and whose minimum length L_t depends on this diameter and the Reynolds number R_e,[7]

$$L_t \geqq 0.04\, dR_e$$

$$R_e = \frac{\langle v \rangle_{\text{av}} \cdot d}{\nu}$$

d = lens diameter (=0.635 cm)
$\langle v \rangle_{\text{av}}$ = average gas velocity
ν = kinematic viscosity (0.15 cm^2/s for air).

For an air flow rate of 3 liters per minute and for a resulting R_e of 670, the minimum length is at least 27 lens diameters. However, this value is only approximate since the stratification of the flow depends on the particular type of injection used and the initial amount of eddies present. Furthermore, it depends on the degree of beam stability required.

Another reason that focusing fluctuations occurred was the presence of flow separation in the fast expanding exhaust region, which again resulted in a fluctuating temperature distribution. Because of its attractively simple design, we tried to retain the axial exhaust of the continuous flow system, and we attempted to improve its shortcomings. Through the selection of a longer transition tube, laminar flow can be obtained in the entrance region. However, in order to avoid flow separation, expansion of the diffuser behind the lens has to occur very gradually. Eventually the gas is not cooled back to its original temperature by mere heat exchange with an insulated diffuser wall and the diffuser would have to serve as a heat sink as well. During the cooling process the temperature profile becomes distorted. Thus, any cooling

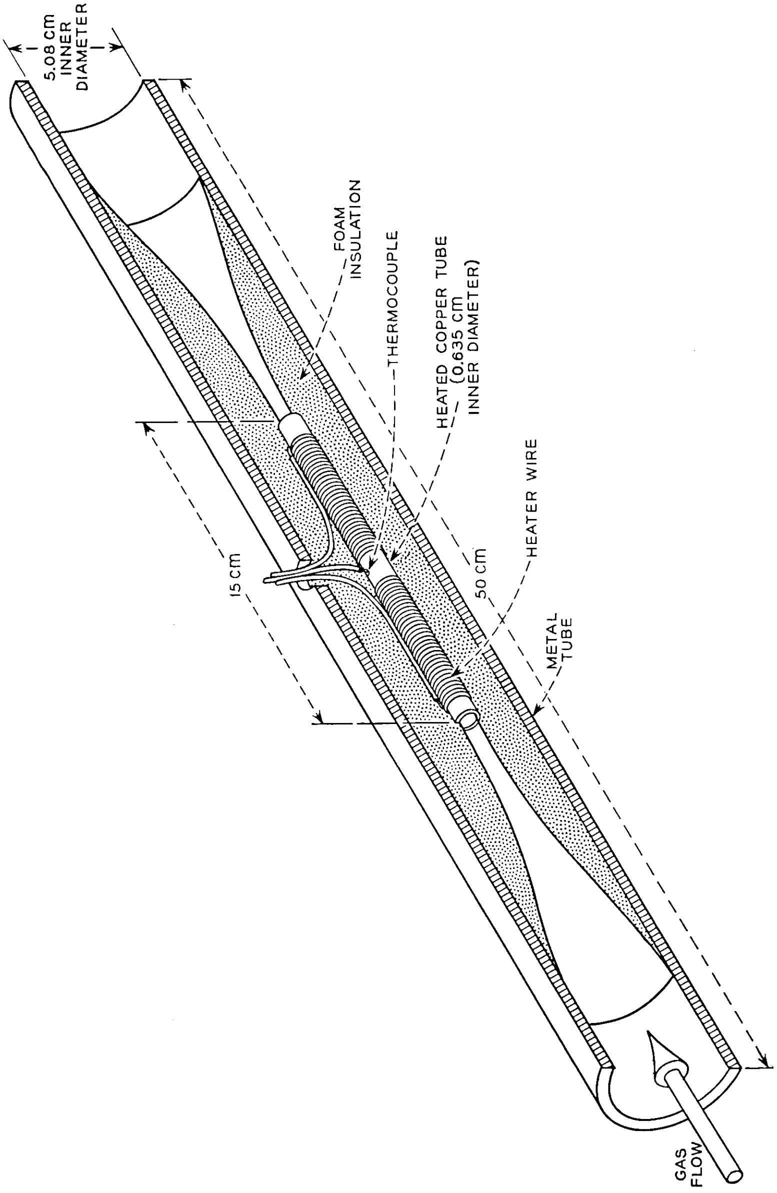

5.08 cm
INNER
DIAMETER
FOAM
INSULATION
THERMOCOUPLE
HEATED COPPER TUBE
(0.635 cm
INNER DIAMETER)
HEATER WIRE
50 cm
15 cm
METAL
TUBE
GAS
FLOW

Fig. 1 — Thermal gas lens with axial exhaust.

of the heated gas in the path of the light beam appears to be associated with large aberrations. With our present knowledge, therefore, their practical usefulness seems to be rather limited.

We can avoid the lens deteriorations which are associated with an axial exhaust by exhausting the heated gas radially. This type of exhaust is suitably created through the confluence of two opposing flows in the counter-flow lens (CFL). This lens consists of two tubular lens sections which are separated by a small gap (Fig. 2). Laminar flows of cold gas enter the counter-flow lens from both sides and exhaust through the gap in the center via a free stagnation flow. The flow pattern and optical properties of the counter-flow lens are the subject of the following study.

