THE BELL SYSTEM TECHNICAL JOURNAL

VOLUME XLV NOVEMBER 1966 NUMBER 9

Programming and Control Problems Arising from Optimal Routing in Telephone Networks*

By V. E. BENEŠ

(Manuscript received June 10, 1966)

In many circumstances a telephone call can be completed through a connecting network in several ways. Hence, there naturally arise problems of optimal routing, that is, of making the choices of routes so as to achieve extrema of one or more measures of system performance, such as the loss (probability of blocking) or the carried load.

As is customary in traffic theory, a Markov process is used to describe network operation with complete information. The controlled system is described by linear differential equations with the control functions (expressing the routing method being used) among the coefficients. Restricting attention to asymptotic behavior leads to a problem of maximizing a bilinear form subject to a linear equality constraint whose matrix is itself constrained to lie in a given convex set. An alternative approach first shows that minimizing the loss, and maximizing the fraction of events that are successful attempts to place a call, are equivalent. This fact permits a dynamic programming formulation, which, in turn, leads to a very large linear programming problem. Two small examples are treated numerically by this method.

It is particularly important to try to verbalize, and then mechanize, the optimal routing strategies. In this endeavor, the linear programming formulation is of limited usefulness. Therefore, in the latter half of the work we

[*] Presented at the First International Conference on Programming and Control, USAF Academy, Colorado, April 15–16, 1965. A prolonged abstract of the present text appears as part of the proceedings of the Conference in the SIAM Journal on Control, vol. 4, number 1, February 1966, pp. 6–18.

have attempted to use the special combinatorial structure imposed by the telephonic origins of the problem to shed light on the character of the optimal strategies. In particular, we show that for connecting networks with suitable combinatorial properties, the optimal route choices can be very simply described. Some of the results obtained were suggested by, and verify, conjectures from the practical lore of telephone routing.

The problem of routing calls falls into two parts: Which attempted calls should be accepted in which states? What route should an accepted call use? The first problem is very hard, and only sample numerical answers for small networks are obtained. We solve the second problem analytically for a large class of cases by appeal to combinatorial structure in the network. These cases can be described roughly as those in which the relative merit of states (as far as blocking is concerned) is consistent or continuous; i.e., if a state x is "better" than another y, then the neighbors of x are in the same sense "better" than the corresponding neighbors of y. An abundance of examples indicates that these cases are numerous and so warrant attention. In a network with this kind of combinatorial property, a policy which rejects no unblocked calls and minimizes the number of additional calls that are blocked by completing an attempted call differs from an optimal policy only in that the latter may reject some calls.

I. INTRODUCTION

A telephone connecting network invariably provides many paths on which a particular telephone call can be completed. One of the operational problems faced by the control unit of a telephone system is then to assign to each accepted and completable call a path and, in particular, to choose these assigned paths in the best way. This is the problem of optimal routing of telephone calls. Thus, in the theory of telephone traffic there naturally arise mathematical problems of optimal routing, that is, of making choices of routes in probabilistic models for operating networks so as to achieve extrema of well-defined measures of system performance, such as the probability of blocking (loss).

Unfortunately, it is not unfair to state that the voluminous probabilistic theory of telephone traffic, now some sixty years old, still has rather little to say about how routes for calls should be chosen. We are speaking here of the *mathematical theory* of traffic. Naturally, a wealth of useful information about routing has accumulated over the years from experience in the telephone field; recently it has been buttressed and extended by many simulation studies. This information, nevertheless, still lies largely outside the province of the existing theory of telephone traffic.

It is the aim of this work to formulate, study, and (in part) solve a general class of optimal routing problems for telephone networks. The formulation of these problems is undertaken insofar as possible within the classical dynamical theory of telephone traffic initiated by A. K. Erlang, that is, in terms of Markov processes based on the assumptions of (*i*) negative exponential distributions for mutually independent holding-times, and (*ii*) randomly originating traffic. To these assumptions is added a description of how attempted calls are accepted and assigned routes.

We conclude this introduction with a brief summary of the entire paper. A complete summary appears later (Section IX) after concepts for formulating the problem have been discussed. As is customary in telephone traffic theory, we use a Markov process to describe the operation of the connecting network under study. The Kolmogorov equations for this process then constitute a set of linear differential equations describing the controlled system; in these the control functions expressing the routing method being used appear among the coefficients. It is natural to restrict attention to asymptotic behavior; this leads to a problem of maximizing a bilinear (or linear fractional) form subject to linear constraints; this problem is equivalent to a linear programming problem. An alternative approach first shows that minimizing the probability of loss, and maximizing the fraction of events that are successful call attempts, are equivalent. This fact permits a classical dynamic programming approach. The remainder of the paper attempts to use this approach to establish relations between combinatorial properties of the network and the policy(ies) optimal for given criteria of performance. In particular, it is shown that for connecting networks having certain "monotone" properties, optimal policies for minimizing loss correspond closely to the heuristic advice, "Prefer those states in which as few calls are blocked as possible".

II. INFORMATION FOR ROUTING DECISIONS

The problem of choosing "good" routes for information flow in a communications network is vastly complicated by the difficult questions surrounding the collection, updating, and relevance of information (about the state of the system) on the basis of which routing decisions are to be made. Thus, one of the items to be chosen in designing a routing scheme is the information on which the routing is to be based. Indeed there is a whole spectrum of possible choices for this information, from no information at all (except what is unwittingly discovered in making call attempts), to full knowledge of the state of the connecting network. Clearly, a practical compromise between total ignorance and a very expensive, complex scheme based on many data must usually be made.

Our considerations in this work will be limited to the case of perfect information, in which the microscopic state of the connecting network is assumed known and available for making routing decisions. This case is, of course, very far from realistic: few existing or envisaged systems utilize even a small fraction of this possible information for routing. Indeed, much of it is likely to be of very little relevance. Nevertheless, it is important to know what would be good routing if we could implement it and could afford it, so the full information case to be considered here forms at worst a limiting situation for which some theory is available, and a natural starting point for investigation.

III. ACCEPTANCE OR REJECTION OF UNBLOCKED CALLS

In the present discussion of the involved problem of routing calls, one of the difficulties that arises deserves special mention. This difficulty is the problem of deciding whether to accept or reject attempted calls which are not blocked.

At first sight, it might seem that no unblocked call attempt should ever be rejected. The natural argument for this view is that the whole point of a telephone system is to complete calls, and that by rejecting an attempt that could have been completed, the system only lowers its performance. Sensible as this argument sounds, it is unacceptable because it turns out that whether rejection of an unblocked call improves or lowers performance depends on the index of performance, on the distribution of traffic among the sources, on the "community of interest" aspects of the system, etc. If the probability of blocking is used as an index, the "bad" effect of adding a particular call in a given state of the system may be so great and so lasting that it is better to reject the call, and improve the chance of completing many later calls.

To put the matter another way, the problem of routing with full information seems at first to boil down to the question: "Which of the paths available for call c in state x should be used?" This form of the problem overlooks the possibility that perhaps the best thing to do when the state is x and c is attempted is not to complete c at all, but to reject it! In other words, it assumes that, naturally, c will be put up in state x if it is attempted in x and is not blocked. This assumption has always been made in previous applications of the model we use.[1,2]

Conceivably, then, it is better to reject a call c that is not blocked in a state x. Thus the problem of routing should be phrased: "Should a call c, free and not blocked in state x, be completed, and if so, by which route?"

It turns out that answering the first part of the question, as to which calls should be completed in which states, is often the hardest part of the problem. Examples can be given in which it is fairly easy to solve the route selection part of the problem, but for which the question of whether a call should go in or not is not settled. That this question has substantial practical import is apparent from the simulation studies carried out by J. H. Weber,[3] which clearly show how in trunking networks prohibition of circuitous routes (and thus rejection of certain unblocked calls) can improve system performance.

J. H. Weber[4] has also remarked that the problem of deciding whether an unblocked call should be refused is closely related to the distinction between *trunking networks*, used in toll systems to interconnect towns and cities, and *central office networks*, used to interconnect trunks and customers' lines at a single location. An important combinatorial difference between the two types of networks depends on whether all calls use the same number of links. This is usually the case in central office networks, but rarely true in trunk networks. One result suggested by this distinction would be that a call should always be put up when all calls use the same number of links, but that circuitous routes might be profitably disallowed otherwise.

It appears then that network structure bears on the problem of what calls to accept. However, examples can be given which show that even when there is almost no network structure, other factors such as the distribution of traffic and the "community of interest" can make rejection of some calls part of an optimal policy.

For example, if two lines calling at rates λ_1, λ_2, respectively, compete for one trunk, the probability of blocking is

$$\frac{2\lambda_1\lambda_2}{\lambda_1 + \lambda_2 + 2\lambda_1\lambda_2},$$

if no unblocked call is rejected. If the calls of the line calling at rate λ_1 are always rejected, the probability of blocking (with rejected calls included among the blocked) is

$$\frac{\lambda_1\lambda_2 + \lambda_1}{\lambda_1 + \lambda_2 + \lambda_1\lambda_2}.$$

(We have assumed that all calls have unit mean holding-time.) It follows here that if

$$\lambda_2^2 > \lambda_1 + \lambda_2 + \lambda_1\lambda_2$$

then it is better to reject all λ_1 calls than to put them all in! This example, although somewhat unrealistic, illustrates how the distribution of traffic affects the rejection problem, even in the absence of network structure.

For an example involving the "community of interest", consider two disjoint sets of $(n + 1)$ lines communicating over one trunk, with the quirk that each set has a distinguished line which only attempts calls to the distinguished line in the other set, while the other n lines of one set only attempt calls to the n nondistinguished lines of the other set. Let c be the call consisting of the two distinguished lines talking to each other. If c is always rejected, the probability of blocking is

$$\frac{1 + \lambda n^2(n - 1)^2}{n^2 + \lambda n^2(n - 1)^2},$$

where we have assumed that lines which call each other do so at rate λ, and holding-times have unit mean. If c is always accepted when it is not blocked, then the probability of blocking is

$$\frac{2\lambda n^2 + \lambda n^2(n - 1)^2}{2\lambda n^2 + 1 + n^2 + \lambda n^2(n - 1)^2}.$$

From these formulas it follows that it is better to reject c entirely if n is large enough, or if λ is large enough, while if λ is small enough it is better always to accept c.

IV. STATES, EVENTS, AND ASSIGNMENTS

The elements of the mathematical model to be used for our study of routing separate naturally into combinatorial ones and probabilistic. The former arise from the structure of the connecting network and from the ways in which calls can be put up in it; the latter represent assumptions about the random traffic the network is to carry. The combinatorial and structural aspects are discussed in this section; terminology and notation for them are introduced. The probabilistic aspects are considered in a later section.

A *connecting network* ν is a quadruple $\nu = (G,I,\Omega,S)$, where G is a graph depicting network *structure*, I is the set of nodes of G which are *inlets*, Ω is the set of nodes of G that are *outlets*, and S is the set of permitted *states*. Variables x,y,z at the end of the alphabet denote states, while u and v (respectively) denote a typical inlet and a typical outlet. A state x can be thought of as a set of disjoint chains on G, each chain joining I to Ω. Not every such set of chains represents a state: sets with wastefully circuitous chains may be excluded from S. It is possible that $I = \Omega$, that $I \cap \Omega = \theta =$ null set, or that some intermediate condition obtain, depending on the "community of interest" aspects of the network ν.

The set S of states is *partially ordered* by *inclusion* $\leqq$, where $x \leqq y$ means that state x can be obtained from state y by removing zero or more calls. If x and y satisfy the same *assignment* of inlets to outlets, i.e., are such that all and only those inlets $u \in I$ are connected in x to outlets $v \in \Omega$ which are connected to the same v in y (though possibly by different *routes*), then we say that x and y are *equivalent*, written $x \sim y$.

The set S of states determines another set $\mathcal{E}$ of *events*, either *hangups* (terminations of calls), *successes* (successful call attempts), or *blocked or rejected calls* (unsuccessful call attempts). The occurrence of an event in a state may lead to a new state obtained by adding or removing a call in progress, or it may, if it is a blocked call or one that is rejected, lead to no change of state. Not every event can occur in every state: naturally, only those calls can hang up in a state which are in progress in that state, and only those inlet-outlet pairs can ask for a connection between them in a state that are idle in that state. The notation e is used for a (general) event, h for a hangup, and c for an attempted call. If e can occur in x we write $e \in x$. A call $c \in x$ is *blocked* in a state x if there is no $y \in S$ which covers x in the sense of the partial ordering $\leqq$ and in which c is in progress. For $h \in x$, $x - h$ is the state obtained from x by performing the hangup h.

We denote by A_x the set of states that are immediately above x in the partial ordering $\leqq$, and by B_x the set of those that are immediately below. Thus,

$$A_x = \{\text{states accessible from } x \text{ by adding a call}\}$$

$$B_x = \{\text{states accessible from } x \text{ by a hangup}\}.$$

For an event $e \in x$, the set A_{ex} is to consist of those states $y \neq x$ to which the network might pass upon the occurrence of e in x. Thus, if e is a blocked call, $A_{ex} = \{\theta\}$; also

$$\bigcup_{h \in x} A_{hx} = B_x$$

$$\bigcup_{\substack{c \in x \\ c \text{ not blocked in } x}} A_{cx} = A_x .$$

The *number of calls in progress* in state x is denoted by $|x|$. The number of call attempts $c \in x$ which are not blocked in x is denoted by $s(x)$, for "*successes in* x." The functions $|\cdot|$ and $s(\cdot)$ defined on S play important roles in the stochastic process to be used for studying routing.

It can be seen, further, that the set S of states is not merely partially ordered by $\leqq$, but also forms a semilattice, or a partially ordered system with intersections, with $x \cap y$ defined to be the state consisting of those calls and their respective routes which are common to both x and y. (See G. Birkhoff,[5] p. 18, ex. 1 and footnote 6.)

An *assignment* is a specification of what inlets should be connected to what outlets. The set A of assignments can be represented as the set of all fixed-point-free correspondences from I to Ω. The set A is partially ordered by inclusion, and there is a natural map $\gamma(\cdot): S \to A$ which takes each state $x \in S$ into the assignment it realizes; the map $\gamma(\cdot)$ is a semilattice homomorphism of S into A, since

$$x \geqq y \quad \text{implies} \quad \gamma(x) \geqq \gamma(y),$$
$$\gamma(x \cap y) \leqq \gamma(x) \cap \gamma(y).$$

V. ROUTING MATRICES

It will be assumed throughout this work that attempted calls to busy terminals are rejected, and have no effect on the state of the network; similarly, blocked attempts to call an idle terminal are refused, with no change of state. Attempts to place a call are completed instantly with some choice of route, or are rejected, in accordance with some policy of routing.

Two mathematical descriptions of how routes are assigned to calls will be used. The first, the *routing matrix*, is convenient for writing the Kolmogorov equations for the Markov processes representing network operation. The second, called a *policy*, affords a convenient notation for the actual determination of optimal routing methods for various networks to be described in detail later. Either description is a *rule* or *doctrine* for routing.

A routing matrix $R = (r_{xy}),x,y \in S$, has the following properties: for each $x \in S$, let Π_x be the partition of A_x induced by the equivalence relation $\sim$ of "having the same calls up," or satisfying the same assignment of inlets to outlets; then for each $Y \in \Pi_x$, r_{xy} for $y \in Y$ is a *possibly improper* probability distribution over Y, (that is, it may not sum to unity over Y),

$$r_{xx} = s(x) - \sum_{y \in A_x} r_{xy},$$

and $r_{xy} = 0$ in all other cases.

The interpretation of the routing matrix R is to be this: any $Y \in \Pi_x$ represents all the ways in which a particular call c not blocked in x

(between an inlet idle in x and an outlet idle in x) *could* be completed when the network is in state x; for $y \in Y$, r_{xy} is the chance that if this call c is attempted in x, it will be completed by being routed through the network so as to take the system to state y. That is, we assume that if c is attempted in x, then with probability

$$1 - \sum_{y \in A_{cx}} r_{xy} \tag{1}$$

it is rejected (even though it is not blocked), and with probability r_{xy} it is completed by being assigned the route which would change the state x to y, for $y \in A_{cx}$. The possibly improper distribution of probability $\{r_{xy}, y \in Y\}$ indicates how the calling rate λ due to c is to be spread over the possible ways of putting up the call c, while the improper part (1) is just the chance that it is rejected outright.

This description of routing matrices is a generalization of that used in Refs. 1 and 2 in that it permits, in the nonvanishing of (1), the rejection of unblocked calls forbidden in the cited references.

Thus, a routing matrix R is any function on S^2 with $r_{xy} \geqq 0$, $r_{xy} = 0$ unless $y \in A_x$ or $y = x$, and such that

$$r_{xx} = s(x) - \sum_{y \in A_x} r_{xy}$$

and

$$\sum_{y \in A_{cx}} r_{xy} \leqq 1,$$

for all $c \in x$ not blocked in x. A routing matrix corresponds to a *fixed rule* if $r_{xy} = 0$ or 1 for $x \neq y$; otherwise it corresponds to a *randomized rule*. The convex set of all possible routing matrices is denoted by C.

A *policy* is a function φ: $\mathcal{E} \times S \rightarrow S$ such that $c,h \in x$ imply

$$\varphi(c,x) \in A_{cx} \cup \{x\}$$

$$\varphi(h,x) = x - h.$$

It is apparent that a policy is equivalent to a fixed rule; the circumstance that $\varphi(\cdot,x)$ is defined also for hangups h is useful in the sequel. Variables φ,ψ are used to denote policies.

The routing rules and doctrines that might be considered here are of course more numerous by far than those we have introduced above. In particular, time-dependent rules and history-dependent rules are natural generalizations. However, since we will be considering only time-invariant traffic and ergodic Markov processes as representations of operating networks, such generalizations add little of significance.

An important point, however, is that the routing methods here considered are based on a complete knowledge of the state of the system, i.e., we postulate that we are in the case of "perfect information." This postulate is grossly unrealistic for present day electromechanical telephone systems; for an electronic system with a very large and very cheap memory, it becomes realistic: the state of the network can actually be stored and the routing rule in use represented by a giant translator. Such a procedure overcomes the obvious impracticality of determining the state by examination of the actual network, and is actually used in the Bell System's No. 1 ESS (Electronic Switching System).[6]

The routing matrices R used in Refs. 1 and 2 had the property that if a call is not blocked in a state, then it is completed in *some* way; *only* blocked attempts or attempts to busy terminals are rejected. Thus none of these rules for routing resembles the methods that are at present likely to be used in practice. However, since C contains rules that reject certain calls in certain states, even though these calls are not blocked, it turns out that a large class of routing rules which do mirror what might happen in practice is included in C.

Some of the simplest routing rules are not based on any knowledge about the current state of the network. Given a call c that has been attempted, they provide a list of routes to be tried in order; the first route found available is used for the call. The list may include all possible routes for c, or only some of them. It is easy to construct a routing matrix to represent such a rule. Let $r_1, r_2, \cdots, r_n$ be the routes to be tried for a call c. For each state x in which c can occur, let $r_{xy} = 1$ if use of the first r_i that is available in x takes the system from x to y, and let $r_{xy} = 0$ for all other $y \in A_{cx}$. If no route for c that is available in x is among $r_1, \cdots, r_n$, then c is rejected in x even though it may not be blocked, simply because the "sieve" for finding routes is too coarse.

It was assumed in the previous paragraph that no information about the state was used. If it is known, e.g., in which element A of a partition Π of S the state currently is, a similar rule can be represented by a class of lists (of routes to be tried in order), one for each $A \in \Pi$. The same kind of construction then yields the appropriate R. Here the A such that $x_t \in A$ is acting as the "information state."

Thus, many R from C which reject certain calls in certain states describe a rule which closely resembles what is done in practice, e.g., in the translator of the Bell System No. 4A crossbar switching system.

VI. PROBABILISTIC ASSUMPTIONS AND STOCHASTIC PROCESSES

A Markov stochastic process x_t taking values on S is used as a mathematical description of an operating connecting network subject to random traffic. It is assumed that this operation is in accordance with one of the routing matrices R of Section V. The rest of the process x_t is based on two simple probabilistic assumptions:

(*i*) Holding-times of calls are mutually independent variates, each with the negative exponential distribution of unit mean.

(*ii*) If u is an inlet idle in state x, and $v \neq u$ is any outlet, there is a (conditional) probability

$$\lambda h + o(h), \qquad \lambda > 0$$

that u attempt a call to v in $(t, t + h)$ if $x_t = x$, as $h \to 0$.

The choice of unit mean for the holding-times merely means that the mean holding-time is being used as the unit of time, so that only the traffic parameter λ needs to be specified.

It is convenient to collect these assumptions and the chosen routing matrix R into one transition rate matrix $Q = (q_{xy})$ characteristic of x_t: this matrix is given by

$$q_{xy} = \begin{cases} 1 & \text{if} \quad y \in B_x \\ \lambda r_{xy} & \text{if} \quad y \in A_x \\ -\mid x \mid - \lambda[s(x) - r_{xx}] & \text{if} \quad y = x \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

In terms of the transition rate matrix Q, it is possible to define an ergodic stationary Markov stochastic process $\{x_t, t \text{ real}\}$ taking values on S. The matrix $P(t)$ of transition probabilities

$$P_{xy}(t) = Pr\{x_t = y \mid x_0 = x\}$$

satisfies the equations of Kolmogorov

$$\frac{d}{dt} P(t) = QP(t) = P(t)Q, \qquad Q(0) = I,$$

and is given formally by the formula

$$P(t) = \exp tQ.$$

Since the zero state (the state with no calls in progress) is accessible from any state in a finite number of steps with positive probability, the process has only one ergodic class, and there exists a unique nonnegative row-vector

$$p = \{p_x, x \in S\}$$

such that as $t \to \infty$

$$P(t) \to \begin{pmatrix} p \\ \vdots \\ p \end{pmatrix},$$

and p satisfies the "statistical equilibrium" or stationarity condition $p'Q = 0$, which can be written out in full in the simple form

$$[|x| + \lambda s(x) - \lambda r_{xx}]p_x = \sum_{y \in A_x} p_y + \lambda \sum_{y \in B_x} p_y r_{yx}, \qquad x \in S.$$

It is possible that a confusion arises in the mind of the reader as to whether we are talking about central office connecting networks or large trunk networks such as the toll system. For in telephone traffic theory these two areas of application are often described by different models: a "finite-source" model like the present one, in which the conditions of the inlets and outlets form a significant part of the state of the system, is commonly used for the former; an "infinite source" model, with groups of customer's lines reduced to Poisson sources of traffic, is frequently used for the latter. The reason for this difference is that it has simply turned out to be sufficient, in the toll case, to restrict attention to the trunking network as the object of principal interest, and to use the simpler Poisson description of sources.

In principle, of course, the model to be used here serves to describe either area listed above, although in the toll case it naturally demands use of a very large number of states. Thus, in the sequel we make no attempt to distinguish the toll case from the central office case. This viewpoint is justified by the fact that the results to be obtained are robust under passage from finite- to infinite-source models, or they can be reformulated and reproved in the infinite-source context.

VII. FORMULATION OF THE ROUTING PROBLEM

The most common figure of merit used by telephone traffic engineers for evaluating connecting networks is the probability of blocking, the fraction of call attempts that are blocked. It is natural, therefore, to use this quantity as the objective function in our optimization problem of routing. It has been shown[2] for the process x_t to be studied here that if no unblocked call is rejected the probability of blocking (in the mnemonic form $Pr\{bl\}$) is given in terms of the stationary state probability vector p by the formula

$$Pr\{bl\} = \frac{\sum_{x \in S} p_x \beta_x}{\sum_{x \in S} p_x \alpha_x} = \frac{p'\beta}{p'\alpha},$$

where

β_x = number of idle inlet-outlet pairs that are blocked in state x,

α_x = number of idle inlet-outlet pairs in state x.

By the same methods it follows that for a process x_t defined in terms of an $R \in C$ the fraction of attempted calls which are not completed (are "lost"), be it because they were blocked or simply rejected, is given by

$$\frac{p'(\beta + r)}{p'\alpha},$$

where $r = \{r_{xx}, x \in S\}$ is the diagonal of the routing matrix R.

We can now replace the informal problem of minimizing, by suitable routing, the fraction of call attempts that are lost by a precise problem of mathematical programming, as follows: Choose $R \in C$ so as to achieve

$$\min \frac{p'(\beta + r)}{p'\alpha}$$

subject to $p'Q = 0$, $p'1 = 1$, and $p \geqq 0$. (The '1' in '$p'1$' is the vector with all components 1.) Of the constraints, the first is the equilibrium condition on p, the second states that the components of p sum to one, and the third says that p is nonnegative. It is understood, of course, that Q is to be related to R by (2) or, what is the same, by

$$Q = H + \lambda R - \operatorname{diag}(|x| + \lambda s(x) + 2\lambda r_{xx}) = Q(R),$$

where $H = (h_{xy})$ is the "hangup matrix" such that $h_{xy} = 1$ or 0 according as $y \in B_x$ or not.

Several authors have formulated routing problems for communications systems. Many of these problems have dealt with systems of the store-and-forward type, in which information is alternately stored at and transmitted from a node in the network without setting up a "continuous path" from source to destination. Such formulations are inapplicable to telephone systems. A possible exception, though, is that[7] of R. Kalaba and M. Juncosa which, for a given amount of traffic between each specified source and destination, and a given network having capacity constraints, attempts to find continuous routes that are best in the sense of maximizing the delivered traffic by solving a linear programming problem.

In its possible application to telephony, this model envisions a given traffic pattern (i.e., a description of who wants to talk to whom) to be satisfied at a particular moment, and tries to find a way of routing as much of this traffic as possible through the network. In our terminology, a traffic pattern is an *assignment* $a(\cdot)$, and satisfying it means finding an $x \in S$ such that $\gamma(x) = a$. The amount of traffic carried is simply the number $|x|$ of calls in progress. Of course, it is not always possible to satisfy an assignment. Thus, Kalaba's and Juncosa's formulation translates into our setup as follows: Given an assignment $a(\cdot)$ either find $x \in S$ with $\gamma(x) = a$, or else if $a(\cdot)$ is unrealizable, find $x \in S$ which realizes as much of $a(\cdot)$ as possible, i.e., such that $\gamma(x) \leqq a$ and $|x|$ is a maximum. This can be rephrased as follows: If $a(\cdot)$ is given, form the cone

$$K = K(a) = \{a_1 : \quad a_1 \leqq a\},$$

and within $\gamma^{-1}(K)$ pick a state x that is maximal in that $|x| \geqq |y|$ for each $y \in \gamma^{-1}(K)$.

It is to be emphasized that this problem is markedly different from our form of the routing problem. The former is purely combinatorial in character. There is no parameter such as the traffic λ per inlet-outlet pair, so the problem involves no probability, and can have nothing to do with the "grade of service" as customarily employed by telephone engineers. Furthermore, the whole formulation overlooks the fact that in present systems call completions must be made without disturbing calls already in progress.

VIII. PRINCIPLES OF ROUTING

It is important to distinguish *methods* of routing from *principles* of routing. A method of routing is a specific way of accepting or rejecting attempted calls and choosing routes in a particular system, e.g., that implicit in the translator of the Bell System No. 4A crossbar switching system. A *principle* of routing is a kind of general prescription of what constitutes* "good" or "optimal" routing; it is the backbone of many routing methods that might be based on it.

A principle of routing is particularly useful if it has two properties:

(*i*) It is relatively simple and intuitive to state.

(*ii*) There is a substantial class of systems for which it describes the (or part of the) optimal routing method.

In our mathematical setting a method of routing corresponds roughly to a rule $R \in C$. We shall see that the "best" rule $R \in C$ can be obtained by solving a linear programming problem. Now if it should happen that for an interesting class of networks the solutions of these linear programs had some common characteristic, some combinatorial property

* Or, more usually, of what someone's intuition tells him constitutes.

of the sets of states of the networks that served as an alternate description of the linear program solution, then this characteristic or property could be abstracted into a genuine *principle of routing*.

Alternatively, one could formulate as conjectures some intuitive principles of routing, and then try to determine for what classes of networks (if any!) these principles did, in fact, describe the optimum routing methods. This second approach will be followed in the present work; the rest of this section is devoted to a discussion of some *a priori* reasonable candidates for "good" routing rules. All of these candidates are expressions of one and the same idea, namely, that one routing rule is better than another if it avoids more "bad" states, where a "bad" state x is one for which β_x is high. This idea is not just an attractive first approximation to "good" or even optimal routing; it leads at once to conjectures for which our results later in the paper provide strong support in precise ways.

In spite of the lack of general theoretical knowledge about routing, traffic engineers have developed various conjectures and intuitive ideas about what might constitute "good" methods for choosing routes. These conjectures are a natural starting place for any rigorous approach to routing, because the formulation of precise theoretical models in which routing can be studied at once raises the question, "Which of these methods, conjectured to be good, can be proved to be optimal in some theoretical model?" Since many of these methods are relatively simple to describe, and hence to mechanize, established answers to this question would have immediate practical applications. Some of these conjectures will now be discussed.

It is apparent that in a telephone system, putting up a new call can only increase the number of idle pairs that are already blocked. Another way of saying this is that in giving service, i.e., in realizing an attempted call in a connecting network, one is possibly denying service to certain inlets and outlets presently idle, who might attempt a call in the very immediate future. This observation has given rise to a number of routing rules (for systems with blocked attempts refused) of great intuitive appeal, which can be described collectively by the admonition: To decrease (minimize?) the probability of blocking, put in new calls in such a way as to minimize the additional congestion resulting from the new calls.

It is illuminating to discuss particular forms of this advice. One form is this: Route new calls through the most heavily loaded part of the network that will accept them. Another is: Put in a given new call so as to minimize the chance that the next attempt to place a call be blocked.

Or: Avoid blocking states, that is, prefer states in which fewer idle pairs are blocked.

For all the intuitive appeal possessed by these rules, rather little is known about them. Nevertheless, they provide conjectures that will be examined in the precise setting of our theoretical model to yield, we hope, the beginnings of a mathematical theory of optimal routing. Let us see what these rules enjoin in terms of our model. If we put up a call c so as to take the system to a state y, the chance that the next event is a blocked call attempt is

$$\frac{\beta_y}{|y| + \lambda\alpha_y}.$$

Suppose that we just left state x, so that $y \in A_{cx}$. This probability will be smallest if y was chosen according to the "maximum $s(\cdot)$" policy, that is,

$$s(y) = \max_{z \in A_{cx}} s(z),$$

i.e., if we prefer states in which fewer idle pairs are blocked. Thus, in our model the second two forms of the above advice coincide.

Another conjecture arises out of consideration of *gradings* in which calls overflowing certain primary routes are pooled and offered to overflow circuits. Here a natural expectation is that one should always "fill the holes in the multiple," meaning by this that a primary route should be used whenever possible, so that the overflow is left available to as many lines as possible. It will be shown for certain examples that if calls are accepted unless they are blocked, then this rule both describes the optimum routing choices, and is equivalent to the "maximum $s(\cdot)$" policy of the previous paragraph.

IX. SUMMARY AND DISCUSSION

In Sections I to VII the problem of routing calls in a telephone network has been formulated as a mathematical one within Erlang's basic traffic theory. Some routing rules which are intuitively reasonable candidates for "good" or even optimal routing were described in Section VIII.

Since the expansion of $\{p_x, x \in S\}$ such that $p'Q = 0$, $p > 0$, is known,[1,2] it is natural to start in Section X with a consideration of $Pr\{bl\}$ for low traffic: $\lambda \to 0$. We have

$$p_x = p_0 \frac{\lambda^{|x|}}{|x|!} r_x + o(\lambda^{|x|}), \qquad \lambda \to 0,$$

where r_x is the number of strictly ascending (in $\leqq$) paths from $\mathbf{0}$ to x which are permitted by R. If x is a blocking state it contributes a term

$$\frac{p_x\beta_x}{p'\alpha} = p_0 \frac{\lambda^{|x|} r_x \beta_x}{|x|!\, p'\alpha} + o(\lambda^{|x|}), \qquad \lambda \to 0$$

to $Pr\{bl\}$ if no calls are rejected. It follows that for sufficiently low traffic the policy that minimizes r_x is optimal within the policies that reject no calls. In a similar way, it can be shown that *always refusing* a call c cannot be optimal for λ sufficiently small, and that there is never any point in rejecting a call attempt in a state x with

$$|x| < \min\{|y|: \quad y \in S, \beta_y > 0\},$$

for λ small enough.

The *nonlinear* problem of choosing R to minimize $Pr\{bl\}$ is reduced to a *linear programming* problem in Section XI. This reduction substantially facilitates obtaining numerical results, examples of which appear later in this summary.

In an effort to identify optimal routing policies, attention now (Section XII) shifts away from the formal linear programming approach to the underlying Markov process. It is shown that minimizing $Pr\{bl\}$, and maximizing the fraction of events which are successful call attempts, are equivalent; this fact leads to a direct dynamic programming approach, in which

$$\min_{R \in C} \quad Pr\{bl\}$$

and

$$\lim_{n\to\infty} \quad n^{-1} \max E\{\text{number of successful call attempts in } n \text{ events}\}$$

(with the maximum in the second expression over all possible policies for n events) are both achieved by essentially the same stationary policies. The word 'essentially' hides the inherent nonuniqueness of optimal policies due to symmetries in the network and to the possible presence of transient states.

In Section XIII it is shown, following C. Derman, that minimum blocking is achieved by a *fixed* rule.

The mathematical programming problems arising in this new approach are again of the linear programming type, and are similar to those arising in Section XI. Our principal interest, however, does not remain with calculating numerical solutions, but shifts abruptly to the relationships of these solutions to the combinatorial structure of the network. Thus, the second half of this paper consists less of suitable programming problems than of intuition and combinatorics applied to exhibit (*in parte* or *in toto*) the solutions of these problems and their dependence on and origin in network structure.

The attempt to discover and characterize optimal policies in a wholesale way by appeal to network combinatorics (rather than piecemeal by numerical calculation) begins in Sections XIV and XV with consideration of some simple examples; these lead to the introduction of some "monotone" properties (of connecting networks) which impose the condition that (roughly) the relative merit (as far as blocking is concerned) of states is consistent or continuous, i.e., that if a state x is "better" than another y, then the neighbors of x are in the same sense "better" than the corresponding neighbors of y.

Consideration of these properties is justified by the facts that (*i*) they appear in the examples, and (*ii*) they yield a series of closely knit results (Theorems 7–15) that go far to bear out the heuristic guesses in Section VIII about the nature of good routing. In particular, in a network with one of the monotone properties, a policy which rejects no unblocked calls and minimizes the number of additional calls that are blocked by completing an attempted call differs from an optimal policy only in that the latter may reject some calls. In other words, the "max $s(\cdot)$" policy is optimal to within rejection of calls.

Each monotone property gives rise to a corresponding isotony theorem which gives a *numerical* expression to the relative merits of routes for calls that are implicit in the purely *combinatorial* monotone property. The relevance of these isotony theorems to optimal routing is explained heuristically in Section XVI. The theory culminates, in Section XVIII, with two optimal routing theorems based on the monotone properties. When one of these properties obtains, these results completely answer the question: Which route should be used for an accepted call when there is a choice of routes? Determining the extent to which these combinatorial properties occur in networks of interest appears to be the next major problem in any continuation of the present study.

It is to be stressed that the monotone properties we introduce serve only to identify the route that a call should take *if it is to be accepted;* they do not in any way help to decide which calls should be accepted. Except for the low-traffic results of Section X, and the (obvious and easily proved) fact that in a nonblocking network no call should be rejected, the problem of acceptance or rejection of calls remains an enigma. Some light on it is shed by the numerical results that immediately follow this summary.

The paper concludes in Appendix A with the remark that if the performance index is modified so as to put greater emphasis on "early blocked attempts", i.e., ones occurring soon after the system is started, then no calls should be rejected. The result is proved in detail for this index: the expected number of events until the first blocked attempt. Such a criterion corresponds to trying to avoid the undesirable event, the blocked call, as long as possible.

We turn now to numerical results obtained by solving the linear programming formulation of Section XI for two simple networks. The first is the three-stage Clos network with 2×2 switches depicted in Figs. 1 and 2, and already considered as an illustration of routing in Refs. 1 and 2. The second is a 6-line to 4-trunk concentrator in which each line has access to 2 trunks; it is shown in Figs. 3 and 4. In this second case, the probabilistic model was modified to make $\lambda > 0$ the calling-rate per idle line, rather than that per idle inlet-outlet pair.

In each example, both the minimal probability of blocking, and the probability of blocking under random routing, were calculated for several values of λ by use of the LP90 program. To be more precise, two linear programming problems were solved for each example; the first determined the optimal policy, the second determined the optimal policy among those policies which assigned random routes to accepted calls.

Several important qualitative features of the optimal routing policy were the same in both examples and are described together in the following list:

(*i*) The optimal policy rejected no calls.

(*ii*) The routes assigned by the optimal policy coincided with those that keep $s(\cdot)$ as large as possible.

(*iii*) The optimal policy was the same for all values of the traffic parameter λ examined.

(*iv*) The improvement over random routing brought about by optimal routing decreases as the traffic λ increases.

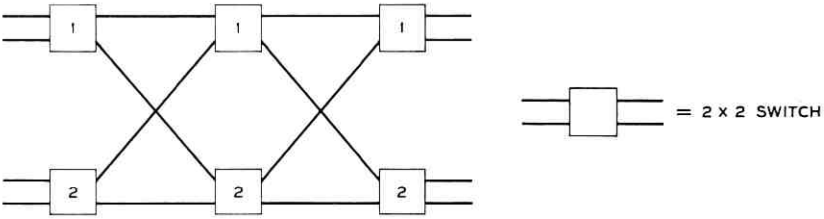


Fig. 1—3-stage Clos network with 2×2 switches.

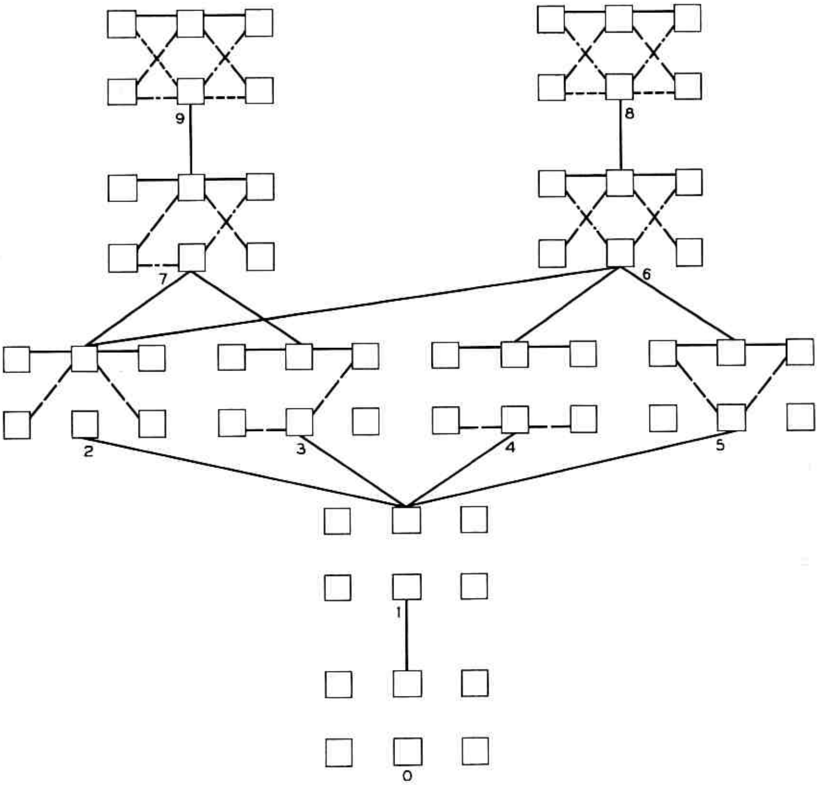


Fig. 2 — States of 3-stage Clos network of Fig. 1.

Under the constraint that accepted calls be routed at random the optimal policy was again to accept all unblocked attempted calls.

Results for the Clos network are given in Fig. 5 and Table I. It is apparent that for low λ optimum routing gives a loss that is easily an *order of magnitude* less than that due to random routing. At high values of λ the difference all but disappears. This behavior is explained in part by the fact that there is no blocking in the "upper" states of Fig. 2; when λ is very large the system spends all its time in these states; when λ is low, however, the occasion for a choice between states 2 and 4 often arises and a correct choice makes a significant difference. (At *very low* values of λ the difference will again decrease because only state 1 will ever be visited with any frequency.)

Results for the concentrator are shown in Fig. 6 and Table II. They include a numerical comparison with hand-calculated loss figures from unpublished work of S. P. Lloyd dated *circa* 1953. At that time Lloyd studied this particular concentrator model, correctly guessed the optimal policy, proved its optimality for low λ, and calculated the loss for some values of λ. This example exhibits the behavior, conjectured in Ref. 2, p. 275, that a good (here, optimal) policy make certain "bad" states *transient* states. The state numbered 9 is such a transient state under the optimal policy found numerically by the linear programming method.

The present study of routing in telephone networks has suggested a number of conclusions and conjectures:

(*i*) The problem of optimal routing of calls in telephone connecting networks (with full information) can be formulated and solved with Erlang's classical theory of traffic. In this endeavor, the contrasting techniques of machine calculation and combinatorial analysis can be employed either as alternative methods or as complementary approaches.

(*ii*) The problem separates into two parts, that of deciding which calls to accept, and that of choosing routes for accepted calls. Analytically, the first part appears to be much harder than the second, which frequently has a simple intuitive solution closely related to the structure of the network.

(*iii*) Posed within Erlang's theory, the routing problem can be reduced to a (usually very large) linear programming problem and attacked numerically, or studied in terms of Markov decision processes and dynamic programming.

(*iv*) In an apparently wide class of connecting networks, certain natural monotone properties and some isotonies based on them

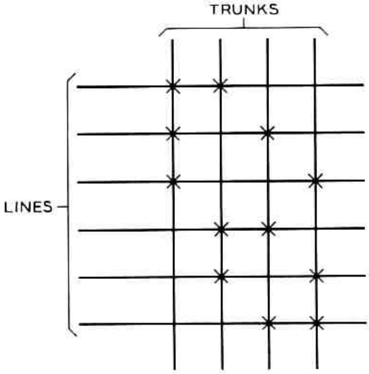


Fig. 3 — 6-to-4, 2 access concentrator.

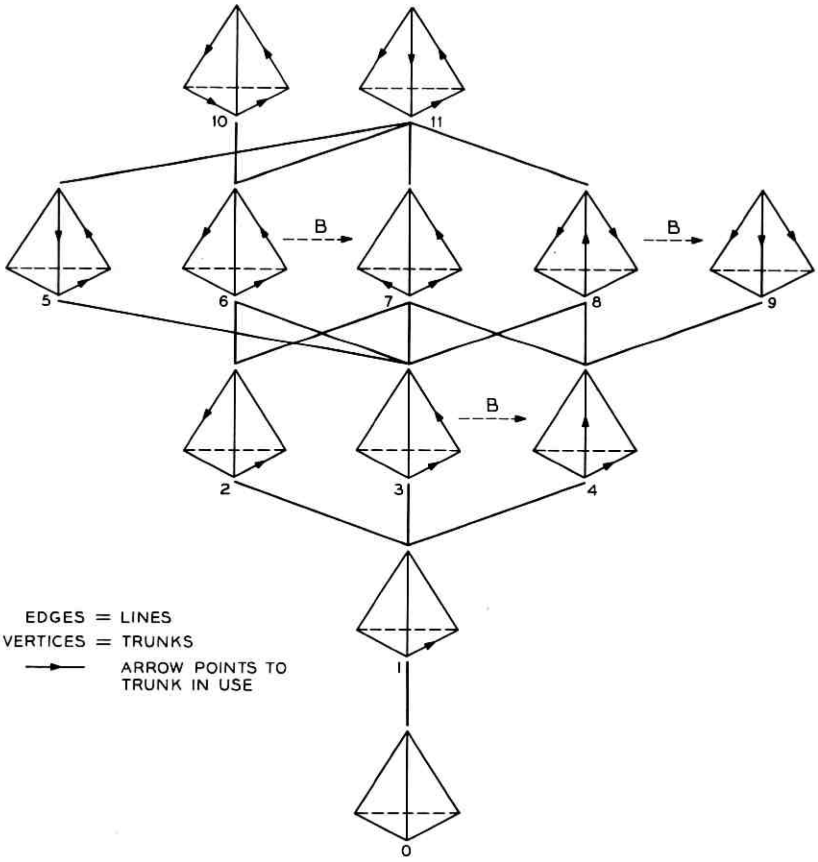


Fig. 4 — States of 6-to-4, 2 access concentrator.

are the key to choosing optimal routes for accepted calls. The resulting optimal policies are remarkably easy to describe and to instrument; they agree fully with some of the conjectures developed over years of practical experience in telephony; they are even robust under changes of performance index. Naturally, each example studied here involves a very small network. Nevertheless, the fact that the monotone properties turned up in each of a substantial number of small networks of diverse structure suggests that they are also present in larger ones. Whether this is so is a topic for future research. In any case, the examples we offer indicate that the theory of routing here developed applies equally well to central office networks and to various gradings and concentrators.

(*v*) In the interesting area of low traffic, optimal routing can be as much as an order of magnitude better than random routing; with high traffic the advantage decreases rapidly. In all the examples studied, the optimal routing policy was independent of the traffic λ; this suggests that in most cases the optimal policy is basically a combinatorial feature of the network alone, and is probably optimal in many probabilistic models of network operation.

(*vi*) There are situations in which attempted calls should be rejected even though they are not blocked. Simple examples of this phenomenon all seem to be rather unnatural; but J. H. Weber[3] has discovered it numerically in trunking networks, and has suggested[4] that it is associated with unequal lengths of paths for calls. The examples we studied numerically in the present work did not show it; but they had the property that all paths for calls were of the same length. We conjecture that there is a large class of "regular, well-behaved, normal, etc." networks in which no optimal policy rejects an unblocked call, and that in general occasions on which such calls should be rejected are rare. Even if they occur in practical central office networks, these occasions

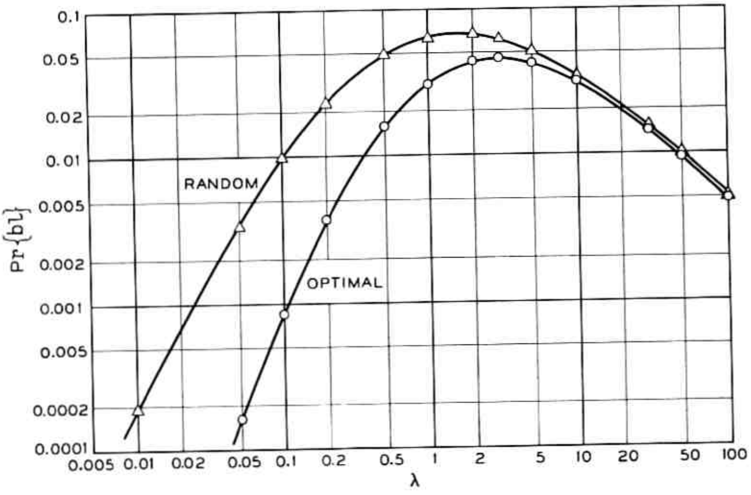


Fig. 5 — $Pr\{bl\}$ for Clos 3-stage network with 2 × 2 switches.

Table I — Probability of Blocking for Clos 3-Stage 2 × 2 Network for Optimal and Random Routing

λ	$Pr\{bl\}$	
	Optimal	Random
0.01	0.00000181	0.00018319
0.05	0.00015926	0.00334468
0.1	0.00087324	0.00960844
0.2	0.00376107	0.02259477
0.5	0.01593861	0.04807122
1.0	0.03146853	0.06360424
2.0	0.04381783	0.06670098
3.0	0.04584041	0.06206897
5.0	0.04233249	0.05152606
10.0	0.03115608	0.03463135
30.0	0.01405820	0.01459520
50.0	0.00901346	0.00922144
100.0	0.00475109	0.00480733

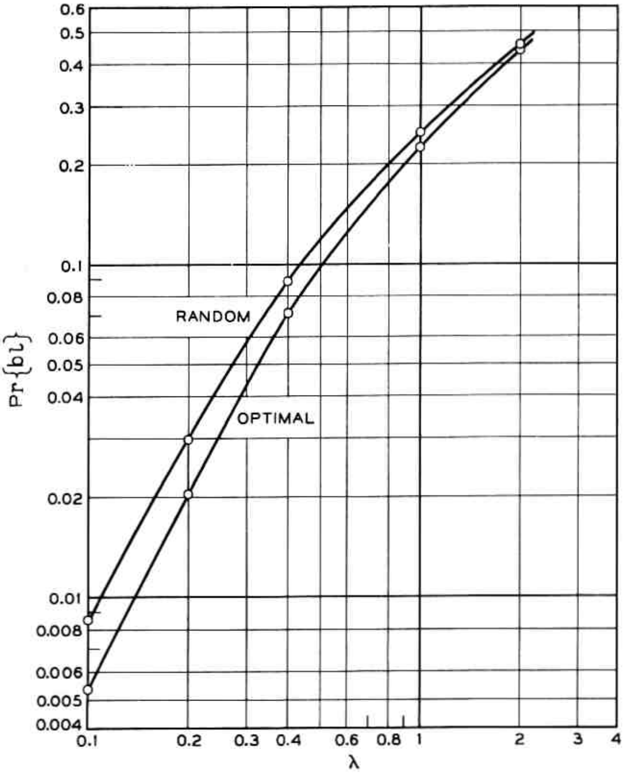


Fig. 6 — $Pr\{bl\}$ for 6-to-4, 2 access concentrator for random and optimal routing.

TABLE II — PROBABILITY OF BLOCKING FOR 6-TO-4, 2 ACCESS CONCENTRATOR FOR OPTIMAL AND RANDOM ROUTING

λ	Optimal		Random
	(S.P. Lloyd)	(Author)	
0.1	0.0049	0.00536231	0.00864729
0.2		0.02093718	0.02972292
0.4	0.0716	0.07170622	0.08856109
0.7	0.1628		
1.0	0.2478	0.23154056	0.24943320
2.0	0.4498	0.44971622	0.46141067

probably should be taken seriously (by a company committed to giving service) only if they are demonstrably associated with large amounts of congestion or a near-breakdown in operation. Hence, finding optimal policies to within rejection of calls may be considered a "practical" solution of the routing problem originally posed.

X. SOME COMPARISON THEOREMS FOR LOW TRAFFIC

There are two ways in which a theoretical analysis can substantially further progress in the problem of routing: (*i*) by means of *local* comparison theorems that establish that one method of routing is better than another, and (*ii*) by means of *global* optimality theorems that exhibit (in part or overall) one or more optimal policies which actually achieve the best possible value of the performance index in use. In this section, we prove some comparison theorems which are valid asymptotically as the traffic parameter approaches zero. At first, we restrict the analysis of the present section to the case[1,2] in which no unblocked call is rejected if it is attempted, so that we avoid the difficult question of deciding whether an attempted call that is not blocked should be completed or rejected.

For a first glimmer of insight, we shall examine the formula

$$Pr\{bl\} = \frac{p'\beta}{p'\alpha}, \qquad Q = Q(R), \qquad R \text{ a fixed rule}$$

valid when no unblocked call is rejected, in the very common situation in which there is an integer greater than zero, n say, such that there is no blocking in states with fewer than n calls in progress, and there are states with n calls in progress in which some calls are blocked. In this case it is known[1,2] that

$$\begin{aligned} p'\beta &= p_0 \frac{\lambda^n}{n!} \sum_{|x|=n} r_x \beta_x + o(\lambda^n), \\ p'\alpha &= p_0 \sum_{k=0}^{n} \frac{\lambda^k}{k!} \prod_{j=0}^{k-1} \alpha_j + o(\lambda^n), \end{aligned} \tag{3}$$

as $\lambda \to 0$, where[1,2]

$$\begin{aligned} r_x &= \text{number of paths on } S \text{ ascending from 0 to } x \text{ and permitted by } R \\ &= (R^{|x|})_{0x} \\ &= \text{the } 0,x \text{ entry of the } |x|\text{-th power of } R, \end{aligned} \tag{4}$$

and

α_j = number of idle inlet-outlet pairs in a state having j calls in progress.

(We recall that for the important cases of one- or two-sided networks $\alpha_x = \alpha_{|x|} = \alpha_j$ for all x with $|x| = j$.) It follows from (3) that for small λ the leading term is critical: the blocking will depend principally on how easy it is to reach a blocking state from the zero state, with this "ease" measured by the number

$$\sum_{|x|=n} r_x \beta_x = (R^n \beta)_0$$

= the number of ways in which a blocked call can arise without having any hangups, starting at zero.

