THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XL JANUARY 1961 NUMBER 1

Proving Theorems by Pattern Recognition — II

By HAO WANG

(Manuscript received July 18, 1960)

Theoretical questions concerning the possibilities of proving theorems by machines are considered here from the viewpoint that emphasizes the underlying logic. A proof procedure for the predicate calculus is given that contains a few minor peculiar features. A fairly extensive discussion of the decision problem is given, including a partial solution of the $(x)(Ey)(z)$ satisfiability case, an alternative procedure for the $(x)(y)(Ez)$ case, and a rather detailed treatment of Skolem's case. In connection with the $(x)(Ey)(z)$ case, an amusing combinatorial problem is suggested in Section 4.1. Some simple mathematical examples are considered in Section VI.

Editor's Note. This is in form the second and concluding part of this paper' Part I having appeared in another journal.[1] However, an expansion of the author's original plan for Part II has made it a complete paper in its own right.

I. A SURVEY OF THE DECISION PROBLEM

1.1 *The Decision Problem and the Reduction Problem*

With regard to any formula of the predicate calculus, we are interested in knowing whether it is a theorem (the problem of provability), or equivalently, whether its negation has any model at all (the problem of satisfiability). Originally this decision problem was directed to the search for one finite procedure which is applicable to all formulae of the predicate calculus. Since it is known that there can be no such omnipotent procedure, the main problem is to devise procedures effective for classes of formulae which satisfy suitable conditions.

The complementary problem of reduction is to give effective procedures which reduce broader classes to narrower ones while preserving provability or satisfiability. In this way, a decision procedure for a smaller class can be made to apply to a larger one. Thus far, most work on the reduction problem has been directed to the special case of finding procedures which reduce all formulae of the predicate calculus to members of some special class (e.g., those in the Skolem normal form). Each such class is called a reduction class relative to satisfiability or provability according to whether satisfiability or provability is preserved by the transformations (Ref. 2, p. 32). It follows automatically that the corresponding decision problem for each reduction class is unsolvable.

The reduction classes and the procedures employed to obtain them are, being concerned with undecidable cases, only of indirect use for the problem of discovering positive results on the decision problem. More directly relevant are reduction procedures which are applicable when the reduced class is not a reduction class and may in particular be a decidable class. Some very preliminary results on this more general aspect of the reduction problem will be described in Section V.

For both the decision problem and the reduction problem, there is, beyond the "yes or no" as to satisfiability, a further question of determining all models and devising transformation procedures which preserve all models. Such questions have been studied to a certain extent (Ref. 3, p. 23), but will be disregarded in what follows.

It is customary to characterize reduction classes and decidable classes in terms of formulae in the prenex normal form, i.e., with all quantifiers at the beginning. Sometimes, with regard to satisfiability (or provability), conjunctions (or disjunctions) of formulae in the prenex normal form are considered. We shall call this the extended prenex form.

In Section V, a procedure will be given for reducing any formula to a finite set of generally simpler formulae in the extended prenex form such that the original formula is provable if and only if all formulae in the reduced set are. In this and the next few sections, we shall only be concerned with formulae in the extended prenex form. Furthermore, we shall give in Section V a proof-decision procedure for the quantifier-free logic, obtained from the propositional calculus by adding equality, function symbols and individual constants. Any theorem in it is called a quantifier-free tautology, as an extension of the notion of a propositional tautology. We shall make use of the fact that we can always decide whether a given formula is a quantifier-free tautology.

1.2 *A Brief Formulation of the Predicate Calculus*

1.2.1 *Primitive Symbols*

1.2.1.1 Variables x, y, z, etc. (an infinite set).
1.2.1.2 Individual constants (a finite or infinite set).
1.2.1.3 Propositional (Boolean) operations: $\sim$, $\vee$, &, $\supset$, $\equiv$.
1.2.1.4 Predicate letters (a finite or infinite set).
1.2.1.5 Function letters (a finite or infinite set).
1.2.1.6 Equality: $=$ (a special predicate symbol).
1.2.1.7 Quantification symbols: ( ), (E).
1.2.1.8 Parentheses.

1.2.2 *Inductive Definition of Terms and Formulae*

1.2.2.1 A variable or an individual constant is a term.

1.2.2.2 A function symbol followed by a suitable number of terms is a term.

1.2.2.3 A predicate followed by a suitable number of terms is a formula (and an atomic formula); in particular, if α, β are terms $=(\alpha,\beta)$ or $\alpha = \beta$ is a formula (and an atomic formula).

1.2.2.4 If φ, ψ are formulae and α is a variable, then $(\alpha)\varphi$, $(E\alpha)\varphi$, $\sim\varphi$, $\varphi \vee \psi$, $\varphi \mathbin{\&} \psi$, $\varphi \supset \psi$, $\varphi \equiv \psi$ are formulae.

1.2.3 *Inductive Definition of Theorems*

1.2.3.1 A quantifier-free tautology is a theorem.

1.2.3.2 If a disjunction D of n alternatives is a theorem, $\varphi\alpha$ is one of the alternatives and β is a variable, then:

(a) If α is a term, then the result of replacing $\varphi\alpha$ by $(E\beta)\varphi\beta$ in D is a theorem;

(b) if α is a variable free in $\varphi\alpha$ but not free in the other alternatives and β is α or does not occur in $\varphi\alpha$, then the result of replacing $\varphi\alpha$ by $(\beta)\varphi\beta$ in D is a theorem.

1.2.3.3 If $\varphi \vee \cdots \vee \varphi$ is theorem, so is also φ.

The above formulation is complete only with respect to formulae in the extended prenex form.

1.3 *The Fundamental Theorem of Logic*

The main purpose of the next few sections is to study the decision problem on the theoretical foundation of the fundamental theorem of logic, an approach initiated by Skolem[4] and Herbrand,[5] and recently revived by Church,[6,7] and by Klaua[8] and Dreben.[9,10]

Suppose $Mxyz$ is a quantifier-free matrix:

1.3.1 $$(x)(Ey)(z)Mxyz,$$

1.3.2 $$(Ex)(y)(Ez) \sim Mxyz.$$

Let now D_n be $M_1 \vee \cdots \vee M_n$ and M_i be $M1ii'$, i' being an abbreviation for $i + 1$. The fundamental theorem, when applied to 1.3.1, states:

1.3.3 The following three conditions are equivalent:
(a) 1.3.1 is a theorem of the predicate calculus; (b) for some n, D_n is a quantifier-free tautology; (c) 1.3.2 is not satisfiable.

If D_n is a quantifier-free tautology, then, by 1.2.3.1, both it and the result of substituting distinct variables for distinct numbers in it are theorems. For example, suppose the result is:

1.3.4 $$Maab \vee Mabc \vee Macd.$$

We have: by 1.2.3.2(b),

$$Maab \vee Mabc \vee (z)Macz;$$

by 1.2.3.2(a),

$$Maab \vee Mabc \vee (Ey)(z)Mayz.$$

Similarly,

$$\begin{gathered} Maab \vee (Ey)(z)Mayz \vee (Ey)(z)Mayz, \\ (Ey)(z)Mayz \vee (Ey)(z)Mayz \vee (Ey)(z)Mayz, \end{gathered}$$

by 1.2.3.3,

$$(Ey)(z)Mayz;$$

by 1.2.3.2(b),

$$(x)(Ey)(z)Mxyz.$$

Hence, condition (b) implies conditions (a) and (c) in 1.3.3.

On the other hand, if no D_n is a quantifier-free tautology, then there is, for each D_n, some interpretation of the function and predicate symbols on the set $\{1, \cdots, n'\}$ which satisfies $\sim D_n$. By a well-known argument, there is then an interpretation on the domain of all positive integers which satisfies $\sim D_1$, $\sim D_2$, etc. simultaneously. This, however, means that under the interpretation each finite segment of the infinite conjunction

1.3.5 $$\sim M112 \ \& \sim M123 \ \& \sim M134 \ \& \cdots$$

is true. But then there is an integer x, viz. 1, such that for every integer y, there is an integer z, viz. y', such that $\sim Mxyz$. In other words, 1.3.2, the negation of 1.3.1, is true under the interpretation. Hence, the negation of condition (b) implies the negations of conditions (a) and (c).

If we take 1.3.5 as a model of 1.3.2, it seems natural to regard y as an independent variable, z as a dependent variable and x as an initial variable (the limiting case of a dependent variable, a function of zero arguments). The general principle of constructing M_n from 1.3.1 may be summarized by saying that each initial variable gets a constant number, the independent variables taking on all possible positive integers as values and the dependent variables always taking on numbers not used before.

In the general case, we must consider a disjunction (for provability) or conjunction (for satisfiability) of formulae with arbitrary strings of quantifiers. Then we can again construct the related quantifier-free formulae in the same way, with the numbers in each clause proceeding independently.

Thus, if we wish to study the satisfiability problem, we consider any formula of the form:

1.3.6 $$\varphi_1 \ \& \cdots \& \ \varphi_n \qquad (n \geqq 1),$$

where each φ_i is of the form, with $d_1 \geqq 0$, $e_c \geqq 0$, $c \geqq 1$, e_1, d_2, e_2, $\cdots$, $d_c \geqq 1$:

1.3.7 $$(Ey_1^1) \cdots (Ey_{d_1}^1)(x_1^1) \cdots (x_{e_1}^1) \cdots (Ey_1^c) \cdots (Ey_{d_c}^c)(x_1^c) \cdots (x_{e_c}^c) My_1^1 \cdots x_{e_c}^c.$$

One familiar way of obtaining M_1, M_2, etc. for the formula 1.3.7 begins by replacing the dependent variables (those with the letter y) each with a function (sometimes called a "Skolem function") of all the preceding independent variables (those with the letter x), and then dropping all the quantifiers. Let the result be M^*. In particular, the initial (dependent) variables are replaced by distinct constants which may be viewed as trivial functions. Suppose $e_1 + \cdots + e_c = p$, $d_1 + \cdots + d_c = q$ in 1.3.7.

The Skolem functions are any functions $g_1, \cdots, g_q$ which, taken together, satisfy the following conditions:

1.3.8 (a) For each g_i, $g_i(u_1, \cdots, u_m) \neq u_j$, $j = 1, \cdots, m$, $i = 1, \cdots, q$.

(b) For each g_i, $g_i(u_1, \cdots, u_m) = g_i(v_1, \cdots, v_m)$ only when $u_1 = v_1, \cdots, u_m = v_m$.

(c) For any g_i, g_j, $i \neq j$, $g_i(u_1, \cdots, u_m) \neq g_j(v_1, \cdots, v_n)$, for all $u_1, \cdots, u_m, v_1, \cdots, v_n$.

Then we can take the smallest domain which contains the constants for the initial (dependent) variables (or an arbitrary constant when there is no such initial variable) and is closed with respect to the Skolem functions. Once such an (enumerable) domain is available, we can somehow enumerate all the p-tuples of members of the domain. Then, for each i, M_i is simply the result obtained from M^* when the independent variables are replaced respectively by members of the ith p-tuple.

The satisfiability problem of 1.3.7 is then reduced to that of the infinite conjunction:

1.3.9
$$M_1 \ \& \ M_2 \ \& \ \cdots$$

Similarly, the satisfiability problem of 1.3.6 can be handled by reducing each φ_i separately and then taking the conjunction of the n infinite conjunctions of the form 1.3.9.

It is customary to use the positive integers as the domain, fix some enumeration of the p-tuples, and specify the Skolem functions in a natural manner. One familiar enumeration of the p-tuples is the following:

1.3.10 $(a_1, \cdots, a_p)$ precedes $(b_1, \cdots, b_p)$. if either

(a) they are permutations of each other but $(a_1, \cdots, a_p)$ precedes $(b_1, \cdots, b_p)$ in the lexicographic order; or

(b) $\max(a_1, \cdots, a_p) = \max(b_1, \cdots, b_p)$, $\Sigma a_i = \Sigma b_i$, but $(a_1, \cdots, a_p)$, rearranged according to nondecreasing magnitude, precedes $(b_1, \cdots, b_p)$, similarly rearranged, in the lexicographic order; or

(c) $\max(a_1, \cdots, a_p) = \max(b_1, \cdots, b_p)$, but $\Sigma a_i < \Sigma b_i$; or

(d) $\max(a_1, \cdots, a_p) < \max(b_1, \cdots, b_p)$.

The Skolem functions are usually chosen by going through the infinite conjunction 1.3.9 from left to right and using each time the smallest unused integer for the next functional expression not yet evaluated. Thus, e.g., $y_1{}^1, \cdots, y_{d_1}{}^1$ in 1.3.7 get the constant values $1, \cdots, d_1$, and M_1 is:

$$M1 \cdots d_1^11 \cdots 1d_1' \cdots (d_1 + d_2) \cdots (q - d_c + 1) \cdots q1 \cdots 1.$$

Each time a functional expression gets a value, the value is substituted in all later occurrences of the same expression.

In this way we arrive at a form of the fundamental theorem of logic as a generalization of 1.3.3.

It is natural to observe that the infinite conjunction 1.3.9 can be divided into sections (Ref. 4, p. 138):

1.3.11 The first section is the set of those M_i's in which the p-tuples replacing the independent variables are made up of integers in the set $\{1, \cdots, d_1\}$, or the set $\{1\}$ if $d_1 = 0$; the $(n+1)$th section is the set of those M_i's not belonging to the nth section in which the p-tuples are made up of integers which occur in the union of the first n sections.

This notion has been used by Skolem in explaining some decision procedures (see Section II below).

1.4 *Special Cases of the Decision Problem*

The principal known decidable classes are, with regard to satisfiability the following:

I. *The monadic case.* The class of all formulae which contain only monadic predicate letters and no function symbols.

II. *The EA satisfiability case (the AE provability case).* The class of all formulae in the prenex form with prefixes of the form $(Ey_1) \cdots (Ey_m)(x_1) \cdots (x_n)$, $m, n \geqq 0$, and no function symbols [or the form $(y_1) \cdots (y_m)(Ex_1) \cdots (Ex_n)$ for provability].

III. *The conjunctive satisfiability case.* Every formula in the prenex form with a matrix which is a conjunction of atomic formulae and their negations. (Equivalently, the disjunctive provability case.)

IV. *The Skolem case.* Every formula in the prenex form with no function symbols such that it has a prefix ending with $(Ey_1) \cdots (Ey_n)$, $n > 0$, and every atomic formula occurring in the matrix contains either one of the variables $y_1, \cdots, y_n$, or all the independent variables. [For provability, $(y_1) \cdots (y_n)$ at the end.]

V. *The EA_2E satisfiability case (the AE_2A provability case).* Every formula containing no function symbols in the prenex form with a prefix $(Ey_1^1) \cdots (Ey_m^1)(x_1)(x_2)(Ey_1^2) \cdots (Ey_n^2)$.

VI. *The Ackermann case.* For satisfiability, every formula which contains no function symbols, no equality sign, only a single dyadic predicate (G say), and has the form $(x)(Ey)Gxy$ & $(x_1) \cdots (x_m)Mx_1 \cdots x_m$, $m \leqq 4$, M quantifier-free.

In addition to these, two other cases may be mentioned:

VII. *The $A_1E_1A_1$ satisfiability case.* Every formula with the prefix $(x_1)(Ey)(x_2)$ and with no function symbols.

VIII. *The Surányi normal form case.* For satisfiability, every formula which has no equality sign, no function symbols, only dyadic predicate symbols, and has the form $(x_1)(x_2)(x_3)Mx_1x_2x_3$ & $(x_1)(x_2)(Ey_3)Nx_1x_2y_3$, M, N quantifier-free.

It may be noted that in all the cases, with the single exception of III, no function symbols are permitted. Indeed, very little is known about the decision problem of formulae containing function symbols (compare Ref. 3, pp. 98–107). Unless otherwise stated, we shall always assume that no function symbols occur.

In what follows, cases I and VI will not be considered. So far as the monadic case without equality (a subcase of I) is concerned, it is possible to obtain a decision procedure from one for case II. Some of the problems suggested by the Ackermann case are also encountered by the $A_1E_1A_1$ case, while other implications of this case seem to call for a closer examination of certain arithmetic predicates.

Formulae under case VIII form a reduction class in the sense that there is an effective procedure by which every formula, possibly containing = and function symbols, can be reduced to one in the class with satisfiability preserved (Ref. 2, p. 60). It follows that there exists no decision procedure for this case. It is, however, desirable to find some "semidecision procedure" for the class which is a decision procedure for some subclass of it that is not specified explicitly in advance. It is thought that such semidecision procedures are a useful way of extending the range of formulae decidable by a predetermined finite set of procedures. A brief discussion is included in Section IV to point to the sort of thing which can be done along this line. It should be of interest to design semidecision procedures for case VIII, as well as for other reduction classes.

The case VII is perhaps the best known unsettled case; it has been mentioned in various connections (see, e.g., Ref. 11, p. 576 and Ref. 12, p. 420). In Section IV a procedure will be given which may be a decision procedure for the whole case but has only been shown to terminate for certain special cases. A proof of finiteness of the procedure is wanting. It is thought that, incomplete as the solution is, it is quite suggestive for further works on the decision problem. Some rather amusing combinatorial problems are also related to the considerations on this case.

An alternative decision procedure for the much-studied case V will be given in Section III in the equivalent form A_2E (for satisfiability).

The Skolem case will be examined in considerable detail in Section II, using ideas proposed by Skolem[4] (p. 138) and Church[6] (p. 264). Remarks relevant to machine realizations of the procedure will also be included.

The Skolem case includes the following special cases:

IVa. The A_1E satisfiability case. Because every atomic formula has to include some variable and there is only one independent variable.

IVb. For satisfiability, every formula whose prefix ends with $(Ey_1) \cdots (Ey_n)$, and in which every atomic formula contains at least one of the variables $y_1, \cdots, y_n$.

IVc. For satisfiability, every formula whose prefix is

$$(Ey_1^1) \cdots (Ey_m^1)(x_1) \cdots (x_n)(Ey_1^2) \cdots (Ey_k^2)$$

and in which every atomic formula contains either all of $x_1, \cdots, x_n$ or at least one of $y_1^2, \cdots, y_k^2$.

IVd. For satisfiability, every formula in the Skolem normal form, i.e., with prefix $(x_1) \cdots (x_m)(Ey_1) \cdots (Ey_n)$, such that every atomic formula contains at least m distinct variables.

For the extensive literature on the decision problem, the reader is referred to the bibliographies in Refs. 2 and 3. The writer has not been able to study carefully much of the relevant literature, and is not certain that the procedures described in Sections II and III may not turn out to be inferior to existing ones. Recently, the writer noticed that ideas along the line of the solution of the E_1A provability case given in Section 3 of Part I[1] are contained in Skolem's writings (e.g., Ref. 4, p. 135).

Of the two remaining cases, II and III, some brief comments will suffice.

1.5 *Two Simple Cases*

The EA satisfiability case II has agreeable decision procedures not dependent on the fundamental theorem of logic (see Ref. 13, p. 13). It is also easy to devise a decision procedure on the basis of the fundamental theorem. Consider

1.5.1 $$(Ey_1) \cdots (Ey_m)(x_1) \cdots (x_n)My_1 \cdots x_n .$$

This is in fact equivalent to:

1.5.2 $$M_1 \ \& \ \cdots \ \& \ M_k, \qquad k = m^n, \qquad \text{or 1 when } m = 0.$$

In fact, this is a limiting case of the fundamental theorem because no Skolem functions are needed, so that the m constants for the initial variables are all we need for fabricating a model. In other words, either the negation of 1.5.2 is a quantifier-free tautology, and the negation of 1.5.1 is a theorem; or 1.5.2 has a model, and 1.5.1 has a model too. The presence of the equal sign is permitted, but the presence of function symbols in 1.5.1 would invalidate the procedure.

The conjunctive satisfiability case III was originally solved by Herbrand (Ref. 5, pp. 44–45). Suppose the matrix is:

1.5.3 $$A_1 \; \& \; \cdots \; \& \; A_m \; \& \sim B_1 \; \& \; \cdots \; \& \sim B_n ,$$

or, in a different notation:

1.5.4 $$A_1 , \cdots , A_m \nrightarrow B_1 , \cdots , B_n .$$

Assume first that neither equality nor function symbols occur. If no predicate letter occurs both on the left side and on the right side, then we can simply choose to make all predicates occurring on the left side true of all numbers and those on the right false for all numbers, and then the infinite conjunction corresponding to the given formula is true under the interpretation.

Whenever there is one clause on the left and one on the right which contain the same predicate letter, e.g., A_i is *Gabc* and B_j is *Guvw*, we compare them and ask whether it is possible to assign the same integers to their arguments in some M_s and M_t respectively. If the answer is yes, the original formula can have no model, because the infinite conjunction must be always false. If the answer is no for every such pair, then the original formula has a model.

To compare A_i and B_j , we examine the three pairs of corresponding variables. If both variables in some pair are distinct dependent variables, then the two clauses A_i and B_j can never get the same numbers. When this is the case for none of the pairs, we can decide the question by asking whether there are positive integers s, t such that $a(s) = u(t)$, $b(s) = v(t)$ and $c(s) = w(t)$, where, for each variable α in the original formula, $\alpha(n)$ is a function giving the number which replaces α in M_n . It is possible to give a scheme to generate such function for each given formula. When there are solutions for some pair of clauses, the original formula is not satisfiable.

If the formula 1.5.4 contains function symbols but not $=$, then the comparison of A_i and B_j has to take functions into considerations sometimes. We may have to ask whether $f(a(s)) = g(u(t))$, instead of $a(s) = u(t)$, has a solution. In such cases, there is a solution only when f and g are the same function, because otherwise we can always give different values to $f(a(s))$ and $g(u(t))$ to avoid the incompatibility of M_s and M_t .

When the equals sign also occurs, we have to list all the equations among $A_1 , \cdots , A_m$, if there is any, and complete the list by using transitivity. If there are none, we need only to proceed as before, except that we can also reject satisfiability on the ground of, e.g., having an equation $u = v$ among $B_1 , \cdots , B_n$, and $u(p) = v(p)$ has a solution in

p. In the general case, we must compare A_i and B_j, which have the same predicate letter, in a more complicated manner. One way to do this is to give an effective survey of all the equalities obtainable in $M_1, \cdots, M_t$, for every t. And then the question of comparing $Gabc$ and $Guvw$ is reduced to the following: whether there are p, q, t such that, with the help of the equalities obtainable from $M_1, \cdots, M_t$, we have $a(p) = u(q)$, $b(p) = v(q)$, $c(p) = w(q)$. Since these considerations are only subsidiary for the main purpose of the paper, details for this and other steps sketched above will not be supplied.

II. THE SKOLEM CASE

2.1 *Outline of a General Method*

The subcase IVb, where every atomic formula contains at least one of the last string of dependent variables, is particularly simple. Thus, in every M_k, each such variable always gets replaced by some new number so that no atomic formula in M_k can have occurred in any of $M_1, \cdots, M_{k-1}$. Hence, a formula of such a form is satisfiable if and only if $\sim M_1$ is not a quantifier-free tautology.

In the general Skolem case, we make use of the definition of sections given above in 1.3.11. Let $(a_1^k, \cdots, a_p^k)$ be the p-tuple which replaces the dependent variables in M to get M_k.

Given any member M_i of the nth section, the only related instances in the nth section are those M_k for which $(a_1^k, \cdots, a_p^k)$ is a permutation of $(a_1^i, \cdots, a_p^i)$, and the only related instances in the $(n + 1)$th section are those M_j for which $(a_1^j, \cdots, a_p^j)$ include only numbers occurring in M_i and at least one number not in the set $\{a_1^i, \cdots, a_p^i\}$.

Hence, it is possible to get a decision procedure by determining whether there exists any set of possibilities which includes models for the instances of the first section, as well as models for all related instances M_k and M_j for every model for M_i in the set.

When the formula is in the Skolem normal form or the form of IVc, somewhat more is true:

2.1.1 If M_j belongs to the $(n + 1)$th section, then it can have common atomic formulae with only at most one M_i in the nth section.

This is so because each atomic formula in M_j either contains a new number not occurring in any member of the nth section, or otherwise contains all of $\{a_1^j, \cdots, a_p^j\}$ with at least one number (say a_t^j) which appeared for the first time in one specific member (say M_i) of the nth section. In the first case the atomic formula in M_j does not occur in any member of the *n*th section. In the second case, M_j can contain no common atomic formula with any member of the *n*th section except possibly M_i, since a_t^j does not occur in any of the other members of the *n*th section.

Detailed considerations will be confined to the treatment of a simple special case.

2.2 *An Explicit Procedure for a Special Case*

We consider a very simple special case in which the matrix contains no equals sign (and of course no function symbols), and a single dyadic predicate G:

2.2.1
$$(x)(y)(Ez)Mxyz.$$

As an illustration, we use the negation of Example (2) of Part I:[1]

2.2.2
$$(x)(y)(Ez)[(Gxy \;\&\; Gyx \;\&\; {\sim}Gxz \;\&\; {\sim}Gzy \;\&\; {\sim}Gzz) \\ \vee \;(Gxz \;\&\; Gzy \;\&\; Gzz \;\&\; {\sim}Gxy \;\&\; {\sim}Gyx)].$$

In an alternative notation, the matrix is:

2.2.3
$$Gxy,Gyx \nrightarrow Gxz,Gzy,Gzz; \\ Gxz,Gzy,Gzz \nrightarrow Gxy,Gyx.$$

We construct a truth table of all the possibilities which can satisfy the above matrix:

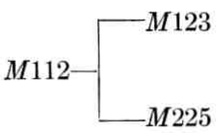
2.2.4

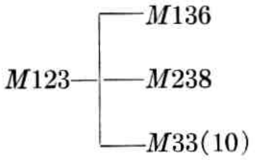
Gxy	*Gyx*	*Gxz*	*Gzx*	*Gyz*	*Gzy*	*Gzz*
t	t	f			f	f
f	f	t			t	t

The blanks may take either t or f as values. Hence, there are eight rows in all.

For the prefix $(x)(y)(Ez)$, the numbers to substitute for (x,y,z) in M_1, M_2, M_3, M_4, etc., are (1,1,2), (1,2,3), (2,1,4), (2,2,5), etc. In order to decide whether a formula of the form 2.2.1 has a model, we ask whether it is possible to make $M112$, $M123$, $M214$, etc., simultaneously true, or, in other words, whether we can find for each M_i one row from the above table according to which M_i is true, such that these infinitely many rows are all compatible in the sense that the same atomic formula always gets the same truth value (t or f).

Among the number triples we can distinguish two classes, those in which x and y get the same numbers, such as (1,1,2), and those in which they get different numbers, such as (2,1,4). The conditions under which a model is possible are roughly: (i) to satisfy $Maab$, a row in the truth table has to behave in a way that x and y are interchangeable; (ii) for each row satisfying $Mabc$, there must be a related row satisfying $Mbac$; (iii) for the two types of row, two corresponding patterns of continuation must be possible, e.g.,





These conditions can be formalized more exactly and applied, in particular, to show that 2.2.2 has a model, and therefore its negation is not a theorem. For this purpose, we assume a formula of the form 2.2.1 for which a truth table T like 2.2.4 is constructed. When, for example, Gxy in a row R of T gets the same value as Gzz in a row S of T, we shall use the brief notation $R_{xy} = S_{zz}$.

2.2.5 A row S in the table T is a uniform row if $S_{xy} = S_{yx}$, $S_{xz} = S_{yz}$, $S_{zx} = S_{zy}$.

Clearly, for a row to satisfy $M112$, it is necessary that it be uniform. If there is no uniform row, then there is no model for the original formula.

2.2.6 A row S in the table T is an heir of a row R in T if S is a uniform row and $R_{zz} = S_{xy}$.

2.2.7 A row in T is trivial if it has no heir.

Since a row having no heir cannot be continued, we may cross out all trivial rows and be concerned only with nontrivial rows. This is not theoretically necessary because further requirements would cross out trivial rows anyhow, but it makes for efficiency.

2.2.8 A row R in the table T is an ordinary row if there is a row S such that $R_{xy} = S_{yx}$, $R_{yx} = S_{xy}$, $R_{xz} = S_{yz}$, $R_{zx} = S_{zy}$, $R_{yz} = S_{xz}$, $R_{zy} = S_{zx}$. R and S are said to be mates of each other.

This is the condition under which R and S can satisfy ($M123$, $M214$) or ($M214$, $M123$) respectively.

In the table 2.2.4 for the formula 2.2.2, it is easily verified that only the two following rows are uniform rows or ordinary rows:

	Gxy	Gyx	Gxz	Gzx	Gyz	Gzy	Gzz
α	t	t	f	f	f	f	f
β	f	f	t	t	t	t	t

In fact, α and β are the only uniform rows, as well as the only ordinary rows. Each of α and β is only a mate of itself.

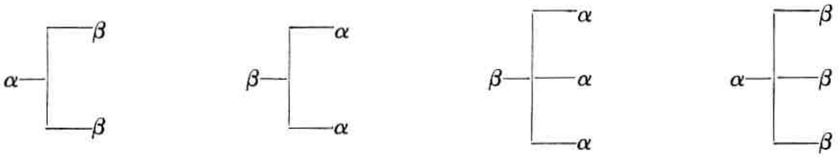
2.2.9 A uniform row R is permanent if (i) it has an heir which is permanent, and (ii) there is a permanent ordinary row S such that $R_{yz} = S_{xy}$, $R_{zy} = S_{yx}$. S is said to be a subordinate of R.

2.2.10 An ordinary row R is permanent if (i) it has an heir which is a permanent (uniform) row, (ii) it has a mate that is a permanent ordinary row, and (iii) there are two permanent ordinary rows P and S such that $R_{xz} = P_{xy}$, $R_{zx} = P_{yx}$, $R_{yz} = S_{xy}$, $R_{zy} = S_{yx}$. P and S are said to be a pair of subordinates of R.

The two definitions 2.2.9 and 2.2.10 embody a simultaneous recursion. Condition (ii) in 2.2.9 is necessary, if, e.g., R is to satisfy $M112$ and S is to satisfy $M123$. Condition (iii) in 2.2.10 is necessary if, e.g., R is to satisfy $M123$, P is to satisfy $M136$ and S is to satisfy $M238$.

2.2.11 The formula 2.2.1 has a model if and only if its truth table T contains a permanent uniform row.

This assertion will be justified in 2.3. We observe first that both α and β are permanent uniform rows for the example 2.2.2. In fact, we have various models for the formula, which are determined, in outline, by the following patterns of continuation:



More exactly, choose, e.g., α as a model of $M112$. As a continuation of this, β satisfies $M123$ and $M225$; since β is its own mate in the sense of 2.2.8, β also satisfies $M214$. Similarly, since α is its own mate, as a continuation of β satisfying $M123$, α satisfies $M136$, $M317$, $M238$, $M329$, and $M33(10)$. In this particular case, the model β of $M214$ can be continued in the same way. Moreover, the model β of $M225$ can be continued by the row α, and, e.g., the model α of $M136$ can be continued by the row β, and so on.

In the general case, a symmetry argument is needed to show that if a model of, e.g., $M123$ can be continued, then a model of $M214$ can also be continued. For example, if (R, S) satisfy $(M123, M214)$ respectively, and (A, B, C, D) satisfy respectively the continuation $(M136, M317, M238, M329)$ of $M123$, then it is easy to see that (B, A, D, C) satisfy the corresponding extension of $M214$. This means that condition (ii) of 2.2.10 can be weakened to require a mate that is an ordinary row with a permanent heir.

The decision procedure implicit in the above definitions may be described explicitly thus:

2.2.12 The decision procedure:

1. Construct a truth table T.
2. Find all uniform rows.
3. Cross out all trivial rows.

Let U_0 be the set of remaining uniform rows, V_0 be the set of remaining ordinary rows. Each time, assume U_n and V_n are given and continue the following four steps:

4. Eliminate every uniform row from U_n which has no subordinate row in V_n, thus obtaining U_{n+1} from U_n and V_n.

5. Eliminate from V_n every ordinary row which has no mate or no pair of subordinate rows in V_n, thus obtaining V_{n+1} from V_n.

6. Eliminate every uniform row from U_{n+1} which has no heir in U_{n+1}, thus obtaining U_{n+2} from U_{n+1}.

7. Eliminate every ordinary row from V_{n+1} which has no heir in U_{n+2}, thus obtaining V_{n+2} from V_{n+1} and U_{n+2}.

8. The steps 4 through 7 are repeated until one of two things happens: either at some stage we obtain an empty U_i and an empty V_i, then we stop and conclude that the original formula 2.2.1 has no model; or else, after a whole round of the steps 4 and 7, we find U_{n+2} and V_{n+2} remain the same as U_n and V_n, then we stop and conclude that the original formula 2.2.1 has a model.

In practice, it is more efficient to perform, if possible, each of the steps 4 through 8 repeatedly, before going to the next step.

The procedure is clearly finite, since U_0 and V_0 are finite, and each round of steps 4 through 8 must reduce the size of U_n or V_n if the procedure has not come to a stop yet. Moreover, the final sets U_i and V_i must be both empty or both nonempty.

2.3 *Justification of the Procedure*

As a Skolem case, the formula 2.2.1 must not contain Gxx and Gyy. It is, however, not obvious that we are justified in not including two columns Gxx and Gyy in the truth tables such as 2.2.4. For a model constructed on the basis of such reduced tables, it is not evident that, for some positive integer a, Gaa might not be compelled to take on the value t at one place, and the value f at another. However, we can prove the following:

2.3.1 In every model obtained on the basis of a truth table not including columns for Gxx and Gyy, for every number a, Gaa is never compelled to take on two different values.

Take, for example, $G22$. If Gzz occurs in the original formula, $G22$ is compelled to take a fixed value in a model with a row R for $M112$. In the same model, if S is the row for $M225$, then $R_{zz} = S_{xy} = S_{yx}$. Hence, it is harmless that S_{xx} and S_{yy} are compelled to take the same value as both R_{zz} and S_{xy} (or S_{yx}). In all other cases, the values for $G22$ can always be given the value of R_{zz} because there is no other place where $G22$ is independently compelled to take a certain truth value.

For the same reason, if neither an atomic formula nor any one obtainable from it by permuting the variables occurs, we may leave out the columns for them. For example, if Gzz does not occur, we can leave it out. If neither Gxy nor Gyx occurs, we can leave both of them out.

On the other hand, if, e.g., Gxy and Gzy occur but Gyx does not, we still must include a column for Gyx. Otherwise, since we do not record the value of Gyx, it may happen that R satisfies $M112$, with R_{zy} = t, and S satisfies $M214$ with S_{xy} = f. Then no row P can satisfy $M123$, because P_{yx} is compelled to take both the value t and the value f, and this is not recorded without a column for Gyx.

To prove 2.2.11, we remark first that there are three types of instances illustrated by $M112$, $M123$, $M214$. For the first kind, an M_i of the form $Maab$, the only M_j, $j > i$, which have common atomic formulae with M_i are $Mbbc$, $Mabd$, $Mbae$, because these are the only ways in which both the independent variables x and y can be replaced by numbers occurring in M_i, and having only one of the two arguments from M_i yields no common atomic formula. Similarly, if M_i is $Mabc$, $a < b$, there are only five M_j, $j > i$ which have common atomic formula with M_i. By the symmetry argument preceding 2.2.12, the mate $Mbae$ is also taken care of.

Hence, if there is any permanent uniform row, we can find a model for all instances M_1, M_2, etc., such that each has some common atomic formula with an earlier one, or, in other words, all those occurring on an infinite tree beginning at M_1. This does not exhaust all the instances. For example, M_{14} and M_{15} [i.e., $M34(15)$ and $M43(16)$] are not included. Since, however, they contain no common atomic formulae with the instances already interpreted, we can take two permanent ordinary rows which are mates and get a model for another sequence of instances. In this way, it is seen that, if there is a permanent uniform row in the table T, then one can so interpret the predicate G in the domain of the positive integers that the whole sequence M_1, M_2, etc., are simultaneously satisfied.

The converse is quite obvious. If there is no permanent uniform row, then no interpretation of $M112$ can be continued indefinitely, and there is an i. such that M_1 & $\cdots$ & M_i is true under no interpretation.

2.4 *Questions of Efficiency*

When doing an example by hand, there are shortcuts we find natural to use. These may be viewed as more refined methods which can be mechanized by additional efforts. We give some informal illustration of the type of quick method we tend to use.

Consider the negation of Example (3) given in Part I:[1]

2.4.1 $$(x)(y)(Ez)\{[Gxy \;\&\; (\sim Gyz \vee \sim Gzz)] \vee [(Gxy \;\&\; Hxy) \;\&\; (\sim Hxz \vee \sim Hzz)]\}.$$

In the alternative notation, the matrix of the above formula is:

2.4.2 $$Gxy \nrightarrow Gyz;\; Gxy \nrightarrow Gzz;\; Gxy, Hxy \nrightarrow Hxz;\; Gxy, Hxy \nrightarrow Hzz.$$

The truth table for this is:

2.4.3	Gxy	Hxy	Gyx	Hyx	Hxz	Hzx	Gyz	Gzy	Gzz	Hzz
α	t						f			
β	t								f	
γ	t	t			f					
δ	t	t								f

Although the formula contains two predicates instead of just one, it is easy to see that the procedure described above can be extended to cover the case in a very straight-forward manner.

Since there are many blanks in the table, it is essential for efficiency that we do not expand the table by filling in the blanks (there would be 2^{24} rows), until we are compelled to do so. In other words, we try to carry out the decision procedure by treating each row containing blanks as a single row and make expansion only when we are not able to eliminate them as single rows.

We observe that for every row, in particular, every uniform row, Gxy gets the value t. It follows that row β, or more exactly, all the 2^7 rows obtainable from β are trivial by 2.2.7, since an heir of β must have Gxy take the value of Gzz in β, which is f. Hence we may delete row β altogether.

In order that row α, or any specification R of α, be permanent (uniform or ordinary), it is necessary, by 2.2.9 and 2.2.10, that there is a subordinate row S, such that Gxy gets the same value in S as Gyz in R, or $R_{Gyz} = S_{Gxy}$. But this is impossible because R_{Gyz} is f in every row obtainable from α, but S_{Gxy} is t in every row. Hence, we can delete row α altogether, and be concerned only with the rows γ and δ.

Since Hxy gets t in all the remaining rows and Hzz gets the value f in δ, every row obtainable from δ has no heir, and the whole row δ can be deleted.

However, no permanent ordinary row can be obtained from γ alone because, by 2.2.10, for any such row R there must be a subordinate row P such that $R_{Hxz} = P_{Hxy}$, but in row γ, Hxz is always f and Hxy is always t. Hence, there can also be no permanent uniform row, and, by 2.2.11, the formula 2.4.1 has no model. Therefore, Example (3) in Part I,[1] the negation of 2.4.1, is a theorem.

Another method of deciding 2.4.1 is the following. We begin with M_1, which is a disjunction of conjunctions, and choose M_i, M_j, etc., which contain common atomic formula with M_1, in the hope that M_1 & M_i & M_j & $\cdots$ as multiplied out into a disjunction of conjunctions will include in each conjunction some atomic formula and its negation. The process may have to be continued.

As we observed before, only M_2, M_3, M_4 can have common atomic formulae with M_1. Of these three, on account of the special structure of 2.4.1, M_3 has no common part with M_1. Hence, we need to consider, to begin with, only M_1, M_2, M_4:

	(i)	(ii)	(iii)	(iv)
$M112$	$G11 \nrightarrow G12$;	$G11 \nrightarrow G22$;	$G11, H11 \nrightarrow H12$;	$G11, H11 \nrightarrow H22$
$M123$	$G12 \nrightarrow G23$;	$G12 \nrightarrow G33$;	$G12, H12 \nrightarrow H13$;	$G12, H12 \nrightarrow H33$
$M225$	$G22 \nrightarrow G25$;	$G22 \nrightarrow G55$;	$G22, H22 \nrightarrow H25$;	$G22, H22 \nrightarrow H55$

By the row for $M123$, (i) of $M112$ can be deleted because (i) contains $\sim G12$ (i.e., after $\nrightarrow$), while each clause in the row for $M123$ contains $G12$. It can be seen then that every row in column (i) can be deleted in the same way. Similarly, (ii) of the row for $M112$ can be deleted because it contains $\sim G22$, while each clause in the row for $M225$ contains $G22$; therefore, the whole column (ii) can be deleted eventually, and we need only consider the columns (iii) and (iv). But then (iii) of the row for $M112$ can also be deleted because it contains $\sim H12$, and all the remaining columns of the row for $M123$ contain $H12$. Finally, we have only column (iv) left. Now, however $\sim H22$ occurs in the row for $M112$ and $H22$ occurs in the row for $M225$. Hence, the conjunction of the three rows of column (iv) is a contradiction, and 2.4.1 has no model.

2.5. *The Inclusion of Equality*

The decision procedure in 2.2 can be extended to deal with cases where the equal sign occurs in the given formula:

2.5.1 $(x)(y)(Ez)Mxyz$, with $=$ occurring.

Additional considerations are needed to take care of the special properties of $=$. First we bring $Mxyz$ into a disjunction of conjunctions of atomic formulae and their negations, in the usual manner. Then we modify the resulting matrix to take care of the properties of $=$. (a) Each conjunction that contains an inequality of the form $v \neq v$, v being x or y or z, is deleted. (b) In each conjunction, a clause of the form $v = v$ is deleted. (c) Within each conjunction, if $u = v$ is a clause with distinct variables u and v, we add also, as new clauses (if not occurring already), $v = u$ and the result of replacing any number of occurrences of u by v (or v by u) in each clause of the conjunction; this is repeated for every equality until no new clause is generated. (d) Repeat the steps (a) and (b) on the result obtained by step (c); in addition, any conjunction which contains both an atomic formula and its negation is deleted.

We now construct the truth table on the basis of the new matrix (in a disjunctive normal form). Uniform rows, ordinary rows and permanence can be defined in a similar manner as before, except that a uniform row has to satisfy the additional condition that $x = y$ and $y = x$ both get the truth value t (not only that they just get a same value). In this way, we can obtain a decision procedure for all formulae of the form 2.5.1.

It is believed that the same type of consideration can be used to extend all the cases considered in this paper to include also the equal sign. In the next two sections, equality will be left out and attention will be confined to formulae not containing the equals sign (nor, of course, function symbols).

III. THE A_2E SATISFIABILITY CASE

We give an alternative treatment of this case which, it is conjectured, is in general more efficient than the method of Schütte[11] as reformulated by Klaua.[8] The method will be explained with the special case when only one dependent variable and only one dyadic predicate G occur:

3.1 $(x)(y)(Ez)Mxyz.$

The main difference between this case and the case solved in 2.2 above is that Gxx and Gyy are permitted to occur in $Mxyz$. As a result, for example, $M123$ may contain common atomic formula with any $Mabc$ in which a or b is one of 1, 2, 3.

As an example, we choose arbitrarily the following:

3.2 $(x)(y)(Ez)[\sim Gxx \;\&\; (Gxy \supset \sim Gyx) \;\&\; Gxz \;\&\; (Gzy \supset Gxy)].$

The matrix may be rewritten as:

3.3
$$Gxz \nrightarrow Gxx, Gxy, Gzy;\ Gxz, Gxy \nrightarrow Gxx, Gyx;$$
$$Gxz \nrightarrow Gxx, Gyz, Gzy.$$

The truth table is:

3.4

Gxx	Gxy	Gyx	Gyy	Gxz	Gyz	Gzx	Gzy	Gzz
f	f			t			f	
f		f		t			f	
f	t	f		t				

The problem is, as before, to decide whether there is a model that satisfies M_1, M_2, etc., simultaneously. The conditions are rather similar to those in 2.2 except that for any two rows R and S which, say, satisfy $Mabc$ and $Mdef$ in a model, there must be two rows which satisfy $Mcfg$ and $Mfch$ in the model. There is also a related requirement for a row satisfying M_1, because the number 1 is never used to replace a dependent variable. The various conditions may be stated:

3.5 A row R is uniform if $R_{xx} = R_{xy} = R_{yx} = R_{yy}$, $R_{xz} = R_{yz}$, $R_{zx} = R_{zy}$.

3.6 A row S is an heir of a row R if S is uniform and $R_{zz} = S_{xx}$.

3.7 Two rows R and S form a parallel pair if $R_{xx} = S_{yy}$, $R_{xy} = S_{yx}$, $R_{yx} = S_{xy}$, $R_{yy} = S_{xx}$, $R_{xz} = S_{yz}$, $R_{yz} = S_{xz}$, $R_{zx} = S_{zy}$, $R_{zy} = S_{zx}$.

Two rows of a parallel pair are said to be mates of each other.

If R and S are to satisfy $Mabc$ and $Mbae$, it is necessary that they form a parallel pair. In general, for a row satisfying $Mabc$, there must also be two parallel pairs of related rows satisfying $Macd$, $Mcae$, $Mbcf$, $Mcbg$. When $a = b$, the two parallel pairs become one. This, plus the requirement that every row in a model must have an heir may be summarized in the following condition.

3.8 A row R is normal if the following conditions are all satisfied:

3.8.1 It has a normal row as a mate;

3.8.2 It has an heir which is a normal row;

3.8.3 There are two normal rows P and S such that $R_{xx} = P_{xx}$, $R_{xz} = P_{xy}$, $R_{zx} = P_{yx}$, $R_{zz} = P_{yy}$, and $R_{yy} = S_{xx}$, $R_{yz} = S_{xy}$, $R_{zy} = S_{yx}$, $R_{zz} = S_{yy}$. Such rows P and S are said to be subordinates of R.

A uniform row is its own mate, although a self-mated row is not always a uniform row. For a uniform row, 3.8.1 is a redundant condition, and P and S coincide in 3.8.3. The definition 3.8 of normality is clearly recursive.

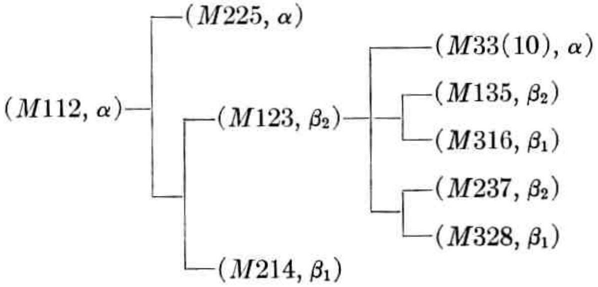
In the table 3.4, we observe that, because Gxx always takes the value f, Gzz can only take the value f in order that the row has an heir. Moreover, since Gxx always gets the value f and Gxz always gets the value t, in order that a row has a mate, Gyy must always take the value f and Gyz always t. Hence, we need consider only the following eight rows which result from filling the remaining gaps:

3.9

	Gxx	Gxy	Gyx	Gyy	Gxz	Gyz	Gzx	Gzy	Gzz
α	f	f	f	f	t	t	f	f	f
β_1	f	f	t	f	t	t	f	f	f
β_2	f	t	f	f	t	t	f	f	f
a	f	f	t	f	t	t	t	f	f
b	f	t	f	f	t	t	f	t	f
c	f	f	f	f	t	t	t	f	f
d	f	t	f	f	t	t	t	f	f
e	f	t	f	f	t	t	t	t	f

Row e has no mate, beacuse of the columns 5 to 8. Rows c and d have no mate, because b, the only row satisfying the condition on Gzx and Gzy, does not satisfy the condition on Gxy and Gyx. Neither row a nor row b has subordinates as required by 3.8.3. Hence, we have only the remaining rows α, β_1, β_2 to consider.

α is the only uniform row, (β_1, β_2) form a parallel pair, and β_2 is both P and S in 3.8.3 for all the three rows α, β_1, β_2. Hence, we have, for example:



In particular, ($M214,\beta_1$) can be continued in the same way as ($M123$, β_2). Indeed, continuation in every branch can be made similarly. In other words, α, β_1, β_2 are all normal by 3.8. This, however, does not yet secure a model for the formula 3.2. There are, for example, those instances in which (1,5), (5,1), (3,4), (4,3), etc., replace (x,y) of $Mxyz$; they also have common atomic formulae with the instances shown in the above graph.

3.10 A formula 3.1 has a model if and only if (a) it has a nonempty table of normal rows, (b) this table has a nonempty subtable T' such that:

3.10.1 For every pair (R,S) in T', there is a parallel pair (P,Q) in T' such that $P_{xx} = R_{zz}$, $Q_{xx} = S_{zz}$.

3.10.2 There is a uniform row R in T' such that for every row S in T', there is a parallel pair (P,Q) in T', for which $P_{xx} = R_{xx}$, $Q_{xx} = S_{zz}$.

These are the additional requirements mentioned after 3.4. In the example under consideration, the table consisting of all the three normal rows α, β_1, β_2 satisfies the requirements on T'. Hence, 3.2 does have models. One model for the predicate G is the relation $<$ among positive integers. That is, however, not the only model, because the model of G does not have to be transitive. For example, $G15$ and $G51$ can be (t,f) or (f,t) or (f,f).

It can be verified that the conditions in 3.10 are indeed necessary and sufficient.

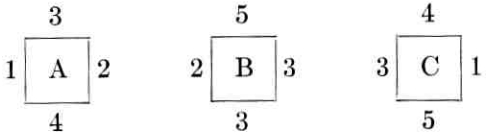
IV. THE $A_1E_1A_1$ SATISFIABILITY CASE

4.1 *A Generalized Game of Dominoes*

The study of the decision problem of the present case has suggested a related abstract mathematical problem which can easily be stated in everyday language. The problem appears to be of interest even to those who are not concerned with questions in mathematical logic.

Assume we are given a finite set of square plates of the same size with edges colored, each in a different manner. Suppose further there are infinitely many copies of each plate (plate type). We are not permitted to rotate or reflect a plate. The question is to find an effective procedure by which we can decide, for each given finite set of plates, whether we can cover up the whole plane (or, equivalently, an infinite quadrant thereof) with copies of the plates subject to the restriction that adjoining edges must have the same color.

For example, suppose a set consists of the three plates:



Then we can easily find an infinite solution by the following argument. The following configuration satisfies the constraint on the edges:

$$\begin{array}{ccc} A & B & C \\ C & A & B \\ B & C & A \end{array}$$

Now the colors on the periphery of the above block are seen to be the following:

$$\begin{array}{ccccc} & 3 & 5 & 4 & \\ 1 & & & & 1 \\ 3 & & & & 3 \\ 2 & & & & 2 \\ & 3 & 5 & 4 & \end{array}$$

In other words, the bottom edge repeats the top edge, and the right edge repeats the left edge. Hence, if we repeat the 3 × 3 block in every direction, we obtain a solution of the given set of three plates. In general, we define a "cyclic rectangle."

4.1.1 Given any finite set of plates, a cyclic rectangle of the plates is a rectangle consisting of copies of some or all plates of the set such that: (a) adjoining edges always have the same color; (b) the bottom edge of the rectangle repeats the top edge; (c) the right edge repeats the left edge.

Clearly, a sufficient condition for a set of plates to have a solution is that there exists a cyclic rectangle of the plates.

What appears to be a reasonable conjecture, which has resisted proof or disproof so far, is:

4.1.2 *The fundamental conjecture*: A finite set of plates is solvable (has at least one solution) if and only if there exists a cyclic rectangle of the plates; or, in other words, a finite set of plates is solvable if and only if it has at least one periodic solution.

It is easy to prove the following:

4.1.3 If 4.1.2 is true, we can decide effectively whether any given finite set of plates is solvable.

Thus, we proceed to build all possible rectangles from copies of the plates of different sizes, using smaller ones first. If 4.1.2 is true, the process will always terminate in one of two ways: either at some stage we arrive at a cyclic rectangle and, therefore, the original set is solvable; or else we arrive at a size such that there is no rectangle of that size in which adjoining edges always have the same colors. The latter alternative is in fact a necessary and sufficient condition under which the original set is not solvable. However, if 4.1.2 is not true, it would be possible that a set has a solution, but we can never see this fact by the latter criterion at any finite stage: there would always be the possibility that for the next size there exist no rectangles with same-colored adjoining edges.

There is a naturally uneasy feeling about the effectiveness of such a procedure. The argument is essentially the familiar one that if a set and its complement are both recursively enumerable, then the set is recursive. It shows that the procedure always terminates (provided 4.1.2 is true) but gives no indication in advance as to how long it might take in each case.

If 4.1.2 is proved, it seems likely that it would be proved in a stronger form by exhibiting some simple recursive function f with the following property. For any set of plates with m distinct colors and n distinct plates, if the set is solvable, there is a cyclic square of the size $k \times k$, where $k = f(m,n)$. If that happens, or even if we have not exhibited such a function f but 4.1.2 can be proved by fairly elementary arguments, we would have some estimate in advance of how long the procedure takes in each case.

As it is, we can make the testing procedure quite systematic even though we do not know whether 4.1.2 is true. The procedure would be a decision procedure and presumably quite an efficient one, if 4.1.2 is true. If 4.1.2 should turn out to be false, then the procedure would only be a semidecision procedure. In fact, it is possible to show that the procedure does work in several classes of cases, e.g., when a set has unique solution apart from translations, or whenever either horizontally or vertically no color can be followed by different colors. But we shall not delay over such partial results.

If 4.1.2 should be false, then there would be two possibilities: either the set of all solvable finite sets of plates is not recursive, or it is recursive but requires a more complex decision procedure.

The problem can clearly be generalized to higher dimensions: for example, to cubes with colored surfaces instead of squares with colored edges.

We return now to the $A_1E_1A_1$ satisfiability case.

4.2 *Preliminary Definitions and an Example*

The general form of the case is:

4.2.1 $$(x)(Ey)(z)Mxyz,$$

where M is a quantifier-free matrix containing neither function symbols nor the equality sign. From the fundamental theorem, it follows that 4.2.1 is satisfiable (solvable) if and only if each finite subset of the infinite set of matrices $Mii'j$ $(i,j = 1, 2, \cdots)$ is solvable (not contradictory). Since the second number is always the successor of the first, we shall write Mij for $Mii'j$.

We illustrate the general case by considering the special case where $Mxyz$ contains only a single dyadic predicate G. The negation of Example (4) given in the introduction of Part I[1] will be the concrete example:

4.2.2 $$(x)(Ey)(z)[\sim Gxx \ \& \ Gxy \ \& \ (Gyz \supset Gxz)].$$

In the alternative notation, the matrix is

4.2.3 $$Gxy, Gxz \nrightarrow Gxx; Gxy \nrightarrow Gyz, Gxx.$$

The truth table is:

4.2.4

Gxx	Gxy	Gyx	Gyy	Gxz	Gzx	Gyz	Gzy	Gzz
f	t			t		t		
f	t			t		f		
f	t			f		f		

Since there are five blank columns, there are altogether 3×2^5 or 96 rows. The problem now is to decide whether we can choose one row for each matrix $Mij(i,j = 1, 2, \cdots)$ such that, taken together, all the matrices come out true. This really involves both the problem of finding the pieces and the problem of putting them together. Thus, if j is distinct from i and i', any row can satisfy Mij alone, if we substitute i, i', j for x, y, z in the truth table; but a row can satisfy Mij when j is i or i' only in case certain related columns get the same truth values. This is the problem of finding the pieces. When there are such pieces, there is the harder problem of putting them together. For example, if there are rows satisfying $M11$ and $M12$ separately, there may yet be no pair of rows which satisfy $M11$ and $M12$ simultaneously because the common atomic formulae in both matrices must get identical values.

Since the putting-together part is quite complex, it seems natural to combine small pieces into blocks first. For this purpose, we consider row pairs and row quadruples (i.e., pairs of pairs).

D4.1 Two rows P,Q in the truth table T form a basic row pair (P,Q) if, for some i, they can simultaneously satisfy Mii' and Mii respectively. More explicitly, the conditions are:

i. $P_{yy} = P_{yz} = P_{zy} = P_{zz}$, $P_{xy} = P_{xz}$, $P_{yx} = P_{zx}$;

ii. $Q_{xx} = Q_{xz} = Q_{zx} = Q_{zz}$, $Q_{xy} = Q_{zy}$, $Q_{yx} = Q_{yz}$;

iii. $P_{xx} = Q_{xx}$, $P_{xy} = Q_{xy}$, $P_{yx} = Q_{yx}$, $P_{yy} = Q_{yy}$.

In the table 4.2.4, it is easy to verify that there are only two basic row pairs (α,β) and (γ,δ):

4.2.5

	Gxx	Gxy	Gyx	Gyy	Gxz	Gzx	Gyz	Gzy	Gzz
α	f	t	f	f	t	f	f	f	f
β	f	t	f	f	f	f	f	t	f
γ	f	t	f	t	t	f	t	t	t
δ	f	t	f	t	f	f	f	t	f

Obviously basic row pairs are necessary for building a model of 4.2.1. In fact, given any formula 4.2.1, if its truth table T contains no basic row pairs, then it has no model and, indeed, the conjunction of $M11$ and $M12$ is a contradiction.

We shall consider pairs of row pairs, called row quadruples, which are useful in chaining row pairs together.

D4.2. Given any two row quadruples $(A,B;\ C,D)$ and $(P,Q;\ R,S)$, if $C = P$, $D = Q$, then the former is a predecessor of the latter and the latter is a successor of the former.

D4.3. Four rows P, Q, R, S form a basic row quadruple $(P,Q;\ R,S)$ if, for some i, they satisfy simultaneously Mii', Mii, $Mi'i''$, $Mi'i'$, respectively, or, more explicitly, if:

i. (P,Q) and (R,S) are basic row pairs;

ii. $P_{yy} = R_{xx}$;

iii. $(P,Q;\ R,S)$ has a successor which is a basic row quadruple.

In the table 4.2.4, there is only one basic row quadruple, viz., $(\alpha,\beta;\ \alpha,\beta)$. The quadruple $(\alpha,\beta;\ \gamma,\delta)$ satisfies i and ii, but not iii. It is easy to see that, given any formula 4.2.1, if its truth table T contains no basic row quadruples, then it has no solution and, indeed, the conjunction of $M12$, $M11$, $M23$, $M22$, $M34$, $M33$ is a contradiction.

Clearly, if a row R satisfies Mij' in a model, then there must be one row S which satisfies Mji, one basic row quadruple $(A,B;\ C,D)$ which satisfies Mii', Mii, $Mi'i''$, $Mi'i'$, and one basic quadruple which satisfies Mjj', Mjj, $Mj'j''$, $Mj'j'$. In particular, when j is i, we get the basic row pairs which occur in some basic quadruple.

D4.4 Two rows R,S form an ordinary row pair (R,S) if

i. $R_{xx} = S_{zz}$, $R_{xz} = S_{zy}$, $R_{zx} = S_{yz}$, $R_{zz} = S_{yy}$;

ii. There is a basic quadruple $(A,B;\ C,D)$ such that $A_{xx} = R_{xx}$, $A_{xy} = R_{xy}$, $A_{yx} = R_{yx}$, $A_{yy} = R_{yy}$;

iii. There is a basic quadruple $(P,Q;\ K,L)$ such that $P_{xx} = S_{xx}$, $P_{xy} = S_{xy}$, $P_{yx} = S_{yx}$, $P_{yy} = S_{yy}$.

In the table 4.2.4, since the only basic quadruple is $(\alpha,\beta;\ \alpha,\beta)$, it is relatively simple to find all the rows which do occur in ordinary row pairs. Since every row which is to satisfy some *Mij* in any solution must occur as one row in some ordinary row pair, we tabulate all such rows together and, from now on, confine our attention to them. It happens in this example that all these rows have in common five columns:

Gxx	*Gxy*	*Gyx*	*Gyy*	*Gzz*
f	t	f	f	f

Therefore, we only have to list the remaining columns:

4.2.6

	Gxz	*Gzx*	*Gyz*	*Gzy*	*ordinary pairs*
α	t	f	f	f	(α,β)
β	f	f	f	t	(β,α)
δ_1	t	t	t	t	(δ_1 , δ_1)
δ_2	f	f	f	f	(δ_2 , δ_2)
δ_3	t	f	f	t	(δ_3 , δ_3)
δ_4	t	f	t	f	(δ_4 , δ_5)
δ_5	f	t	f	t	(δ_5 , δ_4)
δ_6	t	f	t	t	(δ_6 , δ_7)
δ_7	t	t	f	t	(δ_7 , δ_6)

In fact, if only the four columns have to be considered, there are 12 rows in the original table 4.2.4, and the two rows (R,S) in each ordinary row pair satisfy the condition: $R_{xz} = S_{zy}$, $R_{zx} = S_{yz}$. Hence, it is easy to get the above table. Briefly, the relevant information for the example is the nine ordinary pairs given above and the basic quadruple $(\alpha,\beta;\ \alpha,\beta)$.

Thus far we have been concerned only with rather elementary properties of the rows in the truth table. The more involved part is to design a scheme of extending recursively the construction of models. In order to explain how this is done, we introduce a chart.

4.2.7 Chart for $(x)(Ey)(z)Mxyz$:

Basic Pairs *Cyclic Pairs* *Common Row Pairs*

(Mii', Mii) $(Mii'', Mi'i)$ $(Mi(i''+k), M(i'+k)i)$

$$\begin{tikzcd}
(12, 11) \arrow[r] \arrow[d] & (13, 21) \arrow[r] \arrow[d] & (14, 31) \arrow[r] \arrow[d] & (15, 41) \arrow[r] \arrow[d] & (16, 51) \arrow[r] \arrow[d] & \cdots \\
(23, 22) \arrow[r] \arrow[d] \arrow[ur] & (24, 32) \arrow[r] \arrow[d] \arrow[ur] & (25, 42) \arrow[r] \arrow[d] \arrow[ur] & (26, 52) \arrow[r] \arrow[d] \arrow[ur] & \cdots \vdots & \\
(34, 33) \arrow[r] \arrow[d] \arrow[ur] & (35, 43) \arrow[r] \arrow[d] \arrow[ur] & (36, 53) \arrow[r] \arrow[d] \arrow[ur] & \cdots \vdots & & \\
(45, 44) \arrow[r] \arrow[d] \arrow[ur] & (46, 54) \arrow[r] \arrow[d] \arrow[ur] & \cdots \vdots & & & \\
(56, 55) \arrow[r] \arrow[d] \arrow[ur] & \cdots \vdots & & & & \\
\vdots & & & & &
\end{tikzcd}$$

In the chart, the ordinary row pairs satisfying (Mij', Mji) are divided into three classes: basic when $i = j$, cyclic when $i' = j$, common otherwise. The general plan of the procedure is as follows. The existence of basic row quadruples assures that we can find a model for all the matrices $M12$, $M11$, $M23$, $M22$, etc. in the first column. Similarly, we can define cyclic quadruples to give an effective condition for the existence of a model for all matrices appearing in the second column of the chart, and so on. But in order that these models can be combined to give a model for all the matrices and therewith for a given formula 4.2.1, each column must be related to the column on its left in a suitable manner. This situation with two infinite dimensions seems to be the chief cause of the complexity of the $A_1E_1A_1$ case.

In the chart of 4.2.7, each row pair (R,S) that is not basic is subordinate to a quadruple $(A,B; C,D)$ made up of the two row pairs (A,B), (C,D) on its left with arrows leading to it. The quadruple is said to be superior to the pair (R,S).

D4.5 An ordinary row pair (R,S) is a subordinate of a quadruple $(A,B; C,D)$ if

i. $R_{xx} = A_{xx}$, $R_{xy} = A_{xy}$, $R_{yx} = A_{yx}$, $R_{yy} = A_{yy}$, $R_{yz} = C_{xz}$, $R_{zy} = C_{zx}$, $R_{zz} = C_{zz}$;

ii. $S_{xx} = D_{xx}$, $S_{xy} = D_{xy}$, $S_{yx} = D_{yx}$, $S_{xz} = A_{zx}$, $S_{zx} = A_{xz}$.

A quadruple $(R,S; P,Q)$ is subordinate to a row sextuple $(A,B; C,D; K,L)$ if (R,S) is subordinate to $(A,B; C,D)$, and (P,Q) to $(C,D; K,L)$.

D4.6 Two rows R,S form a cyclic row pair (R,S) if

i. (R,S) is an ordinary row pair;

ii. $R_{xy} = S_{zx}$, $R_{yx} = S_{xz}$, $R_{yy} = S_{xx}$, $R_{yz} = S_{xy}$, $R_{zy} = S_{yx}$.

Obviously, given 4.1, if its table contains no two rows forming a cyclic pair, then the conjunction, briefly C_6, of $M12$, $M11$, $M23$, $M22$, $M13$, $M21$ is a contradiction.

In the table 4.2.6, there are, among the nine ordinary row pairs, only one that is cyclic, (δ_4, δ_5). Since there are only one basic quadruple, each has only one superior. This is of course not always the case, it is only due to special features of the example 4.2.2.

In order to find out whether there is any succession of cyclic pairs which will satisfy all rows of the column for cyclic pairs in the chart, we study cyclic quadruples.

D4.7 Four rows P,Q,R,S form a cyclic quadruple $(P,Q; R,S)$ if

i. (P,Q) and (R,S) are cyclic row pairs;

ii. $Q_{xx} = R_{xx}$, $Q_{xy} = R_{xy}$, $Q_{yx} = R_{yx}$, $Q_{yy} = R_{yy}$;

iii. There is a basic sextuple $(A,B; C,D; K,L)$; which is respectively superior to $(P,Q; R,S)$;

iv. $(P,Q; R,S)$ has a successor which is also a cyclic quadruple.

Obviously, given a formula 4.2.1, if its table contains no rows that form a cyclic quadruple, then the conjunction of C_6, $M34$, $M33$, $M24$, $M32$ is a contradiction.

The existence of a cyclic quadruple certainly assures that we can satisfy all the rows of the second column of the chart simultaneously. It assures a bit more: the two pairs (P,Q), (R,S) of a cyclic quadruple are always compatible with any three pairs (A,B), (C,D), (K,L) which form two basic quadruples, respectively superior to them. This is, however, insufficient to secure that all the rows in the first two columns of the chart can be simultaneously satisfied, because it is possible that no cyclic quadruple beginning with (R,S) is subordinate to any quadruple beginning with (K,L). In other words, the blocks might not fit together.

As it happens, this problem does not arise with the example 4.2.2. Since there is only one cyclic pair (δ_4, δ_5), there can be at most one cyclic quadruple, viz, $(\delta_4, \delta_5; \delta_4, \delta_5)$. It can be verified by D4.7 that this is indeed a cyclic row quadruple. Since there is only one basic quadruple $(\alpha,\beta; \alpha,\beta)$, we see immediately that by using (α,β) for (Mii', Mii) $(i = 1, 2, \cdots)$ and (δ_4, δ_5) for $(Mii'', Mi'i)(i = 1, 2, \cdots)$, all these matrices (of the first two columns of the chart) are simultaneously satisfied. Moreover, this is the only possible model for the two initial infinite columns of matrices.

We shall first define common row quadruples, settle 4.2.2, and then come back to the more general question.

D4.8 Two ordinary row pairs (R,S), (P,Q) form a common quadruple $(R,S;\ P,Q)$ of order k [i.e., in the $(2 + k)$th column of the chart] if

i. When $k = 1$, there is a cyclic row sextuple which is superior to $(R,S;\ P,Q)$; or when $k = n + 1$, for some positive integer n, there is a common row sextuple of order n which is superior to $(R,S;\ P,Q)$.

ii. $(R,S;P,Q)$ has a successor which is a common quadruple of order k.

By this definition, we can successively find the common row quadruples of orders 1, 2, etc. In the actual procedure, we examine each time to determine whether we have already enough information to decide the original formula. Only when this is not the case do we find the common quadruples of the next order.

In the case of 4.2.2, since $(\delta_4\,,\ \delta_5\ ;\ \delta_4\,,\ \delta_5)$ is the only cyclic quadruple, it is easy to verify, by 4.2.6 and D4.5 that $(\delta_4\,,\ \delta_5\ ;\ \delta_4\,,\ \delta_5)$ is the only common quadruple of order 1. Thus, by D4.5, if (R,S) is subordinate to the cyclic quadruple $(\delta_4\,,\ \delta_5\ ;\ \delta_4\,,\ \delta_5)$, $R_{yz} = (\delta_4)_{xz} = t$, $R_{zy} = (\delta_4)_{zx} = f$, and $S_{xz} = (\delta_4)_{zx} = f$, $S_{zx} = (\delta_4)_{xz} = t$. By 4.2.6, (R,S) must be $(\delta_4\,,\ \delta_5)$.

From this, it follows that, for every n, there is exactly one common quadruple of order n, viz. $(\delta_4\,,\ \delta_5\ ;\ \delta_4\,,\ \delta_5)$. This is an immediate consequence of D4.8 and the above transition from the cyclic column to the first common column in the chart. Hence, we have obtained a model for 4.2.2. It is easy to verify that the model for G is just the usual ordering relation $<$ among positive integers.

This completes the solution of the example 4.2.2, which, however, is not a sufficient illustration of the general case. We have to discuss a procedure by considering more complex situations.

4.3 *The Procedure*

One possible procedure is to add one infinite column at a time. Thus, it is possible to represent all possible solutions of each column by a graph, and to represent the solutions satisfying all the initial n' columns by a finite set of graphs if it is possible so to represent all solutions satisfying the initial n columns. Since the common columns enjoy a measure of uniformity, simultaneous solutions for all the columns would be assured if suitable repetitions occur. An exact explanation of such a procedure would be quite lengthy. In any case, a successful choice of patterns of repetition has not been found to assure that for every solvable table, such repetition always occurs.

Instead of elaborating the above procedure, we transform the problem to something similar to the abstract question of 4.1. Thus, given any formula of the form 4.2.1, we can, as in 4.2, construct its truth table and find all the common row pairs in the table. Among the common row pairs, some are also cyclic row pairs and some are also basic row pairs.

If now we take the common row pairs a, b, c, d, etc., as elementary units which are to fill up the infinite quadrant as shown in the chart given under 4.2.7, then the following scheme appears to be feasible. Suppose the points in the infinite quadrant are to be filled by a_{ij}, $i,j = 1, \delta, \cdots$, then we may consider instead all the 2 × 2 matrices:

$$\begin{pmatrix} a_{ij} & a_{ij'} \\ a_{i'j} & a_{i'j'} \end{pmatrix}, \qquad \text{for all } i,j = 1,2,\cdots$$

In other words, given the common row pairs, we can form all possible 2 × 2 matrices of them which satisfy the relations of subordination. These 2 × 2 matrices are then the basic pieces from which we are to obtain an infinite solution subject to the conditions: (a) consecutive rows or columns from two matrices are the same; (b) only basic and cyclic row pairs are permitted in the first two columns.

It can be verified that the problem of finding a model for the original formula is equivalent to that of finding a way to fill up the infinite quadrant by such derived 2 × 2 blocks of row pairs.

The abstract problem is: given any finite set of 2 × 2 matrices of the form

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix},$$

to decide whether it is possible to fill up the infinite quadrant with copies of these pieces. This is not quite the same as the problem of colored plates described in 4.1, because here what is done amounts to coloring the corners, or imposing connections between neighboring sides within a same square.

Any set of such 2 × 2 matrices can also be construed as a set of colored plates. Conversely, given any set of colored plates, we can also find in a systematic manner a corresponding set of such matrices such that the solvability problems for them are equivalent. For example, we may replace a colored plate by a block of nine 2 × 2 matrices so that the restriction on neighboring sides no longer operates.

It is possible to use a procedure similar to the one described roughly in 4.1. Some change is needed to take care of the additional conditions on the first two columns. Thus, a sufficient condition is to get a cyclic rectangle $m \times n$ on which we can attach a frill of two columns on the left to obtain a rectangle $m \times (2 + n)$ such that: (a) the tops of the first two columns are the same as the bottoms; (*b*) the additional requirements of being basic or cyclic are satisfied by the frills.

4.4 *Further Problems*

The discussions so far seem to have barely scratched the surface of a group of rather difficult problems, among which the basic one is probably that of measuring the complexity of formulae in the predicate calculus.

One may measure the complexity of a formula in many different ways. The "simplest" model of a formula may be taken as a semantic measure. The quantifier prefix or graph of a formula may be taken as a syntactic measure. In addition, for formulae with a same prefix, we may also classify the possible matrices by the truth tables. Our knowledge on using these criteria to give detailed classifications seems very limited. One example of the ignorance is the following open problem (Ref. 2, p. 177): whether there is any class of formulae which is neither decidable, nor a reduction class. It appears reasonable to conjecture that there must be such classes, although the first examples which one will get are likely to be artificial ones.

Some of the reduction classes are, formally speaking, surprisingly simple. For example, from the Surányi normal form given above as case VIII, it follows that, for satisfiability, one reduction class is:

4.4.1 Formulae with prefix $(x)(y)(Ez)(w)Mxyzw$, where M contains neither function symbols, nor $=$, nor predicate letters which are not dyadic.

Since each matrix M is effectively determined by a truth table on the atomic formulae in M, the class may be viewed as a union of a simple sequence of finite classes C_1, C_2, etc., where C_n is the subclass of formulae each containing exactly n predicates (or, equivalently, the first n predicates in some enumeration). There is a sense in which the decision problem for each finite set of formulae is solvable, and yet usually we as a matter of fact only solve the problem as a corollary to a solution for some infinite class.

To obtain a semidecision procedure for the class VIII or 4.4.1, we need more complicated arrangements of triples or quadruples of positive integers than the case $A_1E_1A_1$. Take, for example, the class in case VIII. We have to consider not only the triples (a,b,c) with $b = a'$, but all the triples for the first half of the formula, and among them those for the A_2E_1 case are used simultaneously for the second half of the formula.

An example is:

4.4.2 $(x)(y)(z)(\sim Gxy \vee \sim Gyz \vee Gxz) \,\&\, (x)(y)(Eu)(\sim Gxx \,\&\, Gyu).$

If we use the Skolem function g of the A_2E_1 case, we can rewrite the above as

4.4.3 $$(\sim Gxy \vee \sim Gyz \vee Gxz) \,\&\, (\sim Gxx \,\&\, Gygxy).$$

In general, we are concerned with deciding the satisfiability of formulae of the form

4.4.4 $$Mxyz \,\&\, Nxygxy.$$

As (x,y,z) runs through all triples of positive integers, we get an infinite sequence from 4.4.4, and a semidecision procedure is to decide, for certain cases, whether such an infinite sequence can be simultaneously satisfied.

For example, we may throw together all permutations of a given triple, and confine ourselves to the triples (a,b,c) with $a \leqq b \leqq c$, assigning each of them a lattice point:

$$\begin{aligned} f(x,y,z) &= (x - 1, z - x, z - y), \\ f^{-1}(x,y,z) &= (x + 1, x + y, x + y + z). \end{aligned}$$

The correlation uses all lattice points (x,y,z) of nonnegative integers. For instance, $(1,3,5)$ gets the point $(0,2,2)$.

We might try to create different types of cubes each with eight vertices from (i,j,k) to (i',j',k') and piece them together. But it is not easy to see how to find a procedure analogous to that described in 4.1 which would at the same time take into consideration the second half of the formula.

V. A PROOF PROCEDURE FOR THE PREDICATE CALCULUS

5.1 *The Quantifier-Free Logic F*

Given the definition of formulae in 1.2, we can define sequents, antecedents, consequents, as in Ref. 13, p. 5. The sequents in F are those containing no quantifiers and the rules for F are exactly the same as those for P_e (Ref. 13, p. 8), except for containing not only variables but also functional expressions as terms.

Example 1. $1 \neq x', x = x + 1 \rightarrow 1 \neq x + 1$

By the rules $P2$a and $P2$b (Ref. 13, p. 5), this is a theorem if the following is:

$$1 = x + 1, x' = x + 1 \rightarrow 1 = x'.$$

This is a theorem by $P7$ and $P8$ (Ref. 13, p. 8).

Example 2. $x + y' = (x + y)', y \neq x + y, y' = v' \supset y = v, v = x + y \rightarrow y' \neq x + y$

By *P*2a, *P*2b, and *P*5b, this is a theorem if the following two sequents are:

i. $x + y' = (x + y)', y = v, v = x + y, y' = x + y \rightarrow y = x + y;$

ii. $x + y' = (x + y)', v = x + y, y' = x + y' \rightarrow y' = v', y = x + y.$

i. is a theorem by *P*7 and *P*8 since we can replace y and $x + y$ by v.
ii. is also a theorem because we can replace v' by $(x + y)'$ and then y' by $x + y'$ in the first clause of the consequent and the result is a theorem by *P*1.

These rules in fact yield a decision procedure for all quantifier-free sequents. In order to see this, we use a more efficient method to speed up applications of P7 and P8.

Given an atomic sequent which contains equality but is not yet a theorem by *P*1 or *P*7. List every pair (a,b) if $a = b$ occurs in the antecedent. Extend repeatedly the set of pairs by symmetry and transitivity. Join each pair by the equals sign and add all of them to the antecedent. Now compare each clause in the antecedent with each clause in the consequent to see whether there is a pair of clauses which can be obtained from each other by substituting equals for equals; moreover, examine each equality in the consequent to see whether it can turn into $\alpha = \alpha$ by substituting equals for equals. If either case occurs, the sequent is a theorem. If neither is the case, then we can find an interpretation of the functions and predicates so that the antecedents are all true but the consequents are all false.

5.2 *The Rules for Quantifiers.*

In the present formulation of the predicate calculus, one emphasis is on separating out reversible rules of proof which serve to supply decision procedures as well, because they have the property that not only the premises imply the conclusion but also conversely.

The rules governing quantifiers were given in Part I.[1]*

* "S4. When the input problem contains quantifiers, the following preliminary simplifications are made: (i) All free variables are replaced by numbers, distinct numbers for distinct variables. (ii) Vacuous quantifiers, i.e., quantifiers whose variables do not occur in their scopes, are deleted. (iii) Different quantifiers are to get distinct variables; for example, if (x) occurs twice, one of its occurrences is replaced by (z), z being a new variable. This last step of modification is specially useful when occurrences of a same quantifier are eliminated more than once at different stages.

The justification of the reduction to subproblems (Part I, T2.1) is obvious because all truth-functional rules are reversible and $(x)(Gx$ & $Hx)$ is a theorem if and only if $(x)Gx$ and $(x)Hx$ both are.

Usually T2.2 (Part I) is true, but restrictions are necessary, as the following example would show:

$$(x)(Ey)[(z)Gyz \text{ \& } Hxy].$$

Although x does not occur in the scope of (z), there is no way to bring (z) out of the scope of (x) because the variable y ties up the two clauses in the formula. There are several possible alternatives: one may make exact the restrictions needed, or record the scope of each quantifier in the usual manner, or use the easy simplification that when a quantifier governs a formula with two halves joined by a logical connective but the variable of the quantifier occurs only in one of the two halves, the scope is just that half.

The test of connectedness of variables and functors (Part I, T2.3) is meant as a device to simplify the interconnections between quantifiers. In particular, the test gives a method for ascertaining that certain apparently complex sequents fall under the AE provability case. In order, however, actually to bring such a set of sequents into the AE form, we need in general transformations similar to those used in reducing a sequent to the miniscope form. Since the process can be tedious, one may prefer an alternative method of not carrying out the transformation but merely determining a bound k such that either the original sequent is a theorem or has a counter-model with no more than k objects. If this alternative is chosen, a method for calculating the bound k has to be devised.

In any case, when we have a finite set of atomic sequents and a set of governing relations among the variables and functors, we should further simplify the matrix, i.e., the set of atomic sequents by the familiar methods of dropping repetitions and immediate consequences.

"S5. After the above preliminary simplifications, each problem is reduced to as many subproblems as possible in the following manner: (i) Eliminate in the usual manner every truth-functional connective which is not governed by any quantifiers. (ii) Drop every initial positive quantifier (i.e., universal in the consequent or existential in the antecedent that is not in the scope of any other quantifier) and treat its variable as free, i.e., replace all its occurrences by those of a new number. (i) and (ii) are repeated for as long as possible. As a final result of this step, each problem is reduced to a finite set of subproblems such that the problem is a theorem if and only if all the subproblems are.

"T2.1 The original problem is a theorem if and only if all its subproblems (in the above sense) are.

"T2.2 We can separate out Q and its scope from those quantifiers whose variables do not occur in the scope of Q.

"T2.3 If two symbols, each a functor or a variable, are not connected in the final matrix, we can always so transform the original sequent as to separate the two quantifiers which give way to them."

If there are two subsets of the set of atomic sequents which contain neither common variables nor common functors, then they can be separated.

Moreover, each atomic formula that contains neither variables nor functors can be eliminated by the familiar method of replacing $F(p)$ by $F(t)$ & $F(f)$. In other words, it can simply be dropped on the ground of the following consideration. E.g., take

$$Guv, G11 \rightarrow Gvk.$$

This is equivalent to the conjunction of:

$$\begin{gathered} Guv, \mathrm{t} \rightarrow Gvk; \\ Guv, \mathrm{f} \rightarrow Gvk. \end{gathered}$$

But the second sequent is always true and can be dropped; the t in the first sequent can be dropped, so that we have

$$Guv \rightarrow Gvk.$$

After all the above steps, we arrive at a finite set of finite sets of atomic sequents which, taken together, are equivalent to the original problem. We may consider each finite set of atomic sequents separately and proceed according to the governing relations between their variables and functors.

We can view the set as a formula in the prenex form with a matrix in a conjunctive normal form. Or, if we prefer, we may replace $\rightarrow$ by $\nrightarrow$ and construe the variables as universal quantifiers, the functors as existential quantifiers. Then we get a negation of the formula in prenex form with a matrix in the disjunctive normal form.

In either case, the remaining problem is to be handled by considerations such as those explained in Sections II through IV.

There is an easily mechanizable procedure by which we can, in theory, not only prove all provable formulae, but also refute all formulae which have finite countermodels. All we have to do is test, besides the sequence M_1, M_2, M_3, etc., whether a formula is satisfiable in a domain with one object, or two objects, or etc. For example, given

$$(x)(y)(Ez)Mxyz, \tag{1}$$

if some of $M112$, $M123$, $\cdots$ is contradictory, then the negation of (1) is a theorem; if relative to some finite domain, (1) can be satisfied, then the negation of (1) is not a theorem. For example, (1) is satisfiable in a domain with one object if and only if $M111$ is satisfiable; with two objects, if and only if

$$(x)(y)(Mxy1 \vee Mxy2)$$

or

$$(x)[(Mx11 \vee Mx12) \,\&\, (Mx21 \vee Mx22)]$$

or

$$[(M111 \vee M112) \,\&\, (M121 \vee M122)] \\ \&\, [(M211 \vee M212) \,\&\, (M221 \vee M222)]$$

is satisfiable.

VI. REMARKS ON MATHEMATICAL DISCIPLINES

Besides the contrast between proving and calculating, there is a contrast between symbol manipulation and number manipulation. There are problems such as proving trigonometric identities, factorization, differentiation and integration, which all appear to be mechanizable. In numerical calculations, it appears likely that the process of choosing one or another method of calculation can also be mechanized in many cases.

There is the problem of applying the methods considered so far to deal with concrete examples.

One example referred to in Part I[1] (p. 231) is Hintikka's derivation of a contradiction from his own formal system.[14] Here, intuitive understanding is required to select from the set of all axioms suitable members which are sufficient to produce contradictions. Experience, however, shows that, even after a reasonable selection is made, to actually give an exact derivation of a contradiction remains quite a dreary affair. In such a case, the sort of procedure discussed in this paper can be useful.

In fact, Hintikka uses five axioms to derive a contradiction. Write briefly:

$$Hayz \quad \text{for} \quad z \neq a \,\&\, z \neq y \,\&\, z \in y \,\&\, y \in z.$$

The conjunction of the axioms is:

$$\begin{aligned}&(Ex)(Ey)(x \neq y) \,\&\\ &\quad (Ea)(Eb)(Ec)(Ed)(y)\,\{[y \neq a \supset (y \in a \equiv (Ez)Hayz)] \,\&\\ &\qquad\qquad [y \neq b \supset (y \in b \equiv \sim(Ez)Hbyz)] \,\& \qquad (2)\\ &\quad [y \neq c \supset (y \in c \equiv (y = a \vee y = b))] \,\&\\ &\qquad\qquad [y \neq d \supset (y \in d \equiv y = c)]\}.\end{aligned}$$

The assertion is that (2) leads to a contradiction. In other words, (2) has no model, and its negation is a theorem of the predicate calculus. To decide whether this assertion is true, we only have to test (2) by essentially the method of Section III because (2) can be transformed into a formula with EA_2E prefix. Such a method yields also a proof or a refutation of the assertion that (2) gives a contradiction.

In a different direction, we may consider some simple examples in the arithmetic of positive integers.

First, we consider the example, $x' \neq x$. We wish, in other words, to prove, with the help of induction, that this is a consequence of the axioms:

$$\begin{aligned} &x' \neq 1, \\ &x' \neq y' \rightarrow x \neq y. \end{aligned}$$

As a general principle, we try to use induction. Since there is only one variable, we reduce the problem to:

$$(x)x' \neq 1,\ (x)(y)(x' = y' \supset x = y) \rightarrow 1' \neq 1, \tag{3}$$

$$(x)x' \neq 1,\ (x)(y)(x' = y' \supset x = y),\ x' \neq x \rightarrow x'' \neq x'. \tag{4}$$

These can be dealt with by the program described in Part I, except that, to avoid confusion, we use now a, b, c, etc., instead of numerals to replace the positive variables. We have:

$$\begin{aligned} 1' = 1,\ u = v &\rightarrow x' = 1, \\ 1' = 1 &\rightarrow x' = 1,\ u' = v', \\ u = v,\ a'' = a' &\rightarrow x' = 1,\ a' = a, \\ a'' = a' &\rightarrow x' = 1,\ a' = a,\ u' = v'. \end{aligned}$$

These sequents are all true by substitution: 1 for x in the first two; a' for u and a for v in the last two.

As a somewhat more complex example, we take the commutativity of addition. In order to prove $x + y = y + x$, we may use induction either on x or on y. We arbitrarily take the earliest variable:

$$1 + y = y + 1, \tag{5}$$

$$x + y = y + x \rightarrow x' + y = y + x'. \tag{6}$$

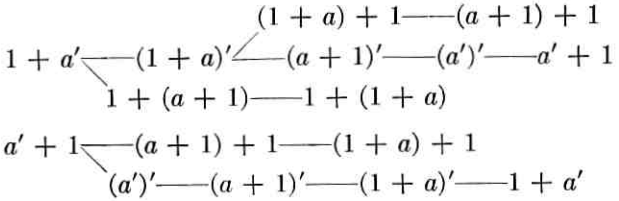
To prove $1 + y = y + 1$, we make induction on y:

$$1 + 1 = 1 + 1,$$
$$1 + a = a + 1 \rightarrow 1 + a' = a' + 1.$$

The first is a theorem by the property of equality. To prove the second, we use another general principle, viz., when a defined symbol occurs, we make use of the definition. In this particular case, we make use of the recursive definition of addition, and try to prove

$$u + 1 = u', u + v' = (u + v)', 1 + a = a + 1 \rightarrow 1 + a' = a' + 1.$$

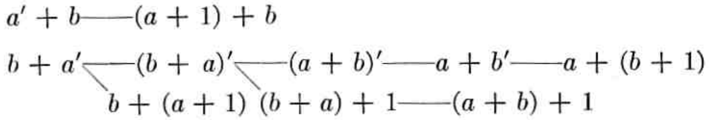
In order to derive the consequent from the antecedent, we start from $1 + a'$ and $a' + 1$, use the equalities in the antecedent to transform them, and attempt to find a chain to join them. Thus, we may try to make all possible applications of the three equalities in the antecedent:



In general, we may begin two trees simultaneously from both sides of the equality, do not write down any term which has already occurred in the same tree, and stop when a common term appears on both trees. When we get to the more complicated situations, we have to investigate two additional things. First, it would take too long to search through trees, so that it is desirable to organize available informations in forms which are more quickly accessible. Second, we may exhaust two trees and still fail to get a common term. Then we need to prove some lemma which would join up the two trees.

For example, the above graphs give us a proof of (5). To prove the other induction hypothesis, viz. (6), we may try to do the same with:

$$u + 1 = u', u + v' = (u + v)', a + b = b + a \rightarrow a' + b = b + a',$$



In this way, we have exhausted the applicable cases of the equalities in the antecedent. Since we have proved the first induction hypothesis (5), we can add it to the antecedent. Then we get some further extensions:

$$\begin{aligned} &(a + 1) + b \text{---} (1 + a) + b, \\ &b + (a + 1) \text{---} b + (1 + a), \\ &a + (b + 1) \text{---} a + (1 + b). \end{aligned}$$

At this stage, we would ask whether any other given theorem can be used to join up the two trees for $a' + b$ and $b + a'$, or, if not, what a reasonable lemma would be. If the associative law has been proved, we may observe that the missing link is supplied by:

$$(a + 1) + b = a + (1 + b). \tag{7}$$

Otherwise we should try to make a "reasonable" selection of some suitable lemma and prove it. If, for example, we have chosen (7), we would try to establish it by induction on a or on b.

It is possible that the quantifier-free theory of positive integers, including arbitrary simple recursive definitions, can be handled mechanically with relative ease, and yield fairly interesting results. The restriction to quantifier-free methods means that we are concerned only with quantifier-free theorems to be proved without using quantifiers in, e.g., applying the principle of mathematical induction. It is clear from works in the literature that this restricted domain of number theory is rather rich in content. It goes beyond logic in an essential way because of the availability of (quantifier-free) mathematical induction.

With regard to the general questions of using machines to assist mathematical research, there is a fundamental contrast between problem and method. While it seems natural to choose first the objective (e.g., number theory or geometry) and then look for methods, it is likely that a more effective approach is to let the methods lead the way. For example, since the known interesting decidable classes of formulae of the predicate calculus either do not contain function symbols or do not contain quantifiers, we are led to the simple examples above: quantifier-free number theory or function-free set theory.

REFERENCES

1. Wang, H., Proving Theorems by Pattern Recognition — I, Comm. Assoc. Comp. Mach., **3,** 1960, p. 220.
2. Surányi, J., *Reduktionstheorie des Entscheidungsproblems*, Budapest, 1959.
3. Ackermann, W., *Solvable Cases of the Decision Problem*, North-Holland, Amsterdam, 1954.
4. Skolem, T., Über die mathematische Logik, Norsk Matematisk Tidsskrift, **10,** 1928, p. 125.

5. Herbrand, J., Sur le problème fondemental de la logique mathématique, Sprawozdania z posiedzen Towarzystwa Naukowego Warszawskiego, Wydz. III, **24,** 1931, p. 12.
6. Church, A., *Introduction to Mathematical Logic*, Vol. I, Princeton Univ. Press, Princeton, N. J., 1956.
7. Church, A., Special Cases of the Decision Problem, Revue philosophique de Louvain, **49,** 1951, p. 203; **50,** 1952, p. 270.
8. Klaua, D., Systematische Behandlung der lösbaren Fälle des Entscheidungsproblems für den Prädikatenkalkül der ersten Stufe, Zeitschrift für mathematische Logik und Grundlagen der Mathematik, **1,** 1955, p. 264.
9. Dreben, B., On the Completeness of Quantification Theory, Proc. Nat. Acad. Sci. U.S.A., **38,** 1952, p. 1047.
10. Dreben, B., Systematic Treatment of the Decision Problem, Summer Institute of Symbolic Logic, Cornell Univ., 1957, p. 363.
11. Schütte, K., Untersuchungen zum Entscheidungsproblem der mathematischen Logik, Mathematische Annalen, **109,** 1934, p. 572.
12. Ackermann, W., Beiträge zum Entscheidungsproblem der mathematischen Logik, Mathematische Annalen, **112,** 1936, p. 419.
13. Wang, H., Toward Mechanical Mathematics, IBM J. Res. Dev., **4,** 1960, p. 2.
14. Hintikka, K. J. J., Vicious Circle Principle and the Paradoxes, J. Symb. Log., **22,** 1957, p. 245.

Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty — I

By D. SLEPIAN and H. O. POLLAK

(Manuscript received August 1, 1960)

A complete set of bandlimited functions is described which possesses the curious property of being orthogonal over a given finite interval as well as over $(-\infty, \infty)$. *Properties of the functions are derived and several applications to the representation of signals are made.*

I. INTRODUCTION

It is pointed out in this paper that the eigenfunctions of the finite Fourier transform are certain prolate spheroidal wave functions. These eigenfunctions properly extended possess properties that make them ideally suited for the study of certain questions regarding the relationship between functions and their Fourier transforms. Here we shall study the functions in some detail and present some applications to the representation of bandlimited functions. The property that we shall be most concerned with is the orthogonality of the functions over two different intervals. The paper[1] by Landau and Pollak which follows draws on this material, establishes other properties of the functions and provides further examples of their application.

After some definitions contained in the next section, we proceed to state without proof in Section III our main results. Certain applications of these results are then given in Section IV. The remaining sections of the paper are devoted to establishing the results already stated.

II. NOTATION

In what follows, we denote by $\mathfrak{L}_\infty^2$ the class of all complex valued functions $f(t)$ defined on the real line and integrable in absolute square. We adopt the notation

$$\| f(t) \|_A^2 = \int_{-A}^{A} |f(t)|^2 \, dt \tag{1}$$

and refer to $\| f(t) \|_\infty^2$ as *the total energy of* $f(t)$ and refer to $\| f(t) \|_A^2$ as *the energy of* $f(t)$ *in the interval* $(-A,A)$. In an analogous manner, we denote by $\mathfrak{L}_A^2$ the class of all complex valued functions $f(t)$ defined for $-A \leqq t \leqq A$ and integrable in absolute square in the interval $(-A,A)$.

Functions in $\mathfrak{L}_\infty^2$ possess Fourier transforms. Upper and lower case versions of a letter will always denote a Fourier pair. We write, for example,

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} \, d\omega, \tag{2}$$

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} \, dt. \tag{3}$$

We refer to t as *time*, ω as *angular frequency* and $\omega/2\pi$ as *frequency*. The functions $F(\omega)$ are also integrable in absolute square. In this notation Parseval's theorem is

$$\int_{-\infty}^{\infty} f(t)\overline{g(t)} \, dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega)\overline{G(\omega)} \, d\omega. \tag{4}$$

We denote by $\mathfrak{B}$ the subclass of $\mathfrak{L}_\infty^2$ consisting of those functions, $f(t)$, whose Fourier transforms, $F(\omega)$, vanish if $|\omega| > \Omega$. Here $\Omega = 2\pi W$ is a positive real number fixed throughout this paper. Every member, $f(t)$, of $\mathfrak{B}$ can be written as a finite Fourier transform of a function integrable in absolute square:

$$f(t) = \frac{1}{2\pi} \int_{-\Omega}^{\Omega} F(\omega) e^{i\omega t} \, d\omega. \tag{5}$$

Functions in $\mathfrak{B}$ are called *bandlimited* and $\mathfrak{B}$ will be referred to as *the class of bandlimited functions*. It follows from (5) that members of $\mathfrak{B}$ are entire functions of the complex variable t.

From any function $f(t)$ in $\mathfrak{L}_\infty^2$ we can obtain a function, $Bf(t)$, contained in $\mathfrak{B}$ by the rule

$$Bf(t) = \frac{1}{2\pi} \int_{-\Omega}^{\Omega} F(\omega) e^{i\omega t} \, d\omega, \tag{6}$$

where $F(\omega)$ is given by (3). We call $Bf(t)$ the *bandlimited version* of $f(t)$. We regard B as an operator whose effect on a function in $\mathfrak{L}_\infty^2$ is to produce its bandlimited version. In electrical engineering terms, $Bf(t)$ results from passing $f(t)$ through an ideal low-pass filter with angular cutoff frequency Ω.

We denote by $\mathfrak{D}$ the subclass of functions, $f(t)$, of $\mathfrak{L}_\infty^2$ each of which vanishes for $|t| > T/2$. Here T is a positive real number fixed throughout this paper. Members of $\mathfrak{D}$ are called *timelimited* and $\mathfrak{D}$ will be referred to as *the class of timelimited functions.*

From any function $f(t)$ in $\mathfrak{L}_\infty^2$ we can obtain a function $Df(t)$ contained in $\mathfrak{D}$ by the rule

$$Df(t) = \begin{cases} f(t), & |t| \leqq T/2 \\ 0, & |t| > T/2. \end{cases} \tag{7}$$

We call $Df(t)$ the *timelimited version* of $f(t)$. We regard D as an operator whose effect on a function of $\mathfrak{L}_\infty^2$ is to produce its timelimited version.

We shall use the notation $f(t) \in \mathfrak{F}$ to mean that the function $f(t)$ belongs to the class $\mathfrak{F}$ of functions.

III. RESULTS

The statements made below are proved in Sections V and VI.

Given any $T > 0$ and any $\Omega > 0$, we can find a countably infinite set of real functions $\psi_0(t),\psi_1(t),\psi_2(t),\cdots$ and a set of real positive numbers

$$\lambda_0 > \lambda_1 > \lambda_2 > \cdots \tag{8}$$

with the following properties:

i. The $\psi_i(t)$ are bandlimited, orthonormal on the real line and complete in $\mathfrak{B}$:

$$\int_{-\infty}^{\infty} \psi_i(t)\psi_j(t)\, dt = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases} \qquad i, j = 0, 1, 2, \cdots. \tag{9}$$

ii. In the interval $-T/2 \leqq t \leqq T/2$, the $\psi_i(t)$ are orthogonal and complete in $\mathfrak{L}_{T/2}^2$:

$$\int_{-T/2}^{T/2} \psi_i(t)\psi_j(t)\, dt = \begin{cases} 0, & i \neq j \\ \lambda_i, & i = j \end{cases} \qquad i, j = 0, 1, 2, \cdots. \tag{10}$$

iii. For all values of t, real or complex,

$$\lambda_i\psi_i(t) = \int_{-T/2}^{T/2} \frac{\sin \Omega(t-s)}{\pi(t-s)}\, \psi_i(s)\, ds, \qquad i = 0, 1, 2, \cdots. \tag{11}$$

Further properties of the ψ's are given in Sections V and VI.

The notation used above conceals the fact that both the ψ's and the λ's are functions of the product ΩT. When it is necessary to make this dependence explicit, we write $\lambda_i = \lambda_i(c)$, $\psi_i(t) = \psi_i(c,t)$, $i = 0,1,2,\cdots$, where $2c = \Omega T$.

Some values of $\lambda_i(c)$ are given in Table I. It is to be noted that for a fixed value of c the λ_i fall off to zero rapidly with increasing i once i has

TABLE I—VALUES OF $\lambda_n(c) = L_n(c) \times 10^{-p_n(c)}$

n	$c = 0.5$		$c = 1.0$		$c = 2.0$		$c = 4.0$		$c = 8.0$	
	L	p	L	p	L	p	L	p	L	p
0	3.0969	1	5.7258	1	8.8056	1	9.9589	1	1.0000	0
1	8.5811	3	6.2791	2	3.5564	1	9.1211	1	9.9988	1
2	3.9175	5	1.2375	3	3.5868	2	5.1905	1	9.9700	1
3	7.2114	8	9.2010	6	1.1522	3	1.1021	1	9.6055	1
4	7.2714	11	3.7179	8	1.8882	5	8.8279	3	7.4790	1
5	4.6378	14	9.4914	11	1.9359	7	3.8129	4	3.2028	1
6	2.0413	17	1.6716	13	1.3661	9	1.0951	5	6.0784	2
7	6.5766	21	2.1544	16	7.0489	12	2.2786	7	6.1263	3
8	1.6183	24	2.1207	19	2.7768	14	3.6066	9	4.1825	4

exceeded $(2/\pi)c$. (The significance of this will be discussed in detail in a later paper.) Because of (9) and (10), namely $\| \psi_i \|_\infty^2 = 1$, $\| \psi_i \|_{T/2}^2 = \lambda_i$, a small value of λ_i implies that $\psi_i(t)$ will have most of its energy outside the interval $(-T/2,T/2)$ whereas a value of λ_i near 1 implies that $\psi_i(t)$ will be concentrated largely in $(-T/2,T/2)$. This behavior of the ψ's can be clearly seen in Figs. 1 through 5. Figs. 1 through 4 show $\psi_0(c,t)$, $\psi_1(c,t)$, $\psi_2(c,t)$ and $\psi_3(c,t)$ for several different values of c. For $c = 0.5$, or $(2/\pi)c = 0.3183$, as shown on Fig. 1, ψ_2 and ψ_3 are practically zero in the interval $(-T/2,T/2)$. For $c = 4$, or $(2/\pi)c = 2.546$, as shown on Fig. 4, ψ_0 is largely concentrated in the interval $(-T/2, T/2)$. Fig. 5 compares $\psi_0(c,t)$ for several different values of c.