The possibility of reducing the spherical aberrations of the thermal gas lens by using two such lenses with opposing flows back to back was originally suggested by Marcuse.[1] His theoretical analysis, however, neglected gravity forces as well as effects in the exhaust region. An experimental comparison between a counter-flow lens and a lens of identical length with unidirectional flow, whose radial exhaust at the end of the lens was created with a glass plate, revealed no discernible difference between their focusing qualities. The theoretically proven difference in their spherical aberrations is obviously small and becomes apparent only after the passage of many lenses. Nevertheless, it adds to the attractive features of the counter-flow lens. Berreman had already used the counter-flow principle in his "chimney lens", in which the flow velocity was determined by free convection.[8]

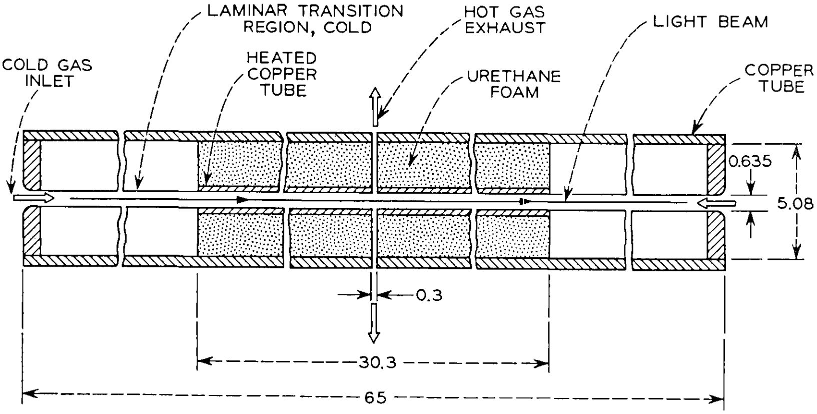


Fig. 2 — Thermal counter-flow gas lens with radial exhaust. (All dimensions are in centimeters.)

III. FLOW CHARACTERISTICS OF THE COUNTER-FLOW LENS

The counter-flow lens used in our experiments consisted of two identical lens halves which were mounted independently (see Figs. 2 and 7). Air was injected into the lens from both sides via 5.08 cm diameter porous tubes (2.54 cm long). These were followed by 0.635 cm diameter flow transition tubes and copper tubes of identical diameter. The assembly was surrounded by a 5.08 cm diameter copper tube for mechanical rigidity. The use of porous tubes allowed us to reduce the length of the transition tubes and still achieve laminar flow. Their omission caused unstable focusing and forced us to increase the length of the transition tubes beyond that indicated in Fig. 2. We concluded at this point that we can shorten the transition length by laminar injection of the gas through porous tubes having identical diameters as the transition tube and lens. This assumption proved to be true in later experiments.

In a typical mode of operation, the counter-flow lens is closed on both sides with glass windows in order to prevent leakage of the gas. In a simplified mode of operation, we can omit the windows if we suck air from the surrounding atmosphere into the lens by means of a vacuum pump connected to the exhaust gap. Aside from its simplicity, the latter method allows us to determine the focusing characteristics of the counter-flow lens more accurately and without the interference of end windows. Note that the length of the transition tubes in this case had to be increased to more than twice the length indicated in Fig. 2 in order to avoid beam fluctuations.

A properly chosen width and shape of the exhaust gap proved to be important for avoiding flow instabilities. When the gap was either too small or too large, we observed fluctuations of the focused beam, particularly for slightly misaligned lens halves and asymmetric flows. We observed most stable focusing when the width of the gap assumed approximately the same value as the lens radius. In this case, the radial exhaust area amounted to twice the cross-sectional area of one lens. As might be imagined, the flaring of the inner diameter near the exhaust gap was also noticed to have some influence on the beam stability. Tubes having inner diameters rounded off and with a radius whose value corresponded to that of the wall thickness (1.6 mm) proved to be inferior to a sharp-edged exhaust. But we believe that some rounding of the edges helps to avoid flow instabilities in that region. Similarly, lack of smoothness of the end faces of both lens halves in the exhaust gap was also found to be conducive to instabilities.

As a result of this study, we recognize that flow instabilities pose a major problem in the operation of the counter-flow lens. This is particularly true when we remember a fact well known in the field of fluid amplification: depending on the boundary conditions, the flow in a rapidly expanding channel can assume several flow patterns. Thus, slight perturbations of the flow or the boundary conditions can cause it to change, for example, from an axially symmetric pattern to an asymmetric pattern, or to oscillate between two asymmetric patterns. Therefore, it seems to be advisable to introduce deliberately a certain asymmetry into the exhaust gap. Its specific character must be left to future investigation, but it is desirable that it does not adversely affect the lens quality.

In studying the instabilities, we limited ourselves to low frequent beam fluctuations detectable by visual observation and by an ordinary x-y recorder.

IV. OPTICAL CHARACTERISTICS OF THE COUNTER-FLOW LENS

The counter-flow lens is put into operation by establishing the proper gas flow rate and heating the lens elements to a temperature which yields the desired focal length. The temperature difference between wall and inlet air (at room temperature) was limited to approximately 100°C because the foam insulation (Nopcofoam H402N) could not tolerate higher temperatures.