If the matrix R is not fixed, but allows some random choices of route, then this quantity can still be viewed as the "expected number of ways in which a blocked call can arise without having any hangups, starting at zero." It is apparent that this number is given by f_0, where the numbers $\{f_x, |x| \leqq n\}$ are defined by the nonlinear recurrence

$$f_x = \begin{cases} \beta_x & |x| = n \\ \displaystyle\sum_{\substack{c \in x \\ c \text{ not blocked in } x}} \min_{y \in A_{cx}} f_y & |x| < n. \end{cases}$$

Indeed we have the result:

Lemma 1:

$$\sum_{|x|=n} r_x \beta_x \geqq f_0 \quad \textit{for} \quad R \in C$$

Proof: Let R be given and let

$$d_x = \begin{cases} \beta_x & |x| = n \\ \sum_{y \in A_x} r_{xy} d_y & |x| < n. \end{cases} \tag{5}$$

We prove the stronger result that $d_x \geqq f_x$. It is clear that

$$d_0 = \sum_{|x|=n} r_x \beta_x, \qquad d_x = f_x \quad \text{for} \quad |x| = n.$$

If $d_y \geqq f_y$ for $|y| = k + 1$, then for $|x| = k$

$$\begin{aligned} d_x = \sum_{y \epsilon A_x} r_{xy} d_y &\geqq \sum_{y \epsilon A_x} r_{xy} f_y \\ &\geqq \sum_{\substack{c \text{ idle in } x \\ c \text{ not blocked in } x}} \min_{y \epsilon A_{cy}} f_y = f_x . \end{aligned}$$

We shall say that $R \in C$ puts $x \in S$ on an ascending path to a state z if and only if $\exists y_0, \cdots, y_{|z|}$ with $y_0 = 0$, $|y_i| = i$, $y_{|z|} = z$, and $r_{y_i y_{i+1}} = 1$ for $i = 0, \cdots, |z| - 1$, and $x > 0$ is among $y_1, \cdots, y_{|z|}$. Let D be the subset of all fixed rules $R \in C$ such that if $|z| = n$, and if R puts x,y with $y \in A_x$ on an ascending path to z, then $r_{xy} = 1$ only if, with $c = \gamma(y - x)$,

$$f_y = \min_{w \varepsilon A_{cx}} f_w .$$

The numbers $\{f_x, |x| \leqq n\}$ are the key to optimal routing for low values of λ, or to put it more picturesquely, they are the key to staying as far away as possible from the blocking states in $\{x: \ |x| = n\}$, which are the ones that provide the leading term in $Pr\{bl\}$ as $\lambda \to 0$. We have

Theorem 1: Let $R \in D$ and $R^* \in C - D$. Then for all λ small enough

$$Pr\{bl\}_R < Pr\{bl\}_{R^*} .$$

Proof: Let d_x^* be defined in terms of R^* according to (5) used in Lemma 1. Since $R^* \notin D$, there exist x,y,c, and $\varepsilon > 0$, such that

$$\begin{aligned} y \in A_x, \quad \gamma(y - x) = c, \quad r_{xy}^* = 1 \\ f_y \geqq \min_{z \in A_{cx}} f_z + \varepsilon, \end{aligned} \tag{6}$$

and a maximal chain $0 = y_0, y_1, y_2, \cdots, y_{|x|-1}, y_{|x|} = x$ ascending in $\leqq$ such that

$$r_{y_i y_{i+1}}^* = 1. \qquad i = 0, \cdots, |x| - 1.$$

Now, using $d^* \geqq f$,

$$\begin{aligned} d_x{}^* &= \sum_{z \in A_x} r_{xz}{}^* d_z{}^* \\ &= \sum_{A_x - \{y\}} r_{xz}{}^* d_z{}^* + f_y \\ &\geqq f_x + \varepsilon, \end{aligned}$$

the last inequality a consequence of $d^* \geqq f$ and the definition of f. Similarly, if $d_{y_{i+1}}{}^* > f_{y_{i+1}}$, then

$$\begin{aligned} d_{y_i}{}^* &= \sum_{A_{y_i} - \{y_{i+1}\}} r_{y_i z}{}^* d_z{}^* + d_{y_{i+1}}{}^* \\ &> f_{y_i}{}^*. \end{aligned}$$

Since $y_0 = 0$, we have $d_0{}^* > f_0$.

Setting $a = f_0$, $a^* = d_0{}^*$, and

$$b = \sum_{k=0}^{n} \frac{\lambda^{k-n}}{k!} \prod_{j=0}^{k-1} \alpha_j ,$$

we have the asymptotic forms

$$Pr\{bl\}_R = \frac{a + \varepsilon}{b + \delta}$$

$$Pr\{bl\}_{R^*} = \frac{a^* + \varepsilon^*}{b + \delta^*}$$

with $\varepsilon,\delta,\varepsilon^*,\delta^*$ all $o(1)$ as $\lambda \to 0$, and $a < a^*$. Since b increases as $\lambda \to 0$

$$(a - a^* + \varepsilon - \varepsilon^*)b < a^*\delta - a\delta^* + \varepsilon^*\delta - \varepsilon\delta^*$$

for all λ small enough. This is equivalent to

$$ab + \varepsilon b + a\delta^* + \varepsilon\delta^* < a^*b + \varepsilon^*b + a^*\delta + \varepsilon^*\delta,$$

$$\frac{a + \varepsilon}{b + \delta} < \frac{a^* + \varepsilon^*}{b + \delta^*},$$

and proves the theorem.

Low traffic analyses of the kind just employed can also shed some light on the problem of rejecting or accepting unblocked calls. For example, if a call c is *always* refused in every state, then

$$r_{xx} \geqq 1$$

and

$$Pr\{bl\} = \frac{p'(\beta + r)}{p'\alpha} \geqq \frac{1 + p'\beta}{p'\alpha}$$

$$\rightarrow \frac{1}{\alpha_0} \quad \text{as} \quad \lambda \rightarrow 0.$$

However, if no unblocked call is rejected, then $Pr\{bl\} \rightarrow 0$ as $\lambda \rightarrow 0$. Thus, *always refusing* c cannot be optimal if λ is sufficiently small.

For another example, suppose as before that

$$n = \min_{y \in S} \{|y|: \ \beta_y > 0\} > 0,$$

and let c be a call which is refused by R in some state x with $|x| < n$. It is easy to see that for the rule R

$$Pr\{bl\} \geqq \frac{\frac{\lambda^{|x|}}{|x|!} r_x + o(\lambda^{|x|})}{\alpha_0 + o(\lambda)}.$$

On the other hand, if the rule R_1 refuses no unblocked calls,

$$Pr\{bl\} = \frac{\frac{\lambda^n}{n!} \sum_{|y|=n} r_y^{(1)} \beta_y + o(\lambda^n)}{\alpha_0 + o(\lambda)},$$

where the superscript 1 indicates that R_1 supplants R in (4). For λ small enough, then

$$\frac{\lambda^{|x|}}{|x|!} r_x > \frac{\lambda^n}{n!} \sum_{|y|=n} r_y^{(1)} \beta_y$$

and R_1 is better than R. Thus, there is never any point in refusing an unblocked call attempt made in a state x whose norm or dimension is less than the minimum norm achieved by the blocking states, if λ is small enough.

XI. REDUCTIONS TO LINEAR PROGRAMMING PROBLEMS

Our effort to choose, with full information about the state of the network, routes for new calls so as to minimize the probability of blocking has led, upon the assumption of a simple probabilistic description for the traffic, to this problem of mathematical programming: To minimize

$$\frac{p'(\beta + r)}{p'\alpha} \tag{7}$$

subject to $p \geqq 0$, $p'1 = 1$, $p'Q = 0$, $Q = Q(R)$, $R \in C$.

It is relatively easy to see that this problem can be formulated as one that has a bilinear (or linear fractional) objective function, and linear constraints. We change variables to $U = (u_{xy})$ and u_{cx} defined by

$$\begin{aligned} u_{xy} &= p_x r_{xy} && x,y \in S, \quad y \in A_x \\ u_{cx} &= p_x - \sum_{y \in A_{cx}} u_{xy} && c \in x, \quad A_{cx} \neq \theta, \\ u_{xx} &= \sum_{\substack{c \in x \\ c \text{ not blocked in } x}} && u_{cx}\,. \end{aligned}$$

Conversely, we introduce p in terms of U by setting

$$p_x = \begin{cases} \dfrac{\lambda}{|x|} \displaystyle\sum_{y \in B_x} u_{yx} & \text{if} \quad s(x) = 0, \\[2ex] \dfrac{u_{xx} + \displaystyle\sum_{y \in A_x} u_{xy}}{s(x)} & \text{if} \quad s(x) > 0. \end{cases}$$

If c is a call which can be completed in state x, then $A_{cx} \neq \theta$, and

$$\lambda \sum_{y \in A_{cx}} u_{xy}$$

is the equilibrium rate at which c is completed in state x, and

$$\lambda u_{cx} = \lambda p_x - \lambda \sum_{y \in A_{cx}} u_{xy}$$

is the equilibrium rate at which c is rejected in state x.

The transformation of variables from p to $\{U, u_{cx}\}$ necessitates adding additional constraints if a sensible problem is to result. Evidently, for $c \in x$ not blocked in x

$$p_x = u_{cx} + \sum_{y \in A_{cx}} u_{xy}\,.$$

The left-hand side does not depend on c. For different $c \in x$ not blocked in x all these formally different ways of calculating p_x must agree, and it is, therefore, necessary to impose the additional constraint that

$$c,c' \in \gamma(A_{cx} - x) = \gamma\{y\colon \quad y = z - x \quad \text{for} \quad z \in A_{cx}\} \text{ implies}$$
$$u_{cx} + \sum_{y \in A_{cx}} u_{xy} = u_{c'x} + \sum_{y \in A_{c'x}} u_{xy}\,.$$

The condition $p'Q = 0$ then gives the condition, for $s(x) > 0$,

$$\frac{|x|}{s(x)} \Big(u_{xx} + \sum_{y \in A_x} u_{xy} \Big) + \sum_{y \in Ax} u_{xy} = \sum_{y \in A_x} \Big(u_{yy} + \sum_{z \in A_y} u_{yz} \Big) + \lambda \sum_{y \in B_x} u_{yx}$$

to be satisfied by U. Naturally, the condition $U \geqq 0$ is imposed. We define R in terms of U by

$$r_{xy} = \begin{cases} 0 & \text{unless} \quad y \in A_x \quad \text{or} \quad y = x, \\ \dfrac{u_{xy}}{u_{xx} + \sum\limits_{z \in A_x} u_{yz}} & \text{if} \quad y \in A_x, \\ s(x) - \sum\limits_{y \in A_x} r_{xy} & \text{if} \quad y = x. \end{cases}$$

The normalization condition $p'1 = 1$, finally, amounts in terms of U to

$$\lambda \sum_{s(x)=0} \frac{1}{|x|} \sum_{y \in B_x} u_{yx} + \sum_{s(x)>0} \frac{\left(u_{xx} + \sum\limits_{y \in A_x} u_{xy}\right)}{s(x)} = 1.$$

In terms of U the objective function is

$$\frac{\lambda \sum\limits_{s(x)=0} \dfrac{\beta_x}{|x|} \sum\limits_{y \in B_x} u_{yx} + \sum\limits_{s(x)>0} \dfrac{\beta_x}{s(x)} \left(u_{xx} + \sum\limits_{y \in A_x} u_{xy}\right) + u_{xx}}{\lambda \sum\limits_{s(x)=0} \dfrac{\alpha_x}{|x|} \sum\limits_{y \in B_x} u_{yx} + \sum\limits_{s(x)>0} \dfrac{\alpha_x}{s(x)} \left(u_{xx} + \sum\limits_{y \in A_x} u_{xy}\right)}.$$

It is possible to describe *linear* programming problems which are equivalent to our nonlinear problem of optimal routing. Two ways of reducing (7) to a linear programming problem will now be discussed. The first is due to A. Charnes and W. W. Cooper.[8] Let $q = tp$, where the scalar $t \geqq 0$ is to be chosen so that $q'\alpha = a$, with $a > 0$ a specified real number. Consider now the "adjoined" *linear* programming problem of finding q,t,r minimizing $q'(\beta + r)$, subject to $q,t \geqq 0$, $q'1 - t = 0$, $q'Q = 0$, $q'\alpha = a$, $Q = Q(R)$, $r = r(R) = \{r_{xx}, x \in S\}$, $R \in C$. (The argument just described shows that the constraints are linear.)

Theorem 2: For any $a > 0$, if q,t,r is a solution of the "adjoined" linear problem, then $p = q/t$ is a solution of (7).

Proof: It is necessary to show first that indeed $t > 0$. Suppose $q,0,r$ is a solution. Then $q'1 = 0$ and $q \geqq 0$ imply $q = 0$, so that $q'\alpha = 0$; but $q'\alpha = a > 0$. Hence, $t > 0$.

If $p'Q = 0$ and $Q = Q(R)$, we use r_p to mean the vector $\{r_{xx}, x \in S\}$. Now suppose that there is a solution p of (7) for which

$$\frac{p'(\beta + r_p)}{p'\alpha} > \frac{q'(\beta + r)}{q'\alpha} = \frac{q'(\beta + r)}{a}. \tag{8}$$

Now $p'\alpha > 0$, because for any $R \in C$ the corresponding value of p_0

(0 = zero state, with no calls up) is > 0, and $\alpha_0 > 0$. Hence, there is a $\theta > 0$ such that $p'\alpha = \theta a$. Consider $\hat{q} = \theta^{-1}p$, $\hat{t} = \theta^{-1}$. Then

$$\theta^{-1}p'\alpha = \hat{q}'\alpha = a$$

and $\hat{q},\hat{t}$ satisfy $\hat{q},\hat{t} \geqq 0$, $\hat{q}'Q = 0$, $\hat{q}'1 - \hat{t} = 0$. But,

$$\frac{p'(\beta + r_p)}{p'\alpha} = \frac{\theta^{-1}p'(\beta + r_p)}{\theta^{-1}p'\alpha} = \frac{\hat{q}'(\beta + r)}{\hat{q}'\alpha} = \frac{\hat{q}'(\beta + r)}{a}.$$

Hence, (8) implies $\hat{q}'(\beta + r_p) > q'(\beta + r)$, because $a > 0$. This contradicts the optimality of q,t,r for the "adjoined" problem.

A cognate reduction to a linear programming problem can be obtained from a lemma of C. Derman,[9] included for completeness:

Lemma 2: The nonlinear function

$$g(x) = \frac{c'x}{d'x}$$

can be minimized subject to $x \geqq 0$, $Ax = b$, *by solving a linear programming problem if* (i) $Ax = 0$, $x \geqq 0$ *imply* $x = 0$ *and* (ii) $x \geqq 0$, $Ax = b$ *imply* $d'x > 0$.

Proof: Conditions (i) and (ii) imply that the transformation

$$z = \begin{pmatrix} \dfrac{x}{d'x} \\ \dfrac{1}{d'x} \end{pmatrix}$$

is one-to-one between $\{x \geqq 0, Ax = b\}$ and z satisfying $z \geqq 0$, $d'z = 1$. and $Bz = 0$, where

$$B = (Ab).$$

Under the transformation $g(x)$ becomes a linear function. It can be verified that (i) and (ii) of Derman's lemma apply to the routing problem (7).

XII. REFORMULATION AS A MARKOV DECISION PROCESS

In Section VII the problem of optimal routing was cast as that of minimizing the probability of blocking, a *bilinear* or *linear fractional* functional of the equilibrium probability vector p, subject to linear constraints. In Section XI it was shown how this problem could be reduced to a linear programming problem which, however, is at best only suggestive in identifying optimal policies. We shall now state an elementary probabilistic result which implies that minimizing the probability of blocking, and maximizing the fraction of events that are successful attempts, are equivalent. This fact permits a direct dynamic programming approach through Markov decision processes, and again leads to a linear programming problem, with the difference, though, that it actually enables us to study optimal policies for many cases, to be described.

Theorem 3: Let p be an equilibrium probability vector for a process x_t resulting from use of some rule $R \in C$. Let

$$m = \sum_{x \in S} |x| p_x = \textit{average number of calls in progress}$$

then both

$$1 - Pr\{bl\} = \frac{1}{1 + \lambda \dfrac{p'(\beta + r)}{m}},$$

and

$$\begin{matrix}\textit{Fraction of events that} \\ \textit{are successful attempts}\end{matrix} = \frac{1}{2 + \lambda \dfrac{p'(\beta + r)}{m}}.$$

Proof: For the first formula with $s = \{s(x), x \in S\}$

$$Pr\{bl\} = \frac{p'(\beta + r)}{p'\alpha} = \frac{p'(\beta + r)}{p'(s - r) + p'(\beta + r)}$$

and $\lambda p'(s - r) = m$, since the average rate of successes must equal that of hangups, in equilibrium, and $\alpha = \beta + s$.

The second quantity is

$$\frac{\text{average rate of successes}}{\text{average rate of events}} = \frac{\lambda p'(s - r)}{m + \lambda p'\alpha} = \frac{m}{2m + \lambda p'(\beta + r)}.$$

An immediate consequence is:

Theorem 4: Maximizing the fraction of events that are successful attempts is equivalent to minimizing the probability of blocking.

The value of the preceding observations is that we can now reformulate the routing problem as an effort to maximize

$$\lim_{n \to \infty} \frac{1}{n} E\{\text{number of successful attempts in } n \text{ events}\},$$

the asymptotic rate of successful attempts when time is counted discretely, by events.

Since only events are at issue, and the epochs at which they occur are irrelevant, we can discard the continuous parameter Markov process $\{x_t, t \text{ real}\}$ in favor of a Markov chain $\{x_n, n \text{ an integer}\}$, with a transition matrix $A = (a_{xy}) = A(R)$ given by

$$[\,|x| + \lambda\alpha_x]a_{xy} = \begin{cases} \lambda(\beta_x + r_{xx}) & x = y, \\ 1 & y \in B_x, \\ \lambda r_{xy} & y \in A_x, \\ 0 & \text{otherwise.} \end{cases}$$

The stationary vector q satisfying $q = q'A$ is related to p by

$$p_x = (\text{constant}) \frac{q_x}{|x| + \lambda\alpha_x}.$$

Then

$$E\{\text{number of successful attempts in } n \text{ events}\} = \sum_{j=0}^{n-1} A^j v$$

where $A = A(R)$ and $v = v(R)$ given by

$$\begin{aligned} v_x &= \frac{\lambda s(x) - \lambda r_{xx}}{|x| + \lambda\alpha_x}, \\ &= \text{chance that first event to occur} \\ &\quad\ \text{starting in } x \text{ is a successful call.} \end{aligned} \tag{9}$$

Thus, the problem of optimal routing can be cast in the form of the Markov decision processes studied by e.g., R. Bellman[10] and R. Howard:[11] For $R \in C$ and $A = A(R) = (a_{xy})$ given by

$$(\,|x| + \lambda\alpha_x)a_{xy} = \begin{cases} \lambda(\beta_x + r_{xx}) & x = y, \\ 1 & y \in B_x, \\ \lambda r_{xy} & y \in A_x, \\ 0 & \text{otherwise,} \end{cases}$$

the minimum

$$\min_{R\in C} Pr\{bl\} = \min_{R\in C} \frac{p'(\beta + r)}{p'\alpha}$$

subject to $p'Q = 0$, $p \geqq 0$, $p'1 = 1$ is achieved by the R which maximizes the scalar ρ such that

$$\rho 1 = \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} A^j v \quad v = v(R), \quad A = A(R), \quad R \in C$$

with v given by (9).

The results of Bellman in Ref. 10 were derived under the strong positivity condition $a_{xy} \geqq d > 0$ on the matrices A; this condition is of course not met in our routing problem, since many a_{xy} necessarily vanish. However, since our matrices have only one ergodic set it is still possible to obtain results like Bellman's provided only that a little care is taken with the transient states.

Lemma 3: Let ρ be the scalar defined by

$$\rho 1 = \max_{R\in C} \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} A^j v, \tag{10}$$

let R^+ achieve the maximum in (10), and let g be the vector determined[11] *up to a multiple of (the vector) 1 by the equation*

$$\rho 1 + g = v(R^+) + A(R^+)g.$$

Let R^* achieve the maximum in

$$\max_{R\in C} \{v(R) + A(R)g\}$$

Let F be the transient set of states relative to $A(R^*)$. Then the restriction of g to $S - F$ satisfies the nonlinear equation

$$\rho + g_x = \max_{R\in C} \{v_x(R) + \sum a_{xy}(R)g_y\}, \qquad x \in S - F, \tag{11}$$

and the right-hand side of (11) depends in fact only on

$$\{g_y, y \in S - F\}.$$

Further, there is a fixed routing matrix R^{**}, agreeing with R^* on $(S - F)^2$, and a vector g^* agreeing with g on $S - F$, such that R^{**} achieves the maximum in

$$\rho 1 + g^* = \max_{R\in C} \{v(R) + A(R)g^*\}.$$

Proof: If the nonlinear equation given does not hold for some $x \in S - F$.

there exists a vector ζ with $\zeta \neq 0$, $\zeta \geqq 0$ such that on $S - F$

$$\rho + \zeta_x + g_x = \max_{R \in C} \{v_x(R) + \sum_y a_{xy}(R) g_y\},$$

$$= v_x(R^*) + \sum_y a_{xy}(R^*) g_y .$$

Let us restrict all vectors to the $| S - F |$ components present in $S - F$, and the matrix $A\,(R^*)$ to $(S - F)^2$. Then, dropping dependence on R^*

$$\rho 1 + \zeta + g = v + Ag.$$

There exists an integer k such that $A^k > 0$ strictly. Left-multiply by A^k and note that $A1 = 1$ to obtain

$$\rho 1 + A^k(\zeta + g) = A^k v + A^{k+1} g.$$

Since A^k is a positive matrix, and $\zeta \neq 0$, $\zeta \geqq 0$, there exists a scalar ε such that $A^k \zeta \geqq \varepsilon 1$, so that

$$(\rho + \varepsilon)1 + A^k g \leqq A^k v + A^{k+1} g.$$

Iterating this inequality n times we obtain

$$n(\rho + \varepsilon)1 + A^k g \leqq \sum_{i=k}^{k+n-1} A^i v + A^{k+n} g.$$

For n large enough this contradicts the maximal character of ρ. To find R^{**} and g^*, consider the equation

$$g_x^* = -\rho + \max_{R \in C} \{v_x(R) + \sum_{y \in F} a_{xy}(R) g_y^* + \sum_{y \in S-F} a_{xy}(R) g_y\}, \quad x \in F.$$

This represents the expected best possible fortune of a gambler who starts broke in state $x \in F$, plays by choosing a matrix R paying an amount ρ to play, receiving $v_x(R)$ if he plays R in state x, and ending the game with a final payoff of g_y if the system leaves F for the first time by going into $y \in S - F$; i.e., if he passes through $x_1 x_2 \cdots x_n y$ playing $R_1 R_2 \cdots R_n$ (with R_i in x_i), going out to $y \in S - F$ from x_n, then he receives (or owes)

$$-n\rho + \sum_{i=1}^{n} v_{x_i}(R_i) + g_y .$$

It is apparent that $\{g_x^*, x \in F\}$ exist; R^{**} on $F^2 \cup (F \times S - F)$ is determined by the property that it achieves the maximum above, and on $(S - F) \times F$ it is zero.

Lemma 4: Let ρ be the scalar defined by the condition

$$\rho 1 = \max_{R \in C} \lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} A^j v, \tag{12}$$

and let the vector g be a solution of the nonlinear inequality

$$\rho 1 + g \leqq \max_{R \in C} \{v(R) + A(R)g\}. \tag{13}$$

If $R^* \in C$ achieves the maximum on the right of (13), then it also achieves that on the right of (12).

Proof: R^* and g are related by

$$\rho 1 + g \leqq v(R^*) + A(R^*)g,$$

whence, left-multiplying by $A^j = A^j(R^*)$ and summing on j from 0 to $(n - 1)$,

$$n\rho 1 + \sum_{j=0}^{n-1} A^j g \leqq \sum_{j=0}^{n-1} A^j v + \sum_{j=1}^{n} A^j g$$

$$\rho 1 - \frac{1}{n} \sum_{j=0}^{n-1} A^j v \leqq o(1) \qquad A = A(R^*), \qquad v = v(R^*).$$

This implies that R^* achieves the maximum in (12).

XIII. OPTIMALITY OF FIXED RULES

If a routing matrix has any entries other than integers, its use introduces a certain amount of additional randomness into the operation of the network, over and above that due to the random traffic, and may be said to represent a "mixed" strategy. It is a natural intuition that since minimizing the probability of loss is a game played against nature, rather than against an intelligent adversary, there can be no real gain from this additional randomization, i.e., that a fixed rule can be found that is as good as any "mixed strategy". To this effect we formulate

Theorem 5: A fixed rule R achieves

$$min \frac{p'(\beta + r)}{p'\alpha}$$

subject to $R \in C$, $p'Q = 0$, $p'1 = 1$, $p \geqq 0$, $Q = Q(R)$.

This theorem is a consequence of the next two results, which, though they are adapted from work of C. Derman,[9] are included here for completeness.

Lemma 5: Let $\xi(\cdot)\colon C \to E^{|S|}$ be an affine map of C into $|S|$ - dimen-

sional Euclidean space, i.e., one such that for real scalars $a_1, a_2 \geqq 0$ *with* $a_1 + a_2 = 1$, *and* $R_1, R_2 \in C$,

$$\xi(a_1R_1 + a_2R_2) = a_1\xi(R_1) + a_2\xi(R_2),$$

and let ξ *be continuous. Then,*

$$\min q'\xi$$

subject to $q \geqq 0$, $q'1 = 1$, $q'A = q$, $A = A(R)$, $\xi = \xi(R)$ *is achieved by a fixed rule* R.

Proof: For $R \in C$ and $A = A(R)$, $\xi = \xi(R)$ set

$$v(R) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=0}^{n-1} A^j\xi.$$

By a known Markov chain limit theorem,[12] $v(R)$ is well-defined. For $\mu \in (0,1)$ let

$$V(R,\mu) = \sum_{j=0}^{\infty} (\mu A)^j\xi.$$

It is clear that for each $\mu \in (0,1)$, and each starting state x, there exists an $R_{\mu x} \in C$

$$V_x(R_{\mu x}, \mu) = \min_{R\in C} V_x(R,\mu).$$

Then

$$V_x(R_{\mu x}, \mu) = \min_{R\in C} \{\xi_x(R) + \mu \sum_{y\in S} a_{xy}(R)V_y(R_{\mu x}, \mu)\}.$$

The right-hand side is an affine functional of R and so assumes a minimum at an extreme point of C, i.e., at a fixed rule R. Thus, we can consider that $R_{\mu x}$ is a fixed rule. Since the fixed rules form a *finite* class, there exists a sequence $\mu_n \to 1$ and a fixed rule R^* such that

$$R_{\mu_n x} = R^* \qquad n = 1,2, \cdots .$$

By a well-known Abelian theorem,[13] for $R \in C$

$$\lim_{\mu\to 1} (1 - \mu)V(R,\mu) = v(R)$$

and also

$$\begin{aligned} v(R) &\geqq \lim_{n\to\infty} (1 - \mu_n)V(R,\mu_n) \\ &\geqq \lim_{n\to\infty} (1 - \mu_n)V(R^*,\mu_n) \\ &\geqq v(R^*). \end{aligned}$$

Thus, R^* is optimal.

Theorem 6: Let ζ, η: $C \rightarrow E^{|S|}$ *be affine maps of* C *into* $|S|$ *- dimensional Euclidean space, and let* ξ *and* η *be continuous, with* $\eta(R) > 0$ *for* $R \in C$. *Then*

$$b = min \frac{q'\zeta}{q'\eta}$$

subject to $q \geqq 0$, $q'A = q$, $q'1 = 1$, $A = A(R)$, $\xi = \xi(R)$, *and* $\eta = \eta(R)$ *is achieved by a fixed rule.*

Proof: Let $b(R)$ be the value of $q'\xi/q'\eta$ for a given choice R, with q determined by the constraints $q \geqq 0$, $q'A = q$, $q'1 = 1$. There exist $R_1, R_2, \cdots \in C$ such that

$$\lim_{n\to\infty} b(R_n) = b.$$

For n fixed, let $\xi(\cdot)$ in Lemma 5 be given by

$$\xi = \zeta - b(R_n)\eta.$$

Then in the notation of Lemma 5, $v(R_n) = 0$. By Lemma 5 there exists a fixed rule R_n^* such that

$$\begin{aligned} v(R_n^*) &\leqq v(R_n) \\ &\leqq 0, \end{aligned}$$

that is, since $q'\eta \neq 0$,

$$b(R_n^*) \leqq b(R_n).$$

Since there is a finite number of fixed rules, there is a subsequence $n_1, n_2, \cdots$ and a fixed rule R^* such that $R_{n_i}^* = R^*$, $i = 1, 2, \cdots$. Then R^* is optimal.

XIV. TRYING TO GET CLOSER TO THE OPTIMAL ROUTING RULES

It is particularly important to try to verbalize, and eventually to mechanize, routing strategies that are optimal, near-optimal, or by some yardstick just "good". In this endeavor, the fact that the original routing problem (7) can be formulated and solved numerically as a linear programming problem, while interesting theoretically and perhaps reassuring, is nevertheless of limited usefulness. For this reason we have attempted to take advantage of some of the special properties of the problem that are due to its telephonic origins, and to describe at least parts of optimal policies in terms of the combinatorial properties of the connecting network upon which they ultimately depend.

In the second half of this paper we introduce some additional notions and assumptions of a combinatorial nature. With their aid we are able to exhibit parts of some actual optimal routing rules. The problem of finding out something concrete about optimal policies has been so difficult that we have quite frankly started with (and so far restricted attention to) cases which can be treated by what T. M. Burford has called "domination" arguments, which depend on or establish isotony[5] properties for certain networks having suitable monotone structures. The word 'monotone' is used loosely here: more specifically, the networks are to have the property that the relative merit of states is consistent or continuous, i.e., that if one state x is "better" than an equivalent state y, then the *neighbors* of x are in the same sense "better" than the corresponding neighbors of y.

Although some of the combinatorial properties (on which the results to be given are based) are strong, we believe that these properties and the optimal policies (or partial policies) they lead to have a definite relevance to the practical aspects of optimal routing, if only because they bear out some of the intuitive conjectures offered in Section VIII. Our results show not only that these conjectures are "in the right ballpark," but also thet in many instances they describe optimal policies.

We start our discussion with four simple examples; once the ideas involved are understood, the principles behind them can be abstracted, and general theorems proved.

It has been shown (Section XII) that minimizing the probability of blocking is equivalent to maximizing the fraction of events that are successful attempts, where an event is either a hangup, a blocked attempt, or a successful one. This maximal fraction is the limit, as n becomes large, of

$$\frac{1}{n} E_x(n),$$

where

$E_x(n)$ = expected number of successful calls in n events, if the network starts in state x and an optimal policy is followed.†

We shall base our approach on the vectors $E(n)$.

First example: Consider the overflow system or grading shown in Figs. 7 and 8. There are two groups of lines, each of two lines; the first has access to both trunks to the destination, but the second has access to the second trunk only. The possible states of this system (reduced under the

† Here an optimal policy is one for which the expected number of successful calls in n steps is a maximum.

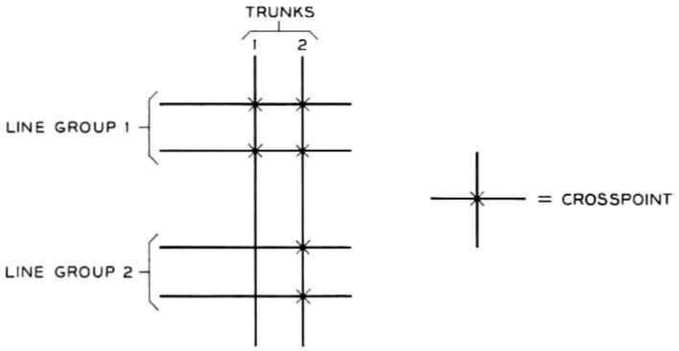


Fig. 7 — Asymmetric grading.

equivalence relation induced by permuting lines within a line group) form the partially ordered system of Fig. 8. There is only one situation which demands a choice between alternative routes for a call; it arises when a call from line group 1 is accepted with no calls in progress. The two alternatives are indicated in Fig. 8 by the notation "ch": one is to put the call on trunk 1, leaving no lines blocked, the other is to put it on trunk 2, leaving 2 lines blocked.

What circumstances make one choice of a route better than another? In the present instance it is clear that use of trunk 1 for a group 1 call in state 0 leaves the "high access" trunk 2 free to serve group 2. Thus, at first glance a route whose use blocked the smallest possible number of additional calls (over and above those that are already blocked) seems to be best. It is natural to expect that in state 0 a new call from group 1 should be routed on trunk 1 and not on trunk 2. Indeed, it can be shown that if such a call should be accepted then it should be placed on trunk 1. (For small λ it *should* always be accepted, as was proved in Section X.) Thus, a policy which routes a group 1 call on trunk 1 in state 0 can differ

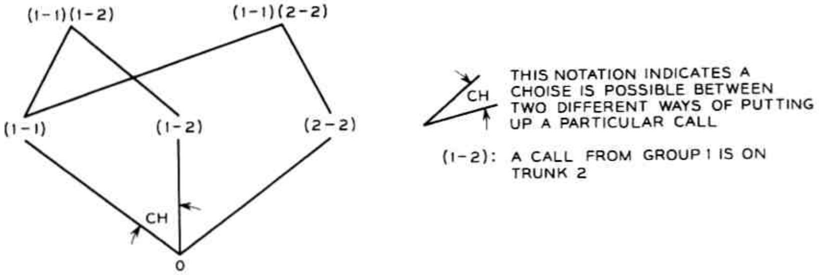

(1-1)(1-2)
(1-1)(2-2)
(1-1)
(1-2)
(2-2)
CH
0
CH
THIS NOTATION INDICATES A CHOISE IS POSSIBLE BETWEEN TWO DIFFERENT WAYS OF PUTTING UP A PARTICULAR CALL
(1-2): A CALL FROM GROUP 1 IS ON TRUNK 2

Fig. 8 — States of the grading of Fig. 7.

from an optimal policy only in that it might accept some calls which the other rejected, and *vice versa*.

Rather than proving the result stated above, we shall discuss other examples, involving different kinds of network: it will turn out that similar circumstances arise. Indeed, we shall claim that the particular circumstance on which the result is based is no isolated happenstance, but a phenomenon common enough to be relevant to the theory of routing. All examples discussed here, as well as many others, will be covered by a general result (Theorem 14) proved later.

Second example: Referring to Fig. 2, which shows the reduced state diagram of the three-stage Clos network of Fig. 1, we observe that only in the state numered 4 are there any blocked calls. State 4 realizes the same assignment of inlets to outlets as state 2, which has no blocked calls. The difference between the two is that in state 2 all the traffic passes through one middle switch, leaving the other entirely free for any call that may arise. This difference illustrates the intuitive rule that one should always put a call through the most heavily loaded part of the network that will still accept it. This example was discussed in Refs. 1, 2 where it was shown (rather laboriously) that if no calls are rejected, then preferring state 2 to state 4 in state 1 is optimal. This result will be an instance of Theorem 14.

Third example: It is to be expected that in some instances a choice of route for a call is immaterial. The concentrating switch depicted in Figs. 9 and 10 is a simple example of this phenomenon. It is intuitively obvious that, because of the symmetries of the network, it makes no difference which of the two trunks a call could use when the system is empty is assigned to it. This insensitivity of performance to routing choices can actually be deduced from Theorem 7.

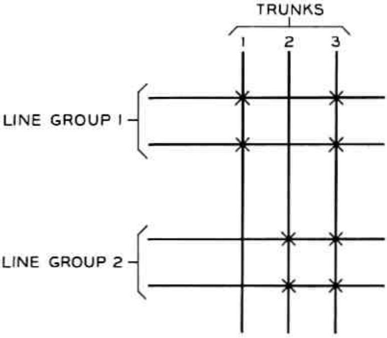


Fig. 9 — Symmetric grading.

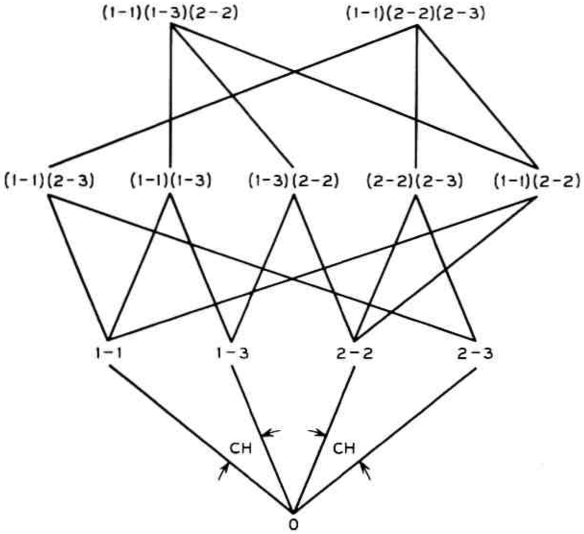


Fig. 10 — States of the grading of Fig. 9.

Fourth example: Figs. 11 and 12 show the structure and (reduced) state diagram for another simple Clos network made of 3 × 3 inlet and outlet switches, and 2 × 2 middle switches. Again, from scrutiny of the state diagram we guess that optimal routing will result if no empty middle switches are used when partially filled ones are available. The notations '*B*' in Fig. 12, intended to suggest that the states to the left of the *B*'s are "better" than those on the right, constitute an expression of the corresponding policy, and are explained in the next paragraphs.

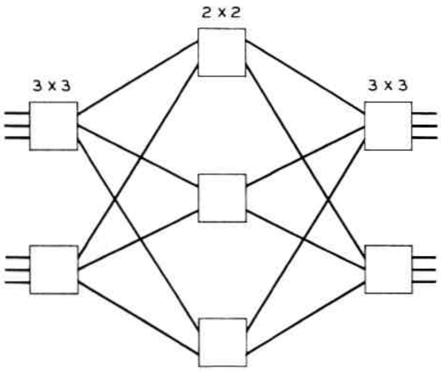


Fig. 11 — 3-stage Clos network with 3 × 3 outer switches.

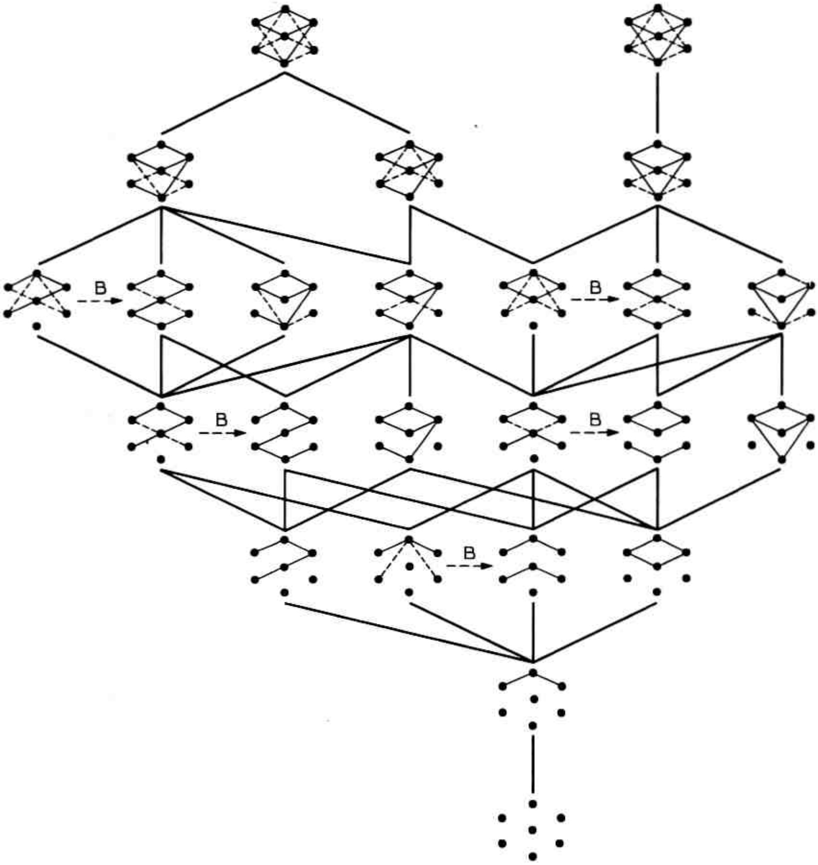


Fig. 12 — States of 3-stage Clos network of Fig. 11.

To abstract the essential features of the preceding examples into a general theorem, we start with the observation that in choosing to enter a state x rather than another y in putting up a call we have always to choose between *equivalent* states ($x \sim y$, in the sense of Section III), in which the same events e can occur. In particular, the same new calls c can arise. If it now happens that every new call blocked in x is also blocked in y, let us regard this as *prima facie* evidence that x is somehow "better" than y, and define a relation $B \subseteq S^2$ by the condition

$$xBy \text{ if and only if } x \sim y \text{ and}$$

$$c \in x,\, c \text{ blocked in } x \quad \text{imply} \quad c \text{ blocked in } y.$$

The relation B is a partial ordering.

In the first example considered above, (1-1)B(1-2), and B obtains between no other distinct states; in the second, 2B4, and again B obtains between no other distinct states.

Let us now suppose (for a general network with state set S) that the network is run according to a policy φ, and ask what happens to B under φ. That is, more specifically, we look at states x,y such that xBy, and we consider, for events e that are either hangups or new calls blocked in neither x nor y, whether or not

$$\varphi(e,x)B\varphi(e,y).$$

If e occurs and φ is used for decisions, then the system moves from x to $\varphi(e,x)$ and from y to $\varphi(e,y)$. If $\varphi(e,x)B\varphi(e,y)$ for all $e \in x$ that are either hangups or new calls blocked in neither x nor y, whenever xBy, we say that φ *preserves* B. Formally,

φ preserves B if and only if xBy implies $\varphi(e,x)B\varphi(e,y)$ for $e \in x$ which are either hangups or new calls blocked in neither x nor y.

In the first example (Fig. 8) there are no new calls c which can be put up in both (1-1) and (1-2), and there is one hangup (say h) which can occur in both. Thus, the set of events to be considered is just $\{h\}$. Clearly, $\varphi(h,1\text{-}1) = \varphi(h,1\text{-}2) = 0$ state for any φ. Since B is reflexive, we conclude that in this case every φ preserves B.

In the second example, a similar situation arises. There are two events to be considered: one is a new call completable in both 2 and 4 leading to state 6, the other is a hangup leading to 1. Again

$$\varphi(e,2) = \varphi(e,4)$$

for all φ and both events e to be considered, and again any φ preserves B.

As noted, routing has no effect in the third example. However, the relation B is defined. It can be verified that any φ preserves B, and that in this case B is a symmetric relation, as it should be, since if routing is to have no effect, then x can only be "just as good" as y if y is "just as good" as x. These facts can be used to *prove* that routing has no effect in this example.

The fourth example, finally, shows the relation B in action. The notations

$$x - - B - \rightarrow y \quad x,y \text{ states}$$

in Fig. 12 show the *irreflexive* part of B. (Obviously xBx for all $x \in S$, and this part of B is not shown in Fig. 12.) The reader is invited to verify that the policy φ of using a partly-filled middle switch whenever possible does indeed preserve B in this example.

The property of a policy φ, that it preserves B, is to be viewed as a kind of *isotony* of φ:

$$xBy \quad \text{implies} \quad \varphi(e,x)B\varphi(e,y), \quad \text{for suitable } e.$$

(See G. Birkhoff,[5] p. 3.) It can also be viewed as a kind of *continuity*, for after all if we think of the set of *neighbors* N_y of y as the states in

$$N_y = A_y \cup B_y,$$

then the property says that if xBy then also zBw where z is a neighbor of x and w a neighbor of y such that $z \sim w$. In other words it states that if xBy then also

$$(N_x \times N_y) \cap (\sim) \subseteq B,$$

i.e., if it holds between x and y then it also holds between equivalent neighbors of x and y.

Note that if φ preserves B, xBy, and φ rejects in x a call c not blocked in y, then it also rejects it in y.

For φ a policy, let

$E_x(n,\varphi)$ = expected number of successful attempts in n events, if the network starts in state x and policy φ is followed.

The isotonic property that φ preserve B has the useful feature that it implies an isotony among the numbers

$$\{E_x(n,\varphi), \quad n \geqq 1, \quad x \in S\}.$$

This is the content of the next result.

Theorem 7: (First Isotony Theorem): If φ preserves B, then xBy implies

$$E_x(n,\varphi) \geqq E_y(n,\varphi), \qquad n = 1, 2, \cdots .$$

Proof: xBy, $c \in x$, $\varphi(c,y) \neq y$ imply $\varphi(c,x) \neq x$. Hence,

$$\sum_{\substack{c \in x \\ \varphi(c,x)=x}} 1 \leqq \sum_{\substack{c \in y \\ \varphi(c,y)=y}} 1,$$

and $E_x(1,\varphi) \geqq E_y(1,\varphi)$. As a hypothesis of induction assume that xBy implies

$$E_x(n,\varphi) \geqq E_y(n,\varphi)$$

for some $n \geqq 1$. We have

$$E_x(n + 1,\varphi) = \sum_{\substack{c \in x \\ \varphi(c,x) \neq x}} \frac{\lambda}{|x| + \lambda\alpha_x} \{1 + E_{\varphi(c,x)}(n,\varphi)\}$$

$$+ \frac{\lambda}{|x| + \lambda\alpha_x} E_x(n,\varphi) \sum_{\substack{c \in x \\ \varphi(c,x) = x}} 1 + \frac{1}{|x| + \lambda\alpha_x} \sum_{h \in x} E_{x-h}(n,\varphi).$$

Since φ preserves B, it must be true that xBy implies

$$\varphi(c,x)B\varphi(c,y)$$

$$(x - h)B(y - h),$$

whence

$$E_{\varphi(c,x)}(n,\varphi) \geqq E_{\varphi(c,y)}(n,\varphi)$$

$$E_{x-h}(n,\varphi) \geqq E_{y-h}(n,\varphi).$$

Therefore,

$$E_x(n + 1),\varphi) \geqq \sum_{\substack{c \in y \\ \varphi(c,y) \neq x}} \frac{\lambda}{|x| + \lambda\alpha_y} \{1 + E_{\varphi(c,y)}(n,\varphi)\}$$

$$+ \frac{\lambda}{|x| + \lambda\alpha_y} E_y(n,\varphi) \sum_{\substack{c \in y \\ \varphi(c,y) = y}} 1$$

$$+ \frac{1}{|y| + \lambda\alpha_y} \sum_{y \in h} E_{y-h}(n,\varphi)$$

$$\geqq E_y(n + 1,\varphi).$$

The power and utility of the relation B are further illustrated by the following comparison theorem for policies. The partial ordering B on S induces a natural partial ordering B of the policies according to the definition

$$\varphi B\psi \equiv e \in x,\ x \in S \quad \text{imply} \quad \varphi(e,x)B\psi(e,x)$$

for e a hangup or a call not blocked in x. We note that $\varphi B\psi$ implies that φ and ψ embody the same rejection policy.

Theorem 8: If $\varphi B\psi$, and one of φ,ψ preserves B, then xBy implies

$$E_x(n,\varphi) \geqq E_x(n,\psi), \qquad n = 1, 2, \cdots.$$

Proof: φ and ψ have the same rejection policy, so $E(1,\varphi) = E(1,\psi)$, and the theorem holds for $n = 1$. Assume as a hypothesis of induction that xBy implies $E_x(n,\varphi) \geqq E_y(n,\psi)$ for a given value of $n \geqq 1$. We have, with $p_{ex} = \Pr\{e \text{ occurs in } x\}$,

$$E_y(n+1,\varphi) = E_y(1,\varphi) + \sum_{e \in y} p_{ey}E_{\varphi(e,y)}(n,\varphi).$$

But $e \in y$ implies $\varphi(e,y)B\psi(e,y)$, and so by the induction hypothesis

$$E_{\psi(e,y)}(n,\varphi) \geqq E_{\psi(e,y)}(n,\psi).$$

However,

$$E_y(n+1,\psi) = E_y(1,\psi) + \sum_{e \in y} p_{ey}E_{\psi(e,y)}(n,\psi)$$
$$\leqq E_y(n+1,\varphi).$$

Let now xBy, and suppose that φ preserves B. The isotony theorem then implies

$$E_x(n+1,\varphi) \geqq E_y(n+1,\varphi)$$
$$\geqq E_y(n+1,\psi).$$

If, instead, ψ preserves B, then

$$E_x(n+1,\psi) \geqq E_y(n+1,\psi)$$

and a repetition of the first part of the argument above with x instead of y gives

$$E_x(n+1,\varphi) \geqq E_x(n+1,\psi)$$
$$\geqq E_y(n+1,\psi).$$

XV. SECOND INTUITIVE APPROACH

In an effort to develop a more general theory than the one that was begun in the previous two sections, we now make a fresh start at understanding the structure of "good" routing; again, we begin with a special case:

Fifth example: We choose the overflow system or grading depicted in Fig. 13. There are two groups of lines, one of two lines, the other of three lines. Each has access to one primary trunk to which the other does not have access, and they share a single common overflow trunk. The possible states of this system form the partially ordered system shown in Fig. 14. Alternative ways of putting up particular calls are marked with "ch", for "choice".

After inspecting the system and its state diagram, intuition tells us

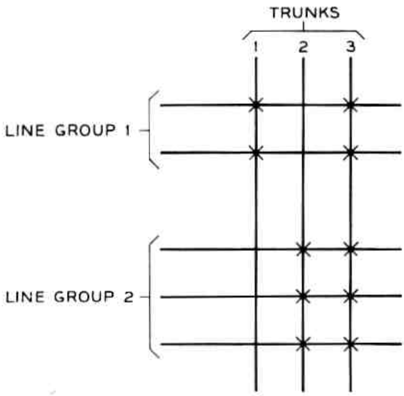


Fig. 13 — Second asymmetric grading.

that, as a first guess, calls should use the primary trunks whenever they can, so as to leave the overflow open as much as possible. Let us, on this basis, formulate some preferences for certain routes.

Clearly, in state 0 a call from group 1 should go on trunk 1, so in state 0 we prefer state (1-1) to (1-3); similarly we prefer (2-2) to (2-3). The same principle should apply if certain calls are already in progress. Thus, in state (2-2) we prefer (1-1) (2-2) over (1-3) (2-2), and in state (1-1) we prefer (1-1) (2-2) to (1-1) (2-3).

If taken seriously and followed, the preferences listed above define a

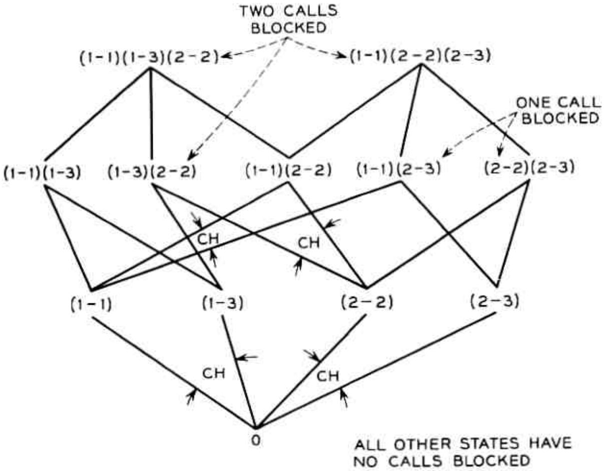


Fig. 14 — States of the grading of Fig. 13.

policy for putting in calls. We shall show that this policy differs from the optimal policy only in that the latter may reject some calls, while the former accepts all unblocked calls. To do this write xPy if state x is preferred to state y. Thus, the relation P is defined by the conditions

$$(1\text{-}1)\ P\ (1\text{-}3)$$

$$(2\text{-}2)\ P\ (2\text{-}3)$$

$$(1\text{-}1)(2\text{-}2)\ P\ (1\text{-}3)(2\text{-}2)$$

$$(1\text{-}1)(2\text{-}2)\ P\ (1\text{-}1)(2\text{-}3).$$

We let

$E_x(n)$ = expected number of successful call attempts in n events, if the system starts in state x and an optimal policy is used.

It must be explained here that by "use of an optimal policy" over n steps we mean simply that we use a policy which will maximize the average number of successful attempts *among those n events*; the policies that achieve this may, for all we know at this point, be different for different n.