IV. SOME APPLICATIONS

4.1 *Extrapolation of a Bandlimited Function*

It is sometimes desired to extrapolate a bandlimited function known only on the interval $(-T/2,T/2)$ to values outside this interval. Since any $f \in \mathcal{B}$ is an entire function, this extrapolation can be done exactly in principle. One could, for example, calculate successive derivatives of f at some point in $(-T/2,T/2)$ and form a Taylor series representation which would converge everywhere. In practice, however, such a Taylor series would necessarily be truncated and the resultant approximation to $f(t)$ would be a polynomial which for sufficiently large values of $|t|$ would give a very poor approximation to f. This approximation is not, of course, bandlimited.

The functions ψ_i provide an alternative approach. Since $f \in \mathcal{B}$, we can write, from i., for all t

$$f(t) = \sum_0^\infty a_n \psi_n(t), \tag{12}$$

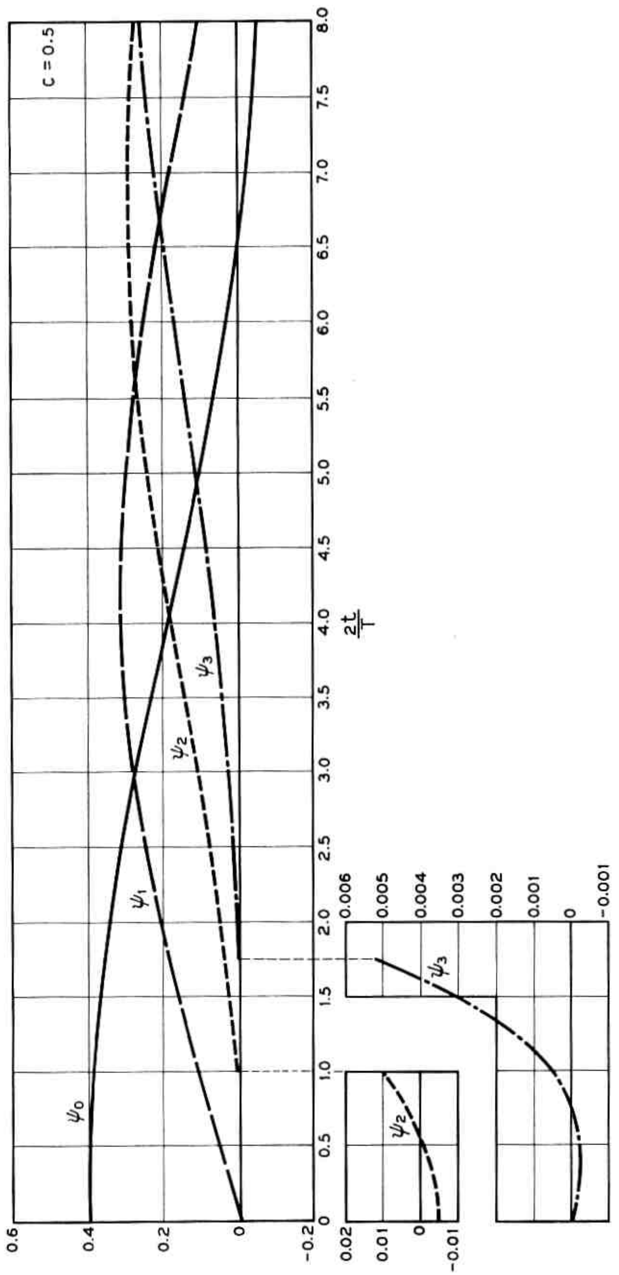


Fig. 1 — $\psi_0(t)$, $\psi_1(t)$, $\psi_2(t)$, $\psi_3(t)$ vs. $2t/T$ for $c = 0.5$.

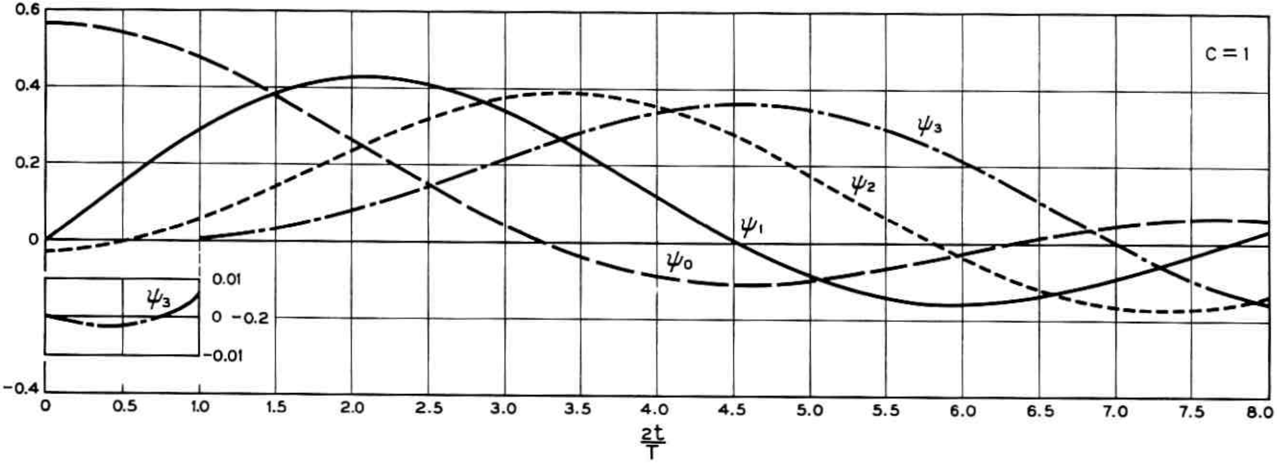


Fig. 2 — $\psi_0(t)$, $\psi_1(t)$, $\psi_2(t)$, $\psi_3(t)$ vs. $2t/T$ for $c = 1.0$.

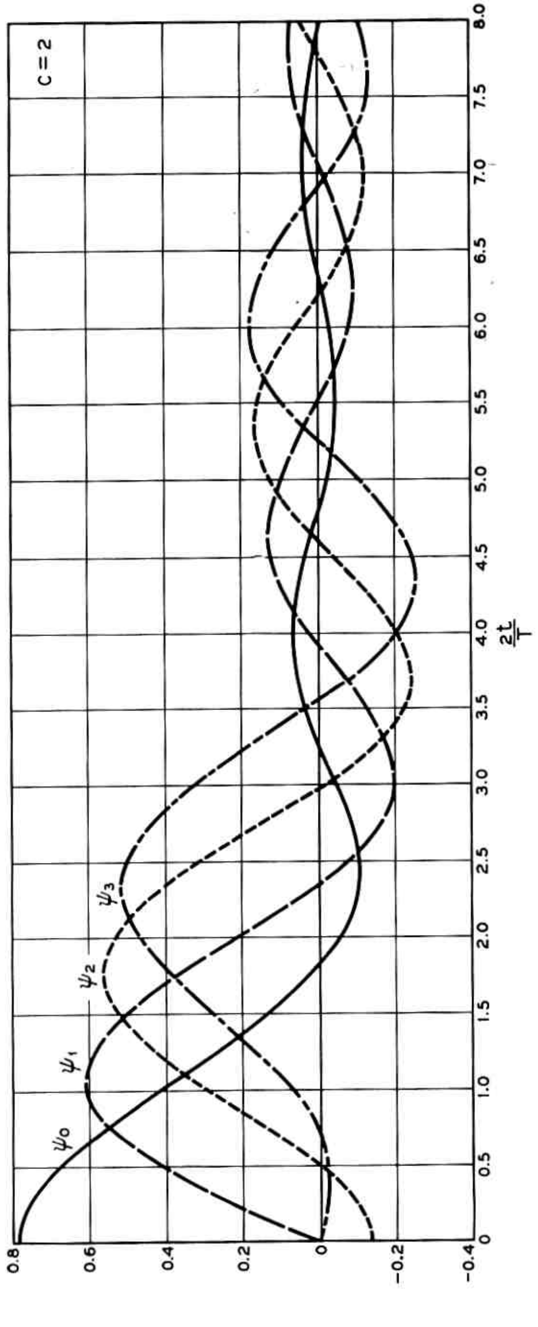


Fig. 3 — $\psi_0(t)$, $\psi_1(t)$, $\psi_2(t)$, $\psi_3(t)$ vs. $2t/T$ for $c = 2.0$.

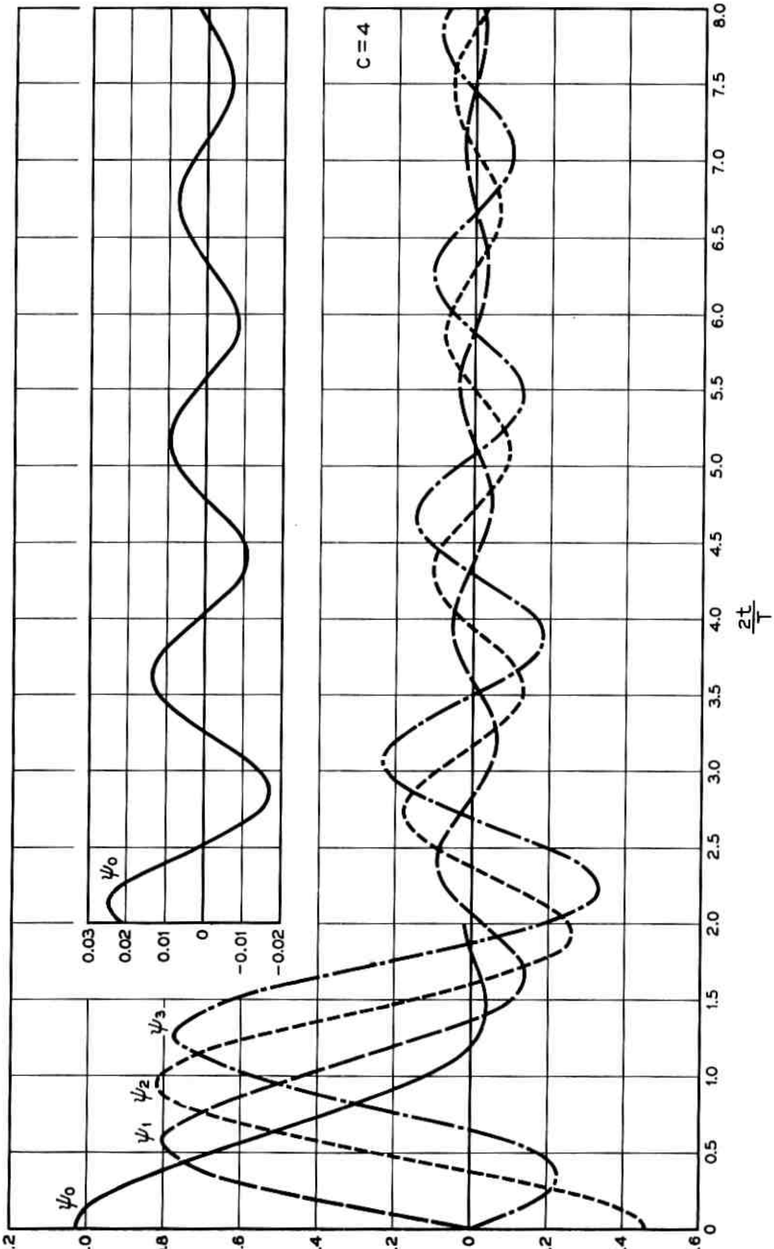


Fig. 4 — $\psi_0(t)$, $\psi_1(t)$, $\psi_2(t)$, $\psi_3(t)$ vs. $2t/T$ for $c = 4.0$.

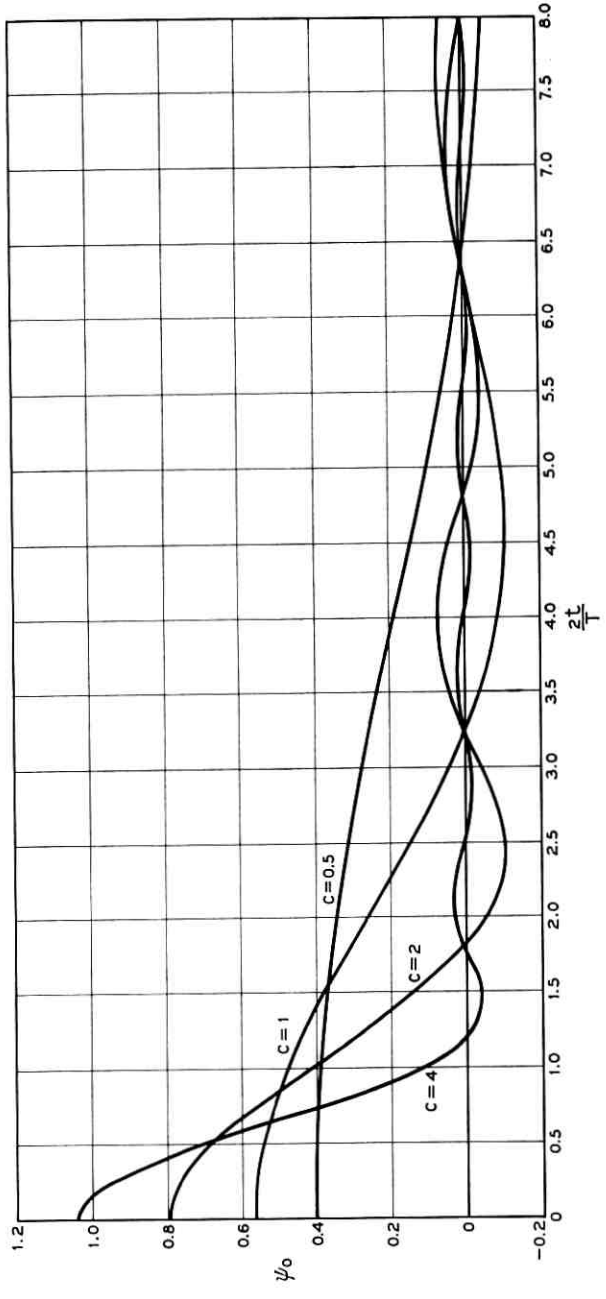


Fig. 5 — $\psi_0(c,t)$ vs. $2t/T$ for $c = 0.5, 1.0, 2.0, 4.0$.

where

$$\begin{aligned} a_n &= \int_{-\infty}^{\infty} f(t)\psi_n(t)\,dt, \\ \sum_0^{\infty} a_n^2 &= \int_{-\infty}^{\infty} f(t)^2\,dt \end{aligned} \tag{13}$$

and the convergence in (12) is in the mean square sense

$$\lim_{N\to\infty} \int_{-\infty}^{\infty} \left[f(t) - \sum^{N} a_n\psi_n(t) \right]^2 dt = 0.$$

Multiply (12) by $\psi_j(t)$, integrate and use (10). There results

$$a_n = \frac{1}{\lambda_n} \int_{-T/2}^{T/2} f(t)\psi_n(t)\,dt. \tag{14}$$

The coefficients in (12) *can be determined by* (14) *from values of $f(t)$ in the interval* $(-T/2,T/2)$.

The above result suggests approximating $f(t)$ for all t by

$$f_N(t) = \sum_0^N a_n\psi_n(t) \tag{15}$$

with the a_n given by (14). The approximation (15) is itself bandlimited. The mean squared error is

$$\int_{-\infty}^{\infty} [f(t) - f_N(t)]^2\,dt = \sum_{N+1}^{\infty} a_n^2 \tag{16}$$

and by (13) can be made as small as desired by making N sufficiently large. In the sense of (16), the extrapolation remains good for all t. The error in the fit of f_N to f in $(-T/2,T/2)$ is given by

$$\int_{-T/2}^{T/2} (f - f_N)^2\,dt = \sum_{N+1}^{\infty} a_n^2\lambda_n\,. \tag{17}$$

As the λ_n approach zero rapidly for sufficiently large n, it may happen that (17) is small for values of N for which (16) is still large. The fit of f_N inside the interval should not be taken as an indication of the fit elsewhere.

4.2 *Approximation in an Interval by a Bandlimited Function*

Suppose now $f(t) \in \mathfrak{L}_{T/2}^2$ is known in the interval $(-T/2,T/2)$ but f is not necessarily a piece of a bandlimited unction. From i. above it follows that $f(t)$ may still be represented by (12) with a's given by (14), but this representation is valid now only for $|t| \leqq T/2$. If indeed f is not a piece of a bandlimited function, the series (12) will certainly not converge in mean square over the whole real line.

The foregoing suggests the utility of finite sums of the form (15) as approximants to bandlimited functions having a prescribed form in the interval $(-T/2,T/2)$. The conditions of bandlimitation and prescribed form in $(-T/2,T/2)$ are, of course, in general incompatible (unless indeed, the prescribed form is a piece of a bandlimited function). However, finite sums of the form (15) taken for all t with a's computed by (14) permit approximations by bandlimited functions to a prescribed $f \in \mathfrak{L}_{T/2}^{2}$. We are assured by ii. that the approximation can be made as good as desired in the sense that the right side of (17) approaches zero for large N. We have, however,

$$\int_{-\infty}^{\infty} f_N^{2}(t)\, dt = \sum_{0}^{N} a_n^{2}$$

and, if f is not a piece of a bandlimited function, $\sum^{N} a_n^{2}$ grows without bound for increasing N. Thus, in approximating a piece of a nonbandlimited function by a bandlimited function, we exchange goodness of fit in $(-T/2,T/2)$ with wildness of behavior outside this interval.

We now impose an energy restriction. Given $f \in \mathfrak{L}_{T/2}^{2}$. What $g \in \mathfrak{B}$ with prescribed energy $\| g \|_{\infty}^{2} = E$ minimizes $\| f - g \|_{T/2}^{2}$? Let

$$f = \sum a_n \psi_n(t), \qquad |t| \leqq T/2,$$
$$g = \sum b_n \psi_n(t), \qquad -\infty < t < \infty.$$

Then a simple argument gives

$$b_n = \frac{a_n \lambda_n}{\mu + \lambda_n},$$

where μ is the unique positive number which satisfies

$$E = \sum \frac{a_n^{2} \lambda_n^{2}}{(\mu + \lambda_n)^2}.$$

If the constraint on g is that the energy outside $(-T/2,T/2)$ is prescribed, $\| g \|_{\infty}^{2} - \| g \|_{T/2}^{2} = E'$, rather than the total energy, the result is

$$b_n = \frac{a_n \lambda_n}{\mu(1 - \lambda_n) + \lambda_n},$$

where μ (again positive) is chosen to satisfy

$$E' = \sum \frac{a_n^2 \lambda_n^2}{[\mu(1 - \lambda_n) + \lambda_n]^2}.$$

4.3 *Some Extremal Properties of $\psi_0(t)$*

The ψ's possess a number of interesting extremal properties. The most important of these, the fact that ψ_0 has the largest energy in $(-T/2, T/2)$ of all function in $\mathcal{B}$ of unit total energy, is discussed in detail by Landau and Pollak.[1] We comment here on two other extremal properties of ψ_0.

Let $f(t) \in \mathcal{L}_\infty^2$ have total energy $E = \| f \|_\infty^2$. The timelimited version of $f(t)$ has total energy $E_D = \| Df \|_\infty^2 = \| f \|_{T/2}^2 \leqq E$. Since Df cannot be bandlimited, its Fourier transform has nonvanishing energy in $| \omega | > \Omega$. The bandlimited version of Df, namely BDf, will therefore have total energy $E_{BD} < E_D \leqq E$. The operation BD transforms a member of $\mathcal{L}_\infty^2$ into a member of $\mathcal{B}$ with smaller total energy. Which members of $\mathcal{L}_\infty^2$ lose the smallest fraction of their energy under such a transformation? That is, for which $f \in \mathcal{L}_\infty^2$ is $\mu \equiv \| BDf \|_\infty^2 / \| f \|_\infty^2$ a maximum?

The answer to this question, unique except for an arbitrary multiplicative constant, is $D\psi_0(t)$. This may be seen as follows. From (3), (6) and the definition (7) of D,

$$\begin{aligned} BDf(t) &= \frac{1}{2\pi} \int_{-\Omega}^{\Omega} d\omega \, e^{i\omega t} \int_{-T/2}^{T/2} ds \, f(s) \, e^{-i\omega s} \\ &= \int_{-T/2}^{T/2} \rho_\Omega(t - s) f(s) \, ds, \end{aligned} \tag{18}$$

where we have written

$$\rho_\Omega(\tau) = \frac{\sin \Omega\tau}{\pi\tau} = \frac{1}{2\pi} \int_{-\Omega}^{\Omega} d\omega \, e^{i\omega\tau}. \tag{19}$$

Note that $\rho_\Omega(\tau)$ is an even function of τ and that from (19) and Parseval's theorem (4) it follows that

$$\int_{-\infty}^{\infty} \rho_\Omega(t - u) \rho_\Omega(u - s) \, du = \rho_\Omega(t - s). \tag{20}$$

Therefore,

$$\begin{aligned}\| BDf(u)\|_\infty^2 &= \int_{-\infty}^{\infty} du\,[BDf(u)]\overline{[BDf(u)]} \\ &= \int_{-\infty}^{\infty} du \int_{-T/2}^{T/2} dt \int_{-T/2}^{T/2} ds\,\rho_\Omega(t-u)\rho_\Omega(u-s)f(t)\bar{f}(s) \qquad (21) \\ &= \int_{-T/2}^{T/2} dt \int_{-T/2}^{T/2} ds\,\rho_\Omega(t-s)\,f(t)\,\bar{f}(s).\end{aligned}$$

Here we have used (20) and the fact that ρ_Ω is real and even.

Since from (21) we see that $\| BDf \|_\infty^2$ depends only on values of f in $(-T/2,T/2)$, it follows that μ is equal to the maximum of

$$\nu = \frac{\| BDf \|_\infty^2}{\| f \|_{T/2}^2} = \frac{\int_{-T/2}^{T/2} dt \int_{-T/2}^{T/2} ds\,\rho_\Omega(t-s)\,f(t)\,\bar{f}(s)}{\int_{-T/2}^{T/2} |\,f(t)\,|^2\,dt}$$

over all $f \in \mathfrak{L}_{T/2}^2$. It is well known that the solution to this problem is $\nu = \lambda_0$, where λ_0 is the largest eigenvalue of the integral equation

$$\lambda f(t) = \int_{-T/2}^{T/2} \rho_\Omega(t-s)\,f(s)\,ds, \qquad |\,t\,| \leqq T/2, \qquad (22)$$

and that ν attains the value λ_0 for f equal to a corresponding eigenfunction. We shall see later that ψ_0 is such an eigenfunction. Thus f agrees with ψ_0 in $(-T/2,T/2)$ and so $D\psi_0$ is a function in $\mathfrak{L}_\infty^2$ for which μ attains it́s maximum value λ_0.

We now ask which $f \in \mathcal{B}$ as opposed to $f \in \mathfrak{L}_\infty^2$ maximizes μ. That is, which *bandlimited* function loses the least (fractional) energy when first timelimited then bandlimited? The answer is ψ_0 and the corresponding value of μ is λ_0^2.

To see this, introduce the representation (5) for $f \in \mathcal{B}$ into the numerator [as given by (21)] of μ. There results

$$\| BDf \|_\infty^2 = \frac{1}{4\pi^2} \int_{-\Omega}^{\Omega} d\omega \int_{-\Omega}^{\Omega} d\omega'\,F(\omega)\overline{F(\omega')}K(\omega,\omega'),$$

where we have set

$$K(\omega,\omega') = \int_{-T/2}^{T/2} dt \int_{-T/2}^{T/2} dt'\,\rho_\Omega(t-t')e^{i\omega t}\,e^{-i\omega' t'}.$$

To transform this expression further, introduce the representation (19) to obtain

$$\begin{aligned} K(\omega,\omega') &= \frac{1}{2\pi}\int_{-\Omega}^{\Omega} d\omega'' \int_{-T/2}^{T/2} dt\, e^{it(\omega-\omega'')} \int_{-T/2}^{T/2} dt'\, e^{-it'(\omega'-\omega'')} \\ &= 2\pi \int_{-\Omega}^{\Omega} d\omega'' \rho_{T/2}(\omega-\omega'')\rho_{T/2}(\omega''-\omega') \\ &\equiv 2\pi\rho_{T/2}{}^{(2)}(\omega,\omega'). \end{aligned}$$

By Parseval's theorem, (4), the denominator of μ can be written as

$$\| f \|_\infty^2 = \frac{1}{2\pi}\int_{-\Omega}^{\Omega} | F(\omega) |^2 \, d\omega.$$

Our task, then, is to maximize

$$\mu = \frac{\| BDf \|_\infty^2}{\| f \|_\infty^2} = \frac{\int_{-\Omega}^{\Omega} d\omega \int_{-\Omega}^{\Omega} d\omega' \rho_{T/2}{}^{(2)}(\omega,\omega')F(\omega)\bar{F}(\omega')}{\int_{-\Omega}^{\Omega} | F(\omega) |^2 \, d\omega}$$

over all $F \in \mathfrak{L}_\Omega^2$. The solution to this problem is $\mu = \mu_0$, where μ_0 is the largest eigenvalue of the integral equation

$$\lambda F(\omega) = \int_{-\Omega}^{\Omega} \rho_{T/2}{}^{(2)}(\omega,\omega')F(\omega')\, d\omega'.$$

Now $\rho_{T/2}{}^{(2)}(\omega,\omega')$ is the first iterate of $\rho_{T/2}(\omega - \omega')$. Therefore, μ_0 is the square of the largest eigenvalue of the integral equation

$$\lambda F(\omega) = \int_{-\Omega}^{\Omega} \rho_{T/2}(\omega - \omega')F(\omega')\, d\omega'.$$

A change of variables reduces this equation to the form of (22) whence it is seen that $\mu = \lambda_0^2$ and that $F(\omega) = \psi_0(\omega T/2\Omega)$ for $| \omega | \leqq \Omega$. From (29), which will be established later, it follows that $f(t) = \psi_0(t)$.

4.4 *Problems Concerning Bandlimited Noise*

Much of the theory of detection, parameter estimation and prediction of signals in noise when observations are made in a finite time is based on the Karhunen-Loève representation of the noise. (See Ref. 2 for such a treatment of these problems.) This representation involves expansions in terms of the eigenfunction solutions of a certain integral equation. When the noise in question is second order stationary and with angular frequency spectral density uniform in $(-\Omega,\Omega)$ and zero elsewhere (bandlimited white noise), the integral equation in question is identical with (19), (22). The function ψ_i and eigenvalues λ_i thus play an important role in numerous questions concerning bandlimited white noise observed for a finite time. Their role in this connection has been pointed out previously in Ref. 3.

V. THE PROLATE SPHEROIDAL WAVE FUNCTION

The functions $\psi_i(c,t)$ are scaled versions of certain of the angular prolate spheroidal wave functions. A number of books[4,5,6,7] treat the prolate spheroidal wave functions in detail. We will draw freely from this literature. We adopt the notation* of Flammer.[4]

When c is real, the differential equation

$$(1 - t^2) \frac{d^2u}{dt^2} - 2t \frac{du}{dt} + (\chi - c^2t^2)u = 0 \tag{23}$$

has continuous solutions in the closed t interval $[-1,1]$ only for certain discrete real positive values $0 < \chi_0(c) < \chi_1(c) < \chi_2(c) < \cdots$ of the parameter χ. Corresponding to each eigenvalue $\chi_n(c)$, $n = 0,1,2, \cdots$ there is a unique solution $S_{0n}(c,t)$ such that $S_{0n}(c,0) = P_n(0)$ where $P_n(t)$ is the nth Legendre polynomial. The functions $S_{0n}(c,t)$ are called *angular prolate spheroidal functions.* They are real for real t, are continuous functions of c for $c \geqq 0$, and can be extended to be entire functions of the complex variable t. They are orthogonal in $(-1,1)$ and are complete in $\mathfrak{L}_1^2$. $S_{0n}(c,t)$ has exactly n zeros in $(-1,1)$, reduces to $P_n(t)$ uniformly in $[-1,1]$ as $c \rightarrow 0$, and is even or odd according as n is even or odd, $n = 0,1,2, \cdots$. The eigenvalues $\chi_n(c)$ are continuous functions of c and $\chi_n(0) = n(n + 1)$, $n = 0,1,2, \cdots$.

A second set of solutions $R_{0n}^{(1)}(c,t)$, $n = 0, 1, \cdots$, called *radial prolate spheroidal functions,* which differ from the angular functions only by a real scale factor,

$$R_{0n}^{(1)}(c,t) = k_n(c) S_{0n}(c,t),$$

are of use in many applications. These radial functions are normalized so that

$$R_{0n}^{(1)}(c,t) \rightarrow \frac{1}{ct} \cos [ct - \tfrac{1}{2}(n + 1)\pi]$$

as $t \rightarrow \infty$.

The equations

$$\frac{2c}{\pi} [R_{0n}^{(1)}(c,1)]^2 S_{0n}(c,t) = \int_{-1}^{1} \frac{\sin c(t - s)}{\pi(t - s)} S_{0n}(c,s)\, ds, \tag{24}$$

$$2i^n R_{0n}^{(1)}(c,1) S_{0n}(c,t) = \int_{-1}^{1} e^{icts} S_{0n}(c,s)\, ds \qquad n = 0, 1, 2, \cdots \tag{25}$$

* The reader should be cautioned that various authors disagree not only on notation for these functions, but also in their method of normalization.

are both special cases of more general integral relations satisfied by prolate spheroidal functions that can be found in the literature. They are valid for all t, real or complex.

Equation (24) shows that $S_{0n}(c,t)$ is a solution of the integral equation

$$\lambda f(t) = \int_{-1}^{1} \rho_c(t - s)f(s)\,ds, \qquad |t| \leqq 1 \tag{26}$$

corresponding to the eigenvalue

$$\lambda_n(c) = \frac{2c}{\pi}[R_{0n}^{(1)}(c,1)]^2, \qquad n = 0, 1, 2, \cdots. \tag{27}$$

Here $\rho_c(\tau)$ is given by (19). Indeed, the completeness of the S_{0n} in $\mathfrak{L}_1^2$ assures us that the quantities (27) are the only eigenvalues of (26) and that if these quantities are distinct, the S_{0n} are (apart from multiplicative constants) the unique $\mathfrak{L}_1^2$ solutions of (26). If several of the quantities (27) are equal for different values of n, then linear combinations of the corresponding S_{0n} will also satisfy (26). Within the sense of this degeneracy, then, the S_{0n} are unique solutions of (26). In Section VI we shall see, indeed, that this degeneracy does not occur.

Equation (19) and Bochner's theorem (Ref. 8, Theorem 23, p. 95) show that the kernel of (26) is positive definite. The quantities (27) are therefore strictly positive. Set

$$[u_n(c)]^2 = \int_{-1}^{1} [S_{0n}(c,t)]^2\,dt.$$

We now finally define

$$\psi_n(c,t) = \frac{\sqrt{\lambda_n(c)}}{u_n(c)} S_{0n}(c,2t/T). \tag{28}$$

Properties ii. of Section III now follow directly from definitions and the orthonormality and completeness of the S_{0n} in $(-1,1)$.

A change of variables and the definitions (27) and (28) convert (24) into (11). A change of variables converts (25) into

$$\frac{i^n \Omega R_{0n}^{(1)}(c,1)}{\pi} \psi_n(c,t) = \frac{1}{2\pi}\int_{-\Omega}^{\Omega} e^{i\omega t}\psi_n(c,\omega T/2\Omega)\,d\omega, \tag{29}$$

which shows $\psi_n \in \mathfrak{B}$. Indeed, since the function $\psi_n(c,\omega T/2\Omega)$ are complete in $-\Omega \leqq \omega \leqq \Omega$, Parseval's theorem shows that the $\psi_n(t)$ are complete in $\mathfrak{B}$. The remaining assertion of i. of Section III, namely (9), follows from a computation. From (11) we have

$$\begin{aligned}
&\int_{-\infty}^{\infty} dt\,\psi_i(t)\psi_j(t)\\
&\quad = \frac{1}{\lambda_i\lambda_j}\int_{-\infty}^{\infty} dt \int_{-T/2}^{T/2} ds\,\rho_\Omega(t-s)\psi_i(s)\int_{-T/2}^{T/2} du\,\rho_\Omega(t-u)\psi_j(u)\\
&\quad = \frac{1}{\lambda_i\lambda_j}\int_{-T/2}^{T/2} du \int_{-T/2}^{T/2} ds\,\psi_i(s)\psi_j(u)\int_{-\infty}^{\infty} dt\,\rho_\Omega(u-t)\rho_\Omega(t-s)\\
&\quad = \frac{1}{\lambda_i\lambda_j}\int_{-T/2}^{T/2} du\,\psi_j(u)\int_{-T/2}^{T/2} ds\,\rho_c(u-s)\psi_i(s)\\
&\quad = \frac{1}{\lambda_j}\int_{-T/2}^{T/2} du\,\psi_j(u)\psi_i(u) = \begin{cases}0, & i\neq j\\ 1, & i=j\end{cases}.
\end{aligned}$$

Here we have used (20) and (10).

All properties of the ψ's asserted in Section III have now been established except for (8). To this end we devote the next section.*

VI. NONDEGENERACY AND ORDERING OF THE EIGENVALUES OF (26)

We have seen that the $S_{0n}(c,t)$ are solutions of (26) with eigenvalues given by (27). We show now that we cannot have two distinct S_{0n} belonging to the same eigenvalue λ if $c > 0$.

Let $f_1(t)$ and $f_2(t)$ be two linearly independent solutions of (26) for the same λ, $c \neq 0$. Then

$$\lambda f_1(t) = \int_{-1}^{1} \rho_c(t-s)f_1(s)\,ds, \tag{30}$$

$$\lambda f_1'(t) = \int_{-1}^{1} \rho_c'(t-s)f_1(s)\,ds, \tag{31}$$

$$\lambda f_1''(t) = \int_{-1}^{1} \rho_c''(t-s)f_1(s)\,ds, \tag{32}$$

$$\lambda f_2(t) = \int_{-1}^{1} \rho_c(t-s)f_2(s)\,ds \tag{33}$$

* Ville and Bouzitat[9] recognized (independently of the earlier Ref. 3) that the solutions of the integral equation (11) are prolate spheroidal functions. They assert that the eigenvalues λ_n are ordered as in (8) when ψ_n is identified with S_{0n} but no proof of this fact appears in their paper or apparently elsewhere in the literature.

and

$$\lambda f_2''(t) = \int_{-1}^{1} \rho_c''(t-s) f_2(s)\, ds. \tag{34}$$

Assume now that f_1 is even and f_2 is odd. Integrate (31) by parts to obtain

$$\lambda f_1'(t) = f_1(1)[\rho_c(-1-t) - \rho_c(1-t)] + \int_{-1}^{1} \rho_c(t-s) f_1'(s)\, ds.$$

Multiply this equation by $f_2(t)$ and integrate to obtain

$$\begin{aligned}\lambda \int_{-1}^{1} f_2(t) f_1'(t)\, dt = {} & \lambda f_1(1)[f_2(-1) - f_2(1)] \\ & + \int_{-1}^{1} dt \int_{-1}^{1} ds\, \rho_c(t-s) f_1'(s) f_2(t).\end{aligned} \tag{35}$$

Now multiply (33) by $f_1'(t)$, integrate and subtract the result from (35). One finds $\lambda f_1(t)[f_2(-1) - f_2(1)] = 0$, or

$$f_1(1) f_2(1) = 0, \qquad f_1 \text{ even}, \quad f_2 \text{ odd}. \tag{36}$$

Assume now that $f_1(t)$ and $f_2(t)$ are of the same parity, i.e., both even or both odd. Multiply (32) by $f_2(t)$, multiply (34) by $f_1(t)$, subtract and integrate. There results

$$\begin{aligned}\lambda \int_{-1}^{1} dt (f_1'' f_2 - f_2'' f_1) &= \lambda \int_{-1}^{1} dt \frac{d}{dt} (f_1' f_2 - f_2' f_1) \\ &= 2\lambda [f_1'(1) f_2(1) - f_2'(1) f_1(1)] = 0\end{aligned}$$

or

$$f_1(1) f_2'(1) = j_2(1) f_1'(1), \qquad f_1 \text{ and } f_2 \text{ of same parity}. \tag{37}$$

For any two linearly independent solutions of (26) belonging to the same eigenvalue we must have either (36) or (37) hold. But we shall show that both of these conditions are impossible for two different S functions, say $S_{0n}(c,t)$ and $S_{0m}(c,t)$. From the differential equation (23), we see that

$$2S_{0n}'(1) = (\chi_n - c^2) S_{0n}(1). \tag{38}$$

If $S_{0n}(1)$ vanishes, then so does $S_{0n}'(1)$. But differentiating (23) shows that if $S_{0n}(1)$ and $S'_{0n}(1)$ vanish so does $S''_{0n}(1)$. Repeated differentiation (which is possible since the S_{0n} are entire) shows that if $S_{0n}(1) = 0$, then $S_{0n}(t) \equiv 0$. Therefore condition (36) cannot hold. On the other hand, since $S_{0n}(1) \neq 0$, $S_{0m}(1) \neq 0$, (37) can be written

$$\frac{S'_{0m}(1)}{S_{0m}(1)} = \frac{S'_{0n}(1)}{S_{0n}(1)}$$

or

$$\frac{\chi_m - c^2}{2} = \frac{\chi_n - c^2}{2} \tag{39}$$

from (38). However, it is known that the eigenvalues of the differential equation (23) are nondegenerate if c is real, so that (39) cannot hold if $m \neq n$. The eigenvalues (27) are thus seen to be distinct.

By their definition, the S_{0n} functions are indexed so that the eigenvalues of the differential equation (23) $\chi_0 < \chi_1 < \chi_2 < \cdots$ are monotone increasing functions of their index. We have defined ψ_n in terms of the S_{0n} by (28) and have labeled the corresponding eigenvalue of (26) λ_n by (27). There remains the task of proving that the λ_n are ordered as in (8).

Our argument makes use of the fact (just demonstrated) that for all real $c \neq 0$ the $\lambda_n(c)$ are nondegenerate and the fact (see for example Ref. 10, vol. I, p. 128) that the eigenfunctions and eigenvalues of (26) are continuous functions of its kernel. Thus if we can prove that for some $c > 0$,

$$\lambda_0(c) > \lambda_1(c) > \lambda_2(c) \cdots,$$

then continuity and nondegeneracy of the λ's allows us to assert this ordering for all positive c.

We now establish this ordering for c sufficiently near zero. Let ψ_n and ψ_{n+1} be successive eigenfunctions of (26), $c \neq 0$. Then

$$\lambda_n \psi'_n(t) = \int_{-1}^{1} \rho'_c(t - s)\psi_n(s)\, ds,$$

$$\lambda_{n+1}\psi'_{n+1}(t) = \int_{-1}^{1} \rho'_c(t - s)\psi_{n+1}(s)\, ds.$$

Multiply the first of these equations by $\lambda_{n+1}\psi_{n+1}(t)$, multiply the second by $\lambda_n\psi_n(t)$, add the results and integrate to obtain

$$\begin{aligned}\lambda_n\lambda_{n+1}\int_{-1}^{1}(\psi_n'\psi_{n+1}+\psi_{n+1}'\psi_n)\,dt \\ &= \lambda_{n+1}\int_{-1}^{1}dt\int_{-1}^{1}ds\;\rho_c'(t-s)\psi_{n+1}(t)\psi_n(s) \\ &\qquad + \lambda_n\int_{-1}^{1}dt\int_{-1}^{1}ds\;\rho_c'(t-s)\psi_n(t)\psi_{n+1}(s) \\ &= (\lambda_n-\lambda_{n+1})\int_{-1}^{1}dt\int_{-1}^{1}ds\;\rho_c'(t-s)\psi_n(t)\psi_{n+1}(s) \\ &= (\lambda_n-\lambda_{n+1})\lambda_{n+1}\int_{-1}^{1}\psi_n(t)\psi_{n+1}'(t)\,dt\end{aligned}$$

or

$$\lambda_n-\lambda_{n+1}=\lambda_n\left(1+\frac{\int_{-1}^{1}\psi_n'\psi_{n+1}\,dt}{\int_{-1}^{1}\psi_n\psi_{n+1}'\,dt}\right). \tag{40}$$

Now as $c \to 0$, $\psi_n \to P_n(t)$, the nth Legendre polynomial, and $\psi_n' \to P_n'(t)$. The denominator of the fraction in (40) approaches

$$\int_{-1}^{1}P_nP_{n+1}'\,dt = P_nP_{n+1}\Big|_{-1}^{1}-\int_{-1}^{1}P_{n+1}P_n'\,dt = 2$$

since the integral on the right vanishes and $P_n(1) = 1$. The numerator approaches

$$\int_{-1}^{1}P_n'P_{n+1}\,dt = 0.$$

By making c sufficiently small, therefore, the fraction on the right of (40) is of absolute value less than unity and $\lambda_n - \lambda_{n+1} = \lambda_n[1 + 0(1)] \geqq 0$. Since for $c \neq 0$ the λ_n are all distinct and positive, the ordering (8) must hold. The limiting eigenvalues for $c \to 0$ are $0 = \lambda_0 = \lambda_1 = \lambda_2 = \cdots$.

VII. COMMENTS

It is worth pointing out that the basic importance of the ψ_n for the study of the relation between functions and their Fourier transforms stems from (25), which shows that the S_{0n} are eigenfunctions of the finite Fourier transform kernel. Indeed, many of the important properties of the ψ's (i. and ii. of Section III, for example) follow directly from (25) or its first iterate (24), without explicit use of (23) or recognition of the S_{0n} as angular prolate spheroidal wave functions.

In the interests of simplicity of presentation, we have not put forth the theme of this work in its most general form. We here make just one comment in this direction and leave other generalizations to the interested reader. The curious orthogonality over two different pointsets of the analytically continued solution of (22) will hold whenever (20) is true and the solutions are in $\mathfrak{L}_\infty^2$. For example, if the kernel $\rho(\tau)$ of (22) is even and has a Fourier transform constant on intervals and zero elsewhere, e.g., $\rho_1(\tau) = \rho_\Omega(\tau) \cos \alpha\tau$, $\alpha > \Omega$, then the double orthogonality maintains. The eigenfunctions for the bandpass kernel $\rho_1(\tau)$ do not seem to be expressible in terms of well-studied functions. Computations in this case indicate the existence of degenerate eigenvalues.

REFERENCES

1. Landau, H. J. and Pollak, H. O., this issue, pp. 65–84.
2. Davenport, W. B. and Root, W. L., *Random Signals and Noise*, McGraw-Hill, New York, 1958.
3. Slepian, D., Estimation of Signal Parameters in the Presence of Noise, IRE Trans., **PGIT-3,** 1954, pp. 68–89.
4. Flammer, C., *Spheroidal Wave Functions*, Stanford University Press, Stanford, Calif., 1957.
5. Stratton, J. A., Morse, P. M., Chu, L. J., Little, J. D. C. and Corbató, F. J., *Spheroidal Wave Functions*, The Technology Press of M.I.T., Cambridge, Mass., and John Wiley and Sons, New York, 1956.
6. Meixner, J. and Schäfke, F. W., *Mathieusche Funktionen und Sphäroidfunktionen*, Julius Springer, Berlin, 1954.
7. Morse, P. M. and Feshbach, H., *Methods of Theoretical Physics*, McGraw-Hill, New York, 1953.
8. Bochner, S., *Lectures on Fourier Integrals*, Annals of Mathematics Studies No. 42, Princeton University Press, Princeton, N. J., 1959.
9. Ville, J. A., and Bouzitat, J., Note sur un signal de durée finie et d'energie filtrée maximum, Cables & Trans., **11,** 1957, pp. 102–127.
10. Courant, R., and Hilbert, D., *Methoden der Mathematischen Physik*, Julius Springer, Berlin, 1931.

Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty — II

By H. J. LANDAU and H. O. POLLAK

(Manuscript received July 6, 1960)

The theory developed in the preceding paper[1] *is applied to a number of questions about timelimited and bandlimited signals. In particular, if a finite-energy signal is given, the possible proportions of its energy in a finite time interval and a finite frequency band are found, as well as the signals which do the best job of simultaneous time and frequency concentration.*

I. INTRODUCTION AND SUMMARY

It is a common experience in the communications field that one cannot simultaneously confine a function $f(t)$ and its Fourier transform $F(\omega)$ too severely. The most familiar statement of this phenomenon is the Heisenberg *uncertainty principle*: If we measure the time-spread T of $f(t)$ by

$$T^2 = \frac{\int_{-\infty}^{\infty} (t - t_0)^2 \,|\, f(t) \,|^2 \, dt}{\int_{-\infty}^{\infty} |\, f(t) \,|^2 \, dt}$$

and the frequency-spread Ω of $F(\omega)$ by

$$\Omega^2 = \frac{\int_{-\infty}^{\infty} (\omega - \omega_0)^2 \,|\, F(\omega) \,|^2 \, d\omega}{\int_{-\infty}^{\infty} |\, F(\omega) \,|^2 \, d\omega}$$

then, for any choice of t_0 and ω_0, $\Omega T \geqq \frac{1}{2}$. Thus T and Ω cannot, for any Fourier transform pair, be both small. Equality will hold if $f(t)$ [and hence $F(\omega)$] are gaussian, and t_0 and ω_0 are chosen as the *means* of $|\, f(t) \,|^2$ and $|\, F(\omega) \,|^2$ (in this case both zero). This result, while demonstrating that our experience with timelimiting and bandlimiting is indeed related to mathematical truth, does not succeed in providing a very good understanding of what is really happening. We should like to know just how close one can come to simultaneous limiting in both time and frequency, and what the price is that one has to pay. We need a sharper measure of the concentrations of $f(t)$ and $F(\omega)$ than that afforded by the above variances of $|f(t)|^2$ and $|F(\omega)|^2$, a measure which, if possible, will depend on the behavior of $f(t)$ in a given finite time interval, and of $F(\omega)$ in a given finite frequency band.

An early attempt to meet this need was made by L. A. MacColl, who around 1940 proved the following previously unpublished form of the uncertainty principle:

If

$$\frac{\int_{t_0}^{t_0+T} |f(t)|^2\,dt}{\int_{-\infty}^{\infty} |f(t)|^2\,dt} = \alpha_1$$

and

$$\frac{\int_{\omega_0}^{\omega_0+\Omega} |F(\omega)|\,d\omega}{\int_{-\infty}^{\infty} |F(\omega)|\,d\omega} = \alpha_2,$$

then

$$\Omega T > 2\pi\alpha_1\alpha_2^2. \tag{1}$$

This theorem does indeed emphasize the behavior of $f(t)$ and $F(\omega)$ in given finite intervals. The quantity α_1, representing the proportion of the total energy of $f(t)$ which is in the time-interval $(t_0, t_0 + T)$, is especially satisfying as a measure of the spread of $f(t)$; on the other hand, α_2 has no immediate physical interpretation. A further difficulty with (1) is that there are no functions for which equality can be achieved, although in practice the estimate is quite good.

A more useful form of the uncertainty principle would replace the above measure α_2 by the proportion of energy of $F(\omega)$ in a frequency band, that is, by a definition similar to that of α_1. This is done in the present paper. We shall see that if

$$\frac{\int_{t_0-T/2}^{t_0+T/2} | f(t) |^2 \, dt}{\int_{-\infty}^{\infty} | f(t) |^2 \, dt} = \alpha^2$$

and

$$\frac{\int_{-\Omega}^{\Omega} | F(\omega) |^2 \, d\omega}{\int_{-\infty}^{\infty} | F(\omega) |^2 \, d\omega} = \beta^2,$$

then

$$\Omega T \geqq \Phi(\alpha,\beta),$$

where $\Phi(\alpha,\beta)$ will be found explicitly, the inequality will be sharp and functions yielding equality will be given. The optimal functions $f(t)$ will always be real if, as in the above statement, the frequency band is centered at zero. The same inequality holds if the frequency band under study is not centered at zero, but then the optimal functions are, in general, complex-valued.

The simplest special case of our result arises if $\beta = 1$, so that *all* of $F(\omega)$ is contained in $| \omega | \leqq \Omega$, and $F(\omega) = 0$ for $| \omega | > \Omega$. The question "if α is given, what is the minimum ΩT?" can now be rephrased "if ΩT is given, what is the maximum α?" Let us introduce the following notation: The *square norm* of f is the total energy of f:

$$\| f \|^2 = \int_{-\infty}^{\infty} | f(t) |^2 \, dt.$$

Timelimiting a function f produces a function Df which is f restricted to $| t | \leqq T/2$:

$$Df \equiv \begin{cases} f & \text{if } | t | \leqq T/2 \\ 0 & \text{if } | t | > T/2. \end{cases}$$

Bandlimiting a function f produces a function Bf whose Fourier transform agrees with the Fourier transform of f for $| \omega | \leqq \Omega$, and vanishes for $| \omega | > \Omega$:

$$Bf = \frac{1}{2\pi} \int_{-\Omega}^{\Omega} F(\omega) e^{i\omega t} \, d\omega.$$

By writing

$$F(\omega) = \int_{-\infty}^{\infty} f(s)e^{-i\omega s}\, ds,$$

we see that an alternative expression for Bf is given by

$$Bf = \frac{1}{\pi}\int_{-\infty}^{\infty} f(s)\,\frac{\sin \Omega(t - s)}{t - s}\, ds.$$

It was shown in the preceding paper[1] that if a function is bandlimited and then timelimited its energy must be reduced by at least a factor λ_0, where λ_0 it the largest eigenvalue of the integral equation

$$\lambda f(t) = \frac{1}{\pi}\int_{-T/2}^{T/2} f(s)\,\frac{\sin \Omega(t - s)}{t - s}\, ds. \tag{2}$$

If, in particular, a function is already bandlimited ($f = Bf$), then by this result $\| Df \|^2 \leqq \lambda_0$. This, now, is just the special case of the uncertainty principle which we have been seeking: If $\beta = 1$, then $\alpha \leqq \sqrt{\lambda_0}$.

In the sequel, we shall take a longer look at this formula and its significance; let us, however, state the full result for all values of α and β:

Theorem: There is a function f such that $\| f \| = 1$, $\| Df \| = \alpha$ and $\| Bf \| = \beta$, under the following conditions, and only under the following conditions:

1. If $\alpha = 0$, when $0 \leqq \beta < 1$.
2. If $0 < \alpha < \sqrt{\lambda_0}$, when $0 \leqq \beta \leqq 1$.
3. If $\sqrt{\lambda_0} \leqq \alpha < 1$, when $\cos^{-1} \alpha + \cos^{-1} \beta \geqq \cos^{-1} \sqrt{\lambda_0}$.
4. If $\alpha = 1$, when $0 < \beta \leqq \sqrt{\lambda_0}$.

The body of the present paper will cover the following sequence of topics: Section II will develop the properties of timelimited and bandlimited functions, and the geometric interpretation of these properties, which we require. Section III contains the proof of the quoted theorem, a discussion of the "best" functions, and a number of pertinent graphs and numerical examples. Section IV indicates possible extensions of the theory, and includes the interesting result that if a timelimited function d and a bandlimited function b are given, it is always possible to find a "smallest" function f so that $Df = d$ and $Bf = b$. Finally, Section V gives applications of the preceding theory to filter theory, data transmission and antenna theory.

II. SPACES OF TIMELIMITED FUNCTIONS AND BANDLIMITED FUNCTIONS

We are concerned, in the present paper, with the collection of functions $f(t)$ which are square-integrable on $(-\infty,\infty)$. These form a Hilbert space, denoted by $\mathfrak{L}^2$, in which the inner product (f,g) is defined by

$$(f,g) = \int_{-\infty}^{\infty} f(t)\overline{g(t)}\, dt,$$

and $\| f \|^2 = (f,f)$ as usual.

The collection of timelimited functions forms a linear subspace $\mathfrak{D}$ of $\mathfrak{L}^2$ so that if f_1 and f_2 are timelimited, so is $af_1 + bf_2$. Furthermore, $\mathfrak{D}$ is *complete*, which means that if we have a sequence of functions $\{f_n\}$, $f_n \in \mathfrak{D}$ and if $\| f_n - f_m \| \to 0$, then there is a function $f \in \mathfrak{D}$ such that $\| f - f_n \| \to 0$.

Exactly the same statements may be made about bandlimited functions; they form a complete linear subspace $\mathfrak{B}$ of $\mathfrak{L}^2$. The latter statement follows from the earlier one through the *Parseval relation* for Fourier transforms: If F and G are the Fourier transforms of f and g respectively, then

$$\int_{-\infty}^{\infty} f(t)\overline{g(t)}\, dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)\overline{G(\omega)}\, d\omega.$$

We shall call two functions f and g *orthogonal* if

$$(f,g) = 0.$$

Notice that Df and $f - Df$ are orthogonal, since each one vanishes where the other one does not; by the Parseval relation, Bf and $f - Bf$ are also orthogonal.

The inner product permits us to define the *angle* between two functions f and g as follows: By the Schwarz inequality, we know that

$$| (f,g) | \leqq \| f \| \cdot \| g \| ;$$

since

$$| \operatorname{Re}(f,g) | \leqq | (f,g) | ,$$

we know that

$$-1 \leqq \frac{\operatorname{Re}(f,g)}{\| f \| \cdot \| g \|} \leqq 1 .$$

We may thus define the *angle* $\theta(f,g)$ between the functions f and g by

$$\theta(f,g) = \cos^{-1} \frac{\operatorname{Re}(f,g)}{\| f \| \cdot \| g \|}.$$

The extreme values 0 and π for $\theta(f,g)$ can be reached only if f and g are proportional (so that equality holds in the Schwarz inequality) and (f,g) is real.

Suppose now that $f \in \mathfrak{B}$ and $g \in \mathfrak{D}$, and that neither function vanishes identically. What can we say about the angle between them? The angle can vanish only if for some constant k, $f = kg$. But since $\mathfrak{B}$ and $\mathfrak{D}$ are linear spaces, this would mean that f is *both* timelimited and bandlimited, and this is known to be impossible.† If, then, the angle cannot vanish, can it be arbitrarily small? This is the key question which shall occupy us for some time. Let us consider, first of all, a fixed function $f \in \mathfrak{B}$, and an arbitrary $g \in \mathfrak{D}$. We know that $\theta(f,g)$ cannot vanish; is $\theta(f,g)$ bounded away from zero? If there is a greatest lower bound for $\theta(f,g)$, is it assumed for some particular functions $g \in \mathfrak{D}$? In this case, the answers are quite simple, and are given by the following:

Lemma 1: If $f \in \mathfrak{B}$ is given, then

$$\inf_{g \in \mathfrak{D}} \theta(f,g) > 0.$$

This infimum equals

$$\cos^{-1} \frac{\| Df \|}{\| f \|},$$

and is assumed by $g = kDf$ for any positive constant k.

Proof: If g is any function in $\mathfrak{D}$, then

$$\operatorname{Re}(f,g) \leqq | (f,g) | = | (Df,g) |$$

since

$$f = f - Df + Df \qquad \text{and} \qquad (f - Df,g) = 0.$$

But

$$| (Df,g) | \leqq \| Df \| \cdot \| g \| ,$$

† For then

$$f(t) = \int_{-\Omega}^{\Omega} F(\omega)\, e^{i\omega t}\, d\omega,$$

since $f \in \mathfrak{B}$, would be an analytic function of the complex variable t whose vanishing for $| t | > T$ would imply $f \equiv 0$.

so that

$$\frac{\operatorname{Re}(f,g)}{\| f \| \cdot \| g \|} \leqq \frac{\| Df \|}{\| f \|} = \frac{\operatorname{Re}(f,Df)}{\| f \| \cdot \| Df \|}.$$

Since $\cos \theta$ is monotone decreasing in $(0,\pi)$, it follows that

$$\theta(f,g) \geqq \theta(f,Df)$$

for any $g \in \mathfrak{D}$, with equality whenever g and Df are proportional. This proves the lemma.

We proceed now to the case of arbitrary $f \in \mathfrak{B}$ and $g \in \mathfrak{D}$. Let us say, for convenience, if

$$\inf_{\substack{f \in \mathfrak{B} \\ g \in \mathfrak{D}}} \theta(f,g)$$

is actually assumed by specific functions, that the spaces $\mathfrak{B}$ and $\mathfrak{D}$ form *a least angle*. We now have the following:

Theorem 1: There exists a least angle between $\mathfrak{B}$ and $\mathfrak{D}$. This angle equals $\cos^{-1} \sqrt{\lambda_0}$, and is assumed by $\psi_0 \in \mathfrak{B}$ and $D\psi_0 \in \mathfrak{D}$, where λ_0 is the largest eigenvalue of (2), and ψ_0 the corresponding eigenfunction.

Proof: By the preceding lemma,

$$\min_{g \in \mathfrak{D}} \theta(f,g) = \cos^{-1} \frac{\| Df \|}{\| f \|},$$

so that

$$\inf_{\substack{f \in \mathfrak{B} \\ g \in \mathfrak{D}}} \theta(f,g) = \inf_{f \in \mathfrak{B}} \cos^{-1} \frac{\| Df \|}{\| f \|} \tag{3}$$

and the infimum on the left of (3) will actually be assumed if the infimum on the right is. It was shown in the preceding paper[1] that any $f \in \mathfrak{B}$ may be expanded in a series, convergent in L^2 mean, of the eigenfunctions ψ_n of (2),

$$f = \sum_{n=0}^{\infty} a_n \psi_n .$$

Then

$$\| f \|^2 = \sum_0^{\infty} | a_n |^2;$$

since

$$Df = \sum_{n=0}^{\infty} a_n D\psi_n ,$$

it follows from the properties of $\{D\psi_n\}$ that

$$\| Df \|^2 = \sum | a_n |^2 \lambda_n .$$

Thus

$$\cos^{-1} \frac{\| Df \|}{\| f \|} = \cos^{-1} \left(\frac{\sum | a_n |^2 \lambda_n}{\sum | a_n |^2} \right)^{\frac{1}{2}} .$$

Since it was shown in the preceding paper[1] that $\lambda_n < \lambda_0$, if $n \geqq 1$, it follows that

$$\max \left(\frac{\sum | a_n |^2 \lambda_n}{\sum | a_n |^2} \right)$$

is achieved if $a_n = 0$ for $n \geqq 1$, so that the minimum possible value of

$$\cos^{-1} \frac{\| Df \|}{\| f \|},$$

namely $\cos^{-1} \sqrt{\lambda_0}$, is actually assumed if $f = \psi_0$, and $g = D\psi_0$. The theorem is proved.

We have thus found that the two subspaces $\mathfrak{B}$ and $\mathfrak{D}$ of $\mathfrak{L}^2$, which have no functions except 0 in common, actually have a minimum angle between them, so that, in fact, a timelimited function and a bandlimited function cannot even be very close together. With the aid of this result, as we shall see, the uncertainty principle which we are seeking will follow.

In preparation for the coming theorems, we must consider one further aspect of the spaces $\mathfrak{B}$ and $\mathfrak{D}$. How close do $\mathfrak{B}$ and $\mathfrak{D}$ together come to filling up all of $\mathfrak{L}^2$? The two specific questions which concern us are the following: (i) if $\{f_n\}$, $f_n = b_n + d_n$ is a Cauchy sequence† of functions in $\mathfrak{B} + \mathfrak{D}$, what can the limiting function f look like; and (ii) do there exist functions $f \in \mathfrak{L}^2$ orthogonal to both $\mathfrak{B}$ and $\mathfrak{D}$ (i.e., to every function in $\mathfrak{B}$ and $\mathfrak{D}$)? The answers to these questions are the subjects of the subsequent two lemmas.

Lemma 2: If $\{f_n\}$ is a Cauchy sequence of functions of the form $f_n =$

† A Cauchy sequence of functions is a sequence such that $\| f_n - f_m \| \to 0$, so that, by the completeness of Hilbert space, there exists a limiting function f such that $\| f - f_n \| \to 0$.

$d_n + b_n$ where $d_n \in \mathfrak{D}$ and $b_n \in \mathfrak{B}$ for each n, then the limiting function f is itself of the form $d + b$, where $d \in \mathfrak{D}$ and $b \in \mathfrak{B}$.

Proof: For each $f_n = d_n + b_n$, we may also write

$$f_n = (b_n - Db_n) + (Db_n + d_n).$$

Here $Db_n + d_n \in \mathfrak{D}$, while $b_n - Db_n \perp \mathfrak{D}$. It now follows from the fact that the f_n form a Cauchy sequence that the functions $b_n - Db_n$ do; for

$$\| f_n - f_m \|^2 = \\ \| b_n - Db_n - (b_m - Db_m) \|^2 + \| Db_n + d_n + Db_m + d_m \|^2,$$

so that

$$\| b_n - Db_n - (b_m - Db_m) \| \leqq \| f_n - f_m \| .$$

But now, since $\{b_n - Db_n\}$ forms a Cauchy sequence, so does $\{b_n\}$ itself. For

$$\| b_n - b_m \|^2 = \| D(b_n - b_m) \|^2 + \| (b_n - b_m) - D(b_n - b_m) \|^2,$$

and by Lemma 1,

$$\| D(b_n - b_m) \| \leqq \sqrt{\lambda_0}\, \| b_n - b_m \| ,$$

so that

$$\| b_n - b_m \|^2 \leqq \frac{\| b_n - Db_n - (b_m - Db_m) \|^2}{1 - \lambda_0}.$$

Since $\{b_n\}$ is now a Cauchy sequence, there is a function $b \in \mathfrak{B}$ such that

$$\| b - b_n \| \to 0.$$

Thus $\{f_n\}$ and $\{b_n\}$ both converge in norm, and hence so does $\{d_n\}$, and to a limiting function $d \in \mathfrak{D}$ for which

$$f = b + d.$$

We have thus shown that taking a limit of sums of functions from $\mathfrak{B}$ and $\mathfrak{D}$ gives us nothing new, but only, once again, a sum of functions in $\mathfrak{B}$ and $\mathfrak{D}$. We may abbreviate this by saying simply that $\mathfrak{B} + \mathfrak{D}$ is *closed.*

Lemma 3: There are infinitely many functions in $\mathfrak{L}^2$ which are orthogonal to $\mathfrak{B} + \mathfrak{D}$.

Proof: The functions

$$f_n = \begin{cases} 1 & \text{if } T/2 + n \leqq |t| \leqq T/2 + n + 1 \\ 0 & \text{elsewhere} \end{cases} \qquad n = 0, 1, 2, \cdots$$

are instances of functions *not* in $\mathcal{B} + \mathcal{D}$, since the portion of f_n in $|t| > T/2$ is not a piece of a bandlimited function. Lemma 2 permits us to write the best approximation to f_n from $\mathcal{B} + \mathcal{D}$ in the form $b_n + d_n$, where $b_n \in \mathcal{B}$ and $d_n \in \mathcal{D}$; then

$$f_n^* = f_n - b_n - d_n$$

are distinct functions in $\mathcal{L}^2$ which are orthogonal to $\mathcal{B} + \mathcal{D}$.

There are in fact, in some sense "many more" functions in $\mathcal{L}^2 - \mathcal{B} - \mathcal{D}$ than in $\mathcal{B} + \mathcal{D}$; we do not know, however, of any really convenient representation for such functions.

III. THE UNCERTAINTY PRINCIPLE

We begin by restating the theorem announced in Section I.

Theorem 2: There is a function $f \in \mathcal{L}^2$ such that $\| f \| = 1$, $\| Df \| = \alpha$ and $\| Bf \| = \beta$, under the following conditions, and only under the following conditions:

1. If $\alpha = 0$, when $0 \leqq \beta < 1$.
2. If $0 < \alpha < \sqrt{\lambda_0}$, when $0 \leqq \beta \leqq 1$.
3. If $\sqrt{\lambda_0} \leqq \alpha < 1$, when $\cos^{-1}\alpha + \cos^{-1}\beta \geqq \cos^{-1}\sqrt{\lambda_0}$.
4. If $\alpha = 1$, when $0 < \beta \leqq \sqrt{\lambda_0}$.

Proof: Let $\mathcal{G}$ be the family of functions $f \in \mathcal{L}^2$ with $\| f \| = 1$ and $\| Df \| = \alpha$, and let us, for each case of α, determine

$$\sup_{f \in \mathcal{G}} \beta = \sup_{f \in \mathcal{G}} \| Bf \| .$$

We shall also show, in each case, that any value of β less than the supremum can be realized by an appropriate function. Whether or not the supremum itself can be realized will vary from case to case.

Case 1. $\alpha = 0$. If $\alpha = 0$, the family $\mathcal{G}$ can contain no function with $\beta = 1$. For if $f \in \mathcal{G}$ with $\beta = 1$ we must have $f \in \mathcal{B}$, whence f is analytic and vanishes for $|t| < T/2$ only if $f \equiv 0$. This is a contradiction.

To show that $\mathcal{G}$ contains functions with values of β arbitrarily close to 1 we set

$$f^* = \frac{\psi_n - D\psi_n}{\sqrt{1 - \lambda_n}},$$

where λ_n and ψ_n are respectively an eigenvalue and corresponding eigenfunction of (2). We observe that $f^* \in \mathcal{G}$ and that $\beta = \| Bf^* \| = \sqrt{1 - \lambda_n}$. Since there exist eigenvalues λ_n arbitrarily small, there exist functions in $\mathcal{G}$ with values of β arbitrarily close to 1.

To find functions in $\mathcal{G}$ with values of β between those already covered, we consider $e^{i\rho t}f^*(t)$, which belongs to $\mathcal{G}$ since $\| e^{i\rho t}f^* \| = \| f^* \| = 1$ and $\| D\, e^{i\rho t}f^* \| = \| Df^* \| = \alpha$. For β we find

$$\beta = \| B\, e^{i\rho t}f^* \| = \left\{ \int_{-\rho-\Omega}^{-\rho+\Omega} | F^*(\omega) |^2 \, d\omega \right\}^{\frac{1}{2}},$$

where F^* is the Fourier transform of f^*. This quantity is continuous in ρ and approaches zero as $\rho \to \infty$, since $F^* \in \mathcal{L}^2$; thus $\mathcal{G}$ contains functions with all smaller values of β, except possibly $\beta = 0$.

A function f in $\mathcal{G}$ for which $\beta = 0$ must have the property that $Df = Bf = 0$; the existence of such functions was demonstrated in Lemma 3.

This completes the proof in Case 1; if we reverse B and D in the preceding arguments, we find that $\beta = 0$ is possible if and only if $0 \leqq \alpha < 1$; thus the minimum β in Cases 2 and 3 has also been established.

Case 2. $0 < \alpha < \sqrt{\lambda_0}$. Since $\lambda_n \to 0$ as $n \to \infty$, we can find an eigenvalue $\lambda_n < \alpha$. Let ψ_n be the corresponding eigenfunction, and consider

$$f^* = \frac{\sqrt{\alpha^2 - \lambda_n}\,\psi_0 + \sqrt{\lambda_0 - \alpha^2}\,\psi_n}{\sqrt{\lambda_0 - \lambda_n}}. \tag{5}$$

We have $f^* \in \mathcal{B}$, and $\| f^* \| = \| Bf^* \| = 1$, while a simple computation shows that $\| Df^* \| = \alpha$. This, then, covers the case $\beta = 1$; by picking $e^{i\rho t}f^*$, as in Case 1, we may obtain any $0 < \beta < 1$, and $\beta = 0$ is covered by the remark immediately preceding Case 2.

Cases 3 and 4. $\sqrt{\lambda_0} \leqq \alpha \leqq 1$. For a function $f \in \mathcal{G}$, let us find the closest point to f on the plane spanned by Df and Bf; we then can write

$$f = \lambda Df + \mu Bf + g, \tag{6}$$

with g orthogonal to both Df and Bf. Taking the inner product of (6) successively with f, Df, Bf and g, and using the fact that $f \in \mathcal{G}$, we obtain

$$\begin{aligned} 1 &= \lambda\alpha^2 + \mu\beta^2 + (g,f), \\ \alpha^2 &= \lambda\alpha^2 + \mu(Bf,Df), \\ \beta^2 &= \lambda(Df,Bf) + \mu\beta^2, \\ (f,g) &= (g,g). \end{aligned}$$

By eliminating (g,f), λ and μ from the above equations we find, for $\alpha\beta \neq 0$,

$$\beta^2 - 2\,\mathrm{Re}\,(Df,Bf) = -\alpha^2 + \left(1 - \frac{|(Df,Bf)|^2}{\alpha^2\beta^2}\right) - \|g\|^2\left(1 - \frac{|(Df,Bf)|^2}{\alpha^2\beta^2}\right). \tag{7}$$

We next set

$$\mathrm{Re}\,\frac{(Df,Bf)}{\|Df\|\cdot\|Bf\|} = \cos\theta.$$

The angle θ is that formed between $Df \in \mathfrak{D}$ and $Bf \in \mathfrak{B}$ so that, by Theorem 1,

$$\theta \geqq \cos^{-1}\sqrt{\lambda_0}\,. \tag{8}$$

Since

$$\alpha\beta\cos\theta = \mathrm{Re}(Df,Bf) \leqq |(Df,Bf)| \leqq \alpha\beta,$$

we have

$$0 \leqq 1 - \frac{|(Df,Bf)|^2}{\alpha^2\beta^2} \leqq 1 - \cos^2\theta. \tag{9}$$

Introducing θ into (7), completing the square on the left-hand side, and applying (9) we obtain

$$(\beta - \alpha\cos\theta)^2 \leqq (1 - \alpha^2)\sin^2\theta, \tag{10}$$

with equality if and only if $g = 0$ and (Df,Bf) is real. From (10) we find immediately

$$\beta \leqq \cos(\theta - \cos^{-1}\alpha),$$

whence by (8)

$$\beta \leqq \cos(\cos^{-1}\sqrt{\lambda_0} - \cos^{-1}\alpha), \tag{11}$$

or

$$\cos^{-1}\alpha + \cos^{-1}\beta \geqq \cos^{-1}\sqrt{\lambda_0}\,.$$

Equality in (11) is attained for the function

$$f^* = p\psi_0 + qD\psi_0\,, \tag{12}$$

with

$$p = \sqrt{\frac{1-\alpha^2}{1-\lambda_0}}$$

and (13)

$$q = \frac{\alpha}{\sqrt{\lambda_0}} - \sqrt{\frac{1 - \alpha^2}{1 - \lambda_0}},$$

since f^* satisfies all the conditions for equality in the above sequence of inequalities; the constants p and q are chosen so that $f \in \mathcal{G}$. As in Case 1, all smaller values of β, except possibly for $\beta = 0$, are attainable by the functions $e^{i\rho t}f^*(t)$ with suitable values of ρ, and, by the argument above, the family $\mathcal{G}$ contains functions with $\beta = 0$ as well, except when $\cos^{-1}\alpha = 0$. Thus, in Case 3, $\mathcal{G}$ is made up of functions for which β takes on all values for which

$$\cos^{-1}\alpha + \cos^{-1}\beta \geqq \cos^{-1}\sqrt{\lambda_0}.$$

If, however, $\alpha = 1$, we must exclude $\beta = 0$, so that we obtain in Case 4

$$0 < \beta < \sqrt{\lambda_0}.$$

The result of Theorem 2 is illustrated in Fig. 1, which shows the permissible region in the (α^2,β^2) plane for various values of $c = \Omega T/2$.

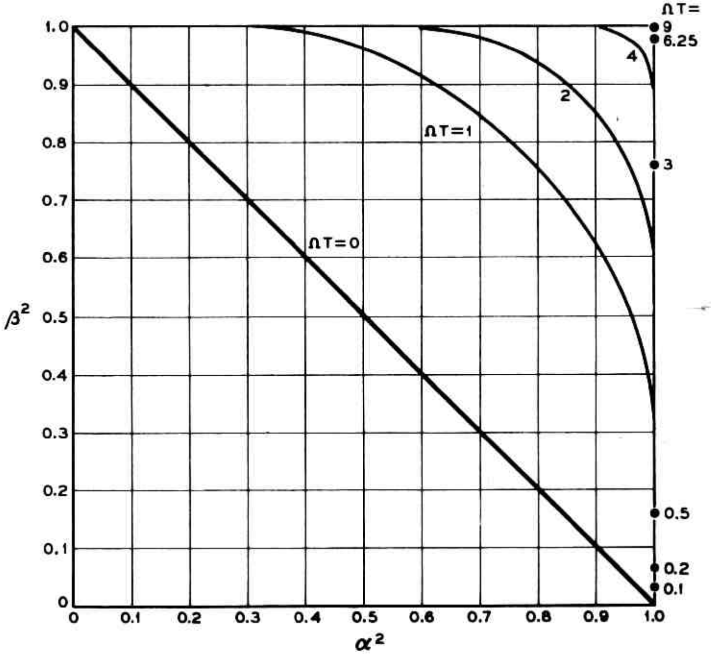


Fig. 1 — Possible combinations of α^2 and β^2 for different ΩT.

For each value of c, this region is bounded by the line segments

$$\begin{aligned} \alpha^2 &= 0 && \text{for } 0 \leqq \beta^2 < 1, \\ \beta^2 &= 0 && \text{for } 0 \leqq \alpha^2 < 1, \\ \alpha^2 &= 1 && \text{for } 0 < \beta^2 \leqq \lambda_0(c), \\ \beta^2 &= 1 && \text{for } 0 < \alpha^2 \leqq \lambda_0(c), \end{aligned}$$

and the curve $\cos^{-1}\alpha + \cos^{-1}\beta = \cos^{-1}\sqrt{\lambda_0(c)}$, which is labeled by the appropriate value of c.

An interesting phenomenon is brought up by the line $\alpha^2 + \beta^2 = 1$, which is labeled with $c = 0$. This labeling agrees with Theorem 2 in the following way:

If $\alpha^2 + \beta^2 \leqq 1$, then $\cos^{-1}\alpha + \cos^{-1}\beta \geqq \pi/2$, which automatically exceeds $\cos^{-1}\sqrt{\lambda_0}$ for any c, no matter how small. In physical terms, this observation states that if the proportions of energy of $f(t)$ in $|t| \leqq T/2$, and of $F(\omega)$ in $|\omega| \leqq \Omega$, add up to less than the total energy of $f(t)$, then we have really put no restraint on Ω and T, and an arbitrarily small ΩT product will still permit this distribution of energy. It is only when $\alpha^2 + \beta^2 > 1$, so that the energies in $|t| < T/2$ and in $|\omega| < \Omega$ add up to more than the total energy, that a nonzero lower bound on ΩT is implied.

Fig. 2 gives a detailed plot of what is essentially the top (or the right) edge of Fig. 1. We plot $\lambda_0(c)$, the maximum of α^2 if $\beta^2 = 1$, against c. We note that $\lambda_0(c) \to 1$ quite rapidly as $c \to \infty$; the approach is exponential, but the exact rate has not been proved. Fig. 2 also gives,

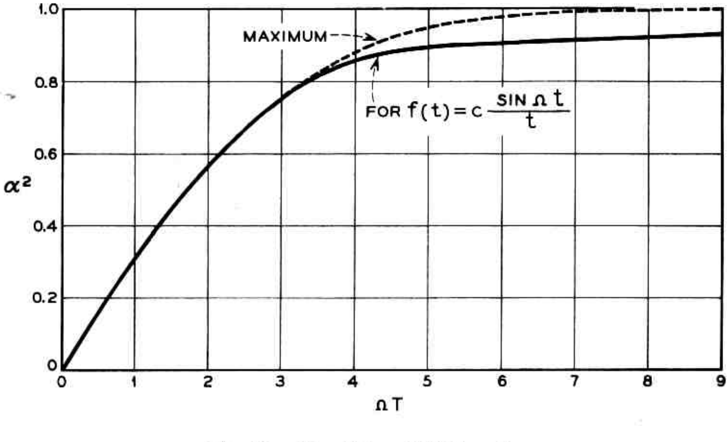


Fig. 2 — Possible α^2 if $\beta^2 = 1$.

for comparison, the proportion of energy in $|t| < T/2$ for the function

$$f(t) = \frac{\sin \Omega t}{t},$$

which has sometimes been "intuitively" considered as the bandlimited function which is as concentrated in time as possible. For small λ_0, it appears, $f(t)$ is indeed essentially as good as the optimal function; if, however, we wish to achieve a proportion of energy like 92 per cent, we see that $\Omega T = 4.5$ suffices, while use of $(\sin \Omega t)/t$ would require $\Omega T = 8.5$. For a proportion of 99 per cent, the minimal ΩT is 6.25, while $(\sin \Omega t)/t$ would require a value of ΩT of about 30.

Let us consider one more numerical example. If values of $\alpha^2 = 0.977$ and $\beta^2 = 0.96$ are desired, what are the minimum ΩT, and the corresponding optimal function? From $\cos^{-1}\alpha + \cos^{-1}\beta = \cos^{-1}\sqrt{\lambda_0}$ we find $\lambda_0 = 0.88$, so that $\Omega T = 4$, or $c = 2$. If, now, $\psi_0(t)$ is the first eigenfunction corresponding to $c = 2$, then, by (12) and (13), the optimal function (see Fig. 3) is $0.578\psi_0 + 0.465D\psi_0$. It is thus not a continuous function of t but has jumps at $t = \pm T/2$; this is characteristic of all of our optimization problems except for the special case $\beta^2 = 1$.

A note on previous work in the direction of Theorem 2. The connection between the extremum problem for $\beta^2 = 1$ and the largest eigenvalue of (2) was noted by Chalk[2] and Gurevich,[3] both of whom found the appearance of the optimal function without analytic solution

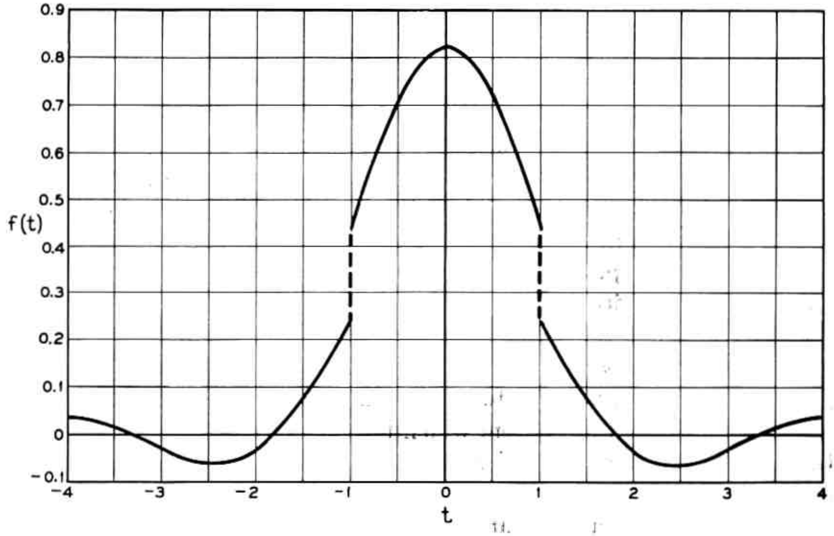


Fig. 3 — Plot of optimal $f(t)$ for $\alpha^2 = 0.977$, $\beta^2 = 0.96$, $T/2 = 1$.

of the integral equation; the latter also plotted the largest eigenvalue. The set of eigenfunctions was recognized in this context by Ville and Bouzitat,[4] who also performed a lot of numerical work. Finally Fuchs[5] has stated, without proof, a theorem equivalent to Theorem 2. He considers n-dimensional spaces and Fourier transforms, and two arbitrary subsets of finite measure in the time- and frequency-spaces respectively. His proof, however, which we have been privileged to see, is quite different, and is not directed towards the properties of $\mathfrak{B}$ and $\mathfrak{D}$ which have been our chief concern. Our present method is capable of broad generalization; some thoughts in this direction are given in the next section.

IV. EXTENSIONS OF THE THEORY

It is quite natural for us to ask what the real essentials of the study up to this point have been, and under what circumstances results similar to Theorems 1 and 2 could be obtained. Such an investigation will be reported in a separate paper;[6] we should, however, note what some of the results are. For the relevant language, we refer the reader to Ref. 6.

We have a Hilbert space $\mathfrak{L}^2$, and two subspaces $\mathfrak{B}$ and $\mathfrak{D}$. The key property we require is that $\mathfrak{B}$ and $\mathfrak{D}$ form a nonzero minimum angle; the latter property is equivalent to requiring that

$$\sup_{f \in \mathfrak{L}^2} \frac{\| BDBf \|}{\| f \|} < 1.$$

It now follows that $\mathfrak{B} + \mathfrak{D}$ is closed, and we can again study the region of possible values of $\| Bf \|$ and $\| Df \|$ if $\| f \| = 1$. We do not, however, obtain eigenfunctions analogous to $\{\psi_n\}$ unless the operator BDB is completely continuous. If, for example, $\mathfrak{L}^2$ is the space of square-integrable functions with respect to Lebesgue measure over n-dimensional Euclidean space R^n, if $\mathfrak{D}$ is the subspace of functions vanishing outside of a bounded subset of R^n of positive measure, and if $\mathfrak{B}$ is the subspace of functions whose Fourier transforms vanish outside of another bounded subset of R^n of positive measure, then BDB is completely continuous, and the full theory applies.

As an example of a theorem which is again true in the general situation, but is of interest also for timelimited and bandlimited functions, let us prove

Theorem 3: Let an arbitrary function $d \in \mathfrak{D}$, and another function $b \in \mathfrak{B}$, be given. Then there exists an infinite collection S of functions $f \in \mathfrak{L}^2$ such that if $f \in S$, then $Df = d$ and $Bf = b$. There is a unique

$f_0 \in S$ of least energy, and there is a unique $f_1 \in S \cap (\mathfrak{B} + \mathfrak{D})$; furthermore, $f_0 = f_1$.

Proof: Let us consider the function

$$f^* = \sum_0^\infty (1 - B)(DB)^m d + \sum_0^\infty (1 - D)(BD)^m b. \qquad (14)$$

The first sum, for example, means

$$d - Bd + DBd - BDBd + DBDBd - BDBDBd + \cdots .$$

Since, for any g, $\| DBg \| < \sqrt{\lambda_0} \, \| g \|$ and $\| BDg \| \leqq \sqrt{\lambda_0} \, \| g \|$, we know that the two series defined on the right side of (14) converge in norm, with their sum defined as the function $f^* \in \mathfrak{L}^2$. Furthermore, since f^* is defined as a limit of functions in $\mathfrak{B} + \mathfrak{D}$, it is, by Lemma 2, itself in $\mathfrak{B} + \mathfrak{D}$. So we may write

$$f^* = d^* + b^*,$$

where $d^* \in \mathfrak{D}$ and $b^* \in \mathfrak{B}$.

Let us next compute Df^* and Bf^*. We have

$$Df^* = \sum_0^\infty (1 - DB)(DB)^m d + \sum_0^\infty (D - D)(BD)^m b;$$

all of the second series, and all but the first half of the first term of the first series, vanish. Hence $Df^* = d$, and similarly $Bf^* = b$. We have thus shown that $f^* \in S \cap (\mathfrak{B} + \mathfrak{D})$; we can complete the proof that $f^* = f_1$ if we can show that $S \cap (\mathfrak{B} + \mathfrak{D})$ contains no other function.

Suppose that $f_i = d_i + b_i$, $i = 1, 2$ are both in $S \cap (\mathfrak{B} + \mathfrak{D})$. Then

$$d = Df_1 = d_1 + Db_1 = d_2 + Db_2 = Df_2 \qquad (15)$$

and

$$b = Bf_1 = Bd_1 + b_1 = Bd_2 + b_2 = Bf_2$$

so that

$$DBd_1 + Db_1 = DBd_2 + Db_2 . \qquad (16)$$

Hence, by subtracting (16) from (15), we have

$$(1 - DB)d_1 = (1 - DB)d_2 ,$$

or

$$(d_1 - d_2) = DB(d_1 - d_2).$$

Since, however, $\| DBg \| \leqq \sqrt{\lambda_0} \, \| g \|$ for any g, we must have $d_1 - d_2 = 0$, so that $d_1 = d_2$. Similarly, $b_1 = b_2$, so that $f_1 = f_2$, and thus f^* is the unique member of $S \cap (\mathfrak{B} + \mathfrak{D})$.

Now suppose x is any other member of S. We may write

$$x = f^* + \varphi,$$

and since $Dx = Df^* = d$ and $Bx = Bf^* = b$, it follows that

$$D\varphi = B\varphi = 0.$$

But

$$\| x \|^2 = \| f^* \|^2 + \| \varphi \|^2 + 2 \operatorname{Re}(f^*,\varphi),$$

and

$$f^* = d^* + b^* \qquad \text{while} \qquad \varphi \perp \mathfrak{D} + \mathfrak{B}.$$

Hence

$$(f^*,\varphi) = 0,$$

and

$$\| x \|^2 = \| f^* \|^2 + \| \varphi \|^2 \geqq \| f^* \|^2,$$

with equality if and only if φ vanishes. Thus f^* is also the unique member of S of minimum norm. An infinite number of other members of S may be formed by adding to f^* any of the functions orthogonal to $\mathfrak{B} + \mathfrak{D}$ whose existence is guaranteed by Lemma 3.

Note: If $d = \sum a_i D\psi_i$ and $b = \sum b_i\psi_i$, then

$$f^* = \sum \frac{a_i - b_i}{1 - \lambda_i} D\psi_i + \sum \frac{b_i - a_i\lambda_i}{1 - \lambda_i} \psi_i ,$$

so that, in particular,

$$\| f^* \| \leqq \frac{1}{\sqrt{1 - \lambda_0}} (\| d \| + \| b \|).$$

V. APPLICATIONS

5.1 *Filter Theory*

Suppose we wish a filter to have an impulse response $f(t)$ which vanishes for $t > T$. Such a filter clearly cannot be strictly bandpass; but how would we select the filter so that as much of the impulse response as possible is contained in $| \omega | < \Omega$ for some given Ω? Suppose, by this, we mean to choose $f(t)$ so that

$$\frac{\int_{-\Omega}^{\Omega} | F(\omega) |^2 \, d\omega}{\int_{-\infty}^{\infty} | F(\omega) |^2 \, d\omega}$$

is as large as possible, where

$$F(\omega) = \int_0^T f(t)\, e^{-i\omega t}\, dt$$

is the Fourier transform of $f(t)$. Then the best choice is

$$f(t) = \psi_0\left(t + \frac{T}{2}, c\right),$$

where $c = \Omega T/2$, and ψ_0 is the prolate spheroidal function of the present and the preceding papers.

If, instead of requiring $f(t)$ to vanish outside of $(0,T)$, we ask that both

$$\int_{|\omega| \geqq \Omega} | F(\omega) |^2 \, d\omega = \beta^2$$

and

$$\int_{-\infty}^{0} + \int_{T}^{\infty} | f(t) |^2 \, dt = \alpha^2$$

be small while the total energy of the impulse response is fixed at unity, then Theorem 2 above gives the complete region of possible (α,β) values.

5.2 *Data Transmission*

When we choose a combination of pulse shape and transmission characteristic for a broadband data transmission system, we are interested in minimizing both the tail of a pulse outside its time slot and its spectrum outside of an assigned frequency band. Once again, it is not possible to make both of these "spillovers" in time and frequency arbitrarily small; the above theory gives some information on interchannel and intersymbol interference. For a theory which is more nearly complete, however, the relation between timelimiting and passbandlimiting (i.e., to $\Omega_1 \leqq | \omega | \leqq \Omega_2$) needs to be better understood; while our general results apply, the identity of the optimal function ψ_0 is not known in the case that B is projection of the transform into such a passband.

5.3 *Antenna Theory*

Let us consider a horizontal (s,t) plane from which the strip $| t | < a$ of width $2a$, to be called the *aperture*, has been removed. If the illumination across the aperture is independent of s, then the amplitude of the field across the aperture may be represented by a function $f(t)$ of one variable, where $| t | < a$. If we consider the resultant pattern of radiation in a distant parallel horizontal plane, then the field at a large distance from the aperture is proportional to

$$\int_{-a}^{a} f(t)\, e^{itu}\, dt = F(u),$$

where $u = k \sin \theta$, $k = 2\pi/\lambda$, θ is an angle measured from the vertical through the center of the aperture, and λ is the wavelength. The Q of the antenna is then defined (equivalent to the definition of Woodward and Lawson;[7] it is given explicitly by Kovács and Solymán[8]) as

$$Q = \frac{\int_{|u|>k} | F(u) |^2\, du}{\int_{-k}^{k} | F(u) |^2\, du}.$$

This may be rewritten as

$$Q = \frac{\int_{-a}^{a} | f(x) |^2\, dx}{\int_{-\infty}^{\infty} | Bf(x) |^2\, dx} - 1,$$

where B means limiting the Fourier transform of f to $| u | \leqq k$. Thus by the previous theory,

$$Q \geqq \frac{1}{\lambda_0} - 1,$$

where $\lambda_0 = \lambda_0(ak/2)$ is the first eigenvalue of (2) as defined in this and the preceding paper. We thus have an absolute lower bound on the Q which can be obtained for given a and k.

REFERENCES

1. Slepian, D. and Pollak, H. O., this issue, p. 43.
2. Chalk, J. H. H., The Optimum Pulse-Shape for Pulse Communication, Proc. I.E.E., **87,** 1950, p. 88.
3. Gurevich, M. S., Signals of Finite Duration, Containing a Maximal Part of Their Energy in a Given Bandwidth, Radiotechnika: Elektronika, **3,** 1956, p. 313.
4. Ville, J. A. and Bouzitat, J., Note sur un signal de durée finie et d'energie filtrée maximum, Cables & Trans., **11,** 1957, p. 102.
5. Fuchs, W. H. J., On the Magnitude of Fourier Transforms, Proc. Int. Math. Cong., Amsterdam, September 1954.
6. Landau, H. J. and Pollak, H. O., Subspaces Forming a Nonzero Angle, with Applications to Fourier Analysis, to be published.
7. Woodward, P. M. and Lawson, J. D., The Theoretical Precision with Which an Arbitrary Radiation Pattern May Be Obtained from a Source of Finite Size, J.I.E.E., **95,** pt. III, 1948, p. 363.
8. Kovács, R. and Solymán, L., Theory of Aperture Aerials Based on the Properties of Entire Functions of Exponential Type, Acta Phys. Budapest, **6,** 1956, p. 161.

Considerations on the Solar Cell

By D. A. KLEINMAN

(Manuscript received September 1, 1960)

The collection efficiency in solar cells is treated by a new method in which all the effects of the solar spectrum and the absorption curve are contained in a single function readily obtained by numerical integration. The method is illustrated by a detailed study of the effects of surface recombination, body recombination and junction depth in silicon cells. The method is also generalized to include built-in electric fields, and calculations are given for silicon. Sufficiently strong fields to improve the collection efficiency markedly can be produced in some compound semiconductors from a gradient in the energy gap. A discussion is given of the dependence of the collection efficiency on the absorption curve of the semiconductor. It is shown that silicon has a very favorable absorption curve in comparison with GaAs or InP. Finally, a treatment is given of the minority carrier collection in a two-junction cell, and calculations are presented for silicon. It is concluded that this structure may be important for cells with high energy gaps and short lifetimes.

I. INTRODUCTION

The considerations reported in this paper have been stimulated by the current interest in the solar battery as a power supply for instruments and transmitters in satellite and space probe vehicles. A number of space vehicles† have contained solar batteries with peak outputs ranging from several watts to several hundred watts. It has been demonstrated that solar battery power supplies are technically feasible, not only in space vehicles but also in terrestrial telephone systems.[2] For the latter type of application, however, the solar battery has been found to be not competitive economically with several other available power supplies.[2] However, it is highly advantageous for space vehicles because of its advanced development and commercial availability, light weight, reliability and long life. The long life is due not only to the ruggedness and permanence of its structure and the absence of moving parts or

† For a review of the use of solar batteries in space see Daniels.[1]

chemically unstable components, but also to its external and independent source of energy, the sun. The belts of particle radiation[3] recently discovered surrounding the earth may materially reduce the life of solar batteries carried into space.[4,5] We shall proceed with our discussion, however, on the assumption that the solar battery can survive the radiation.

The solar battery is an array of hundreds or thousands of individual cells called solar cells. Each cell is a semiconductor slab, typically 1 × 2 cm, containing a p-n junction within a very small distance, typically 2×10^{-4} cm, of the illuminated surface. The front and back surfaces are fitted with "ohmic" contacts for making electrical connection, and the front surface may be specially treated to reduce its reflectivity.[6] Commercially available cells are made of silicon, and have efficiencies of up to 14 per cent for converting the solar radiation incident upon them into electric power.[7] Other methods now known for converting solar radiation are far less efficient.† Thermoelectric converters,[9] for example, on which considerable work has been done, can approach efficiencies of 1 per cent.

The solar cell was invented by Chapin, Fuller, and Pearson,[10] who briefly described its fabrication, its principles of operation, and the limitations on its efficiency. Due to the work of the inventors and later authors, notably Prince,[11] Pfann and Van Roosbroeck,[12] Cumerow,[13] Rittner[14] and Loferski,[15] the solar battery is well understood in terms of concepts familiar in electrical circuits and semiconductor physics.

High efficiency in a solar battery would be desirable in any application, but especially so in space vehicles. The value of the vehicle, launched at considerable expense, depends in large measure on the instruments and transmitters it carries and the power available for this equipment. At the same time, the more equipment that is carried, the less space and weight can be allowed for the power supply. Therefore, in the economy of space vehicles the solar battery should have the maximum possible efficiency irrespective of the costliness of the improvements.

The operation of the solar cell and the losses of efficiency can best be described in five steps:

(a) Radiation is incident upon the surface and some is *reflected* without entering the cell. This reflection can be a very important loss, since the reflectivity[16] of clean silicon is about 30 per cent in the wavelength range of interest (0.4 to 1 μ), and other materials that might be used also have high reflectivities. In practice it is found that the processing in the manufacture of silicon cells leaves the surface with quite a low reflectivity.[6] A recent study[16] has shown that the best antireflection

† For a review of the utilization of solar energy, see Ref. 8, especially Vol. V.

treatments can increase the short circuit current of silicon solar cells by 20 to 25 per cent compared with clean surface cells. It is reasonable to expect that the reflection loss from any material that might be used in the solar cell could be minimized in the same way.

(b) The light enters the cell and some is absorbed by the intrinsic absorption process in which a hole-electron pair is created and a photon is destroyed.† The light absorbed in this way is the useful light in the solar cell. Light of wavelength longer than the intrinsic absorption edge cannot produce hole-electron pairs and is wasted in the solar cell. There is a further waste of energy when hole-electron pairs are produced by photons with more than the minimum required energy, since the excess energy is transferred in a very short time to the semiconductor lattice in the form of heat. These losses may be ascribed to the *spectrum* of solar radiation, since they would not occur if the radiation were monochromatic at the wavelength of the absorption edge. For silicon the losses due to the solar spectrum are about 53 per cent of the energy which enters the cell.[15] By choosing a semiconductor with a somewhat higher energy gap‡ this loss can be considerably reduced.[13,14,15] The energy gap is therefore very important in considering materials for the solar cell.§

(c) Some of the minority carriers produced by the light flow by diffusion to the p-n junction. These are the carriers which contribute to the output current of the cell. Other carriers diffuse away from the junction and *recombine* at the surface or deep inside the cell. The percentage of minority carriers which contribute to the current is called the *collection efficiency*. In a typical commercial silicon solar cell of 9 per cent over-all efficiency the collection efficiency is about 60 per cent.[18] Minority carrier recombination is, therefore, a serious loss of efficiency in the solar battery at the present time. From the spectral response[6,10,18] of silicon solar cells we can infer that the surface recombination velocity is very high, probably greater than 10^5 cm/second. From the analysis to be presented we can also infer that the body lifetime is about 10 microseconds, the diffusion length‖ about 10^{-2} cm. In high-purity silicon the lifetime can be several milliseconds[20,21] and with certain surface treatments the recom-

† For a review of the optical properties of semiconductors, see Hrostowski.[17]

‡ For the purposes of this paper the energy gap and the intrinsic absorption edge are essentially the same thing, except that the former is expressed as an energy while the latter is the equivalent photon wavelength.

§ We shall not consider in this paper composite cells such as those suggested by E. Jackson (Ref. 8, Vol. V), in which one attempts to reduce spectrum losses by stacking several thin cells of different energy gaps. Cells of this type appear to be somewhat impractical from the standpoints of mechanical construction and of providing a suitably matched electrical load.

‖ Recently a measured value of $\sim 5 \times 10^{-3}$ cm for the diffusion length has been reported by Vavilov, Smirnov and Patskevitch.[19]

bination velocity[22] can be as low as 40 cm/second. The recombination, therefore, is due almost entirely to the degradation[20] of surface and body lifetime that occurs in the manufacturing process. In compound semiconductors the recombination losses are likely to be considerably greater than in silicon. In gallium arsenide, for example, the lifetime seems at present to be of the order of a millimicrosecond.[23] In general, it appears difficult to obtain lifetimes greater than 0.01 or 0.1 microsecond in the compound semiconductors.†

(d) The diffusion maintains an excess concentration of minority carriers on both sides of the junction. The voltage developed by the solar cell is due to these excess concentrations of minority carriers. This voltage, however, is considerably less than the energy (in units of electron-volts) of a hole-electron pair in the semiconductor. The latter, for our purposes, may be taken to be the energy gap, which in silicon is 1.2 volts.[25] The voltage of a silicon solar cell in full sunlight under maximum power conditions is about 0.4 volts.[10,18] Therefore, the cell is able to convert only a portion of the energy stored as hole-electron pairs into electrical work. The loss may be referred to as the *junction loss*. The junction loss should vanish and the voltage should approach the energy gap when the minority carrier density approaches the majority carrier density, a limit corresponding to infinite light intensity. In the other limit of zero light intensity, the junction loss causes the efficiency of the solar cell to approach zero. According to the equivalent circuit point of view of Pfann and Van Roosbroeck,[12] the short-circuit current of the cell flows partly through the load and partly through the junction in the forward direction. The voltage and the junction loss therefore depend upon the forward current-voltage characteristic of the junction. The theory of p-n junctions[26,27] predicts that the forward current should decrease exponentially with increasing energy gap. Therefore, insofar as actual junctions obey the ideal junction theory, the junction loss can be reduced by increasing the energy gap. A number of authors[13,14,15] have considered the spectrum loss and junction loss as a function of energy gap. If all other losses are neglected, the maximum efficiency is obtained for an energy gap of about 1.6 volts.[15] It is now possible with mixed semiconductors[24,28] to obtain nearly any desired energy gap from 0.7 volt (germanium) to 2.4 volts (GaP), which completely covers the range of interest. It should be kept in mind that silicon p-n junctions show a large contribution to the current from thermal generation and recombination through traps in the junction region.[29] The short lifetimes in the compound semiconductors suggest that trap effects may be even more important in those

† Ref. 24, p. 58

materials. Therefore, it may not be possible to obtain large reductions in the junction loss by increasing the energy gap.

(e) Finally there is the loss due to *resistance* of the very thin side of the junction next to the surface and of the contact to the surface. This resistance places a lower limit on the depth of the junction. Prince[11] has considered the optimum depth for silicon solar cells taking into account the resistance loss and the collection loss. In practice, the internal resistance of silicon cells[11] is between one and two ohms and the junction depth is between 1×10^{-4} and 2×10^{-4} cm. The elimination of resistance loss would increase the efficiency of a 9 per cent cell to about 11 or 12 per cent.[18] Unlike the other losses considered here, the resistance loss is not characteristic of the material used in the solar cell, and the methods for reducing it will be similar for any material.