Coaxial alignment of the counter-flow lens with a slightly diverging gaussian light beam was accomplished with the aid of a photoelectric probe containing a pinhole. The radius of the unfocused beam at the center of the lens was 0.66 mm. Supported by theoretical analysis, we assumed the center of the counter-flow lens to be the location of the principal plane.[1] We measured the profile of the focused beam at an arbitrary distance of approximately 4 meters from the lens with the aid of an automatic recording system, consisting of a motor-driven photosensor with a pinhole, movable in two perpendicular directions, an amplifier, and an x-y recorder.

The focal length was determined by a substitution method: The half power width of the focused beam was compared with that of a set of glass lenses with known focal lengths. These were placed into the beam instead of the counter-flow lens, yielding a calibration curve beamwidth versus focal length. The focal length f as function of the temperature difference ΔT between wall and inlet air for different flow rates F (with $F/2$ flowing in either lens half) is shown in Fig. 3.

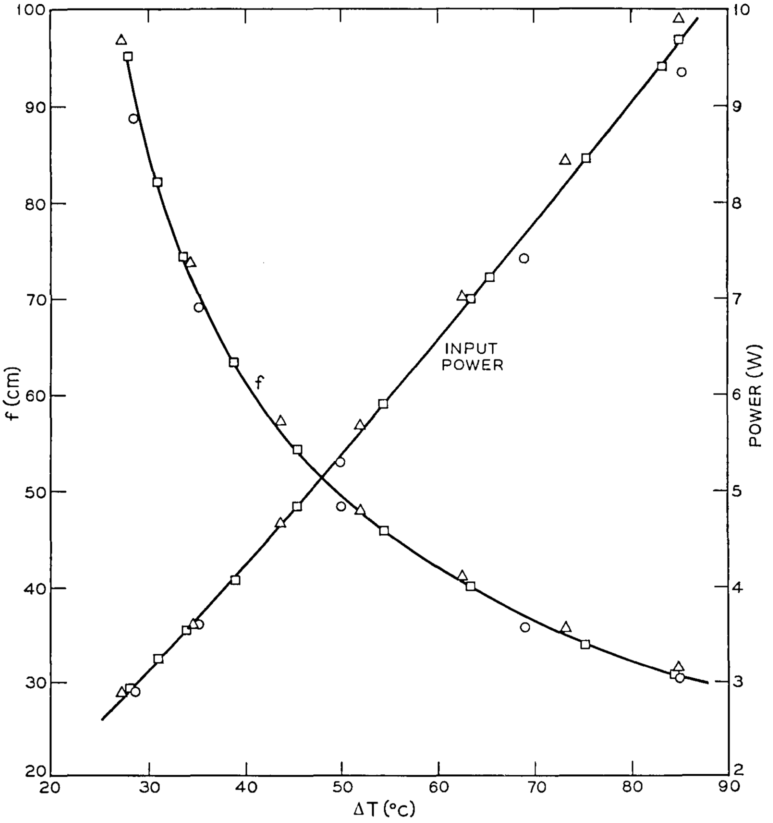


Fig. 3 — Focal length and power consumption of the counter-flow lens as function of the temperature rise for different air flow rates, F. (○ — 5.25 L/min; □ — 5.75 L/min; △ — 6.25 L/min.)

We selected the respective flow rates because we found them to be associated with smallest aberrations as we will see below. We notice that at above values the focal length is essentially independent of the flow rate and depends mainly on the temperature rise. Theoretical analysis[9] has shown that also without the consideration of free convection effects, minimized focal length distortions coincide with flow rates for which the focal length assumes minimum values.

The power consumption P of the counter-flow lens increases as function of the flow rate as shown in Fig. 3. For a focal length of 0.40 m, the required input power was 7.08 W (F = 6.0L/min, ΔT = 61.8°C).

The same temperature difference without gas flow requires an input power of 2.48 W. The difference of 4.60 W is thus necessary to heat the gas. The theoretical value for the heat transferred to the gas at above ΔT is given by[10]

$$P_{th} = \pi k L \ \Delta T \frac{v_o}{V} \left[1 - 0.820 \exp \left(-7.316 \frac{V}{v_o} \right) \right]$$

with

k = thermal conductivity of gas
L = length of one heated lens element
v_o = maximum gas velocity
$V = \dfrac{kL}{a^2 \rho \, c_p}$
a = lens radius
ρ = gas density
c_p = specific heat at constant pressure

and is calculated to be $4.65W$ ($k = 6.28 \cdot 10^{-5}$ calories/cm second degree, $L = 15$ cm, $\Delta T = 61.8$°C, $v_o = 318$ cm/s, $a = 0.3175$ cm, $\rho = 1.21 \cdot 10^{-3}$ gram/cm^3, $c_p = 0.240$ calories/gram degree).

Making use of the fact that the temperature profile of the gas flow is still maintained for some distance behind the heated tube, we could reduce the power consumption of the thermal gas lens by substituting the latter portion of the heated tubes with insulating tubes.