A slight departure from the probabilistic model of Section VI is necessary here: we assume that an idle line generates calls to the trunk destination at a rate $\lambda > 0$, instead of assuming that an idle inlet-outlet pair generates calls at λ. Also, we let α_x be the number of idle lines in x, rather that than that of idle inlet-outlet pairs, and $s(x)$ that of idle lines that are not blocked.

Theorem 9: If xPy, then

$$E_x(n) \geqq E_y(n) \qquad n = 1, 2, 3, \cdots .$$

Proof:

$$E_x(1) = \frac{\lambda s(x)}{|x| + \lambda \alpha_x},$$

and xPy implies $s(x) \geqq s(y)$, so the theorem is true for $n = 1$. Assume that the theorem holds for some $n \geqq 1$. There are four cases, corresponding to the four conditions defining P. We shall give the argument for the case where

$$x = (1\text{-}1)(2\text{-}2)$$

$$y = (1\text{-}3)(2\text{-}2),$$

and (as we know) xPy; the others are similar.

Now apparently

$$E_{(1\text{-}1)(2\text{-}2)}(n+1) = \frac{1}{2+3\lambda}\{E_{(2\text{-}2)}(n) + E_{(1\text{-}1)}(n)\}$$
$$+ \frac{\lambda}{2+3\lambda}\max\{E_{(1\text{-}1)(2\text{-}2)}(n),\ 1 + E_{(1\text{-}1)(1\text{-}3)(2\text{-}2)}(n)\}$$
$$+ \frac{2\lambda}{2+3\lambda}\max\{E_{(1\text{-}1)(2\text{-}2)}(n),\ 1 + E_{(1\text{-}1)(2\text{-}2)(2\text{-}3)}(n)\}$$

and

$$E_{(1\text{-}3)(2\text{-}2)}(n+1) = \frac{1}{2+3\lambda}\{E_{(2\text{-}2)}(n) + E_{(1\text{-}3)}(n)\}$$
$$+ \frac{\lambda}{2+3\lambda}\max\{E_{(1\text{-}3)(2\text{-}2)}(n),\ 1 + E_{(1\text{-}1)(1\text{-}3)(2\text{-}2)}(n)\}$$
$$+ \frac{2\lambda}{2+3\lambda}E_{(1\text{-}3)(2\text{-}2)}(n).$$

By the induction hypothesis,

$$E_{(1\text{-}1)}(n) \geqq E_{(1\text{-}3)}(n)$$
$$E_{(1\text{-}1)(2\text{-}2)}(n) \geqq E_{(1\text{-}3)(2\text{-}2)}(n);$$

hence,

$$E_x(n+1) \geqq E_y(n+1)$$

for the given x and y.

The point is that each event that can occur leads to a "worse" state in y than it does in x. Thus, the hangup of the group 1 call leads both to the state 2-2, a standoff; hangup of the group 2 call takes x into (1-1) and y into (1-3), and (1-1)P(1-3); one of the possible new calls leads both x and y to the state (1-1)(1-3)(2-2), another standoff; the other two possible new calls are blocked in y but not in x, so that by the induction hypothesis, rejecting one of them and staying in x is at least as good as having one of these blocked calls make an attempt in y.

We conclude from Theorem 9 that in an optimal policy the calls which are not rejected are put on the primary trunks if these are available, and on the overflow only if the primary trunk appropriate to the call is already busy. This result is entirely in agreement with our original intuition.

Another example of the same kind is shown in Figs. 15 and 16: the intuitive preferences shown in Fig. 16 by 'P' are optimal to within rejection of unblocked calls.

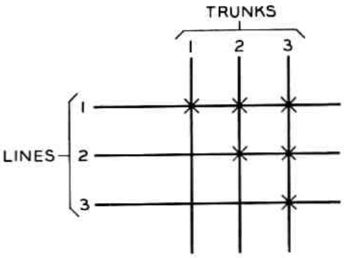


Fig. 15 — Third asymmetric grading.

We now formalize the principles behind the intuitions that led to Theorem 9.

Let P be a relation on S, i.e., a subset of S^2. We may as well put our cards on the table and indicate that P is to be interpreted as a relation of "preference", with xPy meaning "x is preferred to y". If μ is a function, and X,Y are sets, the (customary) notation

$$\mu: \quad X \leftrightarrow Y$$

means that μ takes X into Y in a one-one manner, while

$$\mu: \quad X \rightarrow Y$$

means that the μ-image of X is contained in y.

We say that P has the *strong monotone property* if xPy implies

(*i*) $|x| = |y|$
(*ii*) $\exists\mu$: $B_x \leftrightarrow B_y$ such that $z \in B_x$ implies $zP\mu z$
(*iii*) $\exists\nu$: $A_y \rightarrow A_x$ such that

$$\begin{aligned} \nu(A_{cy}) \subseteq A_{cx} \quad &\text{for} \quad c \in y, \\ z \in A_y \quad \text{implies} \quad &\nu zPz. \end{aligned} \tag{14}$$

Let us denote by F_x the set of all calls which are *free* or idle in x, i.e.

$$\begin{aligned} F_x &= \{c\colon \ c \text{ is idle in } x\} = \{\gamma(y - x)\colon \ y \in A_x\} \\ &= \{c\colon \ c = \{(u,v)\} \subseteq I \times \Omega \text{ with } u,v \text{ both idle in } x\}. \end{aligned}$$

We say that a relation P on S has the *weak monotone property* if xPy implies

(*i*) $|x| = |y|$
(*ii*) $\exists\mu$: $B_x \leftrightarrow B_y$ and $z \in B_x$ implies $zP\mu z$
(*iii*) $\exists\nu$: $F_y \rightarrow F_x$ and $c \in F_y$, $z \in A_{cy}$

$$\text{imply } \exists w \in A_{(\nu c)x} \quad \text{with} \quad wPz. \tag{15}$$

To get the weak monotone property from the strong, define ν on F_y by

$$\nu\gamma(z - y) = \gamma(\nu z - x), \qquad z \in A_y ;$$

then $z \in A_{cy}$ implies $\nu z \in A_x$, and

$$\nu c = \gamma(\nu z - x);$$

thus,

$$\nu z \in A_{(\nu c)x} \quad \text{and} \quad \nu z P z.$$

Keeping in mind the interpretation that 'xPy' means that x is in some sense better than y, we see that: condition (*i*) restricts P to hold only between states of the same norm or dimension, because we are interested only in choosing between states with the same number of calls in progress; condition (*ii*) says roughly that to every hangup leading out of state y there corresponds a hangup in x leading to a state which is at least as "good" (as the one reached by the hangup in y); condition (*iii*) says that for any way of completing a new call c in y there is a way of completing the *same* call c in x which leads to at last as "good" a state (as the one reached by completing that call in y).

It is easily seen that P has one of the monotone properties if and only if xPy implies that P holds between "corresponding respective" neigh-

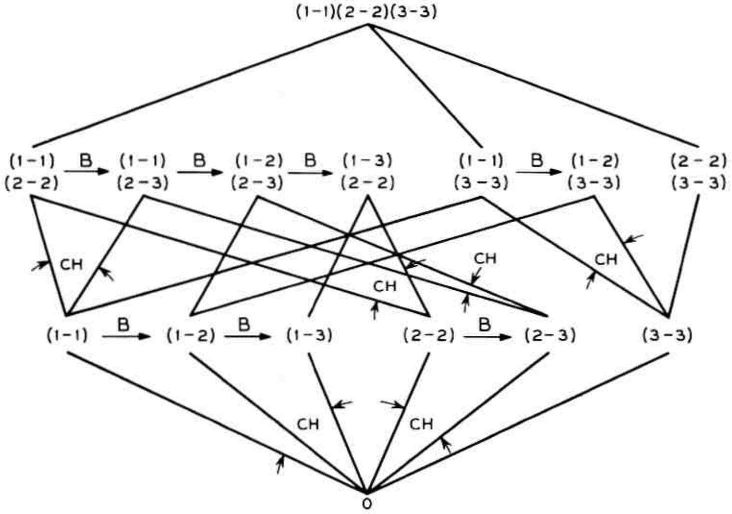


Fig. 16 — States of the grading of Fig. 15.

bors of x and y. Thus, the monotone properties are similar to the property of a policy φ that it preserve B. The principal differences are that here no policy is at issue, and that the meaning of "corresponding neighbor" is weaker than in the definition of preservation. The relationships to the relation B are further clarified in the following remarks.

If P has the weak monotone property, then xPy implies $s(x) \geqq s(y)$. If P has the strong monotone property, then xPy implies that every $c \in x$ blocked in x is blocked in y. Further, since we are primarily interested in comparing *equivalent* states (i.e., x and y such that $x \sim y$), it is natural to restrict attention to preference relations P which are subsets of $\sim$, $P \subseteq \sim$. It can then be verified that if P has either monotone property, and holds only between equivalent states ($P \subseteq \sim$), then $P \subseteq B$.

A "preference" relation should impose at least a partial ordering among the objects for which it is defined, and so it is by nature transitive. The question then arises whether the relations P that have the (strong or weak) monotone property are reflexive and transitive. It is obvious that if P has the monotone property then so does $I \cup P$ where I is the identity relation. Now, as is known, every relation P can be extended uniquely to its *transitive closure* $\bar{P}$, the smallest transitive relation containing P. We shall now prove:

Theorem 10: If $P \subseteq S^2$ has the weak monotone property, then so does its transitive closure $\bar{P}$.

Proof: Clearly $\bar{P} = P \cup P^2 \cup P^3 \cup \cdots$, where the powers represent relative, not Cartesian, products. It is obvious that $x\bar{P}y$ implies $|x| = |y|$, so $\bar{P}$ has property (i) of (14). Next let $x\bar{P}y$, so that there exist $z_1, z_2, \cdots, z_n \in S$ such that $z_1 = x$, $z_n = y$ and

$$z_iPz_{i+1} \qquad i = 1, \cdots, n-1.$$

Thus, there exist maps $\mu_1, \mu_2, \cdots, \mu_{n-1}$ with $\mu_i : B_{z_i} \leftrightarrow B_{z_{i+1}}$ such that $z \in B_{z_i}$ implies

$$zP\mu_iz.$$

Hence, $z \in B_x$ implies

$$zP\mu_1z$$

$$\mu_1zP\mu_2\mu_1z$$

$$\vdots$$

$$\mu_{n-2}\mu_{n-3}\cdots\mu_1zP\mu_{n-1}\mu_{n-2}\cdots\mu_1z,$$

i.e.,

$$z\bar{P}\left(\prod_{j=1}^{n-1} \mu_j\right)z.$$

Thus,

$$\mu = \prod_{j=1}^{n-1} \mu_j$$

has the property that μ: $B_x \leftrightarrow B_y$ and $z \in B_x$ implies $z\bar{P}\mu z$. Hence, $\bar{P}$ has property (ii). Finally, there exist maps $\nu_1, \cdots, \nu_{n-1}$ with ν_{n-i}: $F_{z_{i+1}} \to F_{z_i}$ such that $c \in F_{z_{i+1}}$, $z \in A_{cz_{i+1}}$ implies $w \in A_{(\nu_i c)z_{i+1}}$ with wPz. Let

$$\nu = \prod_{i=1}^{n-1} \nu_i .$$

Hence, for each $c \in F_y$, $w_n \in A_{cy}$ there exist $w_{n-1}, \cdots, w_n \in S$ and $c_{n-1} \cdots c_{n-1}$such that

$$c_i = \nu_i c_{i+1}, \; w_i \in A_{c_i z_i}, \qquad w_i P w_{i+1} \qquad i = 1, \cdots, n-1.$$

It is apparent that $c_1 = \nu c$, $w_1 \in A_{(\nu c)x}$ and $w_i \bar{P} w_n$, so that $\bar{P}$ has property (iii).

The following result is now immediate:

Theorem 11: If P has the weak monotone property, and I is the identity relation, then

$$\overline{(I \cup P)}$$

is a partial ordering relation with the weak monotone property.

Any relation with the weak monotone property can be extended to be a *partial ordering* P that has the weak monotone property. Since $\sim$ is an equivalence relation between states, and P is a partial ordering, it follows that $P \cap \sim$ is also a partial ordering.

Theorem 12: (Second Isotony Theorem): If $P \subseteq S^2$ has the weak monotone property, then

$$xPy \quad \textit{implies} \quad E_x(n) \geqq E_y(n), \qquad n = 1, 2, \cdots .$$

Proof: Property (15) (iii) implies that $s(x) \geqq s(y)$ whenever xPy. Now

$$E_x(1) = \frac{\lambda s(x)}{|x| + \lambda\alpha_x}.$$

Since it is assumed that $\alpha_x = \alpha_{|x|}$ we have, by (15) (*i*),

$$xPy \quad \text{implies} \quad E_x(1) \geqq E_y(1).$$

As an hypothesis of induction assume that xPy implies $E_x(n) \geqq E_y(n)$. We have

$$E_x(n+1) = \frac{1}{|x| + \lambda\alpha_x} \sum_{c \in x} \max \{E_x(n), g(c,x) + \max_{z \in A_{cx}} E_z(n)\}$$
$$+ \frac{\lambda}{|x| + \lambda\alpha_x} \sum_{h \in x} E_{x-h}(n),$$

and a similar expression for $E_y(n+1)$. If now xPy, then $|x| = |y|$ by (15) (*i*), and also

$$E_{x-h}(n) \geqq E_{\mu(x-h)}(n)$$

by (15) (*ii*) and the hypothesis of induction. Similarly,

$$\frac{\lambda\beta_y}{|x| + \lambda\alpha_x} E_x(n) \geqq \frac{\lambda\beta_y}{|y| + \lambda\alpha_y} E_y(n).$$

For c not blocked in y, and $z \in A_{cy}$, xPy implies that there exists $w \in A_{(\nu c)x}$ with wPz, by (15) (*iii*). By the hypothesis of induction, this implies that

$$E_w(n) \geqq E_z(n).$$

Since $z \in A_{cy}$ was arbitrary, we find

$$g(\nu c,x) + \max_{w \in A_{(\nu c)x}} E_w(n) \geqq g(c,y) + \max_{z \in A_{cy}} E_z(n).$$

It follows that xPy implies $E_x(n+1) \geqq E_y(n+1)$.

XVI. RELEVANCE OF THE ISOTONY THEOREMS TO OPTIMAL POLICIES

Let $c \in x$ be a call that is not blocked in state x, so that $A_{cx} \neq \theta$. If the hypotheses of one of the isotony theorems obtain, then it may be possible to single out some of the states $y \in A_{cx}$ as providing ways of completing c in x which are at least as good as certain others. Specifically, the sort of comparison we can make is this: If $y,z \in A_{cx}$ and yBz or yPz, then y is at least as good as z in the sense that

$$E_y(n) \geqq E_z(n), \qquad n = 1, 2, \cdots.$$

Suppose now that there is at least one $y \in A_{cx}$ such that yBz for *all* $z \in A_{cx}$. It then follows that such a y is always at least as good a choice as any other state of A_{cx}, in the above sense. A similar result follows if there is a $y \in A_{cx}$ with yPz for all $z \in A_{cx}$. In such situations a policy that routes c so as to take the system from x to y can differ (so far as x and c are concerned) from an optimal policy only in the respect that an optimal policy might reject c in x. This is the sense in which the isotony theorems can provide the part of the solution of the routing problem which has to do with choosing routes for accepted calls. Two theorems to this effect appear in Section XVIII after an aside about equivalence of decisions and nonuniqueness of optimal policies.

XVII. EQUIVALENCE OF DECISIONS AND NONUNIQUENESS OF OPTIMAL POLICIES

It is natural to expect that there are often several optimal policies, in the sense that, for some c and x with $c \in x$ and $A_{cx} \neq \theta$, there are two choices of a route for c in x which are in some sense distinct routes and yet are both equally "good". For example, in most traffic models for a graded or progressive multiple it often does not make any difference which trunk in a group is used for a call: the possible states resulting from use of one of the trunks in the group are all distinct, yet all are equally "good", being "equivalent" under permutations of trunks within the group. It is intuitively clear that such a nonuniqueness of optimal policies is due in large part to symmetries in the network under study, or more generally, to the presence of various equivalences of states (and hence of routing decisions) under certain groups of permutations of terminals.† Since some of these equivalences appear in a later proof, we digress a little for an account of them, first heuristic, then formal.

As we have seen, one of the principal tools in the description of optimal policies is a combinatorial partial ordering, such as B or P, which implies an ordering in terms of performance. The discussion to follow is based on a general partial ordering R, which the reader can assume is contained in

$$\bigcup_{c \in x} A_{cx}^{2}$$

and which he can interpret as B or P, if he wishes.‡

Let then R be a partial ordering of S and let Y be a subset of S. Cued by the remarks of Section XVI, we want to use R to compare states; in

† It should be noted that the word 'group' is used in this paragraph in two technical senses, the first from traffic theory, referring to a set of trunks, the second from the theory of groups.

‡ This use of 'R' is peculiar to this section, and should not be confused with R as a routing matrix.

particular we wish to talk about elements $y \in Y$ such that yRz for all $z \in Y$. It would be satisfyingly simple if at this point we could introduce the notation

$$\sup_R Y$$

for that element of Y which bears R to every other element of Y. Unfortunately this is usually impossible, because there may be several or many such "suprema" of Y. In this situation the usual mathematical trick to use is to pass to suitable equivalence classes. Use of this procedure is further justified by the fortunate fact that, in the case of several interesting choices of R and Y, there are several senses in which these maximal elements turn out to be equivalent. What is more, there is a natural equivalence based only on R, such that $\sup_R Y$ can, if it exists, be defined in the "quotient" set of the equivalence, i.e., in the image of the semilattice homomorphism that takes each state into the equivalence class to which it belongs.

If $R = P$ and P has the monotone property, then all the P-suprema of A_{cx} are equivalent in this very important sense: If $y_1, \cdots, y_m$ is an enumeration of all the $y \in A_{cx}$ that are best in the sense that yPz for all $z \in A_{cx}$, then

$$y_iPy_j, \qquad 1 \leqq i, j \leqq m$$

and the second isotony theorem gives

$$E_{y_i}(n) = E_{y_j}(n) \qquad n = 1, 2, \cdots, \tag{16}$$

so that *as far as performance is concerned*, $y_1, \cdots, y_m$ are all "equivalent". In many cases, this fact is based on an underlying equivalence of a combinatorial nature, much stronger than (16): e.g., in a trunk group the different states attainable by different choices of a trunk for a call are equivalent in the sense that given any two there is a renaming or permutation of the trunks which carries one into the other.

The isotony theorems provide ways of translating a *combinatorial comparison* of states such as

$$xBy, \quad \text{or} \quad xPy$$

into a *numerical comparison* of the relative merit or value of starting in each state, x or y. In such a setting it is natural to call x and y "equivalent" if the comparison holds both ways, i.e., if, when interpreting 'xBy' as a (rather strong) precise form of 'x is better than y', we have both

$$xBy \quad \text{and} \quad yBx.$$

Lemma 6: Given two states y,z there exists at most one pair c,x such that both

$$y,z \in A_{cx}.$$

Proof: If $y,z \in A_{cx}$ then $x = y \cap z$ in the sense of the semi-lattice of states. Thus, x is unique. If now

$$y,z \in A_{cx} \cap A_{c'x}$$

then $c = \gamma(y - x)$, $c' = \gamma(y - x)$, so $c = c'$.

The foregoing observations are the motivation for the ensuing development. With the partial ordering R we associate the natural equivalence relation $\equiv_R$ defined by

$$z \equiv_R y \quad \text{if and only} \quad zRy \quad \text{and} \quad yRz \quad \text{and} \quad \exists A_{cx} \quad y,z \,\varepsilon\, A_{cx}.$$

The subscript R will usually be dropped as long as it is contextually clear what R is being used to define $\equiv$. Along with $\equiv$ we introduce the semilattice homomorphism

$$r(\cdot)\colon\ S \rightarrow \{\text{equivalence classes of} \equiv\} = S/\equiv$$

$$\text{defined by } \tau(x) = \{z\colon \quad z \equiv x\}.$$

The image $\tau(S)$, i.e., the "quotient" set $S/\equiv$, is partially ordered by the relation R defined by

$$\tau(x)R\tau(y) \quad \text{if and only} \quad u,v \quad u \in \tau(x) \quad \text{and} \quad v \in \tau(y) \quad \text{and} \quad uRv.$$

This is the natural homomorphic "contraction" of R to $S/\equiv$. It can be verified that if $\tau(x)R\tau(y)$ and $\tau(y)R\tau(x)$, then $\tau(x) = \tau(y)$ strictly.

If now Y contained in S is such that there exists a $y \in Y$ with yRz for every $z \in Y$, we use the notation

$$\sup_{R} Y \tag{17}$$

for $\tau(y)$. It is clear that in the "quotient" space, an element maximal with respect to R is unique if it exists at all. Strictly speaking the notation

$$\sup_{\tau R} \tau Y$$

would be better, since it indicates that the supremum operation only makes sense after the homomorphism. However, (17) will be used, with the reminder that it is a set, not a state, and the convention that use of (17) implies the assumed existence of maximal elements.

With the notation (17) we can prove the following natural relationship between the strong monotone property and the notion of preservation of B.

Theorem 13: Let

$$\varphi(e,x) \begin{cases} \in \sup_B A_{e,x} & \textit{for} \quad e = c \\ = x - h & \textit{for} \quad e = h \end{cases}$$

and suppose that φ preserves B. Then B has the strong monotone property.

Proof: xBy implies $x \sim y$ and hence $|x| = |y|$, so B has property (14), (*i*). If xBy, define for $z \in B_x$

$$\mu z = \varphi(\gamma(x - z),y).$$

Then, since φ preserves B

$$\varphi(\gamma(x - z),x)B\varphi(\gamma(x - z),y),$$

$$zB\mu z,$$

and B has property (14) (*ii*). With xBy still, let

$$\nu: \quad A_y \rightarrow A_x$$

be given by $\nu z = \varphi(c,x)$ for $z \in A_{cy}$. Then, since $\varphi(c,y)Bw$ for $w \in A_{cy}$,

$$\varphi(c,x)B\varphi(c,y)$$

$$Bz,$$

so that B has property (14) (*iii*).

XVIII. OPTIMAL ROUTING THEOREMS

This final section contains precise statements showing just how the combinatorial properties introduced in Sections XIV and XV answer the question: "Which route should an accepted call use? "

Two policies φ and ψ will be termed *equivalent with respect to rejections*, written $\varphi \sim \psi$, if they both reject the same calls in the same states, i.e., if $\varphi(c,x) = x$ when and only when $\psi(c,x) = x$ for $c \in x$.

Theorem 14: If φ preserves B, and if $c \in x$ implies

$$\varphi(c,x) \in \sup_B A_{cx}$$

whenever $\varphi(c,x) \neq x$, then

$$E_x(n,\varphi) \geqq E_x(n,\psi) \qquad n = 1, 2, \cdots .$$

for any $\psi \sim \varphi$.

Proof: $E_x(1,\varphi) = E_x(1,\psi)$ by direct calculation. Assume as a hypothesis of induction that $E_x(n,\varphi) \geqq E_x(n,\psi)$ for $x \in S$. We have

$$\begin{aligned} E_x(n+1,\varphi) = & \sum_{\substack{c \in x \\ \varphi(c,x) \neq x}} \frac{\lambda}{|x| + \lambda\alpha_x} \{1 + E_{\varphi(c,x)}(n,\varphi)\} \\ & + \frac{1}{|x| + \lambda\alpha_x} \sum_{\substack{c \in x \\ \varphi(c,x) = x}} E_x(n,\varphi) \\ & + \frac{1}{|x| + \lambda\alpha_x} \sum_{h \in x} E_{x-h}(n,\varphi), \end{aligned}$$

and a similar expression for $E_x(n+1,\psi)$. If now $\varphi(c,x) \neq x$, then $\varphi(c,x)By$ for every $y \in A_{cx}$; in particular, $\psi(x,x) \neq x$ because $\varphi \sim \psi$, and so $\psi(c,x) \in A_{cx}$, whence

$$\varphi(c,x)B\psi(c,x).$$

The first isotony theorem and the induction hypothesis now give

$$\begin{aligned} E_{\varphi(c,x)}(n,\varphi) & \geqq E_{\psi(c,x)}(n,\varphi) \\ & \geqq E_{\psi(c,x)}(n,\psi). \end{aligned}$$

It follows that

$$E_x(n+1,\varphi) \geqq E_x(n+1,\psi).$$

Corollary: If φ preserves B, and

$$\varphi(c,x) \in \sup_B A_{cx}$$

for c $\in$ x not blocked in x, then φ is optimal within the class of policies that reject no unblocked calls.

Theorem 15: If P has the weak monotone property, and

$$\sup_P A_{cx}$$

exists for each c $\in$ x not blocked in x, then there exists an optimal policy R such that c $\in$ x, $y \in A_{cx}$ imply either x is R-transient or else

$$r_{xy} = 0 \quad \textit{unless} \quad y \in \sup_P A_{cx}.$$

Proof: Let ρ be the scalar such that

$$\rho 1 = \max_{R \in C} \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} A^j v \qquad \begin{cases} v = A(R), \\ A = A(R). \end{cases}$$

We first use an argument of R. Bellman[10] to show that the vector sequence

$$E(n) - n\rho 1 = g(n)$$

is bounded in n.

By Lemma 3, there is a vector g^* which satisfies

$$g^* + \rho 1 = \max_{R \in C} \{v(R) + A(R)g^*\} .$$

Choose $K > 0$ so that

$$g^* - K1 \leqq g(1) \leqq g^* + K1.$$

Assume, as an induction hypothesis, that

$$g^* - K1 \leqq g(n) \leqq g^* + K1.$$

We have

$$g(n+1) = -\rho 1 + \max_{R \in C} \{v(R) + A(R)g(n)\} .$$

Hence,

$$\begin{aligned} -\rho 1 - K1 + \max_{R} \{v(R) + A(R)g^*\} &\leqq g(n+1) \\ &\leqq -\rho 1 + K1 + \max_{R} \{v(R) + A(R)g^*\} \end{aligned}$$

$$g^* - K1 \leqq g(n+1) \leqq g^* + K1.$$

Let now

$$g = \limsup_{n \to \infty} g(n),$$

taken componentwise. Let R_n achieve the maximum in

$$\max_{R \in C} \{v(R) + A(R)g(n)\} .$$

Given $\varepsilon > 0$, there exists n_0 such that $n > n_0$ implies

$$g_x(n) \leqq g_x + \varepsilon$$

for all $x \in S$. Thus,

$$\begin{aligned} v(R_n) + A(R_n)g(n) &= v(R_n) + A(R_n)g + A(R_n)[g(n) - \bar{g}, \\ &\leqq v(R_n) + A(R_n)g + \varepsilon \\ &\leqq \max_{R \in C} \{v(R) + A(R)g\} + \varepsilon. \end{aligned}$$

Hence, since $\varepsilon > 0$ was arbitrary,

$$g + \rho 1 \leqq \max_{R \in C} \{v(R) + A(R)g\} . \tag{18}$$

Let R^* achieve the maximum on the right above. By Lemma 3, R^* is optimal. Let F be the set of transient states relative to R^*. The argument used in Lemma 4 shows that equality must obtain in (18) on $S - F$, i.e.,

$$g_x + \rho = \max_{R \in C} \left\{ v_x(R) + \sum_{y \in S-F} a_{xy}(R) g_y \right\}, \qquad x \in S - F.$$

This is equivalent to

$$g_x + \rho = \sum_{\substack{c \in x \\ c \text{ not blocked in } x}} \frac{\lambda}{|x| + \lambda \alpha_x} \max \left\{ g_x, 1 + \max_{z \in A_{cx}} g_z \right\}$$
$$+ \frac{\lambda \beta_x g_x}{|x| + \lambda \alpha_x} + \frac{1}{|x| + \lambda \alpha_x} \sum_{h \in x} g_{x-h}, \qquad x \in S - F.$$

Now the second isotony theorem implies that if xPy, then

$$E_x(n) \geqq E_y(n), \qquad n \geqq 1$$
$$g_x(n) \geqq g_y(n), \qquad n \geqq 1$$
$$g_x \geqq g_y .$$

Thus, if $c \in x$ is not blocked in x

$$\max_{z \in A_{cx}} g_z$$

is achieved by each and any $y \in \sup_P A_{cx}$.

Let R be any routing matrix such that for $y \in A_{cx}$

$$r_{xy} = \begin{cases} 0 & \text{if} \quad y \in S - F, \\ \\ 1 & \text{only if} \quad 1 + g_y \geqq g_x \quad \text{and} \quad y \in \sup\limits_P A_{cx} . \end{cases}$$

Then R achieves the maximum in (18), and so is optimal; it is clear that it also has the property claimed in the theorem.

XIX. ACKNOWLEDGMENT

The author wishes to thank J. H. Suurballe and D. Alcalay for assistance in using the LP 90 program, F. Sinden and E. Wolman for numerous discussions, E. Wolman again for a careful reading of the draft, and S. P. Lloyd for access to unpublished manuscripts.

APPENDIX

Expected Number of Events to the First Blocked Call

The purpose of this appendix is to demonstrate that if the index of performance is changed to one which attaches greater importance (than does $Pr\{bl\}$) to blocked calls occurring soon after the system is started, then no unblocked call should ever be rejected. This result can be obtained for various indices of performance; we obtain it for the expected number of events occurring until the first blocked call. This choice of index of performance has a natural heuristic justification: it corresponds to trying *to put off the undesirable event* (a blocked call) as long as possible. (Time is being measured here in discrete units, by counting events.)

As before we use φ and ψ for policies, but here we limit them to *rejection policies*, or policies for the acceptance or rejection of unblocked calls. We may think of φ as a binary function of c,x with $c \in x$ and c not blocked in x, and interpret $\varphi(c,x) = 1$ as acceptance, and $\varphi(c,x) = 0$ as rejection. A general routing policy, such as described by a fixed routing matrix R, will be said to be *within* φ if it accepts and/or rejects the same calls in the same states.

We first introduce the quantities

$E_x(\varphi)$ = Expected number of events until the first blocked or rejected call under a routing policy optimal within the rejection policy φ, starting in x.†

These satisfy the equations

$$E_x(\varphi) = \frac{|x| + \lambda s(x)}{|x| + \lambda\alpha_x} + \frac{\lambda}{|x| + \lambda\alpha_x} \sum_{\substack{c \in x \\ c \text{ not blocked in } x \\ \varphi(c,x) = 1}} \max_{y \in A_{cx}} E_y(\varphi)$$

$$+ \frac{1}{|x| + \lambda\alpha_x} \sum_{h \in x} E_{x-h}(\varphi).$$

Our object will be to pick the best rejection policy, i.e., to choose φ so as to achieve

$$\max_{\varphi} E_x(\varphi).$$

We next define, for each fixed routing matrix R

$E_x(R)$ = Expected number of events until the first blocked or rejected call, starting in x and using the policy R.

† The word 'optimal' here refers, naturally, to the fact that the (not necessarily stationary) policy followed makes the expected number of events to the first call (rather than $Pr\{bl\}$, or some other index) a maximum.

For fixed φ, let $R^* = R^*(\varphi) = (r_{xy}^*)$ be a routing matrix with the property

$$r_{xy}^* = \begin{cases} 1 & \text{if} \quad c \in x \quad \text{such that} \quad y \in A_{cx}, \qquad \varphi(c,x) = 1, \\ & \text{and} \quad E_y(\varphi) = \max\limits_{z \in A_{cx}} E_z(\varphi), \\ & \\ 0 & \text{otherwise.} \end{cases}$$

It is clear that at least one such R^* exists, that it is within φ, and that it defines a stationary policy for which

$$E(R^*) = E(\varphi).$$

We now partially order all rejection policies thus:

$$\varphi \geqq \psi \quad \text{if and only if} \quad \varphi(c,x) \geqq \psi(c,x) \quad \text{for} \quad c \in x \quad \text{not blocked in } x.$$

Let $\mathcal{R}$ be the set of rejection policies. The principal result is that $E(\cdot)$ is *isotone* on the partial ordering $\geqq$ of $\mathcal{R}$, expressed in

Theorem 16: $\varphi \geqq \psi$ *implies* $E(\varphi) \geqq E(\psi)$.

Proof: For $|S|$-vectors v define the transformations T_φ, $\varphi \in \mathcal{R}$ by

$$(T_\varphi v)_x = \frac{\lambda}{|x| + \lambda\alpha_x} \sum_{\substack{c \in x \\ \varphi(c,x)=1 \\ c \text{ not blocked in } x}} \max_{y \in A_{cx}} v_y + \frac{1}{|x| + \lambda\alpha_x} \sum_{h \in x} v_{x-h}.$$

With

$$b_x = \frac{|x| + \lambda s(x)}{|x| + \lambda\alpha_x},$$

the equation for $E(\varphi)$ becomes

$$E(\varphi) = b + T_\varphi E(\varphi).$$

It is evident that if $v \geqq 0$ and $\varphi \geqq \psi$, then

$$T_\varphi v \geqq T_\psi v.$$

Furthermore, each T_φ, $\varphi \in \mathcal{R}$, is a monotone transformation in that

$$v \geqq w \quad \text{implies} \quad T_\varphi v \geqq T_\varphi w.$$

Hence, $v \geqq w \geqq 0$, $\varphi \geqq \psi$ imply

$$b + T_\varphi v \geqq b + T_\psi w.$$

For $\varphi \geqq \psi$, then, consider the rectangular parallelopiped

$$\mathcal{P} = \{v: \quad 0 \leqq v \leqq E(\varphi)\}.$$

For $v \in \mathcal{P}$ we have

$$E(\varphi) = b + T_\varphi E(\varphi) \geqq b + T_\psi v,$$

so that $T_\psi : \mathcal{P} \rightarrow \mathcal{P}$. It is obvious that $\mathcal{P}$ is closed and that T_ψ is continuous. Hence, by Brouwer's fixed point theorem there is a $v \in \mathcal{P}$ satisfying

$$v = b + T_\psi v.$$

We next show that v is actually the unique solution of this equation, so that $v = E(\psi) \leqq E(\varphi)$. Introduce the norm $\| v \| = \max_{x \in S} v_x$. The case in which the network under study is nonblocking and ψ rejects no calls is trivial. Assume then that there exists a state x and a call $c \in x$ such that either c is blocked in x or c is not blocked in x and is rejected by ψ. This implies that the "matrix" part of T_ψ is strictly substochastic, and hence that for some n

$$\| T_\psi^n \| < 1.$$

Thus, $v = E(\psi)$.

It is an immediate consequence of Theorem 16 that if $\varphi^*(c,x) \equiv 1$ for $c \in x$ not blocked in x, then

$$E(\varphi^*) = \max_{\varphi \in \mathcal{R}} E(\varphi).$$

REFERENCES

1. Beneš, V. E., Markov Processes Representing Traffic in Connecting Networks, B.S.T.J., *42*, November, 1963, pp. 2795–2838.
2. Beneš, V. E., *Mathematical Theory of Connecting Networks and Telephone Traffic*, Academic Press, New York, 1965.
3. Weber, J. H., Some Traffic Characteristics of Communication Networks with Automatic Alternate Routing, B.S.T.J., *41*, March, 1962, pp. 1201–1247.
4. Weber, J. H., private communication.
5. Birkhoff, G., *Lattice Theory*, Revised Edition, Amer. Math. Soc. Coll. Publ., *XXV*.
6. Ketchledge, R. W., The No. 1 Electronic Switching System, IEEE Trans. Comm. Tech., *COM-13*, pp. 38–41, and references therein.
7. Kalaba, R. and Juncosa, M., Optimal Design and Utilization of Communication Networks, Manage. Sci., *3*, 1956, pp. 33–44.
8. Charnes, A. and Cooper, W. W., Programming with Linear Fractional Functionals, Naval Research Logistics Quarterly, *9*, 1962, pp. 181–185.
9. Derman, C., On Sequential Decisions and Markov Chains, Manage. Sci., *9*, 1962, pp. 16–24.
10. Bellman, R., A Markovian Decision Process, J. Math. Mech., *6*, 1957, pp. 679–684.
11. Howard, R., *Dynamic Programming and Markov Processes*, Technology Press and John Wiley & Sons, New York, 1960, p. 62.
12. Chung, K. L., *Markov Chains with Stationary Transition Probabilities*, Springer, Berlin, 1960, p. 32.
13. Widder, D. V., *The Laplace Transform*, Princeton University Press, Princeton, 1946, p. 180.
14. *Ibid.*, p. 181.

Random Tropospheric Angle Errors in Microwave Observations of the Early Bird Satellite

By J. H. W. UNGER

(Manuscript received June 16, 1966)

A simplified analytical model of tropospheric random variations in angle measurements is described. This model is used to predict the minimum and maximum power density spectra between which the tropospheric random angle errors of observations on the Early Bird satellite are expected to lie.

The apparent angular position of the Early Bird satellite was then measured at microwave frequencies with the large horn-reflector antenna at the AT&T station near Andover, Maine. Random variations in the azimuth and elevation angles have been observed and recorded. The analysis of these records results in a description of the observed random angle variations by their power density spectra.

A comparison of the predicted power density spectra from the model with the observed spectra is made. It is concluded that the observed random angle variations are indeed caused by random tropospheric refraction.

The feasibility of acquiring data on atmospheric propagation effects, particularly tropospheric angle errors, with the aid of geo-stationary satellites is therefore also demonstrated.

I. INTRODUCTION

1.1 *Objective of this Paper*

The performance of earth-based radar and optical systems is ultimately limited by temporal and spatial random variations in the refractive index of the tropospheric propagation medium. It is the objective of this paper to present a method for predicting random tropospheric angle errors in such systems, and to compare a prediction with microwave observations made on the Early Bird satellite.

1.2 *Problem Approach*

The scintillation or twinkling of the stars which is experienced in observations through the earth's troposphere is a familiar effect of the random variations in the refractive index of this propagation medium. Astronomers have known for a long time that the troposphere actually causes variations in at least four characteristics of the received star light, namely: (*i*) the intensity, (*ii*) the spectral distribution of the intensity, (*iii*) the shape of the telescopic diffraction image, and (*iv*) the apparent angular position of the star. Scientific studies[1,2] of these effects seem to concentrate mainly on the intensity scintillations. The random variations in the other characteristics, especially in the apparent angular positions of stars, are treated in much less detail.

However, in those radar and optical systems which are used to measure the position (and its time derivatives) of both distant and near objects (such as aerospace vehicles) the random tropospheric angle variations assume great importance. For the analysis and synthesis of these systems, it is valuable to accumulate the knowledge on the random tropospheric errors in form of a sufficiently general model.

Such an analytical model of tropospheric random errors in the position measurements and their time derivatives in radar and optical systems has been developed. Among other capabilities this model also permits the prediction of the random tropospheric angle variations for specified sets of tracking situations and system parameters. The prediction is made in terms of minimum and maximum power density spectra (PDS) between which the observed spectra are expected to lie.

The choice of PDS for the characterization of the random errors is necessary because the relation between errors at two points in this system is usually a function of the error frequency (f). The resulting PDS further permit (*i*) subsequent studies of the effects of frequency dependent data processing operations (smoothing, calculation of derivatives, prediction, etc.). (*ii*) detailed comparison with errors from other sources, and (*iii*) application of the optimization methods described by H. W. Bode, C. E. Shannon, and S. Darlington.[3] Values for the more familiar variance (σ^2) or standard deviation (σ) of these random errors at the output of these processes may then be obtained with a straightforward integration of the output PDS (see Section 1.4 below).

Within the model, the predicted PDS of the random tropospheric angle errors are analytically calculated by operating with certain model functions on a model power density spectrum which is given in the range coordinate. This range model PDS represents the pooled data on tropospheric random refraction. It is based upon observations of random variations in the tropospheric refractive index, and in range and phase measurements mainly made at the National Bureau of Standards.[4,5,6,7]

The successful launch of the Communications Satellite Corporation's Early Bird Satellite on April 6, 1965, and its subsequent stabilization in an almost perfect geo-stationary orbit, provided an opportunity to test the model. For this purpose, azimuth and elevation angle measurements on the microwave beacon of the Early Bird satellite were made with the large horn-reflector antenna at the American Telephone and Telegraph (AT&T) Station near Andover, Maine. Most of this ground equipment was previously described in detail.[9,10,11,12,13,14,15] A brief description of the Early Bird satellite may be found in Ref. 16.

The resulting angle error measurements are particularly valuable for comparison with the theoretical model since they are obtained under two unique conditions provided by a geo-stationary satellite as a target. First, the propagation path goes through the entire atmosphere so that the observed angle errors include possible effects of high altitude turbulence, which are impossible to obtain with Earth based targets. Second, the angular tracking rates are negligible relative to the effective wind in the troposphere with which the refractive index anomalies pass through the propagation path.

Thus, the analysis of the angle error measurements and the prediction of the expected random angle errors for the geo-stationary satellite are considerably simplified compared to the analysis and prediction for the more frequent aerospace targets (aircraft, missiles, low satellites) which have large apparent angular velocities and motion disturbed by forces unknown in the necessary detail. The effects of temporal random variations of the refractive index in the Earth's atmosphere on the angle measurements, integrated along the line-of-sight between a ground antenna and a geo-stationary satellite should be observable in an almost pure form.

1.3 *Scope of this Paper*

In this paper, a simplified version of the model of random tropospheric errors is first described, which permits the calculation of the predicted minimum and maximum PDS of the tropospheric angle errors for tracking tasks involving one almost stationary point target and a single observer (single-site radar).

Next, the presented model is used to calculate the numerical values of the predicted PDS of random tropospheric angle errors for the specific tracking situation of the Early Bird observations.

The methods and specific circumstances of data acquisition for one twenty minute period of observations of the Early Bird satellite from the AT&T ground station near Andover, Maine are then described. Another section is concerned with data processing and analysis; it includes the time series of observed azimuth and elevation angles, the calculations of their power density spectra and confidence limits, and estimates of the manual chart reading error and of the effects of thermal receiver noise.

Finally, a comparison is made between the PDS of random tropospheric angle errors predicted with the model and the PDS of the observed random angle variations.

1.4 *Scaling of Power Density Spectra*

In this paper, the random variations of the observed azimuth and elevation angles will be described by their power density spectra (PDS). The numerical computation of the PDS from the time series of data is made by the indirect method described by Blackman and Tukey.[17] However, the scaling of the PDS in this paper deviates from that of Blackman and Tukey by defining the variance σ^2 of the random error as

$$\sigma^2 = \int_0^\infty P\{f\} df. \tag{1}$$

Thus, the PDS $P\{f\}$ is valid only for positive frequencies, $f \geqq 0$. The power spectral density is

$$P\{f\} = \frac{d(\sigma^2)}{df} \tag{2}$$

of the variance contribution $d(\sigma^2)$ to the random error, per unit frequency bandwidth, df, at the frequency, f.

II. PREDICTION OF RANDOM TROPOSPHERIC ANGLE ERRORS

2.1 *Model Concept*

The analytical model of random tropospheric errors in radar and optical systems which has been developed permits the calculation of the power density spectra, and variances of range, phase, range difference, and angle errors, and their time derivatives from a basic pool of model data with the aid of certain model functions. This general model accommodates many different sets of system parameters, and is flexible enough to allow modification for its continuing improvement based upon the analysis of additional data.

During the development of the model the usual lack of sufficient data, and the non-stationarity of the tropospheric refractivity field soon made themselves felt. It was realized therefore, that only an approximate model of the real troposphere could readily be constructed, which necessarily would yield approximate predictions. However, it was found that this approximate model was good enough to allow the useful prediction of tropospheric errors in several interesting cases of tracking system analysis and synthesis.

In the following part of this paper a simplified version of the general model of random tropospheric errors is described, which is limited to the prediction of the power density spectra of the random tropospheric errors in the *angle* measurements made by a *single-site* radar (or radio tracker) on an (almost) *stationary target.*

The particular coordinate of radar measurements selected for the collective description of the pool of basic model data is the slant-range coordinate. In this approach, all available observations of random tropospheric errors are first normalized to certain model conditions, and transformed into the slant-range coordinate. An analytical power density spectrum (PDS) at the lower limit of these normalized and transformed observations is then defined as the model PDS in range, $P_m\{f\}$, where f is the (error-) frequency.

The derivation of the PDS for the random tropospheric angle errors, and for other than the model conditions, is then achieved by processing the range model PDS, $P_m\{f\}$, with certain power gain functions, called the model functions. These model functions depend on such parameters of the tracking situation as the antenna diameter, slant-range and elevation angle of the target, weather, and wind.

It may be noted particularly, that in this simplified version of the model it is not necessary to do any explicit processing in the space domain. Based upon the assumption of an isotropic, and frozen turbulence field of refractive index anomalies in the troposphere, all processing is confined to the frequency $\{f\}$-domain.

The entire model also can be used in an inversion of the computational flow to yield, from new observations, additional information on the basic range model PDS, $P_m\{f\}$, and on the model functions.

2.2 *Assumptions and Limitations*

The simplified analytical model of tropospheric random errors in radar and optical systems is subject to a number of assumptions and limitations:

(*i*) The model is intended to yield tropospheric errors in tracking tasks where one point target is directly observed within the local horizon along a line-of-sight (LOS) by a single observer.

(*ii*) It is assumed that the random errors are small, thus the model functions are linear in the sense of being independent of the magnitude of the errors.

This assumption is justified by the finding that the random errors in the observed quantities (range, angles) have relative magnitudes of only one part in 10^5, or so.

(*iii*) It is assumed that the random errors are stationary during the calculation, or observation of one PDS. The spatial and temporal nonstationarities of the random tropospheric errors are only considered by the introduction of the "global" weather functions. Local anomalies, as well as diurnal and seasonal variations of the tropospheric random errors thus are not separated here. It is believed that more detailed knowledge in this respect is better obtained by direct measurements under the particular local circumstances of actual tracking situations.

(*iv*) It is assumed that the random errors due to the tropospheric anomalies can be treated as if they were caused by the motion of a locally isotropic field of "frozen" turbulence through the line-of-sight with an effective wind speed (**u**) normal to the LOS.

(*v*) The wavelength (λ) of the transmitted electromagnetic waves is assumed to be small, say $\lambda \leqq 10$ [cm], in order to avoid basic theoretical difficulties which are manageable only if $\lambda \ll l$, where l is the characteristic length of the tropospheric anomalies.[1,2] This assumption is also important in order to avoid the effects of random propagation through the ionosphere.

(*vi*) The size of the antenna system is small with respect to the diameter of the earth (flat earth assumption).

(*vii*) The size of the antenna system is small enough to avoid the lack of correlation between the refractivity anomalies at large distances on the surface of the earth.

2.3 *Model Functions*

2.3.1 *Model Power Density Spectrum in Range, $P_m\{f\}$*

The conditions to which the available data[4,5,6,7] on random tropospheric range, phase, and refractive index variations are normalized are:

(*i*) effective tropospheric path length $= L_m = 15$ [km];

(*ii*) effective wind speed normal to LOS $= u_m = 1$ [m/sec];

(*iii*) surface refractivity $= N_m = 10^6\ (n_m - 1) = 313$, this value is the U.S. average,[8] and n_m is the equivalent refractive index;

(*iv*) known effects of variations of the surface refractivity are not corrected during data acquisition.

After transformation to the selected range coordinate the model power density spectrum, $P_m\{f\}$, is then derived as an analytical approximation to the *lower* limit of all observations.

The derived model PDS in range consists of five branches, which are linear in a log (power density) versus log (frequency) plot, namely:

$$P_m\{f\} = \left\{\begin{array}{l} 9.6 \times 10^{+19} f^{+2}\ [\mathrm{m}^2/\mathrm{Hz}] \\ \qquad\qquad \text{for} \qquad\qquad 0 \leqq f \leqq 2.5 \times 10^{-8}\ [\mathrm{Hz}] \\ 1.5 \times 10^{-3} f^{-1}\ [\mathrm{m}^2/\mathrm{Hz}] \\ \qquad\qquad \text{for}\quad 2.5 \times 10^{-8} \leqq f \leqq 1.0 \times 10^{-5}\ [\mathrm{Hz}] \\ 4.7 \times 10^{-11} f^{-2.5}\ [\mathrm{m}^2/\mathrm{Hz}] \\ \qquad\qquad \text{for}\quad 1.0 \times 10^{-5} \leqq f \leqq 1.0 \times 10^{-3}\ [\mathrm{Hz}] \\ 1.5 \times 10^{-12} f^{-3}\ [\mathrm{m}^2/\mathrm{Hz}] \\ \qquad\qquad \text{for}\quad 1.0 \times 10^{-3} \leqq f \leqq 1.0 \times 10^{+2}\ [\mathrm{Hz}] \\ 1.5 \times 10^{-6} f^{-6}\ [\mathrm{m}^2/\mathrm{Hz}] \\ \qquad\qquad \text{for}\quad 1.0 \times 10^{+2} \leqq f \leqq \infty\ [\mathrm{Hz}] \end{array}\right\} \tag{3}$$

where the frequency f is to be inserted in hertz. This PDS is plotted in Fig. 1.

2.3.2 *Angle Scale Function, S_α*

The PDS of random tropospheric angle (α) errors for a single antenna radar are obtained from the range model PDS by operating on $P_m\{f\}$ with the angle scale function, S_α. The derivation of this function is based upon the fact that refractivity anomalies of characteristic length l, or of wavenumber κ, which drift through the LOS with the effective wind speed u_m cause random error components of frequency

$$f = u_m/l = \kappa u_m/2\pi. \tag{4}$$

To simplify the analysis the circular antenna aperture of diameter d is now approximated by an interferometer system of equal angle accuracy and baseline length

$$B = 0.626\ d \tag{5}$$

which lies in the plane of the angle being measured. The angle measurement is thought to be indirectly obtained by a range-difference (or phase-difference) measurement across the effective baseline length B. The tropospheric refractivity anomalies disturb this range-difference measurement to an amount that depends on the characteristic length l and on the antenna diameter d.

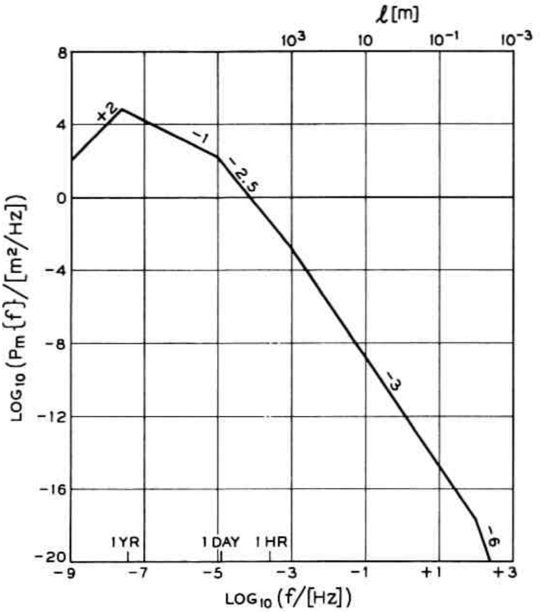


Fig. 1 — Model power density spectrum, $P_m\{f\}$, of tropospheric random errors in the range coordinate versus error-frequency, f, in a log-log plot. $P_m\{f\}$ is valid for the model conditions in Section 2.3.1. The tropospheric anomalies have the characteristic length l.

It is found that for relatively short characteristic lengths

$$l \leqq l_1 = 2d \tag{6}$$

which cause the high error frequencies

$$f \geqq f_1 = u_m/l_1 = 0.5\, u_m/d \tag{7}$$

the random range errors due to these anomalies at the two ends of the effective baseline length B are practically uncorrelated with each other. Thus, they cause a power density of the random error in the range-difference measurement that is twice as large as that of the random error in a single range measurement. Analytically, this finding may be expressed with the aid of a range-difference scale function

$$S_{\Delta R} = 2 \quad \text{for} \quad f \geqq f_1 = 0.5\, u_m/d. \tag{8}$$

The tropospheric refractivity anomalies with characteristic lengths larger than the critical length l_1, namely

$$l \geqq l_1 = 2d \tag{9}$$

cause the lower error frequencies

$$f \leqq f_1 = u_m/l_1 = 0.5\, u_m/d. \tag{10}$$

In this frequency region, the induced random range errors at the two ends of the effective baseline B are more and more correlated as the characteristic length is increased. It is found that with respect to the range-difference errors across B the antenna behaves like a high pass filter with break frequency f_1. Analytically, the resulting reduction in the power density of the low frequency random range-difference errors may be expressed by another branch of the range-difference scale function, namely

$$S_{\Delta R} = 2(f/f_1)^2 \quad \text{for} \quad f \leqq f_1 = 0.5\, u_m/d. \tag{11}$$

The multiplication of the range model PDS, $P_m\{f\}$, with the two branches of $S_{\Delta R}$ in their respective frequency regions would result in a PDS for the random tropospheric range-difference errors across the baseline B under model conditions.[7]

The last step in the derivation of the desired angle scale functions is based upon the assumption of small angular deviations relative to the axis of the antenna system. Then the angle error α is simply related to the range-difference error ΔR and the effective baseline length B by

$$\alpha = \Delta R/B. \tag{12}$$

In terms of power densities, this relation permits the calculation of the angle scale function S_α from the range-difference scale function $S_{\Delta R}$ and B, in general, as

$$S_\alpha = S_{\Delta R}/B^2. \tag{13}$$

The combination of (5), (8), (11), and (13) finally yields the angle scale functions in two branches, namely

$$\left.\begin{aligned} S_\alpha &= 20\ (f/u_m)^2 && \text{for} \quad 0 \leqq f \leqq f_1 \\ S_\alpha &= 5/d^2 && \text{for} \quad f_1 \leqq f \leqq \infty \end{aligned}\right\}. \tag{14}$$

The breakfrequency of the angle scale function is

$$f_1 = 0.5\, u_m/d, \tag{15}$$

where $u_m = 1$ [m/sec] is the model wind speed taken normal to the LOS and in the plane of the angle measurement, and d is the diameter of the circular antenna aperture.