It has been mentioned that the solar spectrum loss and the junction loss have been considered at some length. It is also well understood what to do about reflection and resistance. But the collection loss has not been thoroughly treated in the literature. Several calculations[6] have been made using the approach of Pfann and Van Roosbroeck.[12] This may be called the monochromatic method. Light of a certain wavelength, and therefore having a certain absorption coefficient in the material, enters the solar cell. Solutions are obtained in terms of elementary functions for the minority carrier density, taking into account diffusion, surface and body recombination, the generation of carriers by the light and the boundary conditions at the junction and at infinite depth. The solution gives a certain diffusion current into the junction which is the short-circuit current of the cell. This current must then be averaged over the wavelengths in the solar spectrum to obtain the collection efficiency. The averaging requires a tedious numerical integration, because the monochromatic current is a complicated function of the absorption coefficient, which is in turn a rapidly varying (measured) function of wavelength. The method is sound and can account in a satisfactory way for the collection efficiencies observed in solar cells.[6] The method is not well suited, however, for a systematic discussion of collection efficiency and no such discussion has been given. Several authors[13,14,15] have even argued that the collection efficiency can be considered unity in fundamental considerations on the solar battery, since lifetimes can be expected to increase as technology improves. A review of the history of the photovoltaic effect and its utilization has recently been given by Rappaport.[30]

In this paper we consider the collection efficiency more systematically by a more powerful method than the monochromatic method. This method is based upon obtaining a function, called the *photodensity function*, which includes all the effects of the solar spectrum and the wavelength dependent absorption coefficient. This function is obtained by a relatively easy numerical integration over the solar spectrum. In terms of the photodensity function the solution can be obtained almost immediately to any solar cell collection problem in the approximation in which a single diffusion length describes the minority carriers. Illustrative calculations are presented for silicon solar cells showing how the collection efficiency depends on surface recombination velocity, junction depth and diffusion length. The method is readily generalized to include cases in which different diffusion lengths must be used for electrons and holes. Another generalization is presented which takes into account a "built-in" electric field in the region between the junction and the surface. It is shown that by the use of mixed semiconductors it should be possible to obtain sufficiently large built-in fields to increase the collection efficiency significantly. In ordinary silicon cells, however, one would not expect the field to be large enough to have much effect. The important question of which semiconductors should be best for solar battery applications, already much discussed[11,13,14,15,30] with respect to spectrum and junction losses, is taken up again from the point of view of the collection efficiency. It is pointed out that the absorption coefficient as a function of wavelength is very important in determining the collection efficiency. Silicon has an absorption curve of favorable shape, which in part accounts for its present superiority over other materials of more favorable energy gap. It is possible that for room-temperature use silicon will remain the best material, although higher-gap materials will certainly be needed for use at temperatures above 200°C. Finally, there is presented a discussion of cells containing two junctions to improve the collection efficiency. It is shown that in cells of comparatively low collection efficiency (<50 per cent), considerable improvement can be obtained by the use of a second junction. This construction may prove important in high gap cells which might otherwise have rather low collection efficiencies.

II. FORMULATION OF THE PROBLEM

If the concentration of minority carriers is small compared to the concentration of majority carriers, the equation describing the production, diffusion, and recombination of minority carriers is

$$D\frac{d^2n}{dx^2} - \frac{n}{\tau} + \int_0^{\lambda_G} N(\lambda)\alpha(\lambda)e^{-\alpha x}\,d\lambda = 0. \tag{1}$$

In (1), $n(x)$ is the excess minority carrier concentration over the equilibrium concentration. In a typical case the maximum value of $n(x)$ is of the order 10^{12} cm^{-3}, so that the validity of (1) is assured. The diffusion constant D and lifetime τ will be assumed to apply to all minority carriers whether holes on the n-side or electrons on the p-side of the junction. This assumption should lead to no serious error in silicon, but might have to be modified for some of the III–V semiconductors because of the relatively low mobility of the holes.† The integral term in (1) represents the production of minority carriers by light with a photon distribution $N(\lambda)$ in a material with absorption coefficient $\alpha(\lambda)$ and intrinsic absorption edge λ_G. The total flux of photons capable of producing hole-electron pairs is

$$N = \int_0^{\lambda_G} N(\lambda)\,d\lambda. \tag{2}$$

We shall neglect reflection completely and identify $N(\lambda)$ with the solar photon spectrum with respect to wavelength.‡ The total effective photon flux is $N = 3.3 \times 10^{17}$ cm^{-2} sec^{-1} for silicon.§ The boundary conditions to be imposed on $n(x)$ are

$$n(\infty) = 0, \qquad n'(0) = (s/D)n(0), \tag{3}$$

where s is the recombination velocity of the surface. A solution of (1) and (3) over the whole domain $0 \leqq x < \infty$ represents the minority carrier density in a homogeneous illuminated semiconductor.

If we now locate a junction at depth $x = a$, additional boundary conditions must be satisfied at the junction. In general, these will relate the minority carrier densities on each side of the junction to the operating voltage of the cell. The simplest case is the short-circuit condition, in which the voltage is zero.‖ For this case the excess carrier densities must vanish on each side of the junction.[26,27] The boundary conditions for the short-circuit condition are therefore

$$n(a) = 0. \tag{4}$$

† Ref. 24, p. 12.

‡ Details on the solar spectrum are given by Ref. 31. See also Ref. 15.

§ This number, obtained from Ref. 31, is in substantial agreement with the plot of $\int_0^\lambda N(\lambda)d\lambda$ in Ref. 15. We are considering the solar radiation in space just outside the earth's atmosphere.

‖ We neglect voltage drops due to internal resistance. This causes no loss of generality, since an arbitrary resistance can be included in the equivalent circuit. In the presence of resistance, (4) corresponds to a small forward bias externally applied to the cell. The statement following (4) remains valid, and one identifies this maximum diffusion current, not the short-circuit current, with the current generator.

This corresponds to the maximum current which can be drawn from the cell; smaller currents correspond to nonvanishing values of $n(a)$ on both sides of the junction. However, we may regard any solution of (1) corresponding to arbitrary operating conditions as the *superposition* of the short-circuit solution and an appropriate solution of the homogeneous equation

$$D\frac{d^2n}{dx^2} - \frac{n}{\tau} = 0. \tag{5}$$

According to p-n junction theory,[26,27] these solutions are just those associated with forward currents in an unilluminated p-n junction. The superposition of solutions for the minority carrier density is therefore equivalent to superposing the short-circuit current with an appropriate forward current in the junction. If I_g is the short-circuit current and I_f the forward current, the current in the external circuit is $I_g - I_f$, in accordance with the equivalent circuit of Pfann and Van Roosbroeck.[12]

The collection problem, therefore, is to calculate the short-circuit solution from (1) and (4) and from this the diffusion current density at the junction, taking into account the contributions from each side. The collection efficiency then is given by

$$Q_1 = DN^{-1}(\,|\,n'(a-)\,|\, + \,|\,n'(a+)\,|\,), \tag{6}$$

with the subscript "1" indicating that the expression applies to cells with a single junction. In a many junction cell Q_n would be a sum of terms of the form (6).

The equivalent circuit for a solar cell with load therefore consists of a current generator I_g connected to a junction and a load in parallel, with the currents I_f flowing in the junction (forward) and $I_f - I_g$ in the load. Pfann and Van Roosbroeck[12] obtained the following condition for maximum power delivered to the load:

$$G = z \ln z + z - 1, \tag{7}$$

where

$$z = e^{eV/kT}, \tag{8}$$

V is the voltage, and G is a dimensionless reduced current proportional to I_g. It is assumed that I_f obeys an ideal junction characteristic[26,27]

$$I_f = eAJ_0(e^{eV/kT} - 1), \tag{9}$$

where A is the junction area and J_0 is a characteristic particle current density. The reduced current G is then defined by the relation

$$G = I_g/eAJ_0 = (N/J_0)Q. \tag{10}$$

The over-all efficiency can be written[12]

$$\epsilon = \frac{kT}{W} Qz \frac{(\ln z)^2}{G}, \tag{11}$$

where z is the solution of (7), and

$$W = N^{-1} \int_0^\infty N(\lambda)(hc/\lambda)\, d\lambda. \tag{12}$$

Clearly W is the radiant energy asborbed on the average to produce one hole-electron pair. For large G, (7) becomes

$$G \approx z \ln z \qquad G \gg 1, \tag{13}$$

and (11) becomes

$$\epsilon \approx (eV/W)Q. \tag{14}$$

Here eV is the work done on the load and W/Q_1 is the energy absorbed per carrier flowing in the load. The approximation of large G is fully justified, since in full sunlight G will be of the order 10^7. The efficiency for large G can also be written

$$\epsilon_1 \approx (kT/W)Q_1 \ln G, \tag{15}$$

which exhibits the logarithmic dependence of the efficiency on the light intensity. It will be noted that the over-all efficiency ϵ_1 depends upon the collection efficiency Q_1 through the factor Q_1 and also through G.

III. THE PHOTODENSITY METHOD

The collection problem consists in solving (1) subject to (3) and (4). We begin by observing that the function

$$\int_0^{\lambda_G} N(\lambda)\tau \frac{\alpha(\lambda)}{1 - \alpha^2 L^2} e^{-\alpha x}\, d\lambda, \tag{16}$$

where $L = (D\tau)^{\frac{1}{2}}$ is the diffusion length, is a particular solution of (1). It satisfies the boundary condition at infinity but none of the other boundary conditions. If ambiguity arises due to the pole $\alpha L = 1$ occurring in the range of integration, the integral can be taken as a principal value. The pole is removed if we combine (16) with an appropriate solution of the homogeneous equation (5), and write

$$n(x)_p = \int_0^{\lambda_G} N(\lambda)\tau \frac{\alpha(\lambda)}{1 - \alpha^2 L^2} (e^{-\alpha x} - e^{-x/L}) \, d\lambda. \tag{17}$$

This function is a solution of (1) and satisfies

$$n(\infty) = n(0) = 0, \tag{18}$$

which is the form taken by (3) when the recombination velocity s is very large. Therefore, $n(x)_p$ represents the density of excess minority carriers in a semiconductor without junctions illuminated on a surface with fast recombination. The derivative of $n(x)_p$ at the surface is

$$n'(0)_p = \int_0^{\lambda_G} N(\lambda) \frac{\alpha(\lambda)}{1 + \alpha L} \, d\lambda. \tag{19}$$

We call $n(x)_p$ the *photodensity*. It can be readily evaluated by numerical integration (see Appendix), since the integrand is a simple expression with no poles.

It is convenient now to introduce dimensionless quantities

$$\begin{aligned} \zeta &= x/L, \qquad \beta = \alpha L, \qquad \gamma = a/L, \\ \nu(\lambda) &= N(\lambda)/N, \qquad \omega = D/sL, \end{aligned} \tag{20}$$

and define the function

$$\varphi(z) = (e^z - 1)/z. \tag{21}$$

Then the photodensity can be written

$$n(x)_p = (N\tau/L)F(\zeta,L),$$

where $F(\zeta,L)$ is the *photodensity function*

$$F(\zeta,L) = \zeta e^{-\zeta} \int_0^{\lambda_G} \nu(\lambda) \frac{\beta}{1 + \beta} \varphi[\zeta(1 - \beta)] \, d\lambda. \tag{22}$$

This function is shown in Fig. 1 as a continuous function of ζ for three values of L (10^{-4}, 10^{-3}, 10^{-2} cm) based on the solar spectrum and the absorption curve for silicon.[32]† The calculation of $F(\zeta,L)$ is discussed in the Appendix. For small ζ the photodensity function has the expansion

$$F(\zeta,L) = \zeta \int_0^{\lambda_G} \nu(\lambda) \frac{\beta(\lambda)}{1 + \beta} \, d\lambda - \tfrac{1}{2}\zeta^2 \int_0^{\lambda_G} \nu(\lambda)\beta(\lambda) \, d\lambda + \cdots. \tag{23}$$

† Values of $\alpha > 10^5$ cm^{-1} required as discussed in the Appendix for the calculation of Fig. 5 are from Pfestorf.[33]

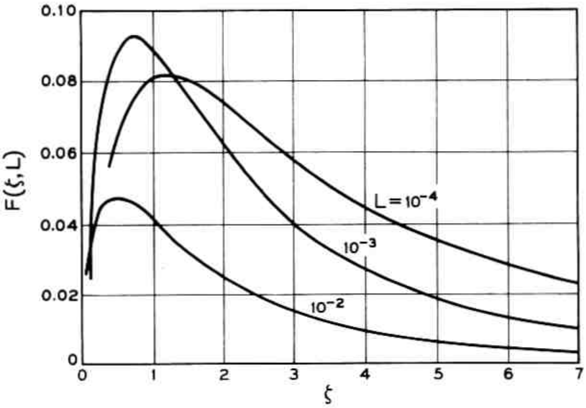


Fig. 1 — The photodensity function (22) for silicon as a function of ζ for $L = 10^{-4}$, 10^{-3} and 10^{-2} cm.

Once $F(\zeta,L)$ has been obtained, the solution $n(x)$ satisfying (1), (3) and (4) is readily constructed from $F(\zeta,L)$ and the functions $e^{\pm\zeta}$, which satisfy the homogeneous equation (5). The solution may be written

$$n(x) = (N\tau/L)[Ae^{\zeta} + Be^{-\zeta} + F(\zeta,L)], \tag{24}$$

where for $0 \leqq x \leqq a$

$$\begin{aligned} A &= -\frac{1}{2}\,\frac{F(\lambda)(1 + \omega) + \omega e^{-\gamma}F'(0)}{\sinh\gamma + \omega\cosh\gamma}, \\ B &= \frac{1}{2}\,\frac{F(\gamma)(1 - \omega) + \omega e^{\gamma}F'(0)}{\sinh\gamma + \omega\cosh\gamma}, \end{aligned} \tag{25}$$

and for $a \leqq x < \infty$

$$\begin{aligned} A &= 0, \\ B &= -e^{\gamma}F(\gamma), \end{aligned} \tag{26}$$

where $F' = dF/d\zeta$ and for brevity L has been omitted as an argument of F. The solution in the region $a \leqq x < \infty$ is, of course, independent of ω. In the region $0 \leqq x \leqq a$ the solutions in the limiting cases $\omega \to 0$ and $\omega \to \infty$ are

$$n(x) \underset{\omega\to 0}{\longrightarrow} \frac{N\tau}{L}\left[F(\zeta) - \frac{\sinh\zeta}{\sinh\gamma}F(\gamma)\right], \tag{27}$$

$$n(x) \underset{\omega\to\infty}{\longrightarrow} \frac{N\tau}{L}\left[F(\zeta) - F(\gamma) + \frac{\sinh(\gamma - \zeta)}{\cosh\gamma}F'(0)\right]. \tag{28}$$

The boundary conditions satisfied by these solutions at $x = 0$ are $n(0) = 0$ and $n'(0) = 0$ respectively.

The collection efficiency for single junction cells can now be obtained from (6) in the following form:

$$Q_1 = \frac{F(\gamma)(1 + \omega)\, e^{\gamma} + \omega F'(0)}{\sinh \gamma + \omega \cosh \gamma}. \tag{29}$$

In the limits $\omega \to 0$ and $\omega \to \infty$ this reduces to

$$\begin{aligned} Q_1 &\underset{\omega \to 0}{\to} F(\gamma)(1 + \coth \gamma), \\ Q_1 &\underset{\omega \to \infty}{\to} F(\gamma)(1 + \tanh \gamma) + F'(0) \operatorname{sech} \gamma, \end{aligned} \tag{30}$$

which refer to fast and slow surface recombination respectively.

Much of the simplicity of (29) and (30) results from the assumption of "effective" values for D and τ which apply to both sides of the junction. The difficulty of measuring these parameters in an actual cell, especially as a function of depth, justifies this simplification for most considerations. The most conspicuous case where this assumption may lead to serious error is in some of the compound semiconductors where the mobility and diffusion constant may be an order of magnitude less for holes than for electrons.† When different L obtain on the two sides of the junction (24), (25) and (26) are still formally valid but (29) is no longer valid. The collection efficiency can then be written $Q_1 = Q_- + Q_+$, where

$$\begin{aligned} Q_- &= \frac{F(\gamma)(\cosh \gamma + \omega \sinh \gamma) + \omega F'(0)}{\sinh \gamma + \omega \cosh \gamma} - F'(\gamma), \\ Q_+ &= F(\gamma) + F'(\gamma) \end{aligned} \tag{31}$$

are the contributions from $0 \leqq x \leqq a$ and $a \leqq x < \infty$ respectively. If the same L is used for both sides of the junction $Q_- + Q_+$ reduces to (29).

The fast surface recombination limit $\omega = 0$ is valid if ω satisfies the two conditions $\omega \ll 1$, $\omega \ll \gamma$. In sillicon solar cells the junction depth is considerably less than the diffusion length ($\gamma \ll 1$) so that the second condition implies the first. Therefore the criterion for high surface recombination velocity is

$$s \gg D/a. \tag{32}$$

The diffusion constant D can be obtained from measured mobilities μ by

† Ref. 24, p. 12.

means of the Einstein relation,[34] which in conventional laboratory units is

$$D \text{ (in cm}^2\text{/second)} = 0.026 \left(\frac{T}{300°\text{K}}\right) \mu \text{ (in cm}^2\text{/volt-second)}, \quad (33)$$

with T the absolute temperature in °K. The mobility in highly doped silicon has been studied by Backenstoss,[35] who finds that for impurity concentrations greater than 10^{19} cm^{-3} the mobilities of electrons and holes approach the limiting values of 80 and 40 cm^2/volt-second respectively. Since $a \sim 10^{-4}$ cm, the criterion (32) gives $s > 10^5$ cm/second in silicon. Although surface treatments are known[21] which can reduce s in pure silicon to below 10^2 cm/second, Prince and Wolf[6] point out that these treatments tend to increase significantly the reflectivity of the surface. If this occurred the treatment might well reduce rather than increase the over-all efficiency of the cell. On the other hand, Malovetskaya, et al.[16] have reported antireflection films that greatly reduce reflection losses without increasing the surface recombination. The effect of surface recombination on Q_1 is shown in Fig. 2 for a silicon solar cell having the typical parameters

$$\begin{aligned} D &= 5 \text{ cm}^2\text{/second}, \\ \tau &= 2 \times 10^{-5} \text{ second}, \\ L &= 10^{-2} \text{ cm}, \\ a &= 10^{-4} \text{ cm}. \end{aligned} \quad (34)$$

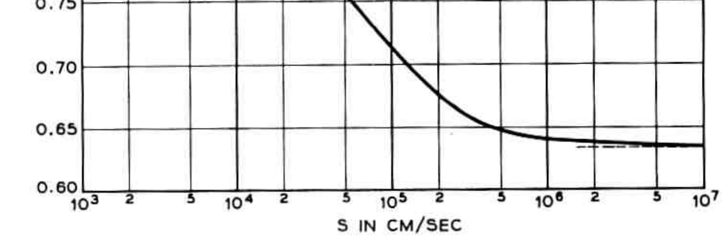


Fig. 2 — The collection efficiency as a function of surface recombination velocity for a silicon cell having the parameters (34).

For this cell Q_1 increases from 63 to 87 per cent as s decreases from 10^7 to 10^3 cm/second. This example shows that considerable improvement in Q_1 would result from the elimination ($s < 10^3$ cm/second) of surface recombination. The presence of surface recombination in commercial silicon cells is indicated by the rapid fall-off in the spectral response[10,18] at short wavelength $\lambda < 6\ \mu$. Little is known about surface recombination in the compound semiconductors. The fact that body lifetimes tend to be short[24] ($\tau \sim 10^{-8}$ second) suggests that surface recombination is probably very fast ($s \sim 10^7$ cm/second). On the basis of these considerations, the most reliable guess is to set $\omega = 0$ except where specific information to the contrary is available.

The dependence of Q_1 upon the junction depth is shown in Fig. 3 for a silicon cell with D, τ and L as given in (34). The curves shown for $\omega = 0$ and $\omega \to \infty$ are envelopes for the entire family $Q_1(\gamma,\omega)$. Although the limiting curves have quite different forms they approach the same form for small and large depths:

$$\begin{aligned} Q_1 \underset{\gamma\to 0}{\longrightarrow} F'(0) &= \int_0^{\lambda_G} \nu(\lambda)\beta(\lambda)(1+\beta)^{-1}\,d\lambda, \\ Q_1 \underset{\gamma\to\infty}{\longrightarrow} 2F(\gamma).& \end{aligned} \tag{35}$$

The monotonic decrease of $Q_1(\gamma,0)$ with increasing γ is typical of silicon cells. In semiconductors of considerably higher energy gap the condition $F'(0) + \frac{1}{2}F''(0) > 0$ may be satisfied, in which case $Q_1(\gamma,0)$ will at first

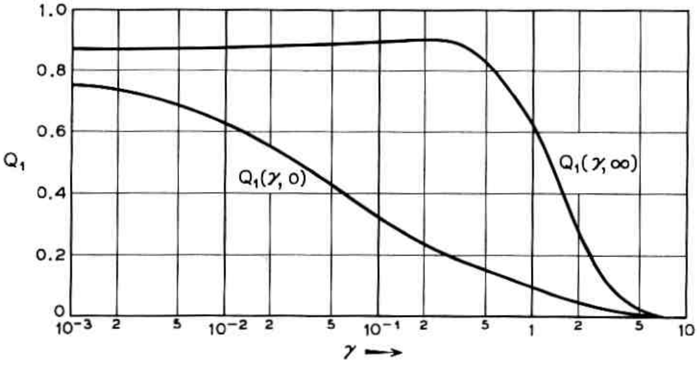


Fig. 3 — The collection efficiency as a function of junction depth for a silicon cell with parameters D,τ,L given by (34). The abscissa is $\gamma = a/L$, where a is the junction depth and L the diffusion length. The curve $Q_1(\gamma,\infty)$ refers to zero and $Q_1(\gamma,0)$ to infinite surface recombination velocity.

increase with γ to reach a maximum and then fall to zero as γ is increased further. A very broad maximum in $Q_1(\gamma, \infty)$ is seen in Fig. 3 near $\gamma = 0.2$. The shape of the curve shows that relatively deep junctions ($\gamma \sim 0.5$) could be used if surface recombination were absent, which would permit the internal resistance of the cell to be greatly reduced. On the other hand, deep junctions cannot be used if the surface recombination is important. This calculation also shows that in a cell with surface recombination reducing a from 10^{-4} ($\gamma = 0.01$) to 5×10^{-5} cm ($\gamma = 0.005$) would improve the collection efficiency only slightly. Therefore, one may regard $a = 10^{-4}$ cm as close to the optimum junction depth for silicon solar cells.

The dependence of Q_1 on diffusion length is shown in Fig. 4 for a silicon cell with a junction depth $a = 10^{-4}$ cm. The curves $Q_1(L,\infty)$ and $Q_1(L,0)$ are envelopes for the family $Q_1(L,\omega)$ considered as continuous functions of L. These calculations show that $L > 0.1$ cm may be considered essentially infinite. The limiting values as $L \to \infty$ are

$$\begin{aligned} Q_1(L,\infty) &\xrightarrow[L\to\infty]{} 1, \\ Q_1(L,0) &\to \int_0^{\lambda_G} \nu(\lambda)\, \frac{1 - e^{-\alpha a}}{\alpha a}\, d\lambda, \end{aligned} \tag{36}$$

which are indicated in the figure. It will be observed that Q_1 increases relatively little as L increases beyond 10^{-2} cm. On the other hand considerable improvement results from increasing L from 10^{-3} to 10^{-2} cm.

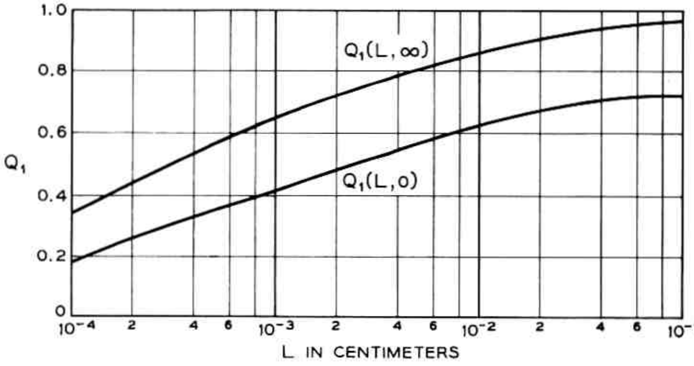


Fig. 4 — The collection efficiency as a function of diffusion length for a silicon cell with the junction depth 10^{-4} cm. The curve $Q_1(L,\infty)$ refers to zero and $Q_1(L,0)$ to infinite surface recombination velocity.

IV. THE EFFECT OF AN ELECTRIC FIELD

In his review, Rappaport[30] has suggested that a "built-in" electric field may exist near the surface of solar cells causing minority carriers to drift toward the junction. If sufficiently strong this electric field would in effect eliminate surface recombination. One would expect that to reduce surface recombination significantly the field would have to be of the order $E \sim s/\mu \sim 10^3$ volts/cm, assuming the typical values $s \sim 2 \times 10^5$ cm/second, $\mu \sim 200$ cm^2/volt-second. The built-in electric field could arise in two ways, from a gradient in the impurity concentration, and from a gradient in the energy gap. In the first case the field is given by[36]

$$E \sim (0.026/a) \ln (N_0/N_a) \quad \text{volts/cm}, \tag{37}$$

where N_0, N_a are the carrier concentrations at $x = 0$ and $x = a$ respectively, and $0.026 = kT/e$ at room temperature. This is just the field required to cancel the diffusion current in equilibrium according to the Einstein relation (33). Actually (37) is an approximation giving an average effective field over the region $0 \leqq x \leqq a$. Consider as an example a cell made by the diffusion of boron from the vapor phase into n-type silicon.[6,10,11] The surface concentration of boron is about[11] $N_0 \sim 18^{18}$ cm^{-3} and $N_a \sim 10^{17}$ cm^{-3}. If $a \sim 10^{-4}$ cm, the built-in field is $E \sim 500$ volts/cm, which is probably much less than s/μ and therefore not large enough to cause much reduction in surface recombination. We may conclude that in commercial solar cells the built-in field can be neglected, as in the treatment of the last section and in the previous literature. On the other hand, the possibility remains that much larger built-in fields could be obtained, since N_0 might be made to approach the solid solubility,[37] which for boron exceeds 10^{20} cm^{-3}. The built-in field would then be about $E \sim 1800$ volts/cm, which might cause a significant improvement in collection efficiency.

It appears that really large fields $E \gg s/\mu$ might be obtained from gradients in the energy gap. Such a gradient could be obtained in GaAs by diffusing in phosphorous from the surface. Mixed crystals[38]† GaAs-GaP can exist in all proportions, and the energy gap varies as a function of composition from 1.4 (GaAs) to 2.4 (GaP). If the impurity concentration is approximately constant, the band edge corresponding to the majority carriers must also be approximately constant. The gradient of the energy gap therefore appears almost entirely in the band edge for the minority carriers, and is equivalent to a built-in electric field acting

† See also Ref. 24, p. 52.

on the minority carriers. It seems quite feasible that the field may be of order $E \sim 10^4$ volts/cm in GaAs-P. In addition to producing the field, the gradient in energy gap causes the absorption coefficient to become a function of position $\alpha(\lambda,x)$. To make the analysis more tractable we may regard $\alpha(\lambda)$ as an "effective" absorption coefficient independent of position.

To take into account a constant electric field it is only necessary to add to the left side of (1) the term $-\mu E(dn/dx)$, giving

$$D\frac{d^2n}{dx^2} - \frac{n}{\tau} - \mu E\frac{dn}{dx} + \int_0^{\lambda_G} N(\lambda)\alpha(\lambda)e^{-\alpha x}\,d\lambda = 0. \tag{38}$$

This is conveniently written in the reduced form

$$\frac{d^2y}{d\zeta^2} - y - 2\mathcal{E}\frac{dy}{d\zeta} + \int_0^{\lambda_G} \nu\beta e^{-\beta\zeta}\,d\lambda = 0, \tag{39}$$

where

$$\mathcal{E} = \frac{E}{2D/\mu L},$$
$$y = \frac{n}{N\tau/L}, \tag{40}$$

and β,ν,ζ are defined in (20). The boundary conditions on $y(\zeta)$ are

$$y'(0) = Sy(0), \qquad y(\gamma) = 0, \qquad y(\infty) = 0, \tag{41}$$

where

$$S = 2\mathcal{E} + \omega^{-1} \tag{42}$$

depends on the electric field as well as the surface recombination velocity. If the last term in (39) is dropped the resulting homogeneous equation has the solution $e^{-\rho\zeta}$, $e^{\sigma\zeta}$, where

$$\rho = (1+\mathcal{E}^2)^{\frac{1}{2}} - \mathcal{E} \leqq 1,$$
$$\sigma = (1+\mathcal{E}^2)^{\frac{1}{2}} + \mathcal{E} \geqq 1, \tag{43}$$
$$\rho\sigma = 1.$$

The general solution can be written

$$y(\zeta) = Be^{-\rho\zeta} + Ae^{\sigma\zeta} + F(\zeta,L,\mathcal{E}), \tag{44}$$

where

$$F(\zeta,L,\mathcal{E}) = \int_0^{\lambda_G} \nu \frac{\beta}{1 + \rho\beta} \frac{e^{-\beta\zeta} - e^{-\rho\zeta}}{1 - \sigma\beta} \, d\lambda. \tag{45}$$

For the region $0 \leqq \zeta \leqq \gamma$,

$$\begin{aligned} B &= \frac{F'(0) + (S - \sigma)F(\gamma)e^{-\sigma\gamma}}{\rho + \sigma e^{-(\rho+\sigma)\gamma} + S(1 - e^{-(\rho+\sigma)\gamma})} \\ A &= -\frac{F'(0)e^{-(\rho+\sigma)\gamma} + (S + \rho)F(\gamma)e^{-\sigma\gamma}}{\rho + \sigma e^{-(\rho+\sigma)\gamma} + S(1 - e^{-(\rho+\sigma)\gamma})} . \end{aligned} \tag{46}$$

As $\mathcal{E}$ approaches zero, $F(\zeta,L,\mathcal{E})$ approaches $F(\zeta,L)$ defined by (22) and A,B of (46) go over into A,B of (25). It will be assumed that the electric field exists only in the region $0 \leqq \zeta \leqq \gamma$. Thus the solution in the region $\zeta \geqq \gamma$ is

$$y(\zeta) = F(\zeta,L) - F(\gamma,L)e^{\gamma-\zeta}. \tag{47}$$

Since y vanishes at the junction, the current is just the diffusion current and the collection efficiency is $Q_1 = Q_- + Q_+$, where

$$\begin{aligned} Q_+ &= y'(\gamma_+), \\ Q_- &= -y'(\gamma_-). \end{aligned} \tag{48}$$

From (47),

$$Q_+ = F(\gamma_+) + F'(\gamma_+) \tag{49}$$

as in (31). From (44) and (46),

$$Q_- = \frac{F(\gamma)[1 - e^{-(\rho+\sigma)\gamma} + S(\sigma + \rho e^{-(\rho+\sigma)\gamma})] + F'(0)e^{-\rho\gamma}(\rho + \sigma)}{\rho + \sigma e^{-(\rho+\sigma)\gamma} + S(1 - e^{-(\rho+\sigma)\gamma})} - F'(\gamma_-), \tag{50}$$

where $F'(\zeta) = (d/d\zeta)F(\zeta,L,\mathcal{E})$. In the limit $\mathcal{E} \to 0$ this goes over into Q_- given in (31). If $2\mathcal{E} \gg \beta$, $2\mathcal{E} \gg 1/\beta$ and $2\mathcal{E} \gg \gamma$, (50) becomes

$$Q_1 \underset{\mathcal{E}\to\infty}{\longrightarrow} \int_0^{\lambda_G} \nu(1 - e^{-\beta\gamma}) \, d\lambda - \frac{1}{2\mathcal{E}} \int_0^{\lambda_G} \nu\beta \, e^{-\beta\gamma} \, d\lambda. \tag{51}$$

Since the first term of (51) is the probability of absorbing a photon in the layer $0 \leqq \zeta \leqq \gamma$, the limit $\mathcal{E} \to \infty$ corresponds to complete collection of minority carriers produced in this layer.

A plot of Q_- as a function of $\mathcal{E}$ is shown in Fig. 5 for a silicon cell with the parameters

$$\begin{aligned} D &= 5 \text{ cm}^2/\text{second}, \\ \mu &= 190 \text{ cm}^2/\text{volt-second}, \\ \tau &= 2 \times 10^{-5} \text{ second}, \\ L &= 10^{-2} \text{ cm}, \\ a &= 10^{-4} \text{ cm}, \\ s &= 2 \times 10^5 \text{ cm/second}, \\ 2D/\mu L &= 5.3 \text{ volts/cm}, \\ s/\mu &= 1050 \text{ volts/cm}. \end{aligned} \tag{52}$$

The electric field is therefore $E = 5.3\mathcal{E}$ volts/cm. The value of Q_+ corresponding to these parameters is 0.52, so that the collection efficiency $Q_1 = Q_- + Q_+$ varies from 0.67 at zero field to 0.87 at infinite field. It will be observed that these values correspond in Fig. 2 to $s = 2 \times 10^5$ cm/second and $s \to 0$ respectively, as one would expect. The critical field s/μ indicated in the figure is the field at which about half the surface

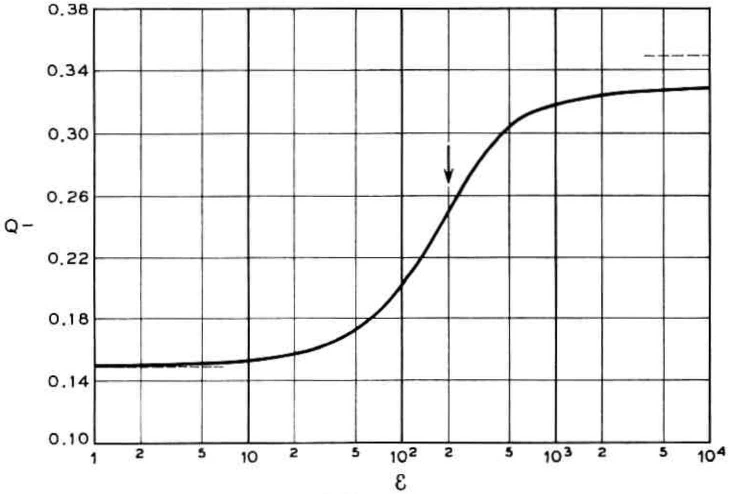


Fig. 5 — The collection efficiency as a function of built-in electric field for a silicon cell with parameters (52). The abscissa is $\mathcal{E} = E\mu L/2D$ and $2D/\mu L = 5.3$ volts/cm. The critical field s/μ and the asymptotic limits of Q_1 are indicated by dotted lines.

recombination has been eliminated. Although it would be difficult to obtain built-in fields much larger than this in silicon, the high-field portion of the curve shows qualitatively the results to be expected from a gradient in the energy gap in a solar cell made of a compound semiconductor such as GaAs-P. The asymptotic formula (51) holds only for much higher fields than $\mathcal{E} = 10^4$ because of the extremely high values[33] of α near $\lambda \sim 0.3\ \mu$. The asymptotic limits at low and high fields are indicated by dotted lines. For estimating the effects of a given field it is convenient to regard the field as effectively lowering the surface recombination velocity. For estimating the effective velocity s^* one may use the recipe

$$s^* = s[1 + (E\mu/s)(1 + sa/D)]^{-1}. \tag{53}$$

The collection efficiency can then be estimated from (29) by using an effective surface parameter $\omega^* = D/Ls^*$ with considerably less labor than by evaluating (50) exactly.

V. THE EFFECT OF THE ABSORPTION CURVE

The absorption curve is the curve obtained by plotting the absorption coefficient $\alpha(\lambda)$ as a function of wavelength λ. Different semiconductors have quite different absorption curves as well as different energy gaps. In the several studies that have been made on the effect of the energy gap no consideration has been given to the absorption curve.[8,11,13,14,15,30] All of these studies have been concerned with spectrum and junction losses rather than collection losses. From the standpoint of collection efficiency, however, the absorption curve is of essential importance if there is any considerable surface recombination in the cell. Although the elimination of surface recombination may be in prospect, it is still of interest to consider the effect of the absorption curve in a cell with essentially infinite surface recombination velocity. This provides another standpoint from which to discuss the advantages of different semiconductors.

In Fig. 6 are shown the absorption curves for silicon,[32] GaAs† and InP,[39] in the range $10^2 \leqq \alpha < 10^5$ cm^{-1}. It will be observed that the curve for silicon differs from the other two in its gradual drop off with increasing wavelength. For $\alpha < 10^2$ cm^{-1} it falls off more steeply, although not as steeply as GaAs or InP. The reciprocal $\alpha(\lambda)^{-1}$ of the absorption coeffi-

† The solid curve for GaAs is based on unpublished transmission measurements by W. G. Spitzer. The dashed part of the curve is an extrapolation based on a single additional reflection measurement by R. J. Archer giving $\alpha \sim (1 \pm 0.3) \times 10^5$ cm^{-1} at 0.546 μ.

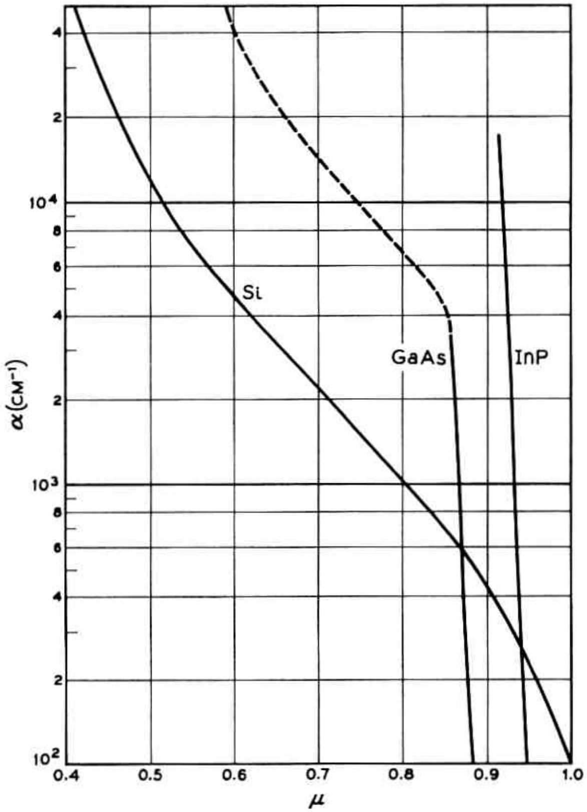


Fig. 6 — The absorption curves for silicon, GaAs and InP.

cient is a measure of the penetration distance of light of wavelength λ. There is some wavelength λ_1, for which

$$\alpha(\lambda_1)^{-1} \sim a, \tag{54}$$

which is the shortest wavelength of light useful in the cell. Light of shorter wavelength that λ_1 is absorbed too close to the surface to be collected at the junction. Similarly, there is a wavelength λ_2, for which

$$\alpha(\lambda_2)^{-1} \sim L, \tag{55}$$

which is the longest wavelength of useful light. If $L = 10^{-2}$ cm or larger λ_2 is nearly equal to the intrinsic absorption edge λ_G corresponding to the energy gap. Light of longer wavelength than λ_2 is absorbed too far from the junction to be collected, or else is not absorbed at all. Since the effective light lies in the range $\lambda_1 \leqq \lambda \leqq \lambda_2$ an effective photon flux can be defined:

$$N_{\text{eff}} = \int_{\lambda_1}^{\lambda_2} N(\lambda)\, d\lambda. \tag{56}$$

The collection efficiency may be written

$$Q_1 = N_{\text{eff}}/N, \tag{57}$$

where the flux N is defined by (2). The limits λ_1, λ_2, and hence Q, evidently depend in an essential way on the absorption curve.

The first equation (30) gives Q_1 for the present case in terms of the photodensity function $F(\zeta,L)$ defined by (22). Therefore, the qualitative considerations given above enter into determining the value of $F(\gamma,L)$. According to (22), $F(\gamma,L)$ contains an integral over λ of the product of two functions $\beta/(1 + \beta)$ and $\varphi[\gamma(1 - \beta)]$. The weighting function $\nu(\lambda)$ representing the solar spectrum is slowly varying and not relevant to the present discussion. Fig. 7 shows plots of the two functions for a silicon cell with the parameters (34). It is clear from this figure that λ_1 and λ_2 correspond to cutoffs in φ and $\beta/(1 + \beta)$ respectively. The integrand in (22) is evidently small except in the region between the vertical lines in the figure corresponding to $\alpha a = 1$ and $\alpha L = 1$.

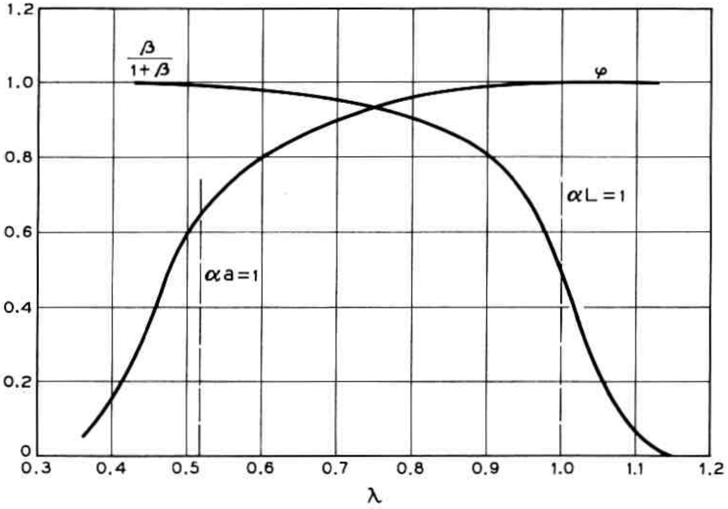


Fig. 7 — The functions $\beta/(1 + \beta)$ and φ entering into the integrand of (22) defining the photodensity function. Calculation refers to a silicon cell with $a = 10^{-4}$, $L = 10^{-2}$ cm. Dashed lines show cutoffs λ_1, λ_2 corresponding to (54) and (55).

It is now clear that when λ_1 lies very close to λ_2 the collection efficiency will be relatively low. This is the case with very steep absorption curves like that of InP. On the other hand, a gradual absorption curve like that of silicon leads to relatively high collection efficiency. If λ_1, λ_2 are defined by $\alpha(\lambda_1) = 10^4$ and $\alpha(\lambda_2) = 10^2 \text{ cm}^{-1}$, (57) predicts the collection efficiencies 0.70, 0.21 and 0.045 for silicon, GaAs and InP respectively. The value for GaAs is subject to considerable uncertainty because it is based on a little-known part of the absorption curve. In any case, all of these numbers are qualitative, and mainly of interest to show that if surface recombination is high silicon is much superior to GaAs or InP. It is important to notice that this should be true in spite of the higher energy gap of the other two materials, because the differences in efficiency predicted for these materials on the basis of their energy gaps are much smaller than the differences in collection efficiency obtained here. For example, Loferski[15] finds that silicon and GaAs have relative efficiencies of 0.21 and 0.24 respectively, whereas silicon may have three times the collection efficiency of GaAs and 15 times that of InP.

Several investigators have studied GaAs photocells. Gremmelmaier[40] obtained an over-all efficiency of about 4 per cent and a collection effidiency of about 0.2, in agreement with the estimates given above. An over-all efficiency of 6.5 per cent for a cell of very small area (0.007 cm^2) has been reported by Jenny, Loferski and Rappaport.[41] These authors give current and voltage information for a somewhat larger cell (0.059 cm^2) having an over-all efficiency of 3.2 per cent, from which it can be deduced that the collection efficiency was about 0.26. It may be significant that the efficiency of 6.5 per cent was apparently obtained using a very low level of illumination (0.0057 watt/cm^2). Also, very low collection efficiencies have been reported for InP cells by Rappaport,[30] although he attributes this to poor ohmic contacts to the cell. According to the present argument GaAs and InP will always give low collection efficiencies unless surface recombination can be eliminated, either by reducing the surface recombination velocity to below 10^4 cm/second, or by means of a built-in electric field of order 10^4 volts/cm.

It should be mentioned that the spectral response for a GaAs cell reported by Jenny et al.[41] is not at all in agreement with the assumption of fast surface recombination. These authors observe an almost constant quantum efficiency from 5 to 9 μ. This differs radically from an earlier measurement by Seraphin,† which showed a pronounced fall-off at short wavelengths characteristic of surface recombination. No explanation has been advanced for this result nor can any be offered here. It may mean

† Ref. 24, p. 65.

that surface recombination can be effectively eliminated in GaAs cells by treatments not yet defined, understood or controlled.

VI. THE TWO-JUNCTION CELL

In the typical silicon solar cell having the parameters (52), about 20 per cent of the minority carriers produced are lost to surface recombination and 13 per cent to body recombination. Therefore large reductions in surface recombination will be more rewarding in terms of over-all efficiency improvement than will proportionate reductions in body recombination. The two preceding secitons have been concerned primarily with surface recombination. It is also worthwhile to consider body recombination in some detail, particularly since high-temperature applications may be in prospect which will require high energy gap material.[42] As discussed in the introduction, body lifetimes tend to be short in these materials. From Fig. 4 one may estimate that Q_1 is likely to be no larger than 0.2 in a cell with short lifetime. How this low collection efficiency may be improved by means of a second junction is discussed in this section.

The collection efficiency for a two-junction cell will be denoted Q_2. If the second junction is located at depth b below the surface, and $\eta = b/L$, the solution $n(x)$ in the region $a \leqq x \leqq b$ is

$$n(x) = \frac{N\tau}{L}\left[F(\zeta) - \frac{\sinh(\eta-\zeta)}{\sinh(\eta-\gamma)}F(\gamma) - \frac{\sinh(\zeta-\gamma)}{\sinh(\eta-\gamma)}F(\eta)\right], \quad (58)$$

which satisfies (1) and the boundary conditions $n(a) = n(b) = 0$. In the region $x \geqq b$ the solution is

$$n(x) = \frac{N\tau}{L}[F(\zeta) - e^{\eta-\zeta}F(\eta)]. \quad (59)$$

The solution in the region $0 \leqq x \leqq a$ is the same as that obtained in Section III and given by (24) and (25). For the case of fast surface recombination $\omega = 0$, this is

$$n(x) = \frac{N\tau}{L}\left[F(\zeta) - \frac{\sinh\zeta}{\sinh\lambda}F(\gamma)\right], \quad (60)$$

satisfying $n(0) = n(a) = 0$. From (6), the total collection efficiency is

$$Q_2 = F(\gamma)[\coth\gamma + \tanh\tfrac{1}{2}(\eta-\gamma)] + F(\eta)[1 + \tanh\tfrac{1}{2}(\eta-\gamma)]. \quad (61)$$

As $\eta \to \gamma$, $Q_2 \to Q_1$ for the case $\omega = 0$ given by (30). The effectiveness of the second junction may be measured by the quantiy

$$\delta = (Q_2 - Q_1)/Q_1, \quad (62)$$

where Q_1 is the collection efficiency for a cell with a single junction at depth a. From (61),

$$\delta(\eta,\gamma,L) = \frac{F(\eta)}{F(\gamma)} \frac{1 + \tanh \frac{1}{2}(\eta - \gamma)}{1 + \coth \gamma} - \frac{1 - \tanh \frac{1}{2}(\eta - \gamma)}{1 + \coth \gamma}. \quad (63)$$

Fig. 8 shows $\delta(\eta,\gamma,L)$ plotted as a function of η for a silicon cell with $a = 10^{-4}$ cm and $L = 10^{-4}$, 10^{-3} and 10^{-2} cm. These calculations show that $\delta(\eta)$ has a well-defined maximum which defines the optimum depth for the second junction for given a,L.

Due to the loading effect of the second junction the relative improvement in over-all efficiency will be somewhat less than δ. This effect may be taken into account by use of the Pfann-Van Roosbroeck[12] efficiency expression (15). For the two-junction cell the reduced current may be written

$$G_2 = \tfrac{1}{2}(Q_2/Q_1)G_1 , \quad (64)$$

where $G_1 = I_g/eAJ_0 = (N/J_0)Q_1$ is the reduced current for the single-junction cell. Thus the efficiency expression for the two-junction cell is

$$\epsilon_2 \approx \frac{kT}{W} Q_2 \ln G_2 . \quad (65)$$

The relative improvement in over-all efficiency can be measured by

$$\delta' = (\epsilon_2 - \epsilon_1)/\epsilon_1 , \quad (66)$$

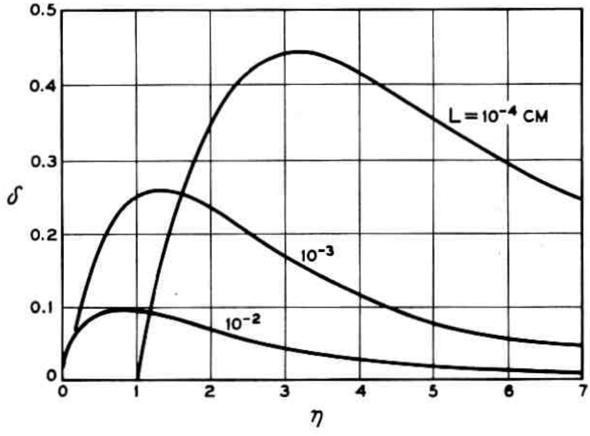


Fig. 8 — The improvement of collection efficiency (defined by (62)) as a function of the depth of the second junction for a silicon cell with first junction at depth 10^{-4} and $L = 10^{-4}$, 10^{-3} and 10^{-2} cm.

where ϵ_1 given by (15) is the efficiency of the single-junction cell with the junction at depth a. Finally, the over-all improvement δ' can be written as a function of the collection improvement δ:

$$\delta' = \delta + (1 + \delta) \frac{\ln \frac{1}{2}(1 + \delta)}{\ln G}. \tag{67}$$

For practical purposes one may set $\ln G = 18$ in (67). The second term gives the effect of the loading due to the second junction. Two extreme cases may be noted: if $\delta > 1$, the second junction is a more effective collector than the first function and $\delta' > \delta$; if $\delta < \ln \frac{1}{2}/\ln G \sim 0.04$, the second junction acts mainly as a load on the first junction and $\delta' < 0$. A summary of calculations for silicon two-junction cells is presented in Table I. In each case it is assumed that the first junction is at $a = 10^{-4}$ cm and the second junction at the optimum depth given in the third column. The collection improvement δ and over-all improvement δ' are given in the fourth and fifth columns. In the second column is given the single junction collection efficiency in agreement with $Q_1(L,0)$ of Fig. 4. The last row with $L = 10^{-2}$ cm is the most typical of good silicon solar cells. For this case the improvement in collection efficiency is only $\delta = 0.092$. On the other hand, for $L = 10^{-4}$ cm, $\delta = 0.44$ and $\delta' = 0.42$. This case applies qualitatively to any short-lifetime cell, and shows that the two-junction structure may be useful for improving the efficiency of high energy gap cells.

VII. SUMMARY

In this paper the present status of solar cell theory has been reviewed, with emphasis on clearly defining the various mechanisms causing losses of efficiency. This review leads to the conclusion that the problem of the collection of minority carriers has not received attention in the literature commensurate with its importance.

The collection problem is formulated and then solved by a new method in which all the effects of the solar spectrum and absorption curve are contained in a single function, the photodensity function. The method is convenient for most solar cell collection problems and especially so when a single diffusion length can be used.

TABLE I

L (in cm)	Q_1	b (in cm)	δ (max)	δ'
10^{-4}	0.186	0.032	0.443	0.417
10^{-3}	0.420	0.013	0.258	0.226
10^{-2}	0.634	0.008	0.092	0.055

The familiar single-junction silicon cell is considered first and calculations are presented to show how the collection efficiency varies with surface recombination velocity, junction depth and diffusion length. It is found that the elimination of surface recombination would not only improve collection efficiency but also permit the internal resistance to be greatly reduced. Little or no improvement in silicon cells is to be expected from making the junction depth less than 10^{-4} cm or the diffusion length longer than 10^{-2} cm.

The theory is extended to include a constant built-in electric field, and calculations are presented for silicon. It is concluded that commercial silicon cells do not have large enough built-in fields to affect the collection efficiency significantly. Sufficiently large fields should be obtainable in some compound semiconductors from gradients in the energy gap. An approximate relation is given for an effective surface recombination velocity less than the true velocity, which takes into account the effect of the built-in field in reducing surface recombination.

A discussion is given of the dependence of the collection efficiency on the absorption curve of the semiconductor. This discussion provides another basis beside the energy gap on which to compare different semiconductors for solar battery use. It is shown that silicon has a very favorable absorption curve in comparison with GaAs and InP. It is suggested that this accounts to a large extent for the continuing superiority of silicon over other materials with larger energy gaps. The theoretical superiority of GaAs and several other higher gap materials over silicon can only be realized if surface recombination can be drastically reduced; specifically, it is necessary that the effective surface recombination velocity be reduced to 10^4 cm/second or lower.

Finally, a two-junction cell is considered in connection with the problem of reducing body recombination. Although body recombination is considerably less important than surface recombination in good silicon cells, it will be probably much more important in high energy gap cells. The illustrative calculations presented for silicon also have qualitative significance for other cells. It is shown that an improvement of about 42 per cent in over-all efficiency may be expected from the two junction structure in material with a diffusion length of 10^{-4} cm. The improvement in a good silicon cell with a diffusion length of 10^{-2} cm would only be 6 per cent.

VIII. ACKNOWLEDGMENT

It is a pleasure for the author to acknowledge the helpful cooperation of W. G. Spitzer, G. L. Pearson, F. M. Smits and G. C. Dacey of Bell Telephone Laboratories.

Note. Since the completion of this work a review paper by Wolf[43] has appeared which discusses some of the topics taken up here.

APPENDIX

In this section some comments are given which may be helpful for the evaluation of the photodensity function (22). A very accurate and convenient method is the Gauss quadrature,[44] which approximates the integral by a summation over the values of the integrand at certain specific (not equidistant) points. The limits $0,\lambda_G$ in (22) can be replaced by λ_1, λ_2 so chosen that the contributions from the regions $0 < \lambda < \lambda_1$ and $\lambda_2 < \lambda < \lambda_G$ are negligible. The approximation is to represent the integral

$$\begin{aligned} F(\zeta,L) &= \int_0^{\lambda_G} g(\lambda)\, d\lambda \\ &\cong \int_{\lambda_1}^{\lambda_2} g(\lambda)\, d\lambda \\ &= \tfrac{1}{2}(\lambda_2 - \lambda_1) \sum_{j=1}^{10} R_j g(\lambda_j). \end{aligned} \tag{68}$$

The choice of a 10-point Gauss quadrature here is arbitrary, but has proved to be a satisfactory compromise between convenience and accuracy. For *silicon* one can choose $\lambda_1 = 0.42\ \mu$, $\lambda_2 = 1.08\ \mu$, and (68) becomes

$$\int g(\lambda)\, d\lambda = \sum (0.33R_j) g(\lambda_j). \tag{69}$$

The wavelengths λ_j and corresponding absorption coefficients[32] $\alpha(\lambda_j)$ and solar spectrum[31] weights $\nu(\lambda_j)$ for this case are given in Table II, along with the quadrature weights $0.33R_j$. The integrand $g(\lambda)$ can be readily evaluated with a slide rule for each λ_j.

TABLE II

$\lambda(\mu)$	$0.33R$	α (in cm^{-1})	$\nu(\mu^{-1})$
0.429	0.0220	3.7×10^4	1.22
0.465	0.0492	2.0(4)	1.50
0.526	0.0723	9.0(3)	1.57
0.607	0.0888	4.3(3)	1.62
0.701	0.0977	2.2(3)	1.51
0.799	0.0977	1.03(3)	1.36
0.893	0.0888	4.5(2)	1.26
0.974	0.0723	1.56(2)	1.12
1.035	0.0492	42.	1.05
1.071	0.0220	17.	1.02

The accuracy of (69) may be tested with an elementary integral having a qualitative behavior similar to $g(\lambda)$ in the range of integration. Such an integral is

$$\int_{0.42}^{1.08} (1.08 - \lambda)^{\frac{1}{2}}(\lambda - 0.42)^{\frac{1}{2}}\, d\lambda = 0.17106.$$

The quardature (69) with λ_j and $0.33R_j$ from Table II gives 0.17111.

The derivative $F'(\zeta,L)$ is defined by the integral

$$\begin{aligned} F'(\zeta,L) &= e^{-\zeta} \int_0^{\lambda_G} \nu(\lambda)\, \frac{\beta}{1+\beta}\, \frac{1 - \beta\, e^{\zeta(1-\beta)}}{1-\beta}\, d\lambda \\ &= e^{-\zeta} \int_0^{\lambda_G} \nu(\lambda)\, \frac{\beta}{1+\beta}\, \{1 - \beta\zeta\varphi[\zeta(1-\beta)]\}\, d\lambda. \end{aligned} \tag{70}$$

This integral may be accurately evaluated by (69), providing $\beta(\lambda_1)\zeta \gg 1$ to insure that the truncation error at λ_1 is small. The photodensity function for a constant electric field defined by (45) can be written

$$F(\zeta,L,\mathcal{E}) = \rho\zeta\, e^{-\rho\zeta} \int_0^{\lambda_G} \nu\, \frac{\beta}{1+\rho\beta}\, \varphi[\rho\zeta(1-\sigma\beta)]\, d\lambda, \tag{71}$$

and its derivative is defined by

$$F'(\zeta,L,\mathcal{E}) = \rho\, e^{-\rho\zeta} \int_0^{\lambda_G} \nu\, \frac{\beta}{1+\rho\beta}\, \{1 - \beta\zeta\varphi[\rho\zeta(1-\sigma\beta)]\}\, d\lambda. \tag{72}$$

These expressions are in the form of (22) and (70) respectively, and can be evaluated by (69) in most cases.

Some of the integrals occurring in the theory cannot be evaluated with sufficient accuracy by (69) because of large contributions from the short wavelengths $\lambda < \lambda_1 = 0.42\ \mu$. As an example of this difficulty consider

$$\int_0^{\lambda_G} \nu(\lambda)\, d\lambda = 1, \tag{73}$$

which follows from the definition of $\nu(\lambda)$ in (20) and of N in (2). The quadrature (69) gives 0.905 for this integral, which shows that

$$\int_0^{0.42} \nu\, d\lambda = 0.095.$$

This result can be used to correct (69) for the evaluation of

$$F'(0,L) = \int_0^{\lambda_G} \nu\, \frac{\beta}{1+\beta}\, d\lambda \tag{74}$$

occurring in (31). If $\beta(\lambda_1) \gg 1$ it is only necessary to add 0.095 to the right side of (69); thus for *silicon*

$$F'(0,L) = 0.095 + \sum_{j=1}^{10} (0.33R_j)\nu_j \frac{\beta_j}{1+\beta_j}. \tag{75}$$

The evaluation of $F'(0,L,\mathcal{E})$ occurring in (50) is more difficult and requires a numerical integration over the larger region $0 < \lambda < \lambda_G$. To carry out the integration the absorption curve must be known for very high absorption[33] $\alpha \sim 10^6$ cm^{-1}. There may also be a significant correction to be added to (69) in the evaluation of $F(\zeta,L,\mathcal{E})$ for very high fields $\mathcal{E} > 100$. These corrections can be readily calculated by integrating from 0.22 to 0.42 μ by the trapezoid rule. The necessary data for silicon is listed in Table III. For large fields (72) becomes

$$F'(\zeta,L,\mathcal{E}) \underset{\mathcal{E}\to\infty}{\longrightarrow} \frac{1}{2\mathcal{E}} \int_0^{\lambda_G} \nu\beta e^{-\beta\zeta}\, d\lambda, \tag{76}$$

which can be evaluated by (69) if $\beta(\lambda_1)\zeta \gg 1$. For silicon with $a = 10^{-4}$ cm the following values were found for the integrals appearing in (23), (36), (51) and (76):

$$\begin{aligned}
\int \nu\alpha\, d\lambda &= 3.7 \times 10^4 \text{ cm}^{-1}, \\
\int \nu \frac{1 - e^{-\alpha a}}{\alpha a}\, d\lambda &= 0.754, \\
\int \nu(1 - e^{-\alpha a})\, d\lambda &= 0.351, \\
\int \nu\alpha e^{-\alpha a}\, d\lambda &= 1490 \text{ cm}^{-1}.
\end{aligned} \tag{77}$$

Table III

$\lambda(\mu)$	$\nu(\mu^{-1})$	α (cm^{-1})
0.22	0.00477	1.2×10^6
0.24	0.0148	1.4
0.26	0.0506	1.5
0.28	0.0625	1.6
0.30	0.178	1.5
0.32	0.306	1.3
0.34	0.448	0.9
0.36	0.538	0.4
0.38	0.654	0.1
0.40	0.801	0.06
0.42	1.14	0.04

REFERENCES

1. Daniels, A. F., Proc. I.R.E., **48,** 1960, p. 636.
2. Smith, D. H., Comm. & Elect., No. 45, 1959, p. 530.
3. Van Allen, J. A., J. Geophys. Res., **64,** 1959, p. 1683.
4. Loferski, J. and Rappaport, P., Phys. Rev., **111,** 1958, p. 432.
5. Brown, W. L. and Pearson, G. L., to be published.
6. Prince, M. B. and Wolf, M., J. Brit. I.R.E., **18,** 1958, p. 583.
7. Hoefler, D. C., Electronic News, **5,** May 23, 1960, p. 4.
8. Carpenter, E. F., ed., *Transactions of the Conference on the Use of Solar Energy — The Scientific Basis*, Vols. I–V, Univ. of Arizona, Tucson, 1955.
9. Telkes, M., J. Appl. Phys., **25,** 1954, p. 765.
10. Chapin, D. M., Fuller, C. S. and Pearson, G. L., J. Appl. Phys., **25,** 1954, p. 676.
11. Prince, M. B., J. Appl. Phys., **26,** 1955, p. 534.
12. Pfann, W. G. and Van Roosbroeck, W., J. Appl. Phys., **25,** 1954, p. 1422.
13. Cumerow, R., Phys. Rev., **95,** 1954, p. 16.
14. Rittner, E. S., Phys. Rev., **96,** 1954, p. 1708.
15. Loferski, J., J. Appl. Phys., **27,** 1956, p. 777.
16. Malovetskaya, V., Vavilov, V. S. and Galkin, G. N., Soviet Physics Solid State (English trans.), **1,** 1960, p. 1099.
17. Hrostowski, H. in Hannay, N. B., ed., *Semiconductors*, Reinhold, New York, 1959, p. 437.
18. Hoffman Electronics Corp., Technical Information Bulletin 32–58, September 1, 1958.
19. Vavilov, V. S., Smirnov, L. S. and Patskevitch, V. M., Soviet Physics Solid State (English trans.), **1,** 1960, p. 1344.
20. Theuerer, H. C., Whelan, J. M., Bridgers, H. E. and Buehler, E., J. Electrochem. Soc., **104,** 1957, p. 721.
21. Theuerer, H. C., J. Electrochem. Soc., **107,** 1960, p. 296.
22. Buck, T. M. and McKim, F. S., J. Electrochem. Soc., **105,** 1958, p. 709.
23. Wertheim, G., private communication.
24. Welker, H. and Weiss, H. in Seitz, F. and Turnbull, D., eds., *Solid-State Physics*, Vol. 3, Academic Press, New York, 1956.
25. Morin, F. and Maita, J., Phys. Rev., **96,** 1954, p. 28.
26. Shockley, W., *Electrons and Holes in Semiconductors*, D. Van Nostrand Co., New York, 1950, Section 12.5.
27. Shive, J., *The Properties, Physics and Design of Semiconductor Devices*, D. Van Nostrand Co., Princeton, N. J., 1959, Section 19.2.
28. Braunstein, R., Moore, A. and Herman, F., Phys. Rev., **109,** 1958, p. 695.
29. Sah, C., Noyce, R. and Shockley, W., Proc. I.R.E., **45,** 1957, p. 1228.
30. Rappaport, P., RCA Rev., **20,** 1959, p. 373.
31. Forsythe, W. E., ed., *Smithsonian Physical Tables*, Smithsonian Institution, Washington, 1954, Tables 812, 813.
32. Dash, W. C. and Newman, R., Phys. Rev., **99,** 1955, p. 1151.
33. Pfestorf, G., Ann. Physik, **81,** 1926, p. 906.
34. Shockley, W., *Electrons and Holes in Semiconductors*, op. cit., Section 12.3.
35. Backenstoss, G., Phys. Rev., **108,** 1957, p. 1416.
36. Krömer, H., Archiv elect. Übertragung, **8,** 1954, p. 223.
37. Trumbore, F. A., B.S.T.J., **39,** 1960, p. 205.
38. Folberth, O. G., Z. Naturforsch., **10a,** 1955, p. 502.
39. Newman, R., Phys. Rev., **111,** 1958, p. 1518.
40. Gremmelmaier, R., Z. Naturforsch., **10a,** 1955, p. 501.
41. Jenny, D., Loferski, J. and Rappaport, P., Phys. Rev., **101,** 1956, p. 1208.
42. Wysocki, J. and Rappaport, P., J. Appl. Phys., **31,** 1960, p. 571.
43. Wolf, M., Proc. I.R.E., **48,** 1960, p. 1246.
44. Lowan, A. N., Davids, N. and Levenson, A., Bull. Am. Math. Soc., **48,** 1942, p. 739.

The Covariance Function of a Simple Trunk Group, with Applications to Traffic Measurement*

By V. E. BENEŠ

(Manuscript received July 6, 1960)

Erlang's classical model for telephone traffic in a loss system is considered: N trunks, calls arriving in a Poisson process and negative exponential holding times; calls which cannot be served at once are dismissed without retrials. Let $N(t)$ be the number of trunks in use at t. An explicit formula for the covariance $R(\cdot)$ of $N(\cdot)$ in terms of the characteristic values of the transition matrix of the Markov process $N(\cdot)$ is obtained. Also, $R(\cdot)$ is expressed purely in terms of constants and the "recovery" function, i.e. the transition probability $Pr\{N(t) = N \mid N(0) = N\}$; $R(\cdot)$ is accurately approximated by $R(0)e^{r_1 t}$, with r_1 the largest negative characteristic value, itself well approximated (underestimated) by $-E\{N(\cdot)\}/R(0)$. Exact and approximate formulas for sampling error in traffic measurement are deduced from these results.

I. INTRODUCTION

A theoretical study of sampling fluctuations in telephone traffic measurements is useful both in designing procedures for measuring traffic loads and in interpreting field observations. Hayward[1] and Palm[2] have given an approximate formula for the sampling error incurred when observations of the numbers of calls in existence are made at fixed intervals of time. Their formula has the disadvantage that it is derived for a probabilistic model (of the traffic) in which there is an infinite number of available trunks. Thus there is no limit to the number of calls which can be in progress at one time, and no congestion. Two important parameters, the number N of trunks in the group, and the probability p_N of loss, are left out of account. For this reason the practical

* This work was completed in part while the author was visiting lecturer at Dartmouth College, Fall-Winter, 1959–60.

application of this model is usually restricted to large groups of trunks which are lightly loaded.

In this paper we derive and study the *covariance function* of the simplest stochastic model of a finite group of N trunks. The sampling error in traffic measurements can be calculated exactly from the covariance. We find formulas for the magnitudes of fluctuations of observed traffic for both periodic and continuous observation. The exact formulas lead to simple approximations similar to Hayward's, which take account of the number of trunks. Our results are summarized and discussed in Section II.

We shall use A. K. Erlang's classical probabilistic model for a group of trunks, described as follows:

i. Holding times of trunks are mutually independent, each with a negative exponential distribution. Time is measured in units of mean holding time.

ii. Epochs at which calls arrive form a Poisson process of intensity $a > 0$, independently of the holding times. The offered load is then a erlangs.

iii. There are $N < \infty$ trunks; calls which find all N of these trunks busy are "lost," and are cleared from the system.

These assumptions determine a Markov stochastic process $N(t)$, $-\infty < t < \infty$, the number of trunks in use at time t. $N(\cdot)$ is a random step-function fluctuating in unit steps between 0 and N. As is well known, $N(\cdot)$ has stationary probabilities $\{p_n, n = 0,1,\cdots, N\}$ given by the (first) Erlang distribution

$$p_n = \frac{\dfrac{a^n}{n!}}{\displaystyle\sum_{k=0}^{N} \frac{a^k}{k!}} \tag{1}$$

$$= \text{equilibrium probability that } n \text{ trunks are busy.}$$

With this choice of absolute probabilities, $N(\cdot)$ is a strictly stationary process, whose mean and variance are respectively

$$m_1 = a(1 - p_N),$$

$$\sigma^2 = m_1 - ap_N(N - m_1).$$

The probability p_N of loss is shown in Fig. 1, the fractional occupancy $N^{-1}m_1$ in Fig. 2, and the variance σ^2 in Fig. 3.

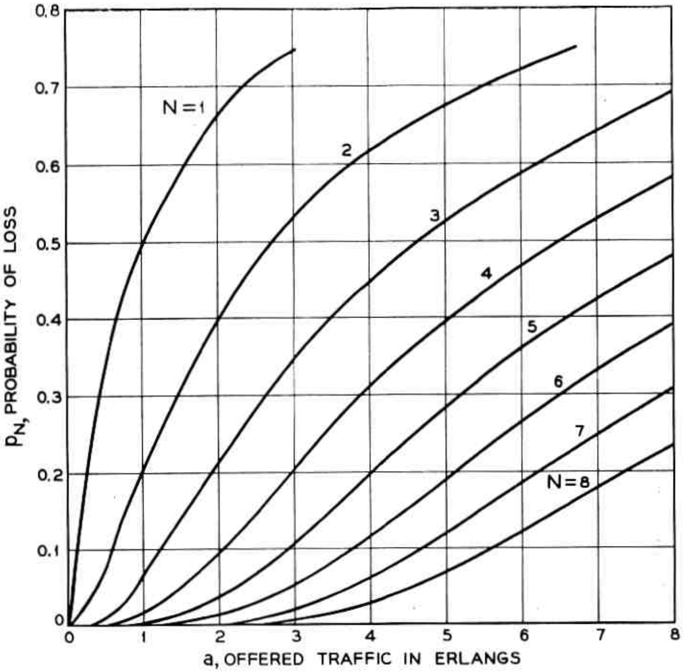


Fig. 1 — Probability p_N of loss.

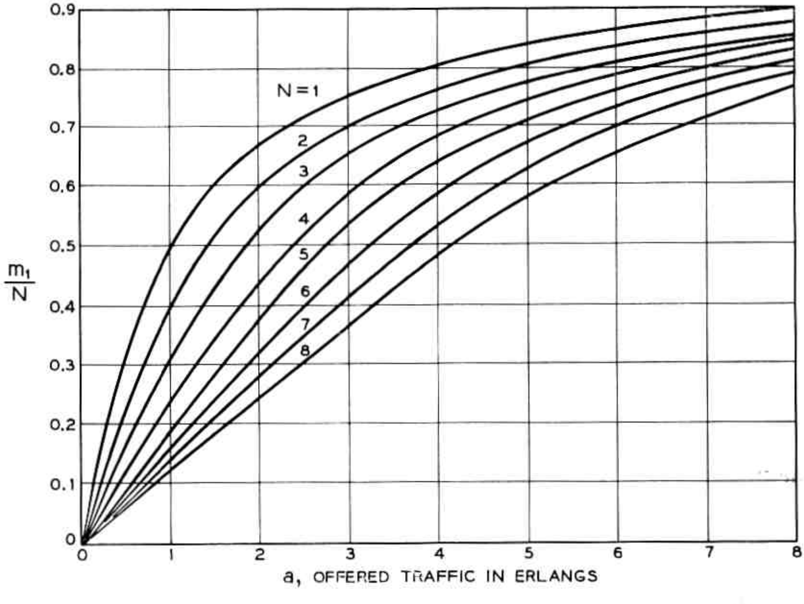


Fig. 2 — Fractional occupancy m_1/N.

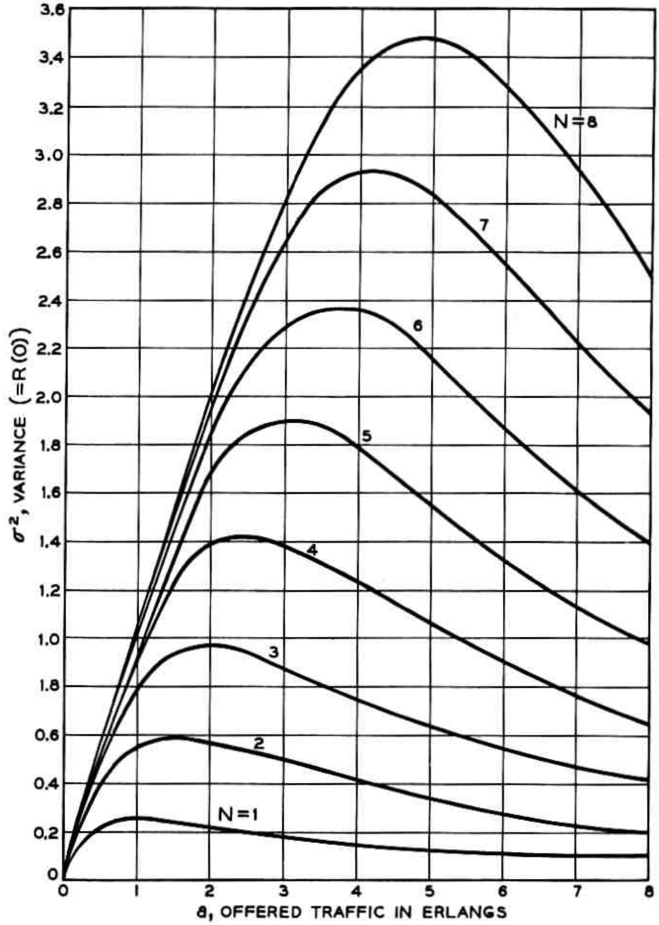


Fig. 3 — Equilibrium variance.

II. DISCUSSION, SUMMARY AND CONCLUSIONS

The covariance $R(t,s)$ between samples $N(t),N(s)$ of the stochastic process $N(\cdot)$ is the average of the product of $N(t)$ and $N(s)$, minus the product of the averages:

$$R(t,s) = E\{N(t)N(s)\} - E\{N(t)\}E\{N(s)\}.$$

Since $N(\cdot)$ is a stationary real process, we have $R(t,s) = R(\mid t - s \mid)$. The function $R(\cdot)$ is called the *covariance* function of the process $N(\cdot)$.

It can be written as

$$R(t) = \lim_{u\to\infty} E\{N(t+u)N(u)\} - E\{N(t+u)\}E\{N(u)\}$$
$$= \sum_{m=0}^{N} mp_m \sum_{n=0}^{N} n \Pr\{N(t) = n \mid N(0) = m\} - \left(\sum_{m=0}^{N} mp_m\right)^2, \tag{2}$$

where $\{p_m\}$ are the stationary (or equilibrium) probabilities given by (1), and

$$\Pr\{N(t) = n \mid N(0) = m\}$$

denotes the transition probability that n trunks are busy at time t if m were busy at time 0. The function $R(\cdot)$ expresses the average dependence or correlation between samples of $N(\cdot)$ taken at times t apart.

The principal practical use of the covariance function $R(\cdot)$ in the theory of telephone traffic is in computing theoretical estimates of sampling error incurred in traffic load measurements. Two methods of measuring traffic, the *switch-count* and the *time-average*, are considered in this paper. In the switch-count, n observations $\{x_1, \cdots, x_n, x_j = N(j\tau), j = 1, \cdots, n\}$ of the random process are made at intervals τ apart; the average

$$\frac{1}{n}\sum_{j=1}^{n} N(j\tau) = \frac{1}{n}\sum_{j=1}^{n} x_j = n^{-1}S_n$$

is then used as an estimate of the carried load $m_1 = a - ap_N$. This method is important economically because it is cheaper to scan trunk groups periodically than to observe them continuously. The number τ is the *scan interval*, and the number $S_n = x_1 + \cdots + x_n$ is called the (total) *number of paths in service*, in n observations. Table I lists actual mean holding times, scan intervals used and resulting values of τ for

TABLE I — HOLDING TIMES, SCAN INTERVALS AND VALUES OF τ

Type of Call	Typical Holding Time (seconds)	Scan Interval (seconds)		Ratio τ of Scan Interval to Holding Time	
		U.S.A.	Europe	U.S.A.	Europe
Local Call	100–200	100	36	1 to $\frac{1}{2}$	$\frac{1}{3}$ to $\frac{1}{6}$
Long Distance Call	200–600	100	36	$\frac{1}{2}$ to $\frac{1}{6}$	$\frac{1}{6}$ to $\frac{1}{20}$
Originating Register Holding Time	10–15	10 or 100	36	1 to $\frac{2}{3}$ or 10 to 7	4 to 2
No. 5 Marker Holding Time	0.25–1.0	10	—	—	—

various types of call. The variance of $n^{-1}S_n$ is expressible in terms of the covariance $R(\cdot)$ as

$$\operatorname{Var}\{n^{-1}S_n\} = n^{-2} \sum_{j=-n}^{n} (n - |j|)R(j\tau). \tag{3a}$$

In the *time-average*, the continuously recorded sample average

$$M(T) = T^{-1} \int_0^T N(t)\, dt,$$

is used to estimate the carried load. The variance of this estimate is

$$\operatorname{Var}\{M(T)\} = 2T^{-2} \int_0^T (T - t)R(t)\, dt. \tag{3b}$$

Thus the mean square error of both these methods of traffic measurement can be calculated theoretically if the covariance $R(\cdot)$ is known.

In formula (2) the covariance function is expressed in terms of the stationary probabilities $\{p_n\}$ given by the Erlang distribution, and the transition probabilities

$$p_{mn}(t) = \Pr\{N(t) = n \mid N(0) = m\}.$$

In the theory of telephone traffic, the particular transition probability

$$p_{NN}(t) = \Pr\{N(t) = N \mid N(0) = N\}$$

has been singled out (in Refs. 3 and 4, for example) as a suitable "recovery" or "relaxation" function that is characteristic of the dynamic behavior of the Markov process $N(\cdot)$ in point of the undesirable "all trunks busy" condition.

We shall show that a much more cogent reason than this can be adduced to support the importance of the recovery function to traffic theory: the covariance function $R(\cdot)$ can be expressed entirely in terms of the recovery function and the offered load a. In other words, a single one of the $(N + 1)^2$ transition probabilities appearing in formula (2) suffices for determining the covariance function, and this one is the recovery function $p_{NN}(\cdot)$. This fact is a theoretical justification of the intuitive view that the recovery function is important, for now the variances of $n^{-1}S_n$ and of $M(T)$ are expressible using only the recovery function.

We next give a summary of the contents of the remaining sections; this is followed by an account of specific results and conclusions.

An exact formula for the covariance $R(\cdot)$ is stated and discussed in Section III, and derived in Section VII. The formula readily yields a rigorous upper bound which appears to give a close approximation to $R(\cdot)$ itself. In Section IV the recovery function $p_{NN}(\cdot)$ is given, and it is shown how the covariance may be expressed in terms of the recovery function by a convolution integral. The variance of $n^{-1}S_n$ is studied in Section V; an exact formula, and an approximating upper bound [based on the upper bound for $R(\cdot)$], are both obtained. The variance of the time-average $M(T)$ is considered in Section VI; again, an exact formula and an approximating upper bound are found.

The covariance function $R(\cdot)$ is bounded from above and closely approximated by a single exponential function

$$R(t) \leqq \sigma^2 e^{r_1 t}, \qquad \sigma^2 = R(0), \qquad r_1 < 0.$$

Here

$$\begin{aligned} \sigma^2 &= R(0) \\ &= \text{equilibrium variance of } N(\cdot) \\ &= (\text{load carried}) - (\text{load lost})(\text{average number of idle trunks}), \end{aligned}$$

and the reciprocal time constant r_1 in the exponent is the dominant* characteristic value of the "rate" or "transition" matrix of the differential equations satisfied by the transition probabilities. Alternately, r_1 is the root of least magnitude of a Poisson-Charlier polynomial. The root r_1 is shown as a function of offered traffic a for $N = 1, \cdots, 8$ in Fig. 4, and is tabulated in Table II.

A lower bound for r_1, *depending only on the mean and variance of* $N(\cdot)$, is derived in Section VIII by making use of the fact that the matrix of the differential equations for the transition probabilities is symmetrizable. For low values of offered traffic per trunk, i.e., $a/N < 1$, this bound can be used to approximate r_1. In any case, the bound is a convenient starting place for the use of Newton's method. The bound is the ratio $-m_1/\sigma^2$, which satisfies the inequality

$$-\frac{m_1}{\sigma^2} \leqq r_1 < -1,$$

with

$$\begin{aligned} m_1 &= \text{equilibrium mean of } N(\cdot) \\ &= \text{load carried} = a(1 - p_N), \\ \sigma^2 &= \text{equilibrium variance of } N(\cdot) \\ &= (\text{load carried}) - (\text{load lost})(\text{average number of idle trunks}). \end{aligned}$$

The approximation $r_1 \cong -m_1/\sigma^2$ is illustrated in Fig. 5.

* I.e., that of least magnitude (among the nonzero characteristic values).

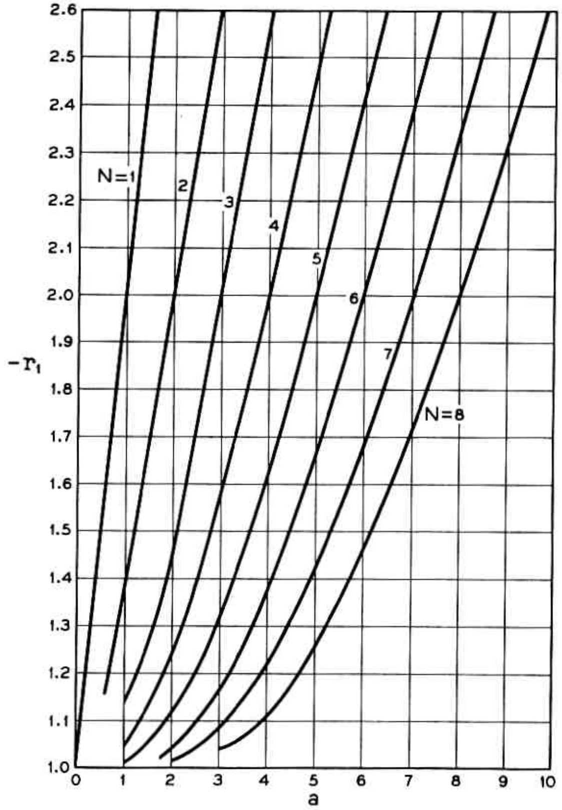


Fig. 4 — Negative of the root r_1 of smallest magnitude as a function of load a for $N = 1, \cdots, 8$.

Table II — Negative of Dominant Characteristic Value r_1

a	$N = 4$	$N = 5$	$N = 6$	$N = 7$	$N = 8$
1	1.043967	1.011448	1.002421	1.000421	1.000062
2	1.249464	1.112166	1.045044	1.015806	1.004800
3	1.582363	1.326321	1.172257	1.084025	1.037229
4	2.000000	1.629624	1.383389	1.222707	1.121762
5	2.477548	2.000000	1.663799	1.427870	1.265214
6	3.000000	2.422137	2.000000	1.689991	1.463798
7	3.557618	2.885474	2.381627	2.000000	1.710891
8	4.143703	3.382497	2.800900	2.350437	2.000000
9	4.753426	3.907677	3.251918	2.735363	2.325514
10	5.383178	4.456828	3.730121	3.150052	2.682770

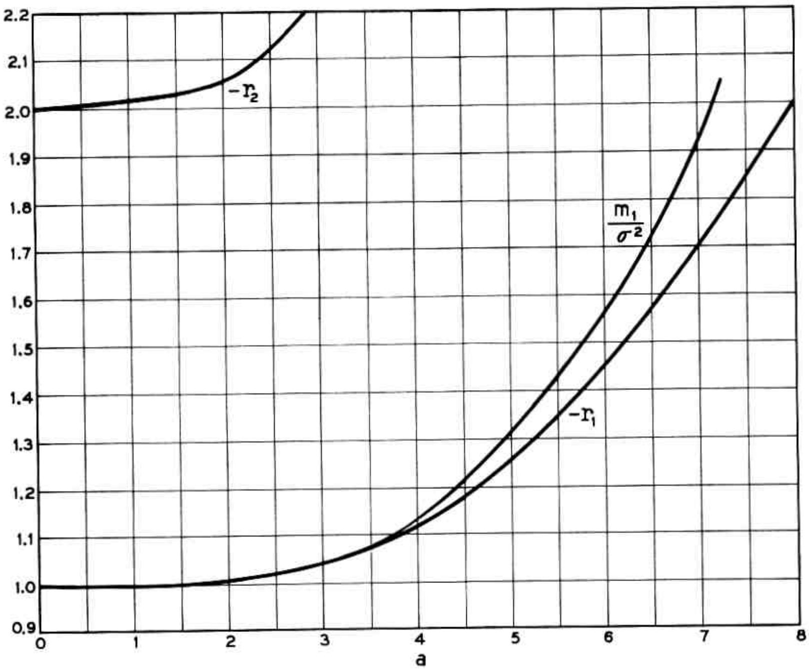


Fig. 5 — Illustration of the approximation $r_1 \cong -m_1/\sigma^2$.

By the "infinite trunk" model we shall henceforth mean the stochastic model for telephone traffic determined by all the same assumptions that we made in the Introduction, except that $N = \infty$; i.e., an unlimited number of trunks is postulated. Riordan[5] and Beneš[6,7] have considered this model; Hayward[1] based his sampling error formula on it.

It is widely believed that the "infinite trunk" model is applicable to large groups of lightly loaded trunks. Such a belief is gratuitous until comparisons with a model having a finite number of trunks are made. Studying the covariance function of the simple finite trunk group enables us to make some of the needed comparisons; e.g., the variances of $n^{-1}S_n$ and $M(T)$ in the two models are of particular interest. Knowledge of the covariance $R(\cdot)$, however, is also relevant to the other three cases to which engineers are loath to apply the "infinite trunk" model, viz.:

i. large groups of heavily loaded trunks,
ii. small groups of lightly loaded trunks,
iii. small groups of heavily loaded trunks.

The variance of $n^{-1}S_n$ is bounded from above and approximated by the formula

$$\mathrm{Var}\{n^{-1}S_n\} \leqq n^{-1}\sigma^2 \left\{ \mathrm{ctnh}\,\lambda - \frac{1 - e^{-2n\lambda}}{2n} \mathrm{csch}^2\,\lambda \right\}, \tag{4}$$

where n is the number of observations, and

$$\lambda = -\frac{\tau r_1}{2} = -\tfrac{1}{2}\text{ (scan interval) (dominant characteristic value).}$$

The *exact* formula for the variance of $n^{-1}S_n$ in the "infinite trunk" model is

$$n^{-1}a \left\{ \mathrm{ctnh}\,\tfrac{1}{2}\tau - \frac{1 - e^{-n\tau}}{2n} \mathrm{csch}^2\,\tfrac{1}{2}\tau \right\}. \tag{5}$$

The upper bound (4) for the finite group is compared with the exact formula (5) for the "infinite trunk" model in Fig. 6, which shows each formula as a function of the scan interval τ for various n, for $a = 20$ erlangs offered to 20 trunks. The curves suggest that the upper bound for $\mathrm{Var}\{n^{-1}S_n\}$ for $N < \infty$ is consistently less than the corresponding variance in the "infinite trunk" model. As might be expected, increasing the scan interval τ improves accuracy for the same number of observations. This is because the covariance function is positive, and monotone in $|t|$.

The variance of $M(T)$ is bounded from above and approximated by

$$\mathrm{Var}\{M(T)\} \leqq 2\sigma^2 \frac{e^{r_1 T} - 1 - r_1 T}{T^2 r_1^2}, \tag{6}$$

where T is the length of the time-interval of continuous observation, and σ^2 and r_1 are, as before, the variance of $N(\cdot)$ and the dominant characteristic value, respectively. The exact formula for the variance of $M(T)$ in the "infinite trunk" model is

$$2a\,\frac{e^{-T} - 1 + T}{T^2}. \tag{7}$$

Since $r_1 < -1$, and σ^2 is always less than a if $N < \infty$, the "infinite trunk" model overestimates the variance of $M(T)$ if applied to a finite group. This conclusion is illustrated in Fig. 7, which shows the formulas (6) and (7) for a load of 20 erlangs offered to 20 trunks. For an observation time of 10 mean holding times the "infinite trunk" formula (7) applied here would overestimate the variance by about 500 per cent. This is about as extreme a case as would occur in practice. Fig. 7 also

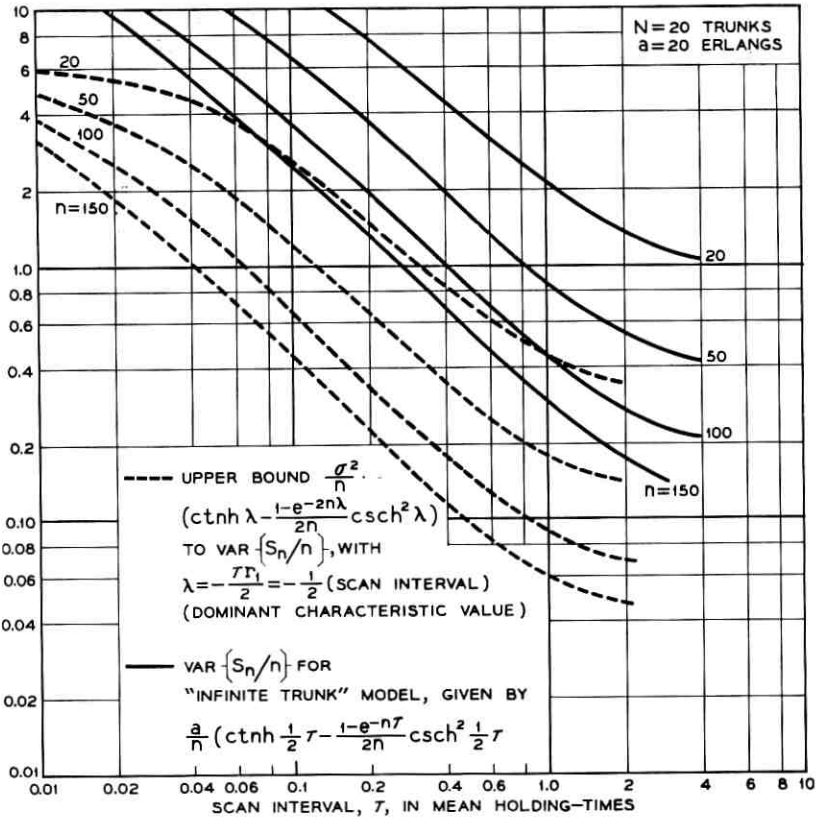


Fig. 6 — Comparison of variance of S_n/n for finite and infinite trunk models.

depicts a "mixed" formula obtained by replacing a by σ^2 in the "infinite trunk" formula (6); for 10 mean holding times the "mixed" formula only overestimates the variance by about 100 per cent. Thus most of the discrepancy is due to the difference between σ^2 and a.

Our conclusions are set down in the following list:

1. The average dynamic behavior of the process $N(\cdot)$, as described by the covariance function $R(\cdot)$, can be adequately determined from the dominant characteristic value r_1 and the variance σ^2 of $N(\cdot)$.

2. The same parameters, r_1 and σ^2, suffice to give simple approximating upper bounds for the sampling error incurred in both periodic and continuous observation of $N(\cdot)$. These bounds depend on the size N of the trunk group.

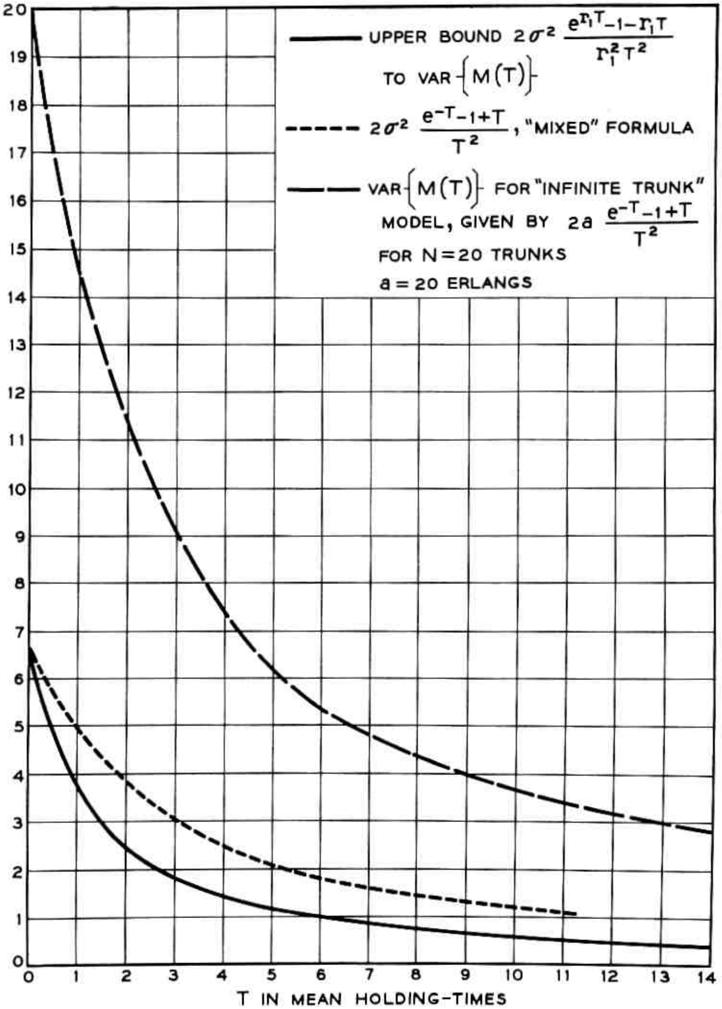


Fig. 7 — Comparison of variance of $M(T)$ for finite and infinite trunk models.

3. In terms of r_1 and σ^2 it is possible to check the applicability, for theoretical estimates of sampling error, of the "infinite trunk" model which assumes $N = \infty$.

4. The "infinite trunk" model, applied to finite trunk groups, consistently and often grossly overestimates the sampling error. The overestimation occurs largely because σ^2 is always less, and for heavy traffic

is much less, than a, the (Poisson) variance of $N(\cdot)$ in the "infinite trunk" model.

5. In terms of r_1 and σ^2 it is possible to design sampling procedures for traffic measurement that depend explicitly on the number N of trunks in the group. By these methods, a given accuracy can be obtained with less observation, and thus at lower cost, than the "infinite trunk" model would require.

6. Hence for finite groups of trunks traffic sampling procedures which are based on the "infinite trunk" model tend to be wasteful, particularly for heavy traffic. The parameters r_1 and σ^2 provide a systematic way of tailoring the measurement procedure to the number of trunks in the group.

III. THE COVARIANCE FUNCTION

To state the formula for $R(\cdot)$ we need the "sigma" functions* defined (see Riordan[9]) as

$$\sigma_0(m) = \frac{a^m}{m!},$$

$$\sigma_k(m) = \sum_{j=0}^{m} \binom{k+j-1}{j} \frac{a^{m-j}}{(m-j)!},$$

with m (but not k) a nonnegative integer. These functions are connected with the Poisson-Charlier polynomials

$$p_n(x) = a^{n/2}(n!)^{\frac{1}{2}} \sum_{j=0}^{n} (-1)^{n-j} \binom{n}{j} j! a^{-j} \binom{x}{j}$$

by the relation

$$\sigma_k(m) = (-a^{\frac{1}{2}})^m (m!)^{-\frac{1}{2}} p_m(-k).$$

(See Ref. 10, p. 33.)

For fixed N and a, let $r_1, r_2, \cdots, r_N$ be (in order of increasing magnitude) the N zeros in the variable s of the polynomial $\sigma_{s+1}(N)$. In Section VII the covariance is shown to be given by (the exact formula)

$$R(t) = -a^2 p_N \sum_{j=1}^{N} \frac{e^{r_j t}}{r_j(1+r_j)^2} \prod_{i \neq j} [1 - (r_j - r_i)^{-1}] \tag{8}$$

where p_N is the probability of loss. It has been shown† that the zeros r_j are all real, negative, and distinct; all are less than -1, and consecu-

* The σ notation is copied from unpublished work of H. Nyquist. The functions themselves were introduced into traffic theory by Palm.[8]

† The earliest reference appear to be Haantjes[11] in 1938. See also Ledermann and Reuter.[12]

tive pairs are separated by at least unity. Fig. 8 shows these roots for $N = 1, 2, 3$ as functions of a.

Now r_j is always negative, and the terms of the product satisfy

$$1 - \frac{1}{r_j - r_i} > 0; \tag{9}$$

hence the sum in (8) has all terms negative, so that

$$R(t) \geqq 0, \qquad \text{all } t.$$

The correlation between successive samples is thus always positive. It is obvious from (8) that

$$-a^2 p_N \sum_{j=1}^{N} \frac{1}{r_j(1 + r_j)^2} \prod_{i \neq j} \frac{r_j - 1 - r_i}{r_j - r_i} = \sigma^2 = R(0). \tag{10}$$

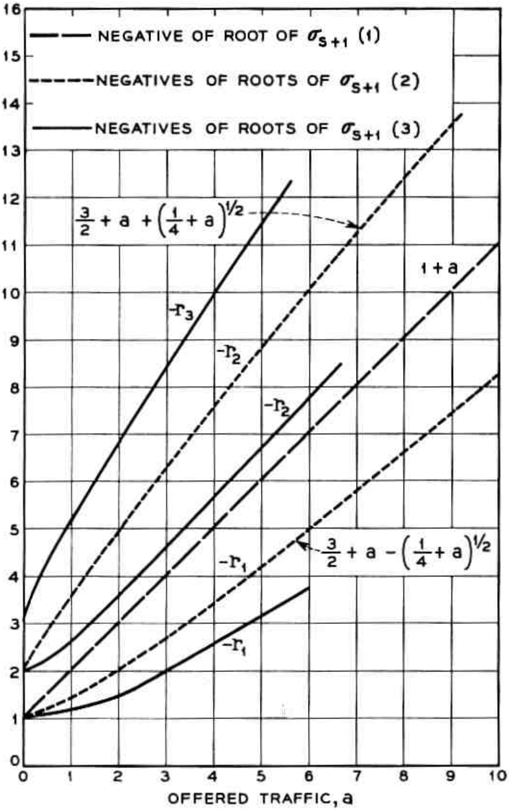


Fig. 8 — Roots of the first three σ-functions.

Since r_1 is the root closest to zero, the value of (8) is only increased if the r_j in the exponents of (8) are replaced by r_1. Using (9) and (10), we conclude that

$$0 \leqq \sigma^2 e^{r_1 t} - R(t) = \xi(t), \tag{11}$$

where

$$\xi(t) = a^2 p_N \sum_{j=2}^{N} \frac{e^{r_j t} - e^{r_1 t}}{r_j(1 + r_j)^2} \prod_{i \neq j} \frac{r_j - 1 - r_i}{r_j - r_i},$$

and

$$\begin{aligned} \xi(t) &\leqq a^2 p_N e^{-t} \sum_{j=2}^{N} (j + 1)^{-2} \leqq a^2 p_N \frac{4\pi^2 - 30}{24} e^{-t} \\ &\leqq (0.3933) a^2 p_N e^{-t}. \end{aligned}$$

The approximation $R(t) \cong \sigma^2 e^{r_1 t}$ is illustrated in Figs. 9 and 10. It appears to be fairly accurate, especially for light loads.

The upper bound $\sigma^2 e^{r_1 t}$ for $R(t)$ should be compared with the rigorous formula (see Riordan[5] and Beneš[7])

$$R(t) = ae^{-t},$$

which holds for the "infinite trunk" model. In this model the equilibrium distribution of occupancy is Poisson, so that

$$R(0) = \sigma^2 = \mathrm{Var}\{N(t)\} = E\{N(t)\} = a,$$

and the "time constant" of the exponential is unity, since time is measured in units of mean holding time.

The difference between the "infinite trunk" model and the "finite trunk" model in point of the covariance can be understood by considering the effect of congestion, which is present in the latter. Congestion affects the upper bound formula most directly through the value of the variance σ^2. It is obvious intuitively, and borne out in Fig. 3, that as a increases σ^2 must eventually decrease to zero. This behavior is not mimicked by the "infinite trunk" model, for which $\sigma^2 = a$.