As mentioned before, for the given geometry, we observed minimized lens aberrations at flow rates near 6 liters per minute. We determined these optimum flow rates in the following way: a gaussian beam was injected into the lens with increasing parallel displacements. Since the focal length in the aberrated gas lens changes with the radius, we used the resulting changes of the beam width as a measure of distortion. Normalized changes of the width as function of off-axis injections for different flow rates are presented in Fig. 4. At flow rates below approximately 6 liters per minute we found pronounced asymmetry for vertically displaced beams due to gravity. At flow rates beyond 6 liter per minute this asymmetry gradually disappeared, but similar symmetric changes of both the vertical and horizontal beam width for growing off-sets indicated the increasing presence of spherical aberrations. Near 6 liters per minute we recorded smallest changes of the width, but still noted a vertical asymmetry.

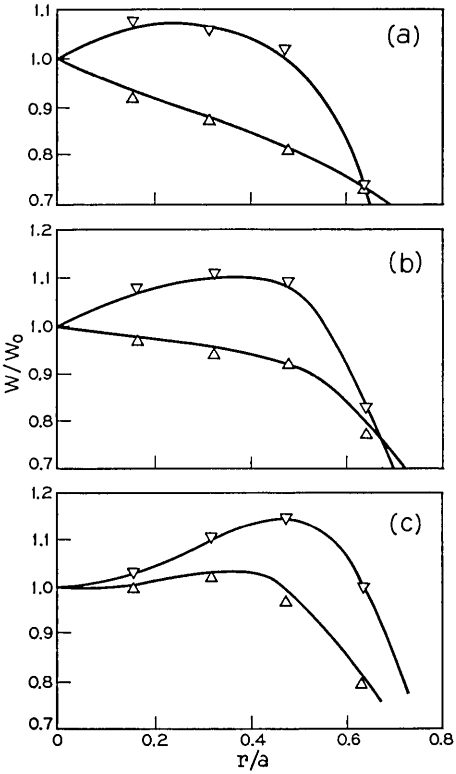


Fig. 4 — Relative change of beam width as function of off-axis injections for beams injected above (△) and below (▽) the optical axis (f = 0.40 m, $\triangle T$ = 62°C, w_o = on-axis beam width, a = lens radius). (a) F = 5.5 L/min; (b) F = 6.0 L/min; (c) F = 6.5 L/min.

The method of off-axis injection permits us to probe exactly the different portions of the counter-flow lens with their aberrations. A somewhat faster, but less detailed method of determining the optimum flow rate is to focus a comparatively large beam which fills a major portion of the lens's cross section. The distorting influence of the lens aberrations is thus magnified. The beam radius chosen for that purpose was 1.10 mm. The temperature at selected flow rates was adjusted so as to result in a focal length of 40 cm, measured as a constant half power width of the horizontal profile. As a measure of distortion, we used the change of the intensity profile at the 1/10 power point, whereby the profile obtained with a high quality glass lens served as reference (Fig. 5). In case of the asymmetrically focused vertical profile, we considered

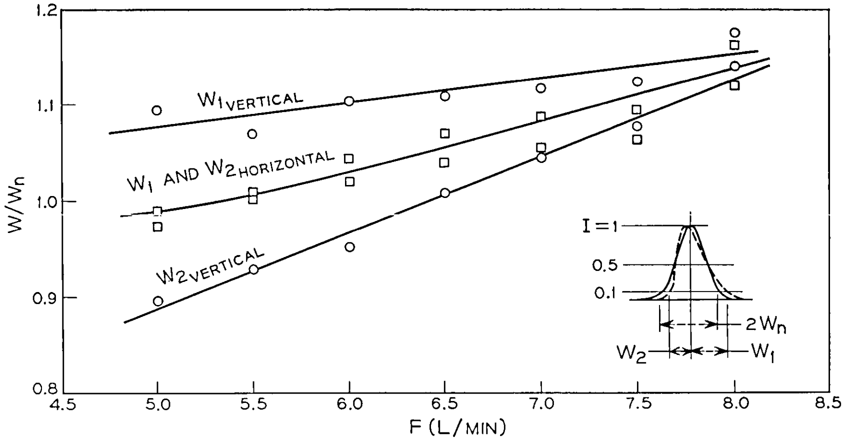


Fig. 5 — Distortion of an oversized beam as function of the flow rate (f = 0.40 m). —— reference beam; — — — — deformed beam; normalized at $I = 1$ and $I = 0.5$ with reference beam.

the changes of the upper and lower half width—counted from the position of the maximum intensity—as a measure of the gravitational aberrations. The approximate symmetric increases for horizontal and vertical profiles at higher flow rates are due to spherical aberrations. As with the method before, the optimum flow rate is again seen to lie somewhere between 5.5 and 6 L/min, where horizontal beam width changes are smallest. Temperature differences and the power consumption at the different flow rates are shown in Fig. 6.

The flow rates associated with minimized total aberrations still have noticeable gravity aberrations which have to be compensated for. For a coaxial alignment of the two lens halves and for the flow rates and temperatures of interest, the difference between the mechanical and optical axes was approximately 0.20 mm. This compares favorably with the 0.5 mm measured with an earlier gas lens with axial exhaust.[4]

If we align the counter-flow lens on its optical axis, the focal length and power consumption are slightly different from those shown in Fig. 3. Note that the optical axis itself depends on the flow rate and the temperature difference.