2.3.3 *Aperture Smoothing Function, Φ_α*

The spatial smoothing on tropospheric random error components that are due to refractivity anomalies of small characteristic length ($l \ll d$) is another function of the antenna diameter, d. Here the combined effects of several small refractivity anomalies tend to cancel across the antenna aperture, hence the antenna acts like a low-pass filter in the error-frequency $\{f\}$-domain. A simplified aperture smoothing function which analytically represents this effect is

$$\left.\begin{aligned} \Phi_\alpha &= 1 && \text{for} \quad 0 \leqq f \leqq f_2 \\ \Phi_\alpha &= (f_2/f)^2 && \text{for} \quad f_2 \leqq f \leqq \infty \end{aligned}\right\}, \tag{16}$$

where

$$f_2 = 2\, u_m/d \tag{17}$$

is the breakfrequency for aperture smoothing in angle measurements.

2.3.4 *Effective Path Length Function, Λ*

From theories of propagation through a uniformly turbulent random medium, as for example given by Chernov,[1] it is known that the variance of phase, range and related errors is proportional to the path length. This proportionality holds in both the near-field and the far-field regions of the "scattering" refractivity anomalies of a given characteristic length, l. Therefore, it is possible to account for an effective tropospheric path length, L, which is different from the model path length L_m, by multiplying the power density with the effective path length function

$$\Lambda = L/L_m \quad \text{for} \quad 0 \leqq f \leqq \infty . \tag{18}$$

The required effective tropospheric path length, L, may be calculated by integrating over the geometrical length differentials along the LOS, which are weighted with the square of the local average refractivity at the height of each layer of the atmosphere. The necessary data on the variation of the refractivity with height have been taken from Bean and Thayer.[8] If the height of the target is $h_2 \geqq 10$ [km] above the surface of the Earth, and the apparent elevation angle is $E_a \geqq 3°$, the effective tropospheric path length becomes

$$L = L_o/\sin E_a , \tag{19}$$

where

$$L_o = \left(\frac{N_s}{N_m}\right)^2 (6.61 - 0.01 N_s)[\text{km}] \tag{20}$$

is the effective height of the troposphere for surface refractivities

$$250 \leqq N_s \leqq 450,$$

and N_m is the model surface refractivity.

In this case, which is relevant to many radar and optical tracking tasks, one thus has the effective path length function as

$$\Lambda = \frac{(N_s/N_m)^2}{L_m \cdot \sin E_a} (6.61 - 0.01N_s)[\text{km}] \quad \text{for} \quad 0 \leqq f \leqq \infty. \tag{21}$$

2.3.5 *Weather Functions, W*

As stated in paragraph 2.3.1 the model power density spectrum, $P_m\{f\}$, is defined as the *lower* limit of the available observations normalized to the model conditions. Essentially all (say 99 percent) normalized observations exhibit larger errors than given by P_m. Consequently, all PDS directly derived from $P_m\{f\}$ for other coordinates and tracking situations also would only give the expected *minimum* errors. Since it is frequently desired to state more about the expected distribution of the derived PDS above the expected minimum level we have introduced certain power gain functions, called weather functions, W_q. These are defined as the maximum weather function $W_{\max}\{f\}$ which covers the maximum errors previously observed, and the median weather function $W_{\text{med}}\{f\}$. On a "global" basis (actually only embracing all circumstances of previous observations entered into the model data), it is expected that 50 percent of the measured PDS will lie above and below the PDS predicted with $W_{\text{med}}\{f\}$, and essentially all (say 99 percent) of the measured PDS will lie below the PDS predicted with $W_{\max}\{f\}$.

The maximum weather function derived from available observations[4,5,6,7] is

$$W_{\max}\{f\} = \left\{ \begin{array}{ll} 6 \text{ for} & 0 \leqq f \leqq 1.00 \times 10^{-5} \text{ [Hz]} \\ 1.89 \times 10^{+8} f^{+1.5} & \\ \quad \text{for } 1.00 \times 10^{-5} \leqq f \leqq 2.23 \times 10^{-5} & \text{[Hz]} \\ 20 \text{ for } 2.23 \times 10^{-5} \leqq f \leqq 1.00 \times 10^{-3} & \text{[Hz]} \\ 6.32 \times 10^{+2} f^{+0.5} & \\ \quad \text{for } 1.00 \times 10^{-3} \leqq f \leqq 1.00 \times 10^{-1} & \text{[Hz]} \\ 200 \text{ for } 1.00 \times 10^{-1} \leqq f < \infty & \text{[Hz]} \end{array} \right\} \tag{22}$$

where f is in hertz.

The median weather function is taken as

$$W_{\text{med}}\{f\} = (W_{\text{max}}\{f\})^{0.5}. \tag{23}$$

Both functions are plotted in Fig. 2. Note that $W_{\text{min}} = 1$ by definition.

2.3.6 *Effective Wind Functions, U*

The purpose of the effective wind functions, U, is to introduce other magnitudes of effective wind speed, $u_\alpha \neq u_m$, into the model. The derivation of these effective wind functions rests upon the assumption that an isotropic, frozen turbulence field of refractivity anomalies exists in the troposphere which moves through the LOS with a constant effective wind speed component, u_α, normal to the LOS and in the plane of the angle coordinate α. With this assumption, a given anomaly causes an angle error of a magnitude that is independent of u_α, and of a frequency that is proportional to u_α. It was found that a PDS that is given in $\{f,u_m\}$-space as a sum of branches of the form

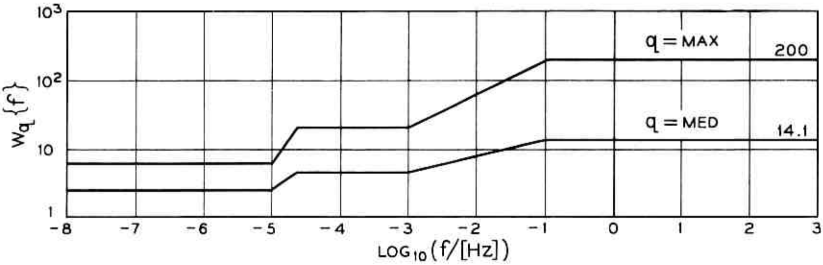


Fig. 2 — Maximum and median weather functions, $W_{\text{max}}\{f\}$ and $W_{\text{med}}\{f\}$, respectively, versus decadic logarithm of the error frequency, f.

$$P_\alpha'\{f,u_m\} = P_o'(f/f_o)^\gamma \quad \text{for} \quad f_{m,\text{min}} \leqq f \leqq f_{m,\text{max}} \tag{24}$$

is transformed into an equivalent PDS in $\{f,u_\alpha\}$-space by the relations

$$\left.\begin{aligned} P_\alpha''\{f,u_\alpha\} &= U_P \cdot P_\alpha'\{f,u_m\} \\ \text{for} \\ U_f \cdot f_{m,\text{min}} = f_{\alpha,\text{min}} &\leqq f \leqq f_{\alpha,\text{max}} = U_f \cdot f_{m,\text{max}} \end{aligned}\right\}, \tag{25}$$

where the effective wind function for the transformation of the power density is

$$U_P = (u_m/u_\alpha)^{\gamma+1} \tag{26}$$

and the effective wind function for the transformation of the frequency regions

$$U_f = u_\alpha / u_m . \tag{27}$$

In these relations the $f_{\alpha,\min}$ and $f_{\alpha,\max}$ are the limits of the frequency region in which P_α'' is valid after the transformation to $\{f,u_\alpha\}$-space.

In the special case of small angular velocity of the LOS, the equivalent wind due to the angular rate is negligible compared to the natural winds in the atmosphere. The effective wind speed is then simply

$$u_\alpha = w_\alpha , \tag{28}$$

where w_α is that component of the natural wind which is normal to the LOS and in the plane of the angle α. In this plane of the angle α, the atmospheric refractivity anomalies, on the average, appear to move through the LOS with the speed w_α.

The calculation of the effective wind speed for azimuth ($\alpha \rightarrow A$) angle errors depends on the geometrical relations between the LOS and the natural average wind vector, $\hat{w}$, Fig. 3. If A is the azimuth angle of the LOS, and δ is the azimuth of the wind vector, their difference

$$\beta = \delta - A \tag{29}$$

can be used to calculate the effective wind speed for azimuth angle errors

$$u_A = | w_A | = | \hat{w} | \cdot | \sin \beta |. \tag{30}$$

With the horizontal LOS component (see Fig. 3)

$$w_1 = | \hat{w} | \cdot \cos \beta \tag{31}$$

the effective wind speed for elevation errors similarly becomes, Fig. 4,

$$u_E = | w_E | = | w_1 | \cdot | \sin E_a | \tag{32}$$

or

$$u_E = | \hat{w} | \cdot | \cos \beta \cdot \sin E_a |. \tag{33}$$

Only the magnitudes of the effective wind speeds are of interest in this special case of small angular velocity of the LOS, and a single-site radar.

2.4 *Computation of the Predicted PDS*

In the prediction of the PDS of random tropospheric angle errors for a particular tracking situation numerical values are inserted for all independent parameters in the model functions given above. The range model PDS, $P_m\{f\}$, is then multiplied by the model functions, within the limits of the stated frequency regions, in the following sequence: angle scale

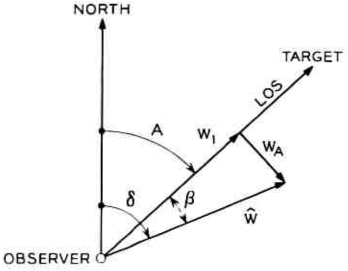


Fig. 3 — Horizontal projection of LOS, wind vector $\hat{w}$, and azimuth angle A.

function, aperture smoothing function, effective path length function, weather function, and effective wind functions.

For the tracking situation of the Early Bird observations on May 7, 1965 the numerical values for the parameters and model functions are given in the Appendix. The operation of these models functions on the range model PDS, $P_m\{f\}$, resulted in four PDS, namely a predicted minimum spectrum and a predicted maximum spectrum for each of the two angle coordinates azimuth and elevation. The resulting PDS $P_{A,\min}$, $P_{A,\max}$ and $P_{E,\min}$, $P_{E,\max}$ are plotted over the interesting frequency range in Figs. 12, and 13.

III. OBSERVATIONS OF RANDOM TROPOSPHERIC ANGLE ERRORS ON THE EARLY BIRD SATELLITE

3.1 *Data Acquisition*

3.1.1 *Method and Equipment*

For the acquisition of the data on random angle variations in azimuth and elevation, the apparent position of the microwave beacon (frequency about 4000 MHz) of the Early Bird satellite was measured with the horn

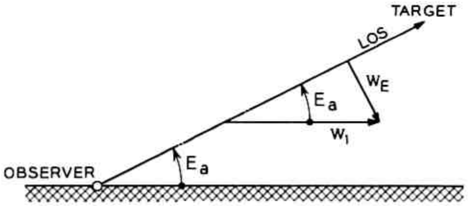


Fig. 4 — Projection of wind vector into vertical plane through LOS with apparent elevation angle E_a.

antenna (aperture diameter $d = 67.7$ ft) and its associated equipment. During these measurements the communications carrier of the satellite was switched off; this resulted in an increase in beacon signal strength to such a level that the thermal receiver noise in the obtained angle measurements was negligible compared to the desired tropospheric random errors.

The signal flow through the major pieces of equipment which were used is illustrated in Fig. 5. After acquisition of the Early Bird satellite beacon in the main beam (beamwidth $\theta = 0.225$ deg) of the horn antenna, the antenna control was turned over to the vernier autotrack system, and the servo loop opened by switching off the hydraulic drive motors. The antenna was now fixed in an orientation indicated by the digital display of the azimuth (A) and elevation (E) angles given in degrees, and derived from digital data pickoff units, which have a precision of encoding[11] of 0.00275[deg].

The satellite now appeared to drift through the fixed horn antenna beam in an irregular motion, which was partially due to motion in its true position (orbit), but also due to the refractive index variations in the intervening atmospheric propagation medium, and possibly other disturbances. The apparent angular position of the satellite relative to the electrical axis of the horn antenna on the ground was determined by the autotrack system which contains angle error sensing and processing equipment. The azimuth and elevation error signals, $\Delta A\{t\}$ and $\Delta E\{t\}$ from the autotrack system were passed through low-pass recording filters before recording either by oscilloscope and camera, or by analog strip chart recorder.

The photographic pictures of the oscilloscope display giving ΔA vs ΔE were only used for inspection. The strip chart recordings giving the $\Delta A\{t\}, \Delta E\{t\}$ time series, however, were used for the more detailed analysis of the data, as described later.

3.1.2 *Propagation Path and Mean Satellite Motion*

During the measurements the propagation path pointed from the horn antenna near Andover, Maine, to the Early Bird satellite approximately at an azimuth angle $A \approx 128.5°$ (southeast), and an elevation angle $E \approx 24.5°$. The slant range between ground antenna and satellite was about 24,300 [statute miles] $\approx$ 39,100 [km]. The terrain surrounding the Earth Station may be described as a shallow bowl of perhaps 10-miles diameter surrounded by hills of up to about 3.5 [deg] elevation.

The mean apparent satellite motion with respect to the azimuth and elevation angles given above consisted of (i) a small linear drift with an

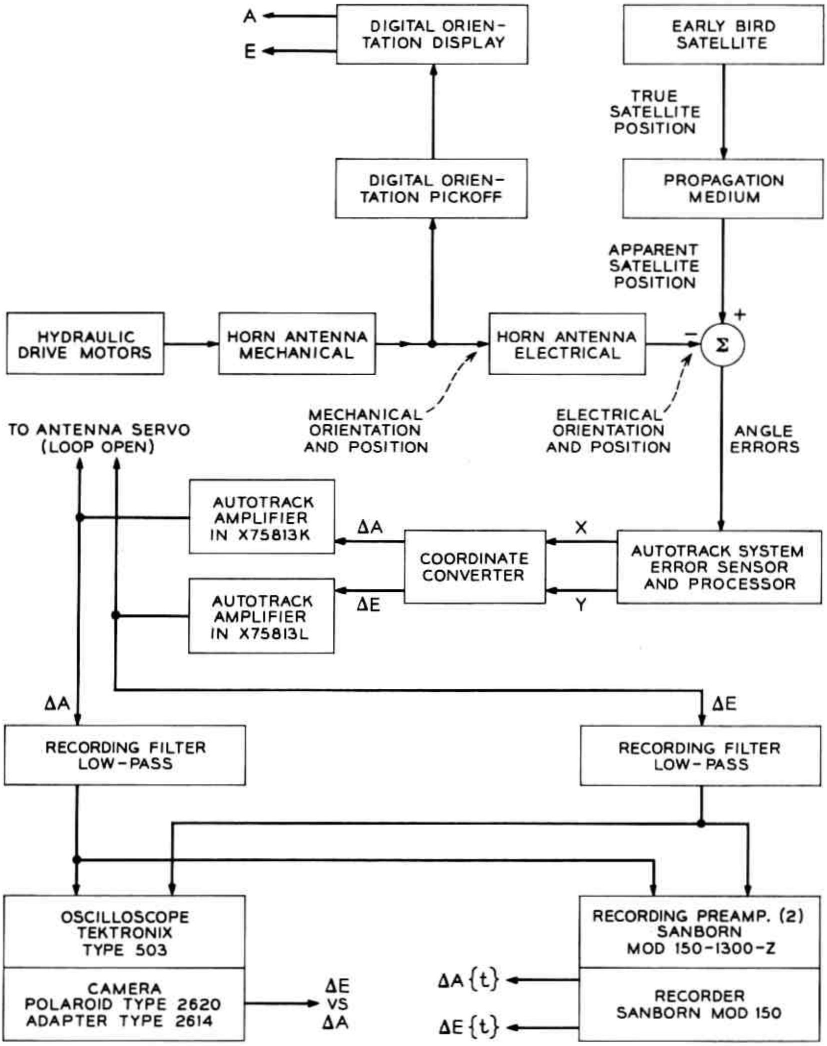

A
E
DIGITAL ORIENTATION DISPLAY
EARLY BIRD SATELLITE
TRUE SATELLITE POSITION
DIGITAL ORIENTATION PICKOFF
PROPAGATION MEDIUM
APPARENT SATELLITE POSITION
+
HYDRAULIC DRIVE MOTORS
HORN ANTENNA MECHANICAL
HORN ANTENNA ELECTRICAL
−
Σ
MECHANICAL ORIENTATION AND POSITION
ELECTRICAL ORIENTATION AND POSITION
TO ANTENNA SERVO (LOOP OPEN)
ANGLE ERRORS
AUTOTRACK AMPLIFIER IN X75813K
ΔA
X
COORDINATE CONVERTER
AUTOTRACK SYSTEM ERROR SENSOR AND PROCESSOR
AUTOTRACK AMPLIFIER IN X75813L
ΔE
Y
ΔA
ΔE
RECORDING FILTER LOW-PASS
RECORDING FILTER LOW-PASS
OSCILLOSCOPE TEKTRONIX TYPE 503
RECORDING PREAMP. (2) SANBORN MOD 150-1300-Z
CAMERA POLAROID TYPE 2620 ADAPTER TYPE 2614
ΔE VS ΔA
ΔA{t}
ΔE{t}
RECORDER SANBORN MOD 150

Fig. 5 — Flow diagram of data acquisition.

azimuth component of $\dot{A}_1 = -2.08 \times 10^{-3}$ [deg/hr], and an elevation component of $\dot{E}_1 = -1.22 \times 10^{-3}$ [deg/hr], plus (*ii*) a diurnal elliptical motion with peak-to-peak amplitudes of $\hat{A} = 0.266$ [deg] in azimuth and $\hat{E} = 0.245$ [deg] in elevation. The net result of these components appears at the Earth Station as a slow motion of the satellite along a helical path seen under an oblique angle. This picture of the mean apparent satellite motion was obtained by plotting the hourly azimuth and elevation angles from the digital display for a few days before the analyzed random angle error data were recorded. The random azimuth and elevation angle errors, ΔA and ΔE, which are subjects of this paper are superimposed on this mean apparent motion.

3.1.3 *Date and Time of Observations*

The random angle error data recorded on strip charts, and analyzed in this paper were taken on May 7, 1965 between about 23 hr:38 min EDT and 23 hr:59 min EDT.

3.1.4 *Weather Conditions*

Weather data at the horn antenna of the Andover Earth Station were not taken. However, the weather data may be estimated from those taken at a private station in nearby Rumford, Me. This estimation yields the following data:[18] Cloud cover 9/10, wind South 19 [statute miles/hr], dry bulb temperature 41.8 [°F], dew point 35 [°F], and pressure 28.5 [inches] = 965.0 [millibars].

3.1.5 *Recording Filters*

The low-pass recording filters mentioned in Section 3.1.1 above were simple two-section RC filters. Since the source impedances feeding these filters are small, and the load impedances connected to their outputs are large compared to the resistances in the RC sections of the filters, their inverse power gain is

$$F = G^{-1} \approx 1 + (\omega T)^2 \, ((\omega T)^2 + 7).$$

In this equation, F is the ratio of input power to output power, $T = RC$ is the time constant of one filter section, and $\omega = 2\pi f$, where f is the frequency.

The power density of the random angle errors before the filters may then be obtained by multiplying the power density of the recorded random angle errors with the inverse power gain F. The filters which were used in these observations allowed a choice between two cutoff frequencies. The results of numerical calculations of the inverse power gains versus frequency for the "LOW", and "HIGH" filters are plotted in Fig. 6.

3.1.6 *Calibration*

The sensitivities of the recorded error voltages (after the recording filters) to errors in the azimuth and elevation angles with respect to the

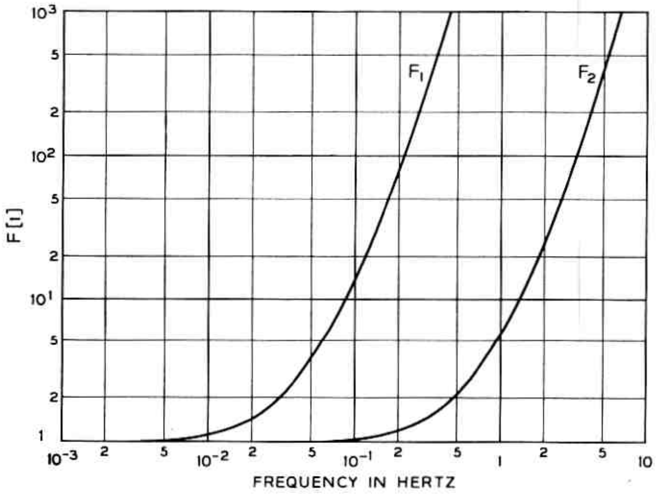


Fig. 6 — Inverse power gain, F, of recording filter versus frequency, f. F_1 for filter in "LOW" range. F_2 for filter in "HIGH" range.

electrical axis of the horn antenna were obtained by direct calibration on the Early Bird satellite. For this purpose, the antenna servo system was disabled, and manual angle offsets were then inserted and their effects on the strip chart records were measured.

3.1.7 *Oscilloscope Displays*

Photographs of oscilloscope displays of the random elevation error (ΔE) versus the simultaneously occurring random azimuth error (ΔA) were also made.

The photo tracing in Fig. 7 was obtained at 22:30 EDT May 7, 1965 while the recording filters were in the "HIGH" range, and the exposure time was five seconds. It is obvious that in this sample of the higher frequency errors the peak-to-peak azimuth variations (about 40 microradians) are considerably larger than those of the elevation errors (20 microradians).

The photo tracing in Fig. 8 was taken at 22:38 EDT on the same date with the recording filters in the "LOW" range. The exposure time was two minutes. In this sample of the lower frequency errors the peak-to-peak azimuth variations (12 microradians) are slightly smaller than the elevation variations (18 microradians).

As mentioned before, these photos were only used for inspection and not for numerical analysis.

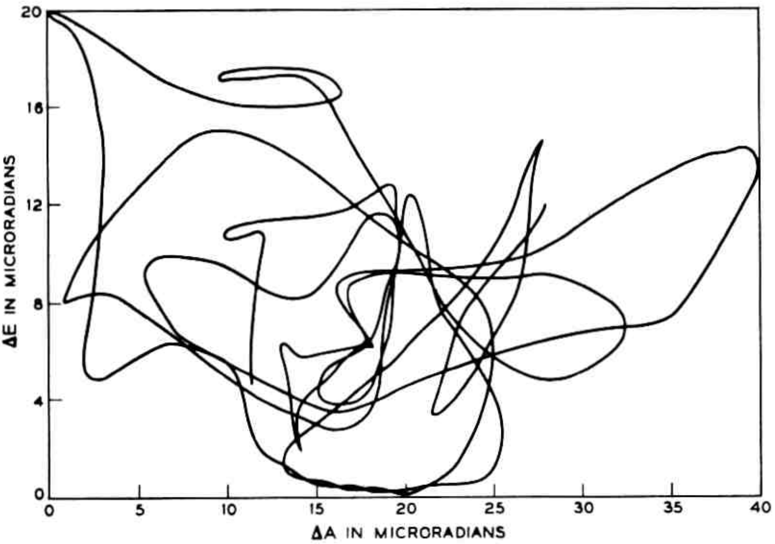


Fig. 7 — Tracing of oscilloscope photograph of random elevation error (ΔE) versus random azimuth error (ΔA). Time is the parameter. Recording filters in "HIGH" range. Exposure time: five seconds.

3.2 *Data Processing and Analysis*

3.2.1 *General Methods and Equipment*

The data on azimuth (ΔA) and elevation angles (ΔE) versus time were recorded on a strip chart recorder with the recording filters in the "LOW" range. The time series of azimuth and elevation angles were manually digitized at two second intervals.

After the manual digitizing process the time series of azimuth and elevation variations were punched into cards for subsequent processing on the 7094 digital computer.

3.2.2 *Time Series of Observed Angle Variations*

The time series of the azimuth (ΔA) and elevation (ΔE) angle variations are shown in Figs. 9 and 10, respectively.

The total observation time was somewhat above twenty minutes. This observation time was limited by the mean apparent drift of the satellite in the fixed antenna beam. This drift resulted in the recording traces going off scale after a certain time.

A total of 720 azimuth data points, and 666 elevation data points were

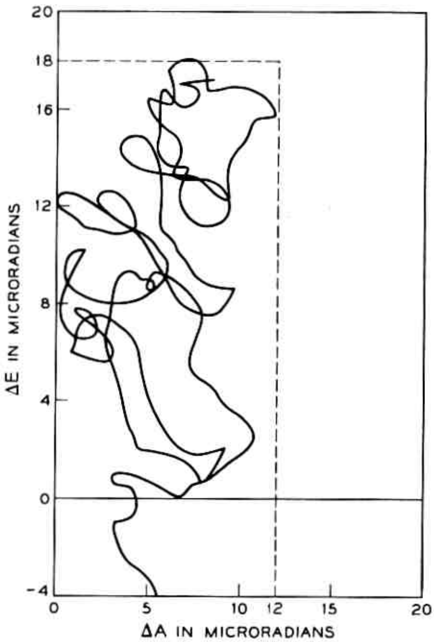


Fig. 8 — Tracing of oscilloscope photograph of random elevation error (ΔE) versus random azimuth error (ΔA). Time is the parameter. Recording filters in "LOW" range. Exposure time: two minutes.

recorded. Due to the systematic drift the elevation record went off the recording scale sooner than the azimuth record.

3.2.3 *Power Density Spectra of Observed Angle Variations*

The random variations of the observed azimuth and elevation angles will also be described by their power density spectra (PDS) for comparison with the predictions. The numerical computation of the PDS from the time series of data is made on a digital computer by the indirect method described by Blackman and Tukey.[17] It proceeded in the following steps: calculation and removal of the mean, and of the linear trend in the series; tapering the first 5 percent (start) and the last 5 percent (end) of the time series with a cosine function; computation of the autocorrelation function versus number r of $0 \leqq r \leqq M$ time lags each of duration of the sampling period Δt; computation of the Fourier transform of the autocorrelation function by a cosine series resulting in a raw power spectrum and subsequent smoothing of the raw spectrum by sliding,

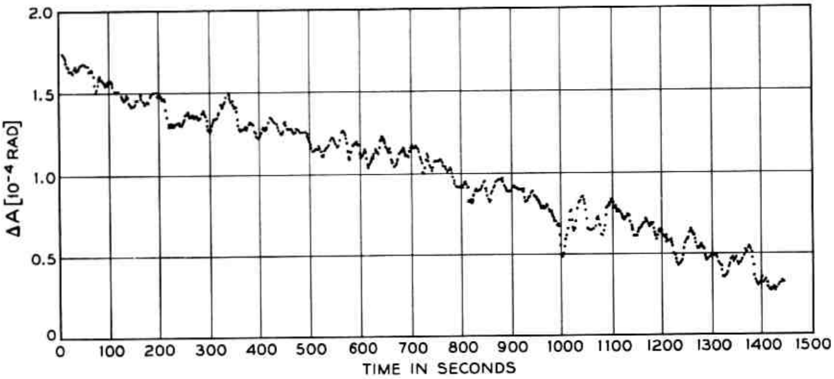


Fig. 9 — Time series of observed azimuth angles, ΔA, versus time, t.

weighted averages of values for three neighboring frequency steps with weights 0.25, 0.50, and 0.25.

The computer program actually calculates a quantity $X'\{f\}$, called "power spectrum", which is related to the usual power density spectrum $P'\{f\}$ by the equation

$$P'\{f\} = X'\{f\}/\Delta f, \tag{34}$$

where

$$\Delta f = f_N/M = 1/(2\cdot\Delta t\cdot M) \tag{35}$$

and

f_N = Nyquist frequency
M = maximum number of lags in autocorrelation
Δt = sampling period.

In this equation, the primed quantities indicate that they still refer to the data at the *output* side of the recording filter. In order to obtain the desired power density spectrum at the *input* of the recording filter, $P'\{f\}$ must be multiplied by the inverse power gain of the filter, $F\{f\}$, yielding

$$P\{f\} = 2\cdot\Delta t\cdot M\cdot F\{f\}\cdot X'\{f\}. \tag{36}$$

Additional smoothing of the power density spectrum is used at the higher error-frequencies, since many cycles of these angle error components have been observed. This is done with a filter of approximately constant relative bandwidth, $\beta = b/f = 0.231$, at the expense of absolute

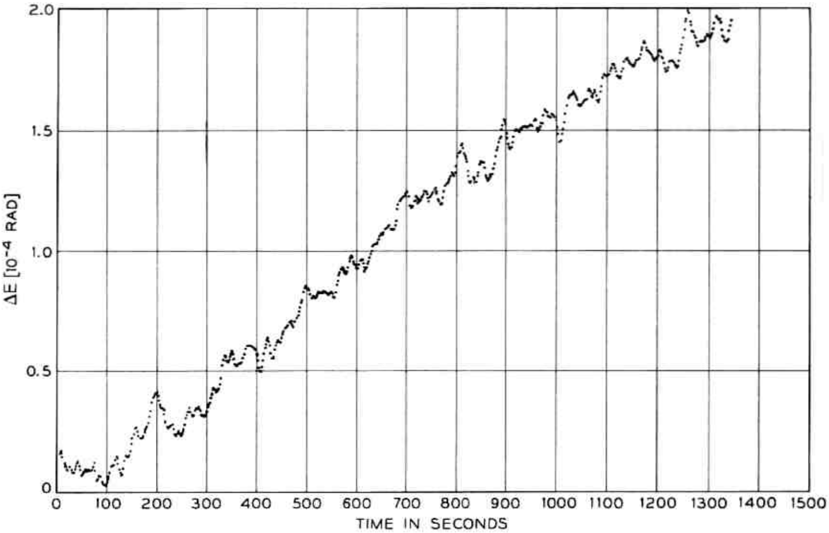


Fig. 10 — Time series of observed elevation angles, ΔE, versus time, t.

frequency resolution. This point will be illuminated again in Section 3.2.4.

The data on azimuth and elevation angle variations given in Section 3.2.2 were analyzed with the methods just described. It was found that the mean linear trends during these observations were in azimuth $+8.1 \times 10^{-8}$ [rad/sec], and in elevation -1.7×10^{-7} [rad/sec]. Even at a distance of 10 [km] along the line-of-sight the magnitude of these angular rates amount to beam sweeping speeds of less than 0.002 [m/sec], which are indeed negligible compared to natural wind speeds in the troposphere.

The power density spectra of the observed azimuth and elevation variations at the input of the recording filters, $P_A\{f\}$ and $P_E\{f\}$, which result from these calculations are plotted in Figs. 12 and 13, respectively. Other spectra also plotted in these figures are explained below.

3.2.4 *Confidence Limits for Power Density Spectra*

In computing confidence limits for the power density spectra it is necessary to distinguish between two error-frequency regions: the low-frequency region in which the absolute analyzing bandwidth of the PDS calculation

$$b = (M \cdot \Delta t)^{-1} \tag{37}$$

is constant, and the high-frequency region in which the relative analyzing bandwidth

$$\beta = b/f \tag{38}$$

is constant.

In the low-frequency region, with constant absolute analyzing bandwidth, b, the number of degrees of freedom in the PDS estimate is approximately

$$k = 2N/M \tag{39}$$

where N is the number of data points observed.

In the high-frequency region, with constant relative analyzing bandwidth, β, the number of degrees of freedom is frequency dependent according to

$$k = 2\beta f N \cdot \Delta t. \tag{40}$$

The confidence limits for the calculated PDS of the observations can now be given, the lower limit being

$$P_1\{f\} = P\{f\}/K_1\{f\} \tag{41}$$

and the upper limit

$$P_2\{f\} = K_2\{f\} \cdot P\{f\}, \tag{42}$$

where $P\{f\}$ is the calculated PDS, and $K_{1,2}\{f\}$ are the confidence factors.

For a confidence level of $\rho = 95$ percent the *upper confidence factor* is approximately

$$K_2 = 1 + \frac{2.77}{\sqrt{k-1}} + \frac{1.30}{k-1}, \tag{43}$$

the *total confidence factor* (here only used as an intermediate to obtain K_1)

$$K_{21} = \text{antilog}_{10}\left(\frac{2.40}{\sqrt{k-1}}\right), \tag{44}$$

and the *lower confidence factor*

$$K_1 = K_{21}/K_2, \tag{45}$$

where k is the number of degrees of freedom given above. For $k \geqq 5$ the stated analytical approximations for the confidence factors have less than 10 percent error.

The resulting numerical values for the upper (K_2) and lower (K_1) confidence factors are plotted in Fig. 11. The results of calculating the

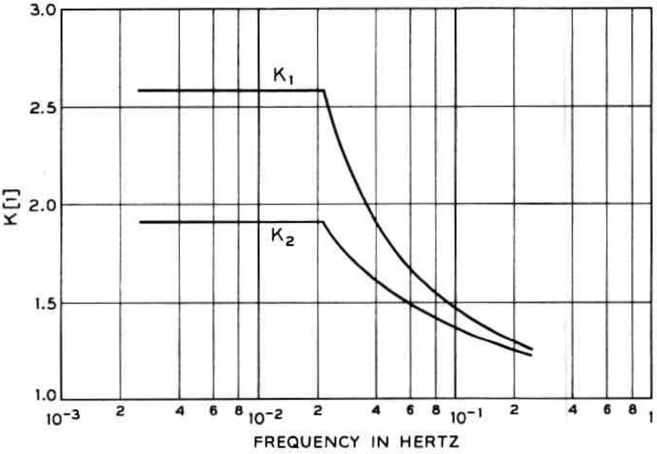


Fig. 11 — Relative confidence factors, K_1 and K_2, for $\rho = 95$ percent confidence level versus frequency, f.

upper confidence limits (P_{A2}, P_{E2}), and the lower confidence limits (P_{A1}, P_{E1}) for the azimuth (P_A) and elevation (P_E) spectra are plotted in Figs. 12 and 13.

3.2.5 *Chart Reading Error*

The errors which are introduced into the data by the manual reading of strip chart records (digitizing) are of the same type as quantization errors. The variance due to a given quantization step size (q) is known to be[19]

$$\sigma_q^2 = q^2/12. \tag{46}$$

If it is now assumed that the quantization noise, which causes this variance, is sharply bandlimited white noise of constant power density (P_q'), and with a cutoff frequency equal to the folding frequency of the digitized time series (f_N), then one also has the variance as

$$\sigma_q^2 = \int_0^{f_N} P_q' \cdot df = P_q' \cdot f_N . \tag{47}$$

Consequently, the noise power density due to the manual chart reading is

$$P_q' = \sigma_q^2/f_N = q^2 \cdot \Delta t/6. \tag{48}$$

As before, this primed power density is taken at the output side of the

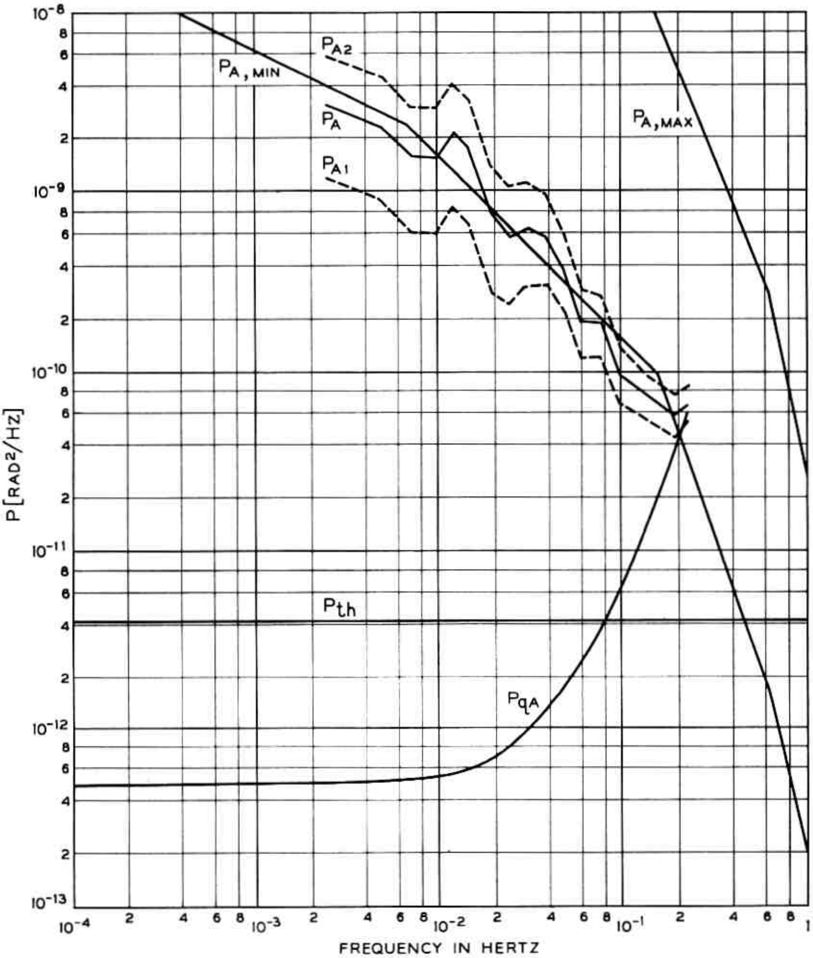


Fig. 12 — Power density of random azimuth angle errors, P, versus frequency, f. P_A, P_{A1}, P_{A2} = observed PDS and 95 percent confidence limits; $P_{A,\min}$, $P_{A,\max}$ = predicted tropospheric PDS limits; P_{th} = thermal receiver noise; P_{qA} = manual chart reading error.

low-pass recording filter. In order to obtain the power density spectrum of the chart reading error referred to the input of the recording filter, it is necessary to multiply P_q' with the inverse power gain $F_1\{f\}$ of the filter, see Section 3.1.5, which yields here for the azimuth coordinate

$$P_{qA}\{f\} = F_1\{f\} \cdot P_{qA}' = F_1\{f\} \cdot q_A^2 \cdot \Delta t/6 \tag{49}$$

and for elevation

$$P_{qE}\{f\} = F_1\{f\} \cdot P_{qE}' = F_1\{f\} \cdot q_E^2 \cdot \Delta t/6. \tag{50}$$

The effective quantization step size for azimuth was $q_A \approx 1.2$ microradians, and for elevation $q_E \approx 0.88$ microradians, the difference being due to different scale factors in the two channels. The sampling period as stated before was $\Delta t = 2$ seconds. The resulting PDS of the manual digitizing process, P_{qA} and P_{qE}, are also plotted in Figs. 12 and 13.

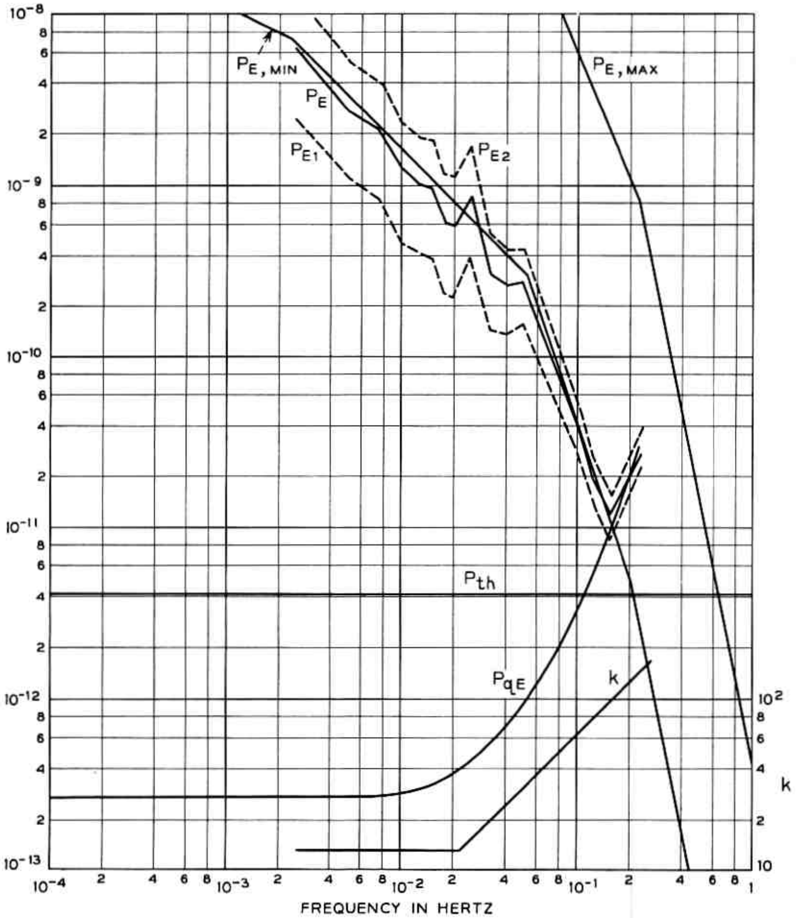


Fig. 13 — Power density of random elevation angle errors, P, versus frequency, f. P_E, P_{E1}, P_{E2} = observed PDS and 95 percent confidence limits; $P_{E,\min}$, $P_{E,\max}$ = predicted tropospheric PDS limits; P_{th} = thermal receiver noise; P_{qE} = manual chart reading error; k = number of degrees of freedom.

3.2.6 *Thermal Angle Errors*

During observations of the tropospheric random angle errors it is important to keep angle errors due to thermal receiver noise at a comparatively low level. The variance of thermal angle errors may be obtained[20] as

$$\sigma_{th}^2 = \frac{\theta^2\left(1 + \frac{S}{N}\right)}{8\left(\frac{S}{N}\right)^2 B\tau} \tag{51}$$

which can be reduced for $S/N \gg 1$ to

$$\sigma_{th}^2 = \frac{\theta^2}{8B\tau(S/N)}, \tag{52}$$

where

θ = antenna beamwidth
S/N = input signal-to-noise power ratio
B = receiver bandwidth
τ = post-detection integration time.

In order to derive the power density (P_{th}) of the white thermal noise spectrum it is first recognized that the variance of the thermal angle error is also

$$\sigma_{th}^2 = \int_0^\infty \frac{P_{th}df}{1 + (f/f_c)^2}, \tag{53}$$

where

$$f_c = 1/2\tau \tag{54}$$

is the cut-off frequency of the post-detection low-pass filter. Equation (53) may be integrated with the substitution $x = f/f_c$; $df = f_c\,dx$ giving

$$\sigma_{th}^2 = P_{th}\cdot f_c\cdot \text{arc tan }(f/f_c)\Big|_0^\infty \tag{55}$$

or

$$\sigma_{th}^2 = \frac{\pi}{2} f_c\cdot P_{th}\,. \tag{56}$$

Combining (52), (54), and (56) then yields the desired density of the thermal angle noise as

$$P_{th} = \frac{\theta^2}{2\pi B(S/N)} \tag{57}$$

independent of the frequency f.

During the observations on May 7, 1965, which are analyzed in this report, the measured signal-to-noise ratio was $(S/N)' = 23$ [dB] = 200[1] while the communications carrier of the Early Bird satellite was switched off. (This ratio was $(S/N)' = 13$ [dB] = 20 [1] due to a weaker beacon signal when the carrier was on.) The primed signal-to-noise ratios stated here are referred to a 3-kHz bandwidth. The effective noise bandwidth, however, is considerably lower due to the employment of a phase-locked tracking loop quite like the one described in Ref. 15. From Fig. 12 in that reference it is seen that the noise bandwidth for

$$(S/N)' = 23 \text{ [dB]} \quad \text{is} \quad B = 390 \text{ [Hz]},$$

which further results in an effective signal-to-noise ratio

$$S/N = 200 \times (3{,}000/390) = 1{,}538 \text{ [1]}.$$

Since the antenna beamwidth was $\theta = 0.225° = 3.94 \times 10^{-3}$ [rad] the desired power density of the thermal receiver noise with (57) here becomes $P_{th} = 4.1 \times 10^{-12}$ [rad^2/Hz] while the communications carrier is switched off, and the beacon signal is strong. This thermal noise level in the angle measurements was low enough to permit observation of the random tropospheric angle variations up to frequencies of a few 0.1 [Hz], see Figs. 12 and 13.

IV. COMPARISON BETWEEN PREDICTED AND OBSERVED ANGLE ERRORS

The comparison between predicted, and observed power density spectra of random tropospheric angle variations may now be made with the aid of Figs. 12 and 13 into which all relevant spectra have been entered. The observed spectra (P_A, P_E) resulted from the analysis of random angle error data taken on the Andover Horn to Early Bird path on May 7, 1965 between about 23 hr:38 min EDT, and 23 hr:59 min EDT. An inspection of Figs. 12 and 13 shows that the PDS of the observations cover about two decades of frequency, namely

$$2.50 \times 10^{-3} \text{ [Hz]} \leqq f \leqq 0.25 \text{ [Hz]}.$$

The comparison of the observed PDS $(P_A; P_E)$ with their respective predicted PDS $(P_{A,\min}, P_{A,\max}; P_{E,\min}, P_{E,\max})$ yields almost identical results for the two angle oocrdinates azimuth (A) and elevation (E). In particular it is found that the PDS of the observed random angle variations (P_A ; P_E), within their respective 95 percent confidence limits (P_{A1}, P_{A2} ; P_{E1}, P_{E2}), lie almost exactly on the predicted *minimum* power spectra ($P_{A,\min}$; $P_{E,\min}$) for random *tropospheric* angle errors.

Thus, the observed PDS match the predicted PDS quite well in the *shape* of their frequency dependence. The *low level* of the observed PDS relative to the predicted range of PDS is thought to be due to the "good tracking weather" at the Andover, Maine site and at the particular time of observation (a quiet night). It must be remembered here that the prediction is based mainly upon the NBS range and phase measurements[4,5,6,7] which were obtained in Hawaii and Colorado. Whether the low level of the random tropospheric errors observed in Maine is a permanent property of the site, or a chance occurrence can be decided by the analysis of additional observations.

It is also possible to compare the observed PDS of the azimuth errors with that of the elevation errors. It is found that the azimuth errors here have a higher level at frequencies $f > 0.1$ Hz than the elevation errors; this is an effect of the higher azimuth wind speed component ($u_A = 6.7$ m/sec versus $u_E = 2.2$ m/sec). Even larger differences between the azimuth and elevation random errors are expected when their effective wind speed components differ by larger amounts. Such wind speed differences may be caused by either peculiar orientation of the natural wind vector relative to the line-of-sight, or also by differences in angular tracking rates.

Near the high-frequency end of the covered band the observed PDS deviate significantly in shape from the predictions. This deviation is particularly evident in the steep increase of the observed elevation PDS above $f = 0.15$ [Hz]. This increase is identified as an effect of quantization errors in the manual digitizing of the analog strip chart records. The transformation of these digitizing errors to the input side of the recording filters results in the steeply rising PDS (P_{qA}, P_{qE}) for these frequencies.

In the frequency band of the observations the angle errors due to thermal receiver noise have a PDS (P_{th}) which is negligible compared to that of the tropospheric angle errors, provided the communications carrier in the Early Bird satellite is turned off.

It is also possible to integrate the predicted PDS of the random tropospheric angle errors over the entire error frequency band, and then to take the square root to obtain the standard deviation

$$\sigma = \left(\int_0^\infty P df \right)^{\frac{1}{2}}.$$

When these integrals are calculated for the predicted minimum and maximum PDS, it is found that the standard deviations of the random tropospheric angle errors are expected to lie between $\sigma_{\min} \approx 10$ [microradians] ≈ 2 [seconds of arc] and $\sigma_{\max} \approx 65$ [microradians] ≈ 13 [seconds of arc]. This range of values compares quite well with Kennedy and Rosson's estimate that the tropospheric angle errors lie between 20 to 50 microradians.[20]

The standard deviation of the expected tropospheric angle fluctuations versus baseline length was previously calculated from NBS data on refractivity and range variations by D. K. Barton.[21] For the equivalent baseline length of the Andover horn antenna of about forty feet, Barton's graph shows a standard deviation of perhaps seventy microradians, a value slightly above our predicted maximum.

Some astronomical observations of random fluctuations in angular star positions, as quoted by Tatarski,[2] show standard deviations of one half to one second of arc. These observations have been made under conditions quite different from those for which our predictions are valid, namely with visible light, in clear weather, with smaller apertures, and probably only over a small fraction of the entire error frequency band. Therefore, it is not too surprising to find that these astronomical measurements lie below our minimum prediction.

Within the limitations of the analyzed observations, and of the described model it is concluded that the observed random angle variations are essentially due to random variations of the refractive index field in the troposphere. The feasibility of acquiring additional data on tropospheric angle errors with the Andover horn antenna on geo-stationary satellites of the Early Bird type therefore was also demonstrated. These data may now be obtained on a routine basis with available and operating equipment.

The comparison of the observations given in this paper with the prediction of random tropospheric angle errors gives some confidence in the described analytical model. Additional observations of random tropospheric angle errors were made with radar and optical equipment over other propagation paths. The comparison of these observations with the relevant predictions from the analytical model (not reported here) are also satisfactory, and have further strengthened the confidence in the model.

V. SUMMARY

Earth-based radar and optical systems which are used to measure the position (and its time derivatives) of both distant and near objects are ultimately limited in accuracy by random angle variations caused by fluctuations of the tropospheric refractive index. For the analysis and synthesis of these systems an analytical model of the random tropospheric errors has been developed.

With this model, the predicted minimum and maximum power density spectra (PDS) between which observed PDS of tropospheric errors are expected to lie can be analytically calculated. The calculation is performed by operating with certain model functions, which depend on the tracking system parameters, on a model PDS (P_m) given in the range coordinate. P_m has been derived from observations of random variations in the tropospheric refractive index, and in range and phase measurements made mainly at the National Bureau of Standards.

A simplified analytical model of random tropospheric angle errors is described here, which is applicable to a tracking situation involving one (almost stationary) target and a single observer. This model is also used to predict the minimum and maximum PDS of the random tropospheric azimuth and elevation angle errors for microwave observations of the Early Bird geo-stationary communication satellite with the large horn-reflector antenna at the AT&T ground station near Andover, Maine.

The general method of data acquisition, Fig. 5, and the specific circumstances of some actual observations on the Early Bird satellite with the Andover horn are then described. Microwave azimuth and elevation angle measurements for an observation time of about twenty minutes were taken on May 7, 1965, while the Early Bird satellite appeared at an elevation angle of about 24.5 degrees.

The analysis of the obtained time series of azimuth and elevation angles results in power density spectra (P_A and P_E) and associated confidence limits which represent the observed random angle variations, see also Figs. 12 and 13. The effect of manual chart reading errors on the observed PDS was also studied. It was shown to consist of a steep increase in the PDS at the high frequency end. The effect of thermal receiver noise on the observed random angle variations was kept at a negligible level.

The comparison of the predicted PDS of the random tropospheric angle errors for the Early Bird observations with the observed PDS leads to the conclusion that the observed random angle variations are indeed caused by the troposphere. In particular it is found that the PDS of the azimuth and elevation observations (P_A and P_E in Figs. 12 and 13), within their respective confidence limits (P_{A1}, P_{A2}; P_{E1}, P_{E2}), lie almost exactly on the predicted *minimum* power density spectra ($P_{A,\,\mathrm{min}}$; $P_{E,\,\mathrm{min}}$) for random tropospheric angle errors.

The feasibility of acquiring additional data on tropospheric propagation effects, especially random angle errors, with the Andover horn antenna on geo-stationary satellites of the Early Bird type, therefore, was also demonstrated.