The finitude of N, i.e., congestion, affects the bound $\sigma^2 e^{\{r_1 t\}}$ in two ways: (a) the "time constant" is not unity but the smaller number $-(r_1)^{-1}$, so that the rate at which dependence between samples of $N(\cdot)$ decreases (as a function of the interval between samples) is larger than in the "infinite trunk" model; this "time constant" decreases as the traffic a increases, because, as illustrated by Fig. 4, r_1 is a monotone

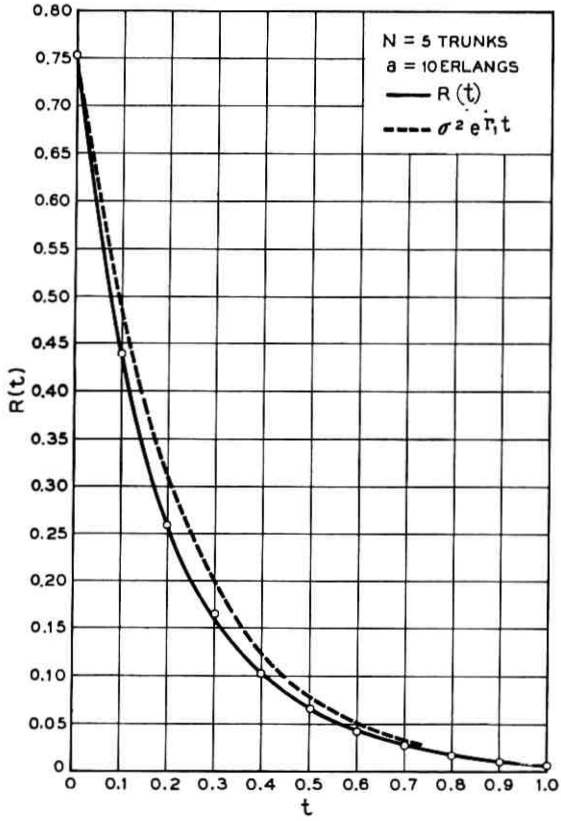


Fig. 9 — The covariance $R(t)$ for $N = 5$ trunks, $a = 10$ erlangs, with the approximate formula $R(t) \sim \sigma^2 e^{r_1 t}$.

decreasing function of a; (b) the value of $R(0)$ $(=\sigma^2)$ is not a but the generally much smaller number

$$\sigma^2 = a(1 - p_N)\left[1 - ap_N\left(\frac{N}{a - ap_N} - 1\right)\right],$$
$$= a[1 - p_N(1 + N - a + ap_N)].$$

The last form shows that $\sigma^2 < a$ for all a and N. In fact, it is obvious intuitively that

$$\sigma^2 = m_1 - ap_N(N - m_1) < m_1 < a.$$

A simple approximation for the dominant root r_1 can sometimes be

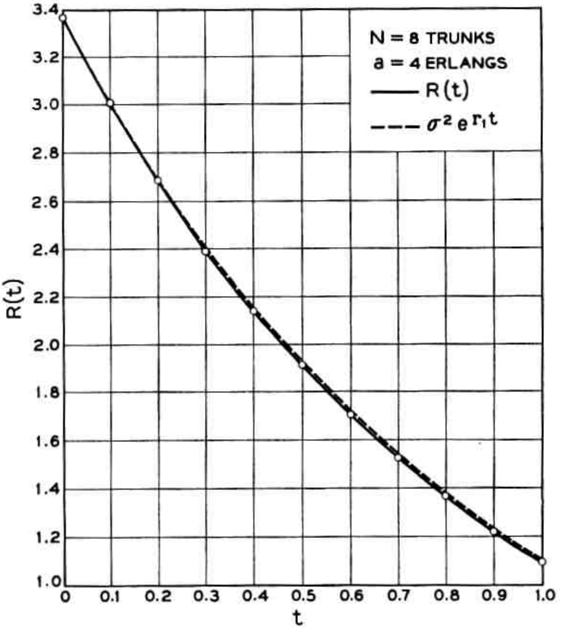


Fig. 10 — The covariance $R(t)$ for $N = 8$ trunks, $a = 4$ erlangs, with the approximate formula $R(t) \sim \sigma^2 e^{r_1 t}$.

used to make the approximation $R(t) \cong \sigma^2 e^{\{r_1 t\}}$ more useful. It is shown in Section VIII that

$$-\frac{m_1}{\sigma^2} = -\frac{\text{carried load}}{\text{load variance}} \leqq r_1 ;$$

i.e., $-m_1/\sigma^2$ is a rigorous lower bound to r_1 . Fig. 5 suggests this bound gives a fairly good approximation to r_1 if $a/N < 1$. Hence a simple approximate formula for $R(\cdot)$, valid for $a/N < 1$, is given by

$$\begin{aligned} R(t) &\cong \sigma^2 e^{\{-m_1 t/\sigma^2\}} \\ &\cong (\text{load variance}) \exp\left\{-\frac{\text{carried load}}{\text{load variance}}\, t\right\}. \qquad (12) \end{aligned}$$

We know that $R(t) \leqq \sigma^2 e^{\{r_1 t\}}$ and that $-m_1/\sigma^2 < r_1$; hence replacing r_1 by $-m_1/\sigma^2$ tends to correct the error in the upper bound formula. The formula (12) is illustrated in Fig. 11.

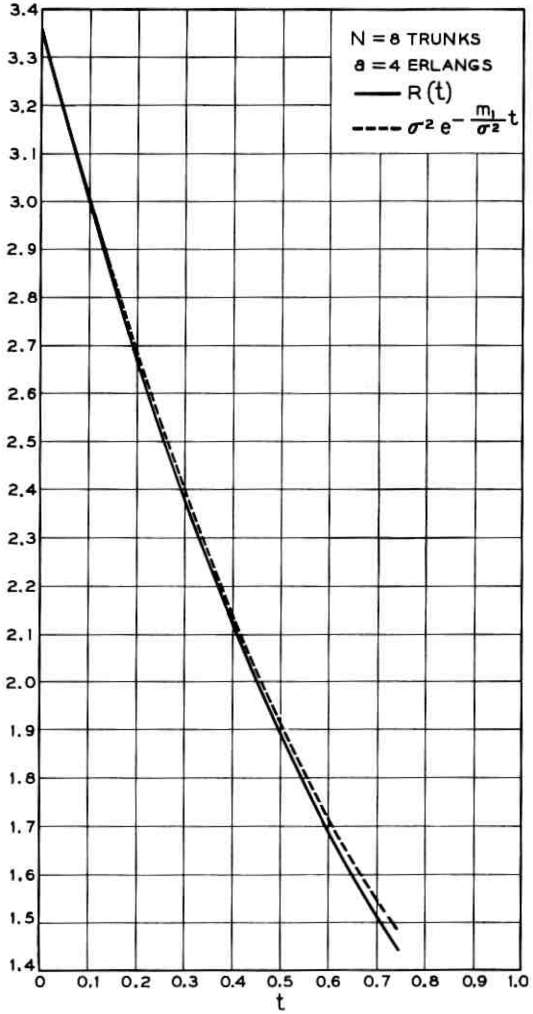


Fig. 11 — Comparison of $R(t)$ with $\sigma^2 e^{-(m_1/\sigma^2)t}$ for $N = 8$ trunks, $a = 4$ erlangs.

IV. THE COVARIANCE IN TERMS OF THE RECOVERY FUNCTION

It has been shown[3] that the Laplace transform of $p_{NN}(\cdot)$ is

$$\frac{\sigma_s(N)}{s\sigma_{s+1}(N)}.$$

By expansion in partial fractions we find that

$$p_{NN}(t) = p_N - \sum_{j=1}^{N} \frac{e^{r_j t}}{r_j} \prod_{i \neq j} \left(1 - \frac{1}{r_j - r_i}\right). \tag{13}$$

The sum assumes only negative values, and so $p_{NN}(\cdot)$ decreases monotonically to the loss probability p_N. The recovery function is illustrated in Fig. 12.

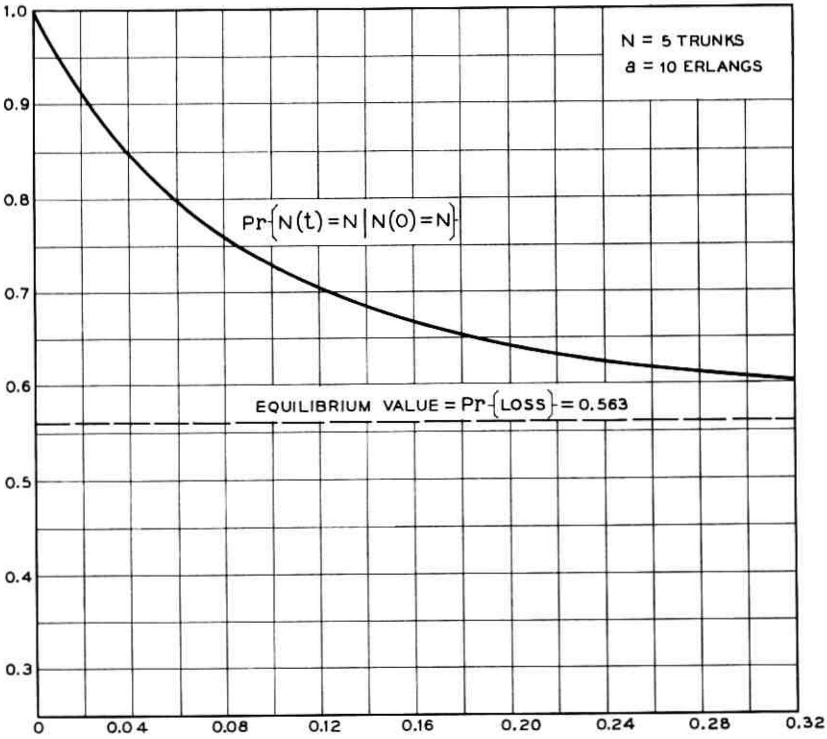


Fig. 12 — Recovery function for $N = 5$ trunks, $a = 10$ erlangs.

We now observe that for each $j = 1, \cdots, \mathrm{N}$,

$$\int_0^t (t - u)\, e^{-(t-u)+r_j u}\, du = \frac{e^{r_j t} - e^{-t}}{(r_j + 1)^2} - \frac{t e^{-t}}{r_j + 1}. \tag{14}$$

By comparison of formulas (8) and (13), and use of (14), one finds that

$$R(t) = a^2 p_N \int_0^t (t - u) e^{-(t-u)} [p_{NN}(u) - p_N]\, du + \sigma^2 e^{-t} + C t e^{-t}, \tag{15}$$

where

$$C = -a^2 p_N \sum_{j=1}^{N} \frac{1}{r_j(r_j + 1)} \prod_{i \neq j} \left(1 - \frac{1}{r_j - r_i}\right).$$

This formula expresses $R(\cdot)$ in terms of $p_{NN}(\cdot)$ by a simple convolution. To evaluate C explicitly we note that

$$C = -a^2 p_N \left[\frac{\sigma_{s+1}(N - 1)}{(1 + s)\sigma_{s+1}(N)} - \frac{a_{-1}}{1 + s}\right]_{s=0},$$

where a_{-1} is the first coefficient in the power series expansion of the left-hand term in the bracket. One finds

$$\begin{aligned} a_{-1} &= \frac{\sigma_0(N - 1)}{\sigma_0(N)} = \frac{N}{a}, \\ C &= a^2 p_N \left(\frac{N}{a} - 1 + p_N\right) \\ &= a p_N (N - m_1) \\ &= \text{(load lost) (average number of idle trunks).} \end{aligned}$$

V. THE VARIANCE OF THE NUMBER OF PATHS IN SERVICE

We assume that n observations $\{x_j, j = 1, \cdots, n\}$ of $N(\cdot)$ are made during an interval of equilibrium, so that

$$\operatorname{Cov}\{x_i, x_j\} = R(|\, i - j\, |\, \tau),$$

where τ is the scan interval. Then with

$$\begin{aligned} S_n &= x_1 + x_2 + \cdots + x_n \\ &= \text{number of paths found in service,} \end{aligned}$$

we find that

$$\begin{aligned}\operatorname{Var}\{S_n\} &= E\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j\right\} - E^2\left\{\sum_{i=1}^{n} x_i\right\} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Cov}\{x_i, x_j\} \qquad (16) \\ &= \sum_{j=-n}^{n} (n - |j|)R(j\tau).\end{aligned}$$

To give an exact formula for $\operatorname{Var}\{S_n\}$ we note that

$$\sum_{m=-\infty}^{\infty} e^{-2|m|u} = \operatorname{ctnh} u,$$

$$\sum_{m=-\infty}^{\infty} |m|\, e^{-2|m|u} = -\frac{d}{du}\sum_{m=1}^{\infty} e^{-2mu} = \tfrac{1}{2}\operatorname{csch}^2 u$$

and

$$\begin{aligned}\sum_{m=n}^{\infty} (m - n)e^{-2mu} &= \tfrac{1}{2}e^{-2nu}\sum_{m=-\infty}^{\infty} |m|\, e^{-2|m|u} \\ &= \tfrac{1}{4}e^{-2nu}\operatorname{csch}^2 u.\end{aligned}$$

Then also

$$\begin{aligned}\sum_{j=-n}^{n} (n - |j|)\, e^{-2|j|u} &= n\sum_{m=-\infty}^{\infty} e^{-2|m|u} - \sum_{m=-\infty}^{\infty} |m|\, e^{-2|m|u} \\ &\quad + 2\sum_{m=n}^{\infty} (m - n)\, e^{-2mu} \qquad (17) \\ &= n\operatorname{ctnh} u - \frac{(1 - e^{-2nu})}{2}\operatorname{csch}^2 u.^*\end{aligned}$$

Since the covariance $R(\cdot)$ is a symmetric function given by (8), it can be seen that

$$\begin{aligned}&\operatorname{Var}\{n^{-1}S_n\} = \\ &\quad -n^{-1}a^2 p_N \sum_{j=1}^{N} \frac{\left[\operatorname{ctnh}\left(-\frac{\tau r_j}{2}\right) - \frac{1 - e^{n\tau r_j}}{2n}\operatorname{csch}^2\left(-\frac{\tau r_j}{2}\right)\right]}{r_j(1 + r_j)^2} \qquad (18) \\ &\qquad \cdot \prod_{i \neq j}\left(1 - \frac{1}{r_j - r_i}\right).\end{aligned}$$

[*] Use of this identity was suggested by unpublished work of J. W. Tukey to which the author had access.

This formula is exact, given the assumptions. It is easily shown from formula (17) that the exact formula for the variance of $n^{-1}S_n$ in the "infinite trunk" model is

$$n^{-1}a\left\{\operatorname{ctnh}\tfrac{1}{2}\tau - \frac{1 - e^{-n\tau}}{2n}\operatorname{csch}^2\tfrac{1}{2}\tau\right\},$$

illustrated in Fig. 6.[13]

Returning to the case of finitely many trunks, we can obtain approximating upper bounds to formula (18) for $\operatorname{Var}\{n^{-1}S_n\}$ by using the results of Section III on the covariance function. It can be seen from the arguments leading to (17) that replacing the roots r_j by r_1 in the hyperbolic functions in (18) increases the values of the expressions in square brackets; this replacement is equivalent to using the upper bound

$$\sigma^2 e^{r_1 t}$$

for $R(t)$ in formula (16). Hence

$$\operatorname{Var}\{n^{-1}S_n\} \leqq n^{-1}\sigma^2\left\{\operatorname{ctnh}\lambda - \frac{1 - e^{-2n\lambda}}{2n}\operatorname{csch}^2\lambda\right\}, \qquad (19)$$

where

$$\lambda = -\frac{\tau r_1}{1} = -\tfrac{1}{2}(\text{scan interval})(\text{dominant characteristic value}).$$

Since $\sigma^2 e^{r_1 t}$ is close to $R(t)$, we may expect that the overestimate (19) gives a good approximation to the actual variance. This approximation is conveniently plotted as a function of λ for various n in Fig. 13.

VI. THE VARIANCE OF TIME AVERAGES

It follows from formulas (3b) and (8) that

$$\operatorname{Var}\left\{\int_0^T N(t)\,dt\right\} = c_0 + c_1 T + o(e^{-T}) \qquad (20)$$

as $T \to \infty$, where

$$c_0 = 2a^2 p_N \sum_{j=1}^{N} \frac{\prod_{i\neq j}\left(1 - \frac{1}{r_j - r_i}\right)}{r_j^{\,3}(1 + r_j)^2}$$

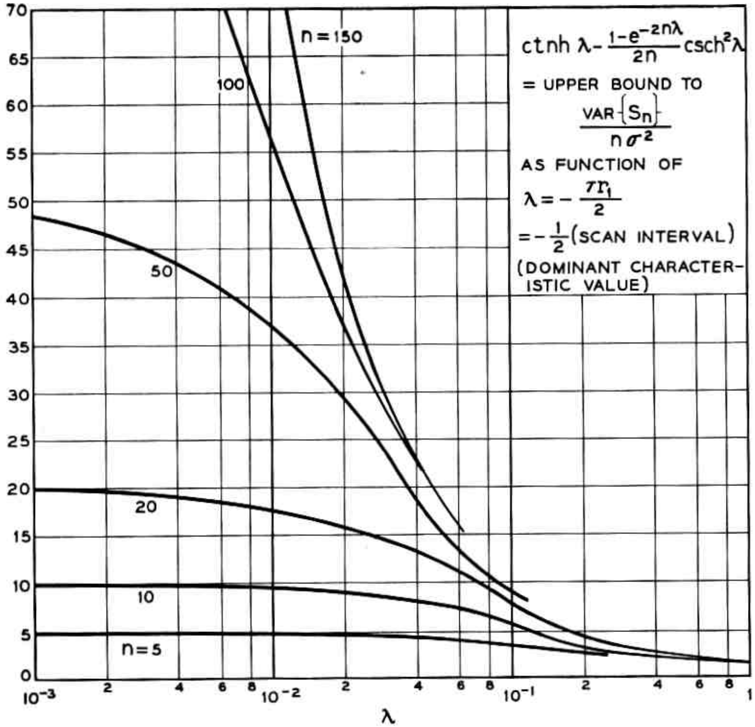


Fig. 13 — Upper bound to $\mathrm{Var}\{S_n\}/n\sigma^2$.

is a negative constant, and

$$c_1 = \int_0^\infty R(u)\,du = 2a^2 p_N \sum_{j=1}^{N} \frac{\prod_{i\neq j}\left(1 - \frac{1}{r_j - r_i}\right)}{r_j^2(1 + r_j)^2}.$$

Note that c_0 and c_1 differ only in the power of r_j that occurs in the denominators. The third term of (20) is positive; is given by

$$o(e^{-T}) = -2a^2 p_N \sum_{j=1}^{N} \frac{e^{r_j T}}{r_j^3(1 + r_j)^2} \prod_{i\neq j}\left(1 - \frac{1}{r_j - r_i}\right);$$

equals $-c_0$ at $T = 0$; and is of smaller order than e^{-T} because $r_1 < -1$. To evaluate c_1 explicitly, we note that

$$c_1 = -2a^2 p_N \left[\frac{\sigma_{s+1}(N-1)}{s(1+s)^2\sigma_{s+1}(N)} - \frac{1 - p_N}{s} - \frac{a_{-1}}{1+s} - \frac{a_{-2}}{(1+s)^2}\right]_{s=0}, \quad (21)$$

where a_{-2}, a_{-1} are respectively the first and second coefficients in the power series expansion of the leftmost term in the bracket of (21); these are given by

$$\begin{aligned} a_{-2} &= \frac{\sigma_0(N-1)}{\sigma_0(N)} = \frac{N}{a}, \\ a_{-1} &= \frac{d}{dx}\frac{\sigma_{x+1}(N-1)}{x\sigma_{x+1}(N)}\bigg]_{x=-1} \\ &= -\frac{N}{a} + \frac{1-p_N}{ap_N}. \end{aligned}$$

To find c_1 we must compute

$$\lim_{s\to 0}\left[\frac{\sigma_{s+1}(N-1)}{s(1+s)^2\sigma_{s+1}(N)} - \frac{1-p_N}{s}\right] = -\frac{d}{dx}\,\frac{\sigma_{x+1}(N-1)}{(1+x)^2\sigma_{x+1}(N)}\bigg]_{x=0}.$$

This equals

$$2(1-p_N) - \frac{d}{dx}\,\frac{\sigma_{x+1}(N-1)}{\sigma_{x+1}(N)}\bigg]_{x=0},$$

or

$$(1-p_N)\left[2 - \frac{d}{dx}\log\sigma_{x+1}(N-1) + \frac{d}{dx}\log\sigma_{x+1}(N)\right]_{x=0}.$$

Now the generating function of the σ-functions is

$$\Phi(s,z) = \sum_{n=0}^{\infty} z^n\sigma_s(n) = (1-z)^{-s}e^{az}$$

so that

$$\begin{aligned} \frac{\partial}{\partial s}\Phi(s,z) &= \Phi(s,z)\sum_{n=1}^{\infty}\frac{z^n}{n}, \\ \frac{d}{ds}\sigma_s(n) &= \frac{\sigma_s(0)}{n} + \frac{\sigma_s(1)}{n-1} + \cdots + \frac{\sigma_s(n-1)}{1}, \\ \xi_n &= \frac{d}{dx}\log\sigma_{x+1}(n)\bigg]_{x=0} = \sum_{j=0}^{n-1}\frac{\sigma_1(j)}{(n-j)\sigma_1(n)}. \end{aligned}$$

It follows that

$$\begin{aligned} c_1 &= 2a^2p_N[a_{-2} + a_{-1} + (1-p_N)(2-\xi_{N-1}+\xi_N)] \\ &= 2\sigma^2 + 2a^2p_N(1-p_N)(1-\xi_{N-1}+\xi_N) + 2aNp_N. \end{aligned}$$

The constant c_0 can be evaluated in a similar fashion.

From the bounds (11) for $R(\cdot)$ we conclude that

$$0 \leqq 2\sigma^2\left(\frac{e^{r_1T} - 1 - r_1T}{T^2r_1^2}\right) - \mathrm{Var}\{M(T)\} \leqq (0.3933)a^2p_N\left(\frac{e^{-T} - 1 + T}{T^2}\right),$$

and since $R(t) \cong \sigma^2 e^{r_1t}$, we may expect that the *overestimate*

$$2\sigma^2\,\frac{e^{r_1T} - 1 - r_1T}{T^2r_1^2} \tag{22}$$

is a good approximation to the variance of $M(T)$. This approximation has the same form as the exact formula (7) for the "infinite trunk" model, because in both cases a single exponential is used for $R(\cdot)$ in formula (8). The overestimate (22) is depicted graphically in Fig. 14. It was convenient to plot the ratio

$$\frac{\mathrm{Var}\{M(T)\}}{\sigma^2}$$

as a function of the single parameter

$$\mu = r_1T = -(\text{dominant characteristic value})\ (\text{observation time}).$$

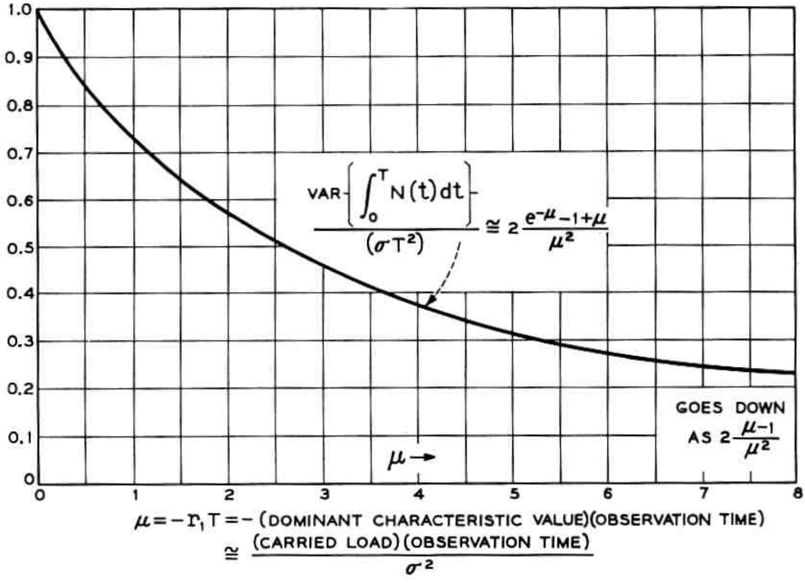


Fig. 14 — Approximation to $\mathrm{Var}\int_0^T\{N(t)dt\}/(\sigma T)^2$.

A simpler form of (22), valid for $a/N < 1$, results when we replace r_1 by its lower bound

$$-\frac{m_1}{\sigma^2} = -\frac{\text{carried load}}{\text{load variance}} \leqq r_1 .$$

This replacement decreases the value obtained, i.e., moves the approximation in the direction of $\operatorname{Var}\{M(T)\}$.

VII. DERIVATION OF THE COVARIANCE

The transition probabilities

$$p_{mn}(t) = \Pr\{N(t) = n \mid N(0) = m\}$$

of $N(\cdot)$ satisfy the Kolmogorov equations

$$\begin{aligned}
&p_{mn}(0) = \delta_{mn} , \\
&\frac{d}{dt} p_{mN} = a p_{m(N-1)} - N p_{mN} , \\
&\frac{d}{dt} p_{mn} = (n+1) p_{m(n+1)} + a p_{m(n-1)} - (a+n) p_{mn} , \qquad 0 < n < N, \\
&\frac{d}{dt} p_{m0} = p_{m1} - a p_{m0} .
\end{aligned} \tag{23}$$

Multiplying the nth equation by n, and summing on the index n, we find

$$\frac{d}{dt} E\{N(t) \mid N(0) = m\} = -E\{N(t) \mid N(0) = m\} + a[1 - p_{mN}(t)],$$

whence

$$E\{N(t) \mid N(0) = m\} = m\, e^{-t} + a \int_0^t e^{-(t-u)}[1 - p_{mN}(u)]\, du.$$

By formula (2), the covariance is then

$$\begin{aligned}
R(t) &= \sum_{m=0}^{N} m p_m E\{N(t) \mid N(0) = m\} - m_1^2 \\
&= m_2 e^{-t} + a m_1 (1 - e^{-t}) - m_1^2 - a \int_0^t e^{-(t-u)} \sum_{m=0}^{N} m p_m p_{mN}(u)\, du,
\end{aligned}$$

where

$$m_i = \sum_{n=0}^{N} n^i p_n$$

for $i = 1, 2$ and $\{p_n\}$ are the stationary probabilities given by (1). In particular,

$$m_1 = a(1 - p_N), \tag{24}$$

$$\sigma = (m_2 - m_1^2)^{\frac{1}{2}} = [m_1 - ap_N(N - m_1)]^{\frac{1}{2}}. \tag{25}$$

The Laplace transform of

$$\Pr\{N(\cdot) = N \mid N(0) = m\}$$

has been determined[3] to be

$$\frac{a^{N-m} m! \sigma_s(m)}{N! s \sigma_{s+1}(N)}.$$

Therefore that of $R(\cdot)$ is

$$\begin{aligned} R^*(s) = \int_0^\infty e^{-st} R(t)\, dt &= \frac{m_2}{1 + s} + \frac{am_1}{s(1 + s)} - \frac{m_1^2}{s} \\ &\quad - \frac{a}{s(1 + s)\sigma_{s+1}(N)} \sum_{m=1}^{N} mp_m \frac{a^{N-m} m! \sigma_s(m)}{N!}. \end{aligned} \tag{26}$$

By (1), the last term of (26) is

$$-\frac{ap_N}{s(1 + s)\sigma_{s+1}(N)} \sum_{m=1}^{N} m\sigma_s(m).$$

It has been shown[9] that the "sigma" functions satisfy the recurrences

$$\sigma_s(m) = \sigma_{s+1}(m) - \sigma_{s+1}(m - 1), \tag{27}$$

$$m\sigma_s(m) = a\sigma_s(m - 1) + s\sigma_{s+1}(m - 1), \tag{28}$$

so that

$$\begin{aligned} \sum_{m=1}^{N} m\sigma_s(m) &= a\sum_{k=0}^{N-1} \sigma_s(k) + s\sum_{k=0}^{N-1} \sigma_{s+1}(k) \\ &= a\sigma_{s+1}(N - 1) + s\sigma_{s+2}(N - 1), \end{aligned}$$

and

$$\frac{\sigma_{s+2}(N - 1)}{\sigma_{s+1}(N)} = \frac{N}{s + 1} \frac{a\sigma_{s+1}(N - 1)}{(s + 1)\sigma_{s+1}(N)}.$$

The foregoing identities yield the following simplified formula for $R^*(s)$:

$$R^*(s) = \frac{m_2}{1+s} + \frac{am_1}{s(1+s)} - \frac{aNp_N}{(1+s)^2} - \frac{a^2 p_N}{1+s}\left[\frac{\sigma_{s+1}(N-1)}{s(1+s)\sigma_{s+1}(N)}\right] - \frac{m_1^2}{s}. \tag{29}$$

From (27) we find that the partial fraction expansion

$$\frac{\sigma_{s+1}(N-1)}{\sigma_{s+1}(N)} = \sum_{j=1}^{N} - \frac{\sigma_{r_j}(N)N!}{(s-r_j)\prod_{i\neq j}(r_j-r_i)} = \sum_{j=1}^{N}(s-r_j)^{-1}\prod_{i\neq j}\frac{r_j-1-r_i}{r_j-r_i},$$

is valid, where $\{r_j\}$ are the zeros of $\sigma_{s+1}(N)$.

By a similar argument, since $p_N = \sigma_0(N)/\sigma_1(N)$,

$$\frac{\sigma_{s+1}(N-1)}{s(1+s)\sigma_{s+1}(N)} = \frac{1-p_N}{s} - \frac{N}{a(1+s)} + \sum_{j=1}^{N}(s-r_j)^{-1}\frac{1}{r_j}\frac{1}{1+r_j}\prod_{i\neq j}\frac{r_j-1-r_i}{r_j-r_i}.$$

Hence formula (29) can be inverted to give, for $t \geqq 0$,

$$\begin{aligned} R(t) &= m_2e^{-t} + am_1[1-e^{-t}] - aNp_Nte^{-t} - m_1^2 \\ &\quad - a^2p_N\int_0^t e^{-(t-u)}\left[1 - p_N - \frac{N}{a}e^{-u} + \sum_{j=1}^{N}\frac{e^{r_ju}}{r_j(1+r_j)} \cdot \prod_{i\neq j}\frac{r_j-1-r_i}{r_j-r_i}\right]du, \\ &= \sigma^2e^{-t} + a^2p_NKe^{-t} - a^2p_N\sum_{j=1}^{N}\frac{e^{r_jt}}{r_j(1+r_j)^2}\prod_{i\neq j}\frac{r_j-1-r_i}{r_j-r_i}, \end{aligned} \tag{30}$$

where

$$\sigma^2 = m_2 - m_1^2 = \text{equilibrium variance},$$

and

$$K = \sum_{j=1}^{N}\frac{1}{r_j}\frac{1}{(1+r_j)^2}\prod_{i\neq j}\frac{r_j-1-r_i}{r_j-r_i}.$$

To evaluate K explicitly we observe that

$$K = -\left[\frac{\sigma_{s+1}(N-1)}{(1+s)^2\sigma_{s+1}(N)} - \frac{a_{-1}}{1+s} - \frac{a_{-2}}{(1+s)^2}\right]_{s=0}, \qquad (31)$$

where a_{-2}, a_{-1} are respectively the first and second coefficients in the power series expansion of the leftmost term in the bracket of (31). Thus

$$K = a_{-1} + a_{-2} - 1 + p_N .$$

Now

$$\frac{\sigma_{s+1}(N-1)}{(1+s)^2\sigma_{s+1}(N)} = (1+s)^{-2}\frac{\sigma_0(N-1)}{\sigma_0(N)} + (1+s)^{-1} \cdot \left[\frac{d}{dx}\frac{\sigma_{x+1}(N-1)}{\sigma_{x+1}(N)}\right]_{x=-1} + \sum_{j=1}^{N}\frac{(s-r_j)^{-1}}{(1+r_j)^2}\prod_{i\neq j}\frac{r_j - 1 - r_i}{r_j - r_i}.$$

From the recurrence (28) for the σ-functions we find that

$$s\,\frac{\sigma_{s+1}(N-1)}{\sigma_s(N)} - N + a\,\frac{\sigma_s(N-1)}{\sigma_s(N)} = 0;$$

differentiating with respect to s and setting $s = 0$, we obtain

$$a_{-1} = \frac{d}{ds}\frac{\sigma_{s+1}(N-1)}{\sigma_{s+1}(N)}\bigg|_{s=-1} = \frac{1}{a}\left(-\frac{\sigma_1(N-1)}{\sigma_0(N)}\right) = -\frac{1-p_N}{ap_N}.$$

Clearly,

$$a_{-2} = \frac{\sigma_0(N-1)}{\sigma_0(N)} = \frac{N}{a},$$

and so

$$K = -\frac{1-p_N}{ap_N} + \frac{N}{a} - 1 + p_N,$$

$$a^2 p_N K = -\sigma^2 .$$

Thus the formula (30) for the covariance function $R(\cdot)$ simplifies to

$$R(t) = -a^2 p_N \sum_{j=1}^{N}\frac{e^{r_j t}}{r_j(1+r_j)^2}\prod_{i\neq j}\frac{r_j - 1 - r_i}{r_j - r_i}. \qquad (32)$$

VIII. APPROXIMATION TO THE DOMINANT CHARACTERISTIC VALUE

The differential equations (23) can be written in the form

$$\frac{d}{dt}P(t) = QP(t),$$

where $P(t)$ is the matrix of transition probabilities $\{p_{mn}(t)\}$ and Q is the matrix of the "transition rates":

$$Q = \begin{pmatrix} -a & 1 & 0 & 0 \;\cdots & 0 & & 0 \\ a & (-a-1) & 2 & 0 & & & 0 \\ 0 & a & (-a-2) & 3 & & & \\ \vdots & & & & & & \vdots \\ & & & & & 0 & \\ & & & & & N-1 & 0 \\ 0 & 0 & \cdots & & a & (-a-N+1) & N \\ 0 & 0 & \cdots & & 0 & a & -N \end{pmatrix}.$$

The characteristic values of Q are $0, r_1, r_2, \cdots, r_N$. We define

$$\mu_n = \frac{1}{p_n} = n!a^{-n} \sum_{j=0}^{N} \frac{a^j}{j!}, \qquad n = 0,1,\cdots,N,$$

and we introduce an inner product for the space $L_2(\mu)$ of $(N+1)$-tuples of complex numbers by the definition

$$(x,y) = \sum_{n=0}^{N} x_n \bar{y}_n \mu_n .$$

The matrix Q represents a symmetric operator on $L_2(\mu)$, i.e.,

$$(Qx,y) = (x,Qy), \qquad x,y \in L_2(\mu).$$

It is easily seen that

$$\sum_{n=0}^{N} \frac{1}{\mu_n} = \sum_{n=0}^{N} p_n = 1, \tag{33}$$

$$Qp = 0, \qquad \text{for} \quad p = (p_0, p_1, \cdots, p_N), \tag{34}$$

$$Q_{ij}\mu_i = Q_{ji}\mu_j, \qquad i,j = 0,1,\cdots,N. \tag{35}$$

The last identity implies that

$$(Qx,y) = -\frac{1}{2} \sum_{i,j} \overline{(y_i\mu_i - y_j\mu_j)} \frac{Q_{ji}}{\mu_i} (\mu_i x_i - \mu_j x_j),$$

$$(Qx,x) \leqq 0,$$

so (as we already know) all characteristic values of Q are nonpositive, being of the form (Qx,x) for some $x \in L_2(\mu)$.

From the extremal properties of the characteristic values of symmetric operators (e.g., Zaanen,[14] p. 383, Theorem 3) we conclude that

$$r_1 = \max(Qx,x),$$

the maximum being over all $x \in L_2(\mu)$ which are not identically zero, and satisfy $(x,x) = 1$, $(x,p) = 0$, p being the vector of stationary probabilities, as in (34).

We can now estimate r_1 from below by choosing an appropriate vector x. We choose

$$x_n = \frac{n - m_1}{\sigma\mu_n}, \qquad n = 0,1,\cdots,N,$$

where m_1 and σ are the mean and standard deviation of $N(\cdot)$ in equilibrium, given by formulas (24) and (25) respectively. Clearly, $(x,x) = 1$ and $(x,p) = 0$, and

$$\begin{aligned}(Qx,x) &= -a \sum_{n=0}^{N-1} p_n \left(\frac{n - m_1}{\sigma} - \frac{n + 1 - m_1}{\sigma}\right)^2 \\ &= -\frac{a(1 - p_N)}{\sigma^2} \\ &= -\frac{m_1}{\sigma^2} \leqq r_1 .\end{aligned}$$

(See Kramer.[15])

This approximation is illustrated in Fig. 5.

IX. ACKNOWLEDGMENTS

The author thanks S. O. Rice and J. W. Tukey for access to unpublished manuscripts and helpful discussions, and A. Descloux for giving the manuscript a careful reading. The suggestions of R. I. Wilkinson and J. Riordan are also appreciated.

REFERENCES

1. Hayward, W. S., Jr., The Reliability of Telephone Traffic Load Measurements by Switch Counts, B.S.T.J., **31,** 1952, p. 357.
2. Palm, C., Tekniska Medelanden fran Kungl. Telegrafstyrelsen, 1941, nr. 7–9.
3. Beneš, V. E., Transition Probabilities for Telephone Traffic, B.S.T.J., **39,** 1960, p. 1297.
4. Kosten, L., Over de invloed van herhaalde oproepen in de theorie der blokkeringskansen, De Ingenieur, **47,** 1947, p. 123.
5. Riordan, J., Telephone Traffic Time Averages, B.S.T.J., **30,** 1951, p. 1129.
6. Beneš, V. E., A Sufficient Set of Statistics for a Simple Telephone Exchange Model, B.S.T.J., **36,** 1957, p. 939.

7. Beneš, V. E., Fluctuations of Telephone Traffic, B.S.T.J., **36,** 1957, p. 965.
8. Palm, C., Calcul exact de la perte dans les groupes de circuits échelonnés, Ericsson Tech., **4,** 1936, p. 41.
9. Riordan, J., appendix to Wilkinson, R. I., Theories for Toll Traffic Engineering in the U.S.A., B.S.T.J., **35,** 1956, p. 507.
10. Szegö, G., *Orthogonal Polynomials*, American Mathematical Society Colloquium Publications, New York, 1938.
11. Haantjes, J., Wiskundige Opgaven, **17,** 1938.
12. Ledermann, W. and Reuter, G. E., Spectral Theory for the Differential Equations of Simple Birth and Death Processes, Phil. Trans. Roy. Soc. (London), **236A,** 1954, p. 321.
13. Olsson, K. M., Calculation of Dispersion in Telephone Traffic Recording Values for Pure Chance Traffic, Tele (Eng. Ed.), **2,** 1959, p. 71.
14. Zaanen, A. C., *Linear Analysis*, Interscience, New York, 1953.
15. Kramer, H. P., Symmetrizable Markov Matrices, Ann. Math. Stat., **30,** 1959, p. 149.

Mode-Conversion Filters

By E. A. MARCATILI

(Manuscript received July 11, 1960)

Resonance of higher-order modes in waveguides can be advantageously used to make band-rejection filters of unusually low loss and simplicity. The region where the resonance takes place can be obtained either by a local change of cross section of the waveguide or by the inclusion of dielectrics. Mode-conversion band-rejection filters can be combined to build channel-dropping filters which are of particular interest in the millimeter wavelength region to separate bands of $TE_{01}{}^{\circ}$ *into* $TE_{10}{}^{\square}$.

In this paper the necessary design relationships for channel-dropping filters using mode-conversion band-rejection filters are derived. It also contains a theoretical derivation of the intrinsic Q *of band-rejection filters in round and rectangular waveguides. Finally, the experimental results obtained with different mode-conversion band-rejection filters at 12 and 56 kmc, and with a channel-dropping filter from* $TE_{01}{}^{\circ}$ *to* $TE_{10}{}^{\square}$ *at 56 kmc, are given.*

I. INTRODUCTION

A large variety of channel-dropping filters operate through the use of band-rejection filters, and since the microwave art is pushing the usable spectrum to higher and higher frequencies, low heat loss and easy-to-build band-rejection filters are important.

This statement is particularly true in the process of separating bands in the long distance waveguide communication system[1] that operates with circular-electric mode in the millimeter wavelength region. The information sent from repeater to repeater through the low-heat-loss multimode circular waveguide must be separated into tens of bands for the purposes of regeneration and amplification at each repeater. Since each repeater operates in single-mode rectangular waveguide, one possible solution is to convert the circular-electric wave to fundamental mode in rectangular waveguide and then to drop the different channels with known techniques. This solution has several disadvantages: the filters are relatively lossy because of the low intrinsic Q of parallelepiped-shaped cavities in the millimeter region, and they are difficult to build. Furthermore, the channels to be dropped last are substantially attenuated because they must travel in a high-loss rectangular waveguide. For instance, the theoretical attenuation in the standard silver waveguide RG98U (50 to 75 kmc) is 0.53 to 0.39 db/foot.

Better solutions are filters that simultaneously drop the channels and make the transfer from circular-electric wave in round waveguide to dominant mode in rectangular waveguide.[2,3]

This paper describes a channel-dropping filter that combines all the desirable features: it filters, it transfers TE_{01}° mode into $TE_{10}^{\square}$, it has low insertion loss and it is extremely easy to build. All this is possible because of the use of mode-conversion band-rejection filters.

II. DESCRIPTION OF THE MODE-CONVERSION BAND-REJECTION FILTER

Consider, for instance, a round waveguide carrying the TE_{01}° mode and barely cut off for the TE_{02}°. If for a length l the diameter of the waveguide is slightly larger, in such a way that TE_{02}° is no longer cut off, the region l becomes a multimode region[4] where the TE_{02}° generated at both diameter discontinuities can resonate and introduce a large insertion loss to the incident TE_{01}°. The bandwidth depends essentially on the amount of mode conversion (size of the discontinuity), and the center frequency depends on the length l. The filter can be made of sliding coaxial tubes because the circumferential cracks do not interrupt the conduction current of circular electric modes.

The intrinsic Q is very high, one order of magnitude better than a $TE_{101}^{\square}$ cavity, because: (a) the resonant mode is essentially a low-loss one,[1] (b) the end walls do not absorb energy since they do not exist,* (c) the tuning mechanism is lossless and (d) the coupling that is provided by the diameter change does not create high-density currents, such as exist in the case of band-rejection cavities coupled through irises to the main waveguide, or in the case of microwave band-rejection filters made of lumped elements.

The reasoning used for circular-electric modes can be generalized, that is, any waveguide that contains a low-loss, multimode region exhibits rejection bands corresponding to the resonances of the confined modes.[4,5,6,7] Thus, a rectangular waveguide cut off for $TE_{20}^{\square}$, except for a length l of slightly larger width capable of generating and supporting $TE_{20}^{\square}$, becomes a mode-conversion band-rejection filter.

* The heat loss due to the penetration of the TE_{02}° mode in the cutoff waveguides is calculated in Section IX.

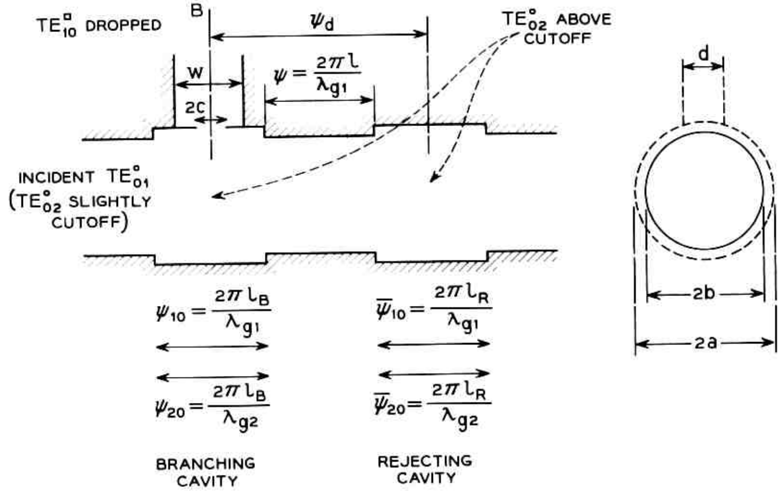
Another band-rejection filter in rectangular waveguide is obtained by building the multimode region with a dielectric slab close to one of the narrow walls, since the dielectric provides an apparent width increase of the waveguide.

All these mode-conversion band-rejection filters have small return loss out of resonance.

III. DESCRIPTION OF THE CHANNEL-DROPPING FILTER

The channel-dropping filter consists of a through waveguide with two multimode regions, one of which is coupled to another waveguide (Fig. 1). For the purpose of fixing ideas, we imagine that the through waveguide is circular with two enlarged regions where the TE_{02}° excited by the incident TE_{01}° can resonate; the dropping arm is a rectangular waveguide. The enlarged regions where the TE_{02}° mode can resonate will be referred to as *cavities*, even if they are not enclosed volumes. The idealized filter must be such that the incident TE_{01}° mode is matched at all frequencies, and at midband all the power flows into the rectangular waveguide.

A low-frequency channel-dropping filter that will be demonstrated to be the equivalent of the microwave one and that satisfies the previous demands is shown in Fig. 2. The resonant circuits are equivalent to the cavities, and the three resistances connected to the circuit through ade-



TE_{10}° DROPPED
B
ψ_d
TE_{02}° ABOVE CUTOFF
W
$\psi = \frac{2\pi l}{\lambda_{g1}}$
2c
d
INCIDENT TE_{01}° (TE_{02}° SLIGHTLY CUTOFF)
$\psi_{10} = \frac{2\pi l_B}{\lambda_{g1}}$
$\overline{\psi}_{10} = \frac{2\pi l_R}{\lambda_{g1}}$
$\psi_{20} = \frac{2\pi l_B}{\lambda_{g2}}$
$\overline{\psi}_{20} = \frac{2\pi l_R}{\lambda_{g2}}$
BRANCHING CAVITY
REJECTING CAVITY
2b
2a

Fig. 1 — Microwave mode-conversion channel-dropping filter.

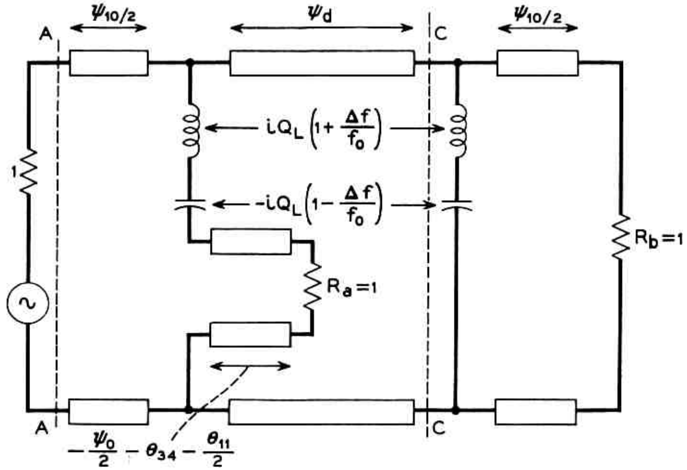


Fig. 2 — Low-frequency channel-dropping filter.

quate transformers (not indicated in the figure for simplicity) are equivalent to the characteristic impedances of the three microwave ports.

In each resonant circuit f_0 is the midband frequency and the loaded Q is defined $Q_L = f_0/(2\Delta f_0)$, where $2\Delta f_0$ is the half power bandwidth of the dropped channel. The normalized reactances are such that the impedance seen toward the right of the plane AA is unity at all frequencies, provided that ψ_d is an odd number of quarter wavelengths. At midband frequency the maximum power available goes to R_a and far from resonance it goes to R_b.

Another microwave circuit equivalent to that of Fig. 1, which may help the reader to understand the behavior of the mode-conversion channel-dropping filter, is shown in Fig. 3. Here, the resonant cavities

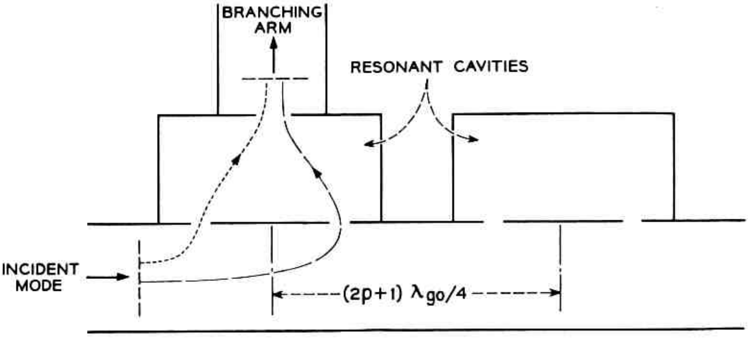


Fig. 3 — Microwave channel-dropping filter; path difference = θ_0.

have been separated from the through waveguide. The incident mode excites each cavity through two coupling holes; in an equivalent way, in Fig. 1, the incident TE_{01}° mode couples to the TE_{02}° of each cavity through two diameter discontinuities. Finally, in Fig. 3, only the resonant mode of the first cavity is assumed to couple to the output load, implying that the coupling between the incident and the branching mode in Fig. 1 is negligible.

The reader who is not interested in the mathematical treatment of the mode-conversion channel-dropping filter may now go directly to the résumé of results in Section VII of this paper. The scattering matrix of the branching cavity, Fig. 1, is studied in Section IV. In Section V, the scattering matrix of the rejecting cavity is derived from that of the branching cavity by reducing to zero the coupling between the round and rectangular waveguides. Then, in Section VI, both cavities are connected through a certain length of single-mode waveguide, and the mathematical description of the channel-dropping filter is completed.

The derivations have been made for single-resonance rejecting and branching filters because these are the building blocks for the design of more complicated filters such as those of the maximally flat type.[8]

IV. SCATTERING MATRIX OF THE BRANCHING CAVITY

Consider the branching cavity of Fig. 1, separated from the rejecting cavity and with all terminals matched. This cavity is represented in Fig. 4 with the elementary components separated. The symbol J_1 represents the junction where port 1′ carrying $TE_{10}^{\square}$ mode couples symmetrically to TE_{01}° and TE_{02}° in the cavity. Since the TE_{02}° mode is almost at cut-off and close to resonance, the coupling to TE_{01}° can be neglected.

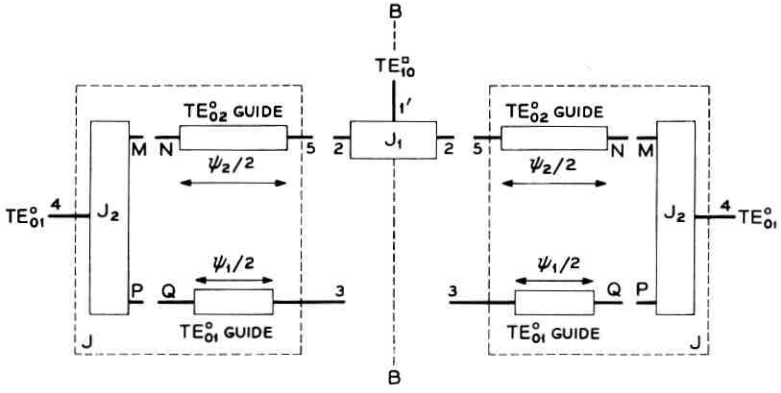


Fig. 4 — Branching cavity with elementary components separated.

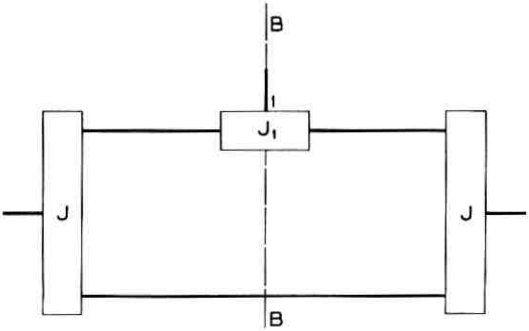


Fig. 5 — Branching cavity.

The symbol J$_2$ represents the junction at a diameter discontinuity; ports 4 and P carry TE$_{01}$° and port M carries TE$_{02}$. Ports M and 2 are connected by a piece of waveguide whose ports are N and 5. This waveguide has a midband electrical length

$$\frac{\psi_{20}}{2} = \frac{\pi l_B}{\lambda_{g2}},$$

where l_B is the distance between diameter discontinuities and λ_{g2} is the midband TE$_{02}$° mode guided wavelength. Likewise, ports P are connected by two pieces of waveguide, each of midband electrical length

$$\frac{\psi_{10}}{2} = \frac{\pi l_B}{\lambda_{g1}},$$

where λ_{g1} is the midband TE$_{01}$° mode guided wavelength in the cavity.

Fig. 4 can be simplified by representing all the elements inside of each of the dotted lines as a single junction J, and so the branching cavity is reduced to the circuit shown in Fig. 5.

Since this three-port structure is symmetric with respect to the plane BB, the scattering matrix can be derived from the scattering matrices of two simpler structures, one derived by making the symmetry plane BB a magnetic plane (open circuit), that is, a surface where the tangential magnetic field is zero, and another obtained by making the symmetry plane an electric one (short circuit), that is, a surface where the electric tangential field is zero.

4.1 *Open-Circuited Half of Branching Cavity*

Assume in Fig. 5 that BB is a magnetic plane. Port 1′ as well as the branching cavity is divided in two symmetrical portions, and each half is shown in Fig. 6.

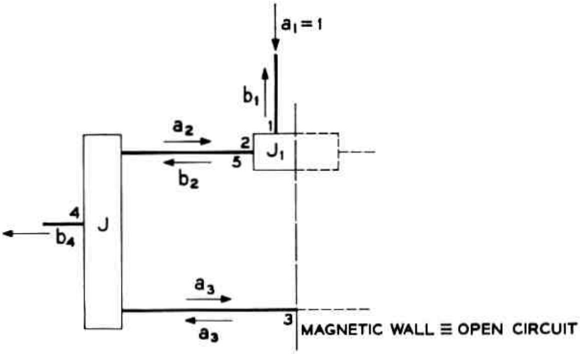


Fig. 6 — Open-circuited half of branching cavity.

If $a_1 = 1$ is the only wave incident in the structure, the outgoing waves from the junctions J and J₁ are

$$b_1 = \Gamma_{11} + a_2\Gamma_{12} , \tag{1}$$

$$b_2 = \Gamma_{12} + a_2\Gamma_{22} , \tag{2}$$

$$a_2 = b_2\Gamma_{55} + a_3\Gamma_{35} , \tag{3}$$

$$b_4 = b_2\Gamma_{45} + a_3\Gamma_{34} , \tag{4}$$

$$a_3 = a_3\Gamma_{33} + b_2\Gamma_{35} , \tag{5}$$

where Γ_{mn} represents the scattering coefficient between ports m and n.

From these five equations the scattered waves

$$b_1 = \Gamma_{11} \frac{\left(1 - \Gamma_{22}\Gamma_{55} + \dfrac{\Gamma_{12}^2\Gamma_{55}}{\Gamma_{11}}\right)(1 - \Gamma_{33}) - \Gamma_{22}\Gamma_{35}^2 + \dfrac{\Gamma_{12}^2\Gamma_{35}^2}{\Gamma_{11}}}{(1 - \Gamma_{22}\Gamma_{55})(1 - \Gamma_{33}) - \Gamma_{22}\Gamma_{35}^2} \tag{6}$$

and

$$b_4 = \Gamma_{12}\Gamma_{45} \frac{1 - \Gamma_{33} + \dfrac{\Gamma_{35}\Gamma_{34}}{\Gamma_{45}}}{(1 - \Gamma_{22}\Gamma_{55})(1 - \Gamma_{33}) - \Gamma_{22}\Gamma_{35}^2} \tag{7}$$

are derived. These expressions can be simplified by assuming

$$\begin{array}{c} |\Gamma_{33}| \\ |\Gamma_{44}| \end{array} \ll \begin{array}{c} |\Gamma_{12}| \\ |\Gamma_{35}| \\ |\Gamma_{45}| \end{array} \tag{8}$$

and

$$\left.\begin{array}{l} \mid \Gamma_{12} \mid \\ \mid \Gamma_{35} \mid \\ \mid \Gamma_{45} \mid \end{array}\right. \ll 1. \tag{9}$$

The first assumption means that the TE_{01}° mode incident on the diameter discontinuity has negligible reflection in the same mode. This is a familiar approximation in multimode waveguide calculations, and it will also be seen later that these reflections are indeed negligible. The second assumption implies that the resonating mode TE_{02}° is loosely coupled to the other modes TE_{01}° and $\text{TE}_{10}^{\square}$.

Substituting (8) and (9) in the conservation of energy relations[9] applicable to junctions J and J_1 of Fig. 6, and neglecting terms of higher order than two, one obtains the following results:

$$\Gamma_{22}\Gamma_{55} = e^{i\varphi}(1 - \mid \Gamma_{35} \mid^2 - \tfrac{1}{2} \mid \Gamma_{12} \mid^2), \tag{10}$$

$$\Gamma_{22}\Gamma_{35}^{\;2} = - \mid \Gamma_{35} \mid^2 e^{i(\theta+\varphi)}, \tag{11}$$

$$\frac{\Gamma_{12}^{\;2}\Gamma_{55}}{\Gamma_{11}} = - \mid \Gamma_{12} \mid^2 e^{i\varphi}, \tag{12}$$

$$\frac{\Gamma_{35}\Gamma_{34}}{\Gamma_{45}} = e^{i\theta}, \tag{13}$$

in which

$$\theta = -\psi_1 + \theta_{35} - \theta_{45} + \theta_{34} \tag{14}$$

and

$$\varphi = -\psi_2 + \theta_{22} + \theta_{55}\,, \tag{15}$$

where ψ_1 and ψ_2 are the electrical distances between the branching cavity discontinuities in terms of the nonresonating and resonating modes, respectively, and θ_{mn} is the phase of the scattering coefficient between ports m and n when the waveguides are reduced to zero length.

The physical meaning of θ and φ will be given later.

Substituting (8), (10), (11), (12) and (13) in (6) and (7), leads to the simplified scattered waves

$$b_1 = \Gamma_{11} \frac{1 - \left[1 - \mid \Gamma_{35} \mid^2 (1 + e^{i\theta}) + \dfrac{\mid \Gamma_{12} \mid^2}{2}\right] e^{i\varphi}}{1 - \left[1 - \mid \Gamma_{35} \mid^2 (1 + e^{i\theta}) - \dfrac{\mid \Gamma_{12} \mid^2}{2}\right] e^{i\varphi}} \tag{16}$$

and

$$b_4 = \Gamma_{12}\Gamma_{45} \frac{1 + e^{i\theta}}{1 - \left[1 - |\Gamma_{35}|^2(1 + e^{i\theta}) - \frac{|\Gamma_{12}|^2}{2}\right] e^{i\varphi}}. \quad (17)$$

The values of θ and φ given in (14) and (15) are frequency-sensitive, so we define

$$\theta = \theta_0 + \Delta\theta \quad (18)$$

and

$$\varphi = \varphi_0 + \Delta\varphi, \quad (19)$$

where θ_0 and φ_0 are the values taken by θ and φ at midband frequency f_0, and $\Delta\theta$ and $\Delta\varphi$ are their small departures when the frequency is

$$f = f_0 + \Delta f \quad (20)$$

and

$$\frac{\Delta f}{f_0} \ll 1. \quad (21)$$

In this paper we choose to have the branching cavity resonating with an odd number of half wavelengths. In order to have the branching cavity resonating with an even number of half wavelengths it would be necessary to make resonant the short-circuited half of the branching cavity.

Resonance of the branching cavity (minimization of the reflected wave b_1) occurs at midband $f = f_0$ when the following relation is satisfied:

$$\varphi_0 = |\Gamma_{35}|^2 \sin\theta_0 - 2\pi s, \quad (22)$$

in which s is an integer.

If one substitutes (18), (19) and (22) in (16) and (17), and again neglects higher-order terms, the scattered waves of the branching cavity become

$$b_1 = e^{i\theta_{11}} \frac{-i\Delta\varphi + |\Gamma_{35}|^2(1 + \cos\theta_0) - \frac{|\Gamma_{12}|^2}{2}}{-i\Delta\varphi + |\Gamma_{35}|^2(1 + \cos\theta_0) + \frac{|\Gamma_{12}|^2}{2}}, \quad (23)$$

$$b_4 = \Gamma_{12}\Gamma_{45} \frac{1 + e^{i\theta_0}}{-i\Delta\varphi + |\Gamma_{35}|^2(1 + \cos\theta_0) + \frac{|\Gamma_{12}|^2}{2}}. \quad (24)$$

Several important results are deduced from the last three equations:

(a) The resonance condition, (22), as well as the scattered waves, (23) and (24), depend on the angle θ_0, which, according to (14) and (18), is the midband electrical length difference between the two possible paths that the waves may follow. These two paths are shown in both Figs. 3 and 7. The last figure is a reproduction of the branching cavity from Fig. 1, and the two paths, as well as the modes with which they are measured, have been indicated in it. If θ_0 is an odd multiple of π, the two waves cancel each other and there is no transmission through the cavity in spite of its resonance.

(b) The transmitted and reflected waves, (23) and (24), depend on $\Delta\varphi$, which is the change with frequency of the electrical length ψ_2 of the distance between diameter changes measured in the resonant mode $\mathrm{TE}_{02}{}^{\circ}$. But the same scattered waves are independent of $\Delta\theta$, which is the change with frequency of the electrical length θ_0.

Now it is possible to express the scattering coefficients of the open-circuited half branching cavity matrix,

$$\begin{vmatrix} S_{11} & S_{14} \\ S_{14} & S_{44} \end{vmatrix}, \tag{25}$$

in terms of b_1 and b_4.

From the definition of scattering coefficients it follows that

$$S_{11} = b_1, \tag{26}$$

$$S_{14} = b_4 \tag{27}$$

and from the conservation of energy relations[9] that

$$S_{11}{}^*S_{14} + S_{14}{}^*S_{44} = 0, \tag{28}$$

where S^* is the complex conjugate of S.

From (26), (27) and (28)

$$S_{44} = -b_1{}^*e^{i2\theta_{14}-i\psi_2}. \tag{29}$$

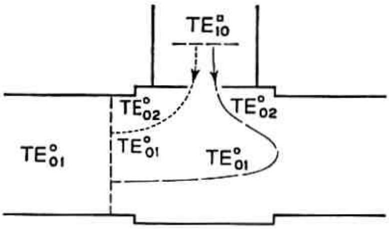


Fig. 7 — Branching cavity; path difference = θ_0.

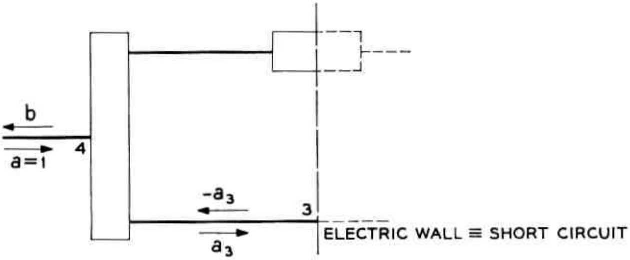


Fig. 8 — Short-circuited half of branching cavity.

4.2 *Short-Circuited Half of Branching Cavity*

Consider again the branching cavity equivalent circuit in Fig. 5, and let BB represent a perfectly conducting surface. One of the halves of the bisected circuit is shown in Fig. 8. A unitary wave fed in port 4 yields, because of the absence of resonance

$$b \cong -e^{-i\psi_1+2i\theta_{34}}. \tag{30}$$

The scattering matrix is $S = b$.

4.3 *Scattering Matrix of the Branching Cavity*

Fig. 9 represents the branching cavity. Considering symmetry and reciprocity, the scattering matrix is

$$\begin{vmatrix} S_{66} & S_{67} & S_{68} \\ S_{67} & S_{77} & S_{67} \\ S_{68} & S_{67} & S_{66} \end{vmatrix}. \tag{31}$$

All the scattering coefficients in (31) are determined as follows:

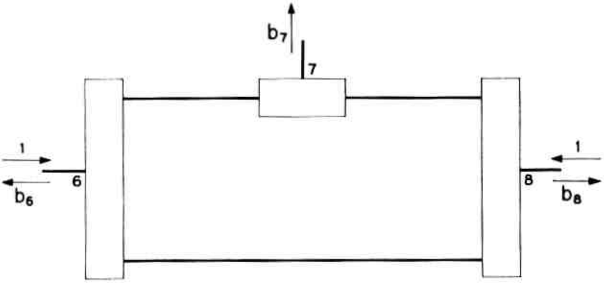


Fig. 9 — Branching cavity with waves fed in phase.

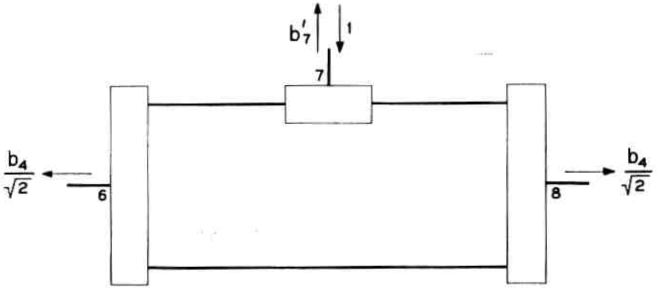


Fig. 10 — Branching cavity with signal fed in branching arm.

Feed unit power into ports 6 and 8 such that the phases of the electric fields are the same. The plane of symmetry becomes a magnetic plane, and the scattered waves derived from (26), (27), (29) and (30) are

$$b_6 = S_{66} + S_{68} = -b_1{}^* e^{i2\theta_{14} - i\psi_2}, \tag{32}$$

$$b_7 = 2S_{67} = \sqrt{2} b_4 . \tag{33}$$

If unit power is fed into port 7, of Fig. 10, the reflection is

$$b_7' = S_{77} = b_1 . \tag{34}$$

Finally, if port 6 and 8 of Fig. 11 are fed 180° out of phase with unit power, the plane of symmetry which has zero tangential electric field becomes a short circuit, and the scattered wave is obtained from (30):

$$b_6' = S_{66} - S_{68} = -e^{-i\psi_1 + i2\theta_{34}}. \tag{35}$$

Substituting the explicit values of b_1 and b_4 given in (23) and (24), in (32), (33), (34) and (35), and, solving this set of equations for the

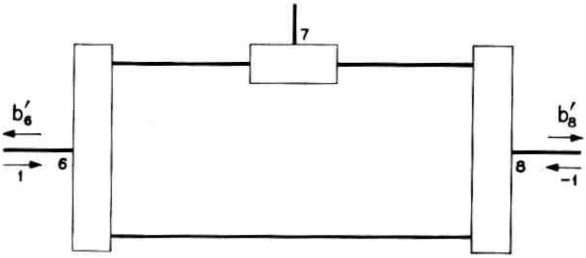


Fig. 11 — Branching cavity with waves fed in opposite phase.

scattering coefficients of the branching cavity, one obtains

$$S_{66} = -e^{i(2\theta_{34}-\psi_1)} \frac{|\Gamma_{35}|^2(1+\cos\theta_0)}{-i\Delta\varphi + \frac{|\Gamma_{12}|^2}{2} + |\Gamma_{35}|^2(1+\cos\theta_0)}, \tag{36}$$

$$S_{67} = e^{i(\varphi/2+\theta_{34}-\psi_1/2+\theta_{11}/2)} \frac{|\Gamma_{12}|\,|\Gamma_{35}|(1+\cos\theta_0)^{\frac{1}{2}}}{-i\Delta\varphi + \frac{|\Gamma_{12}|^2}{2} + |\Gamma_{35}|^2(1+\cos\theta_0)}, \tag{37}$$

$$S_{68} = e^{i(2\theta_{34}-\psi_1)} \frac{-i\Delta\varphi + \frac{|\Gamma_{12}|^2}{2}}{-i\Delta\varphi + \frac{|\Gamma_{12}|^2}{2} + |\Gamma_{35}|^2(1+\cos\theta_0)}, \tag{38}$$

$$S_{77} = e^{i\theta_{11}} \frac{-i\Delta\varphi - \frac{|\Gamma_{12}|^2}{2} + |\Gamma_{35}|^2(1+\cos\theta_0)}{-i\Delta\varphi + \frac{|\Gamma_{12}|^2}{2} + |\Gamma_{35}|^2(1+\cos\theta_0)}. \tag{39}$$

V. SCATTERING MATRIX OF THE REJECTING CAVITY

The elements of the scattering matrix of the rejecting cavity

$$\begin{vmatrix} \bar{S}_{66} & \bar{S}_{68} \\ \bar{S}_{68} & \bar{S}_{66} \end{vmatrix} \tag{40}$$

can be deduced from those of the branching cavity (36) and (38) by eliminating the coupling to the rectangular waveguide, that is, making

$$\Gamma_{12} = 0. \tag{41}$$

The dash over the characters is to distinguish them from those of the branching cavity:

$$\bar{S}_{66} = -e^{i(2\bar{\theta}_{34}-\bar{\psi}_1)} \frac{1}{1 - \frac{i\Delta\bar{\varphi}}{|\bar{\Gamma}_{35}|^2(1+\cos\bar{\theta}_0)}} \tag{42}$$

$$\bar{S}_{68} = e^{i(2\bar{\theta}_{34}-\psi_1)} \frac{\frac{i\Delta\bar{\varphi}}{|\bar{\Gamma}_{35}|^2(1+\cos\bar{\theta}_0)}}{1 - \frac{i\Delta\bar{\varphi}}{(\bar{\Gamma}_{35})^2(1+\cos\bar{\theta}_0)}}. \tag{43}$$

At midband, $\Delta\bar{\varphi} = 0$ and

$$\bar{S}_{66} = -e^{i(2\theta_{34}-\psi_1)}. \tag{44}$$

This means that at resonance the cavity acts as a short circuit located at half the length of the cavity. Again, as in the case of the branching cavity, if

$$\bar{\theta}_0 = (2n + 1)\pi \tag{45}$$

there is no resonance.

VI. SCATTERING COEFFICIENTS OF THE CHANNEL-DROPPING FILTER

Knowing the scattering matrices of the branching and rejecting cavities, we will find the scattering elements of the circuit obtained by joining a branching cavity and a rejecting cavity with a piece of waveguide of electrical length ψ. From the block diagram representation of the branching filter in Fig. 12,

$$B_1 = S_{66} + A S_{68}\,, \tag{46}$$

$$B_2 = S_{67}(1 + A), \tag{47}$$

$$B_3 = Be^{-i\psi}\bar{S}_{68}\,, \tag{48}$$

$$B = S_{68} + A S_{66}\,, \tag{49}$$

$$A = B\bar{S}_{66}e^{-i2\psi}. \tag{50}$$

First it will be demonstrated that, if certain conditions are satisfied, port 1 of the filter (Fig. 12) is matched at all frequencies; then the values of B_2 and B_3 will be ascertained.

From (46), (49) and (50)

$$B_1 = S_{66}\,\frac{1 - \left(S_{66}\bar{S}_{66} - \dfrac{\bar{S}_{66}}{S_{66}}\,S_{68}^{\,2}\right)e^{-i2\psi}}{1 - S_{66}\bar{S}_{66}e^{-i2\psi}}. \tag{51}$$

Replacing S_{66}, S_{68} and $\bar{S}_{66}$ in the numerator with the values given in (36), (38) and (42), one obtains

$$B_1 = S_{66}\,\frac{1 - e^{i2(\theta_{34}+\theta_{34}-\psi-\psi_1/2-\bar{\psi}_1/2)}\,\dfrac{1 - \dfrac{\dfrac{|\,\Gamma_{12}\,|^2}{2} - i\Delta\varphi}{|\,\Gamma_{35}\,|^2\,(1 + \cos\theta_0)}}{1 - \dfrac{i\Delta\bar{\varphi}}{|\,\bar{\Gamma}_{35}\,|^2\,(1 + \cos\bar{\theta}_0)}}}{1 - S_{66}\bar{S}_{66}\,e^{-i2\psi}}. \tag{52}$$

In order to have port 1 of the channel-dropping filter matched at all

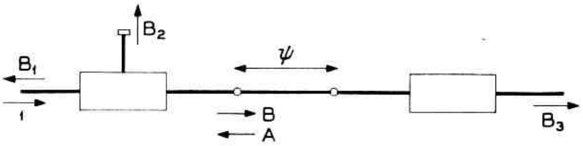


Fig. 12 — Block diagram of the channel-dropping filter.

frequencies, the reflected wave B_1 must vanish. The conditions to be satisfied are

$$\bar{\Gamma}_{35} = \Gamma_{35}\,, \tag{53}$$

$$\bar{\theta}_0 = \theta_0\,, \tag{54}$$

$$\bar{\theta}_{34} = \theta_{34}\,, \tag{55}$$

$$\Delta\bar{\varphi} = \Delta\varphi, \tag{56}$$

$$\theta_{34} + \bar{\theta}_{34} - \psi - \frac{\psi_1}{2} - \frac{\bar{\psi}_1}{2} = -\frac{\pi}{2}(1 + 2p), \tag{57}$$

where p is an arbitrary integer, and

$$|\,\Gamma_{12}\,|^2 = 4\,|\,\Gamma_{35}\,|^2(1 + \cos\theta_0)\,. \tag{58}$$

Conditions (53), (54), (55) and (56) state that, except for a small-length correction due to the effect of the coupling to the dropping waveguide, both resonating cavities must be equal. Condition (57) establishes that the distance between the centers of the resonating cavities must be an odd number of quarters of guided wavelength of the nonresonant mode. Since this condition is fulfilled rigorously only at discrete frequencies, the length of the cavities and the distance between them must be selected as short as possible. Finally, condition (58) states that in the branching cavity, Fig. 7, the power coupled from the resonating mode $TE_{02}°$ to the dropped mode $TE_{10}□$, must be equal to four times the power coupled to each one of the $TE_{01}°$ ports.

Substituting (53) through (58) in the expressions of the branched and through waves (47) and (48), one finds that

$$B_2 = e^{i(\varphi/2+\theta_{34}-\psi_1/2+\theta_{11}/2)}\,\frac{1}{1 - \dfrac{i\Delta\varphi}{2\,|\,\Gamma_{35}\,|^2\,(1 + \cos\theta_0)}}, \tag{59}$$

$$B_3 = -e^{i(-\psi+4\theta_{34}-2\psi_1)}\,\frac{\dfrac{i\Delta\varphi}{2\,|\,\Gamma_{35}\,|^2\,(1 + \cos\theta_0)}}{1 - \dfrac{i\Delta\varphi}{2\,|\,\Gamma_{35}\,|^2\,(1 + \cos\theta_0)}}. \tag{60}$$

At resonance, $\Delta\varphi = 0$ and B_2 and B_3 become

$$B_2 = e^{i(\varphi_0/2+\theta_{34}-\psi_1/2+\theta_{11}/2)}, \tag{61}$$

$$B_3 = 0. \tag{62}$$

Far from resonance,

$$\frac{|\,i\Delta\varphi\,|}{2\,|\,\Gamma_{35}\,|^2\,(1+\cos\theta_0)} \gg 1 \tag{63}$$

and

$$B_2 = 0, \tag{64}$$

$$B_3 = e^{i(-\psi+4\theta_{34}-2\psi_1)}. \tag{65}$$

In order to introduce the concept of loaded Q or Q_L of the channel-dropping filter, the value of $\Delta\varphi$ will be expressed as a function of frequency. $\Delta\varphi$ is the difference between the electrical length of the resonating cavity at midband f_0 and at any other frequency $f_0 + \Delta f$. From (15),

$$\Delta\varphi = \Delta f\,\frac{d\varphi_0}{df_0} \cong -\frac{\Delta f}{f_0}\,\psi_{20}\,\frac{{\lambda_{g2}}^2}{{\lambda_0}^2} + \Delta f\,\frac{d}{df_0}\,(\theta_{22}+\theta_{55}) \tag{66}$$

provided that

$$\left(\frac{\lambda_0}{\lambda_{g2}}\right)^2 \gg \frac{|\,2\Delta f\,|}{f_0}, \tag{67}$$

in which λ_0 and λ_{g2} are the free-space wavelength and guided wavelength of the resonating mode at midband f_0, and ψ_{20} is the midband electrical distance between the diameter discontinuities of the branching cavity measured in terms of the resonating TE_{02}° mode.

Substituting (66) in (59) and (60) leads to

$$B_2 = e^{i(\varphi/2+\theta_{34}-\psi_1/2+\theta_{11}/2)}\,\frac{1}{1+i2Q_L\,\dfrac{\Delta f}{f_0}}, \tag{68}$$

$$B_3 = e^{i(-\psi+4\theta_{34}-2\psi_1)}\,\frac{i2Q_L\,\dfrac{\Delta f}{f_0}}{1+i2Q_L\,\dfrac{\Delta f}{f_0}}, \tag{69}$$

where

$$Q_L = \frac{\psi_{20}\left(\dfrac{\lambda_{g2}}{\lambda_0}\right)^2 - f_0\,\dfrac{d}{df_0}\,(\theta_{22}+\theta_{55})}{4\,|\,\Gamma_{35}\,|^2\,(1+\cos\theta_0)}. \tag{70}$$

The loaded Q is, as expected, inversely proportional to the power coupled into the resonant mode, but that coupling is not enough to insure a finite Q_L. If θ_0, the electrical path difference discussed above and shown in Figs. 3 and 7, is an odd multiple of π, then Q_L becomes infinitely large. Also, as expected, at the frequency at which the resonating mode passes through cutoff, λ_{g2} becomes infinite and Q_L diverges.

It can be shown that (68) and (69) are the transfer coefficients of the low-frequency circuit in Fig. 2, and consequently this circuit is the equivalent to that in Fig. 1.

For the purpose of testing the cavities independently of each other it is important to know their scattering coefficients. They are obtained by substituting (66) and (70) in (36), (37), (38), (39), (42) and (43):

$$S_{66} = -e^{i(2\theta_{34}-\psi_1)} \frac{1}{3 + i4 \dfrac{\Delta f}{f_0} Q_L}, \tag{71}$$

$$S_{67} = e^{i(\varphi/2+\theta_{34}-\psi_1/2+\theta_{11}/2)} \frac{2}{3 + i4 \dfrac{\Delta f}{f_0} Q_L}, \tag{72}$$

$$S_{68} = e^{i(2\theta_{34}-\psi_1)} \frac{2 + i4 \dfrac{\Delta f}{f_0} Q_L}{3 + i4 \dfrac{\Delta f}{f_0} Q_L}, \tag{73}$$

$$S_{77} = e^{i\theta_{11}} \frac{-1 + i4 \dfrac{\Delta f}{f_0} Q_L}{3 + i4 \dfrac{\Delta f}{f_0} Q_L}, \tag{74}$$

$$\bar{S}_{66} = -e^{i(2\theta_{34}-\psi_1)} \frac{1}{1 + i4 \dfrac{\Delta f}{f_0} Q_L}, \tag{75}$$

$$\bar{S}_{68} = e^{i(2\theta_{34}-\psi_1)} \frac{i4 \dfrac{\Delta f}{f_0} Q_L}{1 + i4 \dfrac{\Delta f}{f_0} Q_L}. \tag{76}$$

Considering first the branching cavity, from (71) to (74), it is concluded that at midband the amplitude of the reflection at any port is one-third and the amplitude of the transmission to any other port is two-thirds. Furthermore, the half power band of the reflection characteristic, (71), is

$$\frac{3}{2}\frac{f_0}{Q_L}.$$

For the rejecting cavity, the loaded Q, (75) or (76), is twice the loaded Q of design of the channel-dropping filter.

VII. RÉSUMÉ OF FORMULAS FOR THE DIMENSIONING OF A CHANNEL-DROPPING FILTER

The information given is the midband frequency f_0 and the half power bandwidth of the dropped channel $2\Delta f_0$, defined in terms of the loaded Q:

$$Q_L = \frac{f_0}{2\Delta f_0}. \tag{77}$$

The unknowns are:

ψ_{20}, midband electrical distance between diameter discontinuities of the branching cavity measured in terms of the resonating mode,

$\bar{\psi}_{20}$, midband electrical distance between diameter discontinuities of the rejecting cavity measured in terms of the resonating mode,

ψ_d, midband electrical distance between centers of cavities in terms of the nonresonating mode,

$2\Gamma_{35}(1 + \cos\theta_0)^{\frac{1}{2}}$, coupling coefficient between the resonating mode and the through waveguide,

$\sqrt{2}\Gamma_{12}$, coupling coefficient between the resonating mode and the dropping mode.

From (15) and (22),

$$\psi_{20} = \theta_{22} + \theta_{55} - |\Gamma_{35}|^2 \sin\theta_0 + 2\pi s = \frac{2\pi l_B}{\lambda_{g2}}, \tag{78}$$

where θ_0 reproduced from (14) is

$$\theta_0 = -\psi_{10} + \theta_{34} + \theta_{35} - \theta_{45} \tag{79}$$

and ψ_{10} is the midband electrical distance between diameter discontinuities of the branching cavity measured in terms of the nonresonating mode; θ_{mn} is the phase of the scattering coefficients between ports m and n, Fig. 4, with waveguide lengths reduced to zero; l_B is the length of the branching cavity and λ_{g2} is the midband resonant mode guided wavelength.

For the rejecting cavity $\theta_{22} = 0$; then, from (78),

$$\bar{\psi}_{20} = \theta_{55} - |\Gamma_{35}|^2 \sin \bar{\theta}_0 + 2\pi s = \frac{2\pi l_R}{\lambda_{g2}}, \tag{80}$$

l_R being the length of the rejecting cavity, and

$$\bar{\theta}_0 = -\bar{\psi}_{10} + \theta_{34} + \theta_{35} - \theta_{45}. \tag{81}$$

From (57),

$$\psi_d = \psi_0 + \frac{\psi_{10}}{2} + \frac{\bar{\psi}_{10}}{2} - 2\theta_{34} = \frac{\pi}{2}(1 + 2p), \tag{82}$$

where ψ_0 is the midband electrical distance between cavities and p is an arbitrary integer. From (58) and (70),

$$|\Gamma_{35}|^2(1 + \cos\theta_0) = \frac{|\Gamma_{12}|^2}{4} = \frac{\psi_{20}\left(\frac{\lambda_{g2}}{\lambda_0}\right)^2 - f_0 \frac{d}{df_0}(\theta_{22} + \theta_{55})}{4Q_L}. \tag{83}$$

From the theory of diffraction by small holes,[10]

$$\theta_{22} = \frac{4\pi c^3}{3\lambda_0} \sqrt{\frac{\mu}{\epsilon}} \frac{|H_2{}^2|}{P_2}, \tag{84}$$

$$|\Gamma_{12}| = \frac{2^{\frac{5}{2}}}{3} \frac{\pi c^3}{\lambda_0} \sqrt{\frac{\mu}{\epsilon}} \frac{|H_1||H_2|}{\sqrt{P_1 P_2}}, \tag{85}$$

where c is the radius of the round hole that couples the resonating to the branching mode, and μ and ϵ are the permeability and permittivity of free space. If one considers that the standing resonating field is made of two waves propagating in opposite directions, $|H_2|$ is the absolute value of the magnetic field of one of those waves at the hole and P_2 is the average power carried by such a wave; $|H_1|$ is the absolute value of the magnetic field at the hole of a wave in the branching waveguide and P_1 is the average power carried by such a wave.

The values of θ_{22}, θ_{34}, θ_{55}, θ_0, $|\Gamma_{35}|$, H_1, H_2, P_1, P_2 depend on the particular structure selected for the filter.

7.1 *Channel-Dropping Filter from Mode $TE_{01}{}^\circ$ in Circular Waveguide to $TE_{10}{}^\square$ in Rectangular Waveguide*

In Fig. 1 let us call a and b the radii of the double- and single-mode regions respectively, and W and d the width and height of the rectangular waveguide.

From Ref. 11,

$$\theta_{55} = -2 \operatorname{arctg} \left\{ \frac{Y_{2a}}{|Y_{2b}|} \left[\frac{J_1[v_2(1-\delta)]}{\delta\left(1+\dfrac{\delta}{2}\right) v_2 J_0(v_2)} \right]^2 \right\} \pm \pi, \tag{86}$$

$$\theta_{35} - \theta_{45} = 2\,\frac{|Y_{2b}|}{Y_{1a}} \pm \pi, \tag{87}$$

$$\theta_{34} = 0 \tag{88}$$

and

$$|\Gamma_{35}|^2 \cong \frac{Y_{1a}}{Y_{2a}} \left[\frac{2\delta v_2^2}{v_2^2 - v_1^2} \frac{J_0(v_2)}{J_0(v_1)} \frac{J_1[v_1(1-\delta)]}{J_1[v_2(1-\delta)]} \right]^2 \frac{1 + \left|\dfrac{Y_{2b}}{Y_{1b}}\right|^2}{1 + \left|\dfrac{Y_{2b}}{Y_{2a}}\right|^2}, \tag{89}$$

where

$$Y_{mb} = \sqrt{\frac{\epsilon}{\mu}} \sqrt{1 - \left(\frac{v_m \lambda_0}{2\pi b}\right)^2}, \tag{90}$$

$$Y_{na} = \sqrt{\frac{\epsilon}{\mu}} \sqrt{1 - \left(\frac{v_n \lambda_0}{2\pi a}\right)^2}, \tag{91}$$

$$\delta = 1 - \frac{b}{a}. \tag{92}$$

J_n is the Bessel function of the first kind and order n, v_p is the pth root of the J_1 function.

From Ref. 12, pp. 58–59,

$$\frac{H_2^2}{P_2} = \frac{v_2^2 \sqrt{\dfrac{\epsilon}{\mu}}}{2\pi^3 \sqrt{1 - \left(\dfrac{v_2\lambda_0}{2\pi a}\right)^2}} \frac{\lambda^2}{a^4}. \tag{93}$$

From Ref. 12, p. 55, if the round hole is at the center of the rectangular waveguide cross section,

$$\frac{|H_1|}{\sqrt{P_1}} = 2\left[\frac{\sqrt{\dfrac{\epsilon}{\mu}} \sqrt{1 - \left(\dfrac{\lambda_0}{2W}\right)^2}}{Wd} \right]^{\frac{1}{2}}. \tag{94}$$

Substituting (93) and (94) in (84) and (85) results in

$$\theta_{22} = \frac{2\upsilon_2^2}{3\pi^2 \sqrt{1 - \left(\frac{\upsilon_2\lambda_0}{2\pi a}\right)^2}} \frac{\lambda_0 c^3}{a^4}, \tag{95}$$

$$|\Gamma_{12}| = \frac{8\upsilon_2}{3\pi^{\frac{1}{2}}} \left[\frac{1 - \left(\frac{\lambda_0}{2W}\right)^2}{1 - \left(\frac{\upsilon_2\lambda_0}{2\pi a}\right)^2}\right]^{\frac{1}{2}} \frac{c^3}{(Wd)^{\frac{1}{4}}a^2}. \tag{96}$$

Expressions (78), (79), (80), (81), (82), (83), (86), (89), (95) and (96) are the general relations necessary to determine the dimensions of the filter. As an aid in their solution it is convenient to have the approximate results obtained when the expressions are drastically simplified by the following assumptions: all corrective terms due to coupling effects are neglected; the cutoff radius for $TE_{02}°$ at midband is selected at $a(1 - \delta/2)$; and

$$\upsilon_2\delta \ll 1. \tag{97}$$

These approximate results are

$$\psi_{20} = \bar{\psi}_{20} = \theta_{55} + 2\pi s = \frac{2\pi l_B\sqrt{\delta}}{\lambda_0} = \frac{2\pi l_R\sqrt{\delta}}{\lambda_0}, \tag{98}$$

$$\theta_0 = \bar{\theta}_0 = -\psi_{10} \pm \pi = -\frac{2\pi l_B}{\lambda_0} \pm \pi, \tag{99}$$

$$\psi_d = \psi + \psi_{10} = \frac{\pi}{2}(1 + 2p), \tag{100}$$

$$|\Gamma_{35}|^2 \left(1 + \cos\frac{\pi}{2\sqrt{\delta}}\right) = \frac{|\Gamma_{12}|^2}{4} = \frac{\psi_{20}}{4Q_L\delta}, \tag{101}$$

$$\theta_{55} = \frac{\pi}{2}, \tag{102}$$

$$|\Gamma_{35}|^2 = 2\delta^{\frac{3}{2}} \left(\frac{\upsilon_1\upsilon_2}{\upsilon_2^2 - \upsilon_1^2}\right)^2, \tag{103}$$

$$|\Gamma_{12}| = \frac{8\upsilon_2}{3\pi^{\frac{1}{2}}} \left[\frac{1 - \left(\frac{\lambda_0}{2W}\right)^2}{\delta}\right]^{\frac{1}{4}} \frac{c^3}{(Wd)^{\frac{1}{4}}a^2}, \tag{104}$$

and explicitly,

$$a = \frac{\upsilon_2\lambda_0}{2\pi}\left(1 + \frac{\delta}{2}\right) = 1.117\,\lambda_0\left(1 + \frac{\delta}{2}\right), \tag{105}$$

$$b = \frac{\upsilon_2\lambda_0}{2\pi}\left(1 - \frac{\delta}{2}\right) = 1.117\,\lambda_0\left(1 - \frac{\delta}{2}\right), \tag{106}$$

$$l_B = l_R = \frac{\lambda_0}{4\sqrt{\delta}}\,(1 + 4s), \tag{107}$$

$$l = \frac{\lambda_0}{4}\left(1 + 2p - \frac{1 + 4s}{\sqrt{\delta}}\right), \tag{108}$$

$$\begin{aligned} 2c &= \lambda_0\left\{\frac{9\upsilon_2^{\,2}}{32\pi^2}\,\frac{Wd(1 + 4s)}{\lambda_0^{\,2}Q_L\delta^{\frac{1}{2}}\left[1 - \left(\frac{\lambda}{2W}\right)^2\right]^{\frac{1}{2}}}\right\}^{\frac{1}{6}} \\ &= \lambda_0\left\{1.4\,\frac{Wd(1 + 4s)}{\lambda_0^{\,2}Q_L\delta^{\frac{1}{2}}\left[1 - \left(\frac{\lambda_0}{2W}\right)^2\right]^{\frac{1}{2}}}\right\}^{\frac{1}{6}}, \end{aligned} \tag{109}$$

$$\begin{aligned} \delta\left(1 - \cos\frac{\pi}{2\sqrt{\delta}}\right)^{\frac{2}{3}} &= \left(\frac{\pi}{16}\right)^{\frac{2}{3}}\left[\frac{\upsilon_2^{\,2} - \upsilon_1^{\,2}}{\upsilon_1\upsilon_2}\right]^{\frac{1}{3}}\left(\frac{1 + 4s}{Q_L}\right)^{\frac{2}{3}} \\ &= 0.639\left(\frac{1 + 4s}{Q_L}\right)^{\frac{2}{3}}. \end{aligned} \tag{110}$$

Since Q_L and λ_0 are given, the dimensions of the filter can be obtained by calculating δ from (110) by successive approximations and substituting this value in the relations (105) through (108).

Far from resonance, the amplitude of the reflection of the TE_{01}° mode at each diameter change, derived from Ref. 11, is

$$|\,\Gamma_{11}\,| = |\,\Gamma_{33}\,| = \left(\frac{\upsilon_1}{\upsilon_2}\right)^2\frac{\delta}{2} = 0.15\,\delta. \tag{111}$$

7.2 *Channel-Dropping Filter from Mode $TE_{10}^{\square}$ in Rectangular Waveguide to $TE_{10}^{\square}$ in Rectangular Waveguide*

Calling a and b the widths of the through waveguide in the double and single mode regions, respectively, W the width of the branching rectangular waveguide and d the height of all of them, one obtains from Ref. 11

$$\theta_{55} = -2\,\mathrm{arctg}\,\frac{Y_{2a}}{|\,Y_{2b}\,|}\,\frac{(\sin 2\pi\delta)^2}{(2\pi\delta)} \pm \pi, \tag{112}$$

$$\theta_{35} - \theta_{45} = 2\,\frac{|\,Y_{2b}\,|}{Y_{1b}} \pm \pi, \tag{113}$$

$$\theta_{34} = 0, \tag{114}$$

$$|\,\Gamma_{35}\,|^2 = \frac{Y_{1a}}{Y_{2a}}\left(\frac{8\delta}{3}\,\frac{\sin\pi\delta}{\sin 2\pi\delta}\right)^2 \frac{1 + \left|\dfrac{Y_{2b}}{Y_{1b}}\right|^2}{1 + \left|\dfrac{Y_{2b}}{Y_{2a}}\right|^2}, \tag{115}$$

where

$$Y_{mb} = \sqrt{\frac{\epsilon}{\mu}}\sqrt{1 - \left(\frac{m\lambda_0}{2b}\right)^2}, \tag{116}$$

$$Y_{na} = \sqrt{\frac{\epsilon}{\mu}}\sqrt{1 - \left(\frac{n\lambda_0}{2a}\right)^2}, \tag{117}$$

$$\delta = 1 - \frac{b}{a}. \tag{118}$$

From Ref. 9, p. 55, if the round hole is at the center of the branching rectangular waveguide cross section,

$$\frac{H_2^{\,2}}{P_2} = \frac{4\sqrt{\dfrac{\epsilon}{\mu}}}{\sqrt{1 - \left(\dfrac{\lambda_0}{a}\right)^2}}\,\frac{\lambda_0^{\,2}}{a^3 d}, \tag{119}$$

$$\frac{H_1}{\sqrt{P_1}} = 2\left[\frac{\sqrt{\dfrac{\epsilon}{\mu}}\sqrt{1 - \left(\dfrac{\lambda_0}{2W}\right)^2}}{Wd}\right]^{\frac{1}{2}}. \tag{120}$$

Substituting (119) and (120) in (84) and (85) yields

$$\theta_{22} = \frac{16\pi}{3\sqrt{1 - \left(\dfrac{\lambda_0}{a}\right)^2}}\,\frac{\lambda_0 c^3}{a^3 d} \tag{121}$$

$$|\,\Gamma_{12}\,| = \frac{2^{\frac{9}{4}}\pi}{3}\left[\frac{1 - \left(\dfrac{\lambda_0}{2W}\right)^2}{1 - \left(\dfrac{\lambda_0}{a}\right)^2}\right]^{\frac{1}{4}} \frac{c^3}{(a^3 d^2 W)^{\frac{1}{2}}}. \tag{122}$$

Again, expressions (78), (79), (80), (81), (82), (83), (86), (89), (95) and (96) can be simplified under the following assumptions: All corrective terms due to coupling effects are neglected; the cutoff width for

$TE_{20}^{\square}$ at midband is $a(1 - \delta/2)$; and

$$2\pi\delta \ll 1. \tag{123}$$

Then

$$\psi_{20} = \bar{\psi}_{20} = \theta_{55} = \frac{2\pi l_B\sqrt{\delta}}{\lambda_0} = \frac{2\pi l_R\sqrt{\delta}}{\lambda_0}, \tag{124}$$

$$\theta_0 = -\psi_{10} \pm \pi = -\frac{\pi}{2\sqrt{\delta}} \pm \pi, \tag{125}$$

$$\psi_d = \psi + \psi_{10} = \frac{\pi}{2}(1 + 2p), \tag{126}$$

$$|\Gamma_{35}|^2\left(1 + \cos\frac{\pi}{2\sqrt{\delta}}\right) = \frac{|\Gamma_{12}|^2}{4} = \frac{|\psi_{2B}|}{4Q_L\delta}. \tag{127}$$

$$\theta_{55} = \frac{\pi}{2} + 2s\pi, \tag{128}$$

$$|\Gamma_{35}|^2 = \tfrac{8}{9}\,\delta^{\frac{3}{2}}, \tag{129}$$

$$|\Gamma_{12}| = \frac{2^{\frac{1}{2}}\pi}{3}\left[\frac{1 - \left(\frac{\lambda_0}{2W}\right)^2}{\delta}\right]^{\frac{1}{4}} \frac{c^3}{(a^3W)^{\frac{1}{2}}\,d}, \tag{130}$$

and explicitly

$$a = \lambda_0\left(1 + \frac{\delta}{2}\right), \tag{131}$$

$$b = \lambda_0\left(1 - \frac{\delta}{2}\right), \tag{132}$$

$$l_B = l_R = \frac{\lambda_0(1 + 4s)}{4\sqrt{\delta}}, \tag{133}$$

$$l = \frac{\lambda_0}{4}\left(1 + 2p - \frac{1 + 4s}{\sqrt{\delta}}\right), \tag{134}$$

$$\begin{aligned} 2c &= \lambda_0\left\{\frac{9}{8\pi}\,\frac{Wd^2(1 + 4s)}{Q_L\delta^{\frac{1}{2}}\left[1 - \left(\frac{\lambda_0}{2W}\right)^2\right]^{\frac{1}{2}}\lambda_0{}^3}\right\}^{\frac{1}{6}} \\ &= \lambda_0\left\{0.358\,\frac{Wd^2(1 + 4s)}{Q_L\delta^{\frac{1}{2}}\left[1 - \left(\frac{\lambda_0}{2W}\right)^2\right]^{\frac{1}{2}}\lambda_0{}^2}\right\}^{\frac{1}{6}} \end{aligned} \tag{135}$$

$$\left(1 - \cos\frac{\pi}{2\sqrt{\delta}}\right)^{\frac{2}{3}} = \left[\frac{9\pi(1 + 4s)}{64Q_L}\right]^{\frac{2}{3}} = 0.72\left(\frac{1 + 4s}{Q_L}\right)^{\frac{2}{3}}. \tag{136}$$

Given Q_L and λ_0, the dimensioning of the filter is obtained by calculating δ from (136) and replacing this value in the explicit dimensions (131) through (135).

Far from resonance, the amplitude of the reflection of the $\text{TE}_{10}{}^{\square}$ mode at each width change, derived from Ref. 11, is

$$| \Gamma_{11} | = | \Gamma_{13} | = 0.125\delta.$$

VIII. DESIGN OF MODE-CONVERSION BAND-REJECTION FILTERS

In order to build multipole mode-conversion band-rejection filters, it is necessary to know explicitly the scattering coefficients of a single cavity. These coefficients, given in (42) and (43), can be rewritten with the help of (66) as

$$\bar{S}_{66} = -e^{i(2\theta_{34}-\bar{\psi}_1)} \frac{1}{1 + i2\bar{Q}_L \dfrac{\Delta f}{f_0}}, \tag{137}$$

$$\bar{S}_{68} = e^{i(2\theta_{34}-\bar{\psi}_1)} \frac{i2\bar{Q}_L \dfrac{\Delta f}{f_0}}{1 + i2\bar{Q}_L \dfrac{\Delta f}{f_0}}, \tag{138}$$

where

$$\bar{Q}_L = \frac{\bar{\psi}_{20} \left(\dfrac{\bar{\lambda}_{g2}}{\lambda_0} \right)^2 - f_0 \dfrac{d\bar{\theta}_{55}}{df_0}}{2 \, | \bar{\Gamma}_{35} |^2 (1 + \cos \bar{\theta}_0)}. \tag{139}$$

Comparing this equation with (70), we conclude that the band rejected by the band-rejection filter has half the width of the band dropped by a channel-dropping filter using the same rejection cavity. This coincides with the final remark of Section VI. The formulas that yield the dimensions of the rejecting cavity in Section VII can be used, replacing $\bar{Q}_L$ by $\bar{Q}_L/2$.

IX. INTRINSIC Q OF MODE-CONVERSION BAND-REJECTION FILTERS

By definition, the intrinsic Q of a resonating cavity is

$$Q = \omega \left(\frac{\text{energy stored}}{\text{power dissipated as heat}} \right), \tag{140}$$

where ω is the angular frequency.

Let E be the electric field of the resonating mode at any point at the instant when it passes through a maximum, H the magnetic field at the metallic boundary at the instant when it passes through a maximum, σ the conductivity of the metallic wall, ξ the skin depth and μ and ϵ the permeability and permittivity of free space; then

$$Q = \omega\sigma\xi\epsilon \frac{\displaystyle\int_v E^2 dv}{\displaystyle\int_s H^2 ds}, \tag{141}$$

where v and s are the volume and surface of the waveguide.