V. A BEAM WAVEGUIDE COMPOSED OF COUNTER-FLOW LENSES

For the purpose of examining further the flow pattern and the optical properties, we constructed a beam waveguide consisting of up to eleven counter-flow lenses. We supported the lenses in a U-channel in the same fashion as described earlier.[4] The air was supplied by a 2-inch pipe which ran parallel to the guide and had taps with valves as needed (Fig. 7).

The fundamental mode of the beam waveguide was generated in a laser cavity which consisted of a curved ($R = 0.80$ m) mirror and a plane mirror, and a 17-cm-long He-Ne discharge tube with a 2-mm bore. The ideal beam radius at the center of each lens amounted to 0.382 mm.

The first lens was aligned coaxially with the beam at a distance of one lens spacing from the curved mirror. Near the laser, the lens was closed with a flat, antireflection coated window. On the opposite side (behind feed 2; see Fig. 7) the lens was temporarily closed with a glass window for the purpose of establishing air flows in both lens halves. The lenses were then aligned on their optical axes. The temperatures were adjusted such as to produce the same width of the focused beam as that of the original beam after the target was moved farther away from the laser cavity the equivalent of one lens spacing. By doing this for all lenses, we guaranteed periodicity of the guided beam despite variations existing between their mechanical and optical properties. Thereafter the first half of the second lens was connected with feed 2. We used a flexible coupling between the lenses in order to avoid mechanical feedback. After we removed the alignment-probe, we increased the gas flow into the second feed to 6 liters per minute. Since the temperature of the thermal gas lens is sensitive to changes of the flow rate, and since the

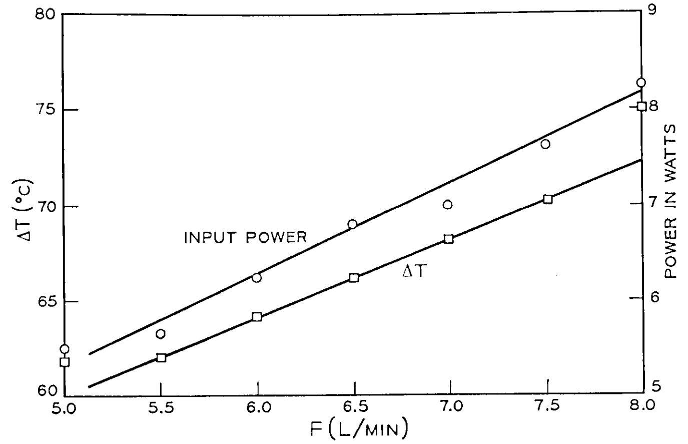


Fig. 6 — Temperature differences and power consumption as function of the gas flow rate for a constant focal length of 0.40 m.

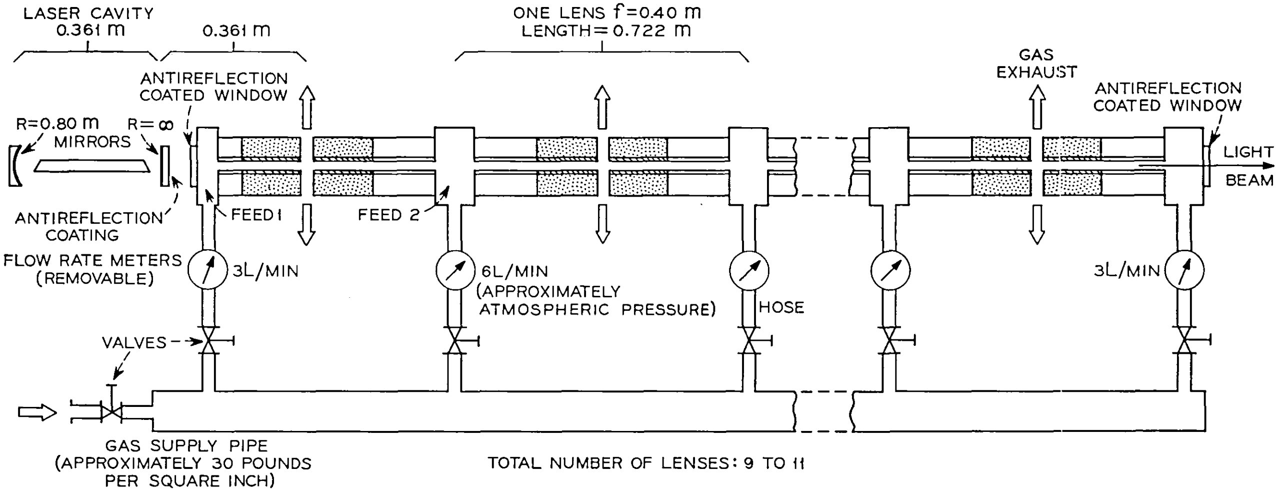

LASER CAVITY
0.361 m
0.361 m
ONE LENS f=0.40 m
LENGTH=0.722 m
GAS
EXHAUST
ANTIREFLECTION
COATED WINDOW
R=0.80 m
R=∞
MIRRORS
ANTIREFLECTION
COATED WINDOW
LIGHT
BEAM
ANTIREFLECTION
COATING
FEED 1
FEED 2
FLOW RATE METERS
(REMOVABLE)
3L/MIN
6L/MIN
(APPROXIMATELY
ATMOSPHERIC PRESSURE)
HOSE
3L/MIN
VALVES
GAS SUPPLY PIPE
(APPROXIMATELY 30 POUNDS
PER SQUARE INCH)
TOTAL NUMBER OF LENSES: 9 TO 11

Fig. 7 — Beam waveguide composed of counter-flow lenses.

temperature of the first lens did not change appreciably after the flow into feed 2 was increased, we had an indication for the symmetric subdivision of the total flow into both directions. Subsequently, the second half of the second lens was aligned and the lens put into operation as before. All other lenses were aligned analogously.