VI. ACKNOWLEDGMENTS

The author gratefully acknowledges the assistance by M. B. Allen, Jr. and W. A. Klute of Bell Laboratories, as well as that by T. F. Palmatier and his personnel of the AT&T Long Lines Department in Andover, Me. in the acquisition of the angle measurements. The manual digitizing of the analog strip chart records was performed with the help of Mrs. C. Unger. Valuable advice and encouragement for this publication were received from C. A. Armstrong, P. V. Dimock, and M. P. Wilson.

APPENDIX

Numerical Calculation of Predicted PDS

The *parameters* of the tracking situation during the Early Bird observations on May 7, 1965, which permit the prediction of the tropospheric angle PDS with the described model are:

transmission frequency $= f_t = 4137.86$ [MHz]
antenna diameter $= d = 67.7$ [ft] $= 20.6$ [m]
beamwidth $= \theta = 0.225$ [deg]
apparent elevation angle $= E_a \approx 24.5$ [deg]
azimuth angle $= A \approx 128.5$ [deg]
altitude of horn antenna $= h_1 = 900$ [ft] $= 274$ [m]
altitude of satellite $= h_2 \approx 22{,}200$ [st. mi.] $\approx 35{,}700$ [km]
slant-range $= R_{12} \approx 24{,}300$ [st. mi.] $\approx 39{,}100$ [km]
wind vector: $|\hat{w}| = 19$ [st. mi/hr]; $\delta = 0°$
surface refractivity $= N_s = 301$

With these parameters the *model functions* for this tracking situation are calculated as follows.

Breakfrequency of the angle scale function, (15):

$$f_1 = 2.43 \times 10^{-2} \text{ [Hz]}.$$

Angle scale function, (14):

$$S_\alpha = 20\,(f/\text{Hz})^2(1/\text{m}^2] \quad \text{for} \quad 0 \leqq f \leqq 2.43 \times 10^{-2} \text{ [Hz]}$$

$$S_\alpha = 1.18 \times 10^{-2}\,[1/\text{m}^2] \quad \text{for} \quad 2.43 \times 10^{-2} \text{ [Hz]} \leqq f \leqq \infty.$$

Breakfrequency for aperture smoothing, (17):

$$f_2 = 9.71 \times 10^{-2} \text{ [Hz]}.$$

Aperture smoothing function, (16):

$$\Phi_\alpha = 1 \quad \text{for} \quad 0 \leqq f \leqq 9.71 \times 10^{-2} \text{ [Hz]}$$

$$\Phi_\alpha = 9.43 \times 10^{-3} \, (f/\text{Hz})^{-2} \quad \text{for} \quad 9.71 \times 10^{-2} \text{ [Hz]} \leqq f \leqq \infty.$$

Effective tropospheric path length, (19) and (20):

$$L = 8.02 \text{ [km]}.$$

Effective path length function, (18):

$$\Lambda = 0.534 \quad \text{for} \quad 0 \leqq f \leqq \infty.$$

Weather functions:

minimum: $W_{\min} = 1$ (by definition of P_m)

maximum: $W_{\max}$ as per (22).

Effective wind speed, azimuth, (30):

$$u_A = 14.9 \text{ [st. mi/hr]} = 6.7 \text{ [m/sec]}.$$

Effective wind speed, elevation, (33):

$$u_E = 4.9 \text{ [st. mi/hr]} = 2.2 \text{ [m/sec]}.$$

Effective wind function for transformation of the power density, azimuth, (26):

$$U_{PA} = 0.149^{\gamma+1}.$$

Effective wind function for transformation of the frequency regions, azimuth, (27):

$$U_{fA} = 6.7.$$

Effective wind function for transformation of the power density, elevation, (26):

$$U_{PE} = 0.455^{\gamma+1}.$$

Effective wind function for transformation of the frequency regions, elevation, (27):

$$U_{fE} = 2.2.$$

The operation with these model functions on the range model PDS $P_m\{f\}$ results in the following four *predicted PDS of tropospheric random angle errors*.

Minimum PDS in azimuth:

$$P_{A,\min} = \left\{\begin{array}{l} 7.57 \times 10^{+16}\ (f/\mathrm{Hz})^{+4}\ [\mathrm{rad}^2/\mathrm{Hz}] \\ \qquad \text{for} \qquad 0 \leqq f \leqq 1.68 \times 10^{-7}\ [\mathrm{Hz}] \\ 3.55 \times 10^{-4}\ (f/\mathrm{Hz})^{+1}\ [\mathrm{rad}^2/\mathrm{Hz}] \\ \qquad \text{for} \quad 1.68 \times 10^{-7} \leqq f \leqq 6.70 \times 10^{-5}\ [\mathrm{Hz}] \\ 1.95 \times 10^{-10}\ (f/\mathrm{Hz})^{-0.5}\ [\mathrm{rad}^2/\mathrm{Hz}] \\ \qquad \text{for} \quad 6.70 \times 10^{-5} \leqq f \leqq 6.70 \times 10^{-3}\ [\mathrm{Hz}] \\ 1.60 \times 10^{-11}\ (f/\mathrm{Hz})^{-1}\ [\mathrm{rad}^2/\mathrm{Hz}] \\ \qquad \text{for} \quad 6.70 \times 10^{-3} \leqq f \leqq 1.63 \times 10^{-1}\ [\mathrm{Hz}] \\ 4.25 \times 10^{-13}\ (f/\mathrm{Hz})^{-3}\ [\mathrm{rad}^2/\mathrm{Hz}] \\ \qquad \text{for} \quad 1.63 \times 10^{-1} \leqq f \leqq 6.51 \times 10^{-1}\ [\mathrm{Hz}] \\ 1.81 \times 10^{-13}\ (\mathrm{f}/\mathrm{Hz})^{-5}\ [\mathrm{rad}^2/\mathrm{Hz}] \\ \qquad \text{for} \quad 6.51 \times 10^{-1} \leqq f \leqq 6.70 \times 10^{+2}\ [\mathrm{Hz}] \\ 5.47 \times 10^{-5}\ (f/\mathrm{Hz})^{-8}\ [\mathrm{rad}^2/\mathrm{Hz}] \\ \qquad \text{for} \quad 6.70 \times 10^{+2} \leqq f \leqq \infty\ [\mathrm{Hz}] \end{array}\right\}$$

Maximum PDS in azimuth:

$$P_{A,\max} = \left\{\begin{array}{l} 4.54 \times 10^{+17}\ (f/\mathrm{Hz})^{+4}\ [\mathrm{rad}^2/\mathrm{Hz}] \\ \qquad \text{for} \qquad 0 \leqq f \leqq 1.68 \times 10^{-7}\ [\mathrm{Hz}] \\ 2.13 \times 10^{-3}\ (f/\mathrm{Hz})^{+1}\ [\mathrm{rad}^2/\mathrm{Hz}] \\ \qquad \text{for} \quad 1.68 \times 10^{-7} \leqq f \leqq 1.49 \times 10^{-4}\ [\mathrm{Hz}] \\ 3.90 \times 10^{-9}\ (f/\mathrm{Hz})^{-0.5}\ [\mathrm{rad}^2/\mathrm{Hz}] \\ \qquad \text{for} \quad 1.49 \times 10^{-4} \leqq f \leqq 1.63 \times 10^{-1}\ [\mathrm{Hz}] \\ 1.03 \times 10^{-10}\ (f/\mathrm{Hz})^{-2.5}\ [\mathrm{rad}^2/\mathrm{Hz}] \\ \qquad \text{for} \quad 1.63 \times 10^{-1} \leqq f \leqq 6.51 \times 10^{-1}\ [\mathrm{Hz}] \\ 4.38 \times 10^{-11}\ (f/\mathrm{Hz})^{-4.5}\ [\mathrm{rad}^2/\mathrm{Hz}] \\ \qquad \text{for} \quad 6.51 \times 10^{-1} \leqq f \leqq 6.70 \times 10^{-1}\ [\mathrm{Hz}] \\ 3.61 \times 10^{-11}\ (f/\mathrm{Hz})^{-5}\ [\mathrm{rad}^2/\mathrm{Hz}] \\ \qquad \text{for} \quad 6.70 \times 10^{-1} \leqq f \leqq 6.70 \times 10^{+2}\ [\mathrm{Hz}] \\ 1.09 \times 10^{-2}\ (f/\mathrm{Hz})^{-8}\ [\mathrm{rad}^2/\mathrm{Hz}] \\ \qquad \text{for} \quad 6.70 \times 10^{+2} \leqq f \leqq \infty\ [\mathrm{Hz}] \end{array}\right\}$$

Minimum PDS in elevation:

$$P_{E,\min} = \begin{cases} 2.01 \times 10^{+19}\ (f/\mathrm{Hz})^{+4}\ [\mathrm{rad}^2/\mathrm{Hz}] & \\ \qquad \text{for} \qquad 0 \leqq f \leqq 5.50 \times 10^{-8}\ [\mathrm{Hz}] & \\ 3.31 \times 10^{-3}\ (f/\mathrm{Hz})^{+1}\ [\mathrm{rad}^2/\mathrm{Hz}] & \\ \qquad \text{for} \quad 5.50 \times 10^{-8} \leqq f \leqq 2.20 \times 10^{-5}\ [\mathrm{Hz}] & \\ 3.41 \times 10^{-10}\ (f/\mathrm{Hz})^{-0.5}\ [\mathrm{rad}^2/\mathrm{Hz}] & \\ \qquad \text{for} \quad 2.20 \times 10^{-5} \leqq f \leqq 2.20 \times 10^{-3}\ [\mathrm{Hz}] & \\ 1.60 \times 10^{-11}\ (f/\mathrm{Hz})^{-1}\ [\mathrm{rad}^2/\mathrm{Hz}] & \\ \qquad \text{for} \quad 2.20 \times 10^{-3} \leqq f \leqq 5.35 \times 10^{-2}\ [\mathrm{Hz}] & \\ 4.57 \times 10^{-14}\ (f/\mathrm{Hz})^{-3}\ [\mathrm{rad}^2/\mathrm{Hz}] & \\ \qquad \text{for} \quad 5.35 \times 10^{-2} \leqq f \leqq 2.14 \times 10^{-1}\ [\mathrm{Hz}] & \\ 2.09 \times 10^{-15}\ (f/\mathrm{Hz})^{-5}\ [\mathrm{rad}^2/\mathrm{Hz}] & \\ \qquad \text{for} \quad 2.14 \times 10^{-1} \leqq f \leqq 2.20 \times 10^{+2}\ [\mathrm{Hz}] & \\ 2.22 \times 10^{-8}\ (f/\mathrm{Hz})^{-8}\ [\mathrm{rad}^2/\mathrm{Hz}] & \\ \qquad \text{for} \quad 2.20 \times 10^{+2} \leqq f \leqq \infty\ [\mathrm{Hz}] & \end{cases}$$

Maximum PDS in elevation:

$$P_{E,\max} = \begin{cases} 1.21 \times 10^{+20}\ (f/\mathrm{Hz})^{+4}\ [\mathrm{rad}^2/\mathrm{Hz}] & \\ \qquad \text{for} \qquad 0 \leqq f \leqq 5.50 \times 10^{-8}\ [\mathrm{Hz}] & \\ 1.99 \times 10^{-2}\ (f/\mathrm{Hz})^{+1}\ [\mathrm{rad}^2/\mathrm{Hz}] & \\ \qquad \text{for} \quad 5.50 \times 10^{-8} \leqq f \leqq 4.91 \times 10^{-5}\ [\mathrm{Hz}] & \\ 6.81 \times 10^{-9}\ (f/\mathrm{Hz})^{-0.5}\ [\mathrm{rad}^2/\mathrm{Hz}] & \\ \qquad \text{for} \quad 4.91 \times 10^{-5} \leqq f \leqq 5.35 \times 10^{-2}\ [\mathrm{Hz}] & \\ 1.95 \times 10^{-11}\ (f/\mathrm{Hz})^{-2.5}\ [\mathrm{rad}^2/\mathrm{Hz}] & \\ \qquad \text{for} \quad 5.35 \times 10^{-2} \leqq f \leqq 2.14 \times 10^{-1}\ [\mathrm{Hz}] & \\ 8.91 \times 10^{-13}\ (f/\mathrm{Hz})^{-4.5}\ [\mathrm{rad}^2/\mathrm{Hz}] & \\ \qquad \text{for} \quad 2.14 \times 10^{-1} \leqq f \leqq 2.20 \times 10^{-1}\ [\mathrm{Hz}] & \\ 4.17 \times 10^{-13}\ (f/\mathrm{Hz})^{-5}\ [\mathrm{rad}^2/\mathrm{Hz}] & \\ \qquad \text{for} \quad 2.20 \times 10^{-1} \leqq f \leqq 2.20 \times 10^{+2}\ [\mathrm{Hz}] & \\ 4.43 \times 10^{-6}\ (f/\mathrm{Hz})^{-8}\ [\mathrm{rad}^2/\mathrm{Hz}] & \\ \qquad \text{for} \quad 2.20 \times 10^{+2} \leqq f \leqq \infty\ [\mathrm{Hz}] & \end{cases}$$

REFERENCES

1. Chernov, L. A., *Wave Propagation in a Random Medium*, McGraw-Hill Book Co., New York, 1960.
2. Tatarski, V. I., *Wave Propagation in a Turbulent Medium*, McGraw-Hill Book Co., New York, 1961.
3. Bode, H. W. and Shannon, C. E., A Simplified Derivation of Linear Least Squares Smoothing and Prediction Theory, Proc. IRE, April, 1950, pp. 417–425.
   Darlington, S., Linear Least-Squares Smoothing and Prediction with Applications, B.S.T.J., *37*, September, 1958, pp. 1221–1294.
4. Thompson, M. C., Jr. and Janes, H. B., Measurements of Phase Stability Over a Low-Level Tropospheric Path, J. Res. NBS, *63D*, 1, July–August, 1959, pp. 54–51.
5. Thompson, M. C., Jr., Janes, H. B., and Kirkpatrick, A. W., An Analysis of Time Variations in Tropospheric Refractive Index and Apparent Radio Path Length, J. Geophys. Res., *65*, 1, January, 1960, pp. 193–201.
6. Norton, K. A., et al., An Experimental Study of Phase Variations in Line-of-Sight Microwave Transmissions, NBS Monograph 33, November 1, 1961.
7. Janes, H. B. and Thompson, M. C., Jr., An Experimental Study of Atmospheric Errors in Microwave Range and Range Difference Measurements, NBS Report 7908, June 25, 1963.
8. Bean, B. R. and Thayer, G. D., Models of the Atmospheric Radio Refractive Index, Proc. IRE, May, 1959, pp. 740–755.
9. Hines, J. N., Tingye Li, Turrin, R. H., The Electrical Characteristics of the Conical Horn-Reflector Antenna, B.S.T.J., *42*, July, 1963, pp. 1187.
10. Githens, J. A., Kelly, H. P., Lozier, J. C., and Lundstrom, A. A., Antenna Pointing System: Organization and Performance, B.S.T.J., *42*, July, 1963, pp. 1213.
11. Githens, J. A. and Peters, T. R., Digital Equipment for the Antenna Pointing System, B.S.T.J., *42*, July, 1963, pp. 1223.
12. Lozier, J. C., Norton, J. A., and Iwama, M., The Servo System for Antenna Positioning, B.S.T.J., *42*, July, 1963, pp. 1253.
13. Cook, J. S. and Lowell, R., The Autotrack System, B.S.T.J., *42*, July, 1963, pp. 1283.
14. Smith, D. H., Carlson, C. P., McCune, R. J., Elicker, R. E., and Sageman, R. E., Planning, Operation, and External Communications of the Andover Earth Station, B.S.T.J., *42*, July, 1963, pp. 1383.
15. Anders, J. V., Higgins, E. F., Jr., Murray, J. L., and Schaefer, F. J., Jr., The Precision Tracker, B.S.T.J., *42*, July, 1963, pp. 1330.
16. Early Bird, TRW Space Log, TRW Systems, Redondo Beach, California, Summer 1965, pp. 33, 34.
17. Blackman, R. B. and Tukey, J. W., *The Measurement of Power Spectra*, Dover Publications, Inc., New York, 1958.
18. Violette, A., personal communication.
19. Bennett, W. R., Spectra of Quantized Signals, B.S.T.J., *27*, July, 1948, pp. 446–472.
20. Kennedy, J. T. and Rosson, J. W., The Use of Solar Radio Emission for the Measurement of Radar Angle Errors, B.S.T.J., *41*, November, 1962, pp. 1799–1812.
21. Barton, D. K., *Radar System Analysis*, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1965.

On the Sensitivity of Channel Capacity for the Gaussian Bandlimited Channel

By I. W. SANDBERG

(Manuscript received June 24, 1966)

It is a classic result of Shannon that binary digits can be communicated with arbitrarily small error probability at any rate less than

$$W \; log_2 \left(1 + \frac{P}{N}\right) \quad (bits/sec)$$

over a channel with bandwidth W and additive Gaussian noise of average power N, using signals of average power at most P. However, in Shannon's proof it is assumed that the input to the receiver is the sum of a linear combination of the bandlimited functions

$$\varphi_0(t - k/2W) \overset{\Delta}{=} \frac{sin\; 2\pi W(t - k/2W)}{2\pi W(t - k/2W)}, \qquad \begin{matrix} -\infty < t < \infty \\ k = 1,2, \cdots \end{matrix}$$

(which are of course of doubly infinite duration) and a sample function from an exactly bandlimited Gaussian random process. The fact that $\varphi_0(k/2W) = 0$ for all integers $k \neq 0$ plays a key role in that it implies the total absence of intersymbol interference.

As a result of these assumptions, there have been some objections to the Shannon model in connection with the notion of rate, the fact that the received signals are entire functions (which are predictable for all time from a knowledge of their values on any interval of nonzero length) and the fact that it is not clear whether the performance of the model is critically dependent on the assumptions that lead to the absence of intersymbol interference.

Since Shannon's model and his associated ingenious arguments are widely known and are of great interest, from the point of view of the system theorist, it is important to be able to prove an "insensitivity theorem" to the effect that if the model is modified to the extent that: (i) $\varphi_0(t)$ is replaced by an approximating function $\varphi(t)$ with the property that the signals are of average power at most $\tilde{P}$ where $\tilde{P}$ is approximately P, and $\varphi(t) = 0$ for $t < t_\varphi$ for some negative number t_φ, and (ii) the noise is approximately

bandlimited with bandwidth W, then, subject to some reasonable qualifications, it is possible to transmit information, with arbitrarily high reliability, at any rate less than

$$W \; log_2 \left(1 + \frac{P}{N}\right).$$

We prove such a theorem in this paper. In fact, we show that if the noise has integrable power spectral density $S(\omega)$ for which

$$0 < \inf_{0 \leqq \omega < 2\pi W} \sum_{p=-\infty}^{\infty} S(\omega + 4\pi W p)$$

and

$$\tilde{N} \overset{\Delta}{=} 2W \sup_{0 \leqq \omega < 2\pi W} \sum_{p=-\infty}^{\infty} S(\omega + 4\pi W p) < \infty$$

(these are very weak assumptions), then any rate

$$R < W \; log_2 \left(1 + \frac{\gamma P}{\tilde{N}}\right)$$

is permissible if $\gamma \; \varepsilon \; (0,1)$ such that [with the understanding that $\varphi(0) = 1$]

$$\sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} | \varphi(k/2W) | < (1 - \gamma^{\frac{1}{2}}) \left(\frac{\tilde{N}\beta}{2WP\gamma}\right)^{\frac{1}{2}}$$

where β is an important positive number that depends on R, $(\tilde{N}/\gamma)$, P, and W.

Observe that if $S(\omega)$ is the ideal spectral density defined by

$$\begin{aligned} S(\omega) &= \frac{N}{2W}, \qquad |\omega| \leqq 2\pi W \\ &= 0, \qquad\quad |\omega| > 2\pi W \end{aligned}$$

then $\tilde{N} = N$.

I. INTRODUCTION

It is a classic result[1] of Shannon that binary digits can be communicated with arbitrarily small error probability at any rate less than

$$W \log_2 \left(1 + \frac{P}{N}\right) \quad \text{(bits/sec)} \tag{1}$$

over a channel with bandwidth W and additive Gaussian noise of average power N, using signals of average power at most P. There are, however, some unrealistic assumptions in Shannon's argument. In particular, there have been some objections[2,3,4] to the Shannon model in connection with, for example, the notion of rate and the fact that the received signals are entire functions (which are predictable for all time from a knowledge of their values on any interval of nonzero length).

The purpose of this paper is to focus attention on Shannon's assumptions[1] and show that they can be modified so that the end result is a quite detailed and informative statement concerned with a much more realistic model* of a communication system.

II. REVIEW OF SHANNON'S ARGUMENT

2.1 *The Capacity of the Time-Discrete Gaussian Channel*

Shannon's result for the bandlimited time-continuous channel follows directly from a result concerned with the following type of time-discrete channel.

The channel receives one of M equally likely inputs (i.e., code words) every T seconds. Each input is a real n-vector $X \triangleq (x_1, x_2, \cdots, x_n)$ which satisfies

$$| X |^2 \leqq \rho T$$

where $| X |$ denotes the Euclidean norm of X and ρ is a positive constant independent of X. It is assumed that there exists a positive constant μ, independent of T, such that $n = 2\mu T$ (with the understanding that we consider only values of T for which $2\mu T$ is an integer).

The channel output (i.e., the receiver input) corresponding to the input X is the n-vector $X + Z$, in which the components of the "noise vector" Z are independent Gaussian random variables with mean zero and variance η. In its attempt to determine which of the M known code words was transmitted, the receiver may make an error, and we shall denote by p_{ei} the probability that an error is made given that code word i is transmitted.

It is assumed that the channel is used to transmit information in the following manner. Let a message source produce independent and equally likely binary digits at the rate R digits per second. Every T seconds,† one of 2^{RT} possible sequences is produced. We set $M = 2^{RT}$ and we represent each of the binary sequences by a particular code word.

* Some different results concerning the significance of the Shannon bound (1) are proved in Ref. 4. In particular, there, for certain models, converse propositions are established.

† We consider only values of T for which RT is an integer.

We say that a rate R is permissible if for each $\epsilon > 0$ there exists a T and a corresponding code such that

$$\max_i p_{ei} \leqq \epsilon.$$

It has been proven that the channel capacity C, the least upper bound of permissible rates, is given by

$$C = \mu \log_2 \left(1 + \frac{\rho}{2\mu\eta}\right) \quad \text{(bits/sec)}.$$

It has also been proven that for $R < C$ there exists a positive number $\beta = \beta(\eta,\rho,\mu,R)$ such that for each $T > 0$ there exists a code with the property that

$$\max_i p_{ei} = \exp\left[-\beta T + o(T)\right].$$

2.2 *The Time-Continuous Bandlimited Channel*

In order to use the ideas and results outlined above in his study of the time-continuous bandlimited channel, Shannon considers the model shown in Fig. 1, with the understanding that **H** represents an ideal low-pass filter with cut-off frequency W, and $z(\cdot)$ denotes a sample function of a bandlimited Gaussian random process with mean zero and power spectral density

$$\begin{aligned} S(\omega) &= \frac{N}{2W}, \qquad |\omega| \leqq 2\pi W \\ &= 0, \qquad |\omega| > 2\pi W, \end{aligned}$$

where N is a positive constant. Clearly the average power of $z(\cdot)$ is N.

As in the time-discrete case, the message source produces R binary digits per second, so that every T seconds one of $M = 2^{RT}$ possible sequences is produced. Consider the ith such sequence. The coder and signal generator associates with this sequence a particular n-vector

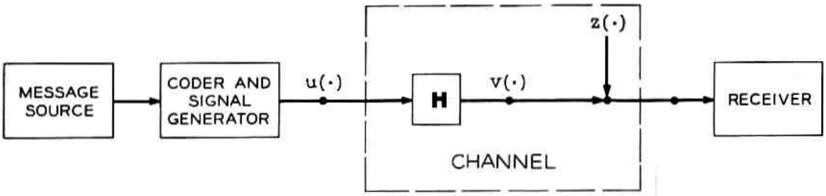


Fig. 1 — Model of a Communication System.

$X \triangleq (x_1, x_2, \cdots, x_n)$, where $n = 2WT$, and a corresponding signal

$$u(t) = \sum_{k=1}^{n} x_k \frac{\sin 2\pi W(t - k/2W)}{2\pi W(t - k/2W)}, \qquad t \, \varepsilon \, (-\infty, \infty)$$

which is transmitted. This process is repeated every T seconds. It is assumed that

$$| X |^2 \leqq 2WPT$$

for each code word, so that, for each signal, as can readily be verified,

$$\frac{1}{T} \int_{-\infty}^{\infty} u(t)^2 dt \leqq P. \tag{2}$$

Insofar as a physical interpretation of (2) is concerned, the object on the left is the total energy of $u(\cdot)$ divided by the length of the interval $[(4W)^{-1}, (4W)^{-1} + T]$ which, considering only the instants $t = k/2W$, contains all of the samples of $u(\cdot)$ that can be made nonzero. If (2) holds, then Shannon says that $u(\cdot)$ has average power at most P.

The received signal due to the noise and only the ith sequence is $u(\cdot) + z(\cdot)$, since the response of $\mathbf{H}$ to $u(\cdot)$ is $u(\cdot)$. The value of this signal at the instant $t = k/2W$ is

$$x_k + z(k/2W) \quad \text{for} \quad k = 1, 2, \cdots, n$$

in which the $z(k/2W)$ are independent* Gaussian random variables with mean zero and variance N. These sample values are the same as those that would have been obtained if we had not ignored the effect at the receiver of transmitted signals due to previous and subsequent sequences, since the values of such signals at $t = k/2W$ vanish for $k = 1, 2, \cdots, n$.

Thus, on the basis of the channel capacity result of the previous section, we see that our continuous channel can process information, with arbitrarily high reliability, at any rate less than the capacity of the time-discrete channel with parameters $\mu = W$, $\rho = 2WP$, and $\eta = N$, that is, at any rate R less than

$$W \log_2 \left(1 + \frac{P}{N}\right).$$

2.3 *Discussion*

The argument of the last section is based on the assumptions that the input to the receiver is the sum of a linear combination of the band-

* The autocorrelation function of the noise vanishes for $\tau = k/2W$, $k \neq 0$.

limited functions

$$\varphi_0(t - k/2W) \triangleq \frac{\sin 2\pi W(t - k/2W)}{2\pi W(t - k/2W)}, \qquad \begin{array}{c} -\infty < t < \infty \\ k = 1,2,\ldots \end{array}$$

(which are of course of *doubly infinite* duration) and a sample function from an *exactly* bandlimited Gaussian random process. The fact that $\varphi_0(k/2W) = 0$ for all integers $k \neq 0$ plays a key role in that it implies the *total* absence of intersymbol interference.

As a result of these assumptions, there have been some objections to the Shannon model in connection with the notion of rate,* the fact that the received signals are entire functions (which are predictable for all time from a knowledge of their values on any interval on nonzero length), and the fact that it is not clear whether or not the performance of the model is critically dependent on the assumptions that lead to the absence of intersymbol interference.

Since Shannon's model and his associated ingenious arguments are widely known and are of great interest, from the point of view of the system theorist, it is important to be able to prove an "insensitivity theorem" to the effect that if the model is modified to the extent that: (*i*) $\varphi_0(t)$ is replaced by an approximating function $\varphi(t)$ with the property that the signals are of average power at most $\tilde{P}$ where $\tilde{P}$ is approximately P, and $\varphi(t) = 0$ for $t < t_\varphi$ for some negative number t_φ, and (*ii*) the noise is approximately bandlimited with bandwidth W, then, subject to some reasonable qualifications, it is possible to transmit information, with arbitrarily high reliability, at any rate less than

$$W \log_2 \left(1 + \frac{P}{N}\right).$$

A quite explicit theorem of this type is stated in the next section.

III. THE MORE REALISTIC MODEL

We now consider the system of Fig. 1 to be an approximation to the Shannon model described in Section 2.2.

Here we assume that $z(\cdot)$ is a sample function from a Gaussian random process with zero mean and integrable power spectral density $S(\omega)$ with the property that

$$\sup_{0 \leq \omega < 2\pi W} \sum_{p=-\infty}^{\infty} S(\omega + 4\pi W p)$$

* Shannon himself has indicated[5] that care must be taken in the physical interpretation of the result of Section 2.2. However, he does not discuss the effect of intersymbol interference or the effect of the departure of the noise spectrum from the ideal spectrum.

is finite. From the engineering viewpoint, this finiteness condition is a very weak assumption; it is certainly satisfied if there exists a constant $K > 0$ such that $S(\omega) \leqq K(1 + \omega^2)^{-1}$ for all real ω.

We again suppose that the message source produces one of $M = 2^{RT}$ equally likely binary sequences every T seconds. We assume that there is a first such sequence and that the coder assigns the code word $(x_1, x_2, \cdots, x_n)$ to it. After T seconds, the second sequence is assigned the code word $(x_{n+1}, x_{n+2}, \cdots, x_{2n})$, and so on. The integer n is equal to $2WT$.

The transmitted signal (i.e., the input to the channel) is assumed to be given by

$$u(t) = \sum_{k=1}^{n} x_k \psi(t - k/2W) + \sum_{k=n+1}^{2n} x_k \varphi(t - k/2W) + \ldots$$

in which $\psi(\cdot)$ is a real-valued function of t defined on $(-\infty, \infty)$ such that there exists a negative constant t_ψ with the property that $\psi(t) = 0$ for $t < t_\psi$. It is evident that each of the signal components (i.e., each sum) is associated with a particular code word, that is, with a particular input sequence to the coder. We note that the first signal component "begins" at $t = t_\psi + (2W)^{-1}$, the second at $t_\psi + (2W)^{-1} + T$, and so on.

The operator **H** in Fig. 1 is assumed here to be causal, linear, and time-invariant. Thus, the output of **H** is

$$v(t) = \sum_{k=1}^{n} x_k \varphi(t - k/2W) + \sum_{k=n+1}^{2n} x_k \varphi(t - k/2W) + \ldots$$

in which $\varphi(\cdot)$ is the response of **H** to $\psi(\cdot)$. Since **H** is causal, there exists a negative constant t_φ such that $\varphi(t) = 0$ for $t < t_\varphi$.

We assume that $\varphi(0) = 1$ and that $\varphi(\cdot)$ belongs to L_2 (i.e., is square integrable). We think of $\varphi(t)$ as being close to

$$\varphi_0(t) \triangleq \frac{\sin 2\pi Wt}{2\pi Wt}$$

in the sense that both $\| \varphi - \varphi_0 \|$ ($\| \cdot \|$ denotes the L_2 norm) and

$$\sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} | \varphi(k/2W) - \varphi_0(k/2W) | = \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} | \varphi(k/2W) |$$

are small. Of course this requires that $-t_\varphi$ be sufficiently large.*

* We may certainly take the view that $\psi(\cdot)$ and **H** are approximations to the ideal signal φ_0 and the ideal bandlimiting filter, respectively. However, the specific nature of these approximations is not pertinent to our development. Observe, in fact, that it makes sense for us to assume here that **H** is an approximation to the ideal bandlimiting filter, but that $\psi(\cdot)$ is an impulse-like function. The response $\varphi(\cdot)$ of **H** to $\psi(\cdot)$ is what we wish to focus attention on.

It is assumed also that

$$\sum_{k=1}^{n} (x_{k+jn})^2 \leqq 2WPT$$

for $j = 0, 1, 2, \cdots$, so that the "average power"

$$\frac{1}{T} \int_{-\infty}^{\infty} \left| \sum_{k=1}^{n} x_{k+jn} \varphi[t - (k + jn)/2W] \right|^2 dt$$

of the jth component of $v(\cdot)$ is bounded from above by $P + \zeta_j$, in which $\zeta_j \to 0$ as $\| \varphi - \varphi_0 \| \to 0$.

The receiver, which is assumed to be in possession of the code, samples the signal $v(\cdot) + z(\cdot)$ at the instants $t = k/2W$, $k = 1, 2, \cdots$, to obtain in succession the "received n-vectors"

$$Y_1 \triangleq (v_1, v_2, \cdots, v_n) + (z_1, z_2, \cdots, z_n)$$

$$Y_2 \triangleq (v_{n+1}, v_{n+2}, \cdots, v_{2n}) + (z_{n+1}, z_{n+2}, \cdots, z_{2n})$$

$$\vdots$$

in which $v_k = v(k/2W)$ and $z = z(k/2W)$. These vectors are used as inputs to a minimum distance decoder. Thus, for example, if

$$| Y_1 - X_i | < \min_{j \neq i} | Y_1 - X_j |,$$

in which $\{X_j\}$ denotes the set of code words, then Y_1 is decoded as X_i. We denote by p_{eij} the maximum probability, over all possible sequences of input code words with the jth code word X_i, that Y_j is not decoded as X_i. We let

$$p_{ei} \triangleq \sup_j p_{eij} .$$

Our result (which is proved in the next section) is

Theorem: Concerning the system described above, let

$$0 < \mathop{inf}_{0 \leqq \omega < 2\pi W} \sum_{p=-\infty}^{\infty} S(\omega + 4\pi W p)$$

and

$$\tilde{N} \triangleq 2W \mathop{sup}_{0 \leqq \omega < 2\pi W} \sum_{p=-\infty}^{\infty} S(\omega + 4\pi W p).$$

Then any rate

$$R < W \, log_2 \left(1 + \frac{\gamma P}{\tilde{N}}\right) \quad (bits/sec)$$

is permissible (in the sense of Section 2.1 with p_{ei} as defined above) provided that $\gamma \; \varepsilon \; (0,1)$ such that

$$\sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} |\varphi(k/2W)| < (1 - \gamma^{\frac{1}{2}}) \left(\frac{\tilde{N}\beta}{2WP\gamma}\right)^{\frac{1}{2}}$$

where $\beta = \beta[(\tilde{N}/\gamma), 2WP, W, R]$ is the number introduced in Section 2.1.

Remarks: Observe that if $S(\omega)$ is the ideal power spectral density defined by

$$\begin{aligned} S(\omega) &= \frac{N}{2W}, & |\omega| &\leqq 2\pi W \\ &= 0, & |\omega| &> 2\pi W \end{aligned}$$

then $\tilde{N} = N$. The condition that

$$0 < \inf_{0 \leqq \omega < 2\pi W} \sum_{p=-\infty}^{\infty} S(\omega + 4\pi Wp)$$

is certainly satisfied if $S(\omega)$ is a reasonable approximation to the ideal spectrum.

If $S(\omega)$ is nonincreasing for $\omega \geqq 0$, then for $p = 1, 2, \cdots,$

$$\sup_{0 \leqq \omega < 2\pi W} S(\omega + 4\pi Wp) \leqq \frac{1}{2\pi W} \int_{4\pi Wp - 2\pi W}^{4\pi Wp} S(\omega) d\omega$$

and

$$\sup_{0 \leqq \omega < 2\pi W} S(\omega - 4\pi Wp) \leqq \frac{1}{2\pi W} \int_{-4\pi Wp + 2\pi W}^{-4\pi W(p-1)} S(\omega) d\omega.$$

Thus, for $S(\omega)$ nonincreasing for $\omega \geqq 0$, we have the bound

$$\tilde{N} \leqq 2W \sup_{0 \leqq \omega < 2\pi W} \sum_{p=-1}^{1} S(\omega + 4\pi Wp) + \frac{1}{\pi} \int_{4\pi W}^{\infty} S(\omega) d\omega.$$

The exponent β has been estimated by Shannon.[5]

The basic idea of the proof of the theorem is, roughly speaking, to (i) treat as an additional "noise source" the *departure* of the samples of $v(\cdot)$ from the corresponding samples in the case of zero intersymbol-interference (Sublemma 1 of Section IV provides an estimate of this departure), and (*ii*) to obtain a lower bound on the channel capacity of the more-realistic model by comparing its error probability performance with that of a model possessing zero intersymbol-interference and independent Gaussian noise samples (this is done in the proof of Sublemma 2 of Section IV).

IV. PROOF OF THE THEOREM

4.1 *The Discrete Channel*

Consider first a discrete channel with memory that receives one of M equally likely inputs (i.e., code words) every T seconds. As in Section 2.1, each input is a real n-vector X which satisfies $|X|^2 \leqq \rho T$, n is equal to $2\mu T$, and each input represents a particular sequence of RT binary digits. Let $(x_1, x_2, \cdots, x_n)$ denote the first code word, $(x_{n+1}, x_{n+2}, \cdots, x_{2n})$ the second code word, and so on.

At time $t = (j-1)T$, the receiver receives the n-vector

$$Y_j \triangleq \{y[1 + (j-1)n], y[2 + (j-1)n], \cdots, y[jn]\}$$

in which

$$y(p) = \sum_{k=1}^{\infty} x_k \varphi(p - k) + z(p), \qquad p = 1,2, \cdots$$

where here $\varphi(\cdot)$ is a function defined on the integers so that $\varphi(0) = 1$ and

$$\sum_{k=-\infty}^{\infty} |\varphi(k)| < \infty,$$

and each $z(p)$ is a Gaussian random variable with zero mean. For each j, let

$$Z_j \triangleq \{z[1 + (j-1)n], z[2 + (j-1)n], \cdots, z[jn]\}$$

and

$$V_j \triangleq \{v_j[1 + (j-1)n], v[2 + (j-1)n], \cdots, v[jn]\}.$$

where

$$v(p) = \sum_{k=1}^{\infty} x_k \varphi(p - k).$$

Then $Y_j = V_j + Z_j$.

We assume that the receiver attempts to determine the jth code word V_j by minimum distance decoding as in Section III. Let p_{ei} denote the error probability associated with the transmission of code word i, as defined in Section III. In Section 4.3 we prove the following result, which we shall exploit here, concerning this channel.

Lemma: Let Z_j, as defined above, possess the property that [with $\mathcal{E}$ the expectation operator and $(\cdot,\cdot)$ denoting the usual inner product of n-vectors] there exist constants ϵ and η such that for every real n-vector U of unit length:

$$0 < \epsilon \leqq \mathcal{E}\,|(U,Z_j)|^2 \leqq \eta$$

uniformly in j and n. Let $\gamma \,\varepsilon\, (0,1)$. Then any rate

$$R < \mu\, log_2\left(1 + \frac{\gamma\rho}{2\mu\eta}\right)$$

is permissible (in the sense of Section 2.1) provided that

$$\sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} |\varphi(k)| < (1 - \gamma^{\frac{1}{2}})\left(\frac{\eta\beta}{\rho\gamma}\right)^{\frac{1}{2}}$$

where $\beta = \beta[(\eta/\gamma),\, \rho,\, \mu,\, R]$ is the number introduced in Section 2.1.

4.2 *Completion of the Proof of the Theorem*

$$\begin{aligned}\mathcal{E}\,|\,(U,Z_j)\,|^2 &= \mathcal{E}\sum_{k,l} u_k u_l z_{[k+(j-1)n]} z_{[l+(j-1)n]} \\ &= \sum_{k,l} u_k u_l R[(l-k)/2W]\end{aligned}$$

for any real n-vector U, in which

$$R(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega)e^{i\omega\tau}d\tau.$$

Thus,

$$\begin{aligned}\mathcal{E}\,|\,(U,Z_j)\,|^2 &= \frac{1}{2\pi}\sum_{k,l} u_k u_l \int_{-\infty}^{\infty} S(\omega)e^{i\omega(l-k)/2W}d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left|\sum_{k=1}^{n} u_k e^{-i\omega k/2W}\right|^2 S(\omega)d\omega \\ &= \frac{1}{2\pi}\sum_{p=-\infty}^{\infty}\int_{-2\pi W+4\pi Wp}^{2\pi W+4\pi Wp}\left|\sum_{k=1}^{n} u_k e^{-i\omega k/2W}\right|^2 S(\omega)d\omega \\ &= \frac{1}{2\pi}\int_{-2\pi W}^{2\pi W}\left|\sum_{k=1}^{n} u_k e^{-i\omega k/2W}\right|^2 \sum_{p=-\infty}^{\infty} S(\omega + 4\pi Wp)d\omega.\end{aligned}$$

It follows at once that

$$\mathcal{E} \mid (U,Z_j) \mid^2 \leqq \sup_{0 \leqq \omega < 2\pi W} \sum_{p=-\infty}^{\infty} S(\omega + 4\pi W p) \frac{1}{2\pi} \int_{-2\pi W}^{2\pi W} \left| \sum_{k=1}^{n} u_k e^{-i\omega k/2W} \right|^2 d\omega,$$

and that

$$\mathcal{E} \mid (U,Z_j) \mid^2 \geqq \inf_{0 \leqq \omega < 2\pi W} \sum_{p=-\infty}^{\infty} S(\omega + 4\pi W p) \frac{1}{2\pi} \int_{-2\pi W}^{2\pi W} \left| \sum_{k=1}^{n} u_k e^{-i\omega k/2W} \right|^2 d\omega.$$

Since

$$\frac{1}{4\pi W} \int_{-2\pi W}^{2\pi W} \left| \sum_{k=1}^{n} u_k e^{-i\omega k/2W} \right|^2 d\omega = \mid U \mid^2,$$

we have

$$\mathcal{E} \mid (U,Z_j) \mid^2 \leqq 2W \sup_{0 \leqq \omega < 2\pi W} \sum_{p=-\infty}^{\infty} S(\omega + 4\pi W p)$$

$$\mathcal{E} \mid (U,Z_j) \mid^2 \geqq 2W \inf_{0 \leqq \omega < 2\pi W} \sum_{p=-\infty}^{\infty} S(\omega + 4\pi W p)$$

for $\mid U \mid = 1$, independent of j and n. Thus, we may view the time continuous system of Section III as a discrete-time communication system of the type described at the outset of this section with $\mu = W$, $\rho = 2WP$,

$$\epsilon = 2W \inf_{0 \leqq \omega < 2\pi W} \sum_{p=-\infty}^{\infty} S(\omega + 4\pi W p),$$

and

$$\eta = 2W \sup_{0 \leqq \omega < 2\pi W} \sum_{p=-\infty}^{\infty} S(\omega + 4\pi W p).$$

This proves the theorem.

4.3 *Proof of the Lemma*

With x_k as defined in Section 4.1, let

$$\tilde{V}_j \triangleq \{x_{[1+(j-1)n]}, x_{[2+(j-1)n]}, \cdots, x_{[jn]}\}.$$

Sublemma 1:

$$\mid V_j - \tilde{V}_j \mid^2 \leqq 2\rho T \left(\sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} \mid \varphi(k) \mid \right)^2$$

Proof:

$$| V_j - \tilde{V}_j |^2 = \sum_{p=1+(j-1)n}^{jn} \left| \sum_{k=1}^{\infty} x_k \varphi(p-k) - x_p \right|^2$$

$$= \sum_{p=1+(j-1)n}^{jn} \left| \sum_{k=-\infty}^{\infty} x_k \tilde{\varphi}(p-k) \right|^2$$

in which $x_k = 0$ for $k < 1$, $\tilde{\varphi}(0) = 0$, and $\tilde{\varphi}(k) = \varphi(k)$ for $k \neq 0$. Therefore,

$$| V_j - \tilde{V}_j |^2 = \sum_p \left| \sum_k x_{(p-k)} \tilde{\varphi}(k) \right|^2,$$

and, by the Schwarz inequality,

$$| V_j - \tilde{V}_j |^2 \leqq \sum_p \sum_k | x_{(p-k)} |^2 \cdot | \tilde{\varphi}(k) | \sum_k | \tilde{\varphi}(k) |$$

$$\leqq \sum_k | \tilde{\varphi}(k) | \sum_p | x_{(p-k)} |^2 \sum_k | \tilde{\varphi}(k) |.$$

Since

$$\sum_p | x_{(p-k)} |^2 \leqq 2\rho T,$$

we have

$$| V_j - \tilde{V}_j |^2 \leqq 2\rho T \left(\sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} | \varphi(k) | \right)^2$$

which is the assertion of Sublemma 1.

Therefore, with Y_j and Z_j as defined in Section 4.1, we have

$$Y_j = \tilde{V}_j + E_j + Z_j$$

in which

$$| E_j |^2 \leqq 2\rho T \left(\sum_{k \neq 0} \varphi(k) \right)^2.$$

This fact when combined with the following result* proves the lemma.

Sublemma 2: Consider a time-discrete channel of the type described in Section 2.1. Replace Z by the n-vector $(E + Q)$ in which E is a fixed vector and the components of Q are Gaussian random variables with zero mean with the property that there exist constants ϵ and η such that for every real n-vector U of unit length:

$$0 < \epsilon \leqq \mathcal{E} | (U,Q) |^2 \leqq \eta$$

* See Ref. 3, Appendix D, for a result related to Sublemma 2.

uniformly in n. Let $\gamma \varepsilon$ (0,1). Then any rate

$$R < \mu \, log_2 \left(1 + \frac{\gamma \rho}{2 \mu \eta} \right)$$

is permissible (in the sense of Section 2.1) provided that

$$| E |^2 \leqq \vartheta T$$

for all $T > 0$, in which

$$\vartheta < 2(1 - \gamma^{\frac{1}{2}})^2 \frac{\eta \beta}{\gamma}$$

where $\beta = \beta(\eta/\gamma, \rho, \mu, R)$ is the number introduced in Section 2.1.

Proof: Let $T_0 \varepsilon (0, \infty)$. Consider the time-discrete channel of Section 2.1 with noise vector Z, but with η replaced with $(1/\gamma)\eta$. Here for

$$R < \mu \log_2 \left(1 + \frac{\gamma \rho}{2 \mu \eta} \right)$$

and $T \geqq T_0$, there exists a code $\{X_j\}$ such that $X_i \neq X_j$ for $i \neq j$, and the error probability (using minimum distance decoding) given that the ith code word was transmitted

$$\hat{p}_{ei} \triangleq Pr \bigcup_{j \neq i} \{ | X_i + Z - X_j | \leqq | Z | \}$$

is at most exp $[- \beta T + \theta(T)]$ independent of i, where

$$\beta = \beta[(\eta/\gamma), \rho, \mu, R]$$

and $\theta(T)/T \to 0$ as $T \to \infty$. For this code, the error probability (using minimum distance decoding) for the channel described in Sublemma 2 is

$$p_{ei} \triangleq Pr \bigcup_{j \neq i} \{ | X_i + E + Q - X_j | \leqq | E + Q | \}.$$

Let $c_{ij} \triangleq | X_i - X_j |$, and let U_{ij} denote the unit-length vector $(X_i - X_j)/c_{ij}$. Then it can easily be shown that

$$| X_i + E + Q - X_j | \leqq | E + Q |$$

if and only if

$$(U_{ij}, Q) \leqq - \tfrac{1}{2} c_{ij} - (U_{ij}, E),$$

in which $(\cdot, \cdot)$ denotes the usual inner product of n-vectors. Thus,

$$p_{ei} = Pr \bigcup_{j \neq i} \{ (U_{ij}, Q) \leqq - \tfrac{1}{2} c_{ij} - (U_{ij}, E) \}$$

and similarly,

$$\hat{p}_{ei} = Pr \bigcup_{j \neq i} \{ (U_{ij}, Z) \leqq -\tfrac{1}{2} c_{ij} \}. \tag{3}$$

Consider (3). Let the n-vector $P \overset{\Delta}{=} (p_1, p_2, \cdots, p_n)$ represent a general point in Euclidean n-space $\mathcal{E}_n$, and let $\mathcal{R}_{ij}$ denote the closed half-space of $\mathcal{E}_n$ throughout which $(U_{ij}, P) \leqq -\frac{1}{2} c_{ij}$. Let $\mathcal{R}_i = \bigcup_{j \neq i} \mathcal{R}_{ij}$. Then

$$\hat{p}_{ei} = (2\pi)^{-n/2} \left(\frac{\eta}{\gamma} \right)^{-n/2} \int_{\mathcal{R}_i} \exp\left[-\frac{1}{2} \frac{\gamma}{\eta} \sum_{i=1}^{n} z_k^2 \right] dz_1 \cdots dz_n .$$

Similarly, let $\mathcal{S}_{ij}$ denote the closed half-space throughout which

$$(U_{ij}, P) \leqq -[\tfrac{1}{2} c_{ij} + (U_{ij}, E)]\gamma^{-\frac{1}{2}},$$

and let

$$\mathcal{S}_i \overset{\Delta}{=} \bigcup_{j \neq i} \mathcal{S}_{ij} .$$

Then, since

$$p_{ei} = Pr \bigcup_{j \neq i} \{ (U_{ij}, \gamma^{-\frac{1}{2}} Q) \leqq -[\tfrac{1}{2} c_{ij} + (U_{ij}, E)]\gamma^{-\frac{1}{2}} \},$$

we have, with Λ the covariance matrix of the random variables $\{ q_i \gamma^{-\frac{1}{2}} \}$,

$$p_{ei} = (2\pi)^{-n/2} (\det \Lambda)^{-\frac{1}{2}} \int_{\mathcal{S}_i} \exp\, [-\tfrac{1}{2} Q^t \Lambda^{-1} Q] dq_1 \cdots dq_n .$$

Let us assume that

$$[\tfrac{1}{2} c_{ij} + (U_{ij}, E)]\gamma^{-\frac{1}{2}} \geqq \tfrac{1}{2} c_{ij} \tag{4}$$

for all $j \neq i$. Then $\mathcal{S}_{ij} \subseteq \mathcal{R}_{ij}$, $\mathcal{S}_i \subseteq \mathcal{R}_i$, and hence

$$p_{ei} \leqq (2\pi)^{-n/2} (\det \Lambda)^{-\frac{1}{2}} \int_{\mathcal{R}_i} \exp\, [-\tfrac{1}{2} Q^t \Lambda^{-1} Q] dq_1 \cdots dq_n .$$

Let $Q = \Xi Y$, where Ξ is the orthogonal matrix such that $\Xi^{-1} \Lambda^{-1} \Xi = \operatorname{diag} (\lambda_1^{-1}, \lambda_2^{-1}, \cdots, \lambda_n^{-1})$, with the understanding that λ_1 and λ_n denote the smallest and largest eigenvalues of Λ, respectively. Then

$$p_{ei} \leqq (2\pi)^{-n/2} (\lambda_1 \lambda_2 \cdots \lambda_n)^{-\frac{1}{2}} \int_{\mathcal{R}_i'} \exp\left[-\frac{1}{2} \sum_{k=1}^{n} \lambda_k^{-1} y_k^2 \right] dy_1 \cdots dy_n$$

in which $\mathcal{R}_i'$ denotes the inverse image of $\mathcal{R}_i$ under the transformation represented by Ξ. Similarly,

$$\hat{p}_{ei} = (2\pi)^{-n/2}\left(\frac{\eta}{\gamma}\right)^{-n/2}\int_{\mathcal{R}_i'} \exp\left[-\frac{1}{2}\frac{\gamma}{\eta}\sum_{k=1}^{n} y_k^2\right] dy_1 \cdots dy_n .$$

Since, by assumption,

$$\epsilon \leqq \mathcal{E}\,|(U,Q)|^2 \leqq \eta$$

for every real n-vector U of unit length and every positive integer n, it follows that $\lambda_1 \geqq \epsilon\gamma^{-1}$ and $\lambda_n \leqq \eta\gamma^{-1}$. We note that for $0 < \lambda_j < \eta\gamma^{-1}$:

$$\lambda_j^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\lambda_j^{-1}y_j^2\right] \leqq \left(\frac{\eta}{\gamma}\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\frac{\gamma}{\eta}\,y_j^2\right]$$

provided that $y_j^2 \geqq \eta/\gamma$. Thus,

$$\begin{aligned}(2\pi)^{-n/2}(\lambda_1\lambda_2 \cdots \lambda_n)^{-\frac{1}{2}} &\int_{(\mathcal{R}_i'-\mathcal{C})} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}\lambda_k^{-1}y_k^2\right] dy_1 \cdots dy_n \\ &\leqq (2\pi)^{-n/2}\left(\frac{\eta}{\gamma}\right)^{-n/2}\int_{(\mathcal{R}_i'-\mathcal{C})} \exp\left[-\frac{1}{2}\frac{\gamma}{\eta}\sum_{k=1}^{n} y_k^2\right] dy_1 \cdots dy_n \\ &\leqq \hat{p}_{ei} ,\end{aligned}$$

in which $\mathcal{C}$ denotes the hypercube in $\mathcal{E}_n$ defined by the inequalities: $p_j^2 \leqq \eta/\gamma$ for $j = 1, 2, \cdots, n$.