For the case of the circular electric filter, the fields inside the cavity are

$$E_{\text{in}} = J_1\left(v_2 \frac{r}{a}\right) \cos \frac{2\pi}{\lambda_{g2}} z, \tag{142}$$

$$H_{\text{in}} = \frac{v_2}{a\omega\mu} J_0(v_2) \cos \frac{2\pi}{\lambda_{g2}} z \tag{143}$$

Outside the cavity, because of the boundary conditions of continuity of the tangential field components, they are

$$E_{\text{out}} \cong J_1\left(v_2 \frac{r}{b}\right) \cos \frac{\pi l}{\lambda_{g2}} e^{-(2\pi/\lambda_{g2\,\text{out}})(|z|-l/2)}, \tag{144}$$

$$H_{\text{out}} \cong \frac{v_2}{b\omega\mu} J_0(v_2) \cos \frac{\pi l}{\lambda_{g2}} e^{-(2\pi/\lambda_{g2\,\text{out}})(|z|-l/2)}. \tag{145}$$

The axial coordinate z has its origin in the center of the cavity; the length of the cavity is l; and

$$\lambda_{g2\,\text{out}} = \frac{\lambda_0}{\sqrt{\left(\dfrac{v_2\lambda_0}{2\pi b}\right)^2 - 1}}. \tag{146}$$

Substituting (142) through (145) in (141) leads to

$$Q_{\text{TE}_{02}{}^\circ} = \omega\sigma\xi\epsilon \frac{\displaystyle\int_0^b J_1^{\,2}\left(v_2 \frac{r}{b}\right) r dr}{\left(\dfrac{v_2}{a\omega\mu}\right)^2 J_0^{\,2}(v_2) a} \cdot \frac{\displaystyle\int_{-l/2}^{l/2} \cos^2 \frac{2\pi z}{\lambda_{g2}} dz + 2\left(\frac{a}{b}\right)^2 \cos^2 \frac{\pi l}{\lambda_{g2}} \int_{l/2}^{\infty} e^{-(4\pi/\lambda_{g2\,\text{out}})(|z|-l/2)} dz}{\displaystyle\int_{-l/2}^{l/2} \cos^2 \frac{2\pi z}{\lambda_{g2}} dz + 2\left(\frac{a}{b}\right) \cos^2 \frac{\pi l}{\lambda_{g2}} \int_{l/2}^{\infty} e^{-(4\pi/\lambda_{g2\,\text{out}})(|z|-l/2)} dz}. \tag{147}$$

Since[13]

$$\int_0^b J_1^2\left(\upsilon_2 \frac{r}{b}\right) r dr = \frac{b^2}{2} J_0^2(\upsilon_2), \tag{148}$$

$$\frac{a}{b} \cong 1 \tag{149}$$

and

$$\frac{\upsilon_2 \lambda_0}{2\pi a} \cong 1, \tag{150}$$

(147) becomes

$$Q_{\mathrm{TE}_{02}{}^{\circ}} \cong \frac{a}{\xi}, \tag{151}$$

where ξ, the skin depth, is

$$\xi = \sqrt{\frac{2}{\omega\mu\sigma}}. \tag{152}$$

The result (151) coincides with the intrinsic Q of an infinitely long cylindrical cavity resonating with $\mathrm{TE}_{02}{}^{\circ}$ at cutoff (Ref. 12, p. 59). Similar reasoning for a mode-conversion band-rejection filter in rectangular waveguide yields for the intrinsic Q of the resonating $\mathrm{TE}_{20}{}^{\square}$ mode:

$$Q_{\mathrm{TE}_{20}\square} = \frac{d}{\xi\left(1 + \frac{2d}{a}\right)}, \tag{153}$$

where a and d are the width and height of the rectangular waveguide.

Typical theoretical values in copper waveguides are the following:

For $\mathrm{TE}_{02}{}^{\circ}$ mode at 5.4 millimeters, $a \cong \upsilon_2(5.4/2\pi) = 6.04$ millimeters, and

$$Q_{\mathrm{TE}_{02}{}^{\circ}} = 21{,}400. \tag{154}$$

This theoretical intrinsic Q is very large compared to the intrinsic Q obtainable in a parallelepiped-shaped cavity. For comparison we calculate the intrinsic Q of a half-wavelength cavity at 5.4 millimeters in the standard RG98U waveguide (0.074 × 0.0148 inches) that we assume to be made out of copper. Using the expression for intrinsic Q given on Ref. 12, p. 55,

$$Q_{\mathrm{TE}_{10}\square} = 3460. \tag{155}$$

For $TE_{20}^{\square}$ mode, assuming $a \cong \lambda_0 = 1.2$ inches and $d = 0.4$ inches, expression (153) yields

$$Q_{TE_{20}^{\square}} = 9000. \tag{156}$$

The intrinsic Q of a half-wavelength resonator in RG52U waveguide (0.4×0.9 inch) that we assume made of copper is

$$Q_{TE_{10}^{\square}} = 7990. \tag{157}$$

X. EXPERIMENTAL RESULTS FOR CHANNEL-DROPPING FILTER FROM TE_{01}° TO $TE_{10}^{\square}$

We shall go first through the detailed design procedure of a channel-dropping filter for which the bandwidth is relatively large in order to show the limiting possibilities of these mode conversion filters; the experimental results are quoted later.

The selected center frequency and bandwidth of the dropped channel are 55.5 kmc, ($\lambda_0 = 5.4$ millimeters) and 500 mc. The loaded Q is therefore

$$Q_L = 110. \tag{158}$$

To dimension the filter roughly, we use (105) through (110). We shall use primes to distinguish the approximate sizes from those that are final. From (110), adopting $s = 0$,

$$\delta' \cong 0.1$$

and from (105) through (108), adopting $p = 3$, we find

$$\begin{aligned} a' &= 0.249 \text{ inch}, \\ b' &= 0.226 \text{ inch}, \\ l_B' &= l_R' = 0.168 \text{ inch}, \\ l' &= 0.204 \text{ inch}. \end{aligned}$$

If the branching rectangular waveguide is RG98U, $W = 0.148$ inch, $d = 0.074$ inch and the diameter of the coupling hole to the branching arm results, from (109)

$$2c = 0.105 \text{ inch}.$$

Since this value is bigger than the 0.074-inch height of the rectangular guide, a round coupling hole can not provide enough coupling. There are many ways to increase the coupling. One is to build the rectangular waveguide with its axis not perpendicular to the axis of the round waveguide but parallel to it, providing the coupling through a hole in the narrow wall. One of the ends of the rectangular waveguide must be short-circuited at an odd number of quarters of guide wavelength from the center of the coupling hole. For a fixed size of the hole, the amount of coupling can be increased by decreasing the width of the waveguide W, because the waveguide gets closer to cutoff. Another solution consists in wrapping around the TE_{02}° cavity a rectangular waveguide and providing the necessary mode selective coupling between them by means of several holes. The details are given in Ref. 2.

The obvious third solution, and the one we adopt, is to increase the size of the coupling hole to the total cross section of the rectangular waveguide. If the coupling is too large, it can be decreased by displacing the hole to one side of the cavity.

The strong perturbation of the field in the branching cavity due to such a large coupling hole modifies the scattering coefficients calculated in previous paragraphs, and the final length of this cavity, as well as the distance to the rejecting one, must be selected experimentally. The discrepancy between theoretical and experimental values is not large.

10.1 *Design of the Rejecting Cavity*

The design of the rejecting cavity requires the simultaneous solution of (80) and (83) for the determination of the three quantities a, b and l_R. Thus, one of those quantities can be selected arbitrarily.

A good criterion for this selection consists in demanding that at midband frequency the cutoff radius for TE_{02}° is

$$\frac{v_2\lambda_0}{2\pi} = \frac{a+b}{2} = a\left(1 - \frac{\delta}{2}\right) \tag{159}$$

because with this selection midband is equally separated from the two extreme frequencies that limit the proper operation of the filter. These are a lower frequency that cuts off the TE_{02}° in the large waveguide, invalidating the inequality (67), and an upper one that cuts off TE_{02}° in the smaller waveguide and above which propagation of TE_{02}° in that waveguide starts.

Incidentally, it is interesting to notice that, for the frequency $f = f_0 + \Delta f$,

$$\left(\frac{\lambda_0'}{\lambda_{g2}'}\right)^2 = \left\{1 - \left[\frac{v_2\lambda_0}{2\pi a\left(1 + \dfrac{\Delta f}{f_0}\right)}\right]^2\right\}\left(1 + \frac{2\Delta f}{f_0}\right) \tag{160}$$

becomes, through the use of (159),

$$\left(\frac{\lambda_0'}{\lambda_{g2}'}\right)^2 = 2\left(\delta + \frac{\Delta f}{f_0}\right), \tag{161}$$

and the inequality (67) can be written

$$\frac{\Delta f}{f_0} \ll \delta. \tag{162}$$

This implies that the approximations hold as long as the relative frequency departure from midband is small compared to the relative diameter change.

Another criterion for the selection of a, b or l_R may arise from the advantage of using standard-size waveguides already available, as long as the limiting frequencies discussed previously are not approached. Following this procedure for a standard laboratory waveguide with

$$2a = \tfrac{7}{16} \text{ inch}, \tag{163}$$

the simultaneous solution of (80) and (83) yields

$$2b = 0.5 \text{ inch and} \tag{164}$$

$$l_R = 0.234 \text{ inch}. \tag{165}$$

The measured performance of this band-rejection filter is shown in Fig. 13. The agreement between theoretical and experimental values is excellent.

10.2 *Design of the Branching Cavity*

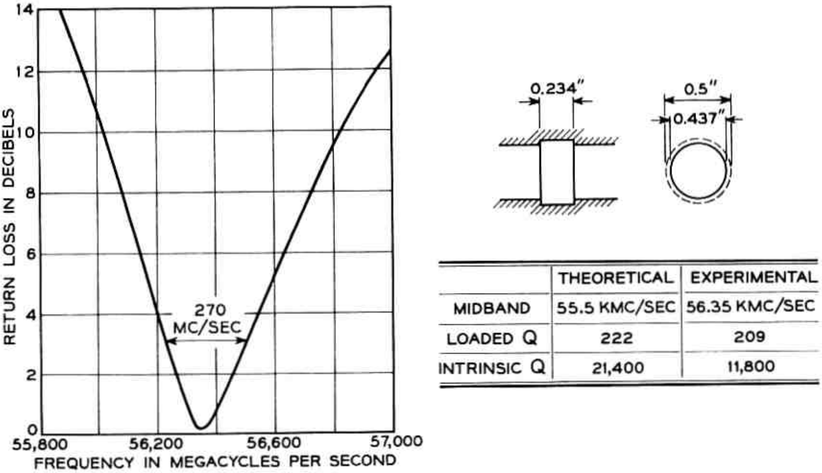
Ignoring the effect of the coupling hole between the branching waveguide and cavity, the dimensions should be those of the rejecting cavity given in (163), (164) and (165).

The distance between centers of the resonating cavities, according to (82) with $p = 4$, should be

$$l_d = 0.572 \text{ inch}.$$

The number of quarters of wavelength between centers of cavities is nine. Experimentally it was found impossible to reduce l_d because the TE_{02}° mode, being close to cutoff in the small waveguide, couples to the other cavity. The final dimensions, as well as the performance of the assembled channel-dropping filter, are shown in Fig. 14.

The relatively high insertion loss for the dropped channel cannot be

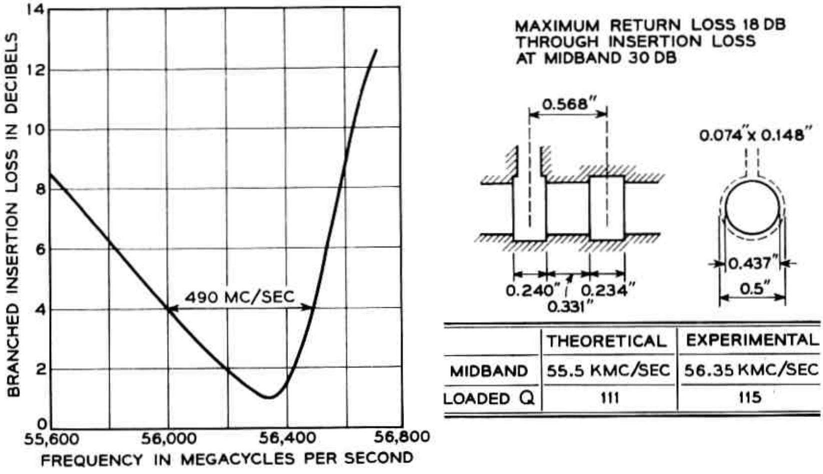


	THEORETICAL	EXPERIMENTAL
MIDBAND	55.5 KMC/SEC	56.35 KMC/SEC
LOADED Q	222	209
INTRINSIC Q	21,400	11,800

Fig. 13 — Performance of circular-electric band-rejection filter.

accounted for by heat losses because of the good performance of the band-rejection filter. Thus, mode conversion due to the asymmetry of the coupling to the rectangular waveguide must be its cause. Loss should be reduced using distributed coupling to the rectangular waveguide, as in Ref. 2.

Pictures of the filter are shown in Figs. 15 and 16.



MAXIMUM RETURN LOSS 18 DB
THROUGH INSERTION LOSS
AT MIDBAND 30 DB

	THEORETICAL	EXPERIMENTAL
MIDBAND	55.5 KMC/SEC	56.35 KMC/SEC
LOADED Q	111	115

Fig. 14 — Performance of mode-conversion channel-dropping filter.

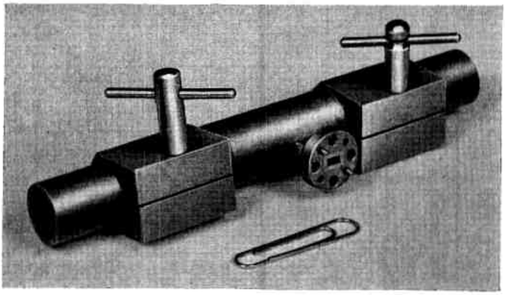


Fig. 15 — Mode-conversion channel-dropping filter.

10.3 *Band Rejection Filters in Different Waveguides*

Figs. 17, 18 and 19 show the geometry and experimental results for different band-rejection filters in round and rectangular waveguide. Those filters that have constant metallic cross section have generation and resonance of a higher-order mode in the region where the dielectric is located. For the case of Fig. 19, it has been shown in Ref. 11 that a rectangular waveguide with a dielectric slab is equivalent to a rectangular waveguide that has a width increase for a length equal to that of the slab. Calling d the width of the slab, ζ the distance to the near narrow wall and ϵ_d the permittivity of the dielectric, the relative apparent in-

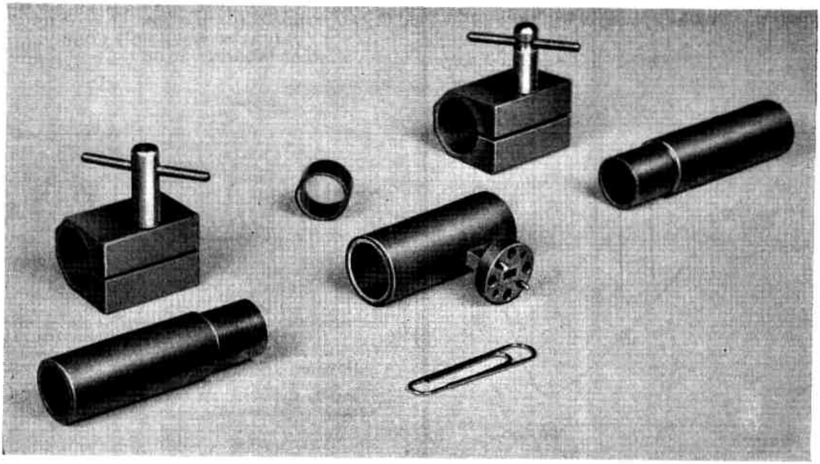
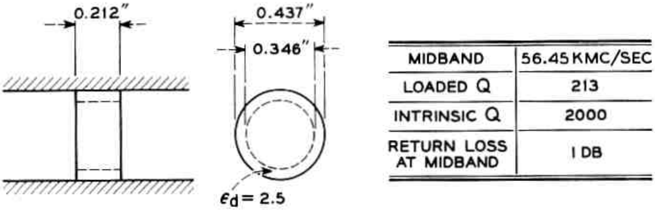


Fig. 16 — Exploded view of mode-conversion channel-dropping filter.



MIDBAND	56.45 KMC/SEC
LOADED Q	213
INTRINSIC Q	2000
RETURN LOSS AT MIDBAND	1 DB

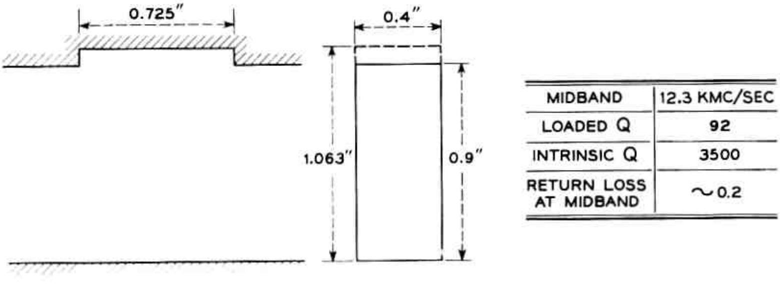
Fig. 17 — Band-rejection filter of TE_{01}° mode (polystyrene ring; $\epsilon_d = 2.5$).

crease of the waveguide width is

$$\delta = \frac{4\pi^2}{3}\frac{d^3}{a\lambda_0^2}\left(\frac{\epsilon_d}{\epsilon} - 1\right)\left[1 + \frac{3\zeta(\zeta + d)}{d^2}\right].$$

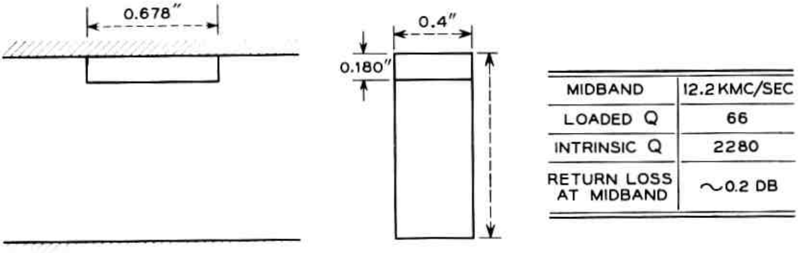
With this value known, all the design formulas in Section VII can be used.

In the round waveguide in which only circular-electric modes are of interest, tuning is available by changing the physical length of the resonating cavity. For that purpose a telescopic type of junction is ideal,



MIDBAND	12.3 KMC/SEC
LOADED Q	92
INTRINSIC Q	3500
RETURN LOSS AT MIDBAND	~0.2

Fig. 18 — Band-rejection filter of $TE_{10}^{\square}$ mode.



MIDBAND	12.2 KMC/SEC
LOADED Q	66
INTRINSIC Q	2280
RETURN LOSS AT MIDBAND	~0.2 DB

Fig. 19 — Band-rejection filter of $TE_{10}^{\square}$ mode (polystyrene ring; $\epsilon_d = 2.5$).

since the cracks do not interrupt the conduction current and since it is very easy to manufacture. One tube of inner diameter $2a$, inside of which two tubes of outer diameter $2a$ and inner diameter $2b$ can slide, will suffice.

The tuning in the case of Fig. 17 can be achieved by trimming the dielectric. For the filters in rectangular waveguide, Figs. 18 and 19, one tuning screw at each one of the electric-field maxima of the resonating mode provide the tuning.

XI. CONCLUSIONS

Resonance of higher-order modes in waveguides has been advantageously used to make very simple band-rejection filters of unusually low loss. In particular, the filter operating with circular-electric modes has an intrinsic Q that is one order of magnitude better than the intrinsic Q of conventional (cavity or lumped) band-rejection filters operating at the same frequency.

Mode-conversion band-rejection filters have been used as building blocks for the construction of channel-dropping filters that simultaneously produce the band separation and the transfer from TE_{01}° mode to $\text{TE}_{10}^{\square}$ required in the long distance waveguide communication system.[1] One model operating at 56.3 kmc has a bandwidth of 490 mc, and the insertion loss for the dropped channel is 1 db.

Although the emphasis in this paper has been on filters operating mainly with circular-electric modes in round waveguides and TE modes in rectangular waveguides, the calculations are quite general and can be applied in any scheme in which mode-conversion filters are used.

XII. ACKNOWLEDGMENT

The author is indebted to D. L. Bisbee for performing the measurements.

XIII. LIST OF SYMBOLS

a = radius of the resonant cylindrical cavity or width of the resonant rectangular waveguide.

b = radius of the through cylindrical waveguide or width of the through rectangular waveguide.

c = radius of the coupling hole between resonant cavity and branching arm.

d = height of any rectangular waveguide.

f = frequency.

f_0 = midband frequency.

l = distance between cavities.

l_B = length of branching cavity.

l_d = distance between centers of cavities.

l_R = length of rejecting cavity.

$2p + 1$ = number of nonresonant mode quarter-wavelengths between centers of branching and rejecting cavities (p is an arbitrary integer).

$2s + 1$ = number of resonant mode half wavelengths in each cavity (s is an arbitrary integer).

Q = intrinsic Q.

Q_L = loaded Q.

S_{mn} = scattering coefficient of a half cavity or a more complicated circuit.

W = width of the branching rectangular waveguide.

Y = admittance.

Γ_{mn} = scattering coefficient of elementary structures.

δ = relative diameter change or width change of through waveguide.

ϵ = permittivity of free space.

ϵ_d = permittivity of dielectric.

θ = electrical difference between two energy paths.

θ_0 = midband electrical difference between two energy paths.

θ_{mn} = phase of the scattering coefficients of junctions with wave guides reduced to zero length.

λ = midband free-space wavelength.

λ_{g1} = midband guided wavelength of the nonresonant mode.

λ_{g2} = midband guided wavelength of the resonant mode.

μ = permeability.

ν_n = nth root of the J_1 function.

σ = conductivity of metal.

φ_0 = midband electrical length of the resonating cavity in terms of the resonating mode.

ψ = midband electrical distance between cavities in terms of the through mode.

ψ_d = midband electrical distance between centers of cavities in terms of the through mode.

ψ_1 = electrical distance between the branching cavity discontinuities in terms of the non-resonating mode.

ψ_{10} = midband electrical distance between the branching cavity discontinuities in terms of the non-resonating mode.

ψ_2 = electrical distance between the branching cavity discontinuities in terms of the resonating mode.

ψ_{20} = midband electrical distance between the branching cavity discontinuities in terms of the resonating mode.

$\bar{\psi}_1$ = electrical distance between the rejecting cavity discontinuities in terms of the nonresonating mode.

$\bar{\psi}_{10}$ = midband electrical distance between the rejecting cavity discontinuities in terms of the nonresonating mode.

$\bar{\psi}_2$ = electrical distance between the rejecting cavity discontinuities in terms of the resonating mode.

$\bar{\psi}_{20}$ = midband electrical distance between the rejecting cavity discontinuities in terms of the resonating mode.

REFERENCES

1. Miller, S. E., Waveguide as a Communication Medium, B.S.T.J., **33,** 1954, p. 1209.
2. Marcatili, E. A., Channel-Dropping Filter in the Millimeter Region Using Circular-Electric Modes, to be published.
3. Marcatili, E. A., A Circular-Electric Hybrid Junction and Some Channel-Dropping Filters, this issue, p. 185.
4. King, A. P. and Marcatili, E. A., Transmission Loss Due to Resonance of Loosely Coupled Modes in Multi-Mode Systems, B.S.T.J., **35,** 1956, p. 899.
5. Marcatili, E. A., Mode Conversion Filters, I.R.E. Wescon Conv. Rec., 1958, part 1.
6. Jaynes, E. T., Ghost Modes in Imperfect Waveguides, Proc. I.R.E., **46,** 1958, p. 416.
7. Vogelman, J. H. Microwave Rejection Filters Using Higher-Order Modes, I.R.E. Trans., **MTT-7,** 1959, p. 461.
8. Goubau, G., *Electromagnetische Wellenleiter und Hohlraume*, Wissenschaftliche Verlagsgesellschaft, Stuttgart, Germany, 1955.
9. Marcuvitz, N., ed., *Waveguide Handbook*, M.I.T. Radiation Laboratory Series, Vol. 10, 1st ed., McGraw-Hill, New York, 1951, p. 108.
10. Bethe, H. A., Theory of Diffraction by Small Holes, Phys. Rev., **66,** 1944, p. 163.
11. Marcatili, E. A., Scattering at a Sudden Change of Cross Section in Multimode Waveguide, to be published.
12. Montgomery, C. G., Dicke, R. H. and Purcell, E. M., eds., M.I.T. Radiation Laboratory Series, Vol. 8, 1st ed., McGraw-Hill, New York, 1948.
13. McLachlan, N. W., *Bessel Functions for Engineers*, Clarendon Press, Oxford, 1934.

A Circular-Electric Hybrid Junction and Some Channel-Dropping Filters

By E. A. MARCATILI

(Manuscript received July 11, 1960)

A $TE_{01}{}^{\circ}$ hybrid junction that operates similarly to the Riblet short-slot hybrid is described, but because the modes involved are circular-electric, the hybrid can be telescopically mounted, allowing for adjustment to almost any power division. The experimental results show that, centered at 55.6 kmc, the frequency range is larger than 20 per cent. Adjusted for equal power division, the balance is better than 0.5 db and the unwanted reflections in the driven and balanced (isolation) arms are at least 23 db below the input signal.

Using the hybrid together with band-reflection, band-transmission or high-pass filters, it is possible to build low-loss channel-dropping filters. In particular, the use of simple cutoff waveguides permits the design of filters with almost rectangular transfer characteristics.

I. INTRODUCTION

The importance of hybrid junctions for many purposes — measuring, filtering, balancing, equalizing, etc. — need hardly be emphasized. The long distance waveguide communication system[1] operating with the low-loss circular-electric $TE_{01}{}^{\circ}$ mode has only two hybrids available: the directional coupler, which has a fixed power division, and the optical hybrid,[2] which requires multimode waveguides. This paper describes a third hybrid, which operates like Riblet's coupler[3] and which adds to the well-known advantages of that coupler the unique property of adjustable power division.

Adjusted for 3-db power division, the hybrid, together with mode-conversion band-rejection filters,[4] band transmission filters or cutoff waveguides,[2] can be used as low-loss components of constant-resistance channel-dropping filters.[5] The scheme that uses high-pass filters (cutoff waveguides) deserves special attention because the amplitude transfer characteristic of the dropped channel can be made to approach a

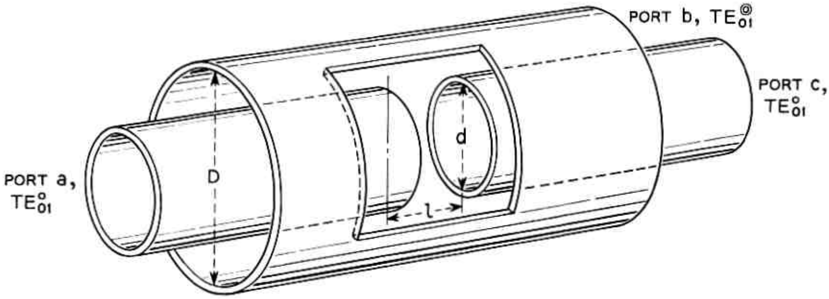

PORT b, $TE_{01}^{\circledcirc}$
PORT c, TE_{01}°
PORT a, TE_{01}°
D
d
l

Fig. 1 — Circular-electric hybrid.

rectangular shape of arbitrary bandwidth. This permits not only relaxing the demands on the guard bands between neighboring channels, but also the multiplexing of bands too broad (extremely short pulses)[6] to be handled by mode-conversion filters.

II. DESCRIPTION OF THE HYBRID

The hybrid consists of two coaxial circular metallic tubes, of which the inner one has a gap l, as shown in Fig. 1. The ratio of diameters selected is equal to the ratio of the second to the first roots of the Bessel function J_1:

$$\frac{D}{d} = \frac{7.016}{3.832} = 1.831. \tag{1}$$

The outer diameter D is chosen so that it cuts off the $TE_{03}{}^{\circ}$ mode at the highest frequency of design of the hybrid.

The hybrid is made of two four-port junctions like the one of Fig. 2. It will be shown that power entering in any port is almost equally divided between the two forward modes. Consequently, going back to

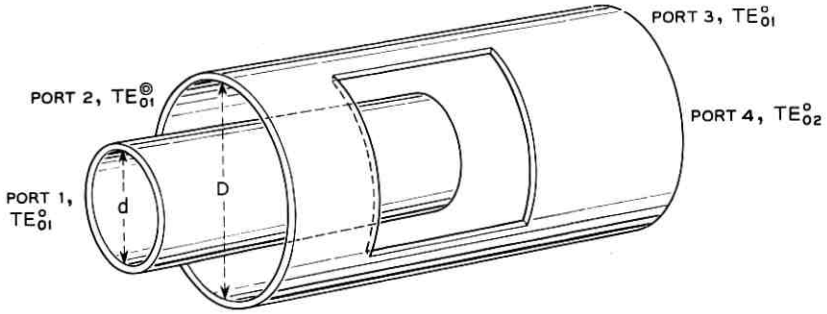

PORT 3, TE_{01}°
PORT 2, $TE_{01}^{\circledcirc}$
PORT 4, TE_{02}°
PORT 1, TE_{01}°
d
D

Fig. 2 — Four-port junction.

Fig. 1, power entering in any port is almost equally divided between $TE_{01}°$ and $TE_{02}°$ in the gap region. Each one of these modes repeats the power division at the end of the gap, so the power collected in each output depends on the relative phases of the modes at the end of the gap. Since the velocities of these modes are different, the relative phase, and consequently the power division, can be selected arbitrarily by changing the length of the gap.

III. PROPERTIES OF THE FOUR-PORT JUNCTION AND THE HYBRID

The most general scattering matrix for the reciprocal four-port device of Fig. 2 is

$$S = \begin{vmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{12} & S_{22} & S_{23} & S_{24} \\ S_{13} & S_{23} & S_{33} & S_{34} \\ S_{14} & S_{24} & S_{34} & S_{44} \end{vmatrix}, \tag{2}$$

Entering port 4 with mode $TE_{02}°$, since the metallic inner tube has its surface where the electric field is zero (first zero of the J_1 function), the boundary conditions are automatically satisfied, the $TE_{02}°$ mode is unperturbed, and consequently the back-scattering

$$S_{34} = S_{44} = 0. \tag{3}$$

Furthermore, the forward-scattering coefficients at the plane where the coaxial waveguide starts are

$$S_{14} = \left| \frac{\int_0^{3.832} J_1^2(\alpha)\alpha d\alpha}{\int_0^{7.016} J_1^2(\alpha)\alpha d\alpha} \right|^{\frac{1}{2}} = 0.733, \tag{4}$$

$$S_{24} = - \left| \frac{\int_{3.832}^{7.016} J_1^2(\alpha)\alpha d\alpha}{\int_0^{7.016} J_1^2(\alpha)\alpha d\alpha} \right|^{\frac{1}{2}} = -0.68. \tag{5}$$

Assuming the junction to be nondissipative, (2) must satisfy, because of conservation of energy, the following unitary relations[7]

$$\sum_{\beta=1}^{4} S_{\beta m} S_{\beta n}{}^* = \begin{cases} 1 & \text{if} \quad m = n \\ 0 & \text{if} \quad m \neq n \end{cases} \tag{6}$$

in which the asterisk means "complex conjugate of."

From (3) and (6),

$$S_{12} = -\frac{S_{14}^*}{S_{24}^*} S_{11}, \tag{7}$$

$$S_{22} = \left(\frac{S_{14}^*}{S_{24}^*}\right)^2 S_{11}, \tag{8}$$

$$S_{13} = |S_{24}| \left(1 - \left|\frac{S_{11}}{S_{24}^2}\right|^2\right)^{\frac{1}{2}} e^{i\theta_{13}}, \tag{9}$$

$$S_{23} = -\frac{S_{14}^*}{S_{24}^*} |S_{24}| \left(1 - \left|\frac{S_{11}}{S_{24}^2}\right|^2\right)^{\frac{1}{2}} e^{i\theta_{13}}, \tag{10}$$

$$S_{33} = -\frac{S_{11}^*}{|S_{24}|^2} e^{i2\theta_{13}}, \tag{11}$$

where θ_{13} is the phase of S_{13}.

Since S_{14} and S_{24} are known from (4) and (5), the five previous expressions become

$$S_{12} = 1.078\, S_{11}, \tag{12}$$

$$S_{22} = 1.163\, S_{11}, \tag{13}$$

$$S_{13} = 0.68(1 - 4.68\, |S_{11}|^2)^{\frac{1}{2}} e^{i\theta_{13}}, \tag{14}$$

$$S_{23} = 0.733(1 - 4.68\, |S_{11}|^2)^{\frac{1}{2}} e^{i\theta_{13}}, \tag{15}$$

$$S_{33} = -2.163\, S_{11}^* e^{i2\theta_{13}}. \tag{16}$$

In the experimental hybrid to be described later on, the modulus of the reflection coefficient is

$$|S_{11}| < 0.05$$

and consequently powers of S_{11} bigger than one can be neglected. With this simplification, the forward transfer elements of the scattering matrix of the hybrid (Fig. 1) are

$$S_{ac} = S_{14}^2 e^{i2\pi l/\lambda_{g2}} \left[1 + \left|\frac{S_{13}}{S_{14}}\right|^2 e^{i2\theta_{13} + i2\pi l/\Lambda}\right], \tag{17}$$

$$S_{ab} = S_{14} S_{24} e^{i2\pi l/\lambda_{g2}} [1 - e^{i2\theta_{13} + i2\pi l/\Lambda}], \tag{18}$$

where

$$\Lambda = \frac{\lambda_{g1}\lambda_{g2}}{\lambda_{g2} - \lambda_{g1}} \tag{19}$$

is the beating wavelength between TE$_{01}$° and TE$_{02}$° in the gap;

$$\lambda_{g1} = \frac{\lambda}{\sqrt{1 - \left(\frac{3.832\lambda}{\pi D}\right)^2}} \tag{20}$$

and

$$\lambda_{g2} = \frac{\lambda}{\sqrt{1 - \left(\frac{7.016\lambda}{\pi D}\right)^2}} \tag{21}$$

are the TE$_{01}$° and TE$_{02}$° guided wavelengths; and λ is the free-space wavelength.

For a given gap l, the power division of the hybrid K, and the phase shift between the two outputs are derived from (4), (5), (14), (15), (17) and (18):

$$K = \left|\frac{S_{ac}}{S_{ab}}\right|^2 = \frac{1.011 + \cos\left(2\theta_{13} + \frac{2\pi l}{\Lambda}\right)}{1 - \cos\left(2\theta_{13} + \frac{2\pi l}{\Lambda}\right)}, \tag{22}$$

$$\theta_{ac} - \theta_{ab} = \pm\pi + tg^{-1} 13.15\sqrt{K - 0.0055}. \tag{23}$$

The possible range of power division K obtained from (22) is

$$0.0055 \leqq K < \infty. \tag{24}$$

For $K = 0.0055$, the power flowing in the inner guide is a minimum and specifically 26 db below the input. For $K = \infty$, the power flowing in the coaxial guide is zero.

Since the beating wavelength Λ, as well as the argument θ_{13}, are frequency-sensitive, the power division K given in (22) also varies with frequency. We have not calculated θ_{13}, but it is known[3] that the frequency dependence of θ_{13} and of Λ tend to cancel each other's effect, allowing the power division K to be constant over a relatively broad band. Furthermore, it is very easy to adjust experimentally the gap l for any allowable power division K because the hybrid can be built with sliding tubes. The modes involved are circular electric and consequently the cracks do not interrupt conduction current lines.

IV. EXPERIMENTAL RESULTS

In order to make available the power from the hybrid a four-port transducer has been electroformed capable of transferring TE$_{01}$° to

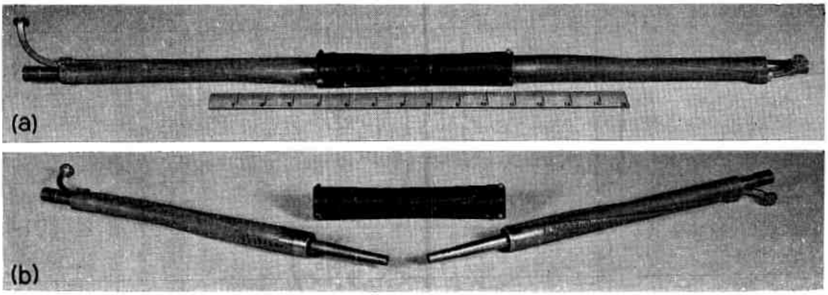


Fig. 3 — (a) Circular-electric hybrid assembled with $\mathrm{TE}_{01}^{\circledcirc}$ to $\mathrm{TE}_{10}^{\square}$ transducers; (b) exploded view.

TE_{01}° and $\mathrm{TE}_{01}^{\circledcirc}$ to $\mathrm{TE}_{10}^{\square}$ (Fig. 3). The last change of modes is obtained by smoothly deforming a rectangular waveguide into a coaxial waveguide. The transducer generates small amounts of unwanted higher order modes, which can resonate[8] and ruin the behavior of the hybrid. The resonances can be damped by using for the external tube of the hybrid a lossy-jacket helix waveguide,[9] which substantially attenuates any mode with axial conduction currents.

Fig. 4 shows the electrical behavior of the hybrid adjusted for equal power division. From 50 to 61.2 kmc the balance is better than 0.5 db and the isolation better than 23 db.

At 55.6 kmc the power lost in the hybrid and transducer is 0.83 db. In order to prove that most of this loss occurs in the $\mathrm{TE}_{01}^{\circledcirc}$ to $\mathrm{TE}_{10}^{\square}$ transducer, the gap was enlarged until, at 55.5 kmc, most of the power was recovered in the inner waveguide ($K = \infty$; $l = 0.906$ inch). The measured insertion loss was then reduced to 0.3 db.

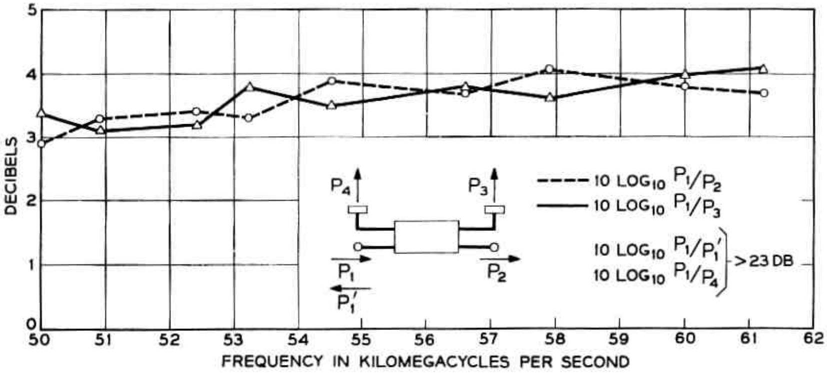


Fig. 4 — Performance of circular-electric hybrid and $\mathrm{TE}_{01}^{\circledcirc}$ to $\mathrm{TE}_{10}^{\square}$ transducers.

No efforts have been made to improve either the hybrid or the transducers. The possible changes for the hybrid are of an experimental nature and consist in varying the diameter of the gap region and including circular symmetric lumped discontinuities to improve the balance and decrease the unwanted reflections. The possible improvement of the transducer consists in passing from the relatively simple-to-build linear taper used in these experiments to more sophisticated designs[10] that reduce mode conversion.

V. CONSTANT-RESISTANCE CHANNEL-DROPPING FILTERS

It is known that a constant-resistance channel-dropping filter[5] (input matched at all frequencies) can be made using two hybrids connected by two filtering paths. The hybrid described in Section III lends itself to use with filters that operate with low-loss circular-electric modes, and is consequently attractive for use in the long distance waveguide communication system.

The filters that most naturally suit the hybrid are those that possess circular symmetry. For example, filters made with inductive irises, mode-conversion filters[4] and cutoff filters. In Fig. 5, two such filters with identical transfer characteristics are located symbolically in the inner and outer waveguides connecting two circular-electric hybrids. $\mathrm{TE}_{01}{}^{\circ}$ power that enters port 1, and is rejected by the filters, recombines as $\mathrm{TE}_{01}{}^{\circledcirc}$ in port 2. The power transmitted through the filters can be made to recombine either in port 3 or in port 4. On one hand, assuming the gaps of both hybrids to be identical, power recombines in port 3 if the inner and outer electrical paths between planes a and b are identical, and power recombines in port 4 if those paths differ by π radians. On the other hand, assuming the two paths to be identical,

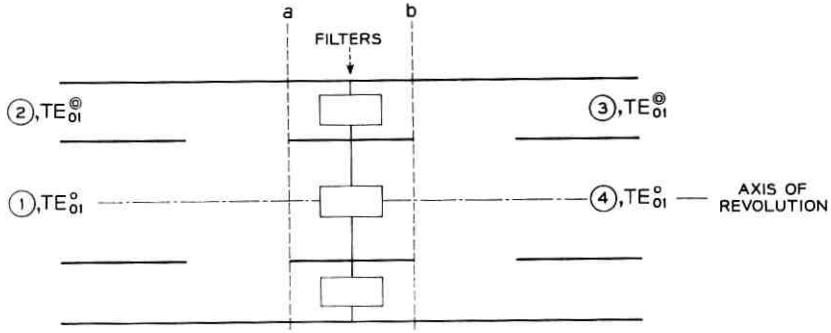

a
b
FILTERS
②,$\mathrm{TE}_{01}^{\circledcirc}$
③,$\mathrm{TE}_{01}^{\circledcirc}$
①,TE_{01}°
④,TE_{01}°
AXIS OF REVOLUTION

Fig. 5 — Channel-dropping filter using circular-electric hybrids

power recombines in port 3 if the hybrids are identical and recombines in port 4 if both gaps differ by half a beating wavelength between the $TE_{01}°$ and $TE_{02}°$ modes.

Probably the most interesting of the channel-dropping filters is obtained by using cutoff waveguides (high-pass filters) in the connecting paths. The interest comes from the fact that the transfer characteristic of the dropped channel can be made to approximate a rectangular shape.

Before considering the actual geometry of these filters we analyze the behavior of a chain of constant resistance filters represented symbolically in Fig. 6(a). The first link consists of two hybrids H connected by two paths of identical transfer and reflection coefficients. Each path includes a high pass filter that cuts off at frequency f_1. The only difference between the successive constant resistance filters is the cutoff frequency of the high-pass filters. Because of the phase-shifts between the different arms of the hybrids, and the similitude of the connecting paths, power entering in port 0 can be recovered only in ports 1, 2, 3 $\cdots$ $(n + 1)$. The power transfers between input and output ports are given in Fig. 6(b); $n - 1$ channels can be dropped out of n constant resistance filters.

The actual geometry of one of the units of the chain is very simple

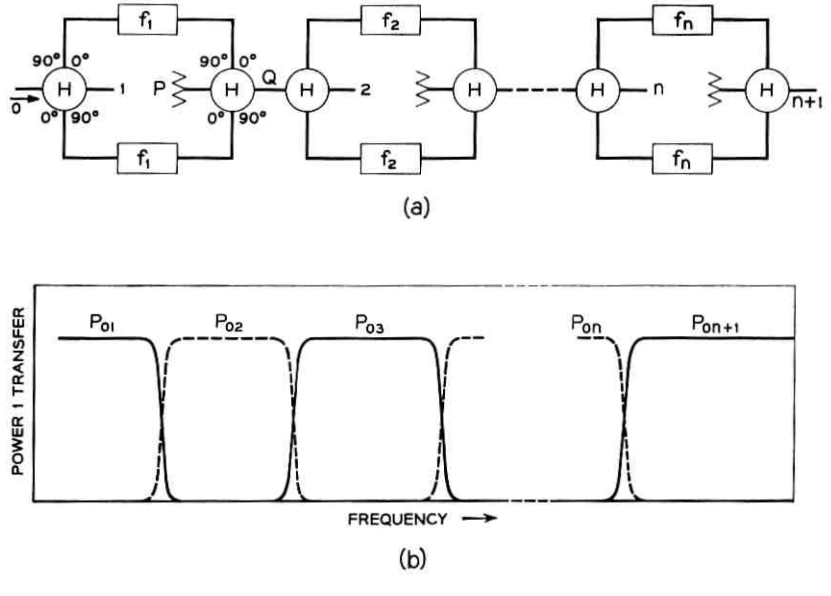


Fig. 6 — (a) Chain of constant-resistance filters; (b) power transfer between ports 0 and 1, 2, 3, $\cdots$, $(n + 1)$.

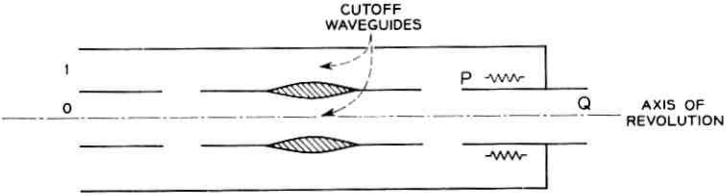


Fig. 7 — Constant-resistance filter.

when circular-electric hybrids and cutoff waveguides are used as shown in Fig. 7. The ports 0, 1, P and Q correspond to those of the first unit in Fig. 6(a). The two hybrids in Fig. 7 are different, in order to recombine the power transmitted through the cutoff sections in the inner waveguide.

Without cavities we have achieved an almost rectangular transfer characteristic of arbitrary width. The guard band between successive channels can be made, at least in principle, arbitrarily small. A working model of cutoff filters has been demonstrated in Ref. 2.

There is another channel-dropping filter worth considering because of its simplicity and because it uses the structure shown in Fig. 2 as a hybrid. [The reader can check that the scattering coefficients of this junction given in (4), (5), (12), (13), (14), (15) and (16) are very close to those of a hybrid when S_{11} is negligibly small.]

Before considering the actual channel-dropping filter we shall describe a microwave equivalent circuit, Fig. 8. It consists of two hybrids, indicated as Riblet couplers, which are connected by two waveguides of equal electrical lengths. These waveguides are also coupled through two identical resonating cavities. The electrical distance between coupling holes in the upper waveguide is an odd multiple of π and in the lower waveguide is an even multiple of π.

Out of resonant frequency of the cavities power entering port 1 splits in equal parts in the first hybrid and recombines in port 5 of the second

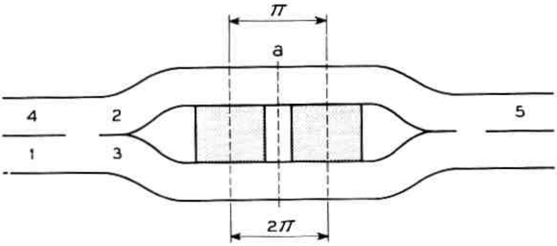
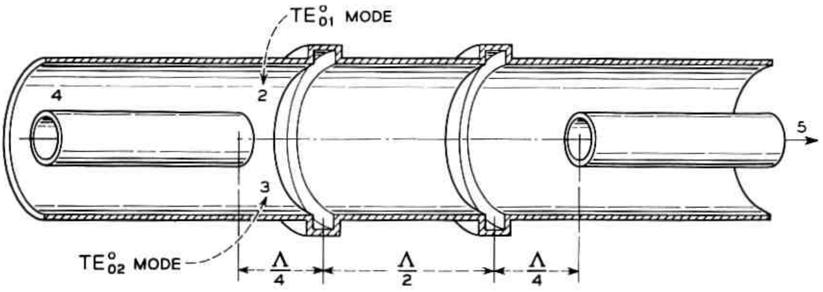


Fig. 8 — Microwave equivalent circuit of channel-dropping filter of Fig. 9.

hybrid. At resonance, power entering port 1 splits in equal parts and each one excites the cavities in different ways. Let us follow the power in the upper path. Because of the distance between coupling holes, the cavities are excited in opposite phase and the reradiation from the cavities is such that all the power flows back toward port 2 as if reflected from an equivalent short circuit located in plane of symmetry *a*. Meanwhile, the power flowing in the lower path from port 3 excites the cavities in phase and again, because of the adequate distance between holes, all the power goes back toward port 3 as if reflected by a short circuit in plane *a*. Recombination of the two waves reflected in plane *a* takes place in port 4 of the first hybrid.

The actual microwave circuit for circular electric waves is shown in Fig. 9. The two hybrids are like those of Fig. 2. Waves flowing in ports 2 and 3 of Fig. 8 are equivalent to the TE_{01}° and TE_{02}° waves in the gap of Fig. 9. The length of the gap region is one beating wavelength between the TE_{01}° and the TE_{02}° modes; the diameter is selected in such a way that the TE_{03}° is cut off except for two enlarged regions where resonance of this mode takes place.[4] These mode-conversion resonant "cavities" couple to both TE_{01}° and TE_{02}° modes and are separated by half a guided wavelength measured in TE_{02}° mode and one guided wavelength measured in TE_{01}° mode. The mode conversion "cavities" are therefore equivalent to the resonant cavities of Fig. 8.

If the coupling between TE_{03}° and TE_{01}° is different from the coupling between TE_{03}° and TE_{02}°, the channel-dropping filter no longer has constant resistance. This can be deduced from Fig. 8 by making the coupling holes in the upper waveguide different from those in the lower one.



$$\Lambda = \frac{\lambda_{g_1}\lambda_{g_2}}{\lambda_{g_2}-\lambda_{g_1}} = \text{BEATING WAVELENGTH BETWEEN } \text{TE}_{01}^\circ \text{ AND } \text{TE}_{02}^\circ \text{ WAVES}$$

Fig. 9 — Channel-dropping filter with TE_{03}° mode-conversion filter.

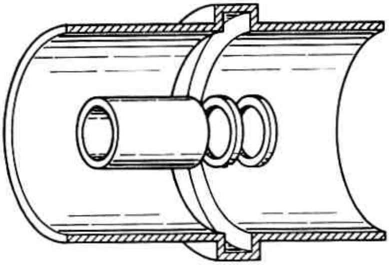


Fig. 10 — Rings to equalize coupling between TE_{03}° and TE_{01}° and between TE_{03}° and TE_{02}°.

To equalize the couplings in Fig. 9, rings like those shown in Fig. 10 can be used.

In all the filters described in this section, the dropped channel appears as $TE_{01}^{\circledcirc}$ mode. It may be necessary to transduce this mode into $TE_{10}^{\square}$. There are essentially two techniques. One consists in using a broadband transducer like the one described in Fig. 3 of Section IV; another consists in using a transmission cavity that resonates with coaxial circular electric mode and that couples to the coaxial waveguide and to a rectangular waveguide.[11] The second approach yields a much shorter transducer but it is not broadband.

VI. CONCLUSIONS

A hybrid capable of dividing TE_{01}° mode into TE_{01}° and $TE_{01}^{\circledcirc}$ has been described. It operates similarly to the Riblet short-slot hybrid, but because the modes involved are circular electric, the hybrid can be made of sliding coaxial tubes that allow adjustment to almost any power division.

The experimental results show that, centered at 55.6 kmc, the frequency range is larger than 20 per cent. Adjusted for 3 db division with the transducers from $TE_{01}^{\circledcirc}$ to $TE_{10}^{\square}$ included, the balance is better than 0.5 db and the unwanted reflections in the driven and balanced (isolation) arms are at least 23 db below the input signal.

No efforts have been made to improve either the hybrid or the transducers. The possible changes for the hybrid are of an experimental nature and consist in varying the diameter of the gap region and including circular symmetric lumped discontinuities to improve the balance and decrease the unwanted reflections. The possible improvement of the transducer consists in passing from the relatively simple-to-build linear taper used in these experiments to more sophisticated designs[10] that reduce mode conversion.

Using the hybrid together with band-reflection, band-transmission or cutoff waveguides, it is possible to build low-loss constant-resistance channel-dropping filters. In particular, the use of cutoff waveguides permits us to design filters with almost rectangular transfer characteristics.

Hybrids and filters described in this paper operate with circular-electric modes, but their equivalents operating with TE modes in rectangular waveguides can be easily derived by the reader. The design of $\text{TE}_{10}^{\square}$ mode conversion filters is given in an accompanying paper.[4]

REFERENCES

1. Miller, S. E., Waveguide as a Communication Medium, B.S.T.J., **33,** 1954, p. 1209.
2. Marcatili, E. A., and Bisbee, D. L., Band-Splitting Filter, this issue, p. 197.
3. Riblet, H. J., The Short-Slot Hybrid Junction, Proc. I.R.E., **40,** 1952, p. 180.
4. Marcatili, E. A., Mode-Conversion Filters, this issue, p. 149.
5. Lewis, W. D. and Tillotson, L. C., A Nonreflecting Branching Filter for Microwaves, B.S.T.J., **27,** 1948, p. 83.
6. Dietrich, A. F. and Goodall, W. M., Solid State Generator for 2 × 10^{-10} Second Pulses, Proc. I.R.E., **48,** 1960, p. 791.
7. Marcuvitz, N., ed., *Waveguide Handbook*, M.I.T. Radiation Laboratory Series, Vol. 10, 1st ed., McGraw-Hill, New York, 1951, p. 108.
8. King, A. P. and Marcatili, E. A., Transmission Loss Due to Resonance of Loosely Coupled Modes in a Multi-Mode System, B.S.T.J., **35,** 1956, p. 899.
9. Morgan, S. P. and Young, J. A., Helix Waveguide, B.S.T.J., **35,** 1956, p. 1347.
10. Unger, H. G., Circular Waveguide Taper of Improved Design, B.S.T.J., **37,** 1958, p. 899.
11. Marcatili, E. A., Channel-Dropping Filter in the Millimeter Region Using Circular-Electric Modes, to be published.

Band-Splitting Filter

By E. A. MARCATILI and D. L. BISBEE

(Manuscript received July 11, 1960)

A constant-resistance filter capable of dividing a very wide band into two subbands is described. It can handle one octave in the millimeter region with only 1.5 db insertion loss for each subband. The splitting transition takes place in a very narrow band (160 mc). Two of its components are important devices: an elbow and a hybrid junction. Both are quasi-optical and work with $TE_{01}{}^{\circ}$ mode in 2 inch diameter waveguide.

I. INTRODUCTION

The long-distance waveguide communication system will handle an extremely broad band extending perhaps from 40 to 80 kmc.[1] For regeneration and amplification this band must be divided into channels around 400 mc apart. Promising filters capable of performing this channel separation have been described elsewhere,[2,3,4] but it is improbable that satisfactory filtering can be obtained if approximately 100 channel-dropping filters are to be stacked one after another. The main reasons for possible trouble are:

(a) Resonance of unwanted modes. This occurs because some of the filters are required to operate over a range of frequencies covering more than one octave.

(b) Multiple reflections. Although the reflection from each filter is small, the combined reflection of as many as 100 may become prohibitively large at discrete frequencies.

Troubles from these sources can be reduced by dividing the broad 40-kmc band into several subbands. The width of the subbands can be adjusted to accommodate a suitable number of channel-dropping filters.

This paper describes a filter capable of dividing a band in two parts. It can easily handle one octave in the millimeter region with low insertion loss because it operates with low loss mode $TE_{01}{}^{\circ}$ mostly in 2 inch diameter waveguide.

The splitting process can be repeated as many times as necessary by cascading similar filters.

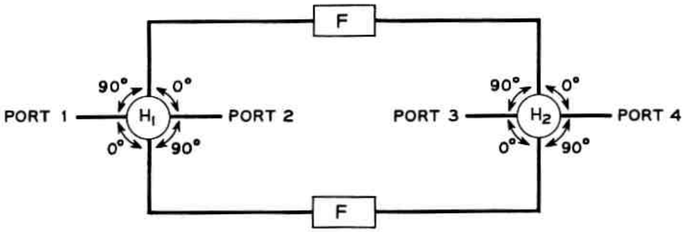


Fig. 1 — Band-splitting filter.

A band-splitting filter has been built and tested. The results are quite promising.

II. BAND-SPLITTING FILTER

The band-splitting filter is a constant-resistance filter, Fig. 1, made essentially of two identical hybrid junctions H_1 and H_2 and two identical high-pass filters F. The phase-shifts between arms of the hybrids are shown in the figure. Power entering in port 1 is divided by the hybrid H_1 into two equal parts that travel through equal electrical paths toward the high pass filters. Frequencies above that of cutoff of the filters keep on traveling and recombine with the same phase in port 4, and the opposite phase in port 3. Frequencies below that of cutoff of the filters are rejected and add in port 2 and subtract in port 1. Consequently, all the power entering in port 1 is recovered in ports 2 and 4.

What happens at frequencies where the hybrids are identical but do not divide power in equal parts? The unitary power entering in port 1 is divided by the hybrid H_1 in two parts, Γ and $1 - \Gamma$, that travel toward the filters. Due to conservation of energy the phase-shifts in the hybrids are independent of the value Γ, and the powers appearing in ports 3 and 4 (above cutoff) are

$$P_3 = 1 - 4\Gamma(1 - \Gamma), \tag{1}$$

$$P_4 = 4\Gamma(1 - \Gamma). \tag{2}$$

Power recovered in the first hybrid (below cutoff) is

$$P_1 = 1 - 4\Gamma(1 - \Gamma), \tag{3}$$

$$P_2 = 4\Gamma(1 - \Gamma). \tag{4}$$

We check immediately that if the hybrids operate ideally splitting power in halves, $\Gamma = \frac{1}{2}$ and

$$P_1 = P_3 = 0,$$

$$P_2 = P_4 = 1.$$

P_2 and P_4 given in (2) and (4) measure the recoverable power when the hybrids are not ideal. Let us plug some numbers into these expressions. For $\Gamma = 0.333 \cdots$ or $\Gamma = 0.666 \cdots$, $P_2 = P_4 = \frac{8}{9}$. In words, even at frequencies where the power division of the hybrids is as bad as two to one, the recoverable power of the band-splitting filter is as high as eight-ninths of the input power (0.5 db loss). This good behavior of the band-splitting filter, even with unequal power division in the hybrid, assures satisfactory operation over an extremely broad band.

An experimental model of a band-splitting filter operating with circular-electric mode is shown in Fig. 2, and its schematic appears in Fig. 3. It is interesting to note that we have used two elbows between the generator and the actual band-splitting filter just for "compactness."

We describe now the experimental technique used to evaluate the band-splitting filter and the results.

Most of the experimental data is presented in oscillographs that carry frequencies in abscissas and power in ordinates. The insertion loss of a device, for example, can be calculated from two oscillographs which show the transmitted powers with and without the device included in the microwave circuitry.

The wide band-sweep displayed in each oscillograph has been achieved through the use of backward wave oscillators, but we had to pay a price in that the output power of these tubes varies rapidly with frequency, and therefore the oscillographs exhibit a fine structure that makes measurement a little cumbersome.

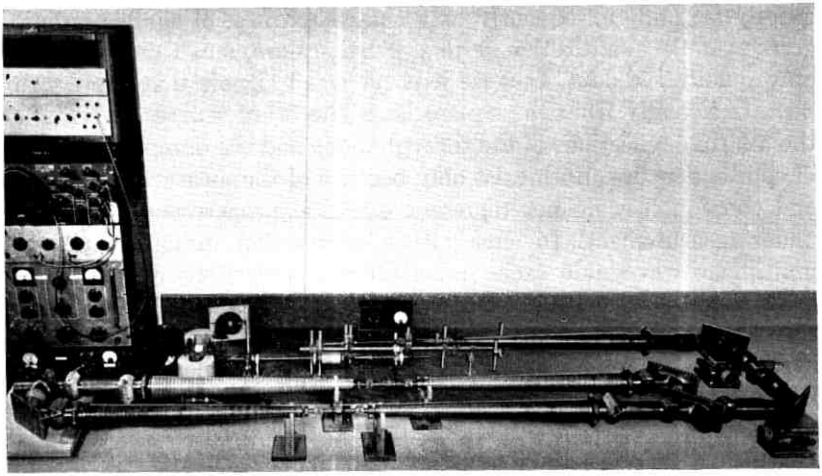


Fig. 2 — Band-splitting filter for $TE_{01}°$.

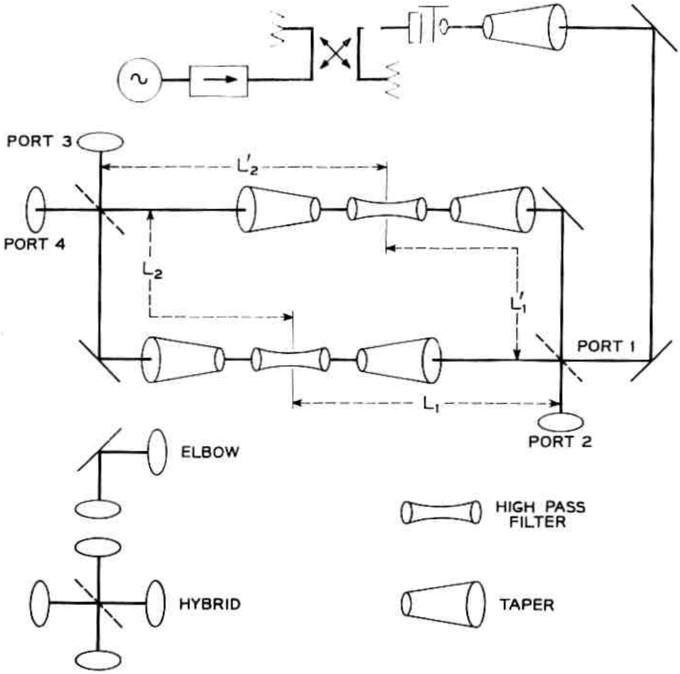


Fig. 3 — Schematic of a band-splitting filter connected to a generator.

There are several sources of errors in our measurements. One stems from the assumption that we have a square-law detector; this is not strictly so, and consequently only the comparison of similar ordinates gives reliable quantitative results. Furthermore, small ordinates yield only qualitative data, because it is for small signals that the detector departs strongly from the square law. The other sources of errors are the multiple reflections of the through mode and the damped resonances of spurious modes. Both exist only because of the measuring technique, and consequently do not represent electrical properties of the devices under measurement. In effect, it is known that discontinuities in a multimode waveguide excite practically no reflections except forward conversion. Now, in order for measurements to be made, the multimode waveguide must be connected to a generator and receiver that operate in single mode rectangular waveguides. It is in the connecting transducers where most of the reflections of the through mode take place and also where the converted modes are cut off and reflected.

Multiple reflections and damped resonances show their presence in the oscillographs as very fine periodic oscillations superimposed on the already jagged backward wave oscillator output [see, for example, Fig. 8(d)]. In transmission, where is the correct reading, at the top of this fine structure, at the bottom, or someplace in between? On one hand, admitting that the fine structure is due exclusively to multiple reflections between the input and output transducers, the correct reading is at the top because the discontinuities in the transducers act like irises located outside of the nonreflecting device being measured, and consequently these irises can only reduce the transmission, never increase it. On the other hand, admitting that the fine structure is due exclusively to resonance of spurious modes, the correct reading is half-way between the top and the bottom since the transmission can be increased or decreased with resonances.[5] A fair compromise between the two extreme readings is the average.

Now we can look at the results. The outputs of the band-splitting filter of Fig. 3 are shown in Figs. 4(a), (b) and (c), together with the reference output of the generator, Fig. 4(d).

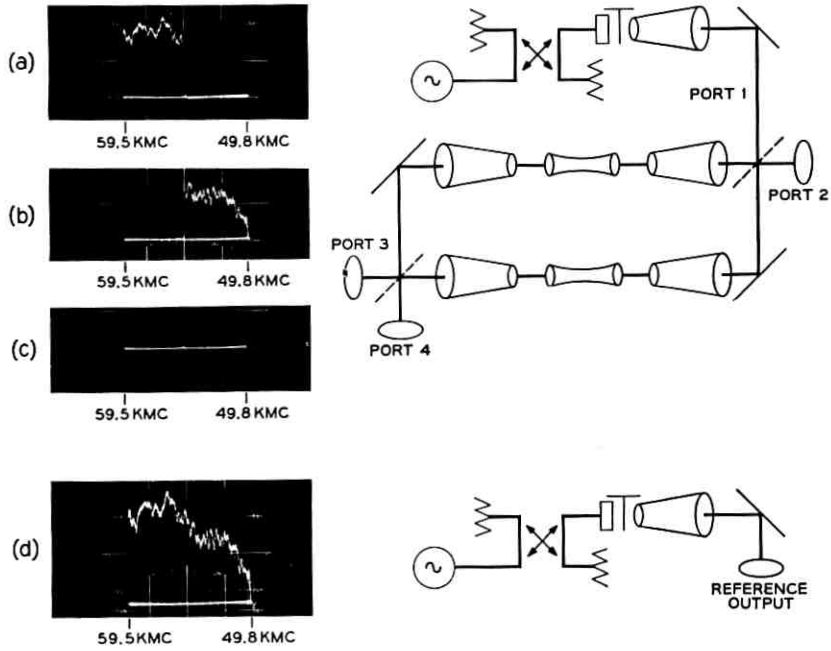


Fig. 4 — (a) Output of port 4; (b) output of port 2; (c) output of port 3 [taken with 6.0 db greater sensitivity than (a), (b) and (d)]; (d) reference output.

Far from cutoff the band-splitting filter has 1.5 ± 0.1 db insertion loss for each subband [compare Figs. 4(a), (b) and (d)]; but even as close to cutoff as 75 mc the insertion loss is increased by only 1 db. The band lost because of the splitting filter is very narrow.

It will be shown in Sections III and V that, of the 1.5 ± 0.1 db insertion loss for either subband, ~0.5 db is lost in each hybrid and ~0.2 db in each elbow. This accounts for ~1.2 db; the remaining 0.3 db must be attributed to losses in the rest of the circuitry. Any substantial reduction of losses will have to come from improvements of the hybrids.

How do we adjust the band-splitting filter? The requirement is that the power recombining at each hybrid must follow equal electrical paths. Frequencies below cutoff require path L_1 in Fig. 3 to be identical to L_1', and frequencies above cutoff must have $L_1 + L_2$ identical to $L_1' + L_2'$. The fact that the adjustment for frequencies below cutoff is independent of L_2 and L_2' suggests a two-step procedure in which the second does not alter the first:

(a) trim L_1 and L_1' for minimum power in port 1 (maximum in port 2);

(b) trim L_2 and L_2' for minimum power in port 3 (maximum in port 4).

We describe next each one of the band-splitting filter components.

III. HYBRID JUNCTION

Consider an infinite volume of metal in which two infinitely long cylindrical holes of equal diameter are bored in such a way that the axes are coplanar and normal to each other, as in Fig. 5. We thus have two cylindrical waveguides making a cross. $TE_{01}°$ mode fed in one of the arms passes straight through the junction almost unperturbed provided the diameter of the waveguide is much larger than the free-space wavelength. The reason is that we are dealing with an almost optical problem.

Now let us include in the junction a thin plane sheet of a material to be described later. The sheet passes through the intersection of the axes of the waveguides and makes an angle of 45° with each of them. This thin layer acts as a semitransparent mirror, and $TE_{01}°$ mode fed in one of the arms is partially transmitted straight through and partially reflected to one of the side arms. If the power division is half and half, the junction becomes a hybrid. If all the power is reflected (sheet of metal) the junction becomes an elbow, as in Fig. 6.

The semitransparent mirror can be obtained with sheets of dielectric, wire mesh, evaporated film, etc.

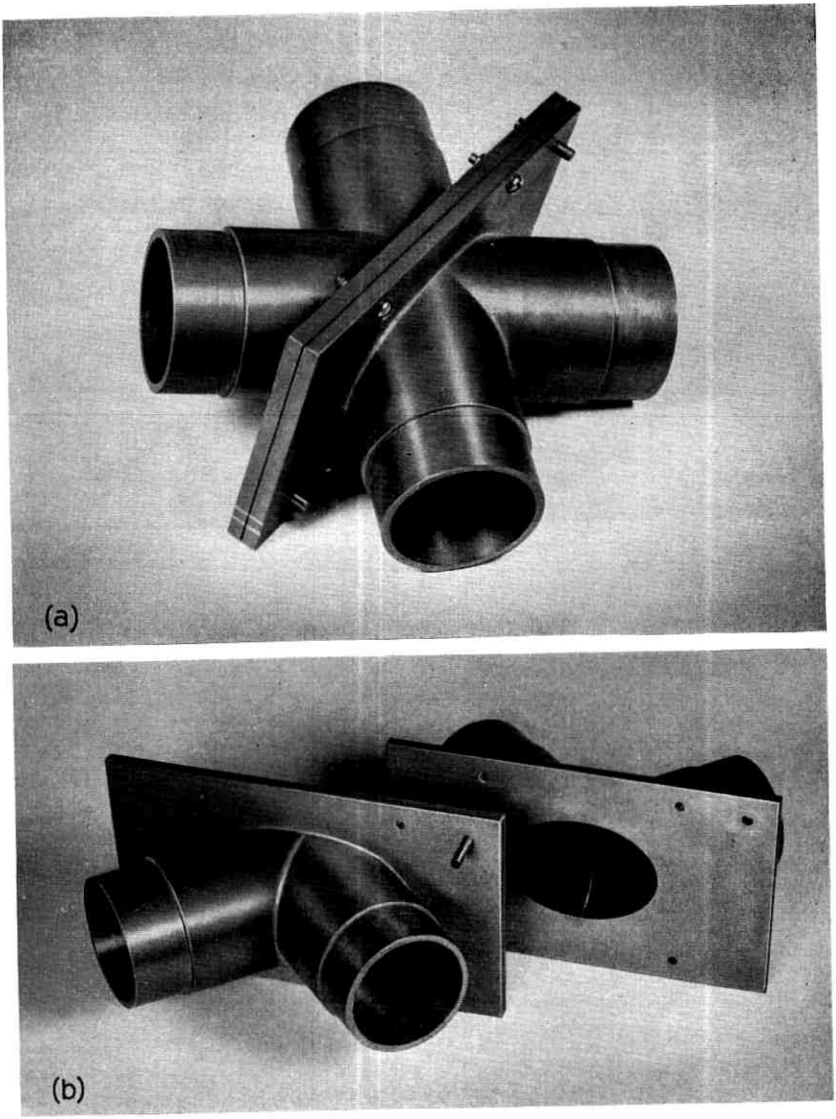


Fig. 5 — (a) TE_{01}° hybrid in 2-inch waveguide; (b) exploded view.

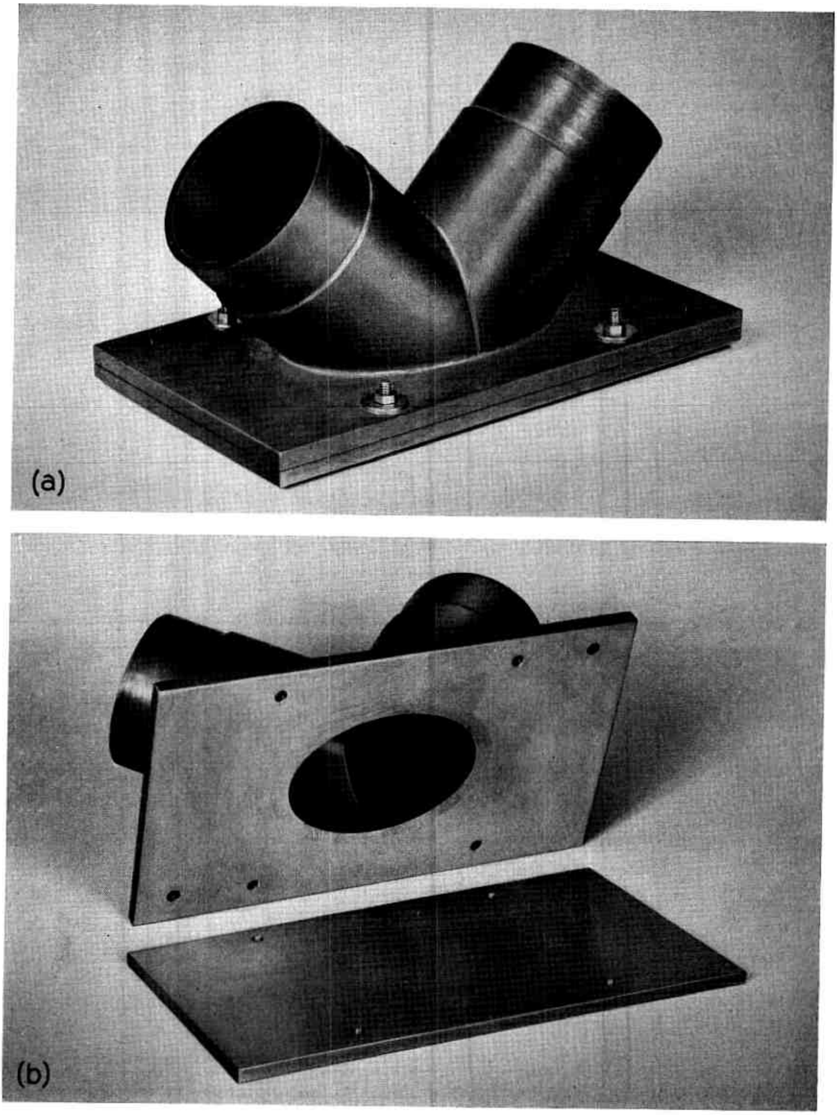


Fig. 6 — (a) TE_{01}° elbow in 2-inch waveguide; (b) exploded view.

Part of the power sent through one of these hybrids is converted to unwanted modes, a good part of which could be recovered by modifying the mirror slightly. In effect, the incident mode can be considered as the superposition of an infinite number of plane wavelets all traveling almost parallel to the axis of the waveguide and each one impinging with different polarization on the semitransparent mirror. As a consequence of the polarization each wavelet has its own reflection coefficient. Mode conversion can be avoided by making the local reflection coefficients identical; this can be achieved with a nonuniform semitransparent mirror — for example, if the sheet is dielectric or wire mesh, the thickness of the dielectric or the density of the holes must be a function of the azimuth.

Let us see the experimental results. Two hybrids were assembled using glass for the semitransparent mirrors. The thicknesses were determined experimentally to provide 3 db power division in the 50 to 60 kmc band. Three sheets of glass with a total thickness of 0.018 inch were assembled into Hybrid No. 1. The other, Hybrid No. 2, was assembled with four sheets of glass with a total thickness of 0.021 inch. The power division did not change rapidly with thickness, and the two mirrors appeared to be the best combinations obtainable with the available glass sheets.

Similar measurements were made to determine the performances of both hybrids and, since the results were very similar, Fig. 7 shows only those for Hybrid No. 1. The outputs from arms 3 and 4, Figs. 7(a) and 7(b), are nearly equal, showing close to 3 db power division.

A method to check the insertion loss and the balance of the hybrid is described next. The four port hybrid is reduced to a two port structure by placing reflecting pistons in arms 3 and 4, Fig. 7(c). By adjusting the relative position of the pistons, the power transmission is maximized. With this scheme the transmitted power crosses the hybrid twice, once going toward the pistons and second, bouncing from them. Figs. 7(c) and (d) show the transmitted and reflected levels. The reference level is shown in Fig. 7(e). The two-way loss of the hybrid is approximately 1 db ± 0.1 db. The reflected power from the hybrid with the pistons is very small [see Fig. 7(d)] because it is quite similar to the power reflected in the internal mismatches of the measuring set, Fig. 7(f). This indicates first that the 1 db insertion loss is not due to reflections but rather to mode conversion; second, that over the range from 51.0 to 60.8 kmc the power division in the hybrid is frequency-insensitive, since otherwise the pistons could not tune out the reflections over this band.

The responses of Hybrids Nos. 1 and 2 from 60 to 68.4 kmc and from

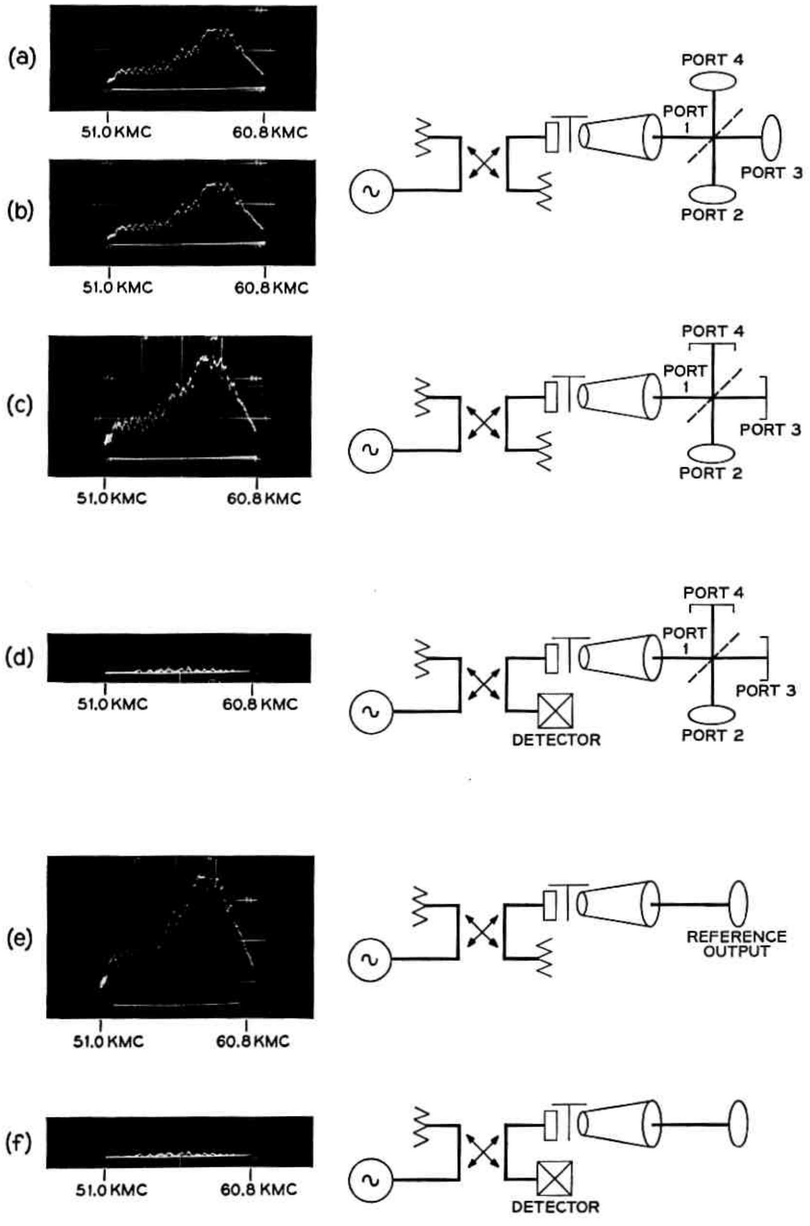


Fig. 7 — Hybrid No. 1: (a) output of port 3; (b) output of port 4; (c) output of port 2 with pistons on ports 3 and 4; (d) output of port 1 with pistons on ports 3 and 4; (e) reference output; (f) reflection from measuring set.

65.9 to 76.4 kmc are shown in Fig. 8. The outputs of arms 3 and 4, which should be equal in the ideal case of 3 db power division, have been superposed for comparison. Figs. 8(a) and 8(b) show that those outputs are very similar in Hybrid No. 1, and consequently power division has little frequency dependence. On the other hand, Figs. 8(c) and 8(d) indicate that power division in Hybrid No. 2 gets worse as the frequency increases. At the highest frequency, 76.4 kmc, the output ratio is close to two to one.

Why the dissimilar frequency behavior? The answer may be that, because of the difference in thickness of the semitransparent mirrors, the hybrids achieve the ideal 3 db power division at different frequencies, and consequently one of them is bound to behave better than the other in the range of our measurements. These hybrids are so broadband that when they were assembled to provide 3 db division in the range from 50 to 60 kmc they looked very similar, and only when they were measured at higher frequencies did the different behavior become apparent.

For the purpose of showing that other semitransparent mirrors different from glass could be adequate to build a hybrid, a copper screen was used. The copper screen is an electroformed mesh 0.0005 inch thick

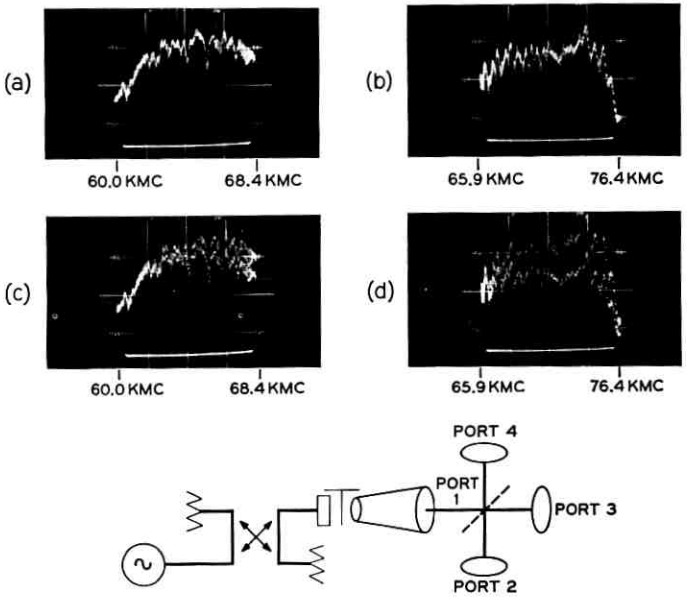


Fig. 8 — Superimposed outputs of ports 3 and 4: (a) Hybrid No. 1, 60.0 through 68.4 kmc; (b) Hybrid No. 1, 65.9 through 76.4 kmc; (c) Hybrid No. 2, 60.0 through 68.4 kmc; (d) Hybrid No. 2, 65.9 through 76.4 kmc.

with 400 square holes per square inch, with the copper separating the holes being 0.0175 inch wide.

The power division of the hybrid is 1.8 to 1, and, with more screen reflectivity, 3 db power division should be achieved. No effort was made in this direction, but a semitransparent mirror of this kind should be weighed carefully against a glass one.

IV. HIGH-PASS FILTER

A TE_{01}° high-pass filter is obtained by reducing the diameter of a circular waveguide. The minimum radius essentially fixes the cutoff frequency and the slope of the tapers determines the steepness of the transfer characteristics.

The high-pass filter, Fig. 9, is made of electroformed round copper pipes. Two relatively smooth tapers connect the cutoff section to $\frac{7}{16}$ inch-diameter waveguide. The cutoff section has a constant diameter of 0.260 inch and is 1 inch long. The over-all length of the filter is 3.8 inches.

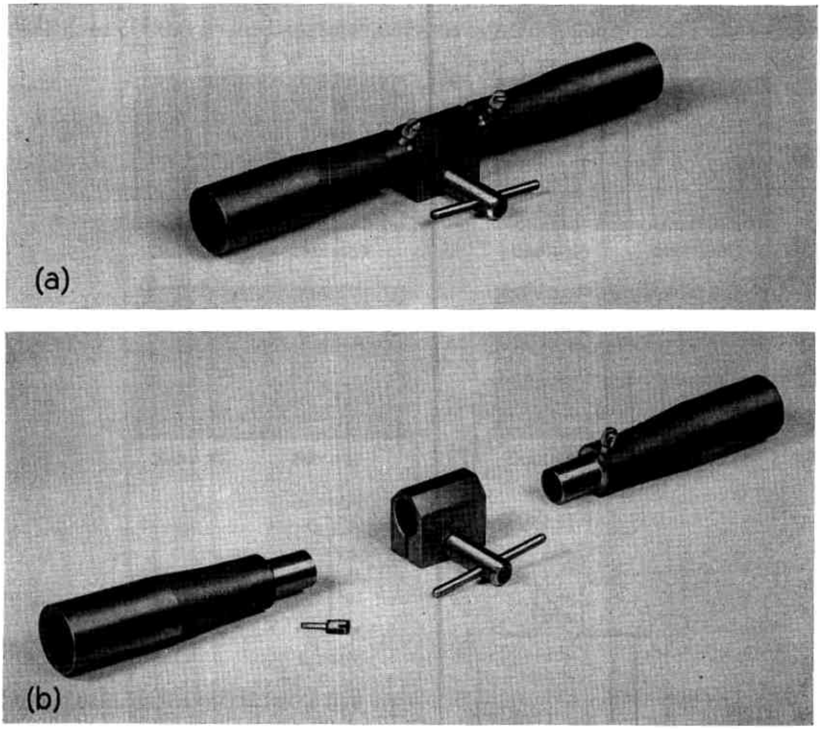


Fig. 9 — (a) TE_{01}° high-pass filter; (b) exploded view.

The end size of the high-pass filter was selected because we already had on hand tapers from $\frac{7}{16}$ to 2 inches in diameter.[6] In a final design of a band-splitting filter it is not necessary to pass through the intermediate size of $\frac{7}{16}$ inch, and consequently the filter can be more compact than the one shown in Fig. 2.

In order to appreciate the high-pass filter behavior at frequencies close to cutoff, transmission and reflection were measured point by point, as shown in Fig. 10. In an ideal high-pass filter, power at frequencies above cutoff should be transmitted completely, but this implies that the taper should match a waveguide of a certain admittance to another

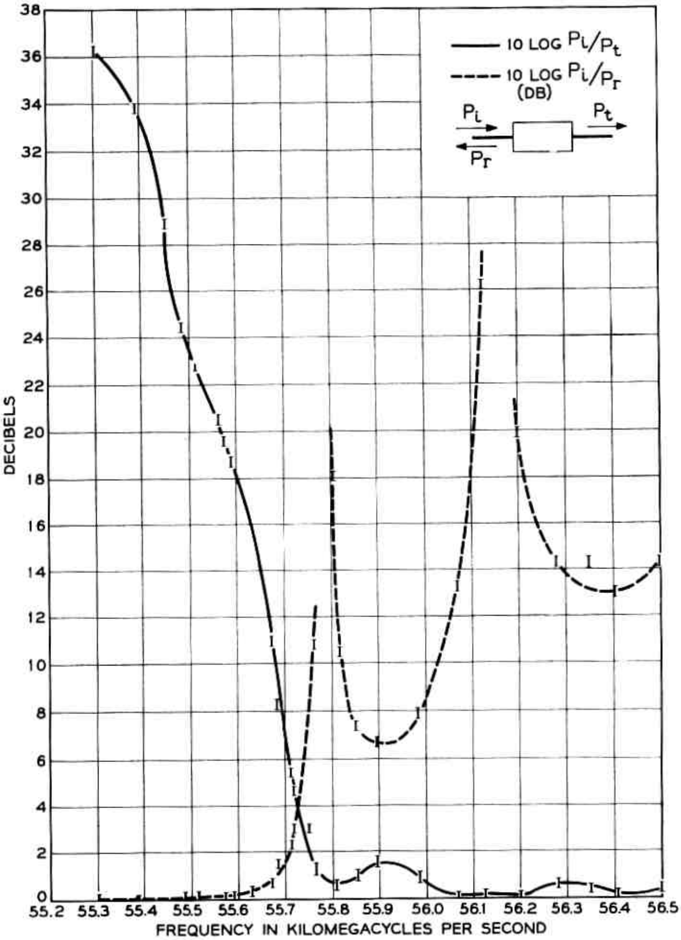


Fig. 10 — Transmission and reflection of high-pass filter (without tuning screws).

of admittance close to zero, which requires extremely long tapers. Our tapers are short, and Fig. 10 shows that the maximum unwanted reflection above cutoff is 6.6 db. This reflection can be reduced by using polyethylene tuning screws. The transmission and reflection of the taper with tuning screws, as seen in Fig. 11, show that the maximum unwanted reflection has been reduced to 16.8 db without changing substantially the transmission characteristic. The difference between the frequencies at which transmission and reflection losses are below 1 db is 160 mc. This is a very sharp cutoff.

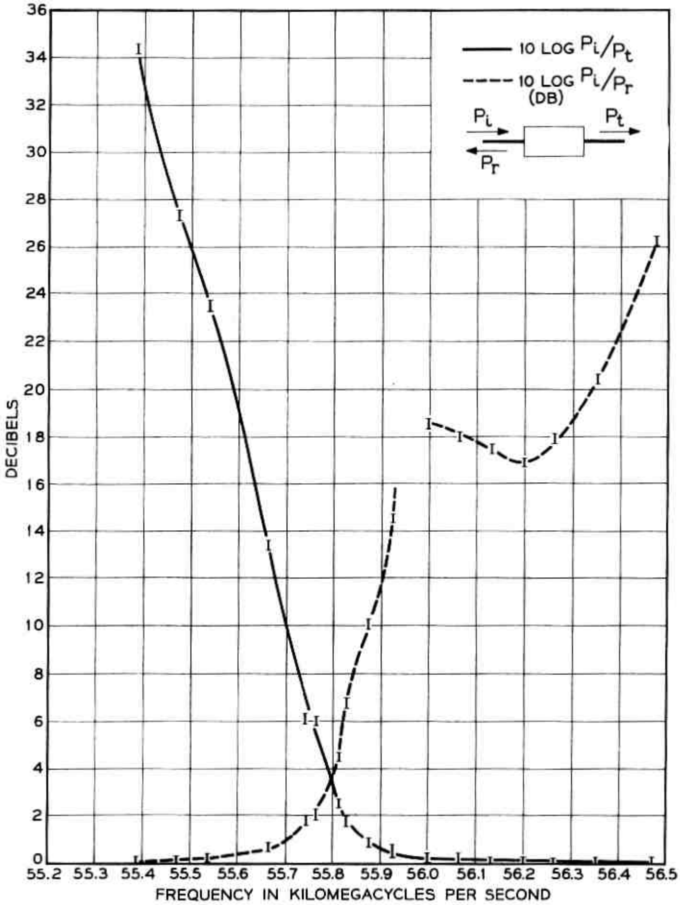


Fig. 11 — Transmission and reflection of high-pass filter (with tuning screws).

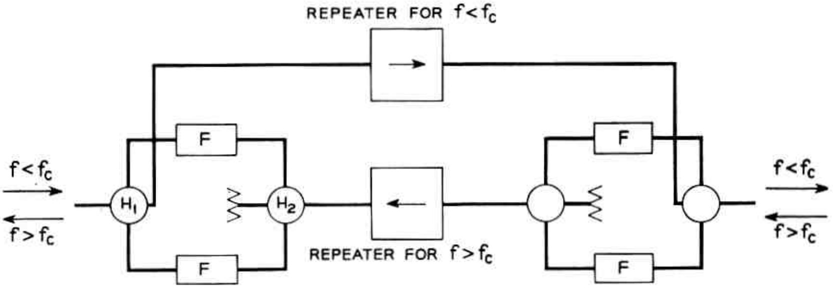


Fig. 12 — Two-way transmission on a single waveguide using two band-splitting filters.

V. ELBOW

The elbow, as mentioned earlier, is derived from the hybrid by replacing the semitransparent mirror with a metal plate. The insertion loss is 0.2 ± 0.05 db.

VI. ANOTHER POSSIBLE USE FOR A BAND-SPLITTING FILTER

This filter allows using a single waveguide to transmit in both directions. Fig. 12 shows one of the possible arrangements. Two identical band splitting filters are used. Calling f_c the cutoff frequencies of the filters F, the reader can check that frequencies $f < f_c$ can travel towards the right and frequencies $f > f_c$ can travel towards the left.

Power leaving the repeater for $f < f_c$ can leak into the repeater for $f > f_c$ only by passing through cutoff filters, and consequently that leakage can be made arbitrarily small. On the other hand, power leaving the repeater for $f > f_c$ can leak into the repeater for $f < f_c$ because of unbalance in the hybrids H_1 and H_2. This leakage can be reduced if necessary by including between hybrid H_1 and the repeater a filter like the one shown in Fig. 13. It consists of a hybrid and two high-pass filters like those described previously.

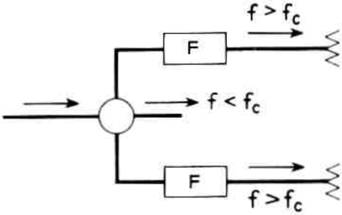


Fig. 13 — Filter to eliminate high frequencies.

VII. CONCLUSIONS

A constant-resistance filter capable of dividing a wide band in two has been described. It can easily handle one octave in the millimeter region with only 1.5 db insertion loss for each subband. The splitting transition takes place in a very narrow band (160 mc). Another use of the filter: it allows using a single waveguide to transmit in both directions.

Two of the components of the filter are important devices: an elbow and a hybrid. Both are quasi-optical and operate easily over one octave in the millimeter region. The elbows allow sudden 90° turns of a 2 inch-diameter multimode waveguide with relatively low insertion loss 0.2 ± 0.05 db. Without them the band-splitting filter would be very bulky. The importance and uses of the hybrid need hardly any comment.

REFERENCES

1. Miller, S. E., Waveguide as a Communication Medium, B.S.T.J., **33**, 1954, p. 1209.
2. Marcatili, E. A., Channel-Dropping Filter in the Millimeter Region Using Circular-Electric Modes, to be published.
3. Marcatili, E. A., A Circular-Electric Hybrid Junction and Some Channel-Dropping Filters, this issue, p. 185.
4. Marcatili, E. A., Mode-Conversion Filters, this issue, p. 149.
5. King, A. P. and Marcatili, E. A., Transmission Loss Due to Resonance of Loosely-Coupled Modes in a Multi-Mode System, B.S.T.J., **35**, 1956, p. 899.
6. Unger, H. G., Circular Waveguide Taper of Improved Design, B.S.T.J., **37**, 1958, p. 899.

Margin Considerations for an Esaki Diode-Resistor OR Gate*

By H. K. GUMMEL and F. M. SMITS

(Manuscript received July 21, 1960)

An Esaki diode-resistor logic, powered from a three-phase supply and involving OR *gates, is analyzed. Practical switching times are of the order of 10 $|R^{-}|$ C. The voltages at which the current maximum and the current minimum occur set an upper limit on the achievable logical gain. For a sum of fan-in plus fan-out of 3, the margins on key diode and circuit parameters must be better than ±2 per cent, with all margins assumed equal. The margins can be ±3.5 per cent for a fan-in plus fan-out of 2, which, however, restricts the applications to shift registers, flip-flops, and the like.*

I. INTRODUCTION

Esaki diodes are being considered for high-speed logic due to their potentially high switching speeds. Several papers have already appeared on the use of Esaki diodes in logic systems.[1,2,3,4,5] In such systems the bistable V-I characteristic of the diodes is utilized to define two logical states ("zero" and "one"). The bias current of the diode, together with a trigger current derived from a previous stage, determines which of the two states will be attained. If the trigger current can be kept small with respect to the output current, logical gain can be achieved. This generally requires that the characteristics of the diodes be well controlled, since the logical gain will primarily depend on the margins of the diodes and of the circuit parameters. Consequently, the considerations of margins become of prime importance in the design of logical systems.

In this paper, worst-case margin considerations are given for one of the simplest types of Esaki diode logic, a diode-resistor logic powered from a three-phase supply.[6] In particular, the discussion is restricted to the least complicated logical element — an OR gate.

* This work was supported in part by the U.S. Army Signal Corps under contract DA-36-039 sc-64618.

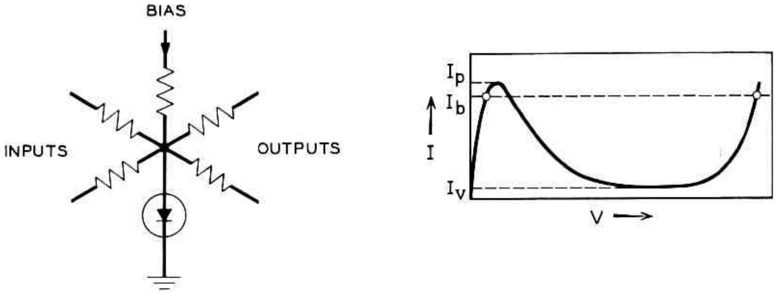


Fig. 1 — Esaki diode-resistor logic.

First, a general description of the system will be given (Section II), followed by a qualitative margin analysis (Section III).* In Section IV it is shown that the present system permits only a finite logical gain, even for zero margins and infinite switching time. The switching speed is analyzed in Section V, followed by the quantitative margin analysis (Section VI). The final result of the quantitative margin analysis is brought into a form corresponding to the qualitative analysis, which permits the reader to follow the discussion (Section VII) and the conclusion (Section VIII) without studying in detail the reasoning in Sections V and VI.

II. ESAKI DIODE-RESISTOR LOGIC

The basic stage in an Esaki diode-resistor logic consists of a series arrangement of a diode and a resistor, R_b, with input and output coupling resistors as shown in Fig. 1. The bias voltage is chosen such that, without any voltage at the far ends of the coupling resistors, it gives rise to a current through the diode which is below the peak current. Consequently, the diode will remain in its low-voltage state. With additional current supplied to the center node through one or more input (or output) resistors, the diode can be made to switch into the high-voltage state. With the bias current only slightly smaller than the peak current, very small "trigger" currents are necessary.

Once the diode is in its high-voltage state it will remain there even if current is now withdrawn at the node. It is only necessary that the current through the diode remain above the valley current I_v. The maximum current that can be withdrawn is, therefore, the difference between the bias current and the valley current. The ratio of this output current to the trigger current constitutes the logical gain.

* The authors are indebted to J. H. Vogelsong, whose unpublished margin studies are incorporated in this section.

By proper choice of the bias current, the current of at least one input or the combined current of several inputs is necessary to trigger the diode. The diode accordingly can act as an OR gate, as an AND gate or as a THRESHOLD gate. The output resistors can be connected to the node of a subsequent stage. Similarly, the input resistors are powered from nodes at previous stages. The extension of this principle leads to a logical network.

In such a network it is, however, necessary to determine the direction in which information will be propagated. One elegant method utilizes a three-phase bias supply[6] as depicted in Fig. 2. Adjacent diode stages are powered from different phases. Thus a diode on phase A, for example, is triggered from a diode on phase C, and it will trigger a diode on phase B.

Even in such a multiphase system "backswitching" can occur.[7] To illustrate this, consider the arrangement of Fig. 3. If the stages represent OR gates, one stage in a high-voltage ("one") condition will trigger a following stage. Stages 1 and 2 are powered from phase A, while stages 3 and 4 are powered from phase B. Assume that stage 1 is in the "one" condition and stage 2 is in the "zero" condition. As soon as phase B is applied, stage 3 will assume the "one" condition. Since stage 3 is coupled to stage 2, it will trigger stage 2 into the "one" condition, resulting in an erroneous "one" in stage 4.

To avoid such backswitching, a system of OR gates must be arranged in such a way that no multiple input is fed from any stage having a multiple output. Logical design then forces the use of "booster" stages, e.g., stages with one output driving stages with one input.

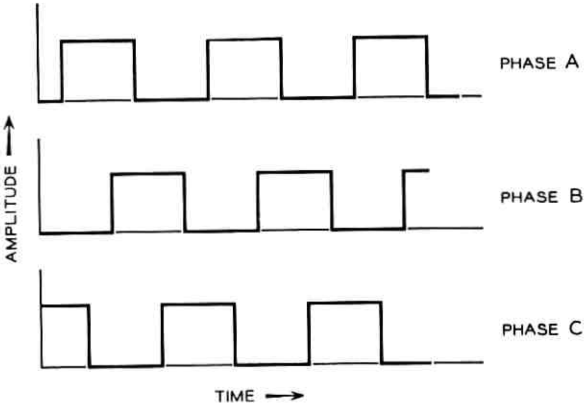


Fig. 2 — Three-phase bias supply.

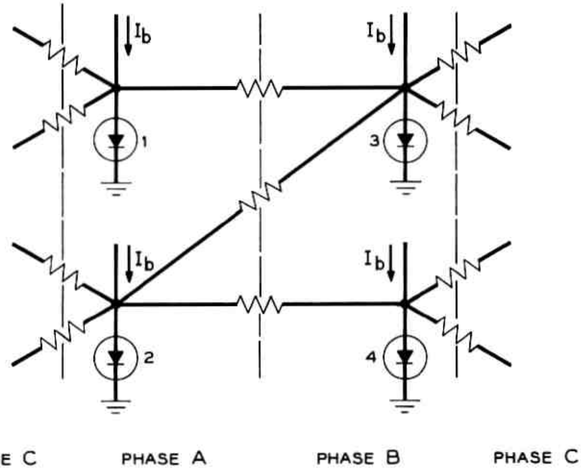


Fig. 3 — Logic network of OR gates leading to backswitching.

III. QUALITATIVE MARGIN ANALYSIS

As pointed out before, the magnitude of the trigger current depends on the difference between peak current and bias current. This difference can be made small (and therewith the logical gain large) if both parameters are tightly controlled. For an OR gate the qualitative effect of a spread in the parameters on the logical gain can be readily demonstrated.

Assume that the peak currents I_p of the devices fall in a range between I_L and I_u, and that the bias currents fall in a range between $I_{b\,\text{min}}$ and $I_{b\,\text{max}}$. One can then introduce such relative variations as:

$$\pi = \frac{I_u - I_L}{I_L} \tag{1}$$

and

$$\beta = \frac{I_{b\,\text{max}} - I_{b\,\text{min}}}{I_L}. \tag{2}$$

With a maximum valley current $I_{v\,\text{max}}$ for all devices one can define a "valley-to-peak" ratio

$$\nu = \frac{I_{v\,\text{max}}}{I_L}. \tag{3}$$

As pointed out before, for triggering a device, the total current through the device must exceed the peak current. It is plausible (and will be discussed in detail in the next section) that an overdrive ΔI is necessary to ensure switching with the required speed. We thus introduce a relative overdrive

$$\delta = \frac{\Delta I}{I_L}. \tag{4}$$

In a worst-case analysis, it must be ascertained that a stage giving a minimum total output current is capable of delivering into each output stage a trigger current which is at least as big as the trigger current required in the worst-case. The logical gain, e.g., the number of stages (n) that can be connected to one output can be found by equating the minimum current that can be delivered into an output stage ($I_{\text{out min}}$) to the worst-case trigger current (I_{tr}). Due to the bilateral nature of the Esaki diode (output and input are identical), this number is the sum of inputs plus outputs ("fan-in" plus "fan-out").