The vertical and horizontal power profiles of the on-axis beam after passing eleven lenses, together with profiles of beams which were injected into the guide with increasing parallel displacements are shown in Fig. 8. For beam displacements amounting up to approximately 45 percent of the lens radius (relative off-axis injection: 0.45), the gaussian profile was maintained and only changes of the beam width took place. One notes that the changes of the beam width for vertical and horizontal off-sets are comparable.

Due to the different phase relationships existing between the fundamental and higher order modes at successive lenses, we obtain a more accurate picture of the higher order mode content by comparing the beam profiles after the ninth, tenth, and eleventh lenses of the guide (Fig. 9). After nine lenses, the beam width first decreased for growing offsets, went through a minimum around 0.3, and approached again its original width near 0.45. Then it rapidly increased, and the profile deteriorated for larger offsets. As could be expected from the phase progression of the second order mode, which is mainly responsible for beam width changes, a similar dependence was observed after eleven lenses. Here, the maximum changes of the beam width were in the order of 10 percent for relative offsets smaller than approximately 0.45. Largest changes of the beam width for up to 11 lenses occurred for 10 lenses and were in the order of 23 percent for relative off-axis injections up to 0.45.

In Fig. 10 we show relative changes of the beam width for a sequence of 10 gas lenses, and also for a single lens with axial exhaust for comparison (see Fig. 1 and Ref. 4).

VI. CONCLUSIONS

The focusing properties of a thermal gas lens with radial exhaust were shown to be distinctly superior to those of a lens with axial exhaust. Laminar flow transition tubes of sufficient length and a properly chosen width and shape of the exhaust gap resulted in a flow pattern with stable focusing action. It also permitted us to increase the gas flow rate to a value at which the combined gravitational and spherical

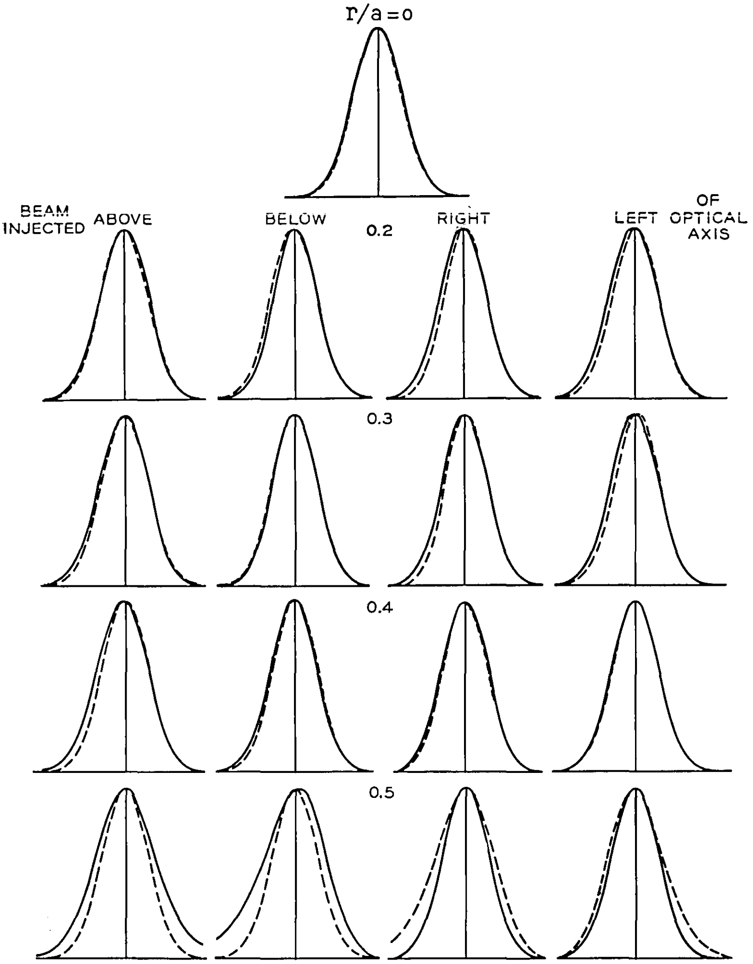


Fig. 8 — Vertical (——) and horizontal (— — — —) power profiles of beams injected off-axis into a beam waveguide composed of 11 counter-flow lenses (f = 0.40 m, distance/focal length = 1.8).

aberrations were minimized. For a total length of the heated lens section of 30 cm (Fig. 2), the optimum flow rate was near 6 liters per minute (diameter of the lens: 0.635 cm). At this flow rate an input power of 7 watts was required to maintain a temperature difference of 62°C between the inlet air and the wall, resulting in a focal length of 0.40 m. An input power of 2.5 watts was required to establish the same wall temperature without gas flow.