Therefore,

$$\begin{aligned}p_{ei} &\leqq \int_{\mathcal{R}_i'} \leqq \int_{(\mathcal{R}_i'-\mathcal{C})} + \int_{\mathcal{C}} \\ &\leqq \hat{p}_{ei} + (2\pi)^{-n/2}(\lambda_1\lambda_2 \cdots \lambda_n)^{-\frac{1}{2}}\int_{\mathcal{C}} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}\lambda_k^{-1}y_k^2\right] dy_1 \cdots dy_n .\end{aligned}$$

However,

$$\begin{aligned}(2\pi)^{-n/2}(\lambda_1\lambda_2 \cdots \lambda_n)^{-\frac{1}{2}} \int_{\mathcal{C}} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}\lambda_k^{-1}y_k^2\right] dy_1 \cdots dy_n& \\ = \prod_{k=1}^{n} (2\pi)^{-\frac{1}{2}}\lambda_k^{-\frac{1}{2}}\int_{-\eta/\gamma}^{\eta/\gamma} e^{-\frac{1}{2}\lambda_k^{-1}y^2} dy& \\ \leqq r^n,&\end{aligned}$$

in which

$$r = (2\pi)^{-\frac{1}{2}}\left(\frac{\gamma}{\epsilon}\right)^{\frac{1}{2}}\int_{-\eta/\gamma}^{\eta/\gamma} \exp\left(-\frac{1}{2}\frac{\gamma}{\epsilon}\,y^2\right) dy.$$

Thus,

$$p_{ei} \leqq \hat{p}_{ei} + r^{n} \leqq \exp\left[-\beta T + \theta(T)\right] + r^{2\mu T}. \tag{5}$$

Since $r < 1$, the right-side of (5) approaches zero as $T \to \infty$. Therefore, to complete the proof of Sublemma 2, it suffices to show that there exist values of T_0 such that (4) is satisfied (for all $j \neq i$) for all $T \geqq T_0$.

We note first that (4) is satisfied if

$$-(U_{ij}, E) \leqq \tfrac{1}{2}(1 - \gamma^{\frac{1}{2}})c_{ij} \tag{6}$$

for all $j \neq i$. Since $-(U_{ij}, E) \leqq |E|$, (6) is satisfied if

$$|E| \leqq \tfrac{1}{2}(1 - \gamma^{\frac{1}{2}})c_{ij} \tag{7}$$

for all $j \neq i$.

We now estimate the numbers c_{ij}. We have,[4] with $a \triangleq \frac{1}{2}c_{ij}(\gamma/\eta)^{\frac{1}{2}}$,

$$\exp\left[-\beta T + \theta(T)\right] \geqq \hat{p}_{ei} \geqq Pr\{(U_{ij},Z) \leqq -\tfrac{1}{2}c_{ij}\}$$
$$= (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{-a} e^{-\frac{1}{2}x^2}\,dx,$$

for any i and any $j \neq i$, since the variance of (U_{ij}, Z) is η/γ. Therefore,

$$\exp\left[-\beta T + \theta(T)\right] \geqq (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{-a} e^{-\frac{1}{2}x^2}\,dx = (2\pi)^{-\frac{1}{2}} \int_{a^2/2}^{\infty} \cdot e^{-y}(2y)^{-\frac{1}{2}}\,dy. \tag{8}$$

Let $\delta > 0$ be a constant, and let $\alpha(\delta)$ denote the smallest nonnegative number such that

$$(2y)^{-\frac{1}{2}} \geqq e^{-\delta y} \quad \text{for all} \quad y \geqq \alpha(\delta).$$

Then

$$\exp\left[-\beta T + \theta(T)\right] \geqq (2\pi)^{-\frac{1}{2}} \int_{a^2/2}^{\infty} \exp\left[-(1+\delta)y\right]dy$$
$$\geqq (2\pi)^{-\frac{1}{2}}(1+\delta)^{-1} \exp\left[-\tfrac{1}{2}(1+\delta)a^2\right]$$

for $a^2 \geqq 2\alpha(\delta)$, from which it follows at once that

$$a^2 \geqq 2(1+\delta)^{-1}\beta T - 2(1+\delta)^{-1}\{\ln\left[(2\pi)^{\frac{1}{2}}(1+\delta)\right] + \theta(T)\}$$

for $a^2 \geqq 2\alpha(\delta)$. Since $\exp\left[-\beta T + \theta(T)\right] \to 0$ as $T \to \infty$, we see from (8) that for each $\alpha(\delta) > 0$, there exists a constant $T_\delta > 0$ such that $a^2 > 2\alpha(\delta)$ for all $T \geqq T_\delta$. Thus, for each $\delta > 0$ there exists a $T_\delta\, \varepsilon\, (0, \infty)$

such that

$$\begin{aligned} c_{ij}^2 \geqq 8(1+\delta)^{-1}\gamma^{-1}\eta\beta T \\ - 8(1+\delta)^{-1}\gamma^{-1}\eta\{\ln[(2\pi)^{\frac{1}{2}}(1+\delta)] + \theta(T)\} \end{aligned} \tag{9}$$

for all $T \geqq T_\delta$.

Inequality (7) is therefore satisfied for all $T \geqq T_0$ if $T_0 > T_\delta$ and

$$\begin{aligned} |E|^2 \leqq 2(1-\gamma^{\frac{1}{2}})^2(1+\delta)^{-1}\gamma^{-1}\eta\beta T \\ - 2(1-\gamma^{\frac{1}{2}})^2(1+\delta)^{-1}\gamma^{-1}\eta\{\ln[(2\pi)^{\frac{1}{2}}(1+\delta)] + \theta(T)\} \end{aligned} \tag{10}$$

for all $T \geqq T_0$. By assumption: $|E|^2 \leqq \vartheta T$ for all $T > 0$, in which

$$\vartheta < 2(1-\gamma^{\frac{1}{2}})^2\gamma^{-1}\eta\beta.$$

Choose $\delta > 0$ so that

$$\vartheta < 2(1-\gamma^{\frac{1}{2}})^2(1+\delta)^{-1}\gamma^{-1}\eta\beta,$$

and then let $T_0 \,\varepsilon\, [T_\delta, \infty)$ be so large that

$$\begin{aligned} \vartheta \leqq 2(1-\gamma^{\frac{1}{2}})^2(1+\delta)^{-1}\gamma^{-1}\eta\beta \\ - 2(1-\gamma^{\frac{1}{2}})^2(1+\delta)^{-1}\gamma^{-1}\eta T^{-1}\{\ln[(2\pi)^{\frac{1}{2}}(1+\delta)] + \theta(T)\} \end{aligned}$$

for all $T \geqq T_0$. Then (10) is satisfied for all $T \geqq T_0$. This completes the proof of Sublemma 2.

V. FINAL REMARKS

The writer is indebted to D. Hamming and L. A. Shepp for discussions concerning this work, and to J. Savage, D. Slepian, and A. Wyner for commenting on the draft.

REFERENCES

1. Shannon, C. E., Communication in the Presence of Noise, Proc. IRE, *37*, January, 1949, pp. 10–21.
2. Slepian, D., Bounds on Communication, B.S.T.J., *42*, 1963, p. 681, (see Appendix B).
3. Ash, R. B., *Information Theory*, John Wiley & Sons, Inc., 1965, p. 256.
4. Wyner, A. D., The Capacity of the Bandlimited Gaussian Channel, B.S.T.J., *45*, March, 1966, p. 364.
5. Shannon, C. E., Probability of Error for Optimal Codes in a Gaussian Channel, B.S.T.J., *38*, May, 1959, p. 611.

Phase Vocoder

By J. L. FLANAGAN and R. M. GOLDEN

(Manuscript received July 18, 1966)

A vocoder technique is described in which speech signals are represented by their short-time phase and amplitude spectra. A complete transmission system utilizing this approach is simulated on a digital computer. The encoding method leads to an economy in transmission bandwidth and to a means for time compression and expansion of speech signals.

I. INTRODUCTION

Analysis-synthesis methods for speech transmission aim at efficient encoding of voice signals. A customary approach is to represent separately the important features of vocal excitation and tract transmission.[1] The well-known channel vocoder of Dudley[2] derives signals which fall into this dichotomy. The tract transmission is described by values of the short-time amplitude spectrum measured at discrete frequencies, and the excitation is described in terms of the fundamental frequency of the voice and the voiced-unvoiced character of the signal. Efforts to solve the long-standing problem of good-quality synthesis from such representations have centered on adequate analysis and specification of the excitation data.

One advance in surmounting the difficulties connected with pitch and voiced-unvoiced extraction is the voice-excited vocoder (VEV).[3] This device relys on transmission of an unprocessed subband of the original speech to carry the excitation information. The spectral envelope information is transmitted as in the channel vocoder by a number of slowly-varying signals. Through accurate preservation of excitation details, a transmission of improved quality and modest bandsaving is achieved.

The present paper proposes another technique for encoding speech to achieve comparable bandsaving and acceptable voice quality. In addition, the technique provides a convenient means for compression and expansion of the time dimension. The method specifies the speech signal in terms of its short-time amplitude and phase spectra. For this reason, it is called phase vocoder. Like the VEV, the phase vocoder does not require the pitch tracking and voiced-unvoiced switching inherent in conventional channel vocoders. Elimination of these decision-making processes and the transmission of excitation information by phase-derivative signals contribute to improved quality in the synthesized signal.

II. PRINCIPLES

If a speech signal $f(t)$ is passed through a parallel bank of contiguous band-pass filters and then recombined, the signal is not substantially degraded. The operation is illustrated in Fig. 1, where BP_1---BP_N represent the contiguous filters. The filters are assumed to have relatively flat amplitude and linear phase characteristics in their pass bands. The output of the nth filter is $f_n(t)$, and the original signal is approximated as

$$f(t) \cong \sum_{n=1}^{N} f_n(t). \tag{1}$$

Let the impulse response of the nth filter be

$$g_n(t) = h(t) \cos \omega_n t, \tag{2}$$

where the envelope function $h(t)$ is normally the impulse response of a physically-realizable low-pass filter. Then the output of the nth filter is the convolution of $f(t)$ with $g_n(t)$,

$$\begin{aligned} f_n(t) &= \int_{-\infty}^{t} f(\lambda)h(t-\lambda) \cos [\omega_n(t-\lambda)]d\lambda \\ &= \mathrm{Re}\left[\exp (j\omega_n t) \int_{-\infty}^{t} f(\lambda)h(t-\lambda) \exp (-j\omega_n\lambda)d\lambda\right]. \end{aligned} \tag{3}$$

The latter integral is a short-time Fourier transform of the input signal $f(t)$, evaluated at radian frequency ω_n. It is the Fourier transform of that part of $f(t)$ which is "viewed" through the sliding time aperture

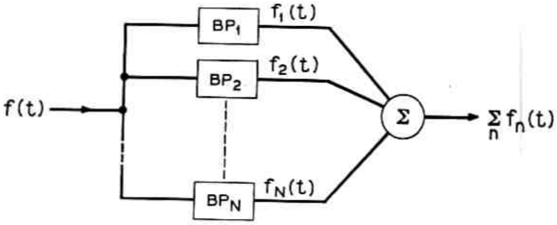


Fig. 1 — Filtering of speech by contiguous band-pass filters.

$h(t)$. If we denote the complex value of this transform as $F(\omega_n, t)$, its magnitude is the short-time amplitude spectrum $| F(\omega_n, t) |$, and its angle is the short-time phase spectrum $\varphi(\omega_n, t)$. Then

$$f_n(t) = \mathrm{Re}[\exp(j\omega_n t) F(\omega_n, t)]$$

or

$$f_n(t) = | F(\omega_n, t) | \cos[\omega_n t + \varphi(\omega_n, t)]. \tag{4}$$

Each $f_n(t)$ may, therefore, be described as the simultaneous amplitude and phase modulation of a carrier ($\cos \omega_n t$) by the short-time amplitude and phase spectra of $f(t)$, both evaluated at frequency ω_n.

Experience with channel vocoders shows that the magnitude functions $| F(\omega_n, t) |$ may be band-limited to around 20 to 30 Hz without substantial loss of perceptually-significant detail. The phase functions $\varphi(\omega_n, t)$, however, are generally not bounded; hence they are unsuitable as transmission parameters. Their time derivatives $\dot{\varphi}(\omega_n, t)$, on the other hand, are more well-behaved, and we speculate that they may be band-limited and used to advantage in transmission. To within an additive constant, the phase functions can be recovered from the integrated (accumulated) values of the derivatives. One practical approximation to $f_n(t)$ is, therefore,

$$\tilde{f}_n(t) = | F(\omega_n, t) | \cos[\omega_n t + \tilde{\varphi}(\omega_n, t)], \tag{5}$$

where

$$\tilde{\varphi}(\omega_n, t) = \int_0^t \dot{\varphi}(\omega_n, t) dt.$$

The expectation is that loss of the additive phase constant will not be unduly deleterious.

Reconstruction of the original signal is accomplished by summing the outputs of n oscillators modulated in phase and amplitude. The oscillators are set to the nominal frequencies ω_n, and they are simultaneously phase and amplitude modulated from band-limited versions of $\dot{\varphi}(\omega_n, t)$ and $| F(\omega_n, t) |$. The synthesis operations are diagrammed in Fig. 2.

These analysis-synthesis operations may be viewed in an intuitively appealing way. The conventional channel vocoder separates vocal excitation and spectral envelope functions. The spectral envelope functions of the conventional vocoder are the same as those described here by $| F(\omega_n, t) |$. The excitation information, however, is contained in a signal which specifies voice pitch and voiced-unvoiced (buzz-hiss) excitation. In the phase vocoder when the number of channels is reasonably

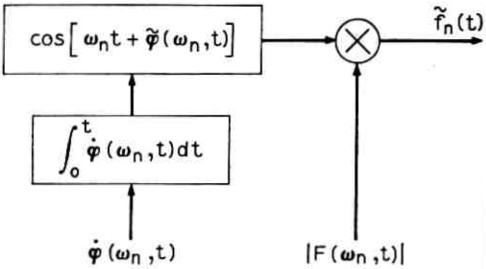

$\cos[\omega_n t + \tilde{\varphi}(\omega_n, t)]$
$\tilde{f}_n(t)$
$\int_0^t \dot{\varphi}(\omega_n, t)dt$
$\dot{\varphi}(\omega_n, t)$
$|F(\omega_n, t)|$

Fig. 2 — Speech synthesis based on the short-time amplitude and phase-derivative spectra.

large, the information about excitation is conveyed primarily by the $\dot{\varphi}(\omega_n, t)$ signals.* In the present technique, and if good quality and natural transmission are requisites, the indications are that the $\dot{\varphi}(\omega_n, t)$ signals may require about the same channel capacity as the spectrum-envelope information. This preliminary impression seems not unreasonable in view of our experience with voice quality in vocoders.

III. COMPUTER SIMULATION

We have simulated a complete phase vocoder analyzer and synthesizer on an IBM 7094 computer. The program, written in the BLODI-B language,[4,5] provides for the processing of any digitalized input speech signal. Flexibility built into the program permits examination of a number of design parameters such as number of channels, width of analyzing pass bands, band center frequencies, and band limitation of the phase and amplitude signals.

In the analyzer, the amplitude and phase spectra are computed by forming the real and imaginary parts of the complex spectrum

$$F(\omega_n, t) = a(\omega_n, t) - jb(\omega_n, t),$$

where

$$a(\omega_n, t) = \int_{-\infty}^{t} f(\lambda)h(t - \lambda) \cos \omega_n \lambda d\lambda$$

and

$$b(\omega_n, t) = \int_{-\infty}^{t} f(\lambda)h(t - \lambda) \sin {}_n\lambda d\lambda. \qquad (6)$$

* At the other extreme, with a small number of broad analyzing channels, the amplitude signals contain more information about the excitation, while the $\dot{\varphi}$ phase signals tend to contain more information about the spectral shape. Qualitatively, therefore, the number of channels determines the relative amounts of excitation and spectral information carried by the amplitude and phase signals.

Then,

$$| F(\omega_n , t) | = (a^2 + b^2)^{\frac{1}{2}}$$

and

$$\dot{\varphi}(\omega_n , t) = \left(\frac{\dot{a}b - \dot{b}a}{a^2 + b^2} \right). \tag{7}$$

The computer, of course, must deal with sampled-data equivalents of these quantities. Transforming the real and imaginary parts of (6) into discrete form for programming yields

$$\begin{aligned} a(\omega_n , mT) &= T \sum_{l=0}^{m} f(lT)[\cos \omega_n lT]h(mT - lT) \\ b(\omega_n , nT) &= T \sum_{l=0}^{m} f(lT)[\sin \omega_n lT]h(mT - lT), \end{aligned} \tag{8}$$

where T is the sampling interval. In the present simulation, $T = 10^{-4}$ sec. From these equations, the difference values are computed as

$$\Delta a = a[\omega_n , (m + 1)T] - a[\omega_n , mT]$$

and

$$\Delta b = b[\omega_n , (m + 1)T] - b[\omega_n , mT]. \tag{9}$$

The magnitude function and phase derivative in discrete form, are computed from (8) and (9) as,

$$| F[\omega_n , mT] | = (a^2 + b^2)^{\frac{1}{2}}$$

$$\frac{\Delta\varphi}{T} [\omega_n , mT] = \frac{1}{T} \frac{(b\Delta a - a\Delta b)}{a^2 + b^2} . \tag{10}$$

Fig. 3 shows a block diagram of a single analyzer channel as realized in BLODI-B. Since this block of coding is required for each channel, it is defined as a new block type and thereafter used as though it were a single block. A parameter associated with the block determines the center frequency for each channel. The time-window analyzing filter, labeled $h(lT)$, is itself a special block and can be changed simply by the substitution of a different block of coding.[6]

In the present simulation, a sixth-order Bessel filter is used for the $h(lT)$ window. Its amplitude, phase, and delay responses are plotted in Figs. 4(a), (b), and (c), respectively. Its impulse and step responses are given in Figs. 4(d) and (e). The present simulation uses 30 channels ($N = 30$) and $\omega_n = 2\pi n(100)$ rad/sec. The equivalent pass bands of the

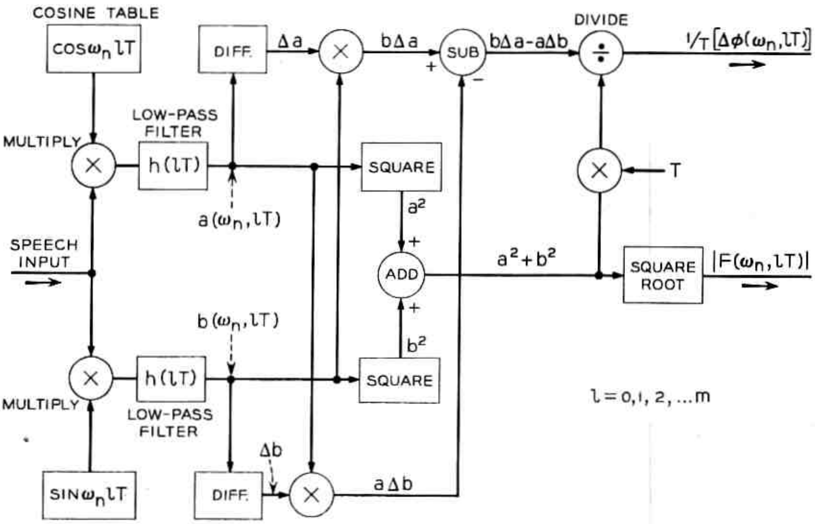


Fig. 3 — Programmed operations for extracting $| F(\omega_n, t) |$ and $\dot{\varphi}(\omega_n, t)$.

analyzing filters overlap at their 6 dB down points and a total spectrum range of 50 to 3050 Hz is analyzed.

Programmed low-pass filtering of any desired form may be applied to the amplitude and phase difference signals as defined by Fig. 3. Simulation of the whole system is completed by the synthesis operations for each channel performed according to

$$\tilde{f}_n(mT) = | F(\omega_n, mT) | \cos\left(\omega_n mT + T\sum_{l=0}^{m}\frac{\Delta\varphi(\omega_n, lT)}{T}\right). \quad (11)$$

Adding the outputs of the n individual channels, according to (1), produces the synthesized speech signal.

IV. TYPICAL RESULTS

As part of the present simulation, identical (programmed) low-pass filters were applied to the $| F(\omega_n, lT) |$ and $(1/T)\Delta\varphi(\omega_n, lT)$ signals delivered by the coding block shown in Fig. 3. These low-pass filters are similar to the $h(lT)$ filters except they are fourth-order Bessel designs. Their response characteristics are shown in Fig. 5. The cut-off frequency is 25 Hz, and the response is −7.6 dB down at this frequency. This filtering is applied to the amplitude and phase signals of all 30 channels in the present simulation. The total bandwidth occupancy of the system is therefore 1500 Hz, or a band reduction of 2:1.

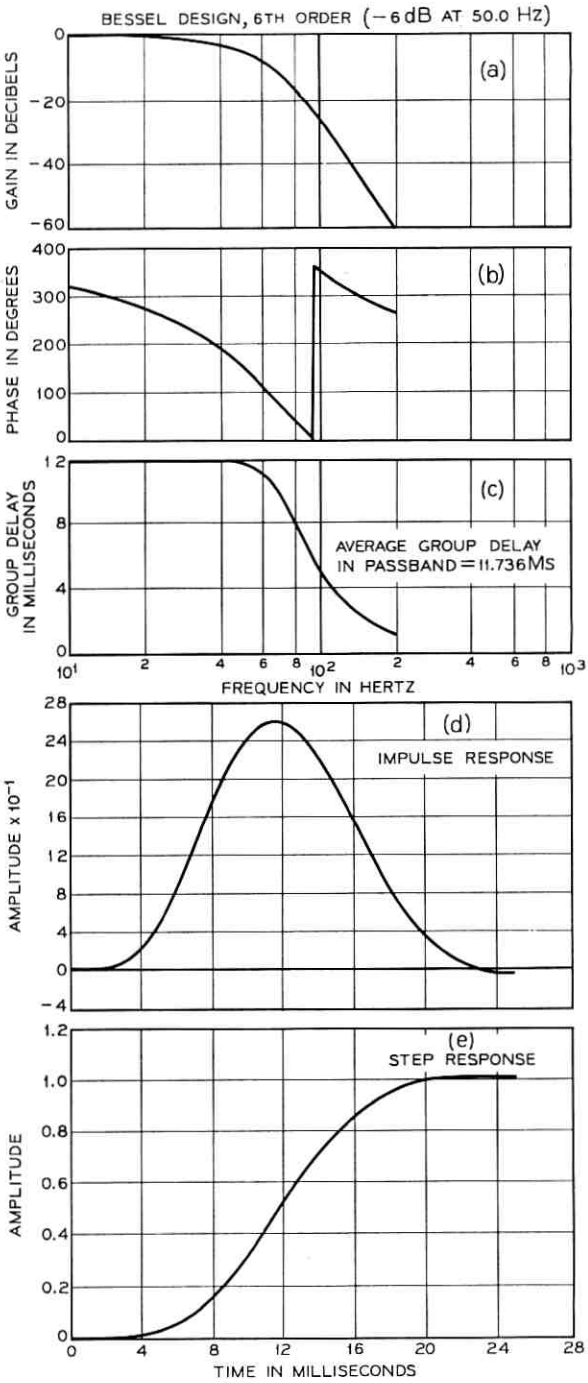


Fig. 4 — $h(t)$ analyzing function and its spectral transform used in one simulation of the phase vocoder. The function is a sixth-order Bessel filter having a −6 dB cut-off of 50 Hz.

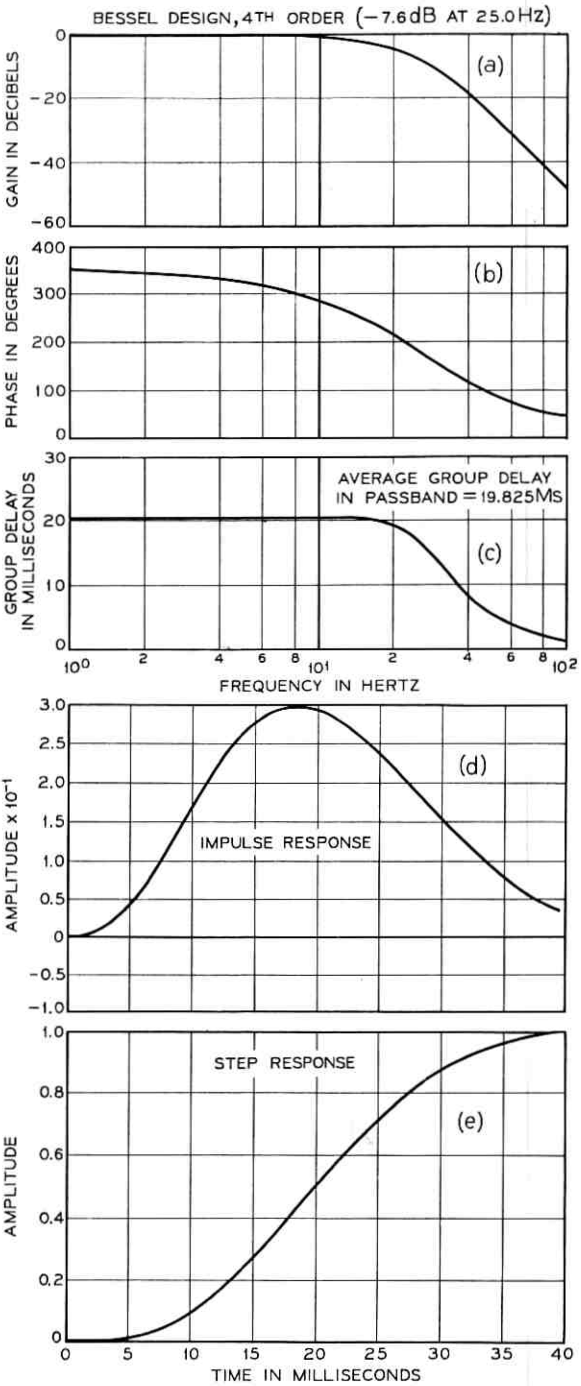


Fig. 5 — Fourth-order Bessel low-pass filter used to smooth the $|F_n|$ and $\dot{\varphi}_n$ signals.

After band-limitation, the phase and amplitude signals are used to synthesize an output according to (11). The result of processing a complete sentence through the programmed system is shown by the sound spectrograms in Fig. 6.* Since the signal band covered by the analysis and synthesis is 50 to 3050, the phase-vocoded result is seen to cut off at 3050 Hz. In this example, the system is connected in a "back-to-back" configuration, and the band-limited channel signals are not multiplexed.

Comparison of original and synthesized spectrograms reveals that formant details are well preserved and pitch and voiced-unvoiced features are retained to perceptually significant accuracy. The quality of the resulting signal considerably surpasses that usually associated with conventional channel vocoders.

V. MULTIPLEXING FOR TRANSMISSION

Besides conventional multiplexing methods for transmitting the band-limited phase and amplitude channel signals (that is, space-frequency or time-division multiplex), the coding technique suggests several other possibilities for transmission in a practicable communication system. As an example, suppose a limited-bandwidth analog channel is the available communication link. One advantageous procedure then is simply to divide (or scale down) all of the phase-derivative signals by some number, say 2 if the available channel has only one-half the conventional voice bandwidth. A synthetic signal of one-half the original bandwidth is then produced by modulating carriers of $\omega_n/2$ by the $\dot{\varphi}_n/2$ and $| F_n |$ signals. The synthetic analog signal now may be transmitted over the half-bandwidth channel.

At the receiver, restoration to the original bandwidth is accomplished by a second sequence of analysis and synthesis operations; namely, amplitude and phase analysis of the half-band signal, multiplication of the phase-derivative signals by a factor of 2, and modulation of ω_n carriers by the restored $\dot{\varphi}_n$ and reanalyzed $| F_n |$ signals. This "self-multiplexing" transmission is illustrated in Fig. 7. Spectrograms of the input signal, the half-band frequency divided signal, and the reanalyzed and resynthesized output are shown. It is clear that two trips through the process introduces measurable degradation, but the intelligibility and quality, particularly for high-pitched voices, remains reasonably good.

In effect, the greatest number q by which the ω_n and $\dot{\varphi}_n$'s may be

* The input speech signal is band limited to 4000 Hz. It is sampled at 10,000 Hz and quantized to 12 bits. It is called into the program from a digital recording prepared previously.

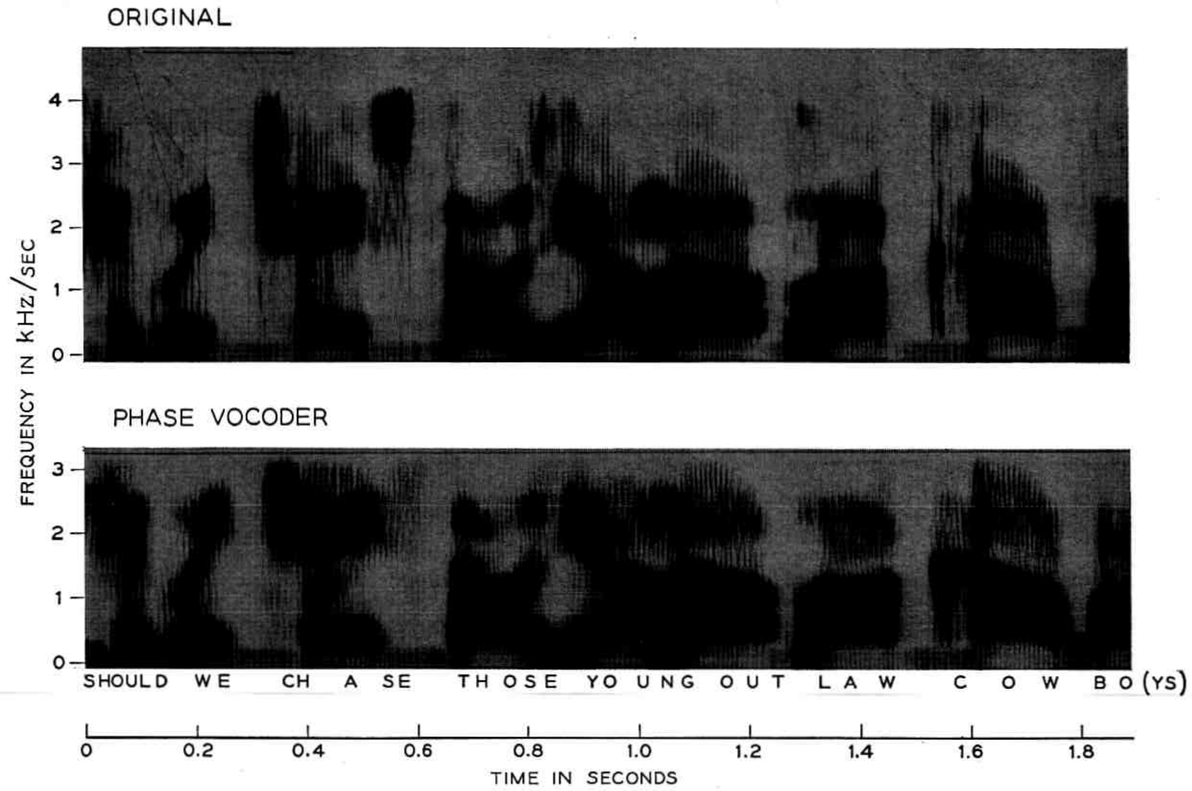


Fig. 6 — Spectrograms illustrating speech transmitted by the phase vocoder ($N = 30$). The band-pass analysis is by sixth-order Bessel filters of 100-Hz band-width. Low-pass filtering of $| F_n |$ and $\dot{\varphi}_n$ is by fourth-order Bessel filters with 25 Hz cut-off. Male speaker A. "Should we chase those young outlaw cowboys."

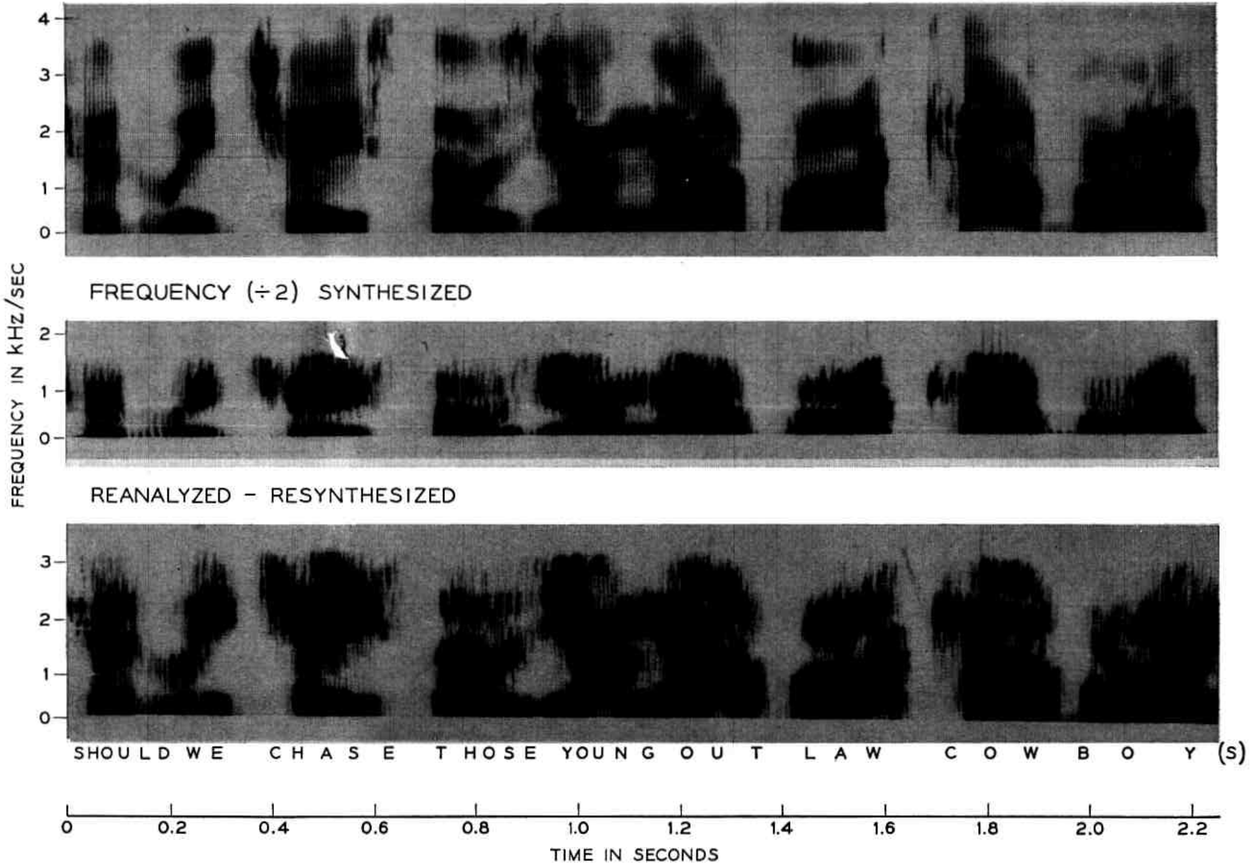


Fig. 7 — Self-multiplexing transmission for a bandwidth reduction of 2:1. (a) Original input; (b) Frequency-divided synthetic signal for analog transmission over one-half bandwidth channel; (c) Synthesized output from the reanalyzed, frequency-multiplied, half-band signal. Male speaker B.

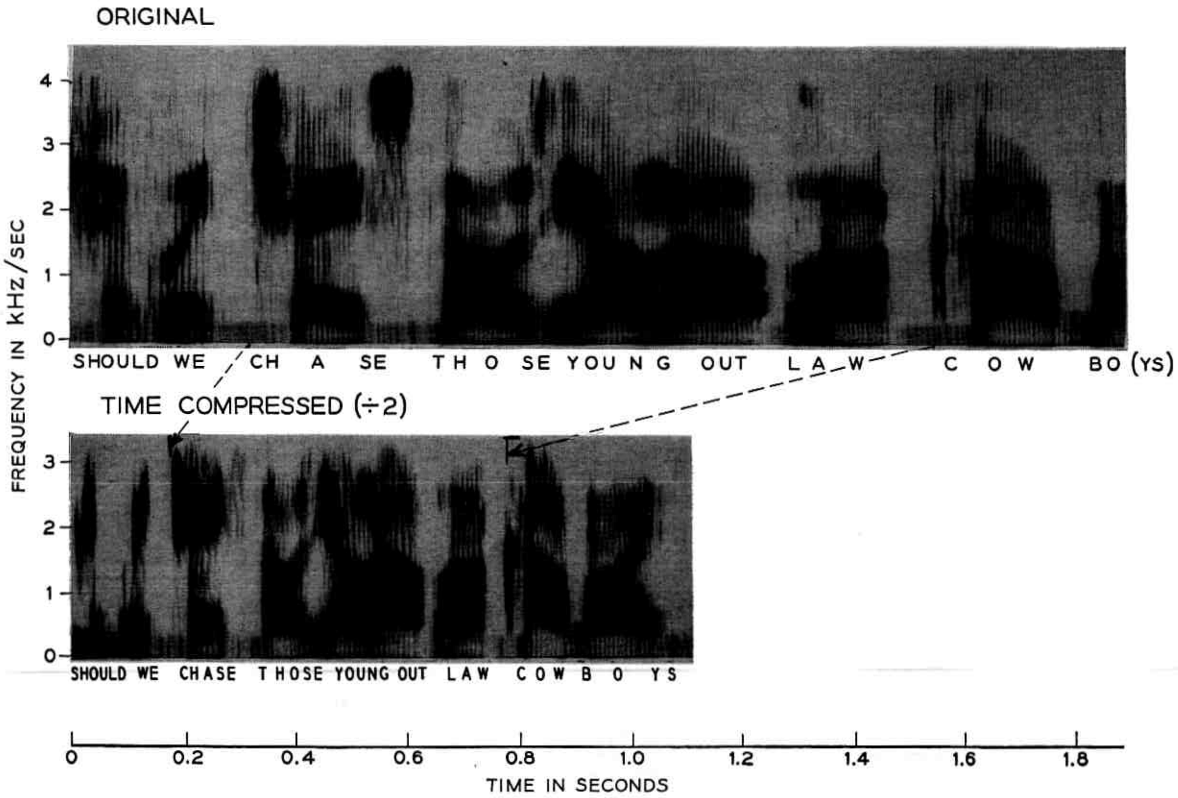


Fig. 8 — Time compression of speech by a factor of 2. Male speaker A. (a) Original input; (b) Time-compressed output.

divided is determined by how distinct the side-bands about each ω_n/q remain, and by how well each $\dot{\varphi}_n/q$ and $|F_n|$ may be retrieved from them.* Practically, the greatest number appears to be about 2 or 3 if transmission of acceptable quality is to be realized.

VI. COMPRESSION AND EXPANSION OF THE TIME SCALE

As mentioned above, a synthetic frequency-divided signal may be produced through division of $[\omega_n t + \int \dot{\varphi}_n dt]$ by some number q. This signal may be essentially restored to its original spectral position by a time speed-up of q. Such a speed-up can be accomplished by recording at one speed and replaying q-times faster. The result is that the time scale is compressed and the message, although spectrally correct, lasts $1/q$th as long as the original. An example of a 2:1 frequency division and time speed-up is shown by the sound spectrograms in Fig. 8. This feature of the phase vocoder is completely parallel to the time-compression feature of the "harmonic compressor" reported earlier.[7] However, the techniques for analysis and synthesis in the two cases are basically different, and the phase vocoder allows compression by non-integer factors.

Time-scale expansion is likewise possible by the frequency multiplication $q[\omega_n t + \int \dot{\varphi}_n dt]$; that is, by recording the frequency-multipied synthetic signal and then replaying it at a speed q-times slower. An example of time-expanded speech is shown by the spectrograms in Fig. 9. The expansion feature provides an interesting "auditory microscope" for directing attention to the spectral properties of specific elements of speech sounds — such as rapidly articulated consonants. In both compression and expansion of the time scale, a perceptual limit exists, of course, to how greatly the time scale may be altered and still have the signal sound like human speech.

An attractive feature of the phase vocoder is that the operations for expansion and compression of the time and frequency scales can be realized by simple scaling of the phase-derivative spectrum. Since the frequency division and multiplication factors can be non-integers, and can be varied with time, the phase vocoder provides an attractive tool for studying non-uniform alterations of the time scale.[8]

* More precisely, the maximum divisor is determined by how closely

$$1/q \int_0^{qt} \dot{\varphi}_n dt$$

represents

$$\int_0^t \dot{\varphi}_n dt .$$

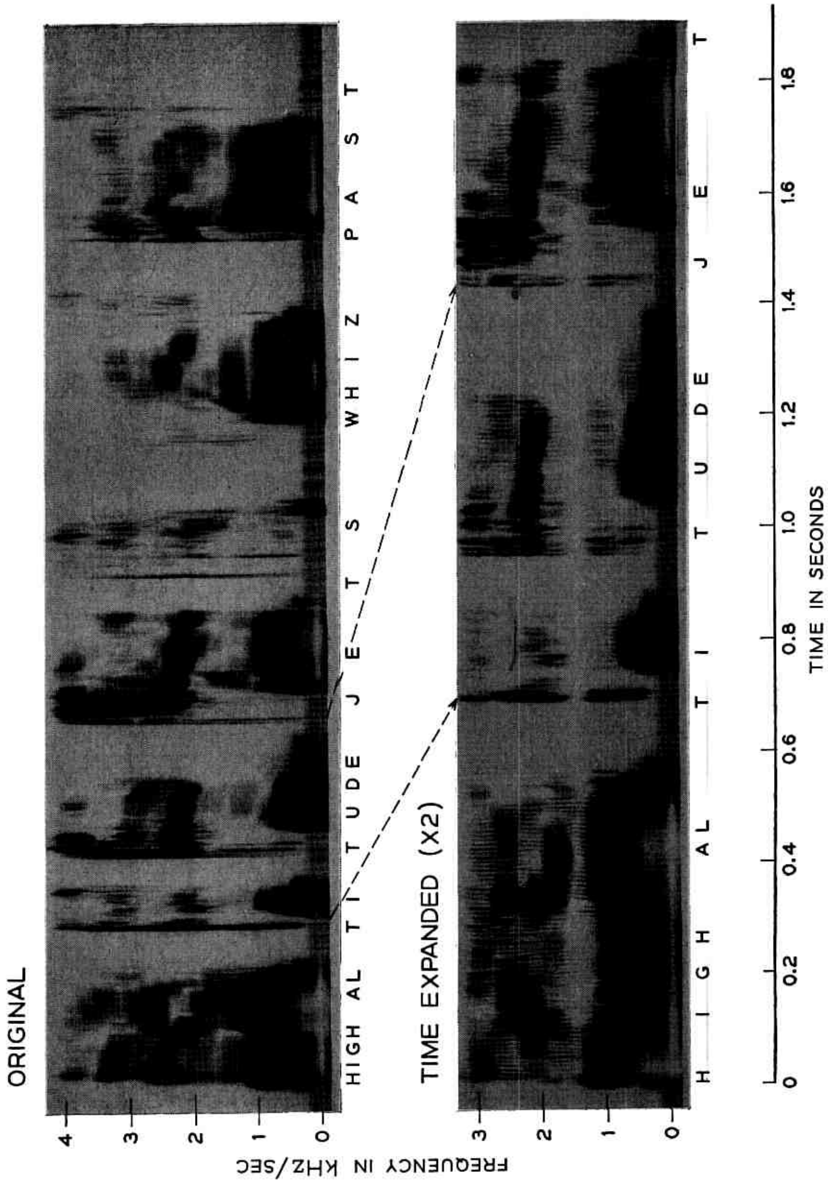


Fig. 9 — Time expansion of speech by a factor of 2. Female speaker. "High altitude jets whiz past screaming." (a) Original input; (b) Time-expanded output.

VII. FURTHER REMARKS ABOUT BAND OCCUPANCY

The possibilities of frequency division imply that the $| F_n |$ and $\dot{\varphi}_n$ signals are, in practical effect, band-limited. As described previously, modest bandwidth reduction of the order of 2:1 can be accomplished by a simple scaling of all the $\dot{\varphi}_n$ signals by $\frac{1}{2}$. (Overt low-pass filtering of the $\dot{\varphi}_n$ signals is not required.) Also, low-pass filtering the analyzed signals to a total band occupancy of one-half the original bandwidth results in relatively good speech quality upon synthesis (Fig. 6). If, however, some further trade between band saving and speech quality is desired, the control signals may be low-passed more severely, with concomitant loss in quality. The impairment resulting from low-passing the $\dot{\varphi}_n$ signals is a comb-filtering, reverberant effect in the reconstituted signals. Qualitatively, low-pass filtering of the $\dot{\varphi}_n$ signals apparently restricts the rate at which pitch changes can be duplicated, and "narrows" the sidebands produced about each ω_n-carrier at the synthesizer.

The discussion connected with (4) has pointed out that each band-pass signal in the phase vocoder may be considered as the simultaneous amplitude and phase modulation:

$$f_n(t) = | F_n | \cos (\omega_n t + \varphi_n),$$

where $| F_n |$ and φ_n are non-band limited, real-valued functions of ω_n and time. Practically, the bandwidth of $f_n(t)$ is confined to $2W$, where W is the cut-off frequency of the low-pass time aperture $h(t)$. This fact does not, however, suggest in an explicit way the band occupancy of the signals $| F_n |$ and φ_n. The experimental results of the present study indicate that each of the latter, at least for practical purposes, can be limited to around $W/2$ or less, but analytical treatment leading to explanation is difficult. Even the inverse problem, that is, calculation of the band occupancy of a simultaneously amplitude and phase modulated carrier, can only be bounded loosely.[9] To apply these bounds requires a precise description of the $| F_n |$ and $\dot{\varphi}_n$ signals. Although these parameters can be measured for a given speech signal, a general mathematical specification is not presently available. It is easy to indicate the difficulties involved. Consider the usual model of voiced speech sounds; that is, a periodic pulse source, whose frequency (pitch) may change with time, supplying excitation to a linear, passive, time-variable network. Variation of the network transmission represents the spectral changes both in the vocal sound source and the vocal tract transmission. For an analysis in terms of narrow pass-bands (large N), the $\dot{\varphi}_n$ signals depend primarily upon voice pitch. The $| F_n |$ signals, on the other hand, depend both upon source spectrum and vocal transmission at any given instant.

VIII. CONSIDERATIONS FOR DIGITAL TRANSMISSION

Applications of the phase vocoder technique to digital transmission are of course obvious. Given an acceptable band-limitation of the $| F_n |$ and $\dot{\varphi}_n$ signals, each may be sampled at its Nyquist rate, or higher, and quantized to an accuracy that is perceptually sufficient. At this writing, optimum parameters for sampling and quantizing the control signals have not been studied in detail. Based upon past experience, however, a nonuniform distribution of the pass bandwidths of the analyzing filters would appear advantageous. For example, center frequencies and bandwidths chosen according to the Koenig scale, the mel (pitch) scale, or the auditory critical-band function should yield dividends.*

All of these bandwidth tapers are characterized by widths which monotonically increase with frequency. In such cases, the low-pass filtering applied to the amplitude signals would have cut-off frequencies also increasing monotonically with frequency. On the other hand, the low-pass filters applied to the phase signals might have cut offs which decrease with frequency. As a result, sampling rates would increase with ω_n for amplitude signals and diminish for phase signals. In addition, quantization levels for all signals might be made more coarse (less numerous) with increasing channel frequency. This is indicated because the ability of the ear to perceive frequency and amplitude changes in the higher end of a complex spectrum is, in general, less acute than for the lower part.

Although detailed study is yet to be made of optimum digital formats, experience in this area with related vocoder devices suggests that transmission at bit rates somewhat less than ten kilobits/sec should be possible without impairment due to digitalization. This rate is several times less than that normally associated with comparable quality PCM encodings of the speech waveform. Besides the questions of design optimization and data format for digital transmission, the trade which may be effected between signal quality and total bit rate is also a subject for further investigation.

IX. CONCLUDING COMMENTS

Because the phase vocoder produces phase derivative signals, it pro-

* Preliminary tests along these lines indicate that a phase vocoder with as few as eight non-uniform channels is capable of relatively good transmission (J. J. Kalsalik, unpublished work).

vides a particularly convenient means for multiplying or dividing the frequency spectrum of a broadband signal. By the same token, it is a convenient method for compressing or expanding the time scale of a signal. Frequency division of speech appears to hold potential as a communication aid for persons with hearing deficient in the high frequencies. Time compression shows promise for auditory "speed-reading" by persons with impaired sight.

Psychoacoustic and physiological studies show that the human ear makes a type of short-time spectral analysis of acoustic signals. This analysis occurs at an early level in the auditory processing; in fact, at a preneural level. It is also clear that the auditory system utilizes information corresponding to smoothed values of the short-time amplitude and phase spectra. The phase vocoder aims to turn these facts to advantage by describing speech signals in terms of band-limited values of the short-time amplitude and phase-derivative spectra. Indications are that band-limited spectral samples, occupying a bandwidth on the order of one half that of the original signal, preserve perceptually-significant features of the signal. Further band conservation can be realized, but at the expense of signal quality. As in many other transmission systems, a continuum of band conservation (or bit rate) versus signal quality exists, and one may choose the point of operation to suit requirements.

REFERENCES

1. Flanagan, J. L., *Speech Analysis, Synthesis and Perception*, Springer Verlag and Academic Press, New York, 1965.
2. Dudley, H., The Vocoder, Bell Labs. Record, *17*, 1939, pp. 122–126.
3. David, E. E., Schroeder, M. R., Logan, B. F., and Prestigiacomo, A. J., New Applications of Voice-Excitation to Vocoders, Proc. Stockholm Speech Comm. Seminar, R.I.T., Stockholm, Sweden, September, 1962.
4. Karafin, B. J., The New Block Diagram Compiler for Simulation of Sampled-Data Systems, AFIPS Conf. Proc., *27*, Pt. 1, 1965, pp. 55–61, Fall Joint Computer Conference, Spartan Books, Washington, D. C.
5. Golden, R. M., Digital Computer Simulation of Sampled Data Communication Systems Using the Block Diagram Compiler, BLODI-B, BSTJ, *45*, March, 1966, pp. 345–358.
6. Golden, R. M. and Kaiser, J. F., Design of Wideband Sampled-Data Filters, BSTJ, *43*, Pt. 2, July, 1964, pp. 1533–1546.
7. Schroeder, M. R., Logan, B. F., and Prestigiacomo, A. J., Methods for Speech Analysis-Synthesis and Bandwidth Compression, Fourth International Congress on Acoustics, Copenhagen, August 21–28, 1962.
8. Hanover, S. L. and Schroeder, M. R., Nonlinear Time Compression and Time Normalization of Speech, 72nd Meeting Acoustical Society of America, November, 1966.
9. Kahn, R. E. and Thomas, J. B., Some Bandwidth Properties of Simultaneous Amplitude and Angle Modulation, IEEE Trans. Inform. Theor., *IT-11*, October, 1965, pp. 516–520.

Theory of Error Rates for Digital FM

By J. E. MAZO and J. SALZ

(Manuscript received June 29, 1966)

A general theory is presented for evaluating the error performance of a digital FM system in the presence of additive noise. The digital system considered is a conventional one employing a voltage-controlled oscillator as the modulator and a limiter-discriminator followed by a low-pass filter as the demodulator. Because of the nonlinear nature of the demodulation process, no adequate analytical techniques have been available to provide a satisfactory treatment. Adopting the notion of "clicks" used by S. O. Rice to study threshold effects in analog FM systems, we have succeeded in evolving a theory capable of predicting performance for a wide range of applications. While our theory reinforces some previously derived results for binary and for narrow-band systems, the results obtained here are not confined to these situations. In particular, the inefficiency of the FM discriminator as a detector for a large number of orthogonal signals is quantitatively evaluated, as well as the role of the post-detection filter. Some qualitative aspects of the error-causing mechanisms discussed in the paper are general, but quantitative results are confined to additive Gaussian noise and large signal-to-noise ratios.

I. INTRODUCTION

Theoretical investigations of FM receivers with analog input signals date back to J. R. Carson and T. C. Fry,[1] and to M. G. Crosby.[2] These investigators and others that followed them[3,4,5] were primarily concerned with the signal-to-noise (S/N) transfer attainable in FM receivers and the determination of threshold effects. Recently S. O. Rice,[6] and previously J. Cohn,[7] attacked the threshold problem in FM receivers from a fresh point of view by using the notion of "clicks." It has been observed that when the noise at the input of an FM receiver is increased beyond some value, the receiver "breaks," that is, for a given (S/N) at the input, a much poorer (S/N) at the output is measured than would be predicted from a linearized analysis of the receiver. Before the breaking point, clicks are heard in the output of an audio receiver. As the input noise is further increased, the clicks merge into a sputtering sound. Rice's approach is to relate this breaking point with the expected number of clicks per second at the output due to the added noise at the input.