For these currents the following normalizations are introduced:

$$\tau = \frac{I_{\text{tr}}}{I_L} \tag{5}$$

and

$$\eta = \frac{I_{\text{out min}}}{I_L}. \tag{6}$$

For a qualitative analysis the magnitude of these currents is found readily by graphical considerations.

Fig. 4 shows a voltage-current characteristic of an Esaki diode including variations. Since the maximum bias current must equal the

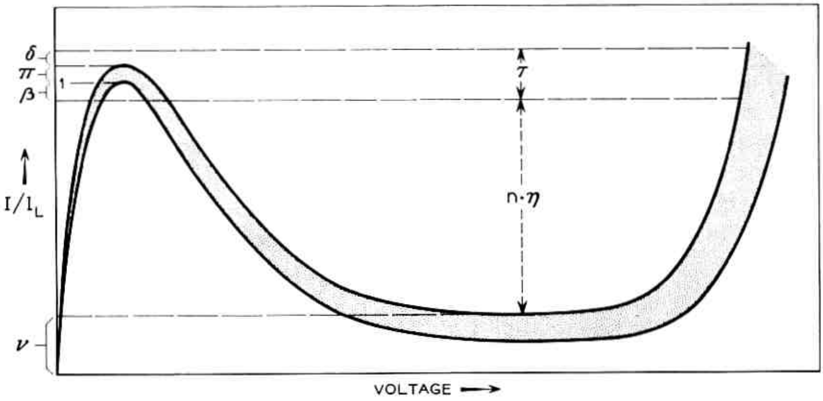


Fig. 4 — Qualitative margin analysis.

minimum peak current, the entire spread in bias currents (β) must lie below the spread in peak currents (π). Since for a unit with the highest peak current a minimum overdrive must be assured, δ must be added to the maximum peak current. From Fig. 4 one obtains for the normalized trigger current

$$\tau = \pi + \beta + \delta, \tag{7}$$

and for the total minimum output current

$$n \cdot \eta = 1 - \nu - \beta. \tag{8}$$

Equating τ and η gives the logical gain as

$$\frac{1 - \nu - \beta}{\pi + \beta + \delta} = n. \tag{9}$$

An evaluation of this equation will give a first-order estimate of the required margins. This analysis, however, neglects the effects of the peak and valley voltages, which even for zero margins will limit the logical gain under certain conditions as will be shown in the next section. For a full evaluation of the margin equations, the relation between relative overdrive δ and the switching speed must be known. This analysis will be given in Section V. The detailed margin analysis in Section VI will not only include the variables considered in (9) and the effect of the peak and valley voltages, but it also will include the variations of these voltages and the variations in the coupling resistors.

IV. LIMITATIONS DUE TO FINITE VOLTAGE LEVELS

The diode-resistor logic discussed here has an upper limit in the logical gain if stages that are to be driven have a fan-out larger than their fan-in. This limitation is determined by the magnitude of the voltages for the current peak and for the current minimum, and exists even if all error margins and the valley current are zero.

To demonstrate this effect, consider the extreme case of a stage having one input and $(n - 1)$ outputs. Assume a device characteristic as shown in Fig. 5. The magnitude of the bias current is determined by the condition that the stage under consideration is not permitted to be triggered into the high-voltage condition if none of the stages connected to either the input or the output resistor is in the high-voltage condition. Due to the bias pulse overlap with the previous and the following stages respectively, the far ends of the coupling resistors can vary in voltage between zero and V_p, the voltage at which the current maximum is reached. The coupling resistors whose far ends are at V_p will

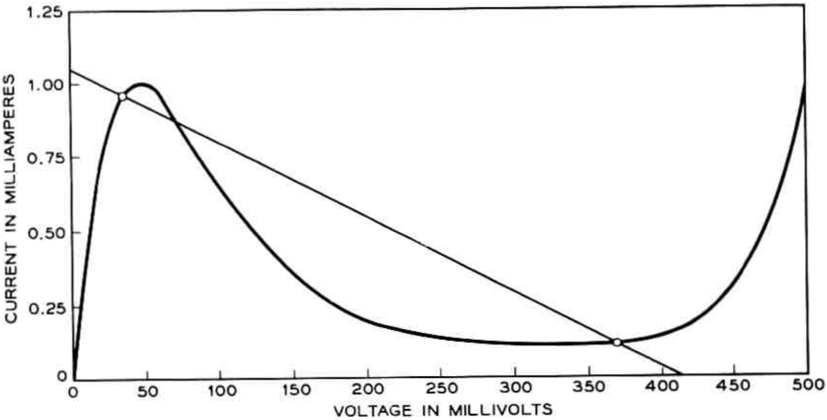


Fig. 5 — Germanium Esaki diode characteristic with load line.

not shunt any current, and it has to be ascertained that the bias current under no condition will supply a current greater than I_p to the diode. With only one coupling resistor acting as a shunt, the maximum bias current becomes

$$I_b = I_p + V_pG, \tag{10}$$

where G is the conductance of each coupling resistor.

A stage is to be triggered into the high-voltage condition at a time when the output resistors are at ground potential, and the trigger current must increase the current through the diode above the peak current. Since, for the case of one input resistor, $(n - 1)$ resistors act as shunts, one obtains for the trigger current

$$I_{\text{tr}} = I_p + (n - 1)GV_p - I_b = (n - 2)GV_p\,. \tag{11}$$

It must be possible to supply this trigger current from the output of one previous stage which is in the high-voltage condition. The far end of this particular coupling resistor thus is at the "valley" voltage V_v, with the near end at v_p. One thus obtains for the output current

$$I_{\text{out}} = (V_v - V_p)G. \tag{12}$$

Equating I_{out} and I_{tr} yields

$$n = \frac{V_v}{V_p} + 1. \tag{13}$$

This demonstrates that the sum of fan-in and fan-out for such an asymmetrical stage remains finite, even in the case of zero tolerances.

V. SWITCHING SPEED

The speed considerations are based on a diode characteristic as shown in Fig. 5, which shows a good but still practical characteristic for a germanium diode. The combined conductances of the input and output resistors are shown as the load line. It is chosen in such a way that while touching the peak point it intercepts the valley. This choice of the load line will allow obtaining a maximum current output.

Assume now that a current of magnitude I_0 ($> I_p$) is applied to the parallel combination of diode and load resistance as shown in Fig. 6. The capacity represents the diode capacity (plus any shunt capacitors in parallel with the diode). Any series inductances have been neglected.

During a transition from a low-voltage state into a high-voltage state, the capacity shunting the diode must be charged. The charging current at a given voltage is the difference between the supplied current I_0 and the sum of the load current and the conductive current through the diode at any given voltage. In Fig. 6 this charging current I_c can be read off as a function of voltage, since it is just the difference between the load line and the static characteristic. The time required to go from voltage V_a to voltage V_b is given by

$$t = \int_{V_a}^{V_b} \frac{C(V)\,dV}{I_c(V)}\,. \tag{14}$$

A numerical integration of this equation thus can give the switching time between two arbitrary points. For the problem on hand, the

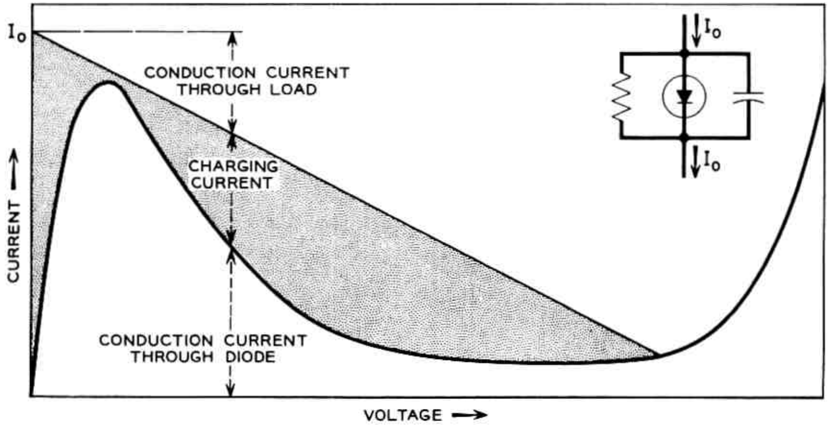


Fig. 6 — Capacitance charging current.

switching time from a low-voltage state to a high-voltage state is of interest.

Prior to switching, the diode voltage corresponds to the stable point (Fig. 5) at low voltage. After switching, the diode voltage corresponds to the stable point at high voltage. Switching is accomplished by the application of a trigger current which lifts the load line above the diode characteristic as depicted in Fig. 6. This lifting of the load line moves the high-voltage intercept to the right, and the diode voltage will approach the voltage corresponding to this intercept. After removal of the overdrive, the voltage would then decrease again. Thus, switching can be considered as completed when, with applied overdrive, the voltage of the stable point prior to the application of the trigger current is reached.

For the analysis to be independent of the particular bias current, it is convenient to consider, as final voltage, the intercept of a load line which just touches the peak. For an analytical treatment the following two simplifying assumptions will be made:

i. The voltage dependent capacity $C(V)$ will be replaced by a constant average capacity C.

ii. The diode characteristic will be approximated by two parabolic sections and a straight section.

Let A and B (Fig. 7) be the points on the diode characteristic at which the slope is equal to that of the load line, and let I_A and I_B be the charging currents at these points. Let C be the point at which one parabolic

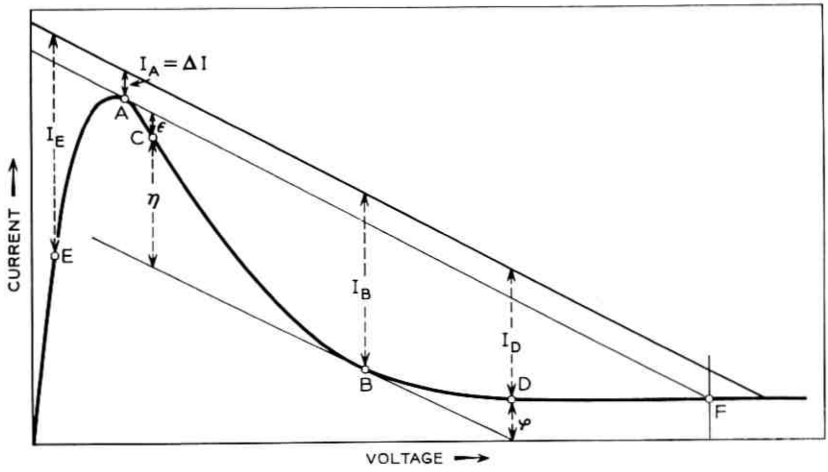


Fig. 7 — Analysis of the switching time.

approximation joins the second approximation and let D be the point where the straight section commences. Then the time needed to switch from an arbitrary point E in the low-voltage part to the final point F is given by

$$t = C\left[\int_{V_E}^{V_C} \frac{dV}{I_A + I_p\left(\frac{V - V_A}{V_2}\right)^2} + \int_{V_C}^{V_D} \frac{dV}{I_B - I_p\left(\frac{V - V_B}{V_2}\right)^2} \right. \\ \left. + \int_{V_D}^{V_F} \frac{dV}{I_A + (I_D - I_A)\,\frac{V_F - V}{V_F - V_D}}\right]. \tag{15}$$

Here the constants describing the curvature of the parabolas are expressed in terms of the peak current, and in terms of constants V_1 and V_2 having the dimension of voltages. Such a presentation is convenient since the constant V_1 is equal to the peak voltage and the constant V_2 is of the order of magnitude of the difference between valley voltage and peak voltage. Note that V_1 and V_2 depend on the diode characteristic only and are independent of the load line. Performing the integration of (15) yields

$$\tau_s = \frac{CV_1}{I_p}\sqrt{\frac{I_p}{I_A}}\left(\tan^{-1}\frac{\epsilon}{I_A} + \tan^{-1}\frac{I_E}{I_A}\right) \\ + \frac{CV_2}{I_p}\sqrt{\frac{I_p}{I_B}}\left(\tanh^{-1}\frac{\eta}{I_B} + \tanh^{-1}\frac{\varphi}{I_B}\right) + C\,\frac{V_F - V_D}{I_D - I_A}\ln\left(\frac{I_D}{I_A}\right). \tag{16}$$

The constants in this equation are defined in Fig. 7. It should be noted that I_A corresponds to the overdrive ($I_A = \Delta I$) and that $I_A/I_p = \delta$ as defined in (4). For an evaluation of this equation, the value of the capacity C must be known. This capacity, however, can be expressed through the characteristic time $\tau_0 = \mid R^- \mid C$, which time is usually considered as the figure of merit for an Esaki diode. It is therefore possible to express the switching time in terms of the time τ_0. With such a normalization, the results of the analysis will be fairly general.

Fig. 8 shows an evaluation of (16) giving the switching times in terms of the characteristic time τ_0 for switching from zero to the final voltage. The assumed value of the load line permits a maximum output current while conforming with the margin considerations. Using zero instead of a finite voltage as the starting point lengthens τ_s only insignificantly.

As can be seen in Fig. 8, the switching behavior can be fairly well approximated by

$$\tau_s/\tau_0 = 2.25\sqrt{I_p/I_A} = 2.25/\sqrt{\delta}. \tag{17}$$

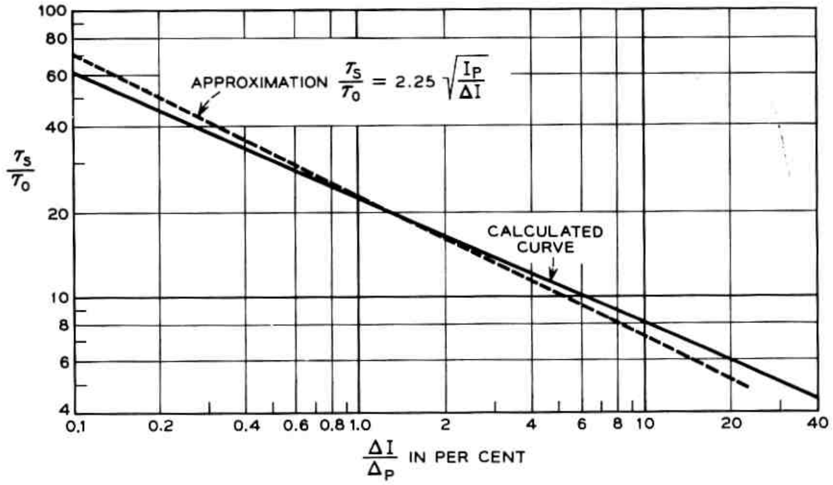


Fig. 8 — Switching speed.

VI. QUANTITATIVE MARGIN ANALYSIS

The qualitative margin analysis of Section III neglected variations of the coupling resistors, the effect of the peak and valley voltage, and the variations of these voltages. As a simplification in the complete analysis it will be assumed that the bias is supplied from a constant current source. Under this assumption the total load line of the diode is given by the sum of the conductances of the input and output resistors. These resistors will terminate at the nodes of adjacent units, and it is necessary to include the voltage of these nodes in the analysis.

As in the qualitative analysis one has to find the current which in the worst case will trigger a stage within a desired time. This current must equal the minimum current that under worst conditions will be delivered into an output resistor.

6.1 *Trigger Current Needed*

The lower the bias current, the larger will be the necessary trigger current. The maximum bias current, however, must be low enough that the stage under consideration will not assume the high-voltage condition unless one driving stage is in such a high-voltage condition. The margin of the bias current then determines the difference between the maximum and the minimum bias current.

The characteristics of all diodes in a system will show a spread, and will fall between two extreme characteristics, which represent the ac-

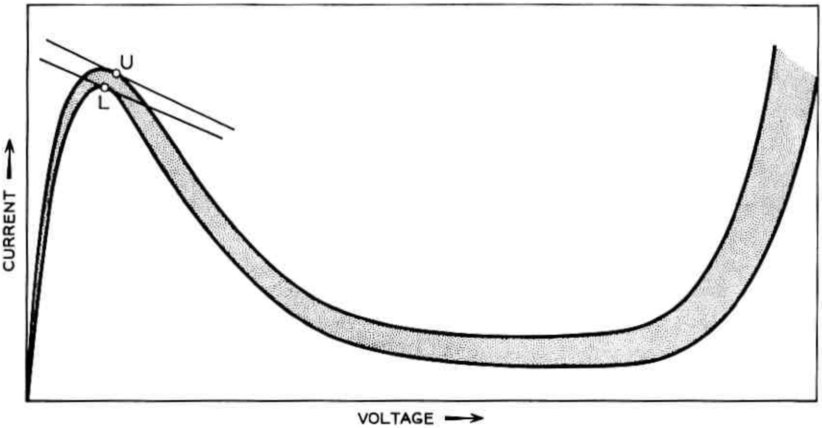


Fig. 9 — Spread in characteristics.

ceptance limits. Such extremes are indicated in Fig. 9. The conductance of the n_i input resistors and the n_0 output resistors will fall between an upper bound of value G_u and a lower bound of value G_L.

For the following analysis one determines the point L (Fig. 9) where a load line of the lowest total conductance $(n_i + n_0)\, G_L$ is tangent to the lower characteristic, and the point U where the load line of largest conductance $(n_i + n_0)\, G_U$ is tangent to the upper curve. The voltages and currents corresponding to these points are V_U, V_L, I_U and I_L respectively.

In a three-phase system as considered here, only two adjacent units will be powered at the same time. Thus one has two extreme conditions under which the unit under consideration should not be triggered:

i. The unit under consideration and the previous unit are powered, in which case the far ends of the input resistors may be as high in voltage as V_L while the output resistors are at ground potential and

ii. The unit under consideration and the following unit are powered, in which case the far ends of the output resistors may be as high as V_L in voltage while the input resistors are at ground.

The maximum permissible bias current should not switch a unit if it has the lowest peak current and if the smallest current is shunted by the coupling resistors. This current is given by

$$I_{b\,\max} = I_L + n_{\min} G_L V_L, \tag{18}$$

where $n_{\min}$ represents the minimum of n_i or n_0. The minimum bias current is below the maximum bias current by the spread in bias currents ΔI_b

$$I_{b\,\min} = I_{b\,\max} - \Delta I_b . \tag{19}$$

In order to trigger a diode it is necessary that a current of at least I_U plus necessary overdrive for speed is delivered to the diode. This current will be the sum of the minimum bias current plus the trigger current minus the current drained through the input and output resistors. However, since at least one stage has to supply the trigger current, only $n_i - 1$ input stages can act as a load. Thus one obtains under worst-case conditions

$$I_U + \Delta I = I_{\text{tr}} + I_{b\,\min} - G_U V_U n_0 - (n_i - 1)\, G_U\, (V_U - V_{\text{off min}})_L . \tag{20}$$

In this equation $V_{\text{off min}}$ is the minimum off-voltage a unit can assume. (It should be realized that the voltage V_L discussed previously can be considered as $V_{\text{off max}}$.)

To bring (20) into normalized form the following quantities are introduced

$$\rho = \frac{G_U - G_L}{G_L}, \tag{21}$$

$$\sigma = \frac{V_U - V_L}{V_L}, \tag{22}$$

$$\varphi = \frac{V_U - V_{\text{off min}}}{V_L}, \tag{23}$$

$$\gamma = \frac{G_L V_L}{I_L}. \tag{24}$$

With these definitions and with the definitions (1), (2), (3), (4) and (5), equation (20) combined with (18) and (19) gives for the normalized trigger current

$$\tau = \pi + \beta + \delta + \gamma[(1 + \rho)(1 + \sigma)n_0 - n_{\min} + (1 + \rho)\varphi(n_i - 1)]. \tag{25}$$

The term multiplied by γ involves the additional terms not present in the equivalent equation (7) of the qualitative analysis.

6.2 *Output Current Available*

In a three-phase diode system as discussed here, the output current must be available while the stage under consideration and the output stages are powered. The nodes of the input stages accordingly are at zero potential while the output stages are at the peak voltage. Certainly the driving stage is in the high-voltage condition and its voltage corresponds to a voltage V_v in the "valley" of the characteristic.

One thus can represent the situation by a circuit as shown in Fig. 10.

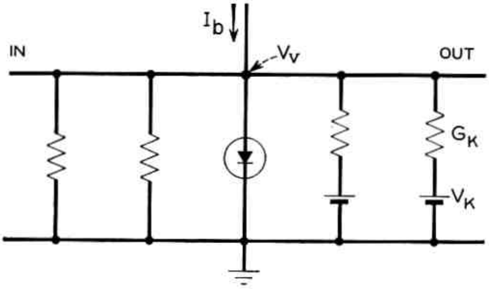


Fig. 10 — Derivation of output current.

The conductances G_k represent the coupling resistors and the voltages V_k fall in the range of the peak voltages. For the analysis it is more convenient to represent the voltages V_k and their associated conductances as current sources feeding the node. This leads to the representation as shown in Fig. 11. Here it must be remembered that $V_k = 0$ for the input conductances. From this representation the valley voltage V_v is readily obtained as

$$V_v = \frac{I_b + \Sigma V_k G_k - I_v}{\Sigma G_k}. \tag{26}$$

If G_j represents one particular coupling resistor to an output stage the current into this stage is readily obtained:

$$I_{\text{out}} = I_j = (V_v - V_j)G_j . \tag{27}$$

Substituting V_v from (26) gives

$$I_{\text{out}} = \frac{I_b + \Sigma V_k G_k - I_v - V_j \Sigma G_k}{\Sigma G_k} G_j . \tag{28}$$

This current has a minimum if

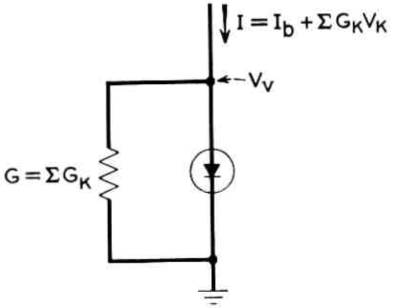


Fig. 11 — Circuit equivalent to Fig. 10.

$$\begin{aligned} I_b &= I_{b\ \min}\,, \\ G_j &= G_L \quad \text{and all other } G_k = G_U\,, \end{aligned} \tag{29}$$

and if for the output stages

$$V_j = V_u \quad \text{and all other } V_k = V_L\,.$$

Introducing these conditions into (28) and substituting, one obtains in normalized form:

$$\begin{aligned} \eta[n + (n-1)\rho] = 1 - \nu - \beta - \gamma[(1+\rho)\,(1+\sigma)n_i - n_{\min} \\ + (1+\rho)\sigma(n_0 - 1)]\,. \end{aligned} \tag{30}$$

Again, this expression has additional terms which are absent in the equivalent equation (8) of the qualitative analysis.

6.3 *Logical Gain*

To determine the sum of fan-in plus fan-out, the normalized trigger current, (25), must be set equal to the normalized output current, (30). For this it is necessary to consider the particular configuration in which stages are interconnected.

For a given n, the trigger current, (25), will have a maximum if $n_i = 1$ and $n_0 = n - 1$. Similarly, the output current, (30), will have a minimum if $n_i = n - 1$ and $n_0 = 1$. Thus the worst combination of two stages is the case in which a stage with a multiple input and a single output drives a stage with a single input and a multiple output. Such a combination represents a "booster" stage, which is an important configuration to avoid backswitching. In the following, the analysis will therefore be given for such a combination. At this point it is convenient to introduce

$$n^* = n + (n-1)\rho. \tag{31}$$

For the worst-case combination of stages, (25) and (30) take then the form

$$\tau = \pi + \beta + \delta + \gamma[(n^* - 1)(1+\sigma) - 1], \tag{25a}$$

$$n^*\eta = 1 - \nu - \beta - \gamma[(n^* - 1)(1+\sigma) - 1]. \tag{30a}$$

Besides the variations of all parameters, these equations involve the value of the coupling resistors in the term γ. Since the load impedance determines the operating point in the high-voltage condition, one can express γ by this point, i.e., by the valley voltage and the valley current. Due to the variations in the load conductances and the bias current, this operating point must be defined for a particular combination of these parameters. The combination to be chosen is the one leading to the minimum output current. One thus can use (26) to express γ in terms of the valley voltage V_v and the valley current I_v. Introducing the condition (29) into (26) (and assuming the particular configuration under discussion) gives:

$$V_v = \frac{I_{b\,\mathrm{min}} + V_U G_L - I_v}{(n - 1)G_U + G_L}. \tag{32}$$

It is plausible (and can be proven readily) that a device characteristic leading to an operating point with a higher voltage V_v or a lower current I_v results in a higher output current. Thus the operating point as defined by (28) is a worst-case condition; such an operating point is schematically indicated in Fig. 12. It specifies an area (shaded) which must be cleared by the high-voltage branch of all diode characteristics.

Introducing a normalized valley voltage

$$y = \frac{V_v}{V_L} \tag{33}$$

and using all previous normalizations permits one to use (32) for the elimination of γ:

$$\gamma = \frac{1 - \nu - \beta}{n^* y - (\sigma + 2)}. \tag{34}$$

Equating τ and η as given in (25a) and (30a) and using the above expression for γ leads to the final result:

$$\frac{1 - \nu - \beta}{\pi + \beta + \delta} = \frac{[n + (n - 1)\rho]y - (\sigma + 2)}{y - [n + (n - 1)\rho](\sigma + 1) + 1}. \tag{35}$$

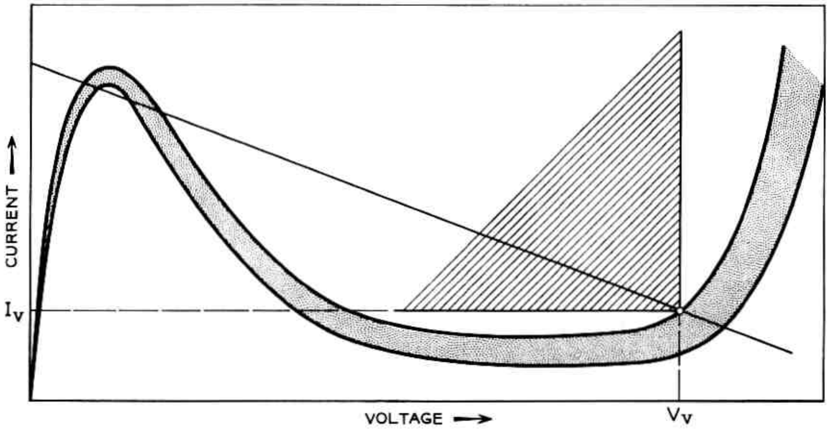


Fig. 12 — Significance of valley voltage and valley current.

VII. DISCUSSION

The final expression of the quantitative margin analysis, (35), differs from the qualitative result obtained in Section III, (9), in the right-hand side only. While in the qualitative expression the right-hand side is the sum of fan-in plus fan-out (n); in the quantitative expression the right-hand side is a function of n and these additional variables: the relative variation in the conductances of the coupling resistors ρ, (21); the relative variation in the voltage for the current peak σ, (22); and the minimum ratio of valley voltage to peak voltage ν, (33).

On account of the similarity in the results it is convenient to introduce a generalized $\tilde{n}$ defined as

$$\tilde{n} = \frac{[n + (n - 1)\rho]y - (\sigma + 2)}{y - [n + (n - 1)\rho](\sigma + 1) + 1}. \tag{36}$$

Even for σ and ρ equal to zero the generalized $\tilde{n}$ becomes infinite for

$$n = y + 1. \tag{37}$$

An infinite $\tilde{n}$ implies zero margins for all variables and zero overdrive. Thus the result of Section IV is recovered.

In considering the effects of finite margins, specific assumptions as to the relative magnitude of the margins on the various variables must be made. For the primary variables in the left-hand side of (9) or (35), the maximum ratio ν of valley current to peak current is assumed as 0.1, since this corresponds to a good ratio achievable in germanium units. The relative overdrive δ will be expressed in terms of the switching speed using the calculated relation obtained in Section V. The result is shown in Fig. 8, which gives the switching time τ_s in terms of the characteristic time $\tau_0 = | R^- | C$ as a function of the relative overdrive δ.

The other quantities of the left-hand side will be treated as independent variables. To keep the discussion fairly general, we assume that all significant parameters are kept within the same relative variation.

In general, the bias current will be determined by a voltage and a resistor. Thus the spread in bias current β is the result of an uncertainty in a voltage and in a resistor. In worst-case analysis the two margins have to be added. Assuming the two margins to be equal and introducing a relative maximum variation, x, from the center value, one can express β as:

$$\beta = 4x. \tag{38}$$

In the analysis, no mention has been made of noise; in particular, the possibility of undesirable crosscoupling between stages. Such "noise" most conveniently can be expressed in an equivalent relative variation of the peak currents. The parameter π accordingly contains noise as a second variable besides the actual variation in peak currents. Equating these variations to the variations in the bias current results in:

$$\pi = 4x. \tag{39}$$

Fig. 13 shows a plot of (9) with β and π expressed by (38) and (39) respectively and with $\nu = 0.1$. It can be seen from the figure that even for small logical gain fairly tight margins are required. It also becomes apparent that switching speeds below 10 τ_0 are impractical.

To evaluate the importance of the margins of the additional variables, which enter in $\tilde{n}$, the relation (36) between actual n and generalized $\tilde{n}$ has to be evaluated. Fig. 14, as an example, shows this relation for an actual $n = 3$, which corresponds to the minimum sum of fan-in plus fan-out required in a logical network. It can be seen that with an increasing ratio of valley voltage to peak voltage (y) the limiting value of $\tilde{n} = n + (n - 1)\rho$ is rapidly approached. It also becomes apparent that, for sufficiently large values of the valley voltage, the spread in peak voltage and in the values of the coupling resistors are only of minor importance.

Since for good germanium Esaki diodes $y \geqq 8$ for $\nu \leqq 0.1$, these

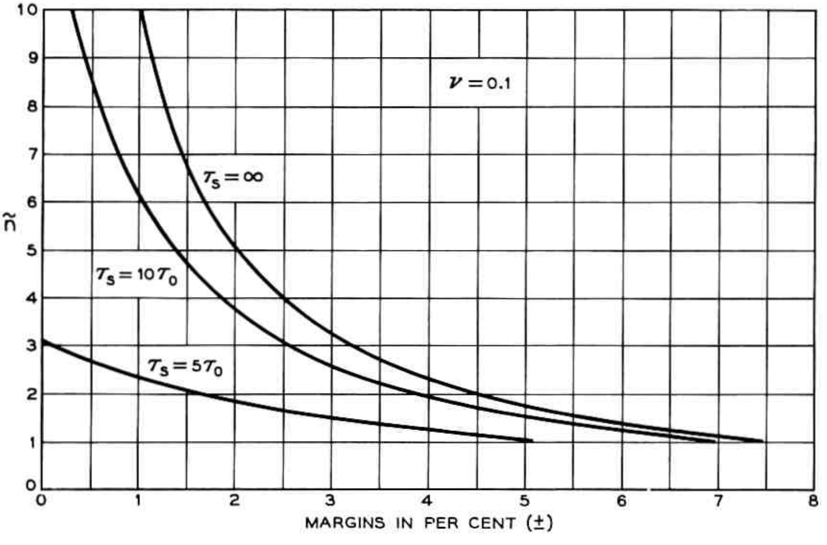


Fig. 13 — Permissible variations from nominal value of the important parameter as a function of generalized $\tilde{n}$.

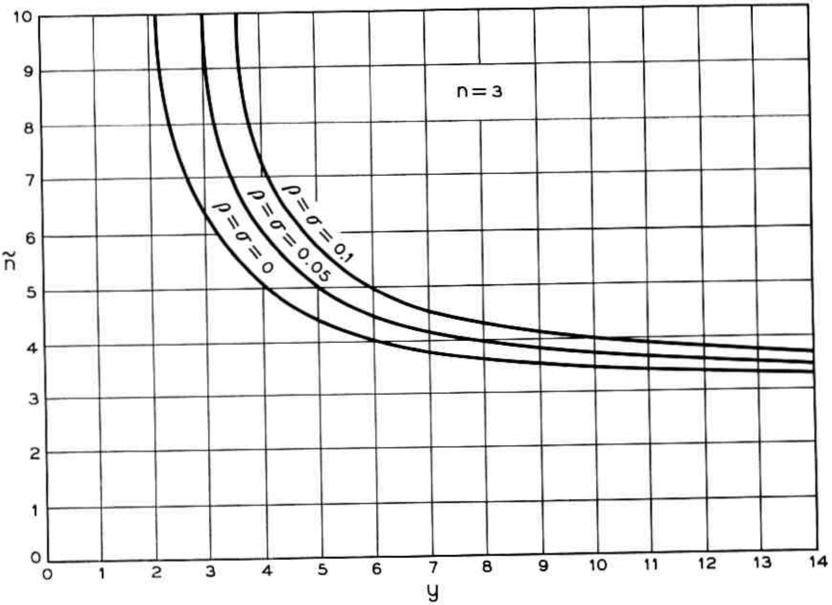


Fig. 14 — Generalized $\tilde{n}$ as a function of $y = V_v/V_L$.

values have been assumed in the construction of Fig. 15, which shows the dependence of the sum of fan-in plus fan-out (n) on the percentage variations x from the correct value of the bias voltage, the bias resistor and the peak current, and a noise equivalent current expressed as a percentage of peak current. Because the effects of the coupling resistors and the peak voltage are relatively small, a fixed relative variation of ±2.5 per cent has been assumed for these quantities.

VIII. CONCLUSIONS

An Esaki diode resistor logic with three-phase power supply shows several basic limitations even if it only involves OR gates.

The possibility of backswitching limits the design of logical networks, requiring the incorporation of "booster" stages in which a device with one output drives a device with one input.

Switching times shorter than 10 $| R^- |$ C are not practical; however, this is not a very severe limitation.

The finite ratio of the voltage for the current minimum to the voltage for the current maximum limits the logical gain even for the case of zero margins.

From Fig. 15 it is apparent that, even for a sum of fan-in plus fan-out

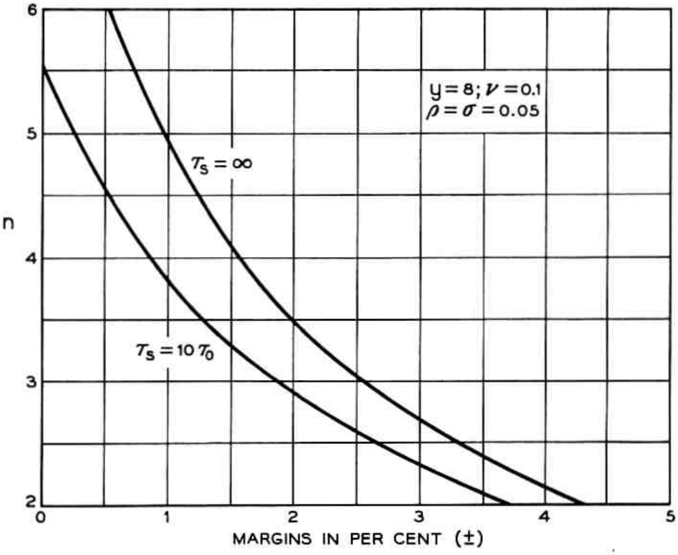


Fig. 15 — Permissible variations from nominal value of the important parameter as a function of fan-in plus fan-out for a specific example.

(n) equal to 3, worst-case margins of less than ±2.5 per cent are required. For an AND gate, a two-sided limit on the trigger current is required, making the margins even tighter. It thus appears questionable that an Esaki diode resistor logic with $n \geqq 3$ is practical, if operation under worst-case conditions is to be guaranteed.

Only for $n = 2$ do the margins appear tolerable under worst-case conditions. However, such a value of n does not permit the construction of a complete logic network, and implies a restriction to applications such as memories, flip-flops, shift register and the like.

REFERENCES

1. Lewin, M. H., Samusenko, A. G. and Lo, A. W., The Tunnel Diode as a Logic Element, Solid State Circuits Conf., Philadelphia, 1960.
2. Neff, G. W., Butler, S. A. and Critchlow, D. L., Esaki (Tunnel) Diode Logic Circuits, Solid State Circuits Conf., Philadelphia, 1960.
3. Miller, J. C., Li, K. and Lo, A. W., The Tunnel Diode as a Storage Element, Solid State Circuits Conf., Philadelphia, 1960.
4. Chow, W. F., Tunnel Diode Digital Circuitry, Solid State Circuits Conf., Philadelphia, 1960.
5. U. S. Navy Department, Bureau of Ships, Electronic Division, *The Dener Diode as a Digital Computer Element* — Phase 1, Project Lightning, Vol. 4, Ch. 8, p. 83.
6. Goto, E., On the Application of Parametrically Excited Nonlinear Resonators, J. Inst. Elect. Comm. Eng. Japan, **38**, 1955, p. 770.
7. Taylor, R. G., unpublished work.

Noncylindrical Helix Waveguide

By H. G. UNGER

(Manuscript received August 9, 1960)

Small uniform deformations of the cross section of helix waveguide perturb the circular electric waves slightly. From these perturbations the added circular electric wave loss is found in a uniformly deformed helix waveguide. For a nonuniformly deformed helix waveguide Maxwell's equations are converted into generalized telegraphist's equations. By an approximate solution for small deformations, mode conversion and circular electric wave loss are found.

Random imperfections with small correlation distance cause an average circular electric wave loss that is nearly independent of the wall impedance which the helix jacket presents to the waveguide interior. It is therefore nearly the same as in metallic waveguide. Near 50 kmc, the rms value of elliptical diameter differences should not be more than 0.0015 inch in order that on the average not more than 10 per cent of TE_{01} loss in a perfect 2-inch inside diameter copper pipe is added to the TE_{01} loss in a helix waveguide of the same inside diameter.

I. INTRODUCTION

Helix waveguide composed of closely wound insulated copper wire covered with a jacket of dielectric material and surrounded by a coaxial metallic shield is a good transmission medium for circular electric waves.[1] In long distance communication with these waves helix waveguide is useful as a mode filter, for negotiating bends and particularly as a transmission line proper. The different applications of helix waveguide require different properties of jacket and shield. Corresponding design rules have been worked out.[2]

The loss of circular electric waves in a metallic waveguide decreases steadily with increasing frequency only if the guide is perfectly round. The same is true for the helix waveguide. To maintain the low-loss properties of the circular electric wave, the helix waveguide must be manufactured to a high degree of roundness and uniformity.

As long as the guide is cylindrical, i.e., any deviation from roundness is independent of distance along the guide, increased circular electric wave loss is the only effect of such deviation from roundness. But if at the same time this deviation changes with length, the transmission characteristics of the guide will be further degraded by mode conversion-reconversion effects. At any change of cross-sectional shape of the guide, power of the circular electric wave will be scattered into unwanted modes, and vice versa. The amount of power scattered depends not only on the magnitude of change but also on the rate of change with length of these deviations from roundness.

Two cases, that of the uniform noncircular helix waveguide and that of the nonuniform helix waveguide, will be analyzed separately. In the first case, a perturbation of the normal modes of the round waveguide will give a simple answer. In the second case, however, Maxwell's equations will be converted into generalized telegraphist's equations,[3] and the results appear to be much more involved.

This paper partly represents an extension of an analysis of noncylindrical metallic waveguide[4] to helix waveguide, and partly uses the results of a mode-conversion analysis which was made more recently.[5,6]

II. THE UNIFORM NONCIRCULAR HELIX WAVEGUIDE

The mathematical model with which helix and surrounding jacket structure is represented in this analysis is an anisotropically conducting sheath. The sheath conducts perfectly in circumferential direction and has a surface impedance Z in longitudinal direction. A cylindrical coordinate system (r,φ,z) will be used, in which $r = 0$ coincides with the axis of the guide. At present the inner radius of the guide is a function of φ only:

$$a = a_0[1 + \delta(\varphi)]. \tag{1}$$

The anisotropic sheath imposes the following boundary conditions at $r = a$:

$$E_\varphi + E_r \frac{d\delta}{d\varphi} = 0, \tag{2}$$

$$E_z = \frac{-Z}{\sqrt{1 + \left(\dfrac{d\delta}{d\varphi}\right)^2}} \left(H_\varphi + H_r \frac{d\delta}{d\varphi}\right). \tag{3}$$

The deviation from the nominal radius a_0 is assumed to be small and smooth:

$$\delta \ll 1 \qquad \text{and} \qquad \frac{d\delta}{d\varphi} \ll 1. \tag{4}$$

Then the electromagnetic field can conveniently be represented as a perturbation of the field in the round guide of radius a_0 :

$$\begin{aligned} E &= E_0 + e, \\ H &= H_0 + h. \end{aligned} \tag{5}$$

Furthermore the fields at $r = a$ can be written in terms of the fields at $r = a_0$:

$$E_0(a,\varphi) = E_0(a_0\,,\varphi) + a_0\delta(\varphi)\,\frac{\partial E_0(a_0\,,\varphi)}{\partial r}. \tag{6}$$

If the unperturbed field is of circular electric form with $E_{0z} = E_{0r} = H_{0\varphi} = 0$, then, upon substituting from (5) into the boundary condition (2), the Taylor series (6) can be used. The perturbation field can then be written in terms of the unperturbed field of the circular electric wave:

$$e_\varphi = -a_0\delta(\varphi)\,\frac{\partial E_{0\varphi}(a_0)}{\partial r}. \tag{7}$$

The boundary condition (3) imposes an additional requirement on the perturbation field

$$e_z = -Zh_\varphi(a_0). \tag{8}$$

Conditions (7) and (8) suffice to calculate the complete perturbation field.

A circular electric wave that carries unit power in positive z direction has an electric field:

$$E_{0\varphi} = -\sqrt{\frac{2\omega\mu}{\pi\beta_0}}\,\frac{J_1(\chi_0 r)}{aJ_0(k_0)}\,e^{-j\beta_0 z}, \tag{9}$$

where

$$k_0 = \chi_0 a_0\,, \qquad J_1(k_0) = 0$$

and

$$\beta_0^2 = \omega^2\mu\epsilon - \chi_0^2.$$

Here, μ and ϵ are permeability and permittivity of the waveguide interior; ω is the angular frequency.

The perturbation δ of the nominal radius is a periodic function of φ. A Fourier expansion is therefore in order:

$$\delta(\varphi) = \sum_p \delta_p \cos p\varphi. \tag{10}$$

Terms with $\sin p\varphi$ have been omitted from (10). They would only add identical perturbations with different polarization. Substituting from (9) and (10) into (7):

$$e_\varphi(a_0) = \sqrt{\frac{2\omega\mu}{\pi\beta_0}}\, \chi_0 e^{-j\beta_0 z} \sum_p \delta_p \cos p\varphi. \tag{11}$$

The expression suggests an expansion of the perturbation fields into terms which individually satisfy Maxwell's equations and have the φ and z dependence of the terms in (11). Such a field is obtained from wave functions

$$\begin{aligned} T_{(p)} &= \sum_p a_{(p)} J_p(\chi_0 r) \sin p\varphi e^{-j\beta_0 z}, \\ T_{[p]} &= \sum_p a_{[p]} J_p(\chi_0 r) \cos p\varphi e^{-j\beta_0 z} \end{aligned} \tag{12}$$

and the following formulae:

$$\begin{aligned} e_r &= -\frac{\beta_0}{\omega\epsilon}\frac{\partial T_{(p)}}{\partial r} - \frac{\partial T_{[p]}}{r\partial\varphi}, \\ e_\varphi &= -\frac{\beta_0}{\omega\epsilon}\frac{\partial T_{(p)}}{r\partial\varphi} + \frac{\partial T_{[p]}}{\partial r}, \\ e_z &= \frac{\chi_0^2}{j\omega\epsilon} T_{(p)}, \\ h_r &= \frac{\partial T_{(p)}}{r\partial\varphi} - \frac{\beta_0}{\omega\mu}\frac{\partial T_{[p]}}{\partial r}, \\ h_\varphi &= -\frac{\partial T_{(p)}}{\partial r} - \frac{\beta_0}{\omega\mu}\frac{\partial T_{[p]}}{r\partial\varphi}, \\ h_z &= \frac{\chi_0^2}{j\omega\mu} T_{[p]}. \end{aligned} \tag{13}$$

Equating $e_\varphi(a_0)$ from (13) with $e_\varphi(a_0)$ from (11) and comparing in this equation the coefficients of $\cos p\varphi$, a relation between $a_{(p)}$, $a_{[p]}$ and δ_p is obtained:

$$-\frac{\beta_0}{\omega\epsilon}\frac{p}{k_0} J_p(k_0)a_{(p)} + J_p'(k_0)a_{[p]} = \sqrt{\frac{2}{\pi}\frac{\omega\mu}{\beta_0}}\, \delta_p. \tag{14}$$

Another relation between $a_{(p)}$ and $a_{[p]}$ is obtained by substituting for e_z and h_φ from (13) into (8):

$$\frac{\chi_0^2}{j\omega\epsilon} J_p(k_0)a_{(p)} = Z\left[\chi_0 J_p'(k_0)a_{(p)} - \frac{\beta_0}{\omega\mu}\frac{p}{a_0} J_p(k_0)a_{[p]}\right]. \tag{15}$$

Equations (14) and (15) can be solved for $a_{(p)}$ and $a_{[p]}$. For example:

$$a_{(p)} =$$

$$-\sqrt{\frac{2\beta_0}{\pi\omega\mu}}\frac{Z\delta_p}{J_p(k_0)} \frac{p}{\dfrac{k_0^2}{j\omega\epsilon a_0}\dfrac{J_p(k_0)}{J_p(k_0)} - k_0 Z\left[\dfrac{J_p'^2(k_0)}{J_p^2(k_0)} - \dfrac{p^2}{k_0^2} - \dfrac{\beta_0^2}{\omega^2\mu\epsilon}\right]}. \tag{16}$$

With $a_{(p)}$ and $a_{[p]}$ the perturbation fields of circular electric waves are known as functions of the δ_p's. Thus the quasi-circular electric waves in any slightly deformed round waveguide can be written in terms of the normal wave and perturbation fields.

The propagation constant remains unchanged and equal to $j\beta_0$ in this first-order approximation. Now it is just the effect of a deformation on the propagation constant and especially on its real part, the attenuation constant, which is most important. Ordinarily a higher order of approximation would be necessary to determine this attenuation. But here, as in all electromagnetic problems where the dissipated energy is small compared to the stored or propagated energy, the losses may be calculated from a lower order of approximation.[7] The attenuation constant is the ratio of power P_d dissipated per unit length to the power carried by the wave:

$$\alpha = \frac{P_d}{2P}.$$

Power is dissipated by the perturbation field through the anisotropic shield into the wall impedance Z:

$$P_d = \tfrac{1}{2}\,\mathrm{Re}\left(\frac{1}{Z}\right)\int_s e_z e_z^*\, ds\,|_{r=a}.$$

This integral along the actual inner radius of the guide is to first order equal to the integral along the nominal radius a_0 :

$$P_d = \frac{\mathrm{Re}\,(Z)}{2\,|\,Z\,|^2}\int_0^{2\pi} e_z e_z^* a_0\, d\varphi. \tag{17}$$

In (9) the power flow of the circular electric wave was assumed to be unity. Substituting for e_z from (13) into (17) and using (16), it is found that each Fourier component of the mechanical deformation contributes α_p to the total loss α:

$$\alpha = \sum_p \alpha_p ,$$

where

$$\alpha_p = \tfrac{1}{2}\,\mathrm{Re}\,(Z)\,\frac{\beta_0}{\omega\mu}\,\frac{\left[p\delta_p\,\dfrac{J_p(k_0)}{J'_p(k_0)}\right]^2}{\left|1 + jZ\,\dfrac{\omega\epsilon a_0}{k_0}\left[\dfrac{\beta_0^{\,2}}{\omega^2\mu\epsilon}\,\dfrac{p^2}{k_0^{\,2}}\,\dfrac{J_p(k_0)}{J'_p(k_0)} - \dfrac{J'_p(k_0)}{J_p(k_0)}\right]\right|^2}. \tag{18}$$

This expression for the added circular electric wave attenuation in a deformed helix waveguide agrees with some obvious facts: Any deformation of a purely reactive wall does not cause any circular electric wave attenuation. δ_0 and δ_1 represent changes in diameter and transverse displacement, respectively, of an otherwise round guide. The circular electric wave configuration is not changed by them. Consequently $\alpha_0 = \alpha_1 = 0$.

Equation (18) is valid for but one special case. The absolute value in the denominator is zero whenever the characteristic equation (61) (of Appendix A) for helix waveguide modes of pth azimuthal order is satisfied by k_0. Whenever a mode of pth azimuthal order has the same propagation constant as the circular electric wave, δ_p, however small it may be, causes a substantial change of the normal circular electric mode that can no longer be described by the perturbation expression of (18).

The propagation constant of any of the asymmetric modes, to be equal to $j\beta_0$, requires a purely reactive wall impedance. Because of finite loss, practical wall impedance values will always be at least slightly resistive; (18) will therefore be valid for all practical cases.

For some typical cross-sectional deviations, (18) can be simplified:

δ_2 represents an elliptical deformation:

$$\alpha_2 a_0 = \tfrac{1}{2}\,\mathrm{Re}\,(Z)\,\frac{\beta_0}{\omega\mu}\,\frac{k_0^{\,2}\delta_2^{\,2}}{\left|1 + j\,\dfrac{2Z}{\omega\mu a_0}\right|^2}; \tag{19}$$

δ_3 represents a trifoil deformation:

$$\alpha_3 a_0 = \tfrac{1}{2}\,\mathrm{Re}\,(Z)\,\frac{\beta_0}{\omega\mu}\,\frac{k_0^{\,2}\delta_3^{\,2}}{\left|\dfrac{k_0^{\,2}}{12} - 1 + j\,\tfrac{1}{2}Z\omega\epsilon a_0\left(1 - \dfrac{k_0^{\,2}}{24} - \dfrac{6}{\omega^2\sqrt{\mu\epsilon}\,a_0^{\,2}}\right)\right|^2}; \tag{20}$$

δ_4 represents a quadrufoil deformation:

$$\alpha_4 a_0 = \frac{1}{2}\,\mathrm{Re}\,(Z)\,\frac{\beta_0}{\omega\mu}\,\frac{k_0^2\delta_4^2}{4\left|\dfrac{k_0^2-12}{24-k_0^2}+j8Z\,\dfrac{\omega\epsilon a_0}{k_0^2}\left[\dfrac{\beta_0^2}{4\omega^2\mu\epsilon}-\left(\dfrac{k_0^2-12}{24-k_0^2}\right)\right]\right|^2}; \tag{21}$$

etc., for any multifoil deformation.

III. NONUNIFORM HELIX WAVEGUIDE

Here the relative deformation δ of the guide radius will not only be a function of φ but it will also change with z. In Appendix A Maxwell's equations are converted into generalized telegraphist's for this structure.

The deformation δ is first assumed to be independent of z. The fields in the deformed but cylindrical waveguide are represented in terms of normal modes of the perfectly round helix waveguide. This series representation for the field components is then substituted into Maxwell's equations. With the boundary conditions (2) and (3) and an orthogonality relation between normal modes of the helix waveguide, a set of simultaneous first-order differential equations is obtained, which determines the z-dependence of the coefficients of this series expansion. If the coefficients are chosen so that they represent amplitudes A and B of forward and backward traveling waves of the round guide modes, then the system of equations for the A's and B's can be written as

$$\begin{aligned}\frac{dA_m}{dz}+jh_mA_m &= -j\sum_n c_{nm}(A_n+B_n),\\ \frac{dB_m}{dz}-jh_mB_m &= +j\sum_n c_{nm}(A_n+B_n).\end{aligned} \tag{22}$$

If the perturbation δ of the nominal radius is expanded into a Fourier series (10), then the coupling coefficients are determined by the coefficients of this Fourier expansion:

$$\begin{aligned} p\neq 0: \qquad c_{[0m][pn]} &= \frac{\sqrt{\pi}}{2}\,N_n\sqrt{\frac{h_{pn}}{h_{0m}}}\,\frac{k_{0m}k_{pn}}{ka_0^2}\,p\,\frac{J_p^2(k_{pn})}{J_p'(k_{pn})}\,\delta_p\,;\\ p=0: \qquad c_{[0m][0n]} &= \frac{k_{0m}k_{0n}}{a_0^2\sqrt{h_{0m}h_{0n}}}\,\delta_0. \end{aligned} \tag{23}$$

The metallic waveguide is the limiting case of the helix waveguide with zero wall impedance. The normal modes of the helix waveguide degenerate into TE_{pn} and TM_{pn}. The separation constant: $k_{pn}=\chi_{pn}a_0$ is the root of $J_p'(k_{pn}) = 0$ for TE_{pn} modes and the root of $J_p(k_{pn}) = 0$ for TM_{pn} modes. The coupling coefficients (23) reduce to $c = 0$ for interaction between TE_{0m} and TM_{pn} modes. For interaction between TE_{0m} and TE_{pn} modes the coupling coefficients are:

$$c_{[0m][pn]} = \frac{k_{0m}k_{pn}}{a_0^2\sqrt{2h_{0m}h_{pn}}} \frac{k_{pn}}{\sqrt{k_{pn}^2 - p^2}} \delta_p . \tag{24}$$

In a nonuniform helix waveguide the coupling coefficients c in (22) are functions of z. Then (22) is a system of first-order linear differential equations with varying coefficients. For small deformations and consequently small coupling coefficients, solutions of (22) can be found by successive approximations. To simplify the representation, the B's of (22) are included in the A's. There are then always pairs of A's associated with propagation constants jh_m and $-jh_m$ and coupling coefficients jc_{nm} and $-jc_{nm}$. Thus the two equations of (22) can be replaced by the first alone. The transformation

$$A_m = e^{-jh_m z} E_m \tag{25}$$

eliminates a common propagation factor:

$$\frac{dE_m}{dz} = -j \sum_n c_{nm} e^{-j(h_n - h_m)z} E_n . \tag{26}$$

The only initial conditions of practical interest are

$$E_1(0) = 1,$$
$$E_n(0) = 0 \qquad \text{for } n \neq 1.$$

A TE_{01} wave of unit amplitude is launched into a nonuniformly deformed helix waveguide. A first-order solution of (26) under these initial conditions is:

$$\begin{aligned} E_1(z) &= 1, \\ E_n(z) &= -j \int_0^z c_{1n} e^{-j(h_1 - h_n)s} \, ds. \end{aligned} \tag{27}$$

The first-order solution is substituted into (26) for a second-order solution:

$$E_1(z) = 1 - \sum_n \int_0^z c_{n1} e^{-j(h_n - h_1)s} \int_0^s c_{1n} e^{-j(h_1 - h_n)t} \, dt \, ds, \tag{28}$$

and so on.

As a typical example, a TE_{01} wave will be launched into a waveguide that has a constant deformation δ between $z = 0 \cdots l$ and is round everywhere else. The waveguide is thus uniform except for two discontinuities at $z = 0$ and $z = l$. The wave amplitudes at any point $z > l$ are, from (27) and (28),

$$E_1(z) = 1 - j \sum_n \frac{c_{1n}^2}{(h_1 - h_n)^2} [(h_1 - h_n)l + j(e^{j(h_1 - h_n)l} - 1)], \quad (29)$$

$$E_n(z) = \frac{c_{1n}}{h_1 - h_n} (e^{-j(h_1 - h_n)l} - 1). \quad (30)$$

The converted wave amplitudes E_n may be regarded as being generated from the TE$_{01}$ wave at the two discontinuities $z = 0$ and $z = l$. Then the conversion at one such discontinuity is:

$$\left| \frac{E_n}{E_1} \right| = \left| \frac{c_{1n}}{h_1 - h_n} \right|. \quad (31)$$

From (23) and (31), with $\delta = \delta_0$, a formula for mode conversion between circular electric waves at diameter changes is obtained:

$$\frac{E_{0n}}{E_{0m}} = \frac{k_{0m} k_{0n}}{a_0^2 \sqrt{h_{0m} h_{0n}} \, (h_{0m} - h_{0n})} \delta_0 . \quad (32)$$

Likewise, a formula for mode conversion in offsets of helix waveguide with $\delta = \delta_1 \cos \varphi$ can be written down. In the case of $Z = 0$, the formula describes mode conversion at offsets of a metallic guide:

$$\frac{E_{1n}}{E_{0m}} = \frac{k_{0m} k_{1n}}{a_0^2 \sqrt{2 h_{0m} h_{1n}} \, (h_{0m} - h_{1n})} \frac{k_{1n}}{\sqrt{k_{1n}^2 - 1}} \delta_1 . \quad (33)$$

Thus, from (31), mode conversion at an arbitrary discontinuity in helix waveguide can be calculated.

Mode conversion at an arbitrary nonuniform deformation of the helix waveguide, however, is found from (27).

IV. TOLERANCES OF HELIX WAVEGUIDE FOR CIRCULAR ELECTRIC WAVE TRANSMISSION

The all-important question may be asked now: What deformations can be tolerated in a helix waveguide without any excessive degradation of the TE$_{01}$ transmission characteristics? There are two factors which degrade the TE$_{01}$ transmission: (a) Additional normal mode loss in a deformed helix waveguide, as calculated in Section II and described by (18), increases the overall TE$_{01}$ transmission loss. (b) Mode conversion and reconversion in nonuniform sections of helix waveguide, as calculated in Section III and described by (27) and (28), cause mode conversion loss and reconversion distortion of the TE_{01} characteristic.

4.1 *Normal Mode Loss*

The normal mode loss of a uniformly deformed waveguide will be considered first. Helix waveguide in current experimental use at the Bell Telephone Laboratories has a nominal inner radius of $a_0 = 1$ inch. A median frequency of the planned operating range is 55.5 kmc. To optimize various transmission characteristics, the surrounding jacket has been made to present a real wall impedance to the interior that is half of free space impedance $Z = \frac{1}{2}\sqrt{\mu/\epsilon}$. For these values, expressions (19), (20), (21) for the added circular electric wave loss have been evaluated:

$$\begin{aligned} \alpha_2 a_0 &= 3.64\ \delta_2^2, \\ \alpha_3 a_0 &= 0.458\ \delta_3^2, \\ \alpha_4 a_0 &= 0.516\ \delta_4^2. \end{aligned} \tag{34}$$

By far the largest losses are caused by an elliptical deformation. The theoretical loss of TE_{01} in a perfect copper waveguide of 2-inch inside diameter at 55.5 kmc is

$$\alpha_0 a_0 = 2.77 \times 10^{-6}.$$

In order that the increase of attenuation be not more than 10 per cent of this theoretical loss, the elliptical deformation should be

$$\delta_2 < 0.276 \times 10^{-3}.$$

The elliptical diameter differences in a 2-inch helix waveguide should not exceed 1 mil. This is quite a strict requirement.

It is interesting to compare these figures with losses in a deformed metallic waveguide:

$$\begin{aligned} \frac{\alpha_2}{\alpha_0} &= \frac{\beta_0^2 a_0^2}{2}\,\delta_2^2, \\ \frac{\alpha_3}{\alpha_0} &= 10\beta_0^2 a_0^2 \delta_3^2, \\ \frac{\alpha_4}{\alpha_0} &= 1.5\beta_0^2 a_0^2 \delta_4^2. \end{aligned} \tag{35}$$

In a metallic waveguide it is the trifoil deformation which causes most loss. In order that such a trifoil deformation not cause more than 10 per cent of the theoretical TE$_{01}$ loss in a 2-inch metallic waveguide at 55.5 kmc, this deformation should be:

$$\delta_3 < 3.42 \times 10^{-3}.$$

4.2 *Mode Conversion Loss*

Equations (34) and (35) describe the added TE$_{01}$ loss correctly only in a waveguide with uniform, z-independent deformation δ. When the deformation is a function of z, as is the case in an imperfect waveguide, the general expression (28) describes the transmission. Changing the order of integration in (28), a more suitable form is obtained:

$$E_1(z) = 1 - \sum_n \int_0^z e^{j(h_1-h_n)u}\, du \int_0^{z-u} c_{1n}(s)c_{1n}(s+u)\, ds. \quad (36)$$

The loss can be expressed in terms of the geometrical imperfections δ with $c_{1n} = C_n\delta$. For sufficiently small δ,

$$| E_1 | = 1 - \Lambda,$$

with the loss

$$\Lambda = \sum_n \int_0^z e^{\Delta\alpha_n u}(P_n \cos \Delta\beta_n u - Q_n \sin \Delta\beta_n u)\, du \cdot \int_0^{z-u} \delta(s)\delta(s+u)\, ds, \quad (37)$$

where

$$C_n^2 = P_n + jQ_n$$

and

$$j(h_1 - h_n) = \Delta\alpha_n + j\Delta\beta_n\, .$$

In general, the geometric imperfections will not be known, only their statistical properties. Rowe and Warters[5] have determined with a relation like (36) the statistics of the loss in terms of the statistics of the guide imperfections. Use of their analysis is made here.

The deformation is assumed to be a stationary random process with covariance $R(u)$ and spectral distribution $S(\zeta)$

$$R(u) = \langle\delta(z)\delta(z+u)\rangle, \quad (38)$$

$$S(\zeta) = \int_{-\infty}^{+\infty} R(u)e^{-j2\pi\zeta u}\, du. \quad (39)$$

In (38), $\langle x\rangle$ is the expected value of x.

Taking the expected value on both sides of (37), the average added loss is obtained in terms of the covariance $R(u)$ is

$$\langle \Lambda \rangle = \sum_n \int_0^z e^{\Delta \alpha u} R(u)(z - u)(P_n \cos \Delta\beta_n u - Q_n \sin \Delta\beta_n u)\, du. \tag{40}$$

For the following analysis, a special form for the covariance must be assumed. Since existing experimental information is rather vague, Rowe[6] assumes $R(u)$ to be exponential as reasonable physically and to simplify the calculation

$$R(u) = \frac{\pi S_0}{L_0} e^{-2\pi|u|/L_0}. \tag{41}$$

Then the spectral distribution of δ becomes

$$S(\zeta) = \frac{S_0}{1 + (L_0\zeta)^2}, \tag{42}$$

where $S(\zeta)$ is nearly flat with spectral density S_0 for mechanical frequencies in distance smaller than

$$\zeta_0 = \frac{1}{L_0}. \tag{43}$$

At ζ_0 the spectral distribution is down 3 db and falls very rapidly above ζ_0 ; L_0 may be regarded as the cutoff mechanical wavelength.

Substituting (41) for the covariance in (40) and performing the integration over a length $z \gg L_0$, the average added loss is:

$$\langle \Lambda \rangle = \pi S_0 z \sum_n \frac{P_n(2\pi - \Delta\alpha_n L_0) - Q_n \Delta\beta_n L_0}{\Delta\beta_n{}^2 L_0^2 + (2\pi - \Delta\alpha_n L_0)^2}. \tag{44}$$

For $\Delta\alpha_n L_0 \gg 2\pi$, (44) reduces to

$$\langle \Lambda \rangle = \frac{\pi S_0 z}{L_0} \sum_n \frac{-P_n \Delta\alpha_n - Q_n \Delta\beta_n}{\Delta\alpha_n{}^2 + \Delta\beta_n{}^2},$$

or with

$$\frac{\pi S_0}{L} = R(0) = \langle \delta^2(z) \rangle$$

the added average loss is for this special case:

$$\frac{\langle \Lambda \rangle}{z} = \langle \delta^2 \rangle \sum_n \operatorname{Re}\left[\frac{-C_n{}^2}{j(h_1 - h_n)}\right]. \tag{45}$$

As seen from (29), a long waveguide with a uniform deformation $\delta =$

$\sqrt{\langle\delta^2\rangle}$ would have the same added loss. Equation (45) then is the added normal mode loss. It is also much simpler than that described by (18). But only when the differential loss $\Delta\alpha$ of every single coupled mode is very large in the cutoff mechanical wavelength L_0 will the added loss be described by (18) with $\delta = \sqrt{\langle\delta^2\rangle}$.

The L_0 for waveguide deformation is probably small, certainly not much larger than 1 foot. Certain coupled modes might have a very high differential loss per foot, but then there would always be coupled modes with low differential loss.

Consequently, the condition leading to (45) is not satisfied for cross-sectional deformation in helix waveguide. Expressions (34) cannot be used to determine cross-sectional tolerances. As shown by Rowe,[6] this conclusion is true for a wide class of covariance functions.

The correct expression for mode conversion in helix waveguide is (44). Written as added loss per wavelength, it reads:

$$\frac{\langle\Lambda\rangle}{z} = \langle\delta^2\rangle L_0 \sum_n \frac{P_n(2\pi - \Delta\alpha_n L_0) - Q_n \Delta\beta_n L_0}{\Delta\beta_n^{\,2} L_0^{\,2} + (2\pi - \Delta\alpha_n L_0)^2}. \qquad (46)$$

For real coupling coefficients in a lossless structure, (46) reduces to

$$\frac{\langle\Lambda\rangle}{z} = \langle\delta^2\rangle L_0 \sum_n \frac{2\pi C_n^{\,2}}{4\pi^2 + \Delta\beta_n^{\,2} L_0^{\,2}}, \qquad (47)$$

and for very small L_0 from (46)

$$\frac{\langle\Lambda\rangle}{z} = \langle\delta^2\rangle \frac{L_0}{2\pi} \sum_n P_n. \qquad (48)$$

For a very short correlation distance, however, a more general expression than (48) for the average added loss can be found. In this case $R(u)$, whatever function it may be, has substantial values only in the immediate vicinity of $u = 0$. Then, instead of (40),

$$\langle\Lambda\rangle = z \int_0^z R(u)\, du \sum_n P_n$$

and, with (39),

$$\frac{\langle\Lambda\rangle}{z} = \tfrac{1}{2} S(0) \sum_n P_n \qquad (49)$$

for any spectral distribution $S(\zeta)$ of geometric imperfections with small correlation distance.

Equation (47) has been evaluated in Appendix B for cross-sectional deformations in a helix waveguide with an infinitely high wall impedance.

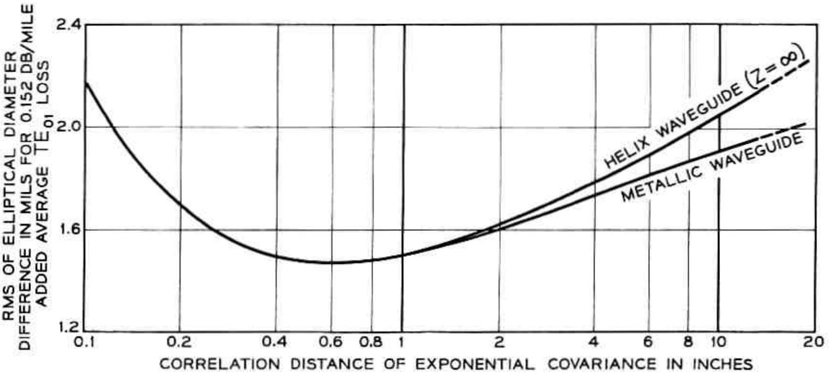


Fig. 1 — TE_{01} loss in round waveguide with random ellipticity; 2 inch inside diameter, at 55.5 kmc.

This particular helix waveguide design minimizes circular electric wave loss and mode conversion in bends.[2] In Fig. 1 is plotted the average ellipticity $\sqrt{\langle\delta_2^2\rangle}$ as a function of the correlation distance L_0 for an additional average loss equal to 10 per cent of the TE_{01} loss in a perfect copper pipe. For comparison, the same curve is plotted for $Z = 0$ representing metallic waveguide.

Both curves coincide for small values of the correlation distance and differ only slighly over the practical range of L_0. Though Fig. 1 is only drawn for a particular helix waveguide and a particular set of covariance functions, it is fairly safe to generalize: Random ellipticity of the cross section causes nearly as much average circular electric wave loss in helix waveguide as it does in metallic waveguide.

A more exact statement has been made for the case of vanishing correlation.[8] When L_0 is small enough for (48) to be valid, the average added TE_{01} loss is independent of the wall impedance and the same as in metallic waveguide.

Manufacturing imperfections usually have a small correlation distance. Therefore helix waveguide has to be manufactured to as close cross-sectional tolerances as metallic waveguide for circular electric wave transmission.

V. CONCLUSIONS

Cross-sectional deformations of the helix waveguide perturb circular electric wave propagation. In a slightly but uniformly deformed helix waveguide circular electric waves propagate with slightly changed field pattern. Power is dissipated into the helix jacket. Consequently, the added circular electric wave loss in a uniformly deformed helix waveguide is considerably larger than it is in a copper waveguide of the same uniform deformation.

Nonuniform deformations cause mode conversion and added TE_{01} loss. Manufacturing imperfections are expected to be random deformations with small correlation distance. Such imperfections increase the average circular electric wave loss nearly independently of the wall impedance which the helix jacket presents to the waveguide interior. The average added loss is therefore nearly the same as it is in metallic waveguide with the same imperfections. For example, ellipticity was assumed to be a stationary random process along the guide with exponential covariance. Then, even at a correlation distance of 1 foot, the added average TE_{01} loss at 55.5 kmc in a 2-inch inside diameter helix waveguide of infinite wall impedance is only 16 per cent smaller than it is in metallic waveguide.

VI. ACKNOWLEDGMENT

I am indebted to H. E. Rowe for many helpful discussions. He especially called my attention to equation (44) and its significance.

APPENDIX A

Generalized Telegraphist's Equations for Deformed Helix-Waveguide

Maxwell's equations in cylindrical coordinates (r,φ,z) are:

$$\frac{1}{r}\frac{\partial E_z}{\partial \varphi} - \frac{\partial E_\varphi}{\partial z} = -j\omega\mu H_r\,, \tag{50}$$

$$\frac{\partial E_r}{\partial z} - \frac{\partial E_z}{\partial r} = -j\omega\mu H_\varphi\,, \tag{51}$$

$$\frac{1}{r}\frac{\partial (rE_\varphi)}{\partial r} - \frac{1}{r}\frac{\partial E_r}{\partial \varphi} = -j\omega\mu H_z\,, \tag{52}$$

$$\frac{1}{r}\frac{\partial H_z}{\partial z} - \frac{\partial H_\varphi}{\partial z} = j\omega\epsilon E_r\,, \tag{53}$$

$$\frac{\partial H_r}{\partial z} - \frac{\partial H_z}{\partial r} = j\omega\epsilon E_\varphi\,, \tag{54}$$

$$\frac{1}{r}\frac{\partial (rH_\varphi)}{\partial r} - \frac{1}{r}\frac{\partial H_r}{\partial \varphi} = j\omega\epsilon E_z\,. \tag{55}$$

The electromagnetic field in the helix waveguide can be derived from two sets of wave functions T_n and T'_n given by

$$\begin{aligned} T_n &= N_n J_p(\chi_n r) \sin p\varphi, \\ T'_n &= N_n J_p(\chi_n r) \cos p\varphi. \end{aligned} \tag{56}$$

The T_n and T'_n satisfy the wave equation

$$\frac{1}{r}\left[\frac{\partial}{\partial r}\left(r\frac{\partial T}{\partial r}\right) + \frac{\partial}{\partial \varphi}\left(\frac{1}{r}\frac{\partial T}{\partial \varphi}\right)\right] = -\chi^2 T, \tag{57}$$

where χ is a separation constant which takes on discrete values for the various normal modes. The transverse field components are written in terms of these functions:

$$\begin{aligned} E_r &= \sum_n V_n \left(\frac{\partial T_n}{\partial r} + d_n \frac{\partial T'_n}{r\partial\varphi}\right), \\ E_\varphi &= \sum_n V_n \left(\frac{\partial T_n}{r\partial\varphi} - d_n \frac{\partial T'_n}{r\partial\varphi}\right), \\ H_r &= \sum_n - I_n \left(\frac{\partial T_n}{r\partial\varphi} - d_n \frac{h_n^2}{k^2}\frac{\partial T'_n}{\partial r}\right), \\ H_\varphi &= \sum_n I_n \left(\frac{\partial T_n}{\partial r} + d_n \frac{h_n^2}{k^2}\frac{\partial T'_n}{r\partial\varphi}\right). \end{aligned} \tag{58}$$

Substituting from (58) into (55) and taking advantage of (57), an expression for the longitudinal electric field is obtained:

$$E_z = j\omega\mu \sum_n I_n \frac{\chi_n^2}{k^2} T_n , \tag{59}$$

where k is the intrinsic propagation constant of the waveguide interior; d_n and the propagation constant h_n are chosen so that the boundary conditions of the round helix waveguide

$$\begin{aligned} E_\varphi(a_0) &= 0, \\ E_z(a_0) &= -ZH_\varphi(a_0) \end{aligned}$$

are satisfied by the individual terms of (58). Only then do the individual terms of (58) represent normal modes of the helix waveguide.

From $E_\varphi(a_0) = 0$:

$$d_n = \left.\frac{\dfrac{\partial T_n}{r\partial\varphi}}{\dfrac{\partial T'_n}{\partial r}}\right|_{a_0} = \frac{pJ_p(k_n)}{k_n J'_p(k_n)}, \tag{60}$$

where $k_n = \chi_n a_0$. The prime at the Bessel function denotes differentiation with respect to the argument. The remaining boundary condition between E_z and H_φ leads to the following (characteristic) equation:

$$\frac{1}{k_n}\frac{J'_p(k_n)}{J_p(k_n)} - \frac{p^2 h_n^2}{k_n^3 k^2}\frac{J_p(k_n)}{J'_p(k_n)} = \frac{-j}{\omega\epsilon a_0 Z}. \tag{61}$$

The characteristic equation, together with

$$k_n^2 = (k^2 - h_n^2)a_0^2,$$

determines the separation constant k_n. The transverse field components of any two different modes are orthogonal to each other in that:

$$\begin{aligned}\frac{1}{V_n I_m}\int_S (E_{tn} \times H_{tm})\, dS = \\ \int_S \frac{\epsilon}{\epsilon_0}\left[\left(\frac{\partial T'_n}{\partial r} + d_n \frac{\partial T'_n}{r\partial\varphi}\right)\left(\frac{\partial T_m}{\partial r} + d_m \frac{h_m^2}{k^2}\frac{\partial T'_m}{r\partial\varphi}\right)\right. \\ \left. + \left(\frac{\partial T_n}{r\partial\varphi} - d_n \frac{\partial T'_n}{\partial r}\right)\left(\frac{\partial T_m}{r\partial\varphi} - d_m \frac{h_m^2}{k^2}\frac{\partial T'_m}{\partial r}\right)\right] dS = \delta_{nm},\end{aligned} \tag{62}$$

where δ_{nm} is the Kronecker symbol. The integration is to be extended over the cross section of the waveguide. For $n = m$ equation (62) determines the normalization factor:

$$\begin{aligned}N_n = \frac{\sqrt{2}}{\sqrt{\pi}\, J_p(k_n)}\left[\frac{h_n^2}{k^2} p^2(k_n^2 - p^2)Y_n^2 + \frac{1}{Y_n^2}\right. \\ \left. + k_n^2\left(1 - \frac{p^2}{k^2 a_0^2}\right) + 2\left(\frac{1}{Y_n} - p^2 Y_n\right)\right]^{-\frac{1}{2}},\end{aligned} \tag{63}$$

with

$$Y_n = \frac{J_p(k_n)}{k_n J'_p(k_n)}.$$

All quantities in (56) and (58) have now been determined except the current and voltage coefficients. To find relations for them the field components from (58) are substituted into Maxwell's equations and these then are converted to generalized telegraphist's equations.

Add

$$-\left(\frac{\partial T_m}{r\partial\varphi} - d_m \frac{h_m^2}{k^2}\frac{\partial T'_m}{\partial r}\right)$$

times (50) and

$$\frac{\partial T_m}{\partial r} + d_m \frac{h_m^2}{k^2}\frac{\partial T'_m}{r\partial\varphi}$$

times (51) and integrate over the cross section. The result is:

$$\begin{aligned}\frac{dV_m}{dz} + j\,\frac{h_m^{\,2}}{\omega\epsilon}\,I_m = \\ &\int_S (\text{grad } E_z)(\text{grad } T_m)\,dS + d_m\,\frac{h_m^{\,2}}{k^2}\int_S (\text{grad } E_z)(\text{flux } T_m')\,dS \\ &- j\omega\mu\sum_n I_n\,\frac{\chi_n^{\,2}}{k^2}\biggl[\int_S (\text{grad } T_n)(\text{grad } T_m)\,dS \\ &\qquad + d_m\,\frac{h_m^{\,2}}{k^2}\int_S (\text{grad } T_n)(\text{flux } T_m')\,dS\biggr],\end{aligned} \tag{64}$$

where the gradient and flux of a scalar are defined by:

$$\begin{aligned}\text{grad}_r\, T &= \frac{\partial T}{\partial r}, & \text{grad}_\varphi\, T &= \frac{1}{r}\frac{\partial T}{\partial \varphi}, \\ \text{flux}_r\, T &= \frac{1}{r}\frac{\partial T}{\partial \varphi}, & \text{flux}_\varphi\, T &= -\frac{\partial T}{\partial r}.\end{aligned} \tag{65}$$

After partial integration on the right-hand side of (64),

$$\begin{aligned}\frac{dV_m}{dz} + j\,\frac{\mathrm{h}_m^2}{\omega\epsilon}\,I_m = \\ &\int_0^{2\pi} E_z\left(\frac{\partial T_m}{\partial r} + \frac{d_m}{a_0}\frac{h_m^{\,2}}{k^2}\frac{\partial T_m'}{\partial\varphi}\right) a_0\,d\varphi + \chi_m^{\,2}\int_S E_z T_m\,dS \\ &- j\omega\mu\sum_n I_n\,\frac{\chi_n^{\,2}}{k^2}\biggl[\int_0^{2\pi} T_n\left(\frac{\partial T_m}{\partial r} + \frac{d_m}{a_0}\frac{h_m^{\,2}}{k^2}\frac{\partial T_m'}{\partial\varphi}\right) a_0\,d\varphi \\ &\qquad + \chi_m^{\,2}\int_S T_n T_m\,dS\biggr].\end{aligned} \tag{66}$$

In special cases when the helix waveguide degenerates into a perfectly conducting metallic waveguide, the individual terms for E_z in (59) are zero for $r = a_0$, while E_z itself, because of the boundary condition (3), is different from zero. Then (59) is a nonuniformly convergent series, which describes E_z only in the open interval $0 \leqq r < a_0$. Term-by-term differentiation will make the series diverge. Therefore the series had not been substituted for E_z in (64). In (66), (59) may now be substituted in the integral over the cross section. In the line integral, E_z from the boundary condition (3) may be substituted. The fields at $r = a$ can by a Taylor series be written in terms of fields at $r = a_0$. Neglecting higher-order terms:

$$E_z(a_0) = -Z\left[H_\varphi(a_0) + \frac{\partial H_\varphi(a_0)}{\partial r}\, a_0\delta + H_r(a_0)\,\frac{d\delta}{d\varphi}\right] - \frac{\partial E_z(a_0)}{\partial r}\, a_0\delta. \tag{67}$$

Thus, instead of (66):

$$\frac{dV_m}{dz} + j\,\frac{h_m^{\;2}}{\omega\epsilon}\, I_m = -a_0 \int_0^{2\pi} \left[a_0\delta\,\frac{\partial}{\partial r}(E_z + ZH_\varphi) + ZH_r\,\frac{\mathrm{d}\delta}{\mathrm{d}\varphi}\right]\left[\frac{\partial T_m}{\partial r} + \frac{d_m}{a_0}\,\frac{h_m^{\;2}}{k^2}\,\frac{\partial T'_m}{\partial\varphi}\right] d\varphi. \tag{68}$$

For the other of the two sets of generalized telegraphist's equations, add

$$-\left(\frac{\partial T_m}{\partial r} + \frac{d_m}{r}\,\frac{\partial T'_m}{\partial\varphi}\right)$$

times (53) and

$$-\left(\frac{1}{r}\,\frac{\partial T_m}{\partial\varphi} - d_m\,\frac{\partial \mathrm{T}'_m}{\partial r}\right)$$

times (54) and integrate over the cross section. The result is:

$$\begin{aligned}\frac{dI_m}{dz} + j\omega\epsilon V_m = {} & -\int_S (\text{grad } H_z)(\text{flux } T_m)\, dS + d_m \int_S (\text{grad } H_z)(\text{grad } T'_m)\, dS \\ & + j\omega\epsilon \sum_n V_n d_n\, \frac{\chi_n^{\;2}}{k^2} \\ & \quad \cdot \int_S [(\text{grad } T'_n)(\text{flux } T_m) - d_m(\text{grad } T'_n)(\text{grad } T'_m)]\, dS.\end{aligned} \tag{69}$$

After partial integration on the right-hand side of (69),

$$\frac{dI_m}{dz} + j\omega\epsilon\, V_m = d_m\chi_m^{\;2} \int_S H_z T'_m\, dS - j\omega\epsilon \sum_n V_n d_n d_m\, \frac{\chi_n^{\;2}\chi_m^{\;2}}{k^2} \int_S T'_n T'_m\, dS. \tag{70}$$

To replace H_z, substitute E_r from (58), in (52), multiply (52) by T'_m and integrate over the cross section. The series (58) for E_φ is nonuniformly convergent and cannot be used in (52). After partial integration,

$$-j\omega\mu \int_S H_z T_m' \, dS = \int_0^{2\pi} E_\varphi T_m' a_0 \partial\varphi + \sum_n V_n d_n \chi_n^2 \int_S T_n' T_m' \, dS. \quad (71)$$

With the boundary condition (2) as Taylor series at $r = a_0$:

$$E_\varphi(a_0) = -E_r(a_0) \frac{d\delta}{d\varphi} - \frac{\partial E_\varphi(a_0)}{\partial r} a_0 \delta.$$

Equation (70) can be written as:

$$\frac{dI_m}{dz} + j\omega\epsilon V_m = -j \frac{d_m \chi_m^2}{\omega\mu} a_0 \int_0^{2\pi} \left(E_r \frac{d\delta}{d\varphi} + \frac{\partial E_\varphi}{\partial r} a_0 \delta \right) T_m' \, d\varphi. \quad (72)$$

Partial integration on the right-hand side,

$$\int_0^{2\pi} E_r \frac{d\delta}{d\varphi} T_m' \, d\varphi = -\int_0^{2\pi} \delta \left(\frac{\partial E_r}{\partial\varphi} T_m' + E_r \frac{\partial T_m'}{\partial\varphi} \right) d\varphi,$$

and substitution of the series expressions (58),

$$-\frac{\partial E_r}{\partial\varphi} + a_0 \frac{\partial E_\varphi}{\partial r} = \sum_n V_n d_n \chi_n^2 a_0 T_n' ,$$

reduces (72) to

$$\begin{aligned} \frac{dI_m}{dz} + j\omega\epsilon V_m = -j \frac{a_0^2}{\omega\mu} \sum_n V_n d_n d_m \chi_n^2 \chi_m^2 \int_0^{2\pi} T_n' T_m' \delta \, d\varphi \\ + j \frac{a_0}{\omega\mu} d_m \chi_m^2 \int_0^{2\pi} E_r \frac{\partial T_m'}{\partial\varphi} \delta \, d\varphi. \end{aligned} \quad (73)$$

The interest is limited here to the propagation characteristics of circular electric waves. Therefore, only terms that describe direct interaction between circular electric and other waves need to be retained in (68) and (73). When V_m and I_m are voltage and current amplitudes of circular electric waves, then T_m and $\partial T_m'/\partial\varphi$, and consequently the right-hand side of (68) and the last term on the right-hand side of (73), are zero. When V_m and I_m are amplitudes of other modes, then the same terms in (68) and (73) are zero, since E_z , H_φ , E_r and $H_r(a_0)$ are zero for circular electric waves. Thus (68) and (73) reduce to:

$$\begin{aligned} \frac{dV_m}{dz} + j \frac{h_m^2}{\omega\epsilon} I_m = 0, \\ \frac{dI_m}{dz} + j\omega\epsilon V_m = -j \sum_n V_n d_n d_m \frac{k_n^2 k_m^2}{\omega\mu a_0^2} \int_0^{2\pi} T_n' T_m' \delta \, d\varphi. \end{aligned} \quad (74)$$

The generalized telegraphist's equations represent an infinite set of coupled transmission lines. It is convenient to write transmission line equations not in terms of currents and voltages but in terms of the amplitudes of forward and backward traveling waves. Thus, let A and B be the amplitudes of the forward and backward waves of a typical mode at a certain cross section. The mode current and voltage are related to the mode amplitudes by

$$\begin{aligned} V &= \sqrt{K}(A + B), \\ I &= \frac{1}{\sqrt{K}}(A - B), \end{aligned} \tag{75}$$

where K is the wave impedance

$$K_m = \frac{h_m}{\omega\epsilon}. \tag{76}$$

If the currents and voltages in the generalized telegraphist's equations (74) are represented in terms of the traveling-wave amplitudes, after some obvious additions and subtractions the following equations for coupled traveling waves are obtained:

$$\begin{aligned} \frac{dA_m}{dz} + jh_m A_m &= -j \sum_n c_{nm}(A_n + B_n), \\ \frac{dB_m}{dz} - jh_m B_m &= +j \sum_n c_{nm}(A_n + B_n). \end{aligned} \tag{77}$$

The c's are coupling coefficients defined by:

$$c_{nm} = \tfrac{1}{2}\sqrt{h_n h_m}\, d_n d_m \frac{k_n^2 k_m^2}{k^2 a^2} \int_0^{2\pi} T'_n T'_m \delta \, d\varphi. \tag{78}$$

To replace the d's and T's in (78), the customary double-subscript notation for the various modes in round helix waveguide is used. Then from (66), (70) and (73) the interaction between circular electric waves and other waves in deformed helix waveguide is described by the coupling coefficients:

$$\begin{aligned} p \neq 0: \quad c_{[0m][pn]} &= \\ & N_n \sqrt{\frac{h_{pn}}{h_{0m}}} \frac{k_{0m}k_{pn}}{ka_0^2} \frac{p}{2\sqrt{\pi}} \frac{J_p^2(k_{pn})}{J'_p(k_{pn})} \int_0^{2\pi} \delta \cos p\varphi \, d\varphi, \\ p = 0: \quad c_{[0m][0n]} &= \frac{k_{0m}k_{0n}}{a_0^2\sqrt{h_{0m}h_{0n}}} \frac{1}{2\pi} \int_0^{2\pi} \delta \, d\varphi. \end{aligned} \tag{79}$$

APPENDIX B

Nonuniform Helix Waveguide with Infinite Wall Impedance

For $Z \to \infty$ the characteristic equation (61) reduces to

$$Y_n = \pm \frac{k}{ph_n} \tag{80}$$

or the two equations:

$$\begin{aligned} \frac{k_n}{p} \frac{J_{p+1}(k_n)}{J_p(k_n)} &= 1 - \sqrt{1 - \frac{k_n^2}{k^2a_0^2}}, \\ \frac{k_n}{p} \frac{J_{p-1}(k_n)}{J_p(k_n)} &= 1 - \sqrt{1 - \frac{k_n^2}{k^2a_0^2}}. \end{aligned} \tag{81}$$

For $k_n < ka$, an approximation for the roots of (81) is furnished by

$$\begin{aligned} J_{p+1}(k_n) &= 0, \\ J_{p-1}(k_n) &= 0. \end{aligned} \tag{82}$$

Equation (81) can be expanded about the roots of (82) to improve the approximations for k_n.

Substituting (80) for Y_n in (63) reduces the normalization factor to:

$$N_n = \frac{1}{\sqrt{\pi}\, k_n J_p(k_n)} \left(1 - \frac{p^2}{k^2a_0^2} \mp \frac{p}{kh_na_0^2}\right)^{-\frac{1}{2}}. \tag{83}$$

Hence the coupling coefficient is, from (23),

$$c_{[0m][pn]} = \pm \frac{1}{2} \frac{k_{0m}k_{pn}}{\sqrt{h_{0m}h_{pn}}\, a_0} \left(1 - \frac{p^2}{k^2a_0^2} \mp \frac{p}{kh_na_0^2}\right)^{-\frac{1}{2}} \frac{\delta_p}{a_0}. \tag{84}$$

In (84) all the subscripts have been included to identify the coupling coefficient properly.

REFERENCES

1. Morgan, S. P. and Young, J. A., Helix Waveguide, B.S.T.J., **35,** 1956, p. 1347.
2. Unger, H. G., Helix Waveguide Theory and Application, B.S.T.J., **37,** 1958, p. 1599.
3. Schelkunoff, S. A., Conversion of Maxwell's Equations into Generalized Telegraphist's Equations, B.S.T.J., **34,** 1955, p. 995.
4. Morgan, S. P., Mode Conversion Losses in Transmission of Circular Electric Waves through Slightly Non-Cylindrical Guides, J. App. Phys. **21,** 1950, p. 329.
5. Rowe, H. E. and Warters, W. D., Transmission Deviations in Waveguide Due to Mode Conversion: Theory and Experiment, Proc. I.E.E., **106,** Part B, Supp. No. 13, 1959, p. 30.
6. Rowe, H. E., to be published.
7. Marcuse, D., Attenuation of the TE_{01} Wave Within the Curved Helix Waveguide, B.S.T.J., **37,** 1958, p. 1649.
8. Unger, H. G., Mode Conversion in Metallic and Helix Waveguide, to be published.

Normal Modes and Mode Conversion in Helix Waveguide

By H. G. UNGER

(Manuscript received August 9, 1960)

Helix waveguide, composed of closely wound insulated copper wire covered with an absorptive or reactive jacket, transmits circular electric waves with low loss. Mechanical imperfections, such as curvature and deformation, cause coupling between the circular electric waves and unwanted modes and degrade the transmission. In designing a helix waveguide for a particular application, a jacket must be found that minimizes the transmission degradation. Unwanted mode characteristics and their coupling coefficients must be known; these quantities are given by the roots of a transcendental equation involving complex Bessel functions.

A program has been set up for automatically finding the complex roots by iterative approximation. Starting from the known roots at infinite jacket conductivity, the characteristic equation is solved for all practical values of wall impedance of the jacket and all modes of interest. The representation of the mode characteristics as a function of wall impedance leads to a definite designation of modes in heterogeneous waveguide. The TE_{pn} modes of helix waveguide with $n \neq 1$ can have only a limited attenuation. These limits determine the design of mode filters. Manufacturing imperfections increase the average TE_{01} loss independently of the wall impedance. Random curvature with large correlation distance is produced by laying tolerances, but its contribution to the average loss is minimized in a helix waveguide with very large wall impedance.

I. INTRODUCTION

Helix waveguide, closely wound from insulated copper wire and covered with an absorptive or reactive jacket, is a good transmission medium for circular electric waves.[1] In long distance communication, waveguide can be designed to act as a mode filter, to negotiate bends or, particularly, to serve as the transmission line proper.[2]

As in metallic waveguide, the loss of circular electric waves decreases steadily with frequency only in a perfect helix waveguide. Any curvature of the guide axis, deformation of the cross section or deviation of the winding from a low and uniform pitch adds to the loss and degrades the transmission characteristics.[3,4]

In a perfect helix waveguide circular electric waves propagate undisturbed. Imperfections cause coupling between circular electric waves and other modes. Power is lost by conversion to unwanted modes and reconversion distorts any smooth transmission characteristics.

In order to control mode conversion and reconversion in helix waveguide with practical imperfections, and also to design helix waveguides for mode filters and intentional bends, the unwanted mode characteristics and unwanted mode coupling must be investigated. Earlier calculations have resulted in a characteristic equation which implicitly determines the properties of helix waveguide-modes,[1] and also in explicit expressions for various coupling coefficients.[2,3,4] Numerical evaluations of these equations have been very informative. They were, however, not complete enough to reveal all the unwanted mode properties and could not serve as a basis for helix waveguide design in every application.

The results of a more exhaustive numerical evaluation of helix waveguide equations will be presented here. In a few typical examples these results will be applied to helix waveguide design problems. First the equations which describe wave propagation in perfect and imperfect helix waveguide will be listed.

II. PERFECT HELIX WAVEGUIDE

A helix waveguide (Fig. 1) will be called perfect when the helix forms a straight circular cylinder and is wound with a low and uniform pitch. The mathematical model which then replaces it is an anisotropic impedance sheet at radius a conducting perfectly in circumferential direction but with a wall impedance Z in axial direction. The Z replaces the jacket surrounding the helix and takes into account the finite size of helix wires. The electromagnetic field components in a cylindrical coordinate system (r,φ,z) are then subject at $r = a$ to the boundary conditions

$$E_\varphi = 0, \tag{1}$$

$$\frac{E_z}{H_\varphi} = -Z. \tag{2}$$

Solutions of Maxwell's equations in cylindrical coordinates are Bessel

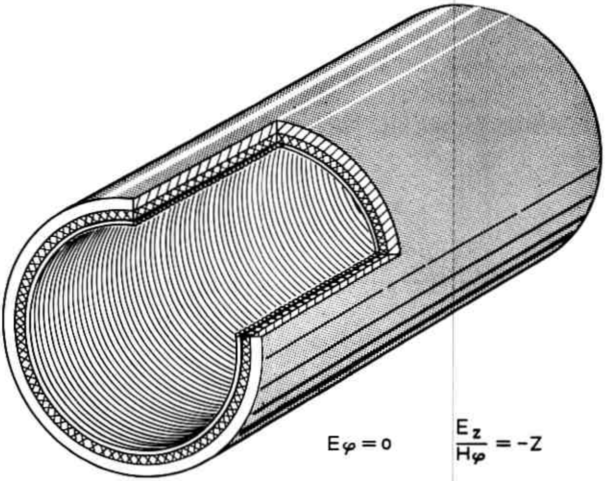


Fig. 1 — Helix waveguide and boundary conditions.

functions J_p of the radius, trigonometric functions of the azimuth and exponential functions of the axial distance:

$$\begin{Bmatrix} J_p\left(\frac{k}{a}r\right) \\ \\ J_p'\left(\frac{k}{a}r\right) \end{Bmatrix} \begin{Bmatrix} \sin p\varphi \\ \\ \cos p\varphi \end{Bmatrix} e^{-\gamma z}, \tag{3}$$

where p is the azimuthal order of the wave. The axial propagation constant γ and radial propagation constant k/a are related with the intrinsic propagation constant $\omega\sqrt{\mu\epsilon}$ of the material filling the waveguide:

$$\left(\frac{k}{a}\right)^2 = \omega^2\mu\epsilon + \gamma^2. \tag{4}$$

When the boundary conditions (1) and (2) are imposed on the solutions (3) of Maxwell's equations the following characteristic equation results:

$$j\omega\epsilon aZ - \frac{kJ_p(k)J_p'(k)}{\frac{p^2}{k^2}\frac{\gamma^2}{\omega^2\mu\epsilon}J_p^2(k) + J_p'^2(k)} = 0. \tag{5}$$

Values of γ that satisfy the characteristic equation (5) are the propagation constants of normal modes of the perfect helix waveguide. They describe wave propagation in a perfect helix waveguide completely.

III. IMPERFECT HELIX WAVEGUIDE

Wave propagation in imperfect helix waveguide has been described by generalized telegraphist's equations.[2,3,4] In a perfect helix waveguide a normal mode n of amplitude $| E_n |$ propagates independently from all other modes m:

$$\frac{dE_n}{dz} = -\gamma_n E_n . \tag{6}$$

Imperfections cause interaction between modes so that the wave amplitudes are mutually coupled:

$$\frac{dE_n}{dz} = -\gamma_n E_n - j \sum_m c_{nm} E_m . \tag{7}$$

The coupling coefficients are determined by the kind and size of the imperfection, but they are also strongly dependent on the wall impedance.

For circular electric wave applications, only coupling between these and other waves is of interest. Coupling coefficients of typical imperfections in helix waveguide will now be listed. The subscript m will refer to the TE_{0m} wave; n will refer to any of the coupled modes. A normalization factor

$$\begin{aligned} N_n &= \frac{\sqrt{2}}{\sqrt{\pi}\, J_p(k_n)} \\ &\cdot \left[\frac{p^2 \gamma_n^2}{\omega^2 \mu \epsilon} (p^2 - k_n^2) Y_n^2 + \frac{1}{Y_n^2} + k_n^2 \left(1 - \frac{p^2}{\omega^2 \mu \epsilon a^2} \right) + 2 \left(\frac{1}{Y_n} - p^2 Y_n \right) \right]^{-\frac{1}{2}} \end{aligned} \tag{8}$$

with

$$Y_n = \frac{J_p(k_n)}{k_n J_p'(k_n)} \tag{9}$$

is used to render the coupling coefficients symmetric, i.e.,

$$c_{nm} = c_{mn} . \tag{10}$$

3.1 *Curvature*[2]

There is only coupling between circular electric modes and modes of first azimuthal order in a curved helix waveguide:

$$c_{nm} = N_n \frac{\sqrt{\pi}\, J_1(k_n)}{2\omega \sqrt{\mu\epsilon}\, a} \sqrt{\frac{\gamma_n}{\gamma_m}} \frac{k_m k_n^2}{k_m^2 - k_n^2} \left[1 + \frac{\gamma_m}{\gamma_n} + \frac{\gamma_m + \gamma_n}{\gamma_m - \gamma_n} Y_n \right] \frac{1}{R}, \tag{11}$$

where R is the radius of curvature.

3.2 *Deformation of the Cross Section*[3]

The radius a_1 of a deformed guide of nominal radius a can be written

$$a_1 = a(1 + \sum_p \delta_p \cos p\varphi). \tag{12}$$

Each component δ_p will cause coupling between circular electric modes and modes of azimuthal order p:

$$c_{mn} = \frac{\sqrt{\pi}}{2} N_n \sqrt{\frac{\gamma_n}{\gamma_m}} \frac{k_m k_n^2}{\omega \sqrt{\mu\epsilon}\, a^2} pJ_p(k_n) Y_n \delta_p . \tag{13}$$

3.3 *Irregular Helix Winding*[4]

In a perfect helix waveguide the angle between a helix wire and the cross section is small enough to be regarded as zero. In an irregular winding this angle can be written

$$\psi = \sum_p \theta_v \sin p\varphi. \tag{14}$$

Each component θ_p causes coupling to modes of azimuthal order p:

$$c_{nm} = \frac{\sqrt{\pi}\, k_m k_n^2 N_n J_p(k_n)}{2\omega\sqrt{\mu\epsilon}\sqrt{\gamma_m \gamma_n}\, a^3} \theta_p . \tag{15}$$

IV. NUMERICAL EVALUATION

The propagation constant of normal modes in helix waveguide is, by (4) and (5), only implicitly given as a function of frequency and waveguide parameters. The problem is to find the complex roots of a transcendental and complex equation.

With (4), γ can be eliminated from (5). Then, for a given frequency and guide radius, the characteristic equation determines k as a function of Z:

$$F(k,Z) = 0. \tag{16}$$

For $Z = 0$ the characteristic equation degenerates into

$$J_p(k) = 0, \qquad J_p'(k) = 0, \tag{17}$$

the roots of which correspond to TM and TE waves respectively of metallic waveguide.

Starting from the known roots of (17) for $Z = 0$, the solutions of (16) for helix waveguide can be traced by gradually increasing the wall impedance. If k_0 and Z_0 are a known solution of (5), then an approximate value for the solution at $Z_1 = Z_0 + \Delta Z$ is given by:

$$k_1 = k_0 - \frac{\dfrac{\partial}{\partial Z}[F(k_0, Z_0)]}{\dfrac{\partial}{\partial k}[F(k_0, Z_0)]} \Delta Z. \tag{18}$$

A better approximation is found by Newton's formula:

$$k_2 = k_1 - \frac{F(k_1, Z_1)}{\dfrac{\partial}{\partial k}[F(k_1, Z_1)]}. \tag{19}$$

For further improvement, the process (19) can be repeated to any desired accuracy.

The final result is the starting point for the next root at the neighboring value of wall impedance. For the numerical evaluation, the wall impedance was related to the impedance of free space $Z_0 = \sqrt{\mu/\epsilon}$:

$$\frac{Z}{Z_0} = \rho e^{j\Phi}. \tag{20}$$

The solutions were traced along lines of constant phase Φ of Z. The increment $\Delta\rho$ was varied and kept sufficiently small to insure continuity of the process.

The evaluation was programmed by Mrs. C. L. Beattie for automatic execution on an IBM 704 Data Processing System.

The characteristic equation was evaluated for all wall impedances with passive phases and amplitudes up to 5000 ohms. All those solutions were traced which for zero wall impedance start as the following metallic waveguide modes:

$$\mathrm{TE}_{11}, \mathrm{TM}_{11}, \mathrm{TE}_{12}, \mathrm{TM}_{12}, \mathrm{TE}_{13}, \mathrm{TM}_{13}.$$
$$\mathrm{TE}_{21}, \mathrm{TM}_{21}, \mathrm{TE}_{22}, \mathrm{TM}_{22}, \mathrm{TE}_{23}.$$
$$\mathrm{TE}_{31}, \mathrm{TM}_{31}, \mathrm{TE}_{32}, \mathrm{TM}_{32}.$$

For some special wall impedance phases the evaluations were extended over many more modes. A value of $a/\lambda = 4.7$ was assumed corresponding to a center frequency of the proposed 35 to 75 kmc frequency band for the 2-inch inside diameter waveguide system.