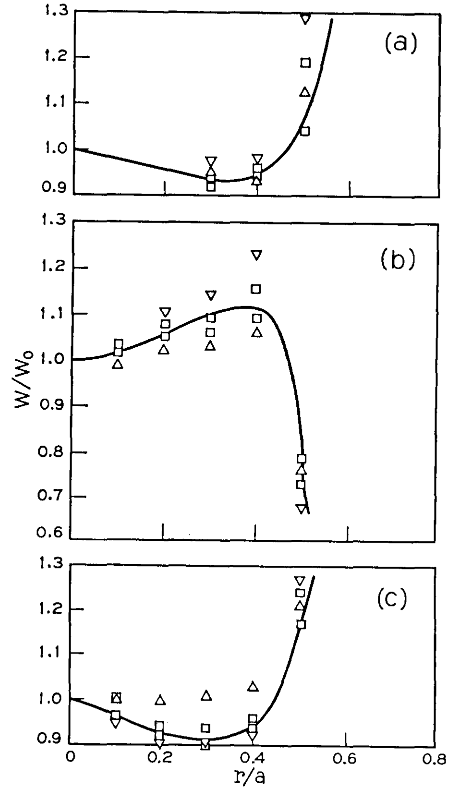


Fig. 9 — Relative change of beam width as function of off-axis injection into a guide composed of (a) 9, (b) 10, and (c) 11 counter-flow lenses ($f = 0.40$ m, $D/f = 1.8$, w_o = on-axis width, a = lens radius). Beam injected above (△), below (▽), and right and left (□) of optical axis.

The guiding properties of a beam waveguide consisting of up to eleven counter-flow lenses were found (focal length 40 cm, lens spacing to focal length ratio 1.8): off-axis injections of the light beam with displacements amounting up to 45 percent of the lens radius resulted only in changes of the beam width (less than 23 percent for up to 11 lenses), whereas the gaussian profile was essentially maintained. In comparison, a beam passing through thermal gas lenses with axial exhaust was completely distorted when injected with similar displacements.

Thus, the counter-flow lens was proven to be a gas lens of superior quality which warrants its use in a lens guide encompassing a larger number of these lenses. In view of this, several improvements and further studies are presently under way which we will report on at a later date.

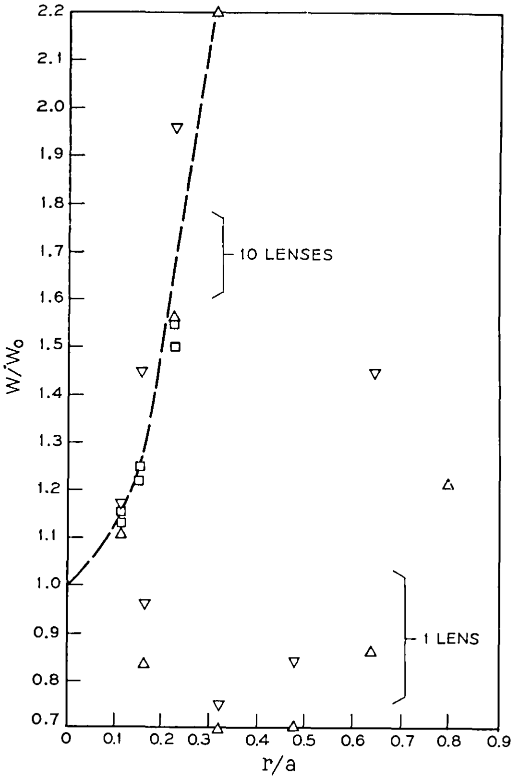


Fig. 10 —Relative increase of beam width for a beam injected off-axis into gas lenses with axial exhaust ($f \cong 0.50$ m, $F = 1.5L$/min, $\Delta T \cong 110$°C). Beam injected above (△), below (▽), and right and left (□) of optical axis.

VII. ACKNOWLEDGMENTS

I gratefully acknowledge the valuable comments and recommendations by E. A. J. Marcatili. I appreciate the assistance of M. D. Divino, who helped in carrying out the experiments. A discussion with I. Pelech helped to clarify some aspects of free stagnation flow.

REFERENCES

1. Marcuse, D., "Deformation of Fields Propagating Through Gas Lenses," B.S.T.J., *45,* No. 8 (October 1966), pp. 1345–1368.
2. Marcatili, E. A. J., "Off-Axis Wave-Optics Transmission in a Lens-like Medium with Aberration," B.S.T.J., *46,* No. 1 (January 1967), pp. 149–166.
3. Gloge, D., "Deformation of Gas Lenses by Gravity," B.S.T.J., *46,* No. 2 (February 1967), pp. 357–365.

4. Kaiser, P., "Measured Beam Deformations in a Guide Made of Tubular Gas Lenses," B.S.T.J., *47,* No. 2 (February 1968), pp. 179–194.
5. Steier, W. H., "Optical Shuttle Pulse Measurements of Gas Lenses," Appl. Opt., *7,* No. 11 (November 1968), pp. 2295–2300.
6. Marcatili, E. A. J., "Effect of Redirectors, Refocusers, and Mode Filters on Light Transmission Through Aberrated and Misaligned Lenses," B.S.T.J., *46,* No. 8 (October 1967), pp. 1733–1752.
7. Schlichting, H., *Boundary Layer Theory,* New York: McGraw-Hill, 1960, p. 171.
8. Berreman, D. W., "Convective Gas Light Guides or Lens Trains for Optical Beam Transmission," J. Opt. Soc. Amer., *55,* No. 3 (March 1965), pp. 239–247.
9. Marcuse, D., "Theory of a Thermal Gradient Gas Lens," IEEE Trans. Microwave Theory and Techniques, *MTT-13,* No. 6 (November 1965), pp. 734–739.
10. Marcuse, D. and Miller, S. E., "Analysis of a Tubular Gas Lens," B.S.T.J., *43,* No. 4 (July 1964), pp. 1759–1782.