While in analog application the criterion of (S/N) transfer is satisfactory, in digital data transmission it does not by itself furnish an adequate performance criterion. Usually performance is judged in terms of error rates which cannot be predicted from the (S/N) transfer for nonlinear receivers. The error rate clearly depends on the statistical distribution of the output noise. In good systems, the errors are very infrequent and are associated with rare peak noise conditions. The statistical structure of the occurrence of infrequent noise peaks and the manner in which they cause errors in FM receivers is the main subject of this paper. Some previous investigations of these effects have been carried out. For example, Bennett and Salz[8] have analyzed binary FM systems, including the effects of distortion. They derived formulas for the error rate without including the post-detection filter in their model. Since the error rates that they obtained for a well-designed *binary* system were close to the optimum obtainable for any receiver, they were able to conclude that the neglect of this filter was justified. Formulas are also available[9,10] for the probability distribution function of the instantaneous frequency of signal plus noise at the input to the post-detection filter for N-ary FM, but these equations are not very useful in predicting the performance of a practical FM system since the task of relating this distribution to the distribution at the output of the post-detection filter is apparently untractable. In a recent paper, Salz[11] considered a multilevel FM narrowband digital communications system where he included the post-detection filter in his analysis. However, the results assume that the post-detection filter did not perform significant selective processing of the detected signal.

In this paper, we shall develop a general theory from which the performance of FM receivers with arbitrary processing gain may be predicted. We shall view the conventional FM receiver, described in Section II, as a device for detecting digital signals and examine its properties in detail. In Section III, after approximating the post-detection filter by an ideal integrator, we show how clicks enter the problem.* Our assumptions and the ensuing mathematical model of the stochastic output are also stated there. The following section supplies the considerable amount of

* Cohn, Ref. (7), has also mentioned the application of the concept of clicks to explain errors in digital FM. Further, D. Schilling of Brooklyn Polytechnic Institute has called to the authors' attention that he is also investigating the relationship between clicks and error rates in FM.

mathematical detail needed to quantitatively substantiate the work of Sections V through VII. In particular, the notion of clicks will be used to explain the poor performance (compared to ideal) of this receiver to detect a large number of orthogonal signals. This phenomenon has also been mentioned by Wozencraft and Jacobs.[12] Another result of the present paper is to establish conditions under which the previous analyses reliably predict the performance of actual FM systems. The work of Refs. 8 and 11 will be supported and it will be shown that for multilevel wideband systems the post-detection filter cannot be ignored. Finally, in Section VIII a discussion is given to suggest circumstances under which successive clicks will not be independent and an instructive example is given showing how this renders ineffective the additional selective filtering possible at the input when the frequencies are very widely spaced.

II. THE DIGITAL FM SYSTEM

A digital FM signal is readily produced by changing the frequency of an oscillator in response to a digital baseband signal. The voltage or current at the output of such an oscillator may be represented as

$$S(t) = A \cos \left[\omega_c t + \int_0^t s(t')dt' + \theta \right], \tag{1}$$

where A is a real amplitude, ω_c the angular center frequency of the oscillator, and θ is an initial phase angle. The digital information-bearing signal $s(t)$ is taken to be a piece-wise constant function of time representable as a random time series of the form

$$s(t) = \omega_d \sum_{n=0}^{n=\infty} a_n g(t - nT), \tag{2}$$

where $\{a_n , n = 0, 1, \cdots\}$ is a sequence of independent and identically distributed integer valued stochastic variables representing the data. For example, one might have $a_n = \pm 1$ with equal probability for binary systems. The function $g(t)$ is a rectangular pulse of unit amplitude and T seconds duration and ω_d is a proportionality constant relating frequency displacement to baseband signal voltage or current. The spectral properties of this FM wave have been extensively analyzed in Refs. 13 and 14.

Transmission and reception of the FM wave is accomplished as follows. The wave $S(t)$ is first processed by a transmitting filter, channel noise is added, and the result is processed again by a receiving filter assumed to be the inverse of the transmitting one. The signal is then detected *via* the limiter-discriminator and filtered at baseband before being synchronously sampled at $t = nT$ (using independent timing information) to de-

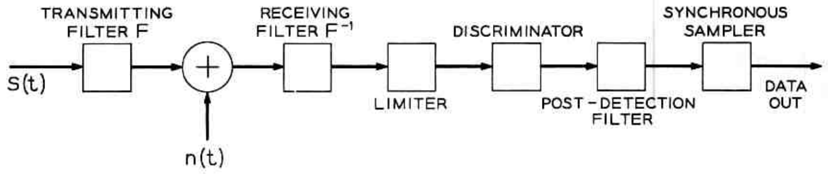


Fig. 1 — Block diagram of a digital FM receiver.

termine sequentially the values of $\{a_n\}$. We have illustrated these operations in block diagram form in Fig. 1. A detailed description of the blocks shown is given in Ref. 15. We shall state here in mathematical terms the assumed operation of the limiter-discriminator. Let the input to the limiter be written in terms of in-phase and quadrature components as

$$x'(t)\cos\omega_c t - y'(t)\sin\omega_c t \equiv R(t)\cos[\omega_c t + \varphi(t)], \tag{3}$$

where

$$R(t) = \sqrt{[x'(t)]^2 + [y'(t)]^2} \tag{4}$$

and

$$\varphi(t) = \tan^{-1} y'(t)/x'(t). \tag{5}$$

Then the output of the discriminator is taken to be

$$\frac{d\varphi}{dt} = \frac{x'(t)\dot{y}'(t) - y'(t)\dot{x}'(t)}{[x'(t)]^2 + [y'(t)]^2}, \tag{6}$$

where the dots denote differentiation with respect to time. The post-detection filter acts upon the quantity (6).

III. FORMULATION OF THE PROBLEM AND A MATHEMATICAL MODEL

We approximate the low-pass filter as an ideal integrator whose impulse response is unity for T' seconds and zero afterward. The duration T' is taken equal to the signaling time T and so no intersymbol interference occurs at the sampling times for a wave described by (1) and (2). The results obtained with this particular choice of filter should be representative of the results one would obtain with any low-pass filter of similar bandwidth.

The sampled output q' of the discriminator low-pass filter output is given by (7)

$$q' = \int_0^T \frac{x'(t)\dot{y}'(t) - y'(t)\dot{x}'(t)}{x'^2(t) + y'^2(t)}\,dt. \tag{7}$$

The in-phase and quadrature components occurring in (7) are now not those of the pure FM wave (1), but have the analogous components of zero mean noise added in as well. One may, by use of a rotating coordinate system, equally consider

$$q = q' - a_n \omega_d T = \int_0^T \frac{x(t)\dot{y}(t) - y(t)\dot{x}(t)}{x^2(t) + y^2(t)}\, dt \tag{8}$$

where $y(t)$ is now a zero mean quadrature noise process, while $x(t)$ is an in-phase noise process with mean A, the amplitude of the noise-free received FM wave. We now proceed formally with (8), defining a quantity

$$r(t) = y(t)/x(t). \tag{9}$$

Equation (8) is then rewritten as a path integral

$$q = \int_{r(0)}^{r(T)} \frac{dr(t)}{1 + r^2(t)} = \int d\varphi(t). \tag{10}$$

In (10) we have written $d\varphi = d(\tan^{-1} y/x)$, but of course we do not mean that φ is evaluated using some fixed branch of $\tan^{-1} y/x$ since this would give φ as a single valued function of y and x and would not allow for the fact that as we circle once about the origin in the xy-plane φ increases by 2π. The noise processes $y(t)$ and $x(t)$ wander about the xy-plane (see Fig. 2), usually staying close to their mean values but occasionally taking large excursions and encircling the origin. Each infinitesimal portion of the path contributes an amount $d\varphi$ volts to the output and all these small amounts from all the small portions of the path must be added together to form the total contribution q. It is easy to see that q depends on the path taken, not just on its endpoints. A simple mathematical reason for this is that the transformation (9) is undefined whenever $x(t) = 0$. Further, the paths taken in the xy-plane are random, and q is therefore, a random variable with some probability density related to the statistics

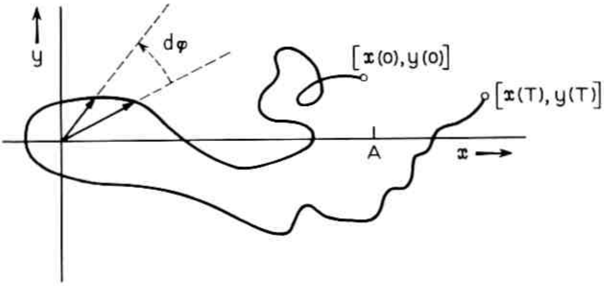


Fig. 2 — A possible path in the xy-plane traced by the noise from $t = 0$ to $t = T$.

of $r(t)$. Unfortunately, this probability density is not determined solely by the elementary statistics of $r(t)$. As will be seen, in addition to the elementary statistics of $r(t)$ the distribution of its singularities on the time axis enters the picture. The singularities of $r(t)$ are determined by the zero-crossings of $x(t)$. Thus, the behavior of FM receivers is intimately related to the structure of the zero crossings of the added noise.[16]

To see how to handle the situation, visualize the following hypothetical state of affairs. Suppose for $0 \leqq t \leqq T$ we have that $y(t) > 0$, and that $x(t)$ is positive for a while, decreases once through zero at $t = t_0$, and then remains negative. A possible plot of $r(t)$ versus t over the time interval is then shown in Fig. 3. For this particular path one has

$$q = \int_{r(0)}^{\infty} \frac{dr}{1 + r^2} + \int_{-\infty}^{r(T)} \frac{dr}{1 + r^2} = \int_{-\infty}^{\infty} \frac{dr}{1 + r^2} + \int_{r(0)}^{r(T)} \frac{dr}{1 + r^2} . \quad (11)$$

In (11) the straightforward interpretation of the integrals is meant. Evaluating the infinite integral one obtains for this path

$$q = \pi + \tan^{-1} r(T) - \tan^{-1} r(0),$$

where $\tan^{-1} x$ means the principal value inverse tangent function, $| \tan^{-1} x | \leqq \pi/2$. In general, one has the result that

$$q = \tan^{-1} r(T) - \tan^{-1} r(0) + n(T)\pi, \quad (12)$$

where $\tan^{-1} x$ again has the principal value interpretation and $n(T)$ is an integer (which may be positive, negative, or zero) which is related to the number of times $x(t)$ vanishes in the interal T and to the sign of $y(t)$ when $x(t)$ vanishes. For large signal-to-noise ratios it is clear that if $x(t)$ vanishes by going to zero from the positive side that it will almost immediately be followed by another vanishing of $x(t)$ in the other direction. If $y(t)$ has not changed, the contribution of the "return trip" to $n(T)$ will cancel the contribution from the previous crossing of the y-axis. On the other hand, if $y(t)$ does change sign so as to cause an encircling of the origin then the contribution to $n(t)$ will be the same as the previous crossing. The net contribution to $n(T)$ of a number of paths is shown in Fig. 4. The paths which have $\Delta n = \pm 2$ are immediately recognized as the "clicks" discussed by Rice.[6] The "clicks" are not the only contribution to $n(T)$ however. There is also a contribution because of the fact that at $t = 0$, when our process begins, we may be in the middle of a large noise fluctuation and be over in the left-half plane. Immediately afterwards, at $t = 0+$, we will experience a contribution of ± 1 to $n(T)$; a similar situation may prevail at time $t = T$, when a possibility exists of stopping the process immediately after we have crossed over to the left-half plane. We will show later that for large signal-to-noise ratios, these

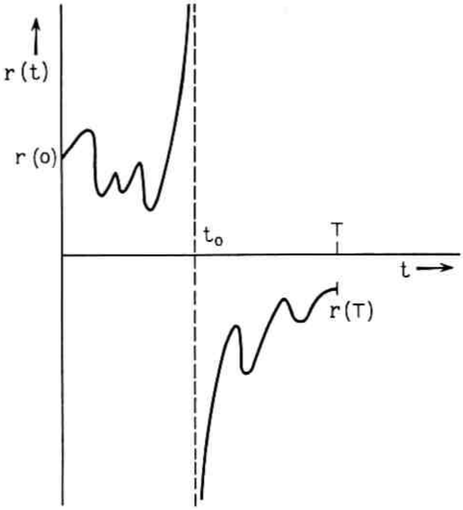


Fig. 3 — A possible sample function of $r(t)$.

end-effects may be neglected because they occur with a probability that is asymptotically negligible compared with the probability of a click.

An important fact to observe before proceeding with the analysis is that q can be decomposed into the sum of three random variables. The first two random variables appearing in (12) are continuous and bounded. Their probability densities are related to the elementary statistics of $x(t)$ and $y(t)$. The third random variable is a discrete one, whose probabilities are determined from the probabilities of zero-crossings of $x(t)$ and $y(t)$.

The remarks made thus far about the effect of noise on FM reception have been general; no assumptions have been made about the statistical nature of the additive disturbance. In order to obtain quantitative results some definite assumptions are necessary. For the remainder of the

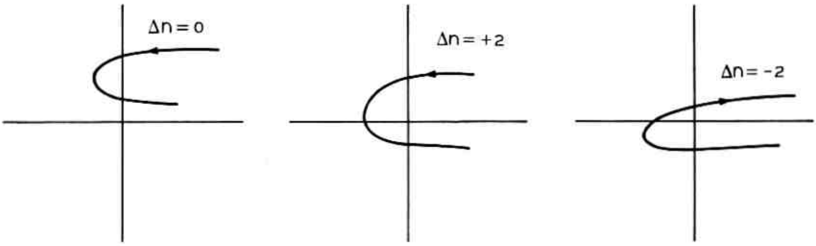


Fig. 4 — Net changes Δn in $n(T)$ caused by some typical paths in the xy-plane.

paper we shall set ourselves the task of studying the structure of the probability distribution of q when the input noise statistics are those of a Gaussian process having a symmetric spectral density about the carrier. From these distributions we determine the error rates as a function of the pertinent system parameters.

No attempt will be made in this paper to derive an exact probability density for the random variable q. This is not a mathematically tractable problem since it requires knowledge of the probability distribution of zero-crossings of random processes. This by itself has been an area of investigation for many years without too much success. The probability distribution of the zero-crossings of most elementary random processes is not currently known.

In order to permit an analysis of the model two assumptions are made, both of which we feel are quite reasonable. These two assumptions taken together state that the three random variables that determine q *via* (12) are all independent. We separate this statement into two assumptions because their individual justification stems from two different physical arguments, one having to do with bandwidth and the other with signal-to-noise ratio. The first assumption states that $\tan^{-1} r(T)$ and $\tan^{-1} r(0)$ are independent. For a flat Gaussian noise input this will be a good approximation if $T \geqq 1/W$, where W is the input noise bandwidth. Since T is also the signaling interval, and the correlation function of the input noise has its first zero at $t \sim 1/W$, the motivation for this assumption is clear. The second assumption, somewhat harder to justify, states that $n(T)$ is independent of the previous two random variables, and the clicks, which comprise $n(T)$, are independent from one another. This is clearly an assumption expressing an intuitive feeling that the clicks occur rarely and of sufficiently short duration. In general, they will be rare if the signal-to-noise ratio is large, and short if the bandwidth satisfies $W \geqq 1/T$ as required above.

These two assumptions plus the identification of crossings of the negative x-axis by the moving point in the xy-plane (as calculated by Rice) with the occurrence of a click shall constitute our working model of the output noise. An indication of how this model must be modified if the input noise spectrum is not relatively flat is given in Section VIII.

IV. THE BASIC DISTRIBUTIONS

Let y be a Guassian variable of zero mean, variance σ^2, and x be another independent Gaussian variable of mean A, variance σ^2.* Then the den-

* Recall that even though our $x(t)$ and $y(t)$ are not independent processes because the noise spectrum will not be symmetrical about $(\omega_c + a_n\omega_d)$, they are independent variables.

sity $\bar{p}(\bar{\varphi})$ where $\tan \bar{\varphi} = y/x$ and $\bar{\varphi}$ has the full range of 2π is well known and is given by Bennett,[17]

$$\bar{p}(\bar{\varphi}) = \frac{\exp(-\rho)}{2\pi} + \tfrac{1}{2}\sqrt{\frac{\rho}{\pi}} \cos \bar{\varphi} \exp(-\rho \sin^2 \bar{\varphi}) \cdot [1 + \operatorname{erf}(\sqrt{\rho} \cos \bar{\varphi})], \tag{13}$$

where $\rho = A^2/2\sigma^2$.

One fact which is implicitly contained in (13) is the probability P_L of finding the signal point in the left half of the xy-plane. However, an easier way to obtain P_L is as follows:

$$P_L = Pr(x < 0) = \tfrac{1}{2} \operatorname{erfc} \frac{A}{\sqrt{2}\,\sigma} \sim \frac{\exp(-\rho)}{2\sqrt{\pi}\sqrt{\rho}} . \tag{14}$$

Equation (14) will be of use in the arguments used to discard the "end effects" at $t = 0$ and $t = T$ spoken of earlier. Equation (13) also immediately yields the probability density $p(\varphi)$ for $\varphi = \tan^{-1}(y/x)$, $-\pi/2 < \varphi < \pi/2$. Indeed, we have

$$\begin{aligned} p(\varphi) &= \bar{p}(\varphi) + \bar{p}(\varphi + \pi) \\ &= \frac{\exp(-\rho)}{\pi} + \sqrt{\frac{\rho}{\pi}} \cos \varphi \exp(-\rho \sin^2 \varphi) \operatorname{erf}(\sqrt{\rho} \cos \varphi) \end{aligned} \tag{15}$$

for $|\varphi| \leqq \pi/2$.

Suppose φ_1 and φ_2 are two independent angles which have the density (15), and define an angle $\Phi = \varphi_1 - \varphi_2$, $|\Phi| \leqq \pi$. It will be of interest for us to determine the probability P_φ that Φ exceeds some angle $\varphi > 0$, i.e., we would like to determine

$$P_\varphi = \int_{-\pi/2}^{(\pi/2)-\varphi} d\varphi_2 \int_{\varphi_2+\varphi}^{\pi/2} d\varphi_1 p(\varphi_1) p(\varphi_2), \qquad \varphi > 0. \tag{16}$$

In general, one is unable to perform these integrations exactly, but since discussion has already been limited to the large S/N region, little will be lost if we make use of this in simplifying the evaluation of (16). The asymptotic evaluation is carried out in detail in the Appendix; we distinguish three cases:*

Case I; $0 < \varphi < \pi/2$:

$$P_\varphi \sim \frac{1}{\sqrt{8\pi}} \frac{\cot(\varphi/2)}{\sqrt{\cos \varphi}} \frac{\exp[-2\rho \sin^2(\varphi/2)]}{\sqrt{\rho}} . \tag{17a}$$

* In (17) the symbol "$\sim$" is used to denote asymptotic equality; this has also been used in (14). Also (17a) and (17c) do not hold if φ gets too close to the end points of the appropriate interval. As a rough rule, φ should not be closer than $1/\sqrt{\rho}$ radians to the end points.

Case II; $\varphi = \pi/2$:

$$P_\varphi \sim (\tfrac{1}{4}) \exp(-\rho). \tag{17b}$$

Case III; $\varphi > \pi/2$:

$$P_\varphi \sim \frac{\exp[-\rho(1 + \cos^2\varphi)]}{2\pi\sqrt{\pi}\,\rho\sqrt{\rho}\sin\varphi\cos^2\varphi}. \tag{17c}$$

The most important characteristic of the result (17) is the dependence of the exponent on angle, since for large ρ the nonexponential factors are relatively slowly varying.

We should remark that for very small angles (15) is well approximated by the Gaussian curve

$$g(\varphi) = \sqrt{\frac{\rho}{\pi}} \exp(-\rho\varphi^2) \tag{18}$$

of zero mean and variance $1/2\rho$. The difference angle Φ would, for very small Φ, be well approximated by the difference of two independent Gaussian variables, each having the density (18). The quantity P_φ calculated on this basis agrees (asymptotically) with the small angle approximation of (17a).

The final item that we discuss in this section is the density of $n(T)$, or rather we discuss the density of that part of $n(T)$ that arises from the clicks ($\Delta n = \pm 2$), ignoring $\Delta n = \pm 1$ contributions. For this we need only take over some ideas and formulas from Rice.[6] We have that (ignoring $\Delta n = \pm 1$)

$$\pi n(T) = 2\pi N(T), \tag{19}$$

where $N(T)$ is the number of clicks that occur in time T. Following Rice, we assume that all clicks are independent and that those tending to increase (decrease) φ by 2π form a Poisson process with rate of occurrence $N_+(N_-)$. In general, with a modulated signal, N_+ and N_- are not equal. The probability density $p(z)$ of $z = N(T)$ is then given by

$$p(z) = \exp[-(N_+ + N_-)T] \sum_{k=-\infty}^{\infty} \delta(z - k) \left(\frac{N_+}{N_-}\right)^{k/2} \cdot I_k(2T\sqrt{N_+N_-}); \tag{20}$$

as may be shown by forming the discrete convolution of the densities of the positive and negative clicks. In (20) $\delta(\cdot)$ is the Dirac delta function and $I_k(\mu)$ is the modified Bessel function of integer order k, behaving for small μ as[18]

$$I_k(\mu) \xrightarrow[\mu\to 0]{} \left(\frac{\mu}{2}\right)^{|k|} \frac{1}{|k|!}; \tag{21}$$

also

$$I_{-k}(z) = I_k(z).$$

The type of modulation that we are concerned with is when the instantaneous frequency deviates by ω_d from the carrier* for a time T, T being the signaling and processing interval. For this situation Rice gives for the average rates N_+ and N_- when the noise at the receiver input is Gaussian

$$\begin{aligned} N_+ = \tfrac{1}{2}\{\sqrt{r^2 + f_d^2}\,[1 - \operatorname{erf}\sqrt{\rho + \rho f_d^2/r^2}] \\ - f_d \exp(-\rho)[1 - \operatorname{erf}(f_d\sqrt{\rho}/r)]\} \end{aligned} \tag{22}$$

and

$$N_- = N_+ + f_d \exp(-\rho), \tag{23}$$

where†

$$\begin{aligned} r &= (1/2\pi)(\dot{\sigma}/\sigma) \\ \sigma^2 &= \operatorname{var} x = \operatorname{var} y \\ \dot{\sigma}^2 &= \operatorname{var} \dot{x} = \operatorname{var} \dot{y}. \end{aligned} \tag{24}$$

Under the assumption that f_d is positive we have asymptotically

$$\begin{aligned} N_+ &\sim \frac{1}{4\sqrt{\pi}} \frac{1}{\rho^{3/2}} \frac{r}{\left(1 + \dfrac{f_d^2}{r^2}\right)\left(\dfrac{f_d^2}{r^2}\right)} \exp[-\rho(1 + f_d^2/r^2)] \\ N_- &\sim N_+ + f_d \exp(-\rho). \end{aligned} \tag{25}$$

Thus, we see that for large ρ an ever greater majority of clicks occur in the negative direction ($f_d > 0$) and for our purposes of computing error rate the clicks in the positive direction may be neglected; i.e., we shall use

$$\left.\begin{aligned} N_+ &\sim 0 \\ \\ N_- &\sim f_d \exp(-\rho) \end{aligned}\right\} \quad \text{for} \quad f_d > 0. \tag{26}$$

* We trust that no confusion will arise between r introduced in (24) and $r(t)$ introduced in (9).

† The case $\omega_d = 0$ corresponds to no modulation. Also, for ease of writing, we no longer explicitly consider the factor a_n.

For $f_d < 0$ the situation is reversed of course. We note that the effect of the clicks on a modulated carrier is to tend to make the measured frequencies appear closer to the carrier frequency than the transmitted frequencies. That is, confining oneself for the moment to *only errors caused by clicks,* frequencies transmitted higher (lower) than the carrier will be measured to be at that frequency or a lower (higher) one, when the noise is small.

Since we shall use approximation (26), the distribution (20) for $z = N(T)$ may be replaced by the simpler Poisson one, where the probability of getting exactly N (negative) clicks in time T is given by*

$$p[N(T)] = \frac{\exp(-N_{-}T)(N_{-}T)^{N(T)}}{[N(T)]!}. \tag{27}$$

Also the probability of getting M or more clicks is, for large signal-to-noise ratios, approximately the probability of getting exactly M clicks.

V. DISTRIBUTION OF OUTPUT AND PROBABILITY OF ERROR

Equations (14), (17), (26), and (27) provide the information required to calculate the distribution of q, (12). In principle we simply convolve the continuous density of $[\tan^{-1} r(T) - \tan^{-1} r(0)]$ with the discrete density of $n(T)\pi$. In Fig. 5, we have given a qualitative sketch of the result, neglecting end effects. This picture is intended to show that the density consists of a central lobe about the transmitted frequency extending to $\pm\pi$ on each side, which is the density of $[\tan^{-1} r(T) - \tan^{-1} r(0)]$, plus identically shaped lobes displaced by integral multiples of 2π toward lower frequencies (assuming $f_d > 0$). These displaced lobes are weighted by the probability of getting the appropriate number of clicks to effect the displacement. Thus, the lobe occupying the space $-2n\pi \pm \pi$ is weighted by the probability of getting exactly n clicks in time T. For $n = 0$ the weighting is essentially one, for large S/N. There are, strictly speaking, similar lobes and weightings on the opposite side as well, but these weights are, for large S/N, negligible compared to the *corresponding* lobe we have drawn. That is to say, the first lobe on the right (not shown in Fig. 5) has small probability compared to the first lobe on the left, but has a large probability compared to the second lobe on the left. Nevertheless, we have neglected to include it because we will generally be concerned with probabilities like $\Pr[\mid q - f_dT \mid > \varphi]$, and thus corresponding weights are important. We dwell on this point be-

* We confine ourselves to $f_d > 0$. Exactly analogous consideration apply to $f_d < 0$. The case $f_d = 0$ occurs if an odd number of frequencies are allowed.

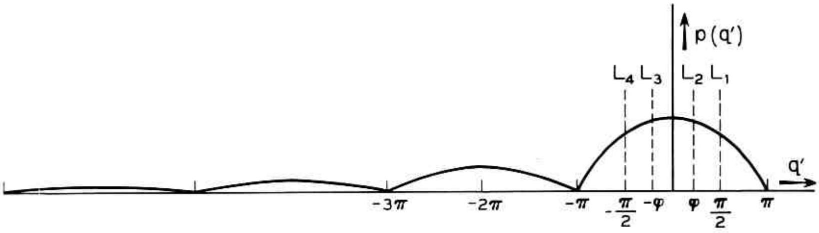


Fig. 5 — Qualitative sketch of density of q' (neglecting end effects) for $f_d > 0$. The dashed lines are for reference in the text.

cause it is conceivable that for some practical or conceptual application the neglect would not be justified.

The discussion given above is still not quite correct; it is modified when we include end effects. The principle correction that inclusion of end effects will cause is to add two more side lobes, one over the interval $[-2\pi,0]$ and the other over the interval $[0,2\pi]$. The weightings of these lobes certainly should not exceed the estimate given in (14), and this will be enough to exclude them for our purposes.

We now apply our results to some typical calculations. Consider the case of narrow band* FM (defined by $\Delta f_d T < \pi$), where one has J equally spaced frequencies of separation Δf_d crowded into a bandwidth W. The probability of error for any *one* of the frequencies† (not situated at the ends) is the area outside of the interval bounded by lines L_2 and L_3 in Fig. 5. If L_2 and L_3 are defined by $| q | = \varphi$ then the probability of error for such a frequency is, from (17a),

$$Pe = \frac{1}{\sqrt{2\pi}} \frac{\cot \varphi/2}{\sqrt{\cos \varphi}} \frac{\exp [-2\rho \sin^2 (\varphi/2)]}{\sqrt{\rho}}, \qquad (28)$$

where, if one assumes that the bandwidth $W = J\Delta f_d$, one would take

$$\varphi = \frac{\pi W T}{J}. \qquad (29)$$

Our requirement that $\Delta f_d T < \pi$ implies $J > 2$ for the narrow-band formula to be applicable (assuming $WT = 1$). Note $\sin^2 (\varphi/2)$ is less than $\frac{1}{2}$, and thus the exponent in (28) is $\exp [-k\rho]$, where $k < 1$. Now the contribution of the clicks to P_e is essentially the area A_L of the first side lobe which is by (26) and (27)

$$A_L = f_d T \exp (-\rho). \qquad (30)$$

* Note the special sense in which the term is uesd here.

† The P_e for a frequency at the end is one-half the expression (28).

But expression (30) is, asymptotically, exponentially small compared to (28). Likewise, the area due to the side lobes caused by end effects is exponentially small, and the probability of error for narrow-band FM is given by (28). The result that the clicks do not asymptotically contribute to errors in narrow-band multilevel FM lends justification to a previous evaluation of this type system by Salz,[11] who considered the special narrow-band system with $WT = 1$. It is both interesting and gratifying that this result is in agreement with the result given in Ref. 11. In a later paper, Salz and Koll[19] report on experimental results which agree with the earlier theoretical results.

Next, consider the asymptotic evaluation of P_e for the case of orthogonal signals; this case corresponds to $(\Delta\omega_d)T = \pi$, and we assume that the thresholds are spaced midway between the frequencies. Thus, (for a frequency not on the edges) we have that the error probability is given by the area outside of that bounded between the lines L_1 and L_4. The contribution from the major lobe is, from (17b),

$$\tfrac{1}{2} \exp(-\rho).$$

In addition, the area of the first side lobe is asymptotically comparable to this and is

$$f_d T \exp(-\rho),$$

being weakly dependent on the frequency sent. In fact, for the nth signal $(J = 2n)$ we have for orthogonal signals that

$$f_d T = \frac{n}{2}, \qquad n = 1, 2, \cdots, \frac{J}{2}.$$

The average error rate is then, for orthogonal signals (J of them, J even, and equally spaced signals and thresholds),

$$P_e = (\tfrac{1}{2}) \exp[-\rho] + (\tfrac{1}{4})(J/2 + 1) \exp(-\rho). \qquad (31)$$

Equation (31) is indeed a surprising result. The first term of (31) is the probability of confusing the transmitted frequency with one of its nearest neighbors. The second term is the (average) probability of confusing it with its second nearest neighbor closest to the carrier. This is because the area from $(-\pi)$ to $(-3\pi/2)$ is, by application of (17b), negligible compared to the area from $(-3\pi/2)$ to $(-5\pi/2)$. Thus, it states that for the multilevel scheme considered here (a not unreasonable one) one is less likely to confuse a transmitted frequency with its nearest neighbors than one is to confuse it with a particular one of its second nearest neighbors. We see from (31) that the error rate from the continuous part of the output is comparable to the error rate caused by clicks.

As a final remark about the orthogonal system we see comparing (31) and (14) why end effects are neglected again.

For a final example, consider the wide-band situation where the signals are loosely packed in the band; i.e., $(\Delta\omega_d)T > \pi$. Now no errors will be caused by the continuous part of the output; only clicks will cause errors. If the frequencies are widely spaced a single click may not cause an error; several clicks during the time interval T may be required. Thus, suppose that the frequencies are spaced so that the phase differences of nearest neighbors is $(\Delta\omega_d)T = 2n\pi$, n being any positive integer. The probability of error will then be the probability of getting n (or more) clicks in time T, which from (26) and (27) behaves as

$$\begin{aligned} \frac{(f_dT)^n \exp(-n\rho)}{n!} &\geqq \frac{(n/2)^n \exp(-n\rho)}{n!} \\ &\geqq \tfrac{1}{2} \exp(-n\rho). \end{aligned} \tag{32}$$

The coefficient in (32) is at least as bad as for the orthogonal case, but the important item is the exponent. Superficially at least it appears that we have gained in performance by spacing the frequencies widely, since the exponential has changed from $e^{-\rho}$ from the minimum orthogonal case ($\Delta\omega_dT = \pi$) to $e^{-n\rho}$. One must realize, however, that one is talking about different ρ's here. The bandwidth for the case under consideration is essentially $2n$ times the minimum orthogonal one and therefore, for the same signal power, the nominal value of ρ has decreased $2n$, and one has in fact not gained a factor of n in the exponent. In addition to the bandwidth penalty, error performance has actually suffered too.

VI. COMPARISON WITH OPTIMUM

One can demonstrate how the FM discriminator compares with the optimum detector when used to detect orthogonal signals; i.e., when $\Delta\omega_dT = \pi$. It is known that when optimum detection is used for any orthogonal set of signals, the (exponential part of the) error rate behaves as $\exp[-E/N_0]$, where E is the signal energy (assumed common to all J levels) and $N_0/2$ is the (two-sided) spectral density of the noise. If we let S denote the average signal power, write $E = ST$, and estimate the total bandwidth W for large J by $W = J/(2T)$, we see that the ideal exponent becomes $\exp[-J\rho/2]$. However, we had seen that, regardless of the number of levels, the discriminator error rate for $\Delta\omega_dT = \pi$ behaves as $\exp[-\rho]$. Thus, we have lost a factor of J in the error exponent by substituting discriminator detection for matched filter detection.

An important conclusion may immediately be drawn concerning the performance of conventional FM receivers or detectors of orthogonal signals. Our results show that the receiver is indeed inferior in performance when compared with the optimum. This fact has been stated by Wozencraft and Jacobs[12] and the reasons are clear from our analysis. The FM receiver admits too much noise at its front-end which cannot be cleaned by the post-detection filter because of the nonlinear anomalies, namely the clicks. As a matter of fact, the amount of noise grows in direct proportion to the number of orthogonal signals, hence the inferior exponent.* The optimum detector is a bank of matched filters. The noise power at the output of each filter does not grow with the number of signals; it is a fixed constant determined by the bandwidth of the filter, which roughly needs be no broader than the symbol rate, $1/T$.

This poor performance of conventional FM receivers when used to detect data *might* be remedied by employing an FM with feedback system such as described in Refs. 20 and 21. The physical argument to support this contention is often stated as follows. In the absence of the feedback loop, the IF filter must be wide enough to pass the total swing of the incoming signal. However, since the feedback loop tracks the incoming frequency, this IF filter, whose width determines the noise variance, could be narrowed and less noise would be admitted.

This possibility of making use of FM with feedback to improve the error rate in digital systems has been suggested by Wozencraft and Jacobs.[12] Unfortunately a mathematical treatment of this difficult problem does not exist at present.

VII. EFFECT OF POST-DETECTION FILTER

In the previous sections we have discussed in detail the performance of an FM discriminator followed by a low-pass filter. The low-pass filter was approximated by an ideal integrator whose integration time was taken to be equal to the duration of an individual signaling interval. Formulas sufficient to evaluate the performance of multilevel FM without the post-detection filter have recently been developed by Mazo and Salz;[10] comparison of the results of the present paper with Ref. 10 will show the influence of filtering.

* Actually, these qualitative conclusions may be arrived at by the Gaussian approximation to the output noise. The reason why this works is apparent from (31) which gives P_e for orthogonal signals. The first term of (31) is not due to clicks but arises from the continuous part of the output noise. This is the part that the Gaussian approximation would tend to duplicate. The second term of (31) is due to clicks and has the same behavior with regard to ρ. Even if one could keep ρ constant as the number of levels J increased, the factor of J in the click contribution to (31) would still degrade performance.

Suppose that the angular frequency $\dot{\psi}$ is sent and we ask for the probability that the observed output is less than z, where $(\dot{\psi} - z) > 0$. It is shown in Ref. 10 that the probability P is essentially given by* (for large ρ)

$$P \cong \exp\left[-\rho\frac{(\dot{\psi} - z)^2}{z^2 + \dot{\sigma}^2/\sigma^2}\right]. \tag{33}$$

Consider the situation for orthogonal signals, or in fact for any signal set where the frequency spacing between the individual frequencies is fixed. One expects the ratio $\dot{\sigma}^2/\sigma^2$ to increase as the square of the total input bandwidth, hence as J^2, the square of the number of levels. Thus, for a large number of orthogonal levels the post-detection filter does very well in improving the error performance, changing the error rate from† (roughly) $\exp(-\rho/J^2)$ to $\exp(-\rho)$. One would certainly expect something like this to be true since, for a large number of levels, the noise bandwidth before the post-detection filter is much greater than the signal bandwidth at that point.

Another qualitative effect of the post-detection filter may be noted. From (33) we see that the distribution of output noise without the post-detection filter depends on the frequency sent, because of the factor $(z^2 + \dot{\sigma}^2/\sigma^2)$ in the exponent; the "spread" of the probability density will be roughly twice as great at the ends of the band than at the center, and thus without a post-detection filter one would not choose the frequencies to be equally spaced. We have seen that there is no such dependence of the error rate exponent on the transmitted frequency when the post-detection filter is present.

VIII. AN APPARENT PARADOX

At this point we have basically concluded our discussion of error rates in digital FM, based in part upon the theory of "clicks" in FM receivers. In particular, we have seen in Section VI that even when frequencies were widely spaced so that $\omega_d T$ is many multiples of 2π the error performance did not improve. The reason was noted to be that although the distance between frequencies increased, the noise admitted to the system increased by a corresponding factor. The latter is predicated on the assumption that the input bandpass filter is essentially a flat filter up to some cutoff frequency determined by the signal spectrum. It may be possible, however, to shape the front-end filter so that increasing the frequency separation does not cause a proportionate increase in the

* Equation (33) represents only the exponential part of P. Also (33) is true [see Ref. 10] only if $(\dot{\psi} - z)^2/[z^2 + \dot{\sigma}^2/\sigma^2] < 1$.

† Set $(\dot{\psi} - z)^2 \approx (\Delta f)^2$, $(\dot{\sigma}^2/\sigma^2) \approx J^2(\Delta f)^2$.

noise power admitted. We know that the power spectrum of the transmitted signal will have peaks at the transmitted frequencies of width of the order $(1/T)$. Suppose we have a notch filter then, with transmittance peaks at the possible frequencies of the appropriate width. The input noise power will be constant and therefore by choosing a large enough separation one can force the probability of error to be arbitrarily small, contradicting optimality considerations for reception of signals against a white Gaussian noise background.

Before giving what we feel is the correct answer to the stated paradox, we wish to explore some other considerations which, on the surface, might resolve the paradox without changing the basic assumptions of the model. One might first object that our argument was too heuristic; is the noise power really constant as the frequency separation increases? To answer this we have performed the following calculations. We have chosen transmitting and receiving filters so that the FM signal is strictly undistorted and then optimized the filters to minimize the variance of the noise admitted. This procedure is discussed in Ref. 11, and the results depend on the power spectrum of the noise. We then specialize to a binary system and, using (48) of Ref. 13 for the spectral density of a binary FSK wave train, calculate the noise admitted. The result shows that while the noise admitted does, in fact, increase as the frequency separation increases, it does so only logarithmically with the separation. Thus, the error probability still will decrease to an arbitrarily small value as the separation increases and from this point of view the question is still unresolved.

A second consideration is the following. The probability of error that we have calculated was based on asymptotic approximations to formulas given in Ref. 6. The results depended only on the amplitude of the received FM wave and the average noise power σ^2 at the input to the limiter-discriminator; if one allows transmitting and receiving filters the more relevant parameters are the average signal power on the line, P_{line}, and σ^2. However, the exact formulas of Rice also involve the quantity $\dot{\sigma}^2$ which is the average power in the derivative of the noise at the input to the limiter-discriminator (after the receiving filter).* Let $S(\omega)$ be the signal spectral density and $F(\omega)$ the transmittance of the receiving filter. Further, let us insist that the signal at the input to the limiter-discriminator be exactly the FSK wave described,† so the transmitting filter is the inverse of the receiving filter. We then have for a white noise background of N_0‡

* Rice, Ref. 6, uses the parameter $r = (1/2\pi)\ (\dot{\sigma}/\sigma)$.

† We emphasize that continuous phase at frequency transition is demanded, but nothing more.

‡ It is for such a noise background that the optimum results are known.

$$\sigma^2 = \frac{N_0}{2\pi} \int_{-\infty}^{\infty} | F(\omega) |^2 \, d\omega \tag{34a}$$

$$\dot{\sigma}^2 = \frac{N_0}{2\pi} \int_{-\infty}^{\infty} \omega^2 \, | F(\omega) |^2 \, d\omega \tag{34b}$$

$$P_{\text{line}} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{S(\omega)}{| F(\omega) |^2} \, d\omega. \tag{34c}$$

When one realizes that the spectrum of an FSK wave decreases at infinity like the fourth power of the frequency,[13] (34b) and (34c) imply that $\dot{\sigma}^2$ and the line power P_{line} cannot both be finite. Thus, suppose $\dot{\sigma}^2$ is finite. The convergence of the integral in (34b) implies that $| F(\omega) |^2$ must decrease at least like $1/\omega^{3+\varepsilon}$, $\varepsilon > 0$. The integral for P_{line} will, for large ω, look like

$$\int \frac{1}{\omega^4} \cdot \omega^{3+\epsilon} d\omega$$

which diverges. Likewise, the assumption of finite line power implies $\dot{\sigma}^2$ is infinite. An infinite $\dot{\sigma}^2$ certainly violates the conditions under which the asymptotic results of Rice's formulas hold. In particular, these formulae show that an infinite $\dot{\sigma}^2$ corresponds to an infinite average number of clicks per second (assuming such a language is still possible) and the FM discriminator will not work, in the strict sense. On the other hand, if we choose the evil of infinite line power then perfect performance is not surprising.

While the above theorem about $\sigma, \dot{\sigma}$, P_{line} is true from a mathematical point of view, it is almost irrelevant from an engineering point of view because it involves discussions of infinitely large frequencies, and does not really eliminate the paradox at all. We need merely precede the limiter-discriminator with a flat filter with a cutoff so high that the signal is *almost* undistorted. Since real discriminators work, this is not an unreasonable thing to assume. Now $\dot{\sigma}$ is finite, and although we may have to go to extremely large S/N ratios, the paradox is as entrenched as ever.

The resolution of the problem lies in a reinterpretation of Rice's calculation of the average number of crossings of the negative x-axis. We had assumed each crossing corresponds to an encirclement of the origin which is independent of all past and future encirclements. This is reasonable when the receiving filter is essentially flat across the whole received spectrum and the correlation time out of the receiving filter is small ($\sim 1/W$). However, if the input noise spectrum is chopped into a few slits or notches, correlations in the noise being processed in the detector will persist for a longer time and multiple encirclements of the origin can occur with essentially the same probability that one would normally associate with a single large excursion close to the origin.

To make our arguments more precise we consider a binary situation at almost zero rate, i.e., we have very narrow filters F_1 and F_2 about the frequencies $(\omega_c + \omega_d)$ and $(\omega_c - \omega_d)$, respectively. The bandwidth of these individual filters is of order $1/T$. The noise out of F_1 and F_2 can be written as

$$\begin{aligned} n_1(t) &= n_{1x}(t) \cos (\omega_c + \omega_d)t - n_{1y}(t) \sin (\omega_c + \omega_d)t \\ n_2(t) &= n_{2x}(t) \cos (\omega_c - \omega_d)t - n_{2y}(t) \sin (\omega_c - \omega_d)t, \end{aligned} \tag{35}$$

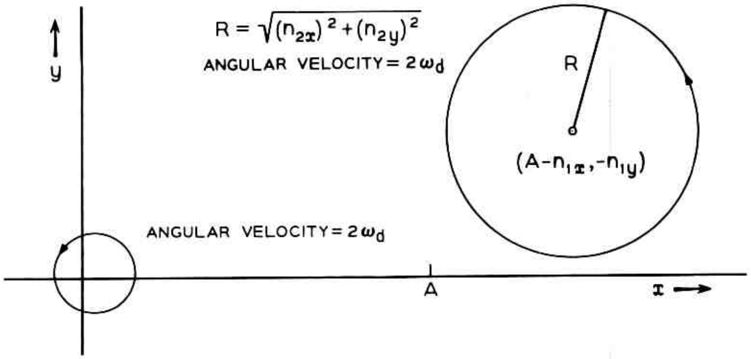
where $n_{1x}(t)$, etc., are independent baseband noise currents. If we assume that the frequency $(\omega_c + \omega_d)$ is being transmitted with amplitude A, then in a "coordinate system" following that frequency we have

$$\begin{aligned} x &= A + X \\ y &= Y, \end{aligned} \tag{36}$$

where

$$\begin{aligned} X &= n_{1x} + n_{2x} \cos 2\omega_d t + n_{2y} \sin 2\omega_d t \\ Y &= n_{1y} + n_{2y} \cos 2\omega_d t - n_{2x} \sin 2\omega_d t. \end{aligned} \tag{37}$$

A typical portion of the path that the noise traces out in the xy plane can be calculated from (36) and (37) and is shown in Fig. 6. Neglecting the time variations of $n_{1x}(t)$, etc., which vary on a time scale comparable



R = $\sqrt{(n_{2x})^2 + (n_{2y})^2}$
ANGULAR VELOCITY = $2\omega_d$
R
$(A-n_{1x}, -n_{1y})$
ANGULAR VELOCITY = $2\omega_d$
y
A
x

Fig. 6 — Small portions of some noise trajectories when receiving filter has two transmittance peaks.

to T, we see the path is a circle centered at $(A - n_{1x}, -n_{1y})$, of radius $\sqrt{n_{2x}^2 + n_{2y}^2}$, and counter-clockwise angular velocity of $(2\omega_d)$. If σ_1^2 and σ_2^2 denote the average noise powers out of F_1 and F_2, respectively, then the probability P that the circle is appropriately situated with a large enough radius to encircle the origin is given exactly by

$$P = 2f_d \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \exp(-\rho), \tag{38}$$

where $\rho = A^2/2(\sigma_1^2 + \sigma_2^2)$. For the case of a symmetrical spectrum about the carrier $(\sigma_1^2 = \sigma_2^2)$, (38) is comparable to (26). However, our circle is rotating with frequency $2f_d$ and will have a constant radius for about T seconds; thus, it will complete $2f_dT$ revolutions in time T. As the frequencies are spread and notched filters are used the noise indeed does not increase proportionally, but the number of multiple encirclements of the origin that a click will make does increase as the separation. Thus, the filter shaping under discussion will affect the statistical structure of the clicks, preventing a violation of optimality.

Note added in Proof. A discussion of the click contribution to the error rate has been given very recently by J. Klapper in the RCA Review, June, 1966.

APPENDIX

Asymptotic Behavior of P_φ

We wish to record here an outline* of the details of the evaluation of (16) for large S/N so as to obtain the results given in (17). If we set

$$p(x) = \frac{1}{\pi} \exp(-\rho) + \sqrt{\frac{\rho}{\pi}} \cos x \exp(-\rho \sin^2 x) \cdot \operatorname{erf}(\sqrt{\rho} \cos x), \tag{39}$$

then according to (16) the required probability is written as

$$P_\varphi = \int_{-(\pi/2)}^{(\pi/2-\varphi)} dy \int_{(y+\varphi)}^{(\pi/2)} dx p(y) p(x). \tag{40}$$

If we define the distribution function

$$P(\xi) = \int_{-(\pi/2)}^{\xi} p(y) dy \tag{41}$$

* We do not explain the techniques used here for the asymptotic evaluation of integrals. The interested reader may wish to consult the subjects "saddle point method," "Laplace's method," "Watson's lemma" in Ref. 22.

and perform and integration by parts, (40) becomes

$$P_\varphi = \int_{-(\pi/2)}^{\pi/2-\varphi} P(y)p(y+\varphi)dy. \tag{42}$$

Our evaluation will be based upon approximating the functions $P(y)$ and $p(y)$ when ρ is large. In particular, from (39) we see

$$p(y) \sim \sqrt{\frac{\rho}{\pi}} \cos y \exp\,(-\rho \sin^2 y), \tag{43}$$

provided y is not close to $\pm\pi/2$. Integrating (43) yields

$$P(y) \sim \tfrac{1}{2}[\operatorname{erf} \sqrt{\rho} + \operatorname{erf}\,(\sqrt{\rho} \sin y)], \tag{44}$$

which will be a good approximation for large ρ except when y is near $-\pi/2$. These exceptional points will receive special consideration.

As a first example consider the case when $\varphi < \pi/2$. The integrand for (42) is shown symbolically in Fig. 7. Consider the contribution first from negative y. This is from (42), (43), and (44)

$$\frac{1}{2}\int_{-(\pi/2)}^{0} dy[\operatorname{erf}\sqrt{\rho} - \operatorname{erf}\,(\sqrt{\rho}\sin|\,y\,|)] \sqrt{\frac{\rho}{\pi}} \cos\,(y+\varphi) \cdot \exp\,[-\rho \sin^2\,(y+\varphi)]. \tag{45}$$

Next, approximate erf $\sqrt{\rho}$ by unity to obtain

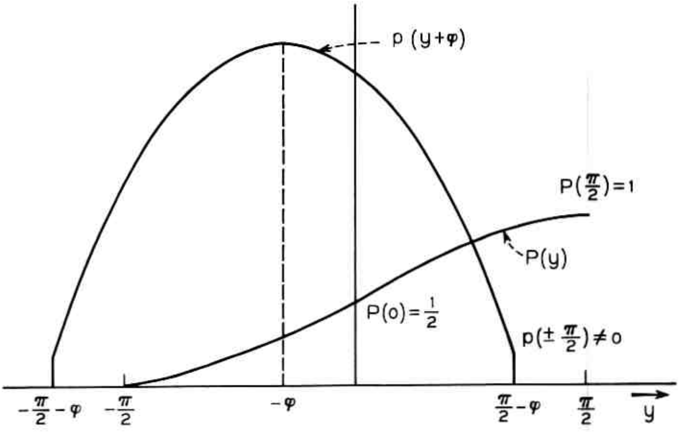


Fig. 7 — Symbolic representation of the factors in the integrand of (42) drawn for $\varphi < \pi/2$.

$$\frac{1}{2}\int_{-(\pi/2)}^{0} dy \operatorname{erfc}\left(\sqrt{\rho}\sin|y|\right)\sqrt{\frac{\rho}{\pi}}\cos(y+\varphi) \cdot\exp\left[-\rho\sin^2(y+\varphi)\right] \tag{46}$$

and use the asymptotic expansion

$$\operatorname{erfc} x \sim \frac{1}{\sqrt{\pi}x}\exp(-x^2). \tag{47}$$

The resultant integrand has a saddle point at $y = -\varphi/2$, and a routine saddle point evaluation will yield (17a) of the text. It is easy to verify that the error made by replacing erf $\sqrt{\rho}$ by unity in (45) creates an asymptotically small error. Likewise the neglect of positive y is asymptotically small for

$$\begin{aligned}\int_0^{(\pi/2)-\varphi} dyP(y)p(y+\varphi) \leqq \int_0^{(\pi/2)-\varphi} dy\,p(y+\varphi) &\leqq \left[\frac{\pi}{2}-\varphi\right]\left(\frac{1}{\pi}\right) \\ &\cdot\exp(-\rho) + \int_0^{(\pi/2)-\varphi} dy\sqrt{\frac{\rho}{\pi}} \\ &\cdot\cos(y+\varphi)\exp\left[-\rho\sin^2(y+\varphi)\right] \\ &\sim \frac{\exp\left[-\rho\sin^2\varphi\right]}{2\sqrt{\pi\rho}\sin\varphi},\end{aligned} \tag{48}$$

by Laplace's method. For $\varphi < \pi/2$ we have

$$2\sin^2(\varphi/2) < \sin^2\varphi$$

which proves our point. The addition of the term $(1/\pi)\exp(-\rho)$ in (48) provides a strict upper bound to $p(y+\varphi)$ and thus takes care of special considerations at the right end of $p(y+\varphi)$. At the left end point of the range of integration, $(-\pi/2)$, $p(y+\varphi)$ is still well approximated. The function $P(y)$ is, however, approximately

$$P(y) \approx \frac{\exp(-\rho)}{\pi}\left[\frac{\pi}{2}+y\right], \qquad y \approx -\frac{\pi}{2}. \tag{49}$$

Using (49) it is easy to obtain an estimate of the contribution of the left end point behaving as $\exp(-\rho)$ and this is asymptotically small. This ends our discussion for $\varphi < \pi/2$.

We give a somewhat more condensed outline for $\varphi = \pi/2$. The contribution of the middle of the range of integration is again approximated by (46) with $\varphi = \pi/2$. Using (47) and (46) immediately evaluates to

$$[\exp(-\rho)]/4.$$

Next, consider the error made at the right end point. Equation (43) holds to within a strip of order $1/\sqrt{\rho}$ from 0, after which $p(y + \varphi)$ behaves like $[\exp(-\rho)]/\pi$. Therefore, the error behaves like

$$\frac{1}{2}\cdot\frac{\exp(-\rho)}{\pi}\cdot\frac{1}{\sqrt{\rho}},$$

which is asymptotically small.