The numerical results were also used to calculate from the separation constant k, the propagation constant γ in its real and imaginary parts. Figs. 2 through 6 are plotted from these results. These diagrams show

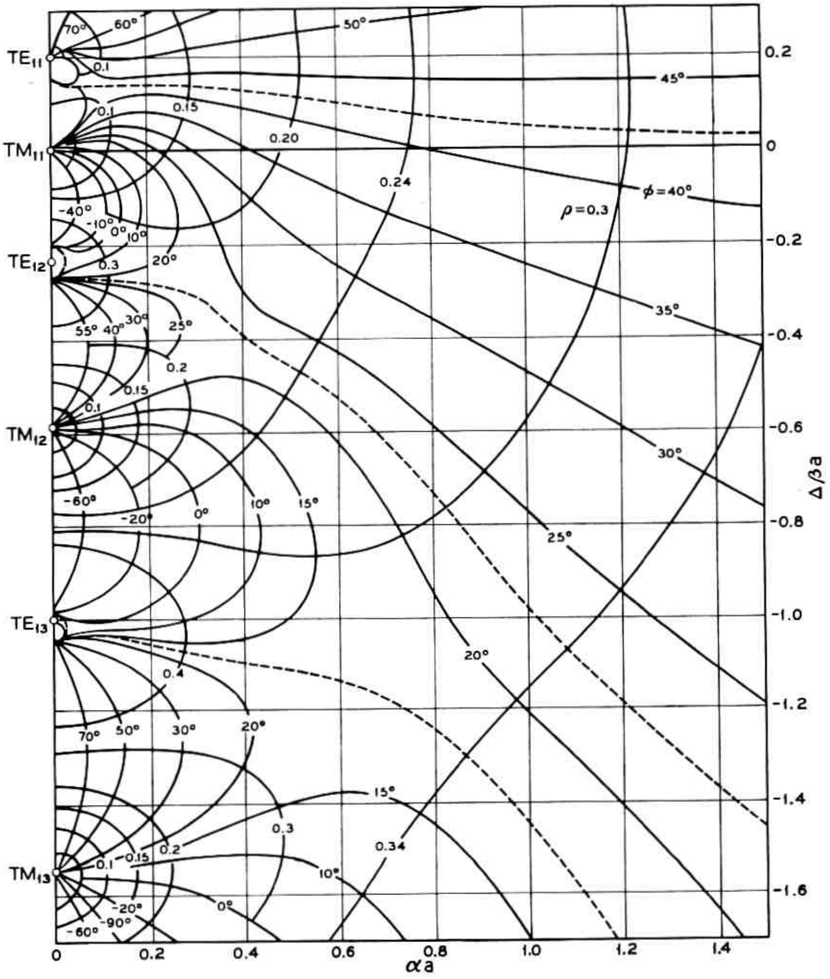


Fig. 2 — Propagation constant $\gamma = \alpha a + j(\beta_{01} + \Delta\beta)a$ in helix waveguide of wall impedance Z; contours in γ-plane of constant magnitude ρ and phase angle Φ of Z/Z_0 ; $a/\lambda = 4.7$, $p = 1$.

contour lines of constant phase Φ and constant amplitude ρ of the wall impedance drawn in the complex plane of propagation constant γ. The scale on the βa-axis has been shifted by the TE$_{01}$ phase $\beta_{01}a = 29.305$ and represents the difference in phase constant between TE$_{01}$ and the plotted mode.

Each diagram is for a particular value of p, specifying the respective

Curves for wall impedances with complex phase fan out into the γ-plane. Some return to the imaginary axis; others continue more and more out to ever increasing values of the attenuation constant.

The propagation constant is a multivalued function of the wall impedance. For any one wall impedance value there are as many different values of the propagation constant as there are points of zero wall im-

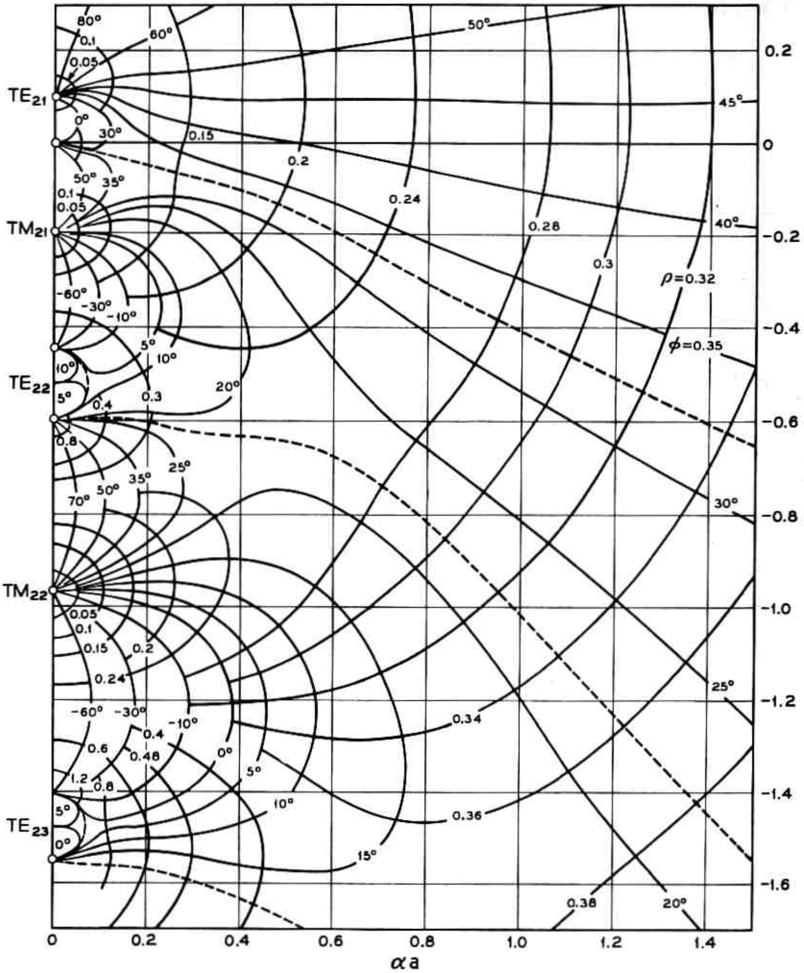


Fig. 5 — Propagation constant $\gamma = \alpha a + j(\beta_{01} + \Delta\beta)a$ in helix waveguide of wall impedance Z; contours in γ-plane of constant magnitude ρ and phase angle φ of Z/Z_0 ; $a/\lambda = 4.7$, $p = 2$.

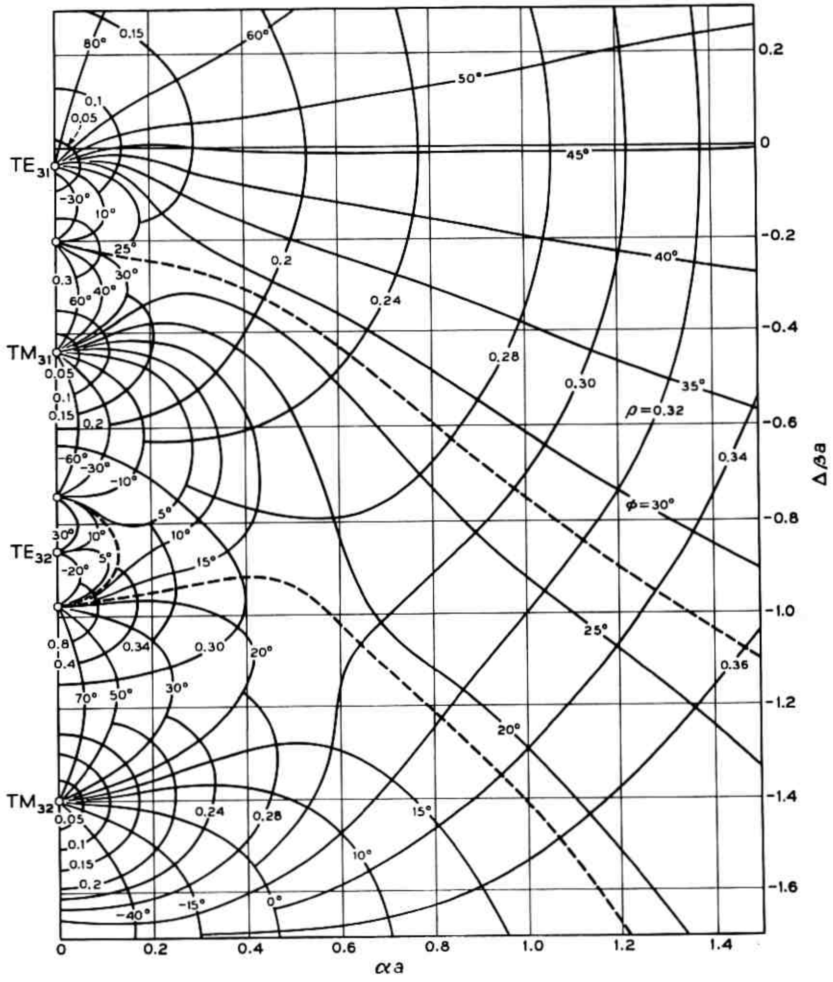


Fig. 6 — Propagation constant $\gamma = \alpha a + j(\beta_{01} + \Delta\beta)a$ in helix waveguide of wall impedance Z; contours in γ-plane of constant magnitude ρ and phase angle φ of Z/Z_0 ; $a/\lambda = 4.7$, $p = 3$.

pedance on the β-axis. Each value of propagation constant corresponds to a normal mode. The designation of these modes is not as simple as in metallic waveguide. The modes of helix waveguide are, in general, neither transverse with respect to any field component nor is their radial order well defined. Therefore, the simple designation of metallic waveguide TE$_{pn}$ or TM$_{pn}$ loses its significance. Nevertheless, the mode designation of metallic waveguide can be extended to helix waveguide or, for that matter, to any heterogeneous waveguide, when the γ-plane is divided up so that in each region the propagation constant is a single-valued function of the critical guide parameter. In the present case the wall impedance is the critical parameter.

The dividing lines in the γ-plane will be branch cuts of γ in the Z-plane. They separate the infinite set of branches of γ from each other, each branch corresponding to a helix waveguide mode. The branch cuts of γ should connect the branch points in the Z-plane. The branch points of γ in the Z-plane are saddle points of Z in the γ-plane. Branch cuts of γ should therefore go through the saddle points of Z in the γ-plane. As many branches will be in contact at the saddle point as is the order of the saddle point. From inspection of the diagrams all saddle points are found to be of second order; therefore, only two branches of γ are in contact at these saddle points and only one dividing line or branch cut must be made through each.

The remaining path of the branch cuts is arbitrary. They should conveniently follow a course that never cuts contour lines of constant phase of the wall impedance and ends either in infinity or on the β-axis at the points of infinite wall impedance.

For example, the branch cut between TE_{11} and TM_{11} starts in Fig. 2 at the corresponding point of infinite wall impedance and separates the contour line ($\Phi = 40°$) coming from TE_{11} from the contour line ($\Phi = 45°$) coming from TM_{11}. Somewhere in the γ-plane the dividing line hits a saddle point of $Z = f(\gamma)$. Beyond this saddle point the branch cut is continued according to the same rule, always separating contour lines of constant phase which originated at different points of zero wall impedance.

Each such region, bounded by the β-axis and the branch cuts (broken lines in the diagrams) is now designated by the metallic waveguide mode located within it. The normal modes of helix waveguide are then defined uniquely, and any further discussions can be made in terms of these modes.

This mode designation in helix waveguide can be defined in fewer words as follows: A mode in helix waveguide of finite wall impedance is identified with the metallic waveguide mode into which it degenerates when the wall impedance phase is kept constant and the wall impedance amplitude made zero.

Modes in any heterogeneous waveguide can correspondingly be identified with metallic waveguide modes when the critical parameters are subjected to the proper limiting process. All critical parameters should be kept constant except that one which in its limit changes the particular heterogeneous waveguide into a metallic waveguide. Requiring the procedure to be most direct will in general eliminate any further ambiguity.

Quite generally it is found that in the γ-plane the regions of all TE_{p1} and TM_{pn} modes in helix waveguide are unbounded while all TE_{pn} modes with $n > 1$ have a bounded region. Thus, the attenuation of all TE_{pn} modes with $n > 1$ is limited and cannot exceed a certain maximum value for any wall impedance. The attenuation constant of any of the other modes can be made arbitrarily high simply by choosing the proper wall impedance.

In most helix waveguide applications unwanted mode loss should be as high as possible. A more detailed discussion of those modes which cannot exceed a certain value of attenuation is therefore in order. A typical mode with limited attenuation is TE_{12} . Fig. 4 shows an enlarged portion of the γ-plane that contains the TE_{12} area. Besides being bounded by the β-axis this area is also bounded by an approximate semicircle as branch cut. The maximum loss of $\alpha a = 0.0363$ for TE_{12} is realized when the wall impedance is chosen

$$\left| \frac{Z}{Z_0} \right| = 0.495 \qquad \text{arc } (Z) = 4.5°, \tag{21}$$

where γ lies on the branch cut at the point of highest α. There is, however, another γ value for this wall impedance on the other side of the branch cut, a γ value that represents a TM_{11} wave. Its real part is $\alpha a = 0.0350$.

For all practical purposes it does not matter which of these points is called TE_{12} and which TM_{11} . The point with lower attenuation α is therefore the decisive one. To render the attenuation of this point as high as possible, it is moved along the branch cut into the saddle point at $\alpha a = 0.0360$. The wall impedance for this condition is

$$\left| \frac{Z}{Z_0} \right| = 0.487 \qquad \text{arc } (Z) = 4.5°. \tag{22}$$

At the same time, the other point moves also into the saddle point, and both modes degenerate into identity.

All other modes with limited attenuation behave similarly.

Using the results for the separation constant k and the propagation constant of helix waveguide modes, the coefficient of curvature coupling between TE_{01} and unwanted modes was computed for modes with first-order ($p = 1$) azimuthal dependence. For the modes of higher order in p the coupling coefficient to TE_{01} in a deformed cross section was computed. For $p = 2$ these coefficients describe coupling in an elliptical pipe. For $p = 3$ it is coupling in a trifoil deformation.

Of greater practical importance are the coefficients of curvature coupling. In Figs. 7 through 12 plots of these coupling-coefficients have been made for the modes TE_{11}, TM_{11}, TE_{12}, TM_{12}. Again, the contour lines of constant phase and the contour lines of constant amplitude of the wall impedance have been plotted as an orthogonal network in the plane of complex coupling coefficient c. Some of the lines of constant phase run out of the diagrams to very large values of c, indicating that the particular coupling coefficient has a pole in their vicinity. Comparison with the propagation constant of the respective modes shows that these poles occur at the saddle points of the $Z = f(\gamma)$ plot. Indeed inspection of (8) and

$$\frac{\partial}{\partial\gamma}\left[F(\gamma,Z)\right]$$

from (16) shows that where $\partial F/\partial\gamma$ is zero and $Z = f(\gamma)$ has a saddle point the normalization factor N_n has a pole.

Poles of the coupling coefficients might cause concern; after all, they represent very strong coupling to unwanted modes. But since the poles coincide with saddle points of $Z = f(\gamma)$ there is always strong coupling to the two degenerate modes at the saddle point. Coupling to each one of these modes is of opposite sign from the other. The total mode conversion stays in quite normal bounds.

It should be recalled on occasions like this that the normal modes of helix waveguide, like modes in any lossy structure, are not orthogonal with respect to power. Suppose, for example, that A_0 is the amplitude normalized with respect to power of a circular electric wave. Then A_0^2 is the power carried by this wave. Let the helix waveguide have a wall impedance near (22). Then the two modes TE_{12} and TM_{11} are nearly degenerate with respect to each other. Curvature will cause coupling as described by (11). Since the coupling coefficients are very large, even a short section of small curvature will generate large amplitudes A_1 of TM_{11} and A_2 of TE_{12}. One of these amplitudes alone, for example A_1, would mean seriously high mode conversion. Since TM_{11} and TE_{12} are not orthogonal with respect to power, both of the amplitudes A_1 and A_2 together compensate each other to a small total effect.

In the plots of Figs. 7 and 9 for TE_{11} and TE_{12} the wall impedance is always a single-valued function of the coupling coefficient. In Figs. 8 and 10 for TM_{11} and TM_{12}, $Z = f(c)$ is multivalued. This observation can be generalized to the following statement: Any coupling coefficient

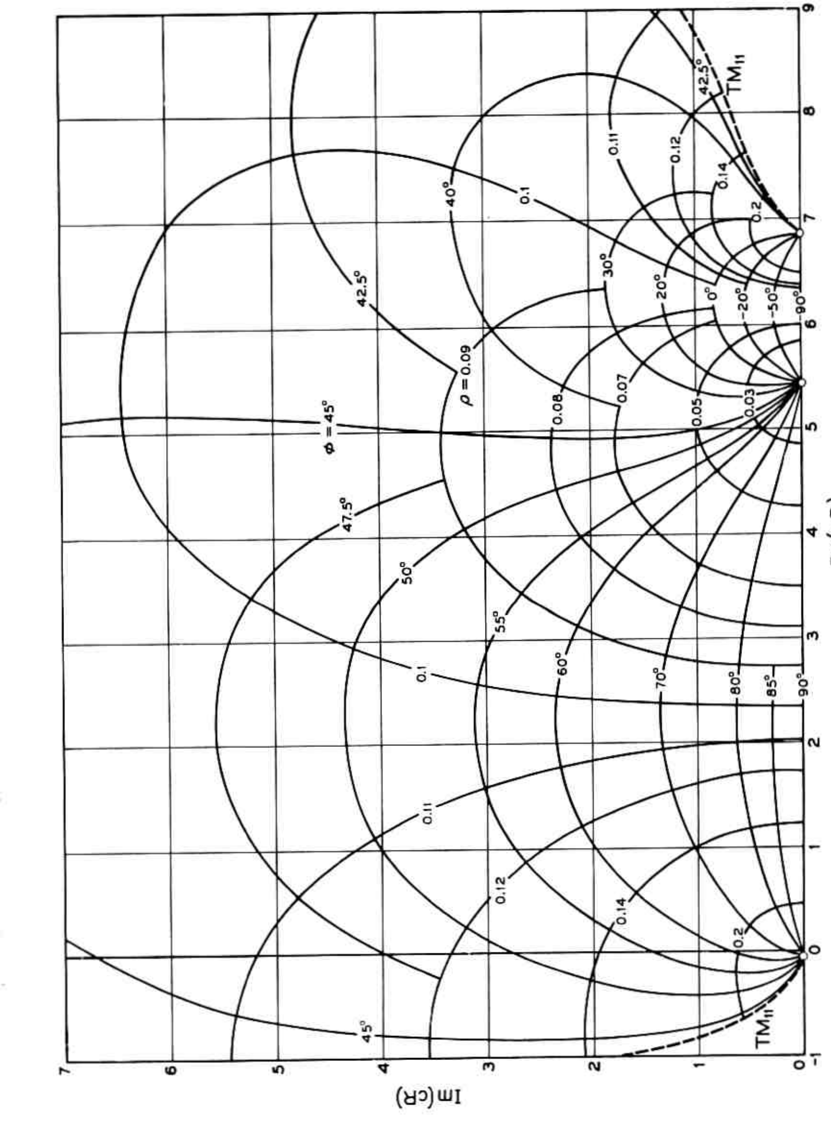


Fig. 7 — Coefficient c of curvature coupling between TE_{01} and TE_{11} in helix waveguide of wall impedance Z. Contours in (cR)-plane of constant magnitude ρ and phase angle Φ of Z/Z_0 ; $a/\lambda = 4.7$.

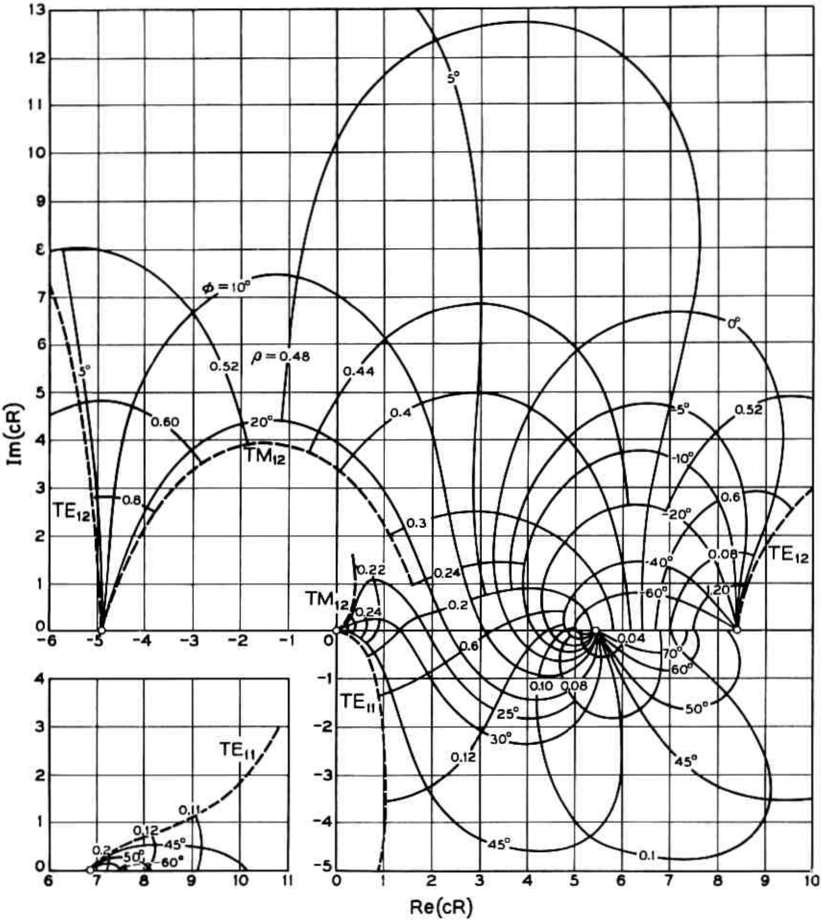


Fig. 8 — Coefficient c of curvature coupling between TE_{01} and TM_{11} in helix waveguide of wall impedance Z. Contours of constant magnitude ρ and phase angle Φ of Z/Z_0 in branches I and II of (cR)-plane; $a/\lambda = 4.7$.

c between circular electric modes and TM modes in helix waveguide of wall impedance Z is a function of Z such that its inversion $Z = f(c)$ is a multivalued function. A sufficient condition for this statement is that $c = g(Z)$ should have more than one pole, for then each of these poles gives a different value $Z = f(c)$ for the same argument $c = \infty$. Inspection of Figs. 2 through 6 shows that the area of every TM mode is adjacent to more than one saddle point of $Z = f(\gamma)$. As stated earlier, a saddle point of $Z = f(\gamma)$ corresponds to a pole of $c = g(Z)$. All TM modes

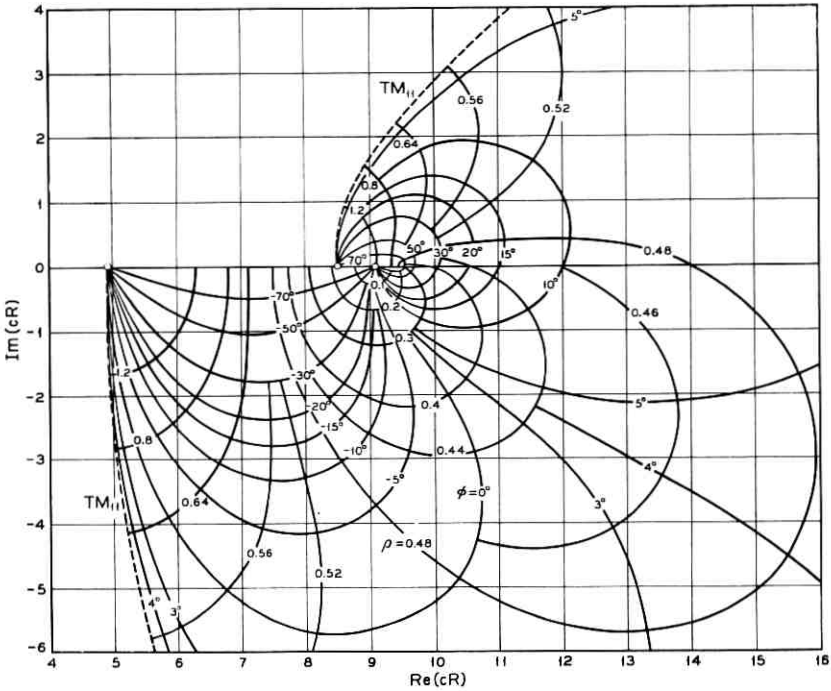


Fig. 9 — Coefficient c of curvature coupling between TE_{01} and TE_{12} in helix waveguide of wall impedance Z. Contours in (cR)-plane of constant magnitude ρ and phase angle Φ of Z/Z_0 ; $a/\lambda = 4.7$.

therefore have more than one pole of $c = g(Z)$ and $Z = f(c)$ is multivalued.

Actually the plots of Figs. 7 and 9 for TE_{11} and TE_{12} might be multivalued too. But when limiting the representation to wall impedance values with positive real part, the plots are single-valued.

To facilitate the representation of the multivalued function $Z = f(c)$ for TM_{11} and TM_{12}, branch cuts have been made in the c-plane and the different branches of c have been plotted in separate planes.

The broken lines indicate the border of a particular mode in the c-plane. They correspond to the branch cuts of γ in Figs. 2 through 6. The adjoining modes are always listed in the corresponding area.

V. APPLICATION

The results of the numerical evaluations have been applied to several problems of helix waveguide design:

5.1 *Mode Filter*

Sections of helix waveguide are inserted at intervals into plain metallic waveguide to absorb unwanted modes. For best absorption of a metallic waveguide mode the attenuation of the corresponding helix waveguide mode should be as high as possible. The most unwanted mode in metallic waveguide is TE_{12}; it most strongly degrades TE_{01} characteristics through mode-conversion effects. A good helix waveguide mode filter should therefore have a wall impedance that makes the attenuation constant of the corresponding TE_{12} mode a maximum. For the present case ($a/\lambda = 4.70$) this wall impedance value is given by (22). As high as the attenuation is for TE_{12} mode for this design, TE_{11} has quite low an attenuation constant

$$\mathrm{TE}_{12}: \qquad \alpha a = 0.0360,$$

$$\mathrm{TE}_{11}: \qquad \alpha a = 0.00686.$$

In metallic waveguide, TE_{11}, although not as objectionable as TE_{12}, is still a serious offender. A mode filter should at least represent moder-

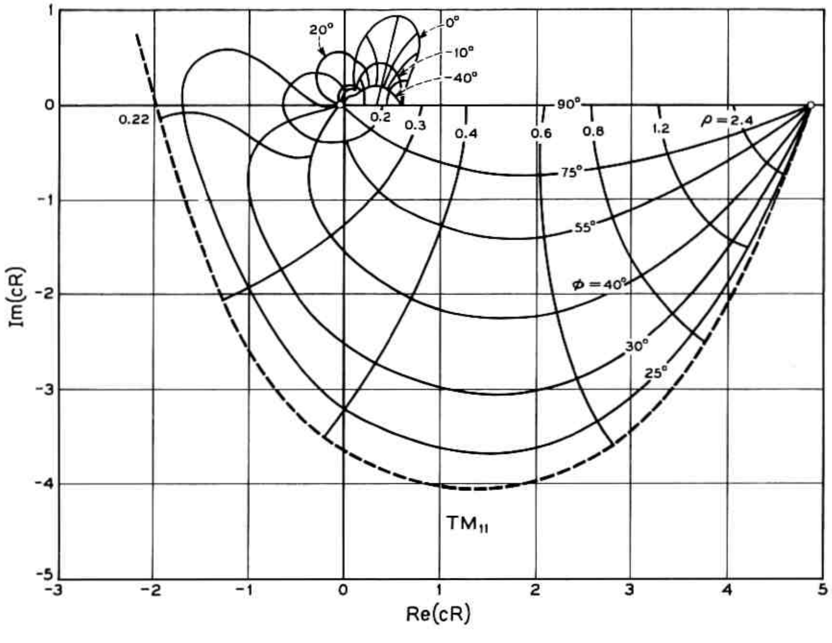


Fig. 10 — Coefficient c of curvature coupling between TE_{01} and TM_{12} in helix waveguide of wall impedance Z, contours of constant magnitude ρ and phase angle Φ of Z/Z_0 in branch I of (cR)-plane; $a/\lambda = 4.7$.

ate absorption to TE$_{11}$. The wall impedance for which the TE$_{11}$ and TE$_{12}$ attenuation are equal and a maximum is

$$\left|\frac{Z}{Z_0}\right| = 0.2975 \qquad \text{arc}\ (Z) = 12.0°, \tag{23}$$

and the corresponding attenuation is:

$$\alpha a = 0.01158.$$

These two wall impedance values are the limits for mode filters. Any practical design will be in between.

5.2 *Random Curvature*

Wave propagation in curved helix waveguide is described by generalized telegraphist's equations as coupling between the modes of the straight guide. For arbitrary but small coupling these equations can be solved approximately. An expression for the added TE$_{01}$ loss can be written in terms of the coupling coefficients and the coupled mode characteristics.

Let the curvature distribution $\kappa(z)$ along the waveguide be a stationary random process with covariance

$$\sigma(u) = \langle\kappa(z)\kappa(z+u)\rangle. \tag{24}$$

According to Rowe[5] (see also Ref. 3), the average added TE$_{01}$ loss can then be expressed in terms of the covariance of the coupling coefficient:

$$\langle\alpha\rangle = \frac{1}{L}\sum_n \int_0^L e^{-\Delta\alpha_n z}\sigma(z)(L-z)(P_n \cos \Delta\beta_n z + Q_n \sin \Delta\beta_n z)\, dz, \tag{25}$$

where L is the length of the line; $(c_nR)^2 = P_n + jQ_n$, the square of the coupling coefficient with c_n from (11); and $\Delta\alpha_n + j\Delta\beta_n = \gamma_n - \gamma_0$, the difference in propagation constant of a coupled mode n to the TE$_{01}$ mode. The summation has to be extended over all coupled modes n.

For a mere estimate of the effects of random curvature the covariance is assumed to be exponential:

$$\sigma(z) = \langle\kappa^2\rangle e^{-2\pi(|z|/L_0)}, \tag{26}$$

where L_0 may be regarded as a correlation distance.

When the correlation distance L_0 is small compared to the total length L of the waveguide, the average added loss is determined by the rms curvature $\sqrt{\langle\kappa^2\rangle}$ and L_0:

$$\langle\alpha\rangle = \langle\kappa^2\rangle L_0 \sum_n \frac{P_n(2\pi + \Delta\alpha_n L_0) + Q_n \Delta\beta_n L_0}{\Delta\beta_n^2 L_0^2 + (2\pi + \Delta\alpha_n L_0)^2}. \tag{27}$$

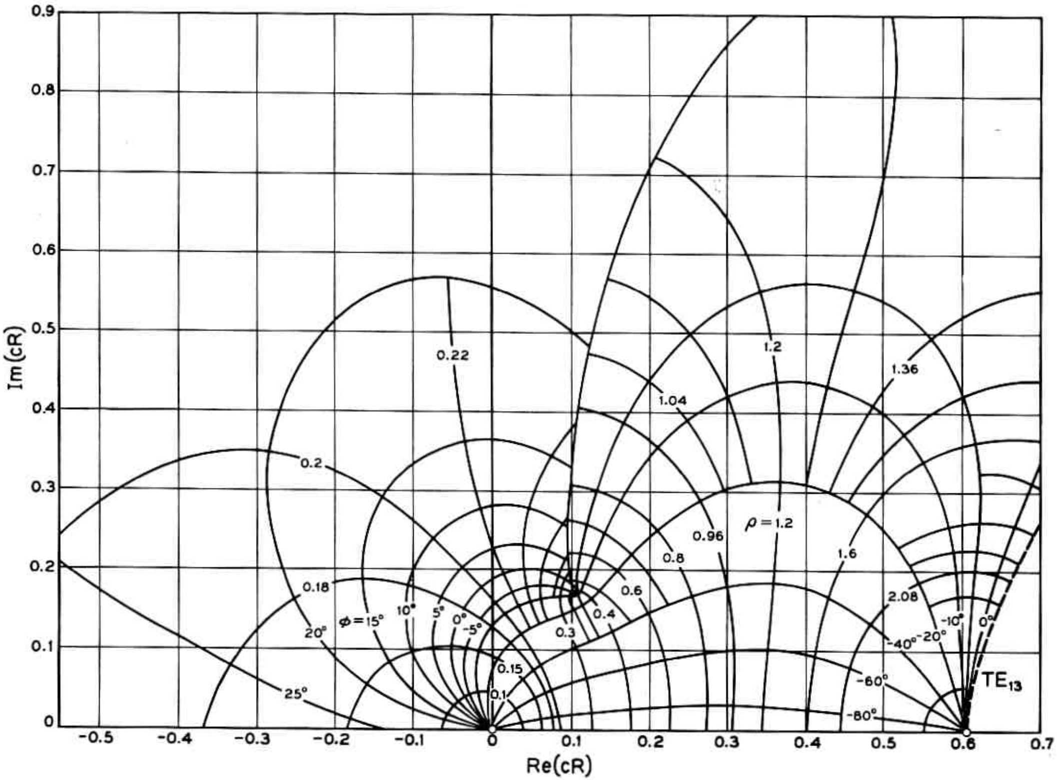


Fig. 11 — Enlarged portion of Fig. 10.

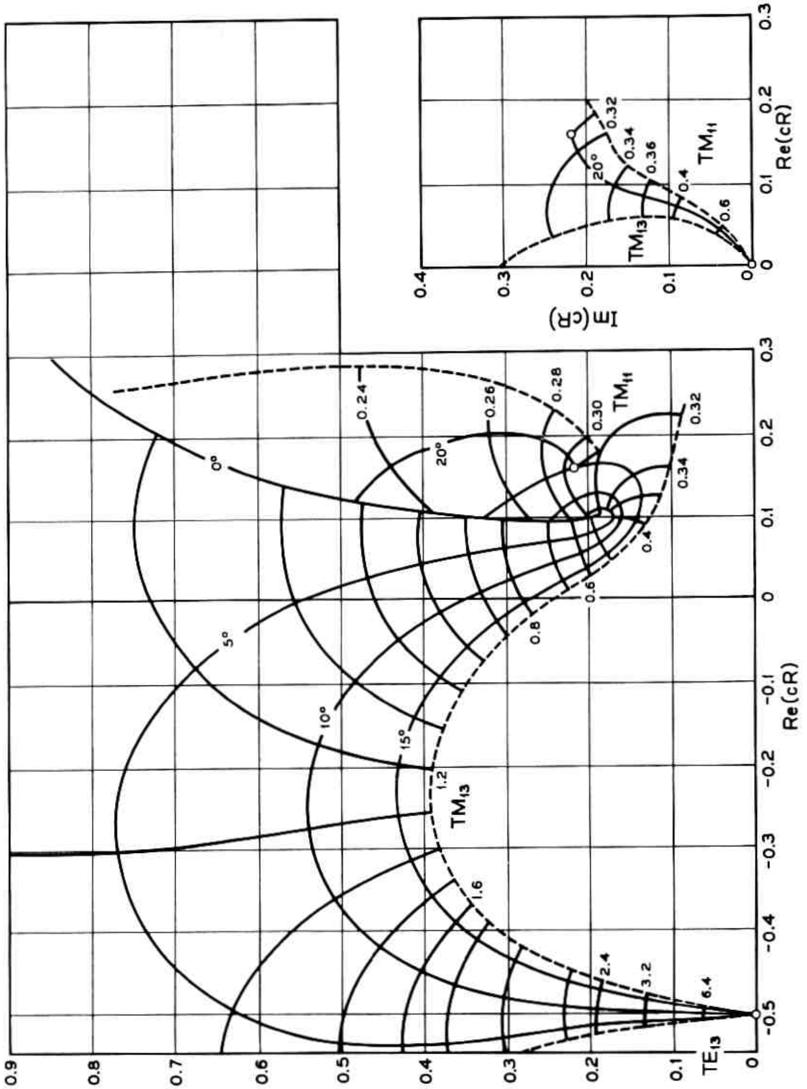


Fig. 12 — Branches II and III of Fig. 10.

Equation (27) has been evaluated for a helix waveguide, the TE_{12} attenuation of which is an absolute maximum and for a helix waveguide with equal and maximum attenuation for TE_{11} and TE_{12}. The results are plotted in Fig. 13. Also plotted in this figure are the corresponding curves for plain metallic waveguide and for helix waveguide with infinite wall impedance. The latter design of helix waveguide minimizes TE_{01} losses in intentional bends.

Shown in Fig. 13 are curves of the rms radius of curvature as a function of correlation distance L_0. This rms value would add 10 per cent of the TE_{01} loss in a perfect copper pipe to the average TE_{01} loss in the respective waveguide.

In calculating the curves of Fig. 13, coupling to the following modes of helix waveguide and metallic waveguide has been taken into account:

$$TE_{11}, \quad TM_{11}, \quad TE_{12}, \quad TM_{12}, \quad TE_{13}, \quad TM_{13}.$$

Contributions from higher-order modes are small enough to be neglected.

One important conclusion can be drawn from Fig. 13. When the correlation distance of random curvature is small enough — smaller than 10 feet in the present case — the added average loss is nearly independent of the wall impedance and nearly the same as in plain metallic waveguide. This independence is not only true for random curvature with exponential covariance but for any random curvature

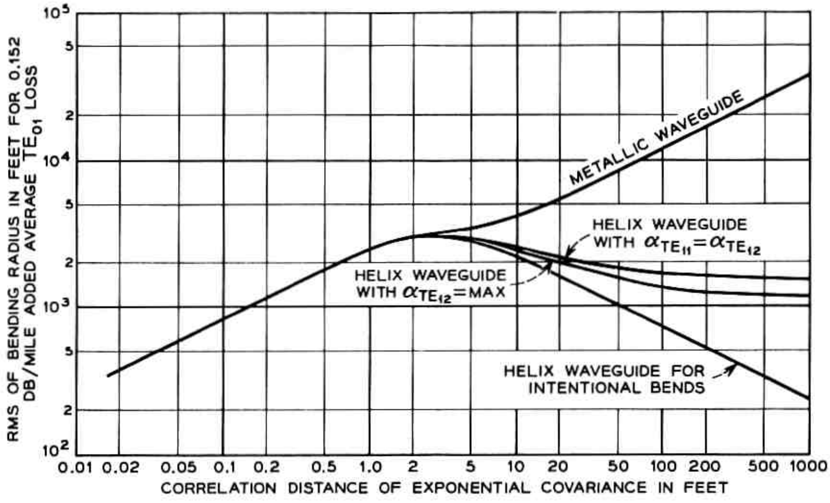


Fig. 13 — TE_{01} loss in round waveguide with random curvature; $a/\lambda = 4.7$.

with sufficiently flat spectral distribution.[6] The curves of Fig. 13 for exponential covariance demonstrate, as a typical example, over what range of correlation distance the average added loss is independent of the particular jacket structure.

Random curvature with a correlation distance smaller than 10 feet can be classified as a manufacturing imperfection. After all, the individual pipe sections which make up the line are usually only 15 feet long. Any particular choice of wall impedance therefore does not relieve the straightness tolerances which should be met in the manufacturing process.

For correlation distances larger than 10 feet the average added loss becomes more and more dependent on the wall impedance. For a specified average loss helix waveguide with infinite wall impedance — for intentional bends — may be bent most strongly. But even a helix waveguide designed optimally as a mode filter — $\alpha_{\text{TE}_{11}} = \alpha_{\text{TE}_{12}}$ or $\alpha_{\text{TE}_{12}} =$ maximum — may be bent much more than plain metallic waveguide.

Random curvature with a correlation distance larger than 10 feet may be classified as a laying imperfection. Its spectral distribution contains mainly mechanical frequencies which correspond to sine waves of 10 feet and more. Such curvature distribution arises from following right of ways or the contour of the landscape or just from not installing the pipe very carefully.

The curves in Fig. 13 have been drawn for a specified average loss. For very large correlation distance they approach asymptotically a constant value. This value corresponds to the normal circular electric mode in the particular helix waveguide with constant curvature. Helix waveguide for intentional bends, since with $Z = \infty$ it is assumed to be lossless, within the limits of the present calculation, may have an arbitrarily small radius of curvature. Uniform curvature causes no loss in this lossless structure. The curve for metallic waveguide goes to infinity. The circular electric mode is not a normal mode of the curved metallic guide.

5.3 *Random Ellipticity*

Wave propagation in elliptical helix waveguide is analyzed in a similar manner to propagation in curved helix waveguide.

Instead of (21) the covariance of the cross sectional deformation

$$\sigma(u) = \langle \delta(z)\delta(z + u) \rangle$$

is introduced and, for $(c_n a/\delta_p)^2 = P_n + jQ_n$, the coupling coefficients

$c_n = c_{n1}$ from (13) of a deformed helix waveguide are substituted. Then the average added TE_{01} loss is given by (25), and for an exponential covariance by (27).

Equation (27) has been evaluated for elliptical deformations of the same waveguides which were analyzed for random curvature before. The result is shown in Fig. 14. The rms of elliptical diameter differences $4(\sqrt{\langle\delta_1^2\rangle})a$, which would add 10 per cent of the TE_{01} loss in a perfect copper pipe to the average TE_{01} loss in the respective waveguide is plotted over the correlation distance L_0. Coupling to all modes which are propagating in the metallic waveguide has been taken into account. For $a/\lambda = 4.70$ there are 17 modes of azimuthal order $p = 2$ propagating. Contributions from higher-order modes are small enough to be neglected.

When the correlation distance is smaller than one foot the average loss is independent of the wall impedance. For larger values of correlation distance the average loss will depend on the wall impedance, but this is hardly of any practical significance. Ellipticity is a typical manufacturing imperfection, and will always have a small correlation distance. For all practical purposes, cross-sectional tolerances in helix

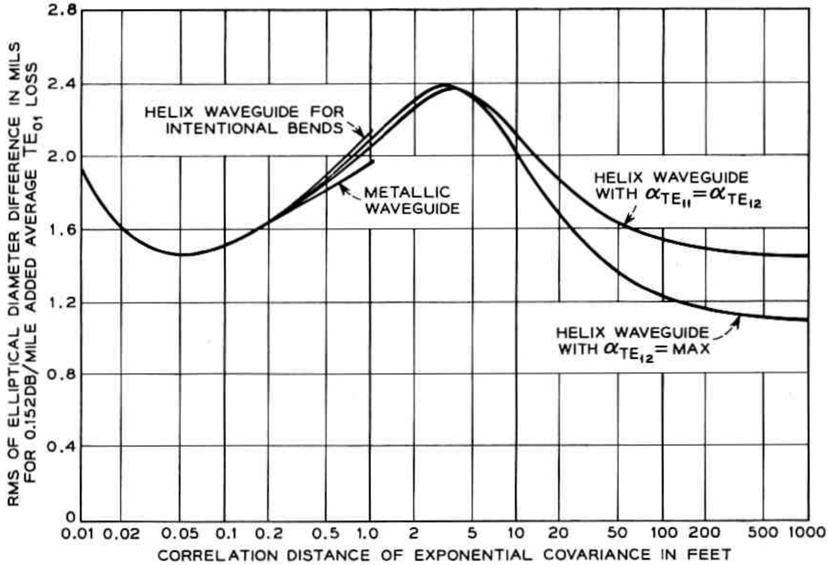


Fig. 14 — TE_{01} loss in round waveguide with random ellipticity; $a/\lambda = 4.7$.

waveguide are independent of the wall impedance and the same as in metallic waveguide.

The curves in Fig. 14 have been drawn for a specified average loss. For very large correlation distance they approach asymptotically a constant value, which corresponds to the normal circular electric mode in the particular helix waveguide with uniform ellipticity.

Metallic waveguide ($Z = 0$) and helix waveguide ($Z = \infty$) — since they are assumed to be lossless — have curves which have a never-leveling slope. Uniform ellipticity causes no loss in these lossless structures.

VI. CONCLUSION

The characteristics of normal modes in helix waveguide can be represented as a function of the wall impedance Z. The propagation constant γ is a multivalued function of the wall impedance, with each value corresponding to a normal mode. But for a specified order of azimuthal dependence the wall impedance is a single-valued function of the propagation constant. The most suitable representation of propagation characteristics of modes in helix waveguide is therefore of contour lines of Z in the γ-plane.

Appropriate branch cuts make γ a single-valued function of Z and lead to a unique mode definition: Any mode of helix waveguide is identified by the mode of metallic waveguide into which it degenerates when the wall impedance phase is kept constant and its amplitude made zero.

The attenuation constant of all TE_{pn} modes with $n \neq 1$ is limited. The attenuation constant of any other mode in helix waveguide can be made arbitrarily high with a proper choice of wall impedance.

Helix waveguide for mode filters should be designed between two extreme rules. One makes the TE_{12} attenuation an absolute maximum and leads to low TE_{11} loss; the other makes TE_{12} and TE_{11} attenuation equal and as high as possible.

Mode conversion between circular electric and other modes in curved or deformed helix waveguide can be calculated from the propagation constants and coupling coefficients of the coupled modes. For random imperfections the added average TE_{01} loss is independent of the wall impedance as long as the correlation distance is small. Manufacturing tolerances for helix waveguide are therefore independent of the particular design.

Laying tolerances produce random curvature of large correlation distance. They depend strongly on the wall impedance. An infinite wall impedance minimizes the average TE_{01} loss in helix waveguide curved randomly in this manner.

REFERENCES

1. Morgan, S. P. and Young, J. A., Helix Waveguide, B.S.T.J., **35,** 1956, p. 1347.
2. Unger, H. G., Helix Waveguide Theory and Application, B.S.T.J., **37,** 1958, p. 1599.
3. Unger, H. G., Noncylindrical Helix Waveguide, this issue, p. 233.
4. Unger, H. G., Winding Tolerances in Helix Waveguide, to be published.
5. Rowe, H. E., to be published.
6. Unger, H. G., Mode Conversion in Metallic and Helix Waveguide, to be published.

Error-Correcting Codes for Multiple-Level Transmission

By JESSIE MacWILLIAMS

(Manuscript received August 2, 1960)

A q-level alphabet is defined as a row vector space over a finite field with q elements. The letters of the alphabet are the rows of the vector space, each consisting of n symbols from the ground field. The weight of a letter is the number of nonzero symbols it contains. The minimum weight of the letters of the alphabet, excluding zero, is denoted by d. A relationship is established between the alphabet and a set of points S in a finite projective space. There is a many-one correspondence between the letters of the alphabet and the hyperplanes of the space. The weight of a letter is simply related to the incidence of the set S with the corresponding hyperplane.

Two sets of points in a finite projective space are called equivalent if they are related by a collineation of the space. Two alphabets are called equivalent if there exists between them, as vector spaces, a weight-preserving semi-isomorphism. It is shown that these definitions mean the same thing and reduce to the usual definition when $q = 2$.

An inequality is established between the dimension of the alphabet and the parameters d, q, n. This gives a lower bound for n in terms of the other parameters. It is shown that this bound cannot be achieved by alphabets with repeated columns. A method is given for constructing a class of alphabets which attain this bound. It is shown that for the case $q = 2$ *these are the only alphabets (in the sense of equivalence) for which the bound is attained.*

I. INTRODUCTION

A great deal of work has been done on error-correcting codes for the binary channel. In this paper we consider codes for a channel that can transmit more than two levels. Multiple-level transmission is practical if the channel is sufficiently quiet, as, for example, the submarine voice cable. It results in a substantial increase in bit rate and in added flexibility in choosing a code. One now has four parameters to adjust — the number of levels of transmission, the number of information symbols, the number of redundant symbols, and the number of errors it is desirable to detect and/or correct. Of course it cannot be decided without detailed analysis whether these advantages will more than compensate for the added complexity of the terminal equipment.

In the binary case, systematic error-correcting codes have certain advantages;[1] in particular, they are amenable to known mathematical techniques. It has been shown by Slepian[2] that the words of a systematic code form a group under place-by-place addition mod 2. The natural generalization of a group code over the field (0,1) appears to be a vector space over a finite field of q elements. We call such vector spaces *alphabets*, and their individual elements are called *letters*. In the general case, a "code" becomes an "alphabet" and a word (unfortunately!) becomes a "letter." Each letter is a row of n symbols picked from the ground field; the alphabet is a space of row vectors of length n. The q different symbols of the ground field correspond to q different transmission levels.

Because only a restricted type of code is considered, some assumptions must be made about the nature of the channel and of the information being transmitted. These are as follows:

(a) The number of transmission levels is a power of a prime number, since the number of elements in a finite field is a power of a prime. In practice this is not a severe restriction; between one and nine we have excluded only the number six.

(b) The channel is "symmetric" in the sense that every symbol has the same chance of getting through correctly, and that the probability of one symbol being changed into another is the same for every pair of symbols.

(c) All errors are equally bad. This might be the case, for example, if one were ordering merchandise from a mail order house by catalog number only.

With these assumptions the principles of error correction by a q-level alphabet are exactly the same as those described by Slepian[2] for a group code (i.e., a two-level alphabet). For convenience, the pertinent results from Slepian's paper are summarized in the Appendix. The parameters of an alphabet, besides n and q are

1. Its dimension as a vector space, denoted by k. The alphabet contains q^k letters; k is also the number of symbols in each letter which can be regarded as carrying information. The remaining $n - k$ symbols are added for the purpose of error detection and/or correction.

2. The minimum weight, d, of the letters of the alphabet other than $(00 \cdots 0)$. (The weight of a letter is the number of nonzero symbols it contains.) The quantity d is closely related to the error-correcting prop-

erties of the alphabet; if an alphabet is to be capable of correcting all occurrences of 1, 2, $\cdots$, e errors in each letter it must have $d = 2e + 1$.

The purpose of this paper is to investigate the properties of vector spaces over finite fields, particularly those properties which are related to the parameter d. The weight of a letter exists only in relation to a particular base of the vector space, which is an awkward situation in modern algebra. Hence our chief mathematical tool is not algebra but finite projective geometry. The connection between binary group codes and finite geometries was pointed out by Bose,[3] and is easily extended to the general case.

We first establish several new definitions of equivalence between alphabets. (Two equivalent alphabets have the same error-correcting properties.) A lower bound for n is found in terms of k, q and d. Clearly it is desirable to have $n - k$ (the number of check symbols) as small as possible. It is shown that this lower bound can be attained, but only by a restricted class of alphabets. These alphabets are, on the whole, not practical for communication purposes unless the expected error rate is extremely high. However, the geometric methods used in the construction of these alphabets can be applied to find useful alphabets for specific cases. The theorems derived for q-level alphabets apply equally well to the case $q = 2$ and contribute to the theory of binary group codes.

II. NOTATION

In this section we define the notation to be used in this paper and introduce Bose's theorem on the relation between alphabets and projective geometries.†

Let $F(q)$ be a finite field with q elements and characteristic p, and let $F^*(q)$ denote the nonzero elements of $F(q)$. We consider a vector space of dimension n over $F(q)$. Let $G_n(q)$ denote the "row space," i.e., that particular representation of the vector space consisting of all possible n-tuples of elements of $F(q)$. For example, $G_2(4)$ consists of the 2-tuples

$$\begin{array}{cccccc} (00) & (10) & (01) & (11) & (1w) & (1w^2) \\ & (w0) & (0w) & (ww) & (ww^2) & (w1) \\ & (w^20) & (0w^2) & (w^2w^2) & (w^21) & (w^2w) \end{array}$$

where w is a primitive cube root of unity.

Clearly $G_n(q)$ has q^n members. The $q^n - 1$ nonzero elements of $G_n(q)$ can be divided, in many ways, into $(q - 1)$ sets $G_1, \cdots, G_{q-1}$ such that $G_i = \lambda G_j$, $\lambda \in F^*(q)$. For our purposes it is usually enough to

† For finite projective geometry, see Carmichael,[4] Ch. 2; for Galois fields, see van der Waerden,[5] Ch. 5, Sect. 37.

examine only one of these sets, for example the first line in the table above.

A subspace of $G_n(q)$ is called an *alphabet over* $F(q)$ and its members are called *letters*. The length of a letter is n and the number of nonzero coordinates in a letter is its weight. Every alphabet contains the letter $(00 \cdots 0)$. The minimum weight of its other letters is denoted by d, and d is also called the weight of the alphabet. The dimension of the alphabet as a vector space over $F(q)$ is k. By $\mathcal{A}(k,d,n)$ we mean an alphabet $\mathcal{A}$ with dimension k, weight d and length (of each letter) n. For example, $G_n(q)$ is $\mathcal{A}(n,1,n)$.

An alphabet $\mathcal{A}(k,d,n)$ contains q^k letters, from which we pick any k independent vectors as generators. We write these as the rows of a $k \times n$ matrix $M(\mathcal{A})$, the *generator matrix* of $\mathcal{A}$. For example,

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

is the generator matrix of an $\mathcal{A}(2,2,3)$. We may assume that no column of a generator matrix consists entirely of zeros, for then the alphabet is isomorphic to a subspace of $G_{n-1}(q)$.

An ordered set of k elements of $F(q)$, not all zero (for example, a column of a generator matrix), may be regarded as the coordinates of a point of a projective space $T_{k-1}(q)$, of projective dimension $(k - 1)$, over $F(q)$. We shall adopt the convention that a k-tuple which refers to a point of $T_{k-1}(q)$ is to be written as a column vector, e.g.,

$$\mathbf{Q}_1 = \begin{bmatrix} q_{11} \\ q_{21} \\ \vdots \\ q_{k1} \end{bmatrix}.$$

$T_{k-1}(q)$ contains $(q^k - 1)/(q - 1)$ points; if $\lambda \in F^*(q)$, $\mathbf{Q}$ and $\lambda\mathbf{Q}$ are the same point. The points of $T_{k-1}(q)$ are in one-to-one correspondence with one-dimensional subspaces through the origin in $G_k(q)$.

Let us now write the generator matrix of $\mathcal{A}(k,d,n)$:

$$M(\mathcal{A}) = \begin{array}{c} \\ R_1 \\ R_2 \\ \vdots \\ R_k \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{Q}_1 & \mathbf{Q}_2 & \cdots & \mathbf{Q}_n \end{array} \\ \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1n} \\ q_{21} & q_{22} & \cdots & q_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ q_{k1} & q_{k2} & \cdots & q_{kn} \end{bmatrix} \end{array}$$

and call the rows $R_1, R_2, \cdots, R_k$ and the columns $\mathbf{Q}_1, \mathbf{Q}_2, \cdots, \mathbf{Q}_n$. Regard the columns as a set of points in $T_{k-1}(q)$. There are exactly k independent columns, so this set of points spans the space $T_{k-1}(q)$. Let ν_i be the number of times which some multiple of the column $\mathbf{Q}_i$ [the multiplier being an element of $F^*(q)$] appears in $M(\mathcal{A})$. The corresponding point in $T_{k-1}(q)$ shall then have multiplicity ν_i. We can now introduce Bose's theorem.†

Theorem 1: Let

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix}$$

be a general point of $T_{k-1}(q)$. Let S denote the set of points $\mathbf{Q}_1, \mathbf{Q}_2, \cdots, \mathbf{Q}_n$ each counted with proper multiplicity. Then the weight of the letter

$$R(\lambda) = \lambda_1 R_1 + \lambda_2 R_2 + \cdots + \lambda_k R_k, \qquad \lambda_i \in F(q)$$

of $\mathcal{A}$ is equal to the number of points of the set S which do not lie on the hyperplane

$$H(\lambda) \equiv \lambda_1 y_1 + \lambda_2 y_2 + \cdots + \lambda_k y_k = 0$$

of $T_{k-1}(q)$.

Proof: If, for example,

$$\lambda_1 q_{11} + \lambda_2 q_{21} + \cdots + \lambda_k q_{k1} = 0,$$

the point $\mathbf{Q}_1$ lies on $H(\lambda)$. The zeros in the letter $R(\lambda)$ arise from the points of S which lie on $H(\lambda)$, and the number of zeros will be the number of such points counted with proper multiplicity. The weight of $R(\lambda)$ is the number of its nonzero coordinates, which is the number of points of S (again counted with proper multiplicity) not lying on $H(\lambda)$. This proves the theorem.

In Fig. 1, the projective plane $T_2(2)$ is over the field (0,1). Note that $\mathbf{Q}_4\mathbf{Q}_5\mathbf{Q}_6$ are also collinear:

$$\mathbf{Q}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{Q}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{Q}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{Q}_4 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix},$$

$$\mathbf{Q}_5 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{Q}_6 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{Q}_7 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

† A different proof of this theorem for the field (01) is given in Ref. 3.

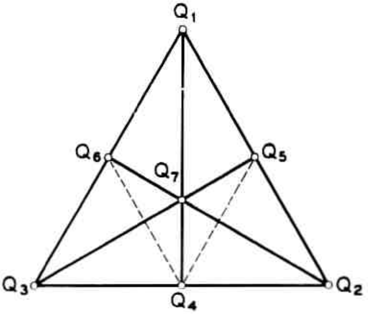


Fig. 1 — Illustration of Theorem 1.

Taking points Q_1,Q_2,Q_3,Q_7 in Fig. 1 as the set N we obtain a generator matrix

$$\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

of an alphabet $\mathfrak{A}(3,2,4)$. It is clear from the figure that there are at least two points of N not on any line of $T_2(2)$.

III. EQUIVALENT ALPHABETS

In this section we take up the question of equivalent alphabets, and show how Slepian's definition of equivalence may be extended to the more general case. First we discuss what properties one would intuitively hope for from such a definition.

We may consider an alphabet as an array of letters arranged one under another in such a way that we can speak of its columns. We know that the operations of permuting the columns, multiplying any column by an element of $F^*(q)$, and interchanging the names of the nonzero symbols will not change the error-correcting properties of the alphabet. The definition of equivalence between alphabets should allow us to do as many of these things as possible.

From Bose's theorem we recall that the weight of every letter of an alphabet is determined by the properties of a set of points in $T_{k-1}(q)$. First we wish that all alphabets derived from the same set of points should be equivalent; secondly, if two sets of points S,S' have, in some sense, the same incidence relations with the hyperplanes of $T_{k-1}(q)$ they should give rise to equivalent alphabets.

Given a set of points S in $T_{k-1}(q)$, we derive an alphabet from them by means of a generator matrix. We obtain the generator matrix by the following steps:

1. Fix a coordinate system in $T_{k-1}(q)$.†

2. Write the coordinates of the points of S as columns of a matrix repeating each column (not necessarily consecutively) with the proper multiplicity.

The order in which we write the columns is immaterial; also if $\mathbf{X}_i$ is such a column, we have the option of using $\lambda\mathbf{X}_i$, $\lambda \in F^*(q)$, instead. Thus it is apparent that a great many different generator matrices may arise from the same set of points.

We shall presently give separate intrinsic definitions of equivalence between two sets of points, two matrices and two alphabets, and show how they are interrelated. First we give a brief description of the collineation group of $T_{k-1}(q)$.‡

A collineation is a mapping of the set of points of $T_{k-1}(q)$ onto itself which preserves all incidence properties; that is, it sends lines into lines, planes into planes, lines through a point into lines through a point, and so on. The collineations of $T_{k-1}(q)$ form a group, denoted by $C(k,q)$. A nonsingular linear projective transformation of coordinates is a collineation; so is the (nonlinear) transformation of coordinates induced by an automorphism of the ground field $F(q)$. Let $P(k,q)$ be the group of linear projective transformations, and $A(k,q)$ the group of transformations induced by automorphisms of the ground field. Then any collineation of $C(k,q)$ can be expressed as the product of a member of $P(k,q)$ and a member of $A(k,q)$. [Although an element of $P(k,q)$ does not in general commute with an element of $A(k,q)$, the two groups commute as subgroups of $C(k,q)$.] We recall that an automorphism of a finite field of $q = p^m$ elements is always of the form $\theta \rightarrow \theta^{p^\nu}$, where θ is a primitive element; and, for a nontrivial automorphism, $0 < \nu < m$. The integers of the field (the elements of the prime subfield) are not changed by such a mapping; hence a prime field has no nontrivial automorphisms, and in this case $C(k,p) = P(k,p)$.

We now make the following definitions of equivalence:

Definition 1: The (unordered) sets of points S,S' are equivalent if there exists a collineation of $T_{k-1}(q)$ which sends S into S'. We write $S' = C(S)$.

† By a fixed coordinate system we mean that the coordinates of every point are fixed, except possibly for multiplication by an element of $F^*(q)$. In the case of finite projective geometries, this involves more than choosing the base points of the system.

‡ The subject is treated in great detail in Carmichael,[4] pp. 355–372.

Definition 2: Two $(k \times n)$ generator matrices M,M' over $F(q)$ are equivalent if

$$M' = g\Phi M^*, \qquad M^* = M\pi\Lambda.$$

Here Φ is an automorphism of the ground field applied to the entries in M^*, g an invertible $(k \times k)$ matrix over $F(q)$, π an $(n \times n)$ permutation matrix, Λ a (nonsingular) diagonal $(n \times n)$ matrix over $F^*(q)$.

Since π has only one nonzero entry in each row and column we can always choose Λ' so that

$$\Lambda'\pi = \pi\Lambda.$$

Definition 3: Two alphabets $\mathfrak{A}$ and $\mathfrak{A}'$ are equivalent if there exists between them a weight-preserving semi-isomorphism.

A semi-isomorphism f between two vector spaces $\mathfrak{A}$, $\mathfrak{A}'$ is uniquely specified by describing what happens to the base vectors $R_1, \cdots, R_k$ of $\mathfrak{A}$, and choosing an automorphism of the ground field. The mapping

$$f(R_i) = R_i', \qquad i = 1, \cdots, k,$$

$$f\left(\sum_{i=1}^{k} \alpha_i R_i\right) = \sum_{i=1}^{k} \Phi(\alpha_i) R_i'$$

is a semi-isomorphism provided that $R_1', \cdots, R_k'$ are linearly independent; any semi-isomorphism can be described in this way.

We note also that a weight-preserving mapping of an alphabet $\mathfrak{A}$ onto an alphabet $\mathfrak{A}'$ is necessarily one-to-one; for only letters of zero weight in $\mathfrak{A}$ can map onto the zero $(00 \cdots 0)$ of $\mathfrak{A}'$.

In all of these definitions, equivalence has its usual properties; i.e., it is symmetric, reflexive and transitive.

We now show that the three definitions are compatible; that is, in a sense to be made precise,

Definition 1 → Definition 2,

Definition 2 → Definition 3,

Definition 3 → Definition 1.

Theorem 2: If S,S' are equivalent in the sense of Definition 1, then the matrices M,M', to which they give rise in a fixed coordinate system, are equivalent in the sense of Definition 2.

Proof: Let $\bar{S}$ be an ordering of the set S, and $\bar{S}'$ the ordering of S'

into which $\bar{S}$ is sent by a collineation $g\Phi$ of $T_{k-1}(q)$. If X_i, X_i' are corresponding points of $\bar{S}, \bar{S}'$ their coordinates are given by

$$\mathbf{X}_i = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \qquad \mathbf{X}_i' = g \begin{pmatrix} \Phi(x_1) \\ \Phi(x_2) \\ \vdots \\ \Phi(x_n) \end{pmatrix}.$$

Let $M(\bar{S})$, $M(\bar{S}')$ denote the matrices with columns

$$\mathbf{X}_1, \cdots, \mathbf{X}_n, \quad \mathbf{X}_1', \cdots, \mathbf{X}_n', \qquad M(\bar{S}') = g\Phi M(\bar{S}).$$

Then there exist permutation matrices such that

$$M'\pi' = M(\bar{S}') = g\Phi M(\bar{S}'), \qquad M(\bar{S}) = M\pi.$$

Hence

$$M' = g\Phi M^*, \quad M^* = M\pi\pi'^{-1} = M\pi^*,$$

where π^* is a permutation matrix.

Theorem 3: If the generator matrices M, M' are equivalent in the sense of Definition 2, then the alphabets $\mathcal{A}, \mathcal{A}'$ derived from them are equivalent in the sense of Definition 3.

Proof: We have

$$M' = g\Phi M^*, \qquad M^* = M\pi\Lambda.$$

Let $\mathcal{A}^*$ be the alphabet derived from M^*. We set up a weight-preserving isomorphism h between $\mathcal{A}$ and $\mathcal{A}^*$, and a weight-preserving semi-isomorphism f between $\mathcal{A}^*$ and $\mathcal{A}'$. We define h as follows: If R is a letter of $\mathcal{A}$ then

$$h(R) = R\pi\Lambda.$$

This is clearly a weight-preserving mapping, since its effect is to permute the entries in R and multiply each entry by an element of $F^*(q)$. It is also linear, for if $R_1, \cdots, R_k$ are the rows of M, and $R_1^*, \cdots, R_k^*$ the rows of M^* we have

$$h(R_i) = R_i\pi\Lambda = R_i^*,$$

$$h\left(\sum_{i=1}^{k} \alpha_i R_i\right) = \sum_{i=1}^{k} \alpha_i R_i \pi\Lambda = \sum_{i=1}^{k} \alpha_i R_i^*.$$

We define f as follows: If $R^* = (r_1, r_2, \cdots, r_n)$ is a letter of $\mathcal{A}^*$, then

$f(R^*) = [\Phi(r_1), \Phi(r_2), \cdots, \Phi(r_n)]$; f is weight-preserving, since $\Phi(r) = 0$ implies $r = 0$.

To show that f is a semi-isomorphism,

$$\mathcal{A}^* \xrightarrow{f} \mathcal{A}',$$

we observe that $g^{-1}M'$ is also a generator matrix of $\mathcal{A}'$. Let $R_1', \cdots, R_k'$ be the rows of $g^{-1}M'$, and let $R_i' = (r_{i1}', r_{i2}', \cdots, r_{in}')$. Let $R_1^*, \cdots, R_k^*$ be the rows of M^*, with $R_i^* = (r_{i1},^* r_{i2},^* \cdots, r_{in}^*)$. Since $g^{-1}M' = \Phi M^*$ we have

$$(r_{i1}', r_{i2}', \cdots, r_{in}') = [\Phi(r_{i1}^*), \Phi(r_{i2}^*), \cdots, \Phi(r_{in}^*)],$$

or

$$f(R_i^*) = R_i'.$$

Then

$$f\left(\sum_{i=1}^{k} \alpha_i R_i^*\right) = \left[\Phi\left(\sum_{i=1}^{k} \alpha_i r_{i1}^*\right), \Phi\left(\sum_{i=1}^{k} \alpha_i r_{i2}^*\right), \cdots, \Phi\left(\sum_{i=1}^{k} \alpha_i r_{in}^*\right)\right].$$

Since Φ is a field automorphism this becomes

$$\begin{aligned} f\left(\sum_{i=1}^{k} \alpha_i R_i^*\right) \\ &= \left[\sum_{i=1}^{k} \Phi(\alpha_i)\Phi(r_{i1}^*), \sum_{i=1}^{k} \Phi(\alpha_i)\Phi(r_{i2}^*), \cdots, \sum_{i=1}^{k} \Phi(\alpha_i)\Phi(r_{in}^*)\right] \\ &= \sum_{i=1}^{k} \Phi(\alpha_i) R_i'. \end{aligned}$$

We then have

$$R_i \xrightarrow{h} R_i^* \xrightarrow{f} R_i', \qquad \Sigma\alpha_i R_i \xrightarrow{h} \Sigma\alpha_i R_i^* \xrightarrow{f} \Sigma\Phi(\alpha_i) R_i',$$

and hf is a weight-preserving semi-isomorphism between $\mathcal{A}$ and $\mathcal{A}'$.

Theorem 4: Let $\mathcal{A}, \mathcal{A}'$ be equivalent alphabets in the sense of Definition 3, and M, M' be any generator matrices of $\mathcal{A}, \mathcal{A}'$. Fix the coordinate system in $T_{k-1}(q)$, and let S, S' be the sets of points whose coordinates are the columns of M and M'. Then S and S' are equivalent in the sense of Definition 1.

Lemma: Let the alphabets $\mathcal{A}, \mathcal{A}^*$ be related by a weight-preserving isomorphism w; M, M^* are generator matrices of $\mathcal{A}$ and $\mathcal{A}^*$ such that $M^* = w(M)$. Then in any coordinate system in $T_{k-1}(q)$ the columns of M and M^* give rise to the same (unordered) set of points S.

Proof of Lemma: If $R_1, \cdots, R_k; R_1^*, \cdots, R_k^*$ are the rows of M and M^* we have

$$w(R_i) = R_i^*, \qquad w\left(\sum_{i=1}^{k} \alpha_i R_i\right) = \sum_{i=1}^{k} \alpha_i R_i^*.$$

Let $(y_1 \cdots y_k)$ be the coordinates of the general point of $T_{k-1}(q)$. Map the letters of $\mathcal{A}$ onto the hyperplanes of $T_{k-1}(q)$ as follows: R_i maps onto $y_i = 0$, $\sum \alpha_i R_i$ maps onto $\sum \alpha_i y_i = 0$. Because of the isomorphism between $\mathcal{A}$ and $\mathcal{A}^*$ we have a similar mapping of the letters of $\mathcal{A}^*$ onto the hyperplanes of $T_{k-1}(q)$: R_i^* maps onto $y_i = 0$, $\sum \alpha_i R_i^*$ maps onto $\sum \alpha_i y_i = 0$.

Let $I = (\delta_{ij})$ be the incidence matrix of points and hyperplanes in $T_{k-1}(q)$, where $\delta_{ij} = 1$ if the ith point lies on the jth hyperplane and is zero otherwise. Each row (column) of I contains $(q^{k-1} - 1)/(q - 1)$ ones and q^{k-1} zeros. The matrix I for the projective plane $T_2(2)$ is illustrated in Table I.

The matrix I is nonsingular. This is easily seen by considering the product $I \cdot I$. In this, all terms on the main diagonal are equal to the number of points, $a = (q^{k-1} - 1)/(q - 1)$, on a hyperplane. All other terms are equal to the number of points, $b = (q^{k-2} - 1)/(q - 1)$, on the intersection of two hyperplanes. The determinant of the matrix is then

$$[a + (\mu - 1)b](a - b)^{n-1}.$$

When we substitute the values for a,b, the first factor becomes

$$\left(\frac{q^{k-1} - 1}{q - 1}\right)^2;$$

hence the determinant is not zero. (We assume $k > 1$.)

Let $P_1, P_2, \cdots, P_\mu$, $\mu = (q^k - 1)/(q - 1)$, be the ordering of the points of $T_{k-1}(q)$ as they appear as columns of I. Let S,S^* be the sets

TABLE I — I = INCIDENCE MATRIX FOR POINTS AND LINES IN $T_2(2)$

	100	010	001	110	101	011	111
$y_1 = 0$	0	1	1	0	0	1	0
$y_2 = 0$	1	0	1	0	1	0	0
$y_3 = 0$	1	1	0	1	0	0	0
$y_1 + y_2 = 0$	0	0	1	1	0	0	1
$y_1 + y_3 = 0$	0	1	0	0	1	0	1
$y_2 + y_3 = 0$	1	0	0	0	0	1	1
$y_1 + y_2 + y_3 = 0$	0	0	0	1	1	1	0

of points whose coordinates are columns of M,M^* respectively. Assign to P_i the multiplicity $n_i(n_i^*)$ with which it appears in the set $S(S^*)$. If P_i does not appear in $S(S^*)$, $n_i = 0$ $(n_i^* = 0)$. Form the column vectors

$$\mathbf{n} = \begin{pmatrix} n_1 \\ n_2 \\ \vdots \\ n_\mu \end{pmatrix}, \quad \mathbf{n}^* = \begin{pmatrix} n_1^* \\ n_2^* \\ \vdots \\ n_\mu^* \end{pmatrix}.$$

The ith term of the matrix product $I\mathbf{n}$ is the sum of the multiplicities of the points of S which lie on the ith hyperplane. By Theorem 1, this is the number of zeros in the corresponding letters of $\mathfrak{A}$.

Since the isomorphism between $\mathfrak{A}$ and $\mathfrak{A}^*$ is weight-preserving we have

$$I\mathbf{n} = I\mathbf{n}^*,$$

or, since I is invertible,

$$\mathbf{n} = \mathbf{n}^*.$$

Hence the set of points S^* is at most a rearrangement of the set S.

Proof of Theorem 4: $\mathfrak{A}$ and $\mathfrak{A}'$ are related by a weight-preserving semi-isomorphism f. Let $R_1, \cdots, R_k$ be the rows of the generator matrix M of $\mathfrak{A}$. $R_1'' = f(R_1), \cdots, R_k'' = f(R_k)$ are k linearly independent letters of $\mathfrak{A}'$, which we may take as the rows of a generator matrix M'' of $\mathfrak{A}'$. We can describe f as follows:

$$f(R_i) = R_i'', \qquad f\left(\sum_{i=1}^{k} \alpha_i R_i\right) = \sum_{i=1}^{k} \Phi(\alpha_i) R_i'',$$

where Φ is an automorphism of the ground field which is uniquely determined by f once we have chosen M.

Let $R_i^* = \Phi^{-1}(R_i'')$, $i = 1, \cdots, k$; $R_1^*, \cdots, R_k^*$ are linearly independent. Let M^* be the generator matrix formed of these rows and $\mathfrak{A}^*$ the alphabet derived from M^*. The mapping h of $\mathfrak{A}'$ onto $\mathfrak{A}^*$ induced by Φ^{-1} is clearly a weight-preserving semi-isomorphism.

Consider the mapping fh between $\mathfrak{A}$ and $\mathfrak{A}^*$. We have

$$R_i \xrightarrow{f} R_i'' \xrightarrow{h} R_i^*,$$

$$\Sigma\alpha_i R_i \xrightarrow{f} \Sigma\Phi(\alpha_i)R_i'' \xrightarrow{h} \Phi^{-1}[\Sigma\Phi(\alpha_i)R_i''] = \Sigma\alpha_i R_i^*.$$

Since f is weight-preserving by hypothesis, fh is a weight-preserving isomorphism between $\mathfrak{A}$ and $\mathfrak{A}^*$; M and M^* are corresponding generator matrices under fh, hence by the Lemma they arise from the same set of points S in $T_{k-1}(q)$.

Let S'' be the points of $T_{k-1}(q)$ corresponding to the columns of M''. If

$$\mathbf{X}_i'' = \begin{pmatrix} x_1'' \\ \vdots \\ x_k'' \end{pmatrix}, \quad \mathbf{X}_i^* = \begin{pmatrix} x_1^* \\ \vdots \\ x_k^* \end{pmatrix}$$

are the ith columns of M'' and M^* respectively, we have

$$\begin{pmatrix} x_1'' \\ \vdots \\ x_k'' \end{pmatrix} = \begin{pmatrix} \Phi(x_1^*) \\ \vdots \\ \Phi(x_k^*) \end{pmatrix}.$$

Hence the set S'' is obtained from the set S^* by a collineation C_1 of $T_{k-1}(q)$.

Let M' be any generator matrix of $\mathfrak{A}'$; then $M' = gM''$. Let S' be the points of $T_{k-1}(q)$ corresponding to the columns of M'. S' arises from S'' by a linear projective transformation, i.e., by a collineation C_2.

We have then

$$S' = C_2(S'') = C_2C_1(S),$$

which proves the theorem.

It can be shown from Theorems 2, 3 and 4 that a complete equivalence class of sets of points gives rise to a complete equivalence class of matrices; a complete equivalence class of matrices gives rise to a complete equivalence class of alphabets; and this in turn gives rise to a complete equivalence class of sets of points. The details of these correspondences are quite complicated, since an unordered set of points can give rise to many matrices, and different generator matrices can produce the same alphabet.

Theorems 2, 3 and 4 are, of course, true over the field (0,1). We rewrite our definitions for this field, since they take a simpler form. Φ is the identity, and the only possible choice for Λ is the unit matrix.

Definition 1': Two sets of points S,S' in $T_{k-1}(2)$ are equivalent if they are related by a linear projective transformation of coordinates.

Definition 2': Two $(k \times n)$ matrices M,M' over $F(2)$ are equivalent if

$$M = gM'\pi,$$

where π is an $(n \times n)$ permutation matrix, and g an invertible $(k \times k)$ matrix over $F(2)$.

Definition 3′: Two alphabets $\mathcal{A},\mathcal{A}'$ over $F(2)$ are equivalent if they are isomorphic as groups in such a way that corresponding elements have the same weight.

It will be recognized that this is, in fact, the familiar definition of equivalence for alphabets over (0,1).

IV. RELATIONS BETWEEN k, d, n

In this section we establish certain relations between the parameters k,d,n, which are necessary conditions for the existence of an alphabet $\mathcal{A}(k,d,n)$. We assume, as before, that the alphabet has no column consisting entirely of zeros.

Define $Z[x]$ to mean the least integer greater than or equal to the rational number x.

Theorem 5:† A necessary condition for the existence of $\mathcal{A}(k,d,n)$ is that

$$n \geqq Z\left[\frac{1}{q^{k-1}}\left(\frac{q^k - 1}{q - 1}\right)d\right].$$

Proof: As before, let I be the incidence matrix of points and hyperplanes in $T_{k-1}(q)$.

Let J be the complement of I obtained by replacing zeros by ones and ones by zeros. J is symmetric; each row (column) contains q^{k-1} ones and $1 + q + \cdots + q^{k-2}$ zeros.

The matrix J for the projective plane $T_2(2)$ over the field (0,1) is illustrated in Table II.

Let

$$\mathbf{n} = \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_\mu \end{bmatrix},$$

where $\mu = 1 + q + \cdots + q^{k-1}$ and n_i stands for the multiplicity of the point P_i of $T_{k-1}(q)$.

Consider the expression $J\mathbf{n}$. The product of the ith row of J with the column of n_i is the sum of the multiplicities of the points P_i which do not lie on the ith hyperplane. By Bose's theorem, this is the weight of the letters of the alphabet corresponding to the ith hyperplane. Now

† This theorem has been obtained for the field (0,1) by many authors in as many ways. See for example, Ref. 6, Theorem 5; Ref. 3, Eq. (52), and other authors quoted in Ref. 3.

TABLE II — J = COMPLEMENT OF I

	100	010	001	110	101	011	111
$y_1 = 0$	1	0	0	1	1	0	1
$y_2 = 0$	0	1	0	1	0	1	1
$y_3 = 0$	0	0	1	0	1	1	1
$y_1 + y_2 = 0$	1	1	0	0	1	1	0
$y_1 + y_3 = 0$	1	0	1	1	0	1	0
$y_2 + y_3 = 0$	0	1	1	1	1	0	0
$y_1 + y_2 + y_3 = 0$	1	1	1	0	0	0	1

define a column vector

$$\mathbf{d} = \begin{pmatrix} d \\ d \\ \vdots \\ d \end{pmatrix}.$$

Since our alphabet is assumed to have minimum weight d, we have the inequalities

$$J\mathbf{n} \geqq \mathbf{d}.$$

Since we may assume $d \geqq 1$, these inequalities imply that there must be at least one point of nonzero multiplicity not lying on any given hyperplane — that is, the points of nonzero multiplicity span the space $T_{k-1}(q)$.

Hence, given k and d, the least value of n for which there exists an alphabet $\mathcal{A}(k,d,n)$ is the minimum value of

$$\sum_{i=1}^{\mu} n_i ,$$

where n_i , $i = 1, \cdots, \mu$ are nonnegative integers which satisfy $J\mathbf{n} \geqq \mathbf{d}$.

By adding all the inequalities of $J\mathbf{n} \geqq \mathbf{d}$, we obtain

$$q^{k-1} \sum_{i=1}^{\mu} n_i \geqq (1 + q + \cdots + q^{k-1})\, d,$$

or, setting

$$n = \sum_{i=1}^{\mu} n_i ,$$

$$n \geqq Z\left[\frac{1}{q^{k-1}}\left(\frac{q^k - 1}{q - 1}\right) d\right].$$

In the case that $d = q^{k-1}$, for any value of k the lower bound becomes

$$n \geqq \frac{q^k - 1}{q - 1} = 1 + q + \cdots + q^{k-1}.$$

In this case the lower bound is the largest possible lower bound, as it is achieved by the alphabet which corresponds to $n_1 = n_2 = \cdots = n_\mu = 1$, that is, the alphabet which results from taking every point of $T_{k-1}(q)$ with multiplicity one.

One has an intuitive feeling that alphabets with the least n for a given k,d are likely to have no repeated columns if this is possible. This is partly justified by the following theorem.

Theorem 6: If a generator matrix of $\mathfrak{A}(k,d,n)$ contains a repeated column [in the sense that $\mathbf{Q}_t = \lambda \mathbf{Q}_s$ for some λ of $F^*(q)$], then

$$n \geqq Z\left[\frac{1}{q^{k-2}}\left(\frac{q^{k-1} - 1}{q - 1}\right) d\right] + 2.$$

For the purposes of this proof and the succeeding lemma we write the above inequality as

$$n \geqq Z\left[\frac{q}{q - 1}\left(1 - \frac{1}{q^{k-1}}\right) d\right] + 2.$$

Proof: Let P be the point of $T_{k-1}(q)$ which corresponds to the repeated column. Choose a coordinate system in which P is one of the base points, say $P = \mathbf{e}_1$. We then have an equivalent alphabet $\mathfrak{A}'$ which may be written

$$M(\mathfrak{A}') = \left\{\begin{array}{ccccc|c} 1 & 1 & 0 & \cdots & 0 & \cdots \\ 0 & 0 & 1 & \cdots & 0 & \cdots \\ \vdots & \vdots & \vdots & & & \\ 0 & 0 & 0 & \cdots & 1 & \cdots \end{array}\right\}.$$

The letters of $\mathfrak{A}'$ to which the first two columns contribute zeros form a vector space $\overline{\mathfrak{A}}$; $\overline{\mathfrak{A}}$ is generated by the rows $2, \cdots, k$ of $M(\mathfrak{A}')$. The minimum weight of the letters of $\overline{\mathfrak{A}}$ is at least as great as the minimum weight of the letters of $\mathfrak{A}'$. Hence the alphabet $\overline{\mathfrak{A}}$ has parameters $k - 1$, d', $n - 2$, with $d' \geqq d$. By Theorem 3 we get

$$n - 2 \geqq Z\left[\frac{q}{q - 1}\left(1 - \frac{1}{q^{k-1}}\right) d'\right],$$

or

$$n \geqq Z\left[\frac{q}{q - 1}\left(1 - \frac{1}{q^{k-1}}\right) d\right] + 2.$$

It is clear that, if $d \leqq 2q^{k-1}$

$$\frac{q}{q-1}\left(1 - \frac{1}{q^{k-1}}\right) d + 2 \geqq \frac{q}{q-1}\left(1 - \frac{1}{q^k}\right) d.$$

We need a little more, namely:

Lemma: If $d \leqq q^{k-1}$, then

$$\frac{q}{q-1}\left(1 - \frac{1}{q^{k-1}}\right) d + 2 \geqq \frac{q}{q-1}\left(1 - \frac{1}{q^k}\right) d + 1.$$

Hence

$$Z\left[\frac{q}{q-1}\left(1 - \frac{1}{q^{k-1}}\right) d + 2\right] > Z\left[\frac{q}{q-1}\left(1 - \frac{1}{q^k}\right) d\right].$$

Proof:

$$\begin{aligned}\frac{q}{q-1}\left(1 - \frac{1}{q^{k-1}}\right) d + 2 &= \left[\frac{q}{q-1}\left(1 - \frac{1}{q^k}\right)\right. \\ &\qquad \left. + \frac{q}{q-1}\left(\frac{1}{q^k} - \frac{1}{q^{k-1}}\right)\right] d + 2 \\ &= \frac{q}{q-1}\left(1 - \frac{1}{q^k}\right) d - \frac{d}{q^{k-1}} + 2 \\ &\geqq \frac{q}{q-1}\left(1 - \frac{1}{q^k}\right) d + 1.\end{aligned}$$

Theorem 7: If $d \leqq q^{k-1}$ the bound given in Theorem 3 cannot be attained by an alphabet with repeated columns.

This result is not surprising in view of the remark at the end of the proof of Theorem 3. If $d > q^{k-1}$, the inequality of Theorem 5 gives $n > (q^k - 1)/(q - 1)$; i.e., n is larger than the total number of points in the space $T_{k-1}(q)$. Thus we must have repeated columns in the generator matrix.

By repeated applications of the procedure of Theorem 6 we can write down lower bounds for the n of alphabets having a given number of columns with given multiplicities. However, this does not seem very interesting; we will first say what we can about alphabets with no repeated columns. We assume from now on that we are dealing with such alphabets.

V. A CLASS OF ALPHABETS

In this section we describe a class of alphabets for which the bound of Theorem 5 is attained, and show how other alphabets which attain this bound may be derived from them.

We can immediately write down the class of alphabets.† Choose a fixed k, and consider the following sets of points in $T_{k-1}(q)$:

(0) — The set S_0 of all points of $T_{k-1}(q)$:

$$n_0 = 1 + q + \cdots + q^{k-1}, \qquad d_0 = q^{k-1}.$$

Every letter of this alphabet has weight d_0.

(1) — The set S_1 of all points but one of $T_{k-1}(q)$:

$$n_1 = q + q^2 + \cdots + q^{k-1}, \qquad d_1 = q^{k-1} - 1.$$

The $(1 + q + \cdots + q^{k-2})$ hyperplanes through the omitted point correspond to letters of weight q^{k-1}, other hyperplanes to letters of weight $q^{k-1} - 1$.

(2) — S_2 = all points of $T_{k-1}(q)$ except for the $(1 + q)$ points of a line L_1.

$$n_2 = q^2 + \cdots + q^{k-1}, \qquad d_2 = q^{k-1} - q.$$

The $(1 + q + \cdots + q^{k-3})$ hyperplanes through L_1 correspond to letters of weight q^{k-1}, others to letters of weight $q^{k-1} - q$.

(3) — S_3 = all points of $T_{k-1}(q)$ except for the $(1 + q + q^2)$ points of a plane P_2.

$$n_3 = q^3 + \cdots + q^{k-1}, \qquad d_3 = q^{k-1} - q^2.$$

The $(1 + q + \cdots + q^{k-4})$ hyperplanes through P_2 correspond to letters of weight q^{k-1}, other hyperplanes to letters of weight $q^{k-1} - q^2$.

⋮

$(k - 1)$ — S_{k-1} = all points of $T_{k-1}(q)$ except for the points of a hyperplane

$$n_{k-1} = q^{k-1}, \qquad d_{k-1} = q^{k-1} - q^{k-2}.$$

The omitted hyperplane corresponds to letters of weight q^{k-1}, all others to letters of weight $q^{k-1} - q^{k-2}$.

It is easy to verify that for these alphabets the bound of Theorem 3 is attained. Consider

† For the case of $q = 2$ some, or all, of these alphabets have been found by other authors by different methods. See, for example, Refs. 3 and 6. They are, of course, picked up by any systematic search, such as linear programming. $\mathfrak{A}_{k-1}$ is the Reed-Muller code for $m = n$, $r = 1$.

$$\begin{aligned} n_i - \frac{1}{q^{k-1}}\left(\frac{q^k-1}{q-1}\right)d_i &= q^i + \cdots + q^{k-1} - \frac{1}{q^{k-1}}\left(\frac{q^k-1}{q-1}\right)(q^{k-1}-q^{i-1}) \\ &= q^i\left(\frac{q^{k-i}-1}{q-1}\right) - \frac{1}{q^{k-1}}\left(\frac{q^k-1}{q-1}\right)(q^{k-i}-1)q^{i-1} \\ &= \frac{q^{k-i}-1}{q-1}\left(q^i - \frac{q^k-1}{q^{k-i}}\right) \\ &= \frac{1}{q^{k-i}}\left(\frac{q^{k-i}-1}{q-1}\right). \end{aligned}$$

Since $q \geqq 2$, this quantity is less than one; i.e.,

$$1 > n_i - \frac{1}{q^{k-1}}\left(\frac{q^k-1}{q-1}\right)d_i > 0.$$

It will appear presently that for $q = 2$ these are the only alphabets which attain the bound of Theorem 5. The case $q > 2$ is more complicated.

Suppose that $\mathcal{A}(k,d,n)$ is an alphabet (with no repeated columns) for which

$$n = Z\left[\frac{1}{q^{k-1}}\left(\frac{q^k-1}{q-1}\right)d\right].$$

Write

$$\frac{1}{q^{k-1}}\left(\frac{q^k-1}{q-1}\right) = 1 + \frac{Q}{q^{k-1}}, \qquad Q = q^{k-2} + \cdots + q + 1 = \frac{q^{k-1}-1}{q-1},$$

$$n = Z\left[d + \frac{Qd}{q^{k-1}}\right].$$

Let

$$Qd = sq^{k-1} - r, \qquad 0 \leqq r \leqq q^{k-1} - 1, \qquad 0 < s \leqq Q, \qquad (1)$$

where r and s are integers; s cannot be zero since d is positive; $s \leqq Q$ since $d \leqq q^{k-1}$. Then

$$n = Z\left[d + s - \frac{r}{q^{k-1}}\right] = d + s.$$

If we remove η columns from the generator matrix of $\mathcal{A}$ in such a way that the remaining matrix is of rank k (this is always possible for $\eta \leqq n - k$), we obtain an alphabet of length $n - \eta$ and minimum weight $\bar{d} \geqq d - \eta$. Let us consider the worst case, i.e., $\bar{d} = d - \eta$.

Lemma: If

$$n = Z\left[\frac{1}{q^{k-1}}\left(\frac{q^k - 1}{q - 1}\right)d\right],$$

then

$$n - \eta = Z\left[\frac{1}{q^{k-1}}\left(\frac{q^k - 1}{q - 1}\right)(d - \eta)\Big)\right]$$

provided that

$$\eta < (q - 1) - \frac{q - 1}{q^{k-1} - 1}(r - 1). \tag{2}$$

Proof:

$$\begin{aligned} Z\left[\left(1 + \frac{Q}{q^{k-1}}\right)(d - \eta)\right] &= Z\left[d + s - \frac{r}{q^{k-1}} - \eta - \frac{\eta Q}{q^{k-1}}\right] \\ &= Z\left[(d + s - \eta) - \frac{Q\eta + r}{q^{k-1}}\right]. \end{aligned}$$

This is equal to $d + s - \eta$ if and only if $(Q\eta + r)/q^{k-1} < 1$; i.e.,

$$\eta < \frac{1}{Q}(q^{k-1} - r) = (q - 1)\frac{q^{k-1}}{q^{k-1} - 1} - (q - 1)\frac{r}{q^{k-1} - 1}$$

or

$$\eta < (q - 1) - (q - 1)\frac{r - 1}{q^{k-1} - 1}.$$

Now suppose that $\bar{d} > d - \eta$, and η satisfies (2):

$$Z\left[\left(1 + \frac{Q}{q^{k-1}}\right)\bar{d}\right] \geqq Z\left[\left(1 + \frac{Q}{q^{k-1}}\right)(d - \eta)\right] = n - \eta.$$

By Theorem 3 applied to the alphabet $\mathcal{A}(k,\bar{d},n - \eta)$, we have

$$n - \eta \geqq Z\left[\left(1 + \frac{Q}{q^{k-1}}\right)\bar{d}\right].$$

Hence only equality is possible.

Theorem 8: If $\mathcal{A}(k,d,n)$ attains the bound of Theorem 5 and $\mathcal{A}(k,d,n - \eta)$ is obtained from it by removing η columns from a generator matrix of $\mathcal{A}$, where η satisfies (2) in such a way that the remaining matrix is of rank k, then the new alphabet also attains the bound of Theorem 5.

We remark that if we select the columns with proper care, it is possible to remove more than the number given by (2) and still obtain an alphabet which attains the bound of Theorem 5. The alphabets $\mathfrak{A}_2, \cdots, \mathfrak{A}_{k-1}$ listed at the beginning of this section are examples.

We now reformulate (2) in a more convenient form. We observe, from (1), that, since d is an integer, so is $(sq^{k-1} - r)/Q$. Subtract from it the integer

$$s(q-1) - \frac{s(q^{k-1}-1)}{Q}$$

and we find that

$$\frac{s-r}{Q} = \frac{s-r}{q^{k-2} + \cdots + q + 1}$$

is also an integer. We have two cases:

i. $s = Q, r = \mu Q \qquad [\mu \leqq q - 1 \text{ from } (1)].$

Then (2) becomes

$$\eta < (q-1) - \mu + \frac{1}{Q}$$

or, since all these symbols represent integers,

$$\eta \leqq (q-1) - \mu. \tag{3}$$

ii. $s < Q, r = \mu Q + s \qquad \left[\mu < \left(1 - \frac{s}{q^{k-1}-1}\right)(q-1) \text{ from } (1)\right].$

Then (2) becomes

$$\eta < (q-1) - \mu - \frac{s-1}{Q}$$

or

$$\eta \leqq q - 1 - \mu - 1 = q - 2 - \mu. \tag{4}$$

In case i we have, from (1),

$$d = q^{k-1} - \mu, \qquad 0 \leqq \mu \leqq q - 1.$$

The alphabet $\mathfrak{A}_0$ corresponds to the case $\mu = 0$, and the alphabet $\mathfrak{A}_1$ to $\mu = 1$. From the alphabet $\mathfrak{A}_0$ we can subtract any number $\eta \leqq q - 1$ of columns and obtain an alphabet which attains the bound of Theorem 5. (The alphabet $\mathfrak{A}_1$ is obtained by subtracting one arbitrary column.)

In case ii we have, from (1),

$$d = \frac{sq^{k-1}}{Q} - \frac{\mu Q + s}{Q} = s(q - 1) - \mu.$$

For the alphabets $\mathfrak{A}_2, \cdots, \mathfrak{A}_{k-1}$ we have $\mu = 0$. This is readily verified by direct calculation:

$$d_i = q^{k-1} - q^{i-1} = q^{i-1}(q^{k-i} - 1) = (q-1)q^{i-1}(q^{k-i-1} + \cdots + q + 1).$$

Thus

$$s = (q^{k-2} + \cdots + q^{i-1}), \qquad \mu = 0.$$

We shall show that these are the only alphabets besides $\mathfrak{A}_0$ for which $\mu = 0$.

The generator matrix of an alphabet $\mathfrak{A}(k,d,n)$ with no repeated columns consists of a subset of the columns of the generator matrix $M(\mathfrak{A}_0)$. Let S be the generating points of $\mathfrak{A}(k,d,n)$ in $T_{k-1}(q)$, and denote by $C(S)$ the points of $T_{k-1}(q)$ which are not in S.

Let ν be the number of points in $C(S)$ and δ the *maximum* number of points of $C(S)$ which do not lie on a hyperplane of $T_{k-1}(q)$. The alphabet $\mathfrak{A}$ then has length $n_0 - \nu$ and weight $d_0 - \delta$, where $n_0 = (q^k - 1)/(q - 1)$, $d_0 = q^{k-1}$ are the parameters of $\mathfrak{A}_0$. Using Theorem 5 on these numbers, we obtain

$$\left(\frac{q^k - 1}{q - 1} - \nu\right)(q - 1) \geqq \frac{q^k - 1}{q^{k-1}}(q^{k-1} - \delta),$$

or

$$\nu(q - 1) \leqq \left(q - \frac{1}{q^{k-1}}\right)\delta.$$

Since $\nu(q - 1)$ is an integer we may replace this by

$$\nu(q - 1) \leqq q\delta - 1. \tag{5}$$

This is the best we can do, since $\delta \leqq q^{k-1}$.

By some further manipulation we find that for the alphabet a, generated by $\mathfrak{A}_0 - C(S)$, to attain the bound of Theorem 5 we must have

$$(q\delta - 1) - (q - 2) \leqq \nu(q - 1) \leqq q\delta - 1. \tag{6}$$

We also wish to have an alphabet with $\mu = 0$; for such an alphabet

$$d = \frac{sq^{k-1} - s}{Q} = s(q - 1),$$

$$\delta = d_0 - d = q^{k-1} - s(q - 1),$$

$$q\delta - 1 = (q^k - 1) - sq(q - 1);$$

i.e., $q\delta - 1$ is divisible by $(q - 1)$. From (6), the only possibility is

$$\nu(q - 1) = q\delta - 1. \tag{7}$$

We also observe from (6) that if $q = 2$ we have $\nu(q - 1) = q\delta - 1$ without any other considerations.

To justify our statement that $\mathcal{A}_2, \cdots, \mathcal{A}_{k-1}$ are the only alphabets besides $\mathcal{A}_0$ for which $\mu = 0$, we prove the following theorem.

Theorem 9: If $C(S)$ is a set of ν points in $T_{k-1}(q)$, with δ defined as above, and $\nu(q - 1) = q\delta - 1$, then $C(S)$ is the set of all points of a linear space in $T_{k-1}(q)$. This, of course, implies that

$$\nu = 1 + q + \cdots + q^s, \qquad \delta = q^s, \qquad 1 \leqq s \leqq k - 2.$$

Conversely, if $C(S)$ is the set of all points of a linear space, then the alphabet $\mathcal{A}$ has $\mu = 0$ and attains the bound of Theorem 5. This we have already verified.

Proof: We have $\nu - \delta = (\nu - 1)/q$, so that $(\nu - 1)$ must be a multiple of q. If $\nu = 1$ the corresponding alphabet is $\mathcal{A}_1$, for which $\mu = 1$.

The proof is by induction on δ; we start by proving the theorem for the case $(\nu - 1)/q = 1$; i.e., $\delta = q$, $\nu = q + 1$.

Lemma: If $\nu = 1 + q$ and $\delta = q$, the $(1 + q)$ points $X_0, X_1, \cdots, X_q$ are collinear, however large the containing space.

An equivalent statement, which is the one we prove, is: If every hyperplane of $T_{k-1}(q)$ contains at least one of the points $X_0, X_1, \cdots, X_q$, then $X_0, X_1, \cdots, X_q$ are the points of a line.

We may assume that there is one hyperplane, say $Y_1 = 0$, which contains exactly one point X_i, which we may call X_1. Pick another point for X_0 and let the coordinates of these two points be $\mathbf{e}_1$, $\mathbf{e}_2$. We assume the coordinate system normalized so that the first nonzero coordinate of every point is unity. Write the coordinates of the X_i as columns of a matrix as follows:

$$\begin{array}{c} \\ Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_k \end{array} \begin{array}{c} \begin{array}{cccccc} \mathbf{X}_0 & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 & \cdots & \mathbf{X}_q \end{array} \\ \left(\begin{array}{cccccc} 1 & 0 & 1 & 1 & \cdots & 1 \\ 0 & 1 & a_2 & a_3 & \cdots & a_q \\ 0 & 0 & b_2 & b_3 & \cdots & b_q \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & f_2 & f_3 & \cdots & f_q \end{array}\right) \end{array} .$$

By Theorem 1, every letter of the form $\alpha Y_1 + \beta Y_2$ must contain at least one zero coordinate. For $\alpha = 0$ $(\beta = 0)$ we always have a zero in the second (first) place of such a letter. In the $(q - 1)$ letters of the form $\alpha Y_1 + Y_2$, $\alpha \in F^*(q)$, the zero must occur in one of the places $2, 3, \cdots, q$. Hence the $a_2, a_3, \cdots, a_q$ above must denote some arrangement of all the elements of $F^*(q)$.

Consider now letters of the form

$$\alpha Y_1 + \beta Y_2 + \gamma Y_3 .$$

Again, the first two coordinate places take care of those letters for which one of α,β,γ is zero. Hence we restrict ourselves to letters

$$\alpha Y_1 + \beta Y_2 + Y_3 , \qquad \alpha,\beta \in F^*(q).$$

Each such letter must have a zero in one of the places $2, 3, \cdots, q$.

We note that there are $(q - 1)^2$ such letters, and $(q - 1)$ coordinate places.

Suppose now that $b_2 \neq 0$. We shall count the number of letters to which the $\mathbf{X}_2$ column contributes a zero. We may choose any α in $F^*(q)$ such that $\alpha \neq b_2$. $\beta(\neq 0)$ is then uniquely determined by the equation [in $F(q)$]

$$\beta a_2 = -(\alpha + b_2).$$

Hence if $b_2 \neq 0$ the $\mathbf{X}_2$ column contributes a zero to only $(q - 2)$ letters.