Contributors to This Issue

SIDNEY DARLINGTON, B.S., 1928, Harvard; B.S.E.E., 1929, Massachusetts Institute of Technology; Ph.D., Columbia University, 1940; Bell Telephone Laboratories, 1929—. Mr. Darlington has been mainly concerned with research in applied mathematics, relating to circuits, systems, and communication theory. Fellow, IEEE; Associate Fellow, American Institute of Aeronautics and Astronautics.

GERARD J. FOSCHINI, B.S.E.E., 1961, Newark College of Engineering; M.E.E., 1963, New York University; Ph.D. (Mathematics), 1967, Stevens Institute of Technology; Bell Telephone Laboratories, 1961—. Mr. Foschini had initially worked on real time program design. Since 1965 he has engaged in analytical work concerning the transmission of signals over stochastic channels. Member, Sigma Xi, Mathematical Association of America.

DETLEF GLOGE, Dipl. Ing., 1961, D.E.E., 1964, Braunschweig Technische Hochschule (Germany); research staff, Braunschweig Technische Hochschule, 1961–1965; Bell Telephone Laboratories, 1965—. In Braunschweig, Mr. Gloge was engaged in research on lasers and optical components. At Bell Telephone Laboratories, he has concentrated on the study of optical transmission techniques. Member, Verband Deutscher Elektroingenieure, IEEE.

J. E. GOELL, B.E.E., 1962, M.S., 1963, and Ph.D. (E.E.), 1965, Cornell University; Bell Telephone Laboratories, 1965—. While at Cornell, Mr. Goell was a teaching assistant and held the Sloan Fellowship and the National Science Cooperative Fellowship. At Bell Telephone Laboratories, he has worked on solid-state repeaters for millimeter wave communication systems and optical integrated circuits. Member, Tau Beta Pi, Eta Kappa Nu, Sigma Xi, Phi Kappa Phi, IEEE.

HERMANN K. GUMMEL, Diplom-Physiker degree (1952), University of Marburg, Germany; M.S. (physics), 1952, Ph.D. (physics), 1957, Syracuse University; Bell Telephone Laboratories, 1957—. He has worked in semiconductor electronics and presently heads a department responsible for design analysis. Member, American Physical Society, Sigma Xi.

W. M. HUBBARD, B.S., 1957, Georgia Institute of Technology; M.S., 1958, University of Illinois; Ph.D., 1963, Georgia Institute of Technology; Bell Telephone Laboratories, 1963—. Mr. Hubbard's work has included analyses related to the design of millimeter-wave solid-state repeaters for use in a waveguide transmission system and the construction of prototype high-speed repeaters for this type of system. Member, Sigma Xi, Tau Beta Pi, Phi Kappa Phi, American Physical Society.

PETER KAISER, Diplom Ingenieur, 1963, Technische Hochschule, Munich, Germany; M.S., 1965, and Ph.D., 1966, University of California, Berkeley; Bell Telephone Laboratories, 1966—. At Berkeley, Mr. Kaiser was working on frequency independent antennas. He now is engaged in optical transmission research with emphasis on gas lens beam waveguides. Member, IEEE.

G. D. MANDEVILLE, 1933–34, Monmouth Junior College; 1935–36, Rutgers University; Western Electric Co., 1939–49; Bell Telephone Laboratories, 1949—. With Western Electric, Mr. Mandeville was concerned with radar development and shop test equipment. He headed the shop test equipment prove-in section for three years. With Bell Laboratories he has been associated with guided-wave research in the areas of waveguide and repeaters.

J. A. MORRISON, B.Sc., 1952, King's College, University of London; Sc.M., 1954, and Ph.D., 1956, Brown University; Bell Telephone Laboratories, 1956—. Mr. Morrison has been doing research in a variety of problems in mathematical physics and applied mathematics. His recent interests have included perturbation techniques for nonlinear oscillations and propagation in random media. He was a visiting professor of mechanics at Lehigh University during the fall semester 1968. Member, American Mathematical Society, SIAM, Sigma Xi.

RAYMOND K. MUELLER, B.S., 1963, M.S., 1965, and D.Sc., 1967, Washington University; Bell Telephone Laboratories, 1967—. Mr. Mueller has worked in operations research and information theory. He is presently working on the calculation of the capacity of telephone cables. Member, Tau Beta Pi, Sigma Xi.

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.

IRWIN W. SANDBERG, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—. Mr. Sandberg has been concerned with analysis of military systems, synthesis and analysis of active and time-varying networks, studies of properties of nonlinear systems, and some problems in communication theory and numerical analysis. He is Head of the Systems Theory Research Department. Member, IEEE, Eta Kappa Nu, Sigma Xi, Tau Beta Pi.