If $b_2 = 0$ we have $(q - 1)$ choices for α, and β is determined by

$$\beta a_2 = -\alpha.$$

In this case the $\mathbf{X}_2$ column contributes a zero to $(q - 1)$ letters.

Hence the only possible choice for b_i is $b_i = 0$ for all i.

The same argument shows that all rows Y_i, $i > 3$, consist entirely of zeros. The coordinates of $X_2, \cdots, X_q$ are linearly dependent on those of X_0,X_1; that is, the points $X_2, \cdots, X_q$ all lie on the line joining X_0,X_1.

Returning now to the main theorem we make the following induction hypothesis:

Let $C(S)$ be a set of ν points in $T_{k-1}(q)$, with δ defined as before, and such that

$$(q - 1)\nu = q\delta - 1. \tag{7}$$

Let $q^{r-2} < \delta \leqq q^{r-1}$, and assume that Theorem 9 is true for values of $\delta \leqq q^{r-2}$. We wish to prove that

i. $\delta = q^{r-1}, \qquad \nu = 1 + q + q^2 + \cdots + q^{r-1}$.

ii. $C(S)$ consists of all points of a linear space of projective dimension $(r - 1)$.

From (7), $\nu - \delta = (\nu - 1)/q = h$, where h is an integer greater than 1. $h = 1$ is the case already considered in the Lemma. Also

$$\delta = \nu - h = hq + 1 - h.$$

An arbitrary space of dimension $(k - 3)$, say D_{k-3}, in $T_{k-1}(q)$ will contain a number α of points of $C(S)$. We wish to find a lower bound $\bar{\alpha}$ for α.

There are $(q + 1)$ hyperplanes of $T_{k-1}(q)$ which pass through D_{k-3}. Denote by $\beta_0, \beta_1, \cdots, \beta_q$ the number of points of $C(S)$, outside of D_{k-3}, contained by these hyperplanes. The hyperplanes through D_{k-3} contain among them all points of $T_{k-1}(q)$, so certainly all of $C(S)$. We have then

$$\begin{gathered} \alpha + \sum_{i=0}^{q} \beta_i = \nu = hq + 1, \\ \alpha + \beta_i \geqq \nu - \delta = h. \end{gathered} \tag{8}$$

A lower bound for α is obtained by making all the β_i equal, $\beta_i = \bar{\beta}$, and replacing "$\geqq$" by "$=$" in (8). Then,

$$\bar{\alpha} + \bar{\beta} = h,$$

$$\bar{\alpha} + (q + 1)\bar{\beta} = hq + 1.$$

Solving these equations,

$$\bar{\alpha} = \frac{h - 1}{q}.$$

$$\bar{\beta} = h - \frac{h - 1}{q} = \frac{\delta}{q}.$$

Let α' be the least integer containing $\bar{\alpha}$. We note that $\alpha' > 0$.

We may assume that some hyperplane, say H_{k-2}, of $T_{k-1}(q)$ contains exactly $\nu - \delta = h$ points of $C(S)$. Call this set of points $C(S')$. Each hyperplane of H_{k-2} is a $(k - 3)$-dimensional subspace of $T_{k-1}(q)$, and so by the previous result it contains at least α' points of $C(S')$.

For $C(S')$ we have

$$\nu' = h, \qquad \delta' \leqq h - \alpha' \leqq h - \frac{h - 1}{q}.$$

Therefore,

$$q\delta' - 1 \leqq qh - h + 1 - 1 = (q - 1)h,$$

or

$$\nu'(q - 1) \geqq q\delta' - 1.$$

Comparing this with (5), only equality is possible; i.e.,

$$q\delta' - 1 = (q - 1)h.$$

This implies that $(h - 1)/q$ is an integer, and

$$\delta' = h - \frac{h - 1}{q} = \frac{\delta}{q}.$$

$C(S')$ is thus a set of points with

$$\nu' = h, \qquad \delta' = \frac{\delta}{q} \qquad \text{and} \qquad (q - 1)\nu' = q\delta' - 1.$$

Since $q^{r-3} < \delta/q \leqq q^{r-2}$ we can apply the induction hypothesis, which gives us

$$\delta' = q^{r-2}, \qquad \nu' = 1 + q + \cdots + q^{r-2}$$

or

$$\delta = q\delta' = q^{r-1}, \qquad \nu = q\nu' + 1 = 1 + q + \cdots + q^{r-1}$$

and the points $C(S')$ are all the points of a linear space B_{r-2} in H_{k-2}.

We can always find in H_{k-2} a $(k - 3)$-dimensional subspace, say D_{k-3}, which intersects B_{r-2} in a space of dimension $(r - 3)$, and thus contains exactly

$$1 + q + \cdots + q^{r-3} = \frac{h - 1}{q} = \bar{\alpha}$$

points of $C(S)$.

Consider the hyperplanes of $T_{k-1}(q)$ which pass through D_{k-3}. From (8), we have for these

$$\beta_i = \bar{\beta} = h - \frac{h - 1}{q},$$

so that the total number of points of $C(S)$ in each hyperplane is

$$\bar{\alpha} + \bar{\beta} = h.$$

By the previous argument the intersection of $C(S)$ with each hyperplane is a linear space of dimension $(r - 2)$. These spaces have in common a linear space of dimension $(r - 3)$, the intersection of B_{r-2} and D_{k-3}. Hence the set of all their points is a linear space of dimension $(r - 1)$. This proves the theorem.

We will now summarize the results of the last section. For $q = 2$, the alphabets $\mathfrak{A}_0, \mathfrak{A}_1, \cdots, \mathfrak{A}_{k-1}$ introduced at the beginning of the section are the only alphabets which attain the bound of Theorem 5. For $q > 2$ these alphabets attain this bound, and have the further property that any k-dimensional alphabet obtained from them by removing up to $(q - 2)$ arbitrary columns of the generator matrix ($q - 1$ for $\mathfrak{A}_0$) also attains this bound. They are the only alphabets with this property. Clearly the alphabets $\mathfrak{A}_0, \mathfrak{A}_1, \cdots, \mathfrak{A}_{k-1}$ are completely determined, up to equivalence, by the values of the parameters k,d,n. For a given k, d and n are restricted to a certain set of values defined at the beginning of this section.

VI. ACKNOWLEDGMENTS

The author would like to thank D. Slepian and J. B. Kruskal of Bell Telephone Laboratories and Professor A. J. Gleason of Harvard University for their careful and critical reading of the manuscript and for several suggestions which resulted in substantial improvements throughout. Previous work in the field of nonbinary error correction, not referred to in the paper, is described in Refs. 7 through 10.

APPENDIX

Slepian's Error-Correction Procedure

Let $F(q)$ denote a finite field, and $G_n(q)$ the group, of order q^n, of all possible rows of n symbols picked from $F(q)$. The group operation is place-by-place addition under the rules prevailing in $F(q)$. Let A be a subgroup of G_n. [For the present purposes A need not be a vector space over $F(q)$; the two concepts are the same if and only if $F(q)$ is a prime field.]

Partition G_n into cosets with respect to A, with an element of least weight in each coset being picked as "coset leader." The element $(00 \cdots 0)$ is, of course, the coset leader of A itself. The cosets are formed into a table as illustrated in Table III. The group A is the first row of the coset table. The first column of the table contains the coset leaders. In the case of Table III these are, besides (0000), all the elements of weight 1 in $G_3(3)$.

The element in the sth row and the tth column of the coset table is obtained by adding the sth coset leader to the element (of A) in the first row and the tth column. The sth row is exactly the coset determined by the sth coset leader, and every element of $G_n(q)$ appears exactly once in the table.

TABLE III — COSETS WITH RESPECT TO A FOR $G_n(q) = G_3(3)$

	1	2	3	4	5	6	7	8	9
1	0000	1011	0112	1120	1202	2022	0221	2210	2101
2	1000	2011	1112	2120	2202	0022	1221	0210	0101
3	2000	0011	2112	0120	0202	1022	2221	1210	1101
4	0100	1111	0212	1220	1002	2122	0021	2010	2201
5	0200	1211	0012	1020	1102	2222	0121	2110	2001
6	0010	1021	0122	1100	1212	2002	0201	2220	2111
7	0020	1001	0102	1110	1222	2012	0211	2200	2121
8	0001	1012	0110	1121	1200	2020	0222	2211	2102
9	0002	1010	0111	1122	1201	2021	0220	2212	2100

The error-correction procedure is as follows: If the received element is a letter of A it is accepted as correct. If not, it is located in the coset table, say in row s, column t, and the letter of A in row 1, column t is substituted.

It is clear that the example of Table III will correct all single errors. Column 2 contains, besides (1011) which belongs to A, all the elements of $G_3(3)$ which differ from (1011) in exactly one place.

In general, if it is required to correct all single, double, etc., errors it is necessary that all elements of $G_n(q)$ of weights 1, 2, etc., appear as coset leaders in the coset table formed by A. Let d be the minimum weight of the letters of A, other than zero. The coset formed by a leader of weight 1 will consist of elements of weight at least $(d - 1)$. Hence all elements of $G_n(q)$ of weight 1 appear as coset leaders if and only if $d \geqq 3$. Similarly, all elements of weight 2 appear as coset leaders if and only if $d \geqq 5$. If it is required to correct all e-fold errors, the alphabet A must have $d \geqq 2e + 1$.

REFERENCES

1. Hamming, R. W., Error Detecting and Error Correcting Codes, B.S.T.J., **29,** 1950, p. 147.
2. Slepian, D., A Class of Binary Signaling Alphabets, B.S.T.J., **35,** 1956, p. 203.
3. Bose, R. C. and Kuebler, R. R., Jr., A Geometry of Binary Sequences Associated with Group Alphabets in Information Theory, Ann. Math. Stat., **31,** 1960, p. 113.
4. Carmichael, R. D., *Introduction to the Theory of Groups of Finite Order*, Dover, New York, 1956.
5. van der Waerden, B. L., *Modern Algebra*, Vol. 1, Ungar, New York, 1949.
6. McClusky, E. J., Jr., Error-Correcting Codes — A Linear Programming Approach, B.S.T.J., **38,** 1959, p. 1485.
7. Zaremba, S. K., Covering Problems Concerning Abelian Groups, J. Lond. Math. Soc., **27,** 1952, p. 242.
8. Ulrich, W., Non-Binary Error Correction Codes, B.S.T.J., **36,** 1957, p. 1341.
9. Golay, M. J. E., Notes on the Penny-Weighing Problem, Lossless Symbol Coding with Nonprimes, I.R.E. Trans., **IT-4,** 1958, p. 103.
10. Cocke, J., Lossless Symbol Coding with Nonprimes, I.R.E. Trans., **IT-5,** 1959, p. 33.

Short-Term Memory in Vision

By E. AVERBACH and A. S. CORIELL

(Manuscript received August 8, 1960)

Experiments are performed that demonstrate some of the functional properties of short-term storage in the visual system, its decay, readout and erasure. Results indicate that the visual process involves a buffer storage which includes an erasure mechanism that is local in character and tends to erase stored information when new information is put in. Storage time appears to be of the order of one-quarter second; storage capacity is more difficult to assess.

I. INTRODUCTION

There can be little doubt that eye movements play an important role in the perception of form, and that perceptions of complicated visual fields are built up from information gathered during many fixations of the eyes. But eye movements over a complicated visual field are unpredictable from subject to subject and from time to time with the same subject. They may therefore be an annoying source of variability in perceptual experiments, and experimenters frequently find it desirable to eliminate them. This is usually done by using a tachistoscope, a device for presenting brief exposures of visual material. The position of the eyes is kept fixed at the crucial time by having the subject fix his eyes *before* exposure of the material, and by using exposures sufficiently brief that the subject cannot change his fixation during the exposure. To accomplish this, exposure times must be less than the reaction time of the eye for a change of fixation (150–200 milliseconds). Actual measurement in a tachistoscopic situation has shown that exposures of 100 milliseconds or shorter satisfy this purpose.[1]

At first thought, it may seem unreasonable to study visual perception under the peculiar condition of tachistoscopic experiments. Some question might be raised about whether data obtained in this way can be generalized to natural perceptual situations. It can be argued, however, that in a very real sense the tachistoscopic situation is not an unnatural one. For it is well established that, in scanning pictorial or printed material, the eyes do not take in information continuously (Ref. 2, p. 493). They fixate first on one point and then move rapidly to another. The fixations are relatively long, but the movements between fixations are so quick that they smear the image drastically during the motion. Thus, normal vision involves the processing of discrete exposures very much like those presented in a tachistoscope. It has been shown, in fact, that reading performance is better if the necessity of moving the eyes is eliminated by presenting reading material serially by means of a tachistoscope.[3]

To anyone who has ever seen objects illuminated briefly by a spark or other kind of brief flash it is evident that the visual impression of the illumination lasts longer than the flash. Even a millisecond flash seems to last a noticeably long time. Because of this persistence, writers on perception, particularly tachistoscopic perception, have assumed the existence of some kind of short-term storage in the visual system. This is implied in their use of such terms as positive afterimage, retinal persistence, persistence of vision, etc., in interpreting tachistoscopic performance. But little work has been done to characterize the functional properties of the storage, its decay, readout and erasure. In this paper we will discuss some of the older studies that bear on these matters and report a few experiments that demonstrate some properties of this visual short-term storage.

II. MEMORY EXPERIMENT

A very old tachistoscopic experiment is the span-of-perception experiment. Its aim is to determine the maximum number of objects a person can take in at a glance, the objects being dots, letters, digits, words, etc. Typically, the experimenter makes up cards having different numbers of items on them. Starting with cards having one item, he keeps increasing the number of items presented until the subject begins to make errors. The perceptual span may be taken to be the maximum number of items that the subject can report perfectly. More usually, the criterion used is the number of objects reported correctly 50 per cent above chance.

Spans of perception measured by different investigators are surprisingly consistent considering the wide range of conditions under which these spans have been measured. The span for letters or words[4,5] is $4\frac{1}{2}$ to 5 and for dots[5,6] about 8. What limits the span of perception? Of course, anything that affects legibility — brightness, contrast, sharpness, etc. — will under some conditions affect the span. But once reasonable legibility is obtained, increasing the brightness and contrast and sharpness does not improve performance. Under conditions of good legibility the limitation is elsewhere.

Two possibilities suggest themselves. First, the span may be limited by the *capacity* of the visual storage. It may be that, as the number of items put into the storage is increased, resolution of the individual items is destroyed. The other possibility is that resolution of the storage is perfectly adequate for numbers of items of the order used in span of perception experiments, but that the *storage time* is too short; i.e., the subject does not have enough time to read more than a few items into his more permanent memory before the decay of the short-term storage.

Selecting between these alternatives presents something of a problem. How does the experimenter determine how many items a subject can store visually if the subject, as shown by span of perception experiments, can report correctly only a limited number (4 or 5) items? This difficulty was circumvented by Sperling[7] who, instead of requiring that his subjects report on the whole of a complicated tachistoscopic presentation, had them report on only a part. He exposed briefly three rows of four randomly chosen letters each. Then, after a variable delay, he presented a tone signal of either high, middle or low pitch which indicated to the subject that he was to report on the upper, middle or lower row of letters — whichever was indicated by the tone. Since the subject was not familiar with the arrays of letters and was not given the instruction tone until the visual stimulus was turned off, he had, in effect, to store the whole array. By this method Sperling was able to show that subjects can store as many as 9 letters of a 12-letter array — and even more when arrays having more letters are used.

The experiment to be described, although conceived independently of Sperling, is of essentially the same form as his. The essential difference lies in the use of a *visual* signal to designate the part of the array to be reported by the subject. This has the virtue that it assures that the array and signal are transmitted to wherever they are processed in the brain at approximately the same rate. It is known that the reaction time to a light is significantly longer than the reaction time to a tone (Ref. 2, p. 16).

A 2 × 8 array of randomly chosen letters is exposed for 50 milliseconds. Then, after a variable delay, a black bar marker is presented for 50 milliseconds either above one of the letter positions of the upper row or below one of the letter positions of the lower row. The subject's task is to name the letter designated by the marker. A typical array and bar marker are shown in Fig. 1. A black circle that is used as a marker in a second experiment to be described later is also shown. The subject, of

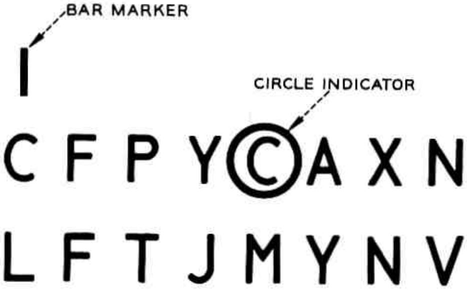


Fig. 1 — Typical array of letters, bar marker and circle indicator.

course, never knows before a given trial which letters will appear in the array, and which of the 16 letter locations will be called for by the marker. Thus, in order to perform well, he is required to store the array until the appearance of the marker. The sequence is illustrated in Fig. 2.

A uniform field of 70 foot-lamberts was maintained constantly throughout the experiment, and the letters and marker appeared black against this background. This test field subtended a visual angle of 4 degrees vertically by 5 degrees horizontally at the viewing distance of 5 feet. It had a small dark fixation point in the center. Surrounding this field was a larger field, 12 degrees on a side, having a luminance of 30 foot-lamberts. Each letter subtended one degree vertically by one-half degree horizontally. The black of the letters had a brightness of less than one foot-lambert.

2.1 *Procedure*

At the beginning of each session subjects were given two or three minutes to adapt to the bright screen. They were then given two warm-up trials with arrays that were not used in the experiment proper. On each trial the subject was given a ready signal and enough time to fix his eyes on the fixation point at the center of the screen. When fixated, he would signal the experimenter and the array and marker were then exposed. The subject was given as much time as he needed to make his response, but was encouraged to use his first guess if he was in doubt. He was given the correct answer after each trial.

During each session 128 array-marker pairs were exposed, covering each of the 16 positions at each of 8 time intervals between array and marker. The same order of presentation of the 128 arrays was used in each session of the experiment, but the time interval and marker posi-

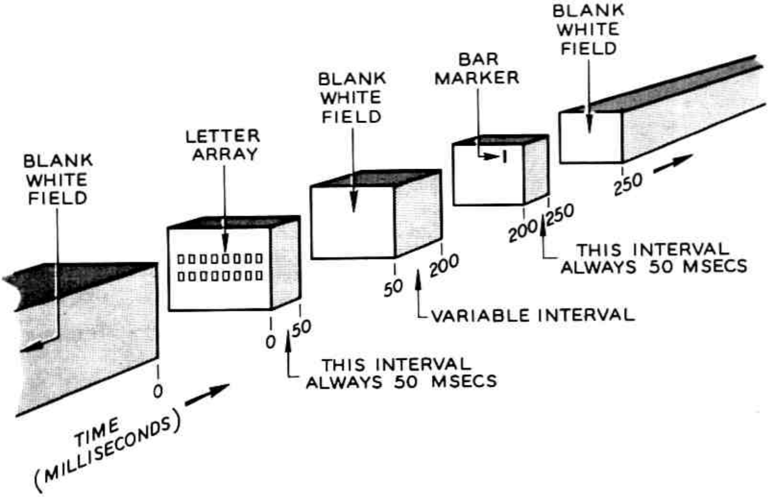


Fig. 2 — Sequence presented in a typical trial.

tions were varied randomly, with the restriction that successive groups of 16 arrays each contained a marker in each of the 16 positions. Three sessions were run with each of the three subjects.

2.2 *Apparatus*

The tachistoscope used is that designed by Nielsen,[8] which uses multichannel television generating equipment and a set of gates controlled by timers for presenting a sequential display of three pictures on a single picture monitor. Each picture can be displayed for a preset time interval of $N/60$ seconds, where N is a number of television fields from 1 to 99. Since all parts of a picture are not exposed simultaneously and a particular point on the monitor is illuminated for only 20 microseconds, exposure times are taken from the time a particular point is first scanned to the time the same point is scanned in a new picture. An exposure of 50 milliseconds, preceded and followed by a white field, is illustrated in Fig. 3, which also shows the brightness of a point near one of the letters. Interlace is ignored since the center-to-center separation of the scanning lines subtended less than one-half minute of arc at the observer's eye.

The time between the onset of an array and the onset of a marker is never an integral number of fields, owing to the fact that the marker does not appear in the same part of the vertical scan as the letter it designates. This small error has been ignored, since it is only -3 milli-

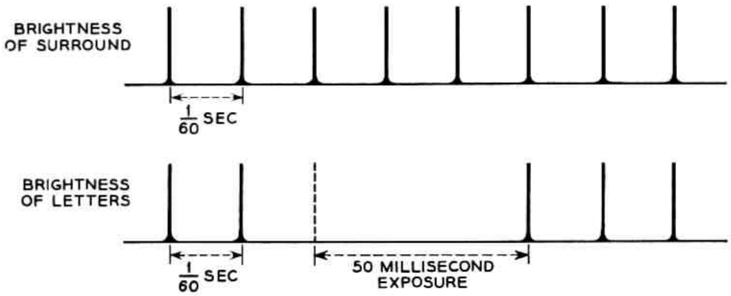


Fig. 3 — Brightness of letter and surround for exposure of 50 milliseconds preceded and followed by a white field.

seconds if the marker falls above the array and +3 milliseconds if the marker falls below.

2.3 *Results*

The results for the three subjects are shown in Fig. 4. The abscissa is the time in milliseconds between the onsets of array and marker. The ordinate is the per cent correct, corrected for chance on the assumption that the subject perceives correctly a certain percentage of the time P_p and guesses randomly from the 26 possible letters when he does not perceive correctly. On this assumption, the measured per cent correct $P_M = P_p + (1 - P_p)\ (1/26)$, which yields the plotted P_p's. Estimates of the standard error of these points, which are a function of the number of trials (48) and the per cent correct, range from 0.07 at 50 per cent to 0.06 at 20 and 80 per cent correct.

The vertical lines through zero and 50 milliseconds represent the onset and offset of the arrays. Negative time means that the marker came before the array. The point at zero time was taken after the rest of the experiment was completed because it required modification of the apparatus.

2.4 *Discussion*

Although it might be assumed that this experiment yields a reasonably good description of the time-decay of the short-term visual storage, the curves obtained cannot be said to represent this decay for two reasons. First, the true storage would be expected to decay to zero for long enough time intervals. But these results decay to a final level of about 35 per cent for two of the subjects and 25 per cent for the third. Second, because the process of detecting a marker and reading a letter

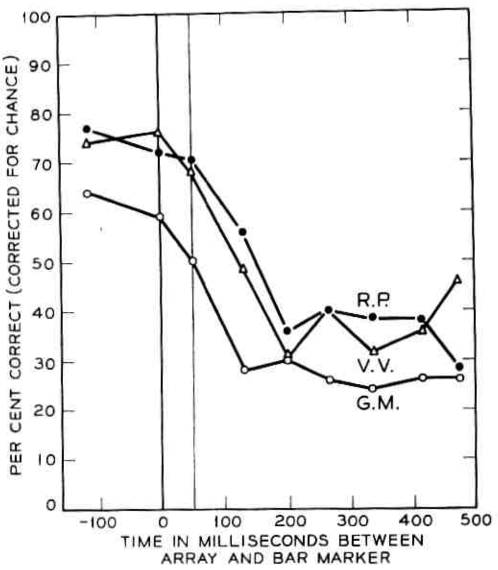


Fig. 4 — Results of memory experiment; "R.P.," "V.V." and "G.M." identify the three subjects participating.

undoubtedly takes time, the measured performance suggests a storage time that is shorter than the true storage time.

The fact that the measured decay curves do not fall to zero suggests that the measured performance contains components of a more permanent memory, as well as the short-term memory component that we would like to measure. In this context, the 25 to 35 per cent final performance level (which represents 4 to 5.6 letters) is attributed to what the subject has read into his more permanent memory.

Maximum performance measured when the marker preceded the array is 65 to 80 per cent. It is obvious, of course, that, if the marker preceded the array by a long enough time, performance would reach 100 per cent. Why, then, doesn't performance reach 100 per cent? The reason is *not* that some letters are outside the fovea, since individual letters exposed in any of the 16 positions of the array are clearly legible. The explanation seems to lie, rather, in the fact that letters in some positions, although perfectly legible by themselves, are not legible in the context of the array. This finding is illustrated by the plot of performance as a function of position shown in Fig. 5. The numbers 1 to 8 represent, from left to right, the positions on the upper line of the array, and 9 to 16 the positions on the lower line. The percentage is based on the pooled

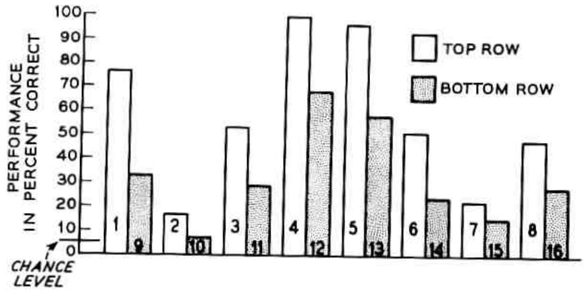


Fig. 5 — Performance by position in array, three-subject average.

data of the three subjects, each point being based on 72 trials taken across all time intervals. All subjects show the same distribution, in which performance is better at the center and ends, and poorer in between. Performance on the upper line is consistently better than performance on the lower.

2.5 *Summary*

In summary, the following can be stated:

1. The visual system can store information for longer than 130 to 200 milliseconds.

2. This storage can be tapped selectively on a signal given by the experimenter.

3. Resolution of the storage — or ease of reading-out — is disturbed when too much data is put in. Sixteen letters in a 2 × 8 matrix is enough to demonstrate this effect.

4. This disturbance does not affect all items of such a stored array equally. It disturbs the center and end items least and those in between most.

5. As an exercise, we estimated the amount of information in the store when the bar marker follows immediately after the array, and obtained the figure of 37 bits for the poorest subject and 54 bits for the other two subjects.

III. ERASURE

If persistence were the only property of the visual storage, it would be difficult to understand how we see at all in our normal, continually changing environment. A storage process normally also involves erasure, to assure that old information is out of the store before new information is put in. Otherwise, new information and old would be inextricably merged in the store. The experiment to be described deals with the erasure properties of the visual storage.

The procedure in this erasing experiment was almost identical to that of the memory experiment just described. The same subjects were used, and the same arrays of letters were presented in the same sequence. The essential difference between the two experiments was in the form of the marker used. In the first experiment the marker consisted of a vertical bar pointing to the letter; in this experiment it consisted of a black circle surrounding the letter, as was illustrated in Fig. 1. Such a circle produces a curious effect upon the letter if the time delay between array and circle is chosen properly. This effect we call *erasure*.

3.1 *Results*

Fig. 6 permits a comparison of the performance by each of our subjects in the bar marker experiment with his performance in the circle experiment. The curves of all three subjects start at a relatively high level, ranging from 70 to 80 per cent, drop sharply to a minimum, ranging from 10 to 20 per cent and rise slowly to an intermediate level of 25 to 40 per cent.

When the circle precedes the array or follows immediately after, performance in the erasure experiment is not greatly different from performance in the first experiment. However, when the circle is delayed by 100 milliseconds, the difference between the curves is quite large. Then, with still longer time delays, performance in the circle experiment rises slowly until it reaches approximately the values obtained in the first experiment. Thus, the curves begin together and end together, with performance in the "circle" experiment significantly poorer between.

3.2 *Discussion*

This experiment shows how a later visual stimulus can drastically affect the perception of an earlier one. This backward-in-time action of the circle implies that the first stimulus is delayed with respect to the second, or, more precisely, that the first stimulus is stored. The process involves more than simple delay, since, as shown by the first experiment, the subject has access to the information during the delay.

The question arises as to why the circle has such a damaging effect on the letters and why the bar does not. The answer lies in their relative distance from the letter. In preliminary experiments it was found that,

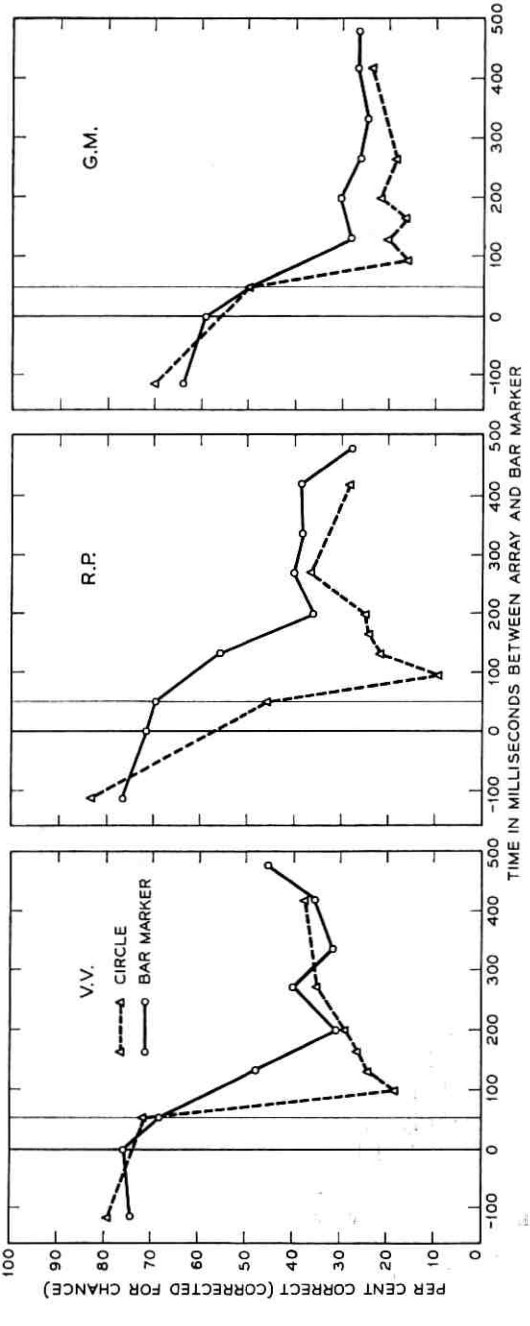


Fig. 6 — Results of erasure experiment compared to those of memory experiment.

when the bar was close to the letters, those parts of the letter near the bar were strongly disturbed. The bar was therefore placed far enough away from the letters to avoid this effect. All parts of the circle, however, are close to the letter.

One can conceive of the observed action of the circle on a preceding letter in many different ways. It is possible, for example, that the main effect on the letter is a "stopping in time", i.e., a quick substitution of the circle for the stored letter. Such a process could function as an erasure mechanism, since it would assure that new and old information are not confused in the store. It is conceivable, on the other hand, that the effect of the circle on the letter is of a different kind. Perhaps the disturbance is a result of the kind of mixing or averaging process that an erasure mechanism seeks to avoid, or perhaps the effect of the circle is primarily to reduce the brightness or contrast of the letter.

We are inclined to reject the latter alternatives. The observed effect is clearly *not* due to averaging in time, for if the circle and the letter it surrounds are presented simultaneously, the legibility of the letter is hardly affected at all. Yet this is just the condition for which averaging should produce the most damaging effect. With regard to the possibility that the circle affects the brightness or contrast of the letters we can say nothing conclusive. Introspectively, however, a change in brightness or contrast does not appear to be the primary effect.

The view that the second stimulus limits the time available for reading-out is more attractive than the other possibilities for several reasons. First, the rise found in the erasure curve with increasing delay of the circle after 100 milliseconds is consistent with the idea that increased delay of the circle allows more time for readout. A second reason for believing that the effect of the circle is primarily to limit readout time stems from our observation that the lowered performance obtained when the circle follows the letter is not independent of the number of letters involved. If a single letter is exposed in any one of the sixteen positions of the array and a circle is presented with a 100-millisecond delay — this is the delay that yields the poorest performance on our curves — this letter will be read correctly by our subjects 100 per cent of the time. On the other hand, if we expose four letters — which can ordinarily be reported perfectly — followed by the circle with this same 100-millisecond delay, performance is disturbed dramatically. Thus, we believe that the effectiveness of the circle in disturbing performance is related to how much reading has to be done before the circle appears. If only one letter must be read, the circle does not affect performance measurably because, according to this interpretation, the subject has

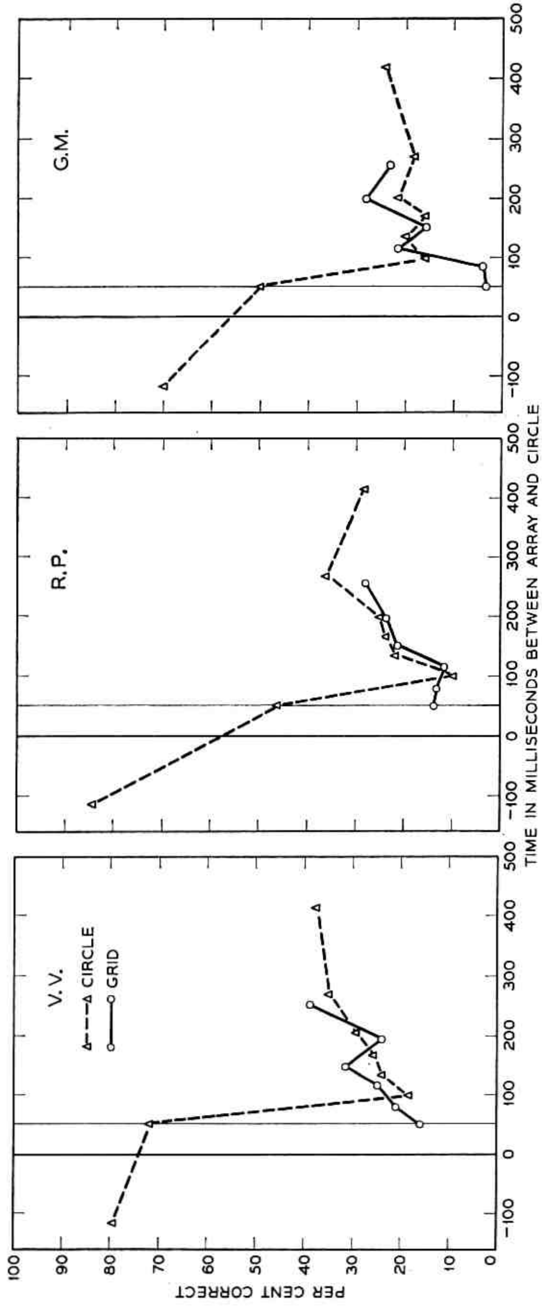


Fig. 7 — Results of erasure experiment, using circle and grid.

enough time to read the letter before the circle appears; if many letters must be read, the subject will not have enough time to read all of the letters before the circle appears. It is then likely that the circle will erase the letter it follows before it is read. Our final reason for thinking that the circle limits the time available is an introspective one. We find that a briefly exposed letter followed by a circle — even when it is seen as perfectly as it is when a single letter is exposed — seems to persist for a much shorter time than it does when it is not followed by a circle.

We are therefore inclined to say that the effect of the circle is to remove previously stored information. On this interpretation, the observed increase in performance with increasing delay is attributed, not to loss of erasing effectiveness of the circle, but to the increased time available for readout.

In the light of the above, the shape of the erasure curve may be interpreted as follows:

1. High performance when the circle follows immediately after the array is due to simple temporal averaging in the visual system. This results in array and circle being effectively superimposed, which does not significantly affect legibility.

2. Decreased performance at slightly longer delays can be attributed to the change from the superposition condition to the erasure condition.

3. The slow rise from the minimum with further increases in delay of the circle is attributable to the increased time available for the subject to read the letter before it is erased. At still longer times, when performance is about the same as in the bar experiment, the circle no longer erases but acts as a marker. This suggests, as outlined in the discussion of the first experiment, that performance at times longer than 200 milliseconds depends not on the contents of the short-term storage at that time, but on the number of letters that had been read into the permanent memory before that time.

The suggestion that two closely timed stimuli are perceived as being superposed is testable. In what was essentially a repeat of the experiment just described, we substituted a circle filled with grid lines for the unfilled erasing circle. When simple superposition holds, such an "eraser" should be much more interfering than a simple circle. The results are shown in Fig. 7.

The dashed lines are plots of the results obtained using an unfilled circle. The solid lines give the results obtained with the filled circle. It is seen that for delays longer than 100 milliseconds the difference between the two curves is small; while for delays of less than 100 milliseconds, the difference is large. This confirms that the disturbing effect

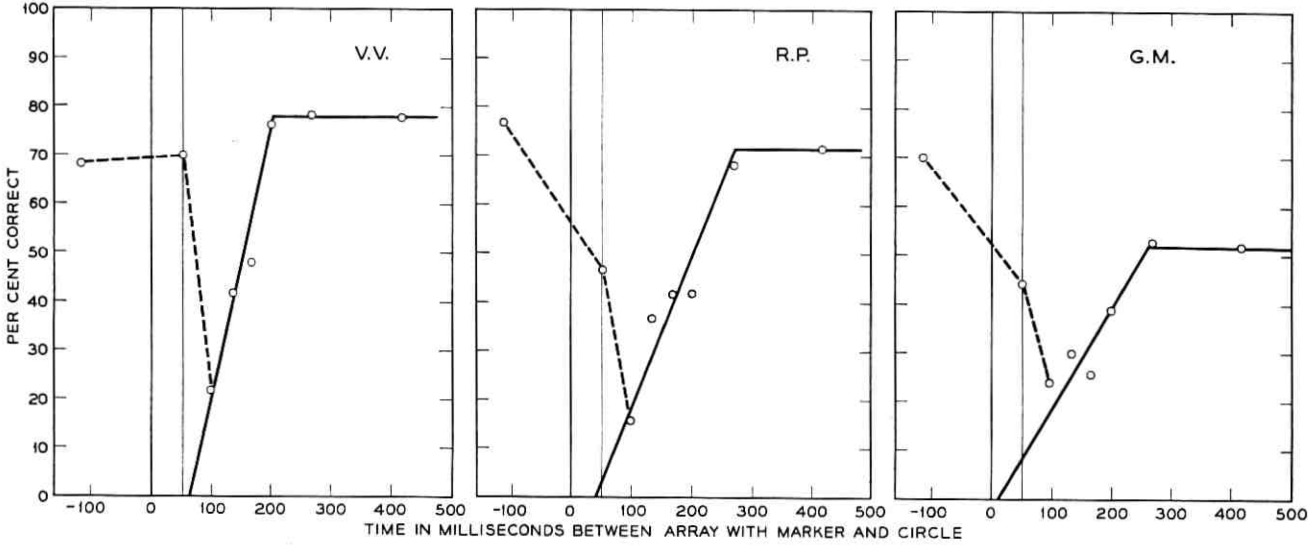


Fig. 8 — Results of readout experiment.

of a later stimulus on a preceding one may be of at least two varieties: the superposition type, which is understandable in terms of averaging; and the other effect, which we have called erasure.

It should be mentioned that preliminary work has been done in which an array of eight letters is presented to one eye and the circle delivered to the other. It is found that erasure occurs under these conditions although it has not yet been determined how this binocular erasure compares with monocular erasure.

IV. READOUT

As we have already pointed out in discussing the bar-marker experiment, the process of detecting the presence of a marker and reading the marked letter undoubtedly takes some time. If the time required for this process could be measured, we would be able to correct the decay characteristic obtained in the bar marker experiment for this time and have a more accurate idea of the duration of the storage. A method for measuring this time is available provided that our conclusions about the action of the circle in erasing a letter are correct. Suppose we present simultaneously an array and a bar marker pointing to one of the letters in the array. Then, a short time afterwards, suppose we present an erasing circle around the marked letter. If the circle indeed removes the marked letter from the subject's storage, his performance under these conditions will measure how well he can detect the marker and read the letter when given only the time interval between the onset of the array and marker, and the onset of the circle. Such an experiment was performed, in fact, using the same subjects and experimental conditions as before.

4.1 *Results*

The results appear in Fig. 8. The abscissa is the time between the onset of the array–bar-marker combination and the onset of the circle. The three curves are similar in form. It is seen that when the circle follows by more than 100 milliseconds performance rises rapidly as a function of the time between array and circle. This is true up to 200 milliseconds for subject "V.V." and 270 milliseconds for the other two subjects, later presentations seeming to have no further effect. Performance when the circle follows by less than 100 milliseconds is very much like that obtained in the erasure experiment using a circle without a bar marker.

4.2 *Discussion*

Results of this experiment indicate that it takes a significant time for subjects to detect a marker and read the designated letter, the level of performance being a function of the time available for detection and reading. Maximum performance requires times of the order of 200 to 270 milliseconds. Thus, the decay curves from the first experiment incorporate two effects: (a) storage time and (b) readout time. As we shall see in the next section, it is possible by means of the readout time measurement to correct for the latter factor and solve for the storage time alone.

V. STORAGE TIME

We indicated in Section III that performance in the bar-marker experiment is probably the result of two different types of performance on the part of the subject. First, there is a nonselective readout, which is independent of appearance of the marker; second, there is a selective process, which occurs only after the marker appears, when the subject has been cued to direct his attention to the single desired letter. The nonselective process is indicated by the finding that the subject's performance does not fall to zero even when the bar marker appears at relatively long times (450 milliseconds) after appearance of the array, at times when, presumably, the short-term storage has already decayed. It was separated out and measured in the erasure experiment, in which it is apparent that, if the subject has not read the designated letter from his short-term storage before the circle appears, he cannot read it later because the letter is erased by the circle.

We have no direct measure of the selective readout component. The effect of this component can be derived from the original bar-marker curves by subtracting out the nonselective component from the whole. Fig. 9 shows the result obtained by subtracting percentages obtained in the erasure experiment from those obtained in the bar-marker experiment. The subtraction is not a simple algebraic one. It is clear that if the subject reads out the correct letter by chance before the marker appears, designation by the marker cannot improve his performance. It therefore seems reasonable to treat the probabilities of reading letters before and after appearance of the marker as independent.

Designating these probabilities as P_B and P_A respectively, the total probability of reading the letter is $P_T = P_B + (1 - P_B)P_A$. P_T is the per cent correct in the first experiment, P_B that in the erasing experiment and P_A, the per cent in Fig. 9, is calculated from $P_A = P_T - P_B/(1 -$

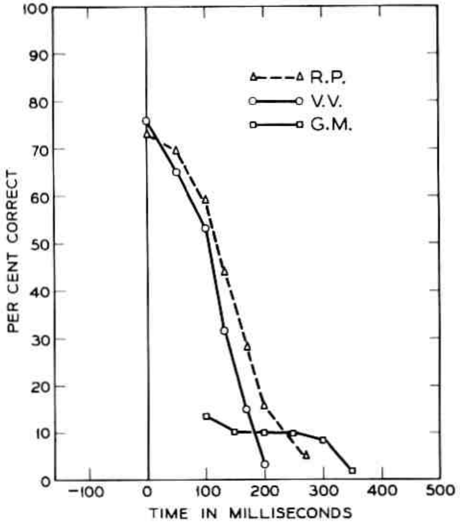


Fig. 9 — Derived "selective readout" performance curves.

P_B). The initial drop in the erasure curve is ignored because, as shown, it involves superposition and not simple erasure.

The derived curves for subjects "V.V." and "R.P." are quite similar. They start at their maxima and drop to zero as would be expected. That of subject "G.M." has the peculiar shape it does because the slope of his circle erasure curve between zero and 100 milliseconds is indeterminate. If this slope is estimated from the "filled circle" erasure curve, the the characteristic shown in Fig. 10 is obtained. This decay is quite similar to that of the other subjects.

Using these derived curves, it is possible to estimate the duration of the short term visual storage. Note that the curves in Figs. 9 and 10 represent that component of the subject's performance accomplished after appearance of the marker. We will assume that this component of performance is limited by the time available to detect the marker and read the letter before decay of the storage. We have already determined experimentally (see Fig. 8) the times required to detect a marker and read a letter to various levels of performance. By adding these readout times at each level of performance to the appropriate times in Figs. 9 and 10, estimates of the storage time are obtained. This process and the result are illustrated in Fig. 11.

The solid lines represent the selective readout components taken from Fig. 9 and, for subject "G.M.," from Fig. 10. The empty points were

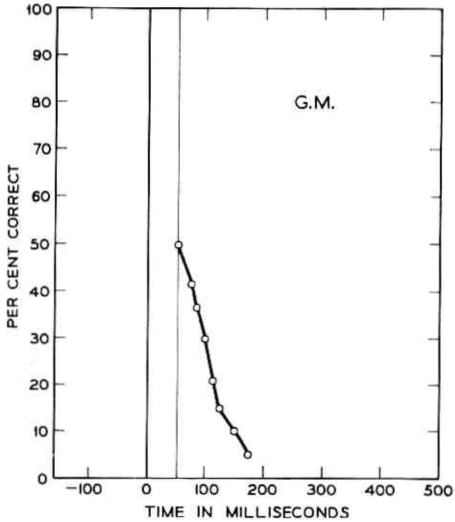


Fig. 10 — "Selective readout" performance curve for subject "G.M." derived from grid experiment.

obtained by adding times at various levels of performance taken from Fig. 8 to the points at the same levels of performance on the solid curve. Each point is therefore an independent estimate of storage time. It is seen that these points approximate vertical lines surprisingly well. The estimated storage times are 300 milliseconds. for "R.P." and 250 milliseconds for the other two subjects.

VI. CONCLUSIONS

In the light of the experiments reported here, the following interpretation seems plausible: The visual process involves a buffer storage whose read-in is very fast and readout relatively slow. The storage includes an erasure mechanism whereby new information put into the storage tends to erase what was previously there. This erasure is local in character, since erasure of a given detail depends on its distance from the areas where new detail is being put in. The storage time is of the order of one-quarter second. The storage capacity is more difficult to assess. A lower bound on the storage capacity, computed from performance obtained when bar marker follows immediately after array, yields a figure of 37 bits for the poorest subject and 54 bits for the other two subjects. This figure seems quite high, considering that the letter arrangement used in the experiment and the sharpness and contrast of the letters

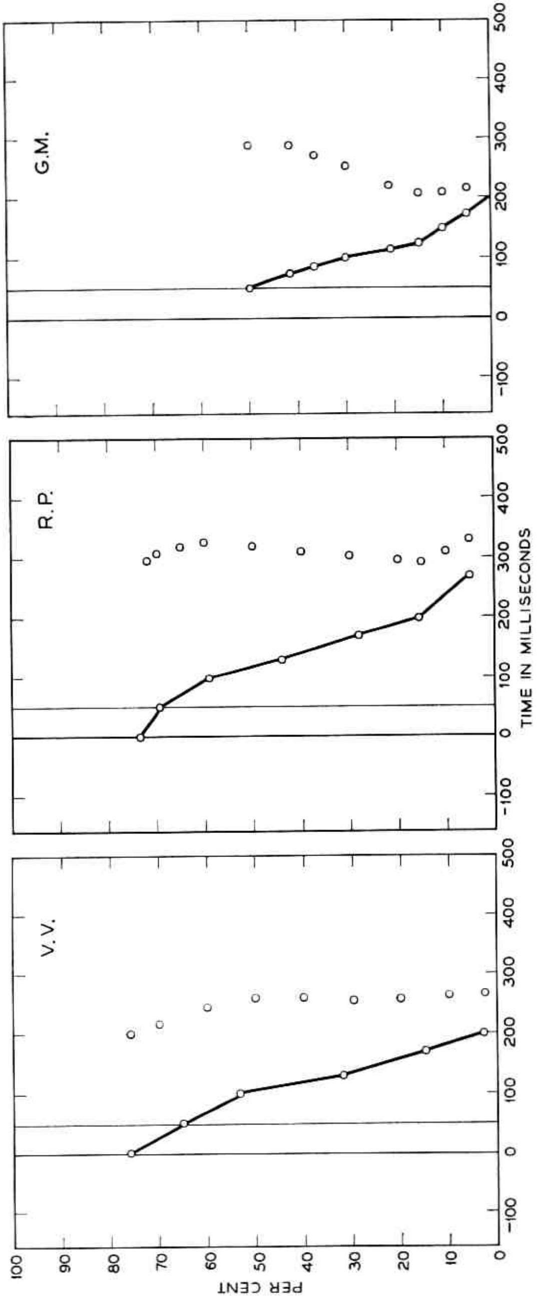


Fig. 11 — Derived storage time.

were not such as to maximize performance. The highest figure, obtained in span-of-perception experiments is 25.36 bits. Sperling,[7] using a technique similar to ours, also not involving complete report, obtained a value of 64 bits.

The interpretation above, of course, is a tentative one. These experiments have raised many more questions about short-term storage than they have answered. Particularly compelling are the questions of how the storage is scanned, and the stimulus factors involved in erasure. It is hoped that further application of the techniques used here will shed more light on these matters.

VII. ACKNOWLEDGMENT

We would like to thank R. E. Graham and E. E. David for their helpful comments during the course of the experiments and the writing of this paper. We would also like to express our appreciation to V. A. Vyssotsky, S. E. Michaels and R. A. Payne, who kindly served as experimental subjects.

REFERENCES

1. Dodge, R., An Experimental Study of Visual Fixation, Psych. Monogr., **8,** 1907, no. 35.
2. Woodworth, R. S. and Schlosberg, H., *Experimental Psychology*, Henry Holt & Co., New York, 1954.
3. Gilbert, L. C., Saccadic Movements as a Factor in Visual Perception in Reading. J. Educ. Psych., **50,** 1959, p. 15.
4. Erdman, B. and Dodge, R., *Psychologische Untersuchungen über das Lesen*, M. Niemeyer, Halle, 1898.
5. Glanville, A. D. and Dallenbach, K. M., The Range of Attention, Amer. J. Psych., **41,** 1929, p. 207.
6. Hunter, W. S. and Sigler, M., The Span of Visual Discrimination as a Function of Time and Intensity of Stimulation, J. Exp. Psych., **26,** 1940, p. 160.
7. Sperling, G., The Information Available in Brief Visual Presentations, Psych. Monogr., **74,** 1960, no. 11.
8. Nielsen, G., unpublished manuscript.

Synthesis of *N*-Port Active *RC* Networks

By I. W. SANDBERG

(Manuscript received August 16, 1960)

The following basic theorem concerning active RC networks is proved: Theorem: An arbitrary N × N matrix of real rational functions in the complex-frequency variable (a) can be realized as the short-circuit admittance matrix of a transformerless active RC N-port network containing N real-coefficient controlled sources, and (b) cannot, in general, be realized as the short-circuit admittance matrix of an active RC network containing less than N controlled sources.

I. INTRODUCTION

It is often desirable to avoid the use of magnetic elements in synthesis procedures, since resistors and capacitors are more nearly ideal elements and are usually cheaper, lighter and smaller. This is especially true in control systems in which, typically, exacting performance is required at very low frequencies. The rapid development of the transistor has provided the network synthesist with an efficient low-cost active element and has stimulated considerable interest in active *RC* network theory during the past decade.

Several techniques have been proposed for the active *RC* realization of transfer and driving-point functions.[1-18] It has been established that any real rational fraction can be realized as the transfer or driving-point function of a transformerless active *RC* network containing one active element. In particular, Linvill's technique[3] has been the basis for much of the later work.

Recently, Sipress[18] has shown that any two of the four short-circuit admittance parameters of a two-port network can be chosen arbitrarily and realized with a structure requiring only one active element. It follows that all four parameters can be realized with three active elements.

The problem of determining the minimum number of controlled sources required to realize all N^2 parameters of an arbitrary N-port immittance matrix is of considerable theoretical importance and has been of interest to network theorists for several years. The solution to this problem is stated in the abstract; its proof is the subject of this paper.

In Section II we derive some fundamental properties of N-port networks containing less than N controlled sources. The results are formulated in terms of inequalities involving the ranks of certain matrices. It follows from this study that at least N controlled sources are required for the realization of an arbitrary $N \times N$ immittance matrix. In Section III we make use of our previous results to establish an approach to the realization problem. This approach leads to a constructive proof that N controlled sources are in fact sufficient. A numerical example illustrating the essential points in the synthesis technique is presented in the Appendix.

II. N-PORT NETWORKS CONTAINING CONTROLLED SOURCES

A controlled source is ordinarily understood to be an ideal two-port network-representation of a single branch-branch constraint. The four types of elementary controlled sources are shown in Fig. 1. Note that the two "hybrid sources" [Fig. 1(a) and (b)] form a complete set, since they can be appropriately connected in cascade to realize each of the other two.

For our purposes it is convenient to generalize the definition of a controlled source to refer to any voltage or current source whose value is a weighted sum of certain prescribed voltages and currents. Specifically, if the value of a controlled voltage or current source is denoted by a_p,

$$a_p = \sum_{i=1}^{j+k} c_{pi} b_i , \tag{1}$$

where $b_1, b_2, \cdots, b_j$ are controlling currents and $b_{j+1}, b_{j+2}, \cdots, b_{j+k}$ are controlling voltages. It is assumed that the a_p, c_{pi} and b_i are Laplace-

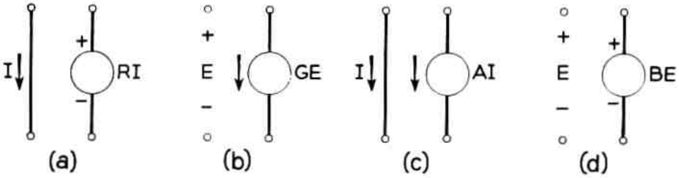


Fig. 1 — The four elementary controlled sources.

transformed quantities and that the c_{pi} are real rational functions of the complex frequency variable s.

2.1 *The Short-Circuit Admittance Matrix of an N-Port Network Containing Controlled Sources*

Consider the evaluation of the short-circuit admittance matrix of an N-port network containing a controlled source subnetwork as shown in Fig. 2. Denote by **E** and **I** respectively the column matrices of voltages and currents at the N accessible ports:

$$\mathbf{E} = \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_N \end{bmatrix}, \qquad \mathbf{I} = \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_N \end{bmatrix}. \tag{2}$$

Let **A** be the column matrix of all l controlled current sources and m controlled voltage sources, and let **B** be the column matrix of all j cur-

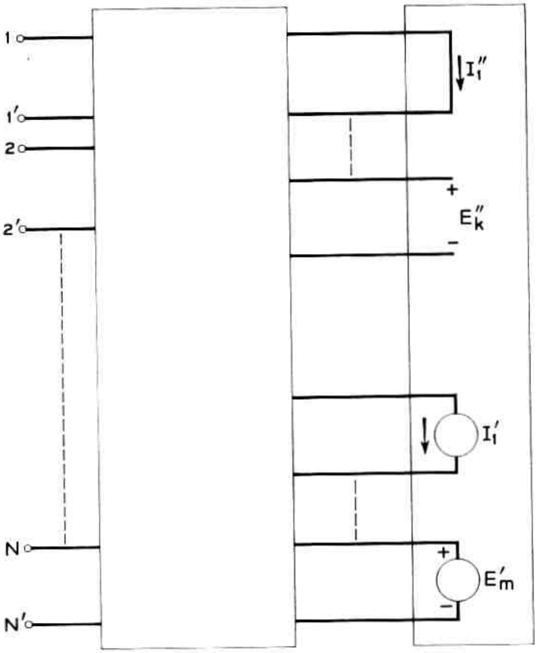


Fig. 2 — N-port network containing a controlled-source subnetwork.

rents and k voltages influencing the controlled sources:

$$\mathbf{A} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_l \\ a_{l+1} \\ \vdots \\ a_{l+m} \end{bmatrix} = \begin{bmatrix} I_1' \\ I_2' \\ \vdots \\ I_l' \\ E_1' \\ \vdots \\ E_m' \end{bmatrix}, \tag{3}$$

$$\mathbf{B} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_j \\ b_{j+1} \\ \vdots \\ b_{j+k} \end{bmatrix} = \begin{bmatrix} I_1'' \\ I_2'' \\ \vdots \\ I_j'' \\ E_1'' \\ \vdots \\ E_k'' \end{bmatrix}. \tag{4}$$

The relationship between **A** and **B** is assumed to be given by

$$\mathbf{A} = \mathbf{CB}, \tag{5}$$

where **C** is a $(l + m) \times (j + k)$ matrix of real rational functions in the complex frequency variable.

With **E** and **A** treated as independent variables, we apply the superposition theorem to obtain

$$\mathbf{I} = \mathbf{Y}_0\mathbf{E} + \mathbf{DA}, \tag{6}$$

where $\mathbf{Y}_0$ and **D** are defined by the equation. In particular, $\mathbf{Y}_0$ is the $N \times N$ short-circuit admittance matrix of the N-port network with the value of all controlled sources set equal to zero.

Similarly, we can express **B** as

$$\mathbf{B} = \mathbf{FE} + \mathbf{GA}, \tag{7}$$

where the matrices **F** and **G** are defined by the equation. From (5) and (7),

$$\mathbf{A} = [\mathbf{U} - \mathbf{CG}]^{-1}\mathbf{CFE}, \tag{8}$$

where $\mathbf{U}$ is the identity matrix of order $l + m$. Using (6),

$$[\mathbf{Y} - \mathbf{Y}_0] = \mathbf{D}[\mathbf{U} - \mathbf{CG}]^{-1}\mathbf{CF}, \tag{9}$$

where $\mathbf{Y}$ and $\mathbf{Y}_0$ are the short-circuit admittance matrices of the N-port network with all controlled sources respectively operative and set equal to zero. In certain degenerate cases, $\mathbf{Y}_0$ and/or the right-hand side of (9) will not exist. In such instances the network can be treated as a limiting case of a structure for which this difficulty does not occur.

2.2 *The Rank of* $[\mathbf{Y} - \mathbf{Y}_0]$

Consider the maximum rank of the $N \times N$ matrix $[\mathbf{Y} - \mathbf{Y}_0]$. Since the rank of a matrix product cannot exceed the rank of any of its constituent factors,[19]

$$\text{rank } [\mathbf{Y} - \mathbf{Y}_0] \leqq \text{rank } [\mathbf{C}] = R_c . \tag{10}$$

The elements of $[\mathbf{Y} - \mathbf{Y}_0]$ are real rational functions in the complex frequency variable. Assuming that this matrix has finite poles at $s = s_1, s_2, \cdots, s_m$ of multiplicity $n_1, n_2, \cdots, n_m$ respectively, it can be expressed as

$$[\mathbf{Y} - \mathbf{Y}_0] = \sum_{k=0}^{p} \mathbf{A}_k s^k + \sum_{l=1}^{m} \sum_{k=1}^{n_l} \mathbf{B}_{-k}{}^{(l)} \frac{1}{(s - s_l)^k}, \tag{11}$$

where the $\mathbf{A}_k$ and $\mathbf{B}_{-k}^{(l)}$ are coefficient matrices and in particular, the $\mathbf{B}_{-1}^{(l)}$ are residue matrices.

From (11), the matrix of coefficients of the first term in the Laurent expansion at the pole $s = s_l$ is

$$\mathbf{B}\,_{-n_l}^{(l)} = (s - s_l)^{n_l}[\mathbf{Y} - \mathbf{Y}_0] \mid_{s=s_l} . \tag{12}$$

In view of (10), we have

$$\text{rank } [\mathbf{B}\,_{-n_l}^{(l)}] \leqq R_c . \tag{13}$$

Similarly, the leading coefficient of the matric polynomial in (11) is given by

$$\mathbf{A}_p = \lim_{s\to\infty} \frac{1}{s^p} [\mathbf{Y} - \mathbf{Y}_0], \tag{14}$$

and hence

$$\text{rank } [\mathbf{A}_p] \leqq R_c . \tag{15}$$

Consequently, when $R_c < N$, all k-rowed minors of the matrices $\mathbf{B}_{-n_l}^{(l)}$ and $\mathbf{A}_p$ vanish, where $k = R_c + 1, R_c + 2, \cdots, N$.

Inequalities (13) and (15) shed considerable light on the fundamental properties of an N-port network containing a controlled source subnetwork. In fact, at poles of $\mathbf{Y}$ which are not poles of $\mathbf{Y}_0$, these conditions yield explicit restrictions on the $\mathbf{Y}$ matrix. For example, let $\mathbf{Y}$ be the admittance matrix of an active RC network and take $\mathbf{Y}_0$ to be the corresponding passive RC matrix obtained from $\mathbf{Y}$ by setting all controlled source coefficients equal to zero. It is well known that $\mathbf{Y}_0$ must be regular everywhere in the complex plane except at infinity and at points on the negative-real axis where only simple poles may occur. Hence $\mathbf{Y}_0$ cannot influence the coefficient matrices $\mathbf{B}_{-n_l}^{(l)}$ at any multiple-order pole or at any pole not on the negative-real axis. In particular, the rank of the residue matrix at any simple complex pole cannot exceed R_c, the rank of the matrix $\mathbf{C}$.

The rank of $\mathbf{C}$, of course, cannot exceed the number of its rows or columns, whichever is smaller. That is,

$$R_c \leqq \min\,[j + k, l + m]. \tag{16}$$

This means that R_c cannot exceed the number of controlled sources or the total number of controlling voltages and currents, whichever is smaller. Consequently, if any of the prescribed $\mathbf{B}_{-n_l}^{(l)}$ are to have full rank at a pole of $\mathbf{Y}$ which is not a pole of $\mathbf{Y}_0$, the controlled source subnetwork must include at least N controlled sources and at least N distinct control ports.†

A similar development, of course, can be carried out in terms of the open-circuit impedance matrices $\mathbf{Z}$ and $\mathbf{Z}_0$. Note that these results are valid for controlled source coefficients c_{pi} which may be any set of real rational functions in the complex-frequency variable. Note also that a driving-point immittance can be regarded as a controlled source, since such immittances impose a constraint which is merely a special case of (1).

III. N-PORT ACTIVE RC REALIZATION

We begin the study of the N-port realization problem by considering an active RC network containing one controlled source. Specifically, consider an $(N + 2)$-port passive RC network characterized by the $(N + 2) \times (N + 2)$ short-circuit admittance matrix $\tilde{\mathbf{Y}}$, and suppose

† The realization of an arbitrary $N \times N$ matrix of constants as the short-circuit admittance matrix of an N-port network containing positive resistors, ideal transformers and controlled sources also requires, in general, at least N controlled sources. This follows from the fact that, in this case, $\mathbf{Y}_0$ is the matrix of a nonnegative quadratic form, and hence it is possible to prescribe constant matrices $\mathbf{Y}$ such that, for the entire class of matrices $\mathbf{Y}_0$, $[\mathbf{Y} - \mathbf{Y}_0]$ is of rank N.

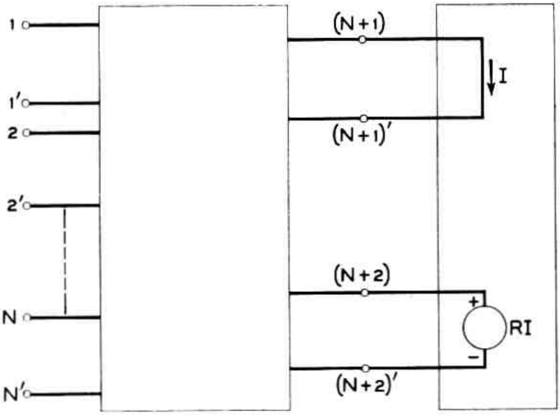


Fig. 3 — Active RC network containing one controlled source—canonical subnetwork.

that a current-controlled voltage source is connected between ports $N + 1$ and $N + 2$ as shown in Fig. 3. Denote by $\mathbf{Y}_0$ and $\mathbf{Y}$ the $N \times N$ short-circuit admittance matrices relating the column vectors

$$\mathbf{E} = \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_N \end{bmatrix} \qquad \text{and} \qquad \mathbf{I} = \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_N \end{bmatrix} \tag{17}$$

when the controlled source coefficient respectively vanishes and is equal to R.

The matrix $\mathbf{Y}$ is given, as a special case of (9), by

$$\mathbf{Y} - \mathbf{Y}_0 = -\frac{R}{1 + R\tilde{y}_{N+1,N+2}} \begin{bmatrix} \tilde{y}_{1,N+2} \\ \vdots \\ \tilde{y}_{N,N+2} \end{bmatrix} [\tilde{y}_{N+1,1} \cdots \tilde{y}_{N+1,N}], \tag{18}$$

where $\mathbf{Y}_0$ is the matrix of elements in the first N rows and columns of $\tilde{\mathbf{Y}}$. It is convenient to express (18) as

$$\mathbf{Y} - \mathbf{Y}_0 = -\frac{R}{q(q + R\tilde{p}_{N+1,N+2})} \begin{bmatrix} \tilde{p}_{1,N+2} \\ \vdots \\ \tilde{p}_{N,N+2} \end{bmatrix} [\tilde{p}_{N+1,1} \cdots \tilde{p}_{N+1,N}], \tag{19}$$

where q and the $\tilde{p}_{jk}$ are polynomials, and

$$\tilde{y}_{jk} = \frac{\tilde{p}_{jk}}{q}. \tag{20}$$

It is evident from (18) or (19) that, as anticipated, $[\mathbf{Y} - \mathbf{Y}_0]$ has unit rank.

3.1 *N-Port Synthesis*

Our objective is to obtain an expression involving $\mathbf{Y}$ similar to (9) with a right-hand side of rank N. We know that a network characterized by such a relationship will require at least N controlled sources.

It is well known that a rank N matrix can be expressed as a sum of N rank 1 matrices.[19] This suggests that the realization of $\mathbf{Y}$ can be accomplished with N networks connected in parallel. We shall specifically consider the parallel connection of N networks of the type shown in Fig. 3.

Assuming that the scalar coefficient on the right-hand side of (19) is the same function of s for each of the N subnetworks, we obtain†

$$\mathbf{Y} - \sum_{i=1}^{N} \mathbf{Y}_{0i} = -\frac{R}{q(q + R\tilde{p}_{N+1,N+2})} \sum_{i=1}^{N} \begin{bmatrix} \tilde{p}^{(i)}_{1,N+2} \\ \vdots \\ \tilde{p}^{(i)}_{N,N+2} \end{bmatrix} [\tilde{p}^{(i)}_{N+1,1} \cdots \tilde{p}^{(i)}_{N+1,N}], \tag{21}$$

where

$$\tilde{\mathbf{Y}}^{(i)} = \frac{1}{q}\,[\tilde{p}^{(i)}_{jk}] \tag{22}$$

and

$$\tilde{p}^{(i)}_{N+1,N+2} = \tilde{p}_{N+1,N+2}\,, \qquad i = 1,2,\cdots,N. \tag{23}$$

The sum of matrix products in (21) can be written as a single matrix with the element in the jth row and kth column given by

$$\sum_{i=1}^{N} \tilde{p}^{(i)}_{j,N+2}\tilde{p}^{(i)}_{N+1,k}\,. \tag{24}$$

This matrix can therefore be written as the product of the following two matrices:

† The networks are assumed to be such that admittance matrices add without the use of ideal transformers. This is justified later by employing balanced structures.

$$\mathbf{P}_1\mathbf{P}_2 = \begin{bmatrix} \tilde{p}_{1,N+2}^{(1)} \cdots \tilde{p}_{1,N+2}^{(N)} \\ \vdots \quad\quad \vdots \\ \tilde{p}_{N,N+2}^{(1)} \cdots \tilde{p}_{N,N+2}^{(N)} \end{bmatrix} \begin{bmatrix} \tilde{p}_{N+1,1}^{(1)} \cdots \tilde{p}_{N+1,N}^{(1)} \\ \vdots \quad\quad \vdots \\ \tilde{p}_{N+1,1}^{(N)} \cdots \tilde{p}_{N+1,N}^{(N)} \end{bmatrix}. \tag{25}$$

From (21) and (25),

$$\mathbf{Y} - \mathbf{Y}_{0T} = -\frac{R}{q(q + R\tilde{p}_{N+1,N+2})} \mathbf{P}_1\mathbf{P}_2, \tag{26}$$

where

$$\mathbf{Y}_{0T} = \sum_{i=1}^{N} \mathbf{Y}_{0i} .$$

Let the prescribed short-circuit admittance matrix $\mathbf{Y}$ be given as

$$\mathbf{Y} = \frac{1}{D} [N_{ij}], \tag{27}$$

where D is the common denominator polynomial of the elements in $\mathbf{Y}$ and $[N_{ij}]$ is a matrix of polynomials. Similarly, write $\mathbf{Y}_{0T}$ as

$$\mathbf{Y}_{0T} = \frac{1}{q} [p_{ij}] = \frac{1}{q} \sum_{k=1}^{N} [p_{ij}^{(k)}]. \tag{28}$$

From (26), (27) and (28),

$$\frac{1}{qD} [Dp_{ij} - qN_{ij}] = \frac{R}{q(q + R\tilde{p}_{N+1,N+2})} \mathbf{P}_1\mathbf{P}_2 . \tag{29}$$

In (29) let terms be identified as follows:

$$\tilde{p}_{N+1,N+2} = \frac{1}{R} (D - q), \tag{30}$$

$$\mathbf{P}_1\mathbf{P}_2 = \frac{1}{R} [Dp_{ij} - qN_{ij}]. \tag{31}$$

At this point we have reduced the synthesis of the N-port admittance matrix $\mathbf{Y}$ to the determination of N realizable $(N + 2) \times (N + 2)$ RC network matrices $\tilde{\mathbf{Y}}^{(i)}$ whose elements satisfy (30) and (31).

3.2 *Sufficient Conditions for the Realization of* $\mathbf{Y}$

The matrices $\tilde{\mathbf{Y}}^{(i)}$ can be expressed as

$$\tilde{\mathbf{Y}}^{(i)} = s\mathbf{K}_\infty{}^{(i)} + \mathbf{K}_0{}^{(i)} + \sum_{j=1}^{\deg q} \mathbf{K}_j{}^{(i)} \frac{s}{s + \sigma_j}, \tag{32}$$

where "deg q" means the degree of the polynomial q, and where the σ_j are real and satisfy

$$0 < \sigma_1 < \sigma_2 \cdots < \sigma_{\deg q}.$$

If the coefficient matrices $\mathbf{K}_\infty$, $\mathbf{K}_0$ and $\mathbf{K}_j$ are "dominant-diagonal" matrices,† (32) can be realized as a transformerless balanced $RC(N + 2)$-port network.[20]

Assume that $\mathbf{Y}_{0T}$ has been chosen so that

(a) its coefficient matrices satisfy the dominant-diagonal condition with the inequality sign;†

(b) the matrix $(1/R)[Dp_{ij} - qN_{ij}]$ can be expressed as the product of two polynomial matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ with the property that $(1/q)\mathbf{P}_1$ and $(1/q)\mathbf{P}_2$ are matrices of realizable RC transfer admittances (these admittances are assumed to have poles at infinity only when $\mathbf{Y}_{0T}$ has a pole at infinity); and

(c) the function $\tilde{p}_{N+1,N+2}$ satisfies the realizability and regularity constraints stated in (b).

If (a) is satisfied, we can write $\mathbf{Y}_{0T}$ as the sum of N matrices $\mathbf{Y}_{0i}$, each of which has coefficient matrices that satisfy the dominant-diagonal condition with the inequality sign. Recall that $\mathbf{Y}_{0i}$ is the matrix of elements in the first N rows and columns of the $(N + 2) \times (N + 2)$ matrix $\tilde{\mathbf{Y}}^{(i)}$. To obtain $\tilde{\mathbf{Y}}^{(i)}$, we border $\mathbf{Y}_{0i}$ with two additional rows and columns of elements. All but three of the required numerator polynomials are determined by the entries in the polynomial matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ which satisfy (b). Of the remaining three polynomials, $\tilde{p}_{N+1,N+2}$ is given by (30) and is assumed to satisfy (c), while $\tilde{p}_{N+1,N+1}^{(i)}$ and $\tilde{p}_{N+2,N+2}^{(i)}$ may be chosen freely to assist realizability, since they are unrestricted by (30) or (31).

The realizability of $\tilde{\mathbf{Y}}^{(i)}$ can be ensured by having it exhibit the dominance characteristic, and this can always be done by choosing the scale factors of the polynomials $\tilde{p}_{N+1,N+1}^{(i)}$ and $\tilde{p}_{N+2,N+2}^{(i)}$ as well as the value of R, the controlled source coefficient, to be sufficiently large.

Hence (a), (b) and (c) are sufficient for the realization of $\mathbf{Y}$. To make further progress, we next establish conditions that permit $\mathbf{P} = (1/R)[Dp_{ij} - qN_{ij}]$ to be written as the product of two matrices with polynomial elements of lower degree.

† A dominant-diagonal matrix $\mathbf{M}$ has elements m_{jk} which satisfy

$$m_{jj} \geqq \sum_{k \neq j} | m_{jk} |.$$

3.3 *Factorization of the Matric Polynomial* **P**

Let L be the degree of the highest degree polynomial in $\mathbf{P} = (1/R)\cdot[Dp_{ij} - qN_{ij}]$, and suppose that the zeros of

$$\det \mathbf{P} = \sum_{k=0}^{NL} a_k s^k$$

include K distinct real zeros at $s = s_i, \quad (i = 1, 2, \cdots, N, \cdots, K)$.

Consider the result of determining a nonsingular matrix **Q** with real constant elements such that every element in the ith column of **PQ** has a zero at $s = s_i, \quad (i = 1, 2, \cdots, N)$. If indeed this can be done, **P** can be written as

$$\mathbf{P} = (\mathbf{PQ})\mathbf{Q}^{-1} = \mathbf{P}'(\mathbf{DQ}^{-1}), \tag{33}$$

where **D** is the diagonal matrix diag $[s - s_1, s - s_2, \cdots, s - s_N]$, and the degree of the highest degree polynomial in $\mathbf{P}'$ is $L - 1$. This is equivalent to removing a linear factor of the matric polynomial **P**:

$$\begin{aligned} \mathbf{P} = \sum_{j=1}^{L} s^j \mathbf{A}_j &= \left[\sum_{j=1}^{L-1} s^j \mathbf{A}_j'\right] \mathbf{DQ}^{-1} \\ &= \left[\sum_{j=1}^{L-1} s^j \mathbf{A}_j' \mathbf{Q}^{-1}\right] \mathbf{QDQ}^{-1} \\ &= \left[\sum_{j=1}^{L-1} s^j \mathbf{A}_j''\right] (s\mathbf{U} - \mathbf{B}), \end{aligned} \tag{34}$$

where **U** is the identity matrix of order N and

$$\mathbf{B} = \mathbf{Q} \text{ diag } [s_1, s_2, \cdots, s_N]\mathbf{Q}^{-1}. \tag{35}$$

If $(N - 1)L < K$, a matrix **Q** having the required properties exists and can be constructed as follows. First, note that at any zero of det **P**, say at $s = s_l$, the column rank of **P** is necessarily less than N, and hence there exists a relationship of the form

$$0 = \sum_{j=1}^{N} \alpha_{jl}[\mathbf{P}_j(s_l)], \tag{36}$$

where $[\mathbf{P}_j(s_l)]$ is the jth column vector of **P** evaluated at $s = s_l$, and the α_{jl} are not all zero. Note also that at no more than $(N - 1)L$ of the zeros of det **P** is it possible to determine alphas, not all zero, which satisfy

$$0 = \sum_{j \neq k}^{N} \alpha_{jl}[\mathbf{P}_j(s_l)], \tag{37}$$

where k is any one of the integers $[1,2,\cdots,N]$. This follows at once from the fact that all nonidentically vanishing determinants, formed from det $\mathbf{P}$ by replacing the kth column of det $\mathbf{P}$ with a column of constants, vanish at most at $(N - 1)L$ points. Therefore, if $(N - 1)L < K$, there must exist at least one equation of the type (36) for a real zero and with $\alpha_{kl} \neq 0$. In other words, there exists a nonsingular matrix of real elements

$$\mathbf{Q}_k = \begin{bmatrix} 1 & & & q_{1k} & & & \\ & \ddots & & & & & \\ & & 1 & \vdots & & & \\ & & & q_{kk} & & & \\ & & & \vdots & 1 & & \\ & & & & & \ddots & \\ & & & q_{Nk} & & & 1 \end{bmatrix} \tag{38}$$

such that every element in the kth column of $\mathbf{PQ}_k$ has a real zero at $s = s_k$. Note that the elements in all columns except the kth remain unchanged. Hence the matrix $\mathbf{Q}$ can be constructed as a product of N matrices $\mathbf{Q}_j$ chosen so that every element in the ith column of

$$\mathbf{P}\prod_{j=1}^{m}\mathbf{Q}_j\,, \qquad i = 1,2,\cdots, m.$$

has a real zero at $s = s_i$.

To summarize, if $(N - 1)L < K$, N distinct real zeros of det $\mathbf{P}$ can be removed as a linear factor of the matric polynomial $\mathbf{P}$. The remaining polynomial is of degree $L - 1$ and all coefficient matrices are real.†

To simplify the discussion, we have not considered certain extensions of the factorization technique. It is possible, for example, to carry out a similar development with respect to the rows of $\mathbf{P}$. This permits the removal of a linear factor that premultiplies the remaining matric polynomial.

3.4 *Consideration of Conditions (a), (b) and (c)*

The admittance matrix $\mathbf{Y}_{0T}$ can be made to have dominant-diagonal coefficient matrices by choosing any $N \times N$ realizable RC admittance

† This implies that the matric polynomial $\mathbf{P}$ can be written as

$$\mathbf{P} = \mathbf{C}\prod_{i=1}^{L}(s\mathbf{U} - \mathbf{B}_i)$$

when det $\mathbf{P}$ has NL distinct zeros. When these zeros are all real the coefficient matrices $\mathbf{C}$ and $\mathbf{B}_i$ are also real.

matrix, with elements of suitable degree as determined subsequently, and multiplying each diagonal entry by a sufficiently large positive real constant ρ. Hence condition (a) is easily satisfied. Denote the matrix determined in this way by

$$\mathbf{Y}_{0T} = \frac{1}{q}\begin{bmatrix} \rho p'_{11} & p_{12} & \cdots & p_{1N} \\ \vdots & \rho p'_{22} & & \\ & & \ddots & \\ p_{N1} & & & \rho p'_{NN} \end{bmatrix}. \tag{39}$$

The polynomial det $\mathbf{P}$ can be written as

$$\det \mathbf{P} = \det \frac{1}{R}\,[Dp_{ij} - qN_{ij}] = \left(\frac{\rho}{R}\right)^N \left\{D^N \prod_{i=1}^{N} p'_{ii} + \frac{R(s)}{\rho^N}\right\}, \tag{40}$$

where $R(s)/\rho^N$ is a polynomial with degree not exceeding NL and with all coefficients that approach zero as ρ approaches infinity. We shall assume that the degree of p_{ii}, deg p_{ii}, has been chosen to be independent of the index i. Note that, as ρ approaches infinity, N deg p_{ii} zeros of det $\mathbf{P}$ approach the zeros of

$$\prod_{i=1}^{N} p'_{ii}.$$

The zeros of this product can be chosen to be distinct and different from those of D. Hence, for a sufficiently large value of ρ, (a) is satisfied and det $\mathbf{P}$ has at least N deg p_{ii} distinct real zeros.

Next, consider condition (b). The degree of the highest degree polynomial in $\mathbf{P}$ is given by

$$\begin{aligned} L &= \max\,[\max \deg p_{ij} + \deg D, \max \deg N_{ij} + \deg q] \\ &= \max\,[\deg p_{ii} + \deg D, \max \deg N_{ij} + \deg q]. \end{aligned} \tag{41}$$

Hence,

$$\begin{aligned} L &= \deg p_{ii} + \max\,[\max \deg N_{ij} - \epsilon, \deg D] \\ &= \deg p_{ii} + L_\epsilon , \end{aligned} \tag{42}$$

where

$$\begin{aligned} \epsilon &= 0, \qquad \deg p_{ii} = \deg q \\ \epsilon &= 1, \qquad \deg p_{ii} = \deg q + 1. \end{aligned} \tag{43}$$

To remove k linear factors of the matric polynomial $\mathbf{P}$ as described in Section 3.3, it is sufficient, after removal of the $(k - 1)$th factor, that

$$(N - 1)[\deg p_{ii} + L_\epsilon - (k - 1)] < N \deg p_{ii} - N(k - 1). \quad (44)$$

If $k = L_\epsilon$ factors are removed, $\mathbf{P}$ could be written as the product of two matrices, one of degree L_ϵ and the other of degree $\deg p_{ii}$. Substituting this value of k into (44) gives the required relationship between L_ϵ and $\deg p_{ii}$:

$$NL_\epsilon - 1 < \deg p_{ii}. \quad (45)$$

Hence conditions (a) and (b) are satisfied† with $\deg p_{ii} = NL_\epsilon$. Finally, it is evident that condition (c) can be satisfied simultaneously, since $\tilde{p}_{N+1,N+2}$ can be chosen to have any degree not exceeding $\deg p_{ii}$.

This proves the theorem stated in the abstract.

IV. CONCLUSION

We have proven that N is the sufficient and, in general, minimum number of controlled sources required to realize an arbitrary $N \times N$ matrix of real rational functions as a transformerless active *RC* N-port network. A canonical structure is a parallel combination of N networks, each containing a single controlled source. The type of controlled source employed is one of the two basic elementary controlled sources. Similar developments can be carried out for other types of controlled sources.

Further work is indicated in several directions. It is desirable to avoid the use of balanced networks. A detailed investigation of matric polynomial factorization may shed some light on this possibility. A major difficulty stems from the fact that relatively little is known about the realization of transformerless passive *RC* networks. Even so, it is almost certain that more practical canonical structures will be discovered.

It is noteworthy that the analytical machinery employed here provides insight into other fundamental questions. For example, it is easy to show that all N resistors in Oono's passive N-port realization[21] are in fact necessary. Similarly all N-negative and N-positive resistors in Carlin's active N-port realization[22] are necessary.

V. ACKNOWLEDGMENT

The author wishes to acknowledge the helpful and stimulating discussions held with S. Darlington and B. McMillan. The technique of

† This is a pessimistic statement of the required degree of p_{ii}, and results from the particular technique employed to factor $\mathbf{P}$. A more detailed study of matric polynomial factorization, as yet incomplete, indicates that the degree of p_{ii} can generally be reduced by a factor of N.

factoring a matric polynomial presented in Section 3.3 is based on a suggestion by S. Darlington.

APPENDIX

Synthesis of a Two-Port Network — A Numerical Example

To illustrate the main points in the synthesis technique presented in Section III, we consider in detail the synthesis of a two-port network. This example demonstrates also that (45) is not a necessary condition.

Let the prescribed 2 × 2 matrix be

$$\begin{aligned}\mathbf{Y} &= \frac{1}{D}[N_{ij}] \\ &= \frac{1}{s^2+s+1}\begin{bmatrix} s^2+s+2 & s^2+s+3 \\ s^2+s+4 & s^2+s+5 \end{bmatrix}.\end{aligned} \tag{46}$$

We choose $\mathbf{Y}_{0T}$ as the following matrix that obviously satisfies the dominance condition with the inequality sign:

$$\begin{aligned}\mathbf{Y}_{0T} &= \frac{1}{q}[p_{ij}] \\ &= \frac{1}{(s+2)(s+4)}\begin{bmatrix} 5(s+1)(s+3) & 0 \\ 0 & 5(s+1)(s+3) \end{bmatrix}.\end{aligned} \tag{47}$$

From (30), (31), (46) and (47),

$$\tilde{p}_{3,4} = -\frac{1}{R}(5s+7), \tag{48}$$

$$\begin{aligned}\mathbf{P}_1\mathbf{P}_2 &= \mathbf{P} = \frac{1}{R} \\ &\cdot\begin{bmatrix} 4s^4+18s^3+24s^2+15s-1 & -s^4-7s^3-17s^2-26s-24 \\ -s^4-7s^3-18s^2-32s-32 & 4s^4+18s^3+21s^2-3s-25 \end{bmatrix}.\end{aligned} \tag{49}$$

Consider the factorization of $\mathbf{P}$ into two matric polynomials of the second degree. The factors of

$$\begin{aligned}R^N \det \mathbf{P} = 15s^8 + 130s^7 + 420s^6 + 555s^5 - 152s^4 \\ - 1629s^3 - 2474s^2 - 1972s - 743,\end{aligned}$$

determined with a digital computer, are

$$\begin{aligned}&(s+1.0707018)(s-1.6223931)(s+3.0014915)(s+2.6871002) \\ &\cdot(s+1.3191886 \pm j1.2215876)\ (s+0.4456939 \pm j0.9460882).\end{aligned} \tag{50}$$

Denote by s_1, s_2, s_3 and s_4 respectively the zeros of the first four factors in (50).

First, we determine a matrix $\mathbf{Q}_1$ such that both elements in the first column of $\mathbf{PQ}_1$ have a zero at $s = s_1$. At $s = s_1$,

$$\alpha_{11}[\mathbf{P}_1(s_1)] + \alpha_{21}[\mathbf{P}_2(s_1)] = 0. \tag{51}$$

By evaluating the pair of polynomials in either row of (49) at $s = s_1$ we obtain:

$$0.76249\ \alpha_{11} + \alpha_{21} = 0.$$

Hence,

$$\mathbf{Q}_1 = \begin{bmatrix} 1 & 0 \\ -0.76249 & 1 \end{bmatrix}. \tag{52}$$

From (49) and (52),

$$\mathbf{PQ}_1 = \frac{1}{R}\,[a_{ij}], \tag{53}$$

where

$$a_{11} = (4.7625s^3 + 18.2382s^2 + 17.4347s + 16.1575)(s + 1.0707),$$

$$a_{21} = -(4.0499s^3 + 16.3886s^2 + 16.4651s + 12.0835)(s + 1.0707),$$

$$a_{12} = -(\mathrm{s}^4 + 7s^3 + 17s^2 + 26s + 24),$$

$$a_{22} = (4s^4 + 18s^3 + 21s^2 - 3s - 25).$$

Next we find a matrix $\mathbf{Q}_2$ such that the first column of $\mathbf{PQ}_1\mathbf{Q}_2$ is identical to that of $\mathbf{PQ}_1$, and both elements in the second column of $\mathbf{PQ}_1\mathbf{Q}_2$ have a zero at $s = s_2$. The evaluation of polynomials as before leads to

$$\mathbf{Q}_2 = \begin{bmatrix} 1 & 0.48643 \\ 0 & 1 \end{bmatrix}. \tag{54}$$

At this point, $\mathbf{P}$ can be expressed as

$$\mathbf{P} = \frac{1}{R}\,[b_{ij}]\ \mathrm{diag}\ [s + 1.0707, s - 1.6223]\mathbf{Q}^{-1}, \tag{55}$$

where $\mathbf{Q}^{-1} = (\mathbf{Q}_1\mathbf{Q}_2)^{-1}$, and

$$b_{11} = 4.7625s^3 + 18.2382s^2 + 17.4347s + 16.1575,$$

$$b_{21} = -(4.0499s^3 + 16.3886s^2 + 16.4651s + 12.0835),$$

$$b_{12} = 1.3166s^3 + 6.4881s^2 + 11.5058s + 9.6064,$$

$$b_{22} = 2.0300s^3 + 11.2124s^2 + 22.6465s + 19.2894.$$

A second linear factor of **P** can be removed by repeating this process. Specifically, if the zeros at $s = s_3$ and $s = s_4$ are removed respectively from the first and second columns of $[b_{ij}]$, **P** can be expressed as

$$\mathbf{P} = \mathbf{P}_1\mathbf{P}_2$$

with

$$\mathbf{P}_1 = \beta \begin{bmatrix} 4.3560s^2+3.1605s+4.3961 & 1.3166s^2+2.9503s+3.5781 \\ -(4.6766s^2+5.8136s+6.0077) & 2.0300s^2+5.7576s+7.1753 \end{bmatrix},$$

$$\mathbf{P}_2 = \frac{1}{\beta R}\begin{bmatrix} 0.6292s^2+2.5622s+2.0221 & -0.4863s^2-1.9803s-1.5629 \\ 0.9568s^2+1.5419s-2.7648 & 0.8499s^2+0.5007s-4.7910 \end{bmatrix}, \tag{56}$$

where β is an arbitrary nonzero real parameter.

To determine $\tilde{\mathbf{Y}}^{(1)}$ and $\tilde{\mathbf{Y}}^{(2)}$, first write $\mathbf{Y}_{0T}$ as the sum of two matrices, $\mathbf{Y}_{01}$ and $\mathbf{Y}_{02}$, that satisfy the dominance condition with the inequality sign. The following choice is clearly acceptable:

$$\mathbf{Y}_{01} = \mathbf{Y}_{02} = \tfrac{1}{2}\mathbf{Y}_{0T}\ \cdot$$

Hence, $\tilde{\mathbf{Y}}^{(1)}$ and $\tilde{\mathbf{Y}}^{(2)}$ are given by

$$\tilde{\mathbf{Y}}^{(i)} = \frac{1}{q}\begin{bmatrix} \tilde{p}_{11}^{(i)} & 0 & \tilde{p}_{13}^{(i)} & \tilde{p}_{14}^{(i)} \\ 0 & \tilde{p}_{22}^{(i)} & \tilde{p}_{23}^{(i)} & \tilde{p}_{24}^{(i)} \\ \tilde{p}_{31}^{(i)} & \tilde{p}_{32}^{(i)} & \tilde{p}_{33}^{(i)} & \tilde{p}_{34}^{(i)} \\ \tilde{p}_{41}^{(i)} & \tilde{p}_{42}^{(i)} & \tilde{p}_{43}^{(i)} & \tilde{p}_{44}^{(i)} \end{bmatrix}, \tag{57}$$

where

$$\tilde{p}_{34}^{(1)} = \tilde{p}_{34}^{(2)} = -\frac{1}{R}(5s+7),$$

$$\tilde{p}_{11}^{(1)} = \tilde{p}_{11}^{(2)} = \tilde{p}_{22}^{(1)} = \tilde{p}_{22}^{(2)} = \tfrac{5}{2}(s+1)(s+3).$$

The polynomials $\tilde{p}_{33}^{(1)}$, $\tilde{p}_{33}^{(2)}$, $\tilde{p}_{44}^{(1)}$, and $\tilde{p}_{44}^{(2)}$ are unrestricted by (31) and hence, for simplicity, can be chosen to be $\frac{5}{2}(s+1)(s+3)$. The remaining polynomials are obtained from (25) with $\mathbf{P}_1$ and $\mathbf{P}_2$ given explicitly in (56). It is evident that finite, nonzero parameters β and R can be determined so that the matrices $\tilde{\mathbf{Y}}^{(1)}$ and $\tilde{\mathbf{Y}}^{(2)}$ satisfy the dominance condition. The realization of each of these matrices takes the form shown

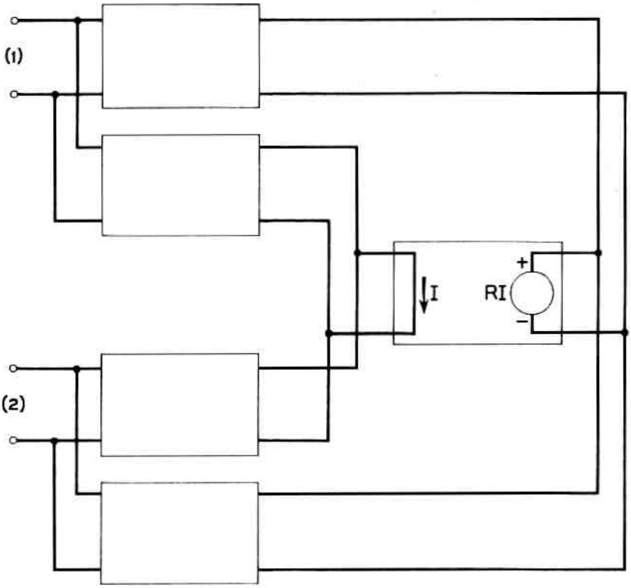


Fig. 4 — Realization of $\tilde{Y}^{(1)}$ or $\tilde{Y}^{(2)}$ for two-port network example.

in Fig. 4, where the rectangles enclose transformerless passive balanced *RC* structures.[20]

REFERENCES

1. Dietzold, R. L., Frequency Discriminative Electric Transducer, U. S. Patent 2,549,965, April 17, 1951.
2. Bangert, J. T., The Transistor as a Network Element, B.S.T.J., **33,** 1954, p. 329.
3. Linvill, J. G., RC Active Filters, Proc. I.R.E., **42,** 1954, p. 555.
4. Armstrong, D. B. and Reza, F. M., Synthesis of Transfer Functions by Active RC Networks, I.R.E. Trans., **CT-1,** 1954, p. 8.
5. Sallen, R. P. and Key, E. L., A Practical Method of Designing RC Active Filters, I.R.E. Trans., **CT-2,** 1955, p. 74.
6. Horowitz, I. M., RC-Transistor Network Synthesis, Proc. Nat. Elect. Conf. October 1956, p. 818.
7. Horowitz, I. M., Synthesis of Active RC Transfer Functions, Research Report R-507-56 PIB-437, Microwave Research Inst., Polytechnic Inst. of Brooklyn, 1956.
8. Sipress, J. M., Active RC Partitioning Synthesis of High-Q Bandpass Filter, thesis, Polytechnic Inst. of Brooklyn, 1957.
9. Bongiorno, J. J., Synthesis of Active RC Single-Tuned Bandpass Filters, I.R.E. Nat. Conv. Rec., March 1958, Pt. 2, p. 30.
10. Sandberg, I. W., Active RC Networks, Research Report R-662-58 PIB-590, Microwave Research Inst., Polytechnic Inst. of Brooklyn, 1958.
11. DeClaris, N., Synthesis of Active Networks — Driving-Point Functions, I.R.E. Nat. Conv. Rec., March 1959, Pt. 2, p. 23.
12. Myers, B. R., Transistor-RC Network Synthesis, I.R.E. Wescon Conv. Rec., August 1959, Pt. 2, p. 65.

13. Kinariwala, B. K., Synthesis of Active RC Networks, B.S.T.J., **38,** 1959, p. 1269.
14. Horowitz, I. M., Optimization of Negative Impedance Converter Synthesis Techniques, I.R.E. Trans., **CT-6,** 1959, p. 296.
15. Blecher, F. H., Application of Synthesis Techniques to Electronic Circuit Design, I.R.E. Nat. Conv. Rec., March 1960, Pt. 2, p. 210.
16. Kuh, E. S., Transfer Function Synthesis of Active RC Networks, I.R.E. Nat. Conv. Rec., March 1960, Pt. 2, p. 134.
17. Sandberg, I. W., Synthesis of Driving-Point Impedances with Active RC Networks, B.S.T.J., **39,** July 1960, p. 947.
18. Sipress, J. M., Synthesis of Active RC Networks, to be published.
19. Halmos, P. R., *Finite-Dimensional Vector Spaces,* D. Van Nostrand Co., New York, 1942, pp. 92, 93.
20. Slepian, P. and Weinberg, L., Synthesis Applications of Paramount and Dominant Matrices, Proc. Nat. Elect. Conf., October 1958, p. 611.
21. Oono, Y., Synthesis of a Finite 2-N Terminal Network by a Group of Networks Each of Which Contains Only One Ohmic Resistance, J. Math. Phys., **29,** 1950, p. 13.
22. Carlin, H. J., General N-Port Synthesis with Negative Resistors, Proc. I.R.E. **48,** 1960, p. 1174.

Contributors to this Issue

EMANUEL AVERBACH, B.A., 1951, University of Pennsylvania; M.A., 1953, Swarthmore College; Ph.D., 1956, Johns Hopkins University; Bell Telephone Laboratories, 1956—. His research has been in the field of perception of picture sharpness and, more recently, in the way humans extract information from their visual environment. He is also engaged in military systems studies. Member A.A.A.S., American Psychological Association, Optical Society of America, Sigma Xi.

VÁCLAV E. BENEŠ, A.B., 1950, Harvard College; M.A., Ph.D., 1953, Princeton University; Bell Telephone Laboratories, 1953—. Mr. Beneš has been engaged in mathematical research on stochastic processes, traffic theory and servomechanisms. In 1959–60 he was visiting lecturer in mathematics at Dartmouth College. Member American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, Society for Industrial and Applied Mathematics, Mind Association, Phi Beta Kappa.

DAN L. BISBEE, Western Electric Co., 1953; Bell Telephone Laboratories, 1955—. He has been engaged in construction and use of microwave test equipment for long-distance waveguide communications experiments. This has included microwave measurements of waveguide components in the millimeter wavelength region.

A. S. CORIELL, B.S., 1947, Lehigh University; Bell Telephone Laboratories, 1958—. He has been engaged in research on short-term storage in vision and in an experiment on the apparent brightness of brief flashes of light. Member Optical Society of America.

HERMANN K. GUMMEL, Dipl. Phys., 1952, Philipps University (Germany); M.S., 1952, and Ph.D., 1957, Syracuse University; Bell Telephone Laboratories, 1956—. His work has been in research and development of semiconductor devices. Member American Physical Society, Sigma Xi.

DAVID A. KLEINMAN, S.B., 1946, and S.M., 1947, Massachusetts Institute of Technology; Ph.D., 1952, Brown University; Brookhaven National Laboratory, 1949–53; Bell Telephone Laboratories, 1953—. He has worked in the areas of neutron scattering in solids, semiconductor electronics, electron energy bands and the infrared properties of crystals. Member American Physical Society.

Henry J. Landau, A.B., 1953, and Ph.D., 1957, Harvard University; teaching fellow, Harvard, 1956–57; Bell Telephone Laboratories, 1957—. He has been engaged in mathematical research in function theory and harmonic analysis. In 1959–60 he was on leave of absence from Bell Laboratories for study at the Institute for Advanced Study. Member American Mathematical Society, Phi Beta Kappa, Sigma Xi.

Jessie MacWilliams, B.A., 1939, and M.A., 1958, Cambridge University (England); Bell Telephone Laboratories, 1956—. Mrs. MacWilliams has been engaged in writing computer programs, relating primarily to network design and analysis. Member Mathematical Association of America.

E. Marcatili, Aeronautical Engineer, 1947, and E.E., 1948, University of Cordoba (Argentina); research staff, University of Cordoba, 1947–54; Bell Telephone Laboratories, 1954—. He has been engaged in theory and design of filters in multimode waveguides. More recently he has concentrated on waveguide systems research. Member I.R.E., Physical Association of Argentina.

H. O. Pollak, B.A., 1947, Yale University; M.A., 1948, and Ph.D., 1951, Harvard University; Bell Telephone Laboratories, 1951—. He has been engaged in mathematical analysis of gunnery and missile systems and in mathematical research in communications. Member American Mathematical Society, Mathematical Association of America.

Irwin W. Sandberg, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—. He has been concerned with analysis of military systems, particularly radar systems, and with synthesis and analysis of active and time-varying networks. Recently he transferred to a group engaged in research on communications fundamentals. Member I.R.E., Eta Kappa Nu, Sigma Xi, Tau Beta Pi.

David Slepian, University of Michigan, 1941–43; M.A., 1947, and Ph.D., 1949, Harvard University; Bell Telephone Laboratories, 1950—.

He has been engaged in mathematical research in communication theory, switching theory and theory of noise, as well as various aspects of applied mathematics. He has been mathematical consultant on a number of Laboratories projects. In 1958 and 1959 he was Visiting Mackay Professor of Electrical Engineering at the University of California at Berkeley. Member A.A.A.S., American Mathematical Society, Institute of Mathematical Statistics, I.R.E., Society of Industrial and Applied Mathematics, U.R.S.I. Commission 6.

Friedolf M. Smits, Dipl. Phys., 1950, and Dr. rer. nat., 1950, University of Freiburg (Germany); research associate, Physikalisches Institut, University of Freiburg, 1950–54; Bell Telephone Laboratories, 1954—. His past work included studies of solid-state diffusion in germanium and silicon, as well as device feasibility and process studies. At present he heads a group concerned with the development of UHF semiconductor devices. Member American Physical Society, German Physical Society.

Hans-Georg Unger, Dipl. Ing., 1951 and Dr. Ing., 1954, Technische Hochschule, Braunschweig (Germany); Siemens and Halske (Germany), 1951–55; Bell Telephone Laboratories, 1956—. His work at Bell Laboratories has been in research in waveguides, especially circular electric wave transmission. He is now on leave of absence from Bell Laboratories as professor of electrical engineering at the Technische Hochschule in Braunschweig. Senior member I.R.E.; member German Communication Engineering Society.

Hao Wang, B.S., 1943, Sinan Lienta (China); Ph.D., 1948, Harvard University; Bell Telephone Laboratories, 1959–60; Reader in the Philosophy of Mathematics, Oxford University (England), 1956—. A logician, Dr. Wang has been engaged in a long-range research project to prove mathematical theorems with the use of modern computers. He pursued this research on a year's leave of absence from Oxford University at Bell Laboratories as a member of a group engaged in research on advanced uses of computers.

Editor's Note: The listing of "Recent Monographs of Bell System Technical Papers" will no longer be published in each issue. Instead, a numerical list of published monographs will be sent semi-annually to Bell System Technical Journal subscribers.