The left end point error is bounded by

$$-\int_{-(\pi/2)}^{0}\frac{\exp(-\rho)}{\pi}\left(\frac{\pi}{2}+y\right)\sqrt{\frac{\rho}{\pi}}\sin y\exp(-\rho\cos^2 y)dy\sim\frac{\exp(-\rho)}{\pi\sqrt{\pi\rho}},$$

which is again asymptotically small.

Our final case is $\varphi > \pi/2$, and this time end point contributions will not be small. The reason is that if one examines the integral representing the contribution from the middle of the range of integration, i.e.,

$$\frac{1}{2}\int_{-(\pi/2)}^{(\pi/2)-\varphi} dy[\operatorname{erfc}\sqrt{\rho}\sin|y| - \operatorname{erfc}\sqrt{\rho}]\sqrt{\frac{\rho}{\pi}}\cos(y+\varphi) \times \exp[-\rho\sin^2(y+\varphi)], \tag{50}$$

it is exponentially dominated by contributions near the end points. But in (50) our approximation to $P(y)$ vanishes faster than the correct $P(y)$ at $y = -\pi/2$, and our approximation to $p(y + \varphi)$ vanishes at $y = (\pi/2 - \varphi)$ while the true $p(y + \varphi)$ does not. This implies that the asymptotic evaluation of (50) will be asymptotically smaller than the correct contributions from the ends of the interval. The contribution from the right end is

$$\frac{1}{2}\int_{-(\pi/2)}^{(\pi/2)-\varphi} dy\,[\operatorname{erf}\sqrt{\rho} - \operatorname{erf}\sqrt{\rho}\sin|y|]\frac{\exp(-\rho)}{\pi} \sim \frac{\exp[-\rho(1+\cos^2\varphi)]}{4\pi\sqrt{\pi}\,\rho\sqrt{\rho}\sin\varphi\cos^2\varphi}. \tag{51}$$

The lower limit of integration (51) is immaterial, as will be the upper limit in (52). For the contribution from the left end we have

$$\frac{e^{-\rho}}{\pi}\int_{-(\pi/2)}^{(\pi/2)-\varphi} dy\left[\frac{\pi}{2}+y\right]\sqrt{\frac{\rho}{\pi}}\cos(y+\varphi)\exp[-\rho\sin^2(y+\varphi)] \sim \frac{\exp -\rho(1+\cos^2\varphi)]}{4\pi\sqrt{\pi}\,\rho\sqrt{\rho}\sin\varphi\cos^2\varphi}. \tag{52}$$

The sum of (52) and (51) yields (17c) of the text.

REFERENCES

1. Carson, J. R. and Fry, T. C., Variable Frequency Electric Circuit Theory with Application to the Theory of Frequency-Modulation, B.S.T.J., *16*, October, 1937, pp. 513–540.
2. Crosby, M. G., Frequency-Modulation Noise Characteristics. Proc. IRE, *25*, 1937, pp. 472–514.
3. Blachman, N. M., The Demodulation of a Frequency-Modulated Carrier and Random Noise by a Limiter and Discriminator. J. Appl. Phys., *20*, 1949, p. 38, 976.
4. Lawson, J. L. and Uhlenbeck, G. E., *Threshold Signals*, MIT Radiation Laboratory Series, *24*, McGraw-Hill Book Co., New York, 1950, Chap. 13.
5. Middleton, David, *Statistical Communication Theory*, McGraw-Hill Book Co., New York, 1960, Chap. 15.
6. Rice, S. O., Noise in FM Receivers, Chapter 25 of *Time Series Analysis*, M. Rosenblatt, (ed.), John Wiley & Sons, Inc., New York, 1963.
7. Cohn, John, Proc. N. E. C., (Chicago) *12*, 1956, pp. 221–236. We would like to thank S. O. Rice for pointing out this reference to us.
8. Bennett, W. R. and Salz, J., Binary Data Transmission by FM Over a Real Channel, B.S.T.J., *42*, 1963, pp. 2387–2426.
9. Salz, J. and Stein, S., Distribution of Instantaneous Frequency for Signal Plus Noise, IEEE Trans., *IT-10*, 1964, pp. 272–274.
10. Mazo, J. E. and Salz, J., Probability of Error for Quadratic Detectors, B.S.T.J., *44*, November, 1965, pp. 2165–2186.
11. Salz, J., Performance of Multilevel Narrow-Band FM Digital Communication Systems, IEEE Trans., *COM-13*, 1965, pp. 420–424.
12. Wozencraft, J. M. and Jacobs, I. M., *Principles of Communication Engineering*, John Wiley & Sons, Inc., New York, 1965.
13. Bennett, W. R. and Rice, S. O., Spectral Density and Autocorrelation Functions Associated with Binary Frequency Shift Keying, B.S.T.J., *42*, September, 1963, pp. 2355–2385.
14. Anderson, R. R. and Salz, J., Spectra of Digital FM, B.S.T.J., *44*, July–August, 1965, pp. 1165–1189.
15. Bennett, W. R. and Davey, J. R., *Data Transmission*, McGraw-Hill Book Co., New York, 1965.
16. Blachman, Nelson M., FM Reception and the Zeros of Narrow-Band Gaussian Noise, IEEE Trans., *IT-10*, 1964, pp. 235–241.
17. Bennett, W. R., Methods of Solving Noise Problems, Proc. IRE, *44*, 1956, pp. 609–638. See (253).
18. National Bureau of Standards, Handbook of Mathematical Functions, U. S. Government Printing Office, Washington, D.C., 1964.
19. Salz, J. and Koll, V. G., An Experimental Digital Multilevel FM Modem, IEEE Trans., *COM-14*, 1966, pp. 259–265.
20. Chaffee, J. G., The Application of Negative Feedback to Frequency Modulation Systems, B.S.T.J., *18*, July, 1939, pp. 403–437.
21. Enloe, L. H., Decreasing the Threshold in FM by Frequency Feedback, Proc. IRE, *50*, 1962, pp. 18–30.
22. Copson, E. T., *Asymptotic Expansions*, Cambridge University Press, 1965.

Noise in an FM System Due to an Imperfect Linear Transducer

By M. L. LIOU

(Manuscript received February 23, 1966)

An approach to the calculation of intermodulation noise in FM systems due to imperfect transmission media is presented in this paper. The technique is essentially that originating with Carson and Fry. In this paper we extend a formulation due to Rice to include an arbitrary continuous pre-emphasis characteristic as well as an arbitrary gain and phase shape transmission medium which are representable by low-order polynomial series in radian frequency. Series approximations are carried out far enough to ensure accurate results for transmission characteristics normally encountered in broadband microwave radio systems. In many cases, only the second- and third-order noise is significant in broadband microwave radio systems. Hence, the analysis carried out in this paper considers only the second- and third-order distortion terms. A digital computer program concerning the intermodulation noise has been written. This analysis and the digital computer program are of aid in the design of microwave radio systems. With a slight modification, the calculation of noise due to AM-to-PM conversion caused by transmission deviation can also be accomplished. An optimum design of the pre-emphasis network may be achieved by using the computer programs through an iterative approach.

I. INTRODUCTION

In the course of designing FM systems, intermodulation noise is an important factor which deserves special attention. Many people have made contributions to this subject.[1-10] In this paper, we extend their results to include an arbitrary continuous pre-emphasis characteristic as well as an arbitrary gain and phase shape transmission medium which are representable by low-order polynomial series in radian frequency. Series approximations are carried out far enough to ensure accurate results for transmission characteristics normally encountered in broadband microwave radio systems. The multichannel baseband signal of an

FM system is represented by a band of gaussian random noise with flat power-density spectrum. The noise due to imperfect transmission medium can be calculated at any frequency in the baseband. In many cases, only the second- and third-order noise is significant in broadband microwave radio systems. Hence, the analysis carried out in this paper considers only the second- and third-order distortion terms. Extension to a higher-order distortion becomes unmanageable. A digital computer program concerning the intermodulation noise has been written. A typical problem can be solved at a very low cost. This analysis and the digital computer program are of aid in the design of microwave radio systems. With a slight modification, the calculation of noise due to AM-to-PM conversion caused by transmission deviation can also be accomplished. An optimum design of the pre-emphasis network may be achieved by using the computer programs through an iterative approach. Several examples are given for illustration.

II. DESCRIPTION OF SYSTEM

A portion of an FM system can be represented by the block diagram shown in Fig. 1. An FM baseband signal, $\varphi_i'(t) = d\varphi_i(t)/dt$, is fed into a pre-emphasis network. The output of the pre-emphasis network is the pre-emphasized signal $\varphi'(t)$. This signal passes through an FM modulator to yield the FM wave, cos $[\omega_c t + \varphi(t)]$ where ω_c is the angular carrier frequency. When this FM wave goes through an imperfect transmission medium, the output is distorted and becomes $V(t)$ cos $[\omega_c t + \varphi_0(t)]$. Let the transfer function of the transmission medium be

$$Y(\omega) = \exp[-\alpha(f) - i\beta(f)], \tag{1}$$

where $\omega = 2\pi f$. Its impulse response is

$$g(t) = \int_{-\infty}^{\infty} Y(\omega) \exp(i\omega t)\, df. \tag{2}$$

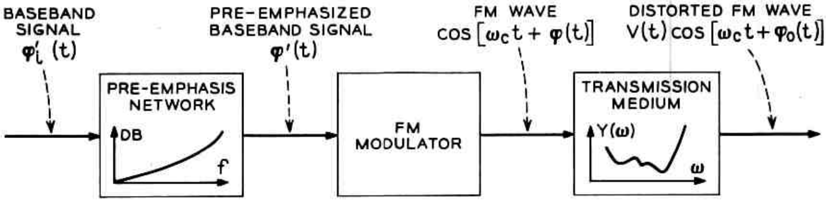

BASEBAND SIGNAL $\varphi_i'(t)$
PRE-EMPHASIZED BASEBAND SIGNAL $\varphi'(t)$
FM WAVE cos $[\omega_c t + \varphi(t)]$
DISTORTED FM WAVE $V(t)$ cos $[\omega_c t + \varphi_0(t)]$
PRE-EMPHASIS NETWORK
DB
f
FM MODULATOR
TRANSMISSION MEDIUM
$Y(\omega)$
ω

Fig. 1 — Block diagram of a portion of an FM system.

Using the complex notation, the input and output of the transmission medium can be written, respectively, as

$$v_i(t) = \exp\,[i\omega_c t + i\varphi(t)] \tag{3}$$

and

$$v_0(t) = V(t)\exp\,[i\omega_c t + i\varphi_0(t)], \tag{4}$$

where $V(t)$ is taken to be positive and $\varphi_0(t)$ is determined to within $2n\pi$, n is an integer.

The output and input of the transmission medium are related by the convolution integral

$$v_0(t) = \int_{-\infty}^{\infty} v_i(t-x)\,g(x)\,dx. \tag{5}$$

From (3), (4), and (5)

$$V(t)\exp\,[i\varphi_0(t)] = \int_{-\infty}^{\infty} \exp\,[i\varphi\,(t-x) - i\omega_c x]\,g(x)\,dx. \tag{6}$$

Let

$$V(t) = \exp\,[a(t)]. \tag{7}$$

Then

$$a(t) = \operatorname{Re}\ln\int_{-\infty}^{\infty} \exp\,[i\varphi(t-x) - i\omega_c x]\,g(x)\,dx \tag{8}$$

$$\varphi_0(t) = \operatorname{Im}\ln\int_{-\infty}^{\infty} \exp\,[i\varphi(t-x) - i\omega_c x]\,g(x)\,dx. \tag{9}$$

The AM distortion term expressed in dB is 20 × 0.4343 $a(t)$ and the PM distortion term expressed in radians (or degrees) is $\varphi_0(t) - \varphi(t)$.

Assume that the transmission medium passes only frequencies in the neighborhood of the carrier frequency, $\pm f_c \pm b$, with $b/f_c \ll 1$. Thus, to a high degree of approximation, we have

$$\exp\,(-i\omega_c x)\,g(x) \cong \int_{-b}^{b} Y(\omega_c + \omega)\exp\,(i\omega x)\,df. \tag{10}$$

Let

$$k(x) = \frac{1}{Y(\omega_c)}\exp\,(-i\omega_c x)\,g(x). \tag{11}$$

Substituting (1) and (11) into (8) and (9) we obtain

$$a(t) = -\alpha(f_c) + \operatorname{Re} \ln \int_{-\infty}^{\infty} \exp [i\varphi(t - x)] \, k(x) \, dx \tag{12}$$

$$\varphi_0(t) = -\beta(f_c) + \operatorname{Im} \ln \int_{-\infty}^{\infty} \exp [i\varphi(t - x)] \, k(x) \, dx. \tag{13}$$

The quantity $k(x)$ may be regarded as the normalized envelope function of the impulse response $g(x)$. We shall discuss now the logarithm of the integral in (12) and (13) in detail.

III. DERIVATION OF DISTORTION TERMS

Let

$$\exp [i\varphi(t - x)] = M(t,x). \tag{14}$$

A delay t_d is often introduced in order to improve the degree of approximation of various series with the first few terms. Then $M(t,x)$ can be expanded about $x = t_d$ as

$$\begin{aligned} M(t,x) = M(t,t_d) &+ \frac{(x - t_d)}{1!} \left[\frac{\partial}{\partial x} M(t,x) \right]_{x=t_d} \\ &+ \frac{(x - t_d)^2}{2!} \left[\frac{\partial^2}{\partial x^2} M(t,x) \right]_{x=t_d} + \cdots. \end{aligned} \tag{15}$$

One choice of t_d is

$$t_d = \operatorname{Re} \int_{-\infty}^{\infty} x k(x) \, dx = \frac{1}{2\pi} \left[\frac{d\beta(f)}{df} \right]_{f=f_c}. \tag{16}$$

Using (14) and (15), the integral in (12) and (13) can thus be expressed as

$$\int_{-\infty}^{\infty} \exp [i\varphi(t - x)] \, k(x) dx = \sum_{n=0}^{\infty} \frac{m_n}{n!} \left[\frac{\partial^n}{\partial x^n} M(t,x) \right]_{x=t_d}, \tag{17}$$

where m_n is the nth moment of $k(x)$ defined as

$$m_n = \int_{-\infty}^{\infty} (x - t_d)^n \, k(x) dx, \qquad m_0 = 1 \tag{18}$$

or equivalently,

$$m_n = \frac{(-1)^n}{Y(\omega_c)} \left[\frac{d^n}{d(i\omega)^n} Y(\omega_c + \omega) \exp (i\omega t_d) \right]_{\omega=0}. \tag{19}$$

The series (17) is equivalent to the Carson-Fry series with delay t_d. It may not converge in certain cases. However, when the characteristic of a transmission medium can be truly represented by a polynomial, from (19), the higher moments become zero and the series reduces to a polynomial. The logarithm of (17) can be written as

$$\begin{aligned}\ln \int_{-\infty}^{\infty} \exp[i\varphi(t - x)]\, k(x)dx &= i\varphi(t - t_d)\\ &\quad + \ln\left[1 + \sum_{n=1}^{\infty} \frac{(-1)^n m_n}{n!} F_n(t - t_d)\right],\end{aligned} \tag{20}$$

where

$$F_n(t - t_d) = (-1)^n \exp\left[-i\varphi(t - t_d)\right]\left[\frac{\partial^n}{\partial x^n} M(t,x)\right]_{x=t_d}.$$

Using Taylor's series expression, we have*

$$\begin{aligned}\ln\left(1 + \sum_1^{\infty} \alpha_n(t)x^n/n!\right) &= \frac{x}{1!}\alpha_1 + \frac{x^2}{2!}(\alpha_2 - \alpha_1^2)\\ &\quad + \frac{x^3}{3!}(\alpha_3 - 3\alpha_1\alpha_2 + 2\alpha_1^3)\\ &\quad + \frac{x^4}{4!}(\alpha_4 - 4\alpha_1\alpha_3 - 3\alpha_2^2 + 12\alpha_1^2\alpha_2 - 6\alpha_1^4) + \cdots.\end{aligned}$$

With $x = -1$ and $\alpha_n(t) = m_n F_n(t - t_d)$, substituting the above series into (20), after considerable algebra, one obtains an asymptotic series which has been written by S. O. Rice in an unpublished work as

$$\begin{aligned}\ln \int_{-\infty}^{\infty} \exp\left[i\varphi(t - x)\right] k(x)dx = i\Big[&\varphi - \frac{m_1}{1!}\varphi' + \frac{m_2}{2!}\varphi'' - \frac{m_3}{3!}\varphi'''\\ &+ \frac{m_4}{4!}\varphi'''' - \cdots + \tfrac{1}{6}\lambda_3\varphi'^3\\ &- \tfrac{1}{4}(m_4 - 2m_1m_3 - m_2^2\\ &+ 2m_1^2m_2)\varphi'^2\varphi'' + \cdots\Big]\\ &+ [-\tfrac{1}{2}\lambda_2\varphi'^2 + \tfrac{1}{2}(m_3\\ &- m_1m_2)\varphi'\varphi'' - \tfrac{1}{6}(m_4\\ &- m_1m_3)\varphi'\varphi''' - \tfrac{1}{8}(m_4\\ &- m_2^2)\varphi''^2 + \tfrac{1}{24}\lambda_4\varphi'^4 + \cdots],\end{aligned} \tag{21}$$

* Notice that this is not the approximation $\ln(1 + y) \cong y$ which can be a poor approximation and yet has been quite widely used in FM work.

where φ stands for $\varphi(t - t_d)$ and λ_n's are the semi-invariants which are related to the moments by

$$\begin{aligned}
\lambda_2 &= m_2 - m_1^2 \\
\lambda_3 &= m_3 - 3m_1m_2 + 2m_1^3 \\
\lambda_4 &= m_4 - 4m_1m_3 - 3m_2^2 + 12m_1^2m_2 - 6m_1^4.
\end{aligned}$$

Taking the real and imaginary parts of (21) and substituting the result into (12) and (13) gives, respectively,

$$\begin{aligned}
a(t) = &- \alpha(f_c) + m_{1i}\varphi' - \frac{m_{2i}}{2!}\varphi'' + \frac{m_{3i}}{3!}\varphi''' - \frac{m_{4i}}{4!}\varphi'''' + \cdots \\
&- \frac{\lambda_{3i}}{6}\varphi'^3 + \frac{l_{1i}}{4}\varphi'^2\varphi'' - \frac{\lambda_{2r}}{2}\varphi'^2 + \frac{l_{2r}}{2}\varphi'\varphi'' \qquad (22) \\
&- \frac{l_{3r}}{6}\varphi'\varphi''' - \frac{l_{5r}}{8}\varphi''^2 + \frac{\lambda_{4r}}{24}\varphi'^4 + \cdots
\end{aligned}$$

$$\begin{aligned}
\varphi_0(t) = &-\beta(f_c) + \varphi - m_{1r}\varphi' + \frac{m_{2r}}{2!}\varphi'' - \frac{m_{3r}}{3!}\varphi''' + \frac{m_{4r}}{4!}\varphi'''' - \cdots \\
&+ \frac{\lambda_{3r}}{6}\varphi'^3 - \frac{l_{1r}}{4}\varphi'^2\varphi'' - \frac{\lambda_{2i}}{2}\varphi'^2 + \frac{l_{2i}}{2}\varphi'\varphi'' - \frac{l_{3i}}{6}\varphi'\varphi''' \qquad (23) \\
&- \frac{l_{5i}}{8}\varphi''^2 + \frac{\lambda_{4i}}{24}\varphi'^4 + \cdots,
\end{aligned}$$

where the subscripts r and i denote the real and imaginary parts of the corresponding coefficients and

$$\begin{aligned}
l_1 &= m_4 - 2m_1m_3 - m_2^2 + 2m_1^2m_2\,, \\
l_2 &= m_3 - m_1m_2\,, \\
l_3 &= m_4 - m_1m_3\,, \\
l_5 &= m_4 - m_2^2.
\end{aligned}$$

From (22) and (23), amplitude and phase distortions are divided into linear, second- and third-order terms and are shown in Table I. The terms $-\alpha(f_c)$ and $-\beta(f_c)$ in (22) and (23) do not appear in Table I. This is due to the fact that they are constants and introduce only constant amounts of amplitude and phase distortions.

IV. PRE-EMPHASIS CHARACTERISTIC

The multichannel baseband signal of an FM system is represented by a band of random noise. Assuming that the bottom baseband fre-

TABLE I — AMPLITUDE AND PHASE DISTORTIONS DUE TO IMPERFECT TRANSMISSION MEDIUM

Order of Distortion	Amplitude Distortion $a(t)$
Linear	$m_{1i}\varphi' - \frac{m_{2i}}{2!}\varphi'' + \frac{m_{3i}}{3!}\varphi''' - \frac{m_{4i}}{4!}\varphi'''' + \cdots$
Second-order	$-\frac{\lambda_{2r}}{2}\varphi'^2 + \frac{l_{2r}}{2}\varphi'\varphi'' - \frac{l_{3r}}{6}\varphi'\varphi''' - \frac{l_{5r}}{8}\varphi''^2 + \cdots$
Third-order	$-\frac{\lambda_{3i}}{6}\varphi'^3 + \frac{l_{1i}}{4}\varphi'^2\varphi'' + \cdots$
	Phase Distortion $\varphi_0(t) - \varphi(t)$
Linear	$-m_{1r}\varphi' + \frac{m_{2r}}{2!}\varphi'' - \frac{m_{3r}}{3!}\varphi''' + \frac{m_{4r}}{4!}\varphi'''' - \cdots$
Second-order	$-\frac{\lambda_{2i}}{2}\varphi'^2 + \frac{l_{2i}}{2}\varphi'\varphi'' - \frac{l_{3i}}{6}\varphi'\varphi''' - \frac{l_{5i}}{8}\varphi''^2 + \cdots$
Third-order	$\frac{\lambda_{3r}}{6}\varphi'^3 - \frac{l_{1r}}{4}\varphi'^2\phi'' + \cdots$

quency is much smaller than the top baseband frequency, the power-density spectrum of the baseband FM signal is expressed as

$$S_{\varphi_i'}(\omega) = P_0, \qquad |f| \leqq f_b,$$

where f_b is the top baseband frequency.

A pre-emphasis network is used in an FM system in order to optimize the noise across the baseband. Let $Z(\omega)$ be the transfer function of the pre-emphasis network, we write

$$|Z(\omega)|^2 = a_0 + a_2f^2 + a_4f^4 + a_6f^6, \qquad |f| \leqq f_b,$$

where the a's are real constants either given *a priori* or determined by the least squares fitting from an actual curve. The power-density spectrum of the pre-emphasized baseband FM signal is

$$\begin{aligned} S_{\varphi'}(\omega) &= |Z(\omega)|^2 S_{\varphi_i'}(\omega) \\ &= P_0(a_0 + a_2f^2 + a_4f^4 + a_6f^6), \qquad |f| \leqq f_b. \end{aligned} \tag{24}$$

In the case when the pre-emphasis coefficients a_0, a_2, a_4, and a_6 are determined by the least squares fitting from an actual curve, the following weighting function of normalized scale has been found useful for better approximations near the bottom channels

$$W(f/f_b) = 10^{-(z/10)[(f/f_b)-1]},$$

where Z is the difference of relative power (in dB) of top and bottom channels of the given pre-emphasis characteristic.

The rms frequency deviation, σ, due to noise loading can be expressed as

$$(2\pi\sigma)^2 = \text{ave}\,[\varphi'^2(t)] = \int_{-\infty}^{\infty} S_{\varphi'}(\omega)\, df.$$

Using (24), the relation between P_0 and σ can be expressed as

$$P_0 = \frac{(2\pi\sigma)^2}{2f_b\left[a_0 + \left(\frac{a_2 f_b^2}{3}\right) + \left(\frac{a_4 f_b^4}{5}\right) + \left(\frac{a_6 f_b^6}{7}\right)\right]}, \qquad (\text{rad/sec})^2/\text{Hz},$$

where the units of σ and f_b are Hz.

V. TRANSMISSION MEDIUM

Within the band of interest, $f_c \pm b$, let the gain and phase of the transmission medium be, respectively,

$$\exp\,[-\alpha(f + f_c)] = 1 + g_1\omega + g_2\omega^2 + g_3\omega^3 + g_4\omega^4 + \sum_{k=1}^{N} u_k \cos\,(p_k\omega + \theta_k) \tag{25}$$

$$-\beta(f + f_c) = b_2\omega^2 + b_3\omega^3 + b_4\omega^4 + \sum_{k=1}^{N} v_k \sin\,(q_k\omega + \sigma_k), \tag{26}$$

where the g's and b's represent coarse shape transmission deviation; the u's and v's represent fine shape transmission deviation. It should be emphasized that the fine shape transmission deviation is restricted to be a slowly varying ripple characteristic, hence this analysis does not apply to the noise due to single echo in general. The fine shape representation is useful when we study the effect on the noise of a given system due to the shift of the carrier. Since a constant delay does not introduce intermodulation noise, the linear term in ω is not included in (26). The transmission medium coefficients g, b, u, v, p, q, θ, and σ in (25) and (26) are either given directly or obtained approximately by a curve-fitting computer program.

Substituting (26) into (16), we have

$$t_d = -\sum_{k=1}^{N} v_k q_k \cos \sigma_k\,.$$

From (19), we can evaluate various moments, hence, the coefficients associated with the distortion terms in Table I in terms of transmission medium coefficients, as given in Appendix A. In the following sections we shall derive expressions for intermodulation noise calculation due to second- and third-order distortion terms.

VI. NOISE POWER DUE TO SECOND-ORDER DISTORTION TERM

Since

$$\frac{d}{dt}\varphi'^2 = 2\varphi'\varphi'',$$

$$\frac{d^2}{dt^2}\varphi'^2 = 2\varphi'\varphi''' + 2\varphi''^2,$$

the second-order PM distortion in Table I can be written approximately as

$$\varphi_2(t) = \left(-\tfrac{1}{2}\lambda_{2i} + \tfrac{1}{4}l_{2i}\frac{d}{dt} - \tfrac{1}{12}l_{3i}\frac{d^2}{dt^2}\right)\varphi'^2 + \tfrac{1}{24}l_{4i}\varphi''^2, \qquad (27)$$

where

$$l_{4i} = 4l_{3i} - 3l_{5i}\,.$$

The second-order PM distortion term of (27) can be represented by the block diagram shown in Fig. 2, where

$$H_1(\omega) = (\tfrac{1}{12}l_{3i}\omega^2 - \tfrac{1}{2}\lambda_{2i}) + i(\tfrac{1}{4}l_{2i}\omega),$$

$$H_2(\omega) = \tfrac{1}{24}l_{4i}\,.$$

The power-density spectrum of $\varphi_2(t)$ is[11]

$$\begin{aligned} S_{\varphi_2}(\omega) = {} & H_1(-\omega)H_1(\omega)S_{\varphi'^2}(\omega) + H_1(-\omega)H_2(\omega)S_{\varphi'^2\varphi''^2}(\omega) \\ & + H_2(-\omega)H_1(\omega)S_{\varphi''^2\varphi'^2}(\omega) + H_2(-\omega)H_2(\omega)S_{\varphi''^2}(\omega), \end{aligned} \qquad (28)$$

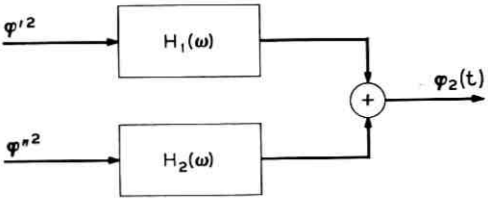


Fig. 2 — Block diagram representation of the second-order PM distortion term.

where $S_{\varphi'^2}(\omega)$ and $S_{\varphi''^2}(\omega)$ are the power-density spectra of φ'^2 and φ''^2, respectively, and $S_{\varphi'^2\varphi''^2}(\omega)$ [or $S_{\varphi''^2\varphi'^2}(\omega)$] is the cross-power-density spectrum of φ'^2 and φ''^2 (or φ''^2 and φ'^2). These quantities can be derived as[12]

$$\begin{aligned} S_{\varphi'^2}(\omega) &= \mathfrak{F}[2R_{\varphi'}^2(\tau) + R_{\varphi'}^2(0)], \\ S_{\varphi'^2\varphi''^2}(\omega) &= \mathfrak{F}[2R_{\varphi'\varphi''}^2(\tau) + R_{\varphi'}(0)R_{\varphi''}(0)], \\ S_{\varphi''^2\varphi'^2}(\omega) &= \mathfrak{F}[2R_{\varphi''\varphi'}^2(\tau) + R_{\varphi'}(0)R_{\varphi''}(0)], \\ S_{\varphi''^2}(\omega) &= \mathfrak{F}[2R_{\varphi''}^2(\tau) + R_{\varphi''}^2(0)], \end{aligned}$$

where $R_{\varphi'}(\tau)$ and $R_{\varphi''}(\tau)$ are the autocorrelation functions of φ' and φ'', respectively, and $R_{\varphi'\varphi''}(\tau)$ [or $R_{\varphi''\varphi'}(\tau)$] is the cross-correlation function of φ' and φ'' (or φ'' and φ'), and $\mathfrak{F}$ stands for "the Fourier transform of".

The Fourier transform of a constant function is a delta function at zero frequency. In the situation of evaluating noise power in the baseband, this quantity is not of interest. Also, it can be shown that

$$S_{\varphi'^2\varphi''^2}(\omega) = S_{\varphi''^2\varphi'^2}(\omega).$$

Thus, (28) can be simplified as

$$\begin{aligned} S_{\varphi_2}(\omega) = 2\,|\,H_1(\omega)\,|^2\mathfrak{F}[R_{\varphi'}^2(\tau)] + 2\,|\,H_2(\omega)\,|^2\mathfrak{F}[R_{\varphi''}^2(\tau)] \\ + 2[H_1(-\omega)H_2(\omega) + H_2(-\omega)H_1(\omega)]\mathfrak{F}[R_{\varphi'\varphi''}^2(\tau)]. \end{aligned} \tag{29}$$

The intermodulation noise to signal power ratio due to the second-order distortion term expressed in dB is, therefore,

$$N_2/S(\text{dB}) = 10 \log [\omega^2 S_{\varphi_2}(\omega)/S_{\varphi'}(\omega)]. \tag{30}$$

VII. NOISE POWER DUE TO THIRD-ORDER DISTORTION TERM

Since

$$\frac{d}{dt}\varphi'^3 = 3\varphi'^2\varphi'',$$

the third-order PM distortion in Table I can be written approximately as

$$\varphi_3(t) = \left(\tfrac{1}{6}\lambda_{3r} - \tfrac{1}{12}l_{1r}\frac{d}{dt}\right)\varphi'^3.$$

The above equation can be represented by the block diagram shown in Fig. 3, where

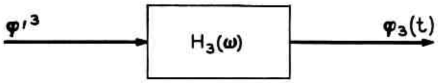


Fig. 3 — Block diagram representation of the third-order PM distortion term.

$$H_3(\omega) = (\tfrac{1}{6}\lambda_{3r}) + i(-\tfrac{1}{12}l_{1r}\omega).$$

The power-density spectrum of $\varphi_3(t)$ is

$$S_{\varphi_3}(\omega) = | H_3(\omega) |^2 S_{\varphi'^3}(\omega),$$

where $S_{\varphi'3}(\omega)$ is the power-density spectrum of φ'^3 which can be derived as[12]

$$S_{\varphi'^3}(\omega) = 6\mathfrak{F}[R_{\varphi'}{}^3(\tau)] + 9R_{\varphi'}{}^2(0)S_{\varphi'}(\omega).$$

In the above equation, the term $9R_{\varphi'}{}^2(0)S_{\varphi'}(\omega)$ is merely a scaled power-density spectrum of the input baseband FM signal, hence, it does not contribute to the intermodulation noise and can be neglected in the computation.

The intermodulation noise to signal power ratio due to the third-order distortion term expressed in dB is, therefore,

$$N_3/S(\text{dB}) = 10 \log [\omega^2 S_{\varphi_3}(\omega)/S_{\varphi'}(\omega)], \tag{31}$$

where

$$S_{\varphi_3}(\omega) = 6 | H_3(\omega) |^2 \mathfrak{F}[R_{\varphi'}{}^3(\tau)]. \tag{32}$$

In (29) and (32), the Fourier transforms of $R_{\varphi'}{}^2(\tau)$, $R_{\varphi''}{}^2(\tau)$, $R_{\varphi'\varphi''}{}^2(\tau)$, and $R_{\varphi'}{}^3(\tau)$ may be obtained by taking the convolutions in the frequency domain. However, this requires numerical integration. For given pre-emphasis characteristics and P_0 (or σ), these Fourier transforms can be expressed in algebraic forms as shown in Appendix B. Hence, no numerical integration is necessary. A digital computer program has been written to calculate the second- and third-order intermodulation noise due to second- and third-order distortion terms in dB. A typical problem can be solved at a very low cost.

VIII. EXAMPLES

Several examples are considered in this paper. Calculated results are compared with measured data when they are available. Expressions for noise calculation are derived for simple cases. For more complicated situations, the noise calculation is best carried out by using a digital computer.

8.1 *Example 1*

In this example, we wish to demonstrate how the intermodulation noise across the baseband can be optimized using an appropriate preemphasis network. To simplify the calculation, we assume that the transmission characteristic consists of linear delay distortion (b_2) only. From Appendix A, we obtain

$$\lambda_{2i} = -2b_2 \, , \; l_{1r} = -8b_2^2$$

and all the other coefficients are equal to zero. Hence,

$$H_1(\omega) = b_2 \, , \qquad H_2(\omega) = 0, \qquad H_3(\omega) = i\tfrac{2}{3}b_2^2\omega.$$

In practical cases, the third-order distortion term $[H_3(\omega)]$ is negligible (say, 50 dB less) compared with the second-order distortion term. From (30) we write

$$\frac{N_2}{S}\,(\text{dB}) = 10 \log \frac{2b_2^{\,2}\omega^2 \, \mathfrak{F}[R_{\varphi'}^{\;2}(\tau)]}{S_{\varphi'}(\omega)} \, .$$

For simplicity, we let

$$S_{\varphi'}(\omega) = P_0(1 + a_2 f^2), \qquad | f \,| \leqq f_b \, .$$

From Appendix B, we obtain

$$\mathfrak{F}[R_{\varphi'}^{\;2}(\tau)] = P_0^2 f_b \eta(f),$$

where

$$\begin{aligned} \eta(f) &= -\tfrac{1}{30}A_2^{\,2}\Omega^5 - \tfrac{2}{3}A_2\Omega^3 + (\tfrac{2}{3}A_2 + 2)A_2\Omega^2 \\ &\qquad - (1 + A_2)^2\Omega + (\tfrac{6}{15}A_2^{\,2} + \tfrac{4}{3}A_2 + 2) \\ A_2 &= a_2 f_b^{\,2} \\ \Omega &= f/f_b \, . \end{aligned}$$

Since

$$P_o = \frac{(2\pi\sigma)^2}{2f_b\left(1 + \dfrac{A_2}{3}\right)} \, ,$$

consequently,

$$\frac{N_2}{S}\,(\text{dB}) = 10 \log \frac{(2\pi)^4(b_2 f_b^{\,2})^2\left(\dfrac{\sigma}{f_b}\right)^2 \Omega^2\eta(f)}{\left(1 + \dfrac{A_2}{3}\right)(1 + A_2\Omega^2)} \, .$$

Specifically, we let

$$b_2 = 7.962 \times 10^{-17} \text{ (linear delay of 1 nanosec/MHz)}$$

$$f_b = 1 \text{ MHz}$$

$$\sigma = 1 \text{ MHz.}$$

After several computer runs, the optimal choice of a_2 is 7. The intermodulation noise with no pre-emphasis and with the optimal pre-emphasis are plotted in Fig. 4. The noise has been reduced to more than 5 dB and is evenly distributed across the baseband by using the optimal pre-emphasis network.

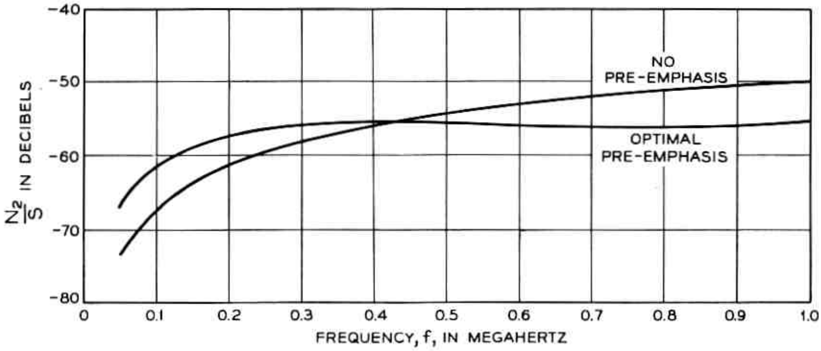


Fig. 4 — Intermodulation noise due to linear delay with and without pre-emphasis.

8.2 *Example 2*

In this example, we use a typical radio system pre-emphasis characteristic shown in Fig. 5. The top baseband frequency is $f_b = 5.772$ MHz. Since the given pre-emphasis characteristic is expressed as relative power in dB versus baseband frequency, it is first converted to ratio versus normalized baseband frequency, $\Omega = f/f_b$.

The weighting function is

$$W(\Omega) = 10^{-(9.5/10)(\Omega-1)}.$$

A least squares approximation program is used to obtain the approximating polynomial

$$a_0 + a_2 f^2 + a_4 f^4 + a_6 f^6,$$

where

$$a_0 = 0.99894166$$

$$a_2 = 11.944252/f_b^2$$

$$a_4 = -5.5771705/f_b^4$$

$$a_6 = 1.4396088/f_b^6.$$

Consider a transmission characteristic consisting of a linear, a parabolic, and a slowly varying sinusoidal delay as shown in Fig. 6. The expressions for the noise due to second- and third-order distortion are too complicated to write down. However, by using a digital computer, the results are plotted in Fig. 7 for $\sigma = 0.771$ MHz. Clearly, a better pre-emphasis network should be used to optimize the noise across the base-band for this particular transmission characteristic.

8.3 *Example 3*

As a final example, we consider a single pole IF filter with

$$Y(\omega + \omega_c) = \frac{1}{1 + i\dfrac{f}{w}},$$

where w is the 3-dB half-bandwidth of the filter. Using a least squares approximation with appropriate weighting function, the magnitude and phase of $Y(\omega + \omega_c)$ are expressed by

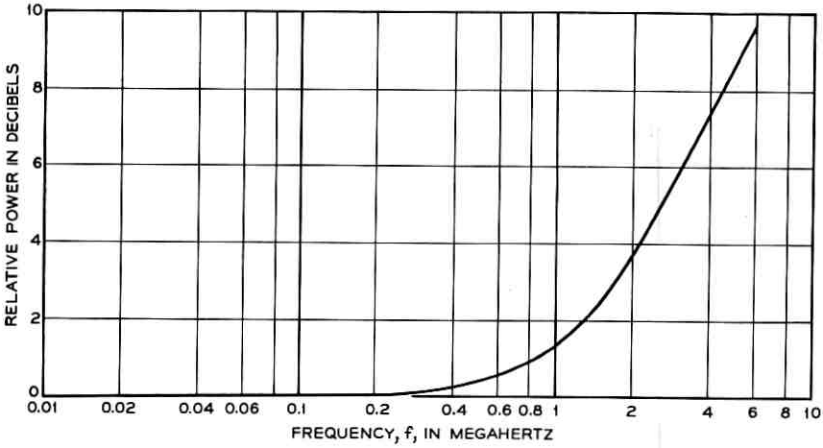


Fig. 5.—Pre-emphasis characteristic of a typical radio system.

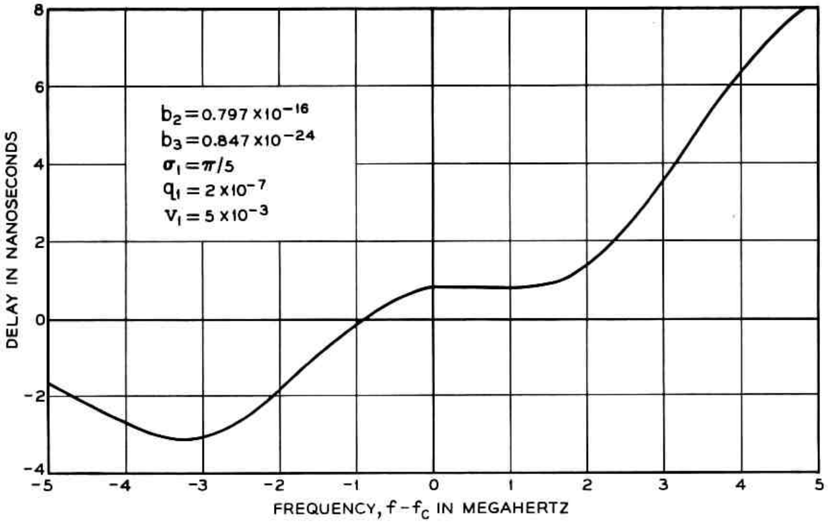


Fig. 6—An arbitrary delay characteristic of a transmission medium.

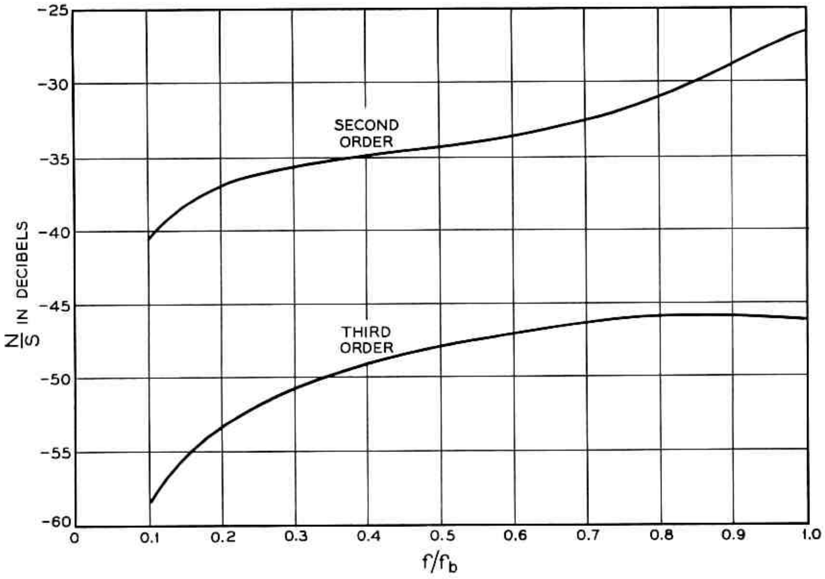


Fig. 7 — Intermodulation noise due to the delay distortion of Fig. 6.

$$\exp\left[-\alpha(f + f_c)\right] \cong 1 + g_2\omega^2 + g_4\omega^4 = 1 + G_2(f/w)^2 + G_4(f/w)^4$$

$$-\beta(f + f_c) \cong b_1\omega + b_3\omega^3 = B_1(f/w) + B_3(f/w)^3,$$

where

$$G_2 = g_2(2\pi w)^2 = -0.4209, \qquad G_4 = g_4(2\pi w)^4 = 0.08027$$

$$B_1 = b_1(2\pi w) = -0.9529, \qquad B_3 = b_3(2\pi w)^3 = 0.1294.$$

The actual and approximated transmission characteristic are plotted in Fig. 8. Since a constant delay does not introduce intermodulation noise,

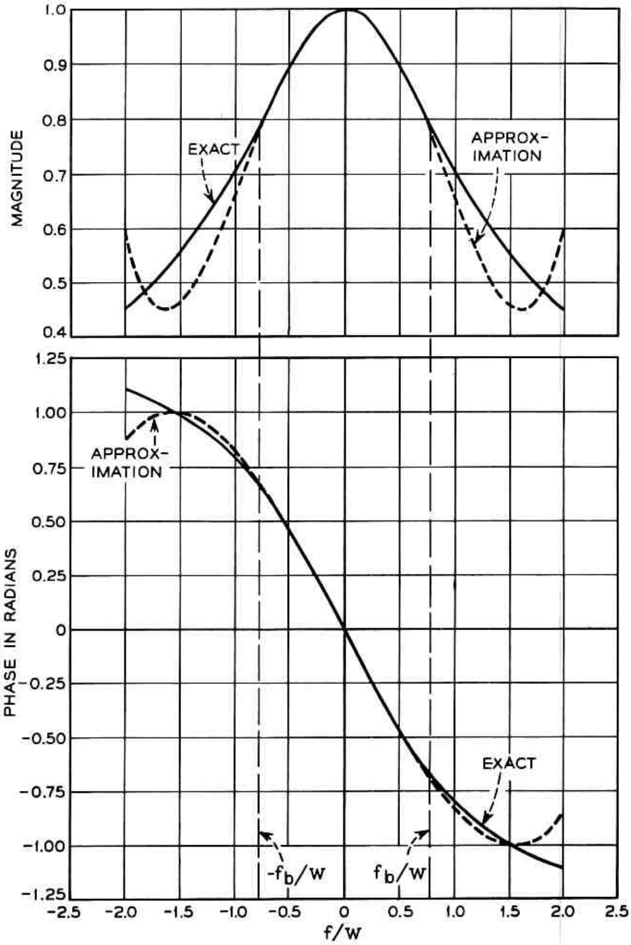


Fig. 8 — Gain and phase characteristic of a single pole filter.

b_1 is not included for calculation. Using the expressions in Appendix A, the coefficients associated with the second-order distortion term are

$$\lambda_{2i} = 0, \qquad l_{2i} = 0, \qquad l_{3i} = 0, \qquad l_{4i} = 0.$$

Hence,

$$H_1(\omega) = H_2(\omega) = 0.$$

The noise contribution due to second-order distortion is, therefore, zero.

The coefficients associated with the third-order distortion term are

$$\lambda_{3r} = 6b_3\,, \qquad l_{1r} = 24g_4 - 4g_2^2.$$

Assuming no pre-emphasis, that is, $a_2 = a_4 = a_6 = 0$, from Appendix B, we have

$$\mathfrak{F}[R_{\varphi'}{}^3(\tau)] = P_0^3 f_b^2 a_0^3 (3 - \Omega^2), \qquad |\,\Omega\,| \leqq 1$$

where

$$\Omega = f/f_b\,.$$

Using the relation

$$P_o = \frac{(2\pi\sigma)^2}{2f_b\, a_o}\,,$$

the intermodulation noise to signal power ratio due to the third-order distortion term can be derived from (31) as

$$\frac{N_3}{S}\,(\text{dB}) = 10\log\frac{3}{2}\left(\frac{\sigma}{f_b}\right)^4\left(\frac{f_b}{w}\right)^6\Omega^2\,(3 - \Omega^2)\left[B_3^2 + \left(\frac{f}{w}\right)^2\left(2G_4 - \frac{G_2^2}{3}\right)^2\right].$$

For $f_b = 1$ MHz and $w = 1.25$ MHz, N_3/S (dB) is calculated at $f = 0.084$, 0.36, and 1 MHz as a function of (σ/f_b). The dotted lines in Fig. 9 represent the calculated value (S/N_3) while the solid curves represent the measured data taken by W. F. Bodtmann.[13] The discrepancy between the measured and calculated values can be attributed to several reasons: (*i*) The power-density spectrum of the multichannel baseband signal used in the experiment was not perfectly rectangular, (*ii*) During the measurement, a non-ideal limiter was used which caused some AM-to-PM conversion, (*iii*) The actual transmission characteristic was approximated by few parameters in a limited region, and (*iv*) the formulas (30) and (31) derived in this paper involved approximations. Nevertheless, the measured and calculated results are close enough to show the utility of the analysis even for cases beyond the application for which it was originally intended.

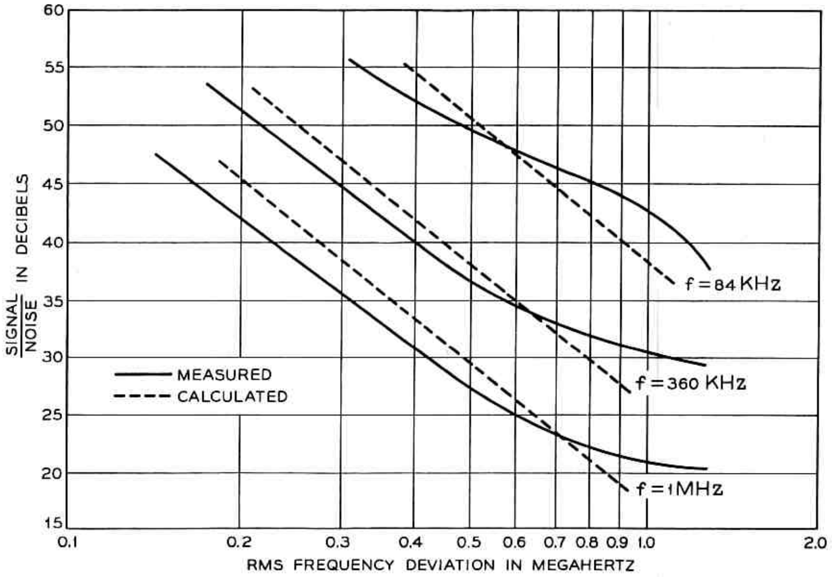


Fig. 9 — Measured and calculated intermodulation noise due to the single pole filter.

IX. DISTORTION DUE TO AM-TO-PM CONVERSION

Let K be the AM-PM conversion factor of a device expressed in degrees/dB, then the phase distortion due to AM-PM conversion is

$$\text{Phase Distortion due to AM-PM Conversion} = 20 \times 0.4343K\, a(t)$$
$$= 8.686K\, a(t) \text{ degrees,}$$

where $a(t)$ is given by (22).

Similarly,

$$\text{Frequency Distortion due to AM-PM Conversion} = \frac{8.686}{2\pi} K \frac{d}{dt} a(t)$$
$$= 1.382\, K \frac{d}{dt} a(t)\ Hz.$$

In Table I, the second- and third-order terms of amplitude distortion cause intermodulation noise while the linear term causes video roll-off or enhancement. Using the same approach discussed previously in this paper, noise calculation due to AM-to-PM conversion which is caused by transmission deviation can also be accomplished.

X. ACKNOWLEDGMENTS

The analysis carried out in this paper is based on an unpublished work of S. O. Rice. His permission to present his result here is greatly appreciated. The author is grateful to J. T. Bangert for his encouragement and direction and also wishes to thank J. J. Lang, F. J. Witt, L. H. Enloe, C. L. Ruthroff, and W. T. Barnett for many constructive suggestions. The author is indebted to Miss J. D. Witkowski for the digital computer programming and to many other colleagues who have helped to carry out this analysis.

APPENDIX A

Coefficients Associated with the Distortion Terms

The moments defined in (19) can be expressed as

$$\begin{aligned}
m_0 &= 1 \\
m_1 &= i\left(\frac{A'}{A}\right) \\
m_2 &= \left(-\frac{A''}{A}\right) + i\,(-B'') \\
m_3 &= \left(B''' + \frac{3A'B''}{A}\right) + i\left(-\frac{A'''}{A}\right) \\
m_4 &= \left(-3B''^2 + \frac{A''''}{A}\right) + i\left(B'''' + \frac{4A'B'''}{A} + \frac{6A''B''}{A}\right),
\end{aligned}$$

where

$$\begin{aligned}
A &= 1 + \sum_{k=1}^{N} u_k \cos\theta_k \\
A' &= g_1 - \sum_{k=1}^{N} u_k p_k \sin\theta_k \\
A'' &= 2g_2 - \sum_{k=1}^{N} u_k p_k^{\,2} \cos\theta_k \\
A''' &= 6g_3 + \sum_{k=1}^{N} u_k p_k^{\,3} \sin\theta_k \\
A'''' &= 24g_4 + \sum_{k=1}^{N} u_k p_k^{\,4} \cos\theta_k
\end{aligned}$$

$$B'' = 2b_2 - \sum_{k=1}^{N} v_k q_k^{\,2} \sin \sigma_k$$

$$B''' = 6b_3 - \sum_{k=1}^{N} v_k q_k^{\,3} \cos \sigma_k$$

$$B'''' = 24b_4 + \sum_{k=1}^{N} v_k q_k^{\,4} \sin \sigma_k \,.$$

The coefficients associated with the second-order distortion are

$$\lambda_{2i} = -B''$$

$$l_{2i} = -\frac{A'''}{A} + \frac{A'A''}{A^2}$$

$$l_{3i} = B'''' + \frac{3A'B'''}{A} + \frac{6A''B''}{A} - \frac{3A'^2B''}{A^2}$$

$$l_{4i} = B'''' + \frac{12A''B''}{A} - 12\,\frac{A'^2B''}{A^2}\,.$$

The coefficients associated with the third-order distortion are

$$\lambda_{3r} = B'''$$

$$l_{1r} = -2B''^2 + \frac{A''''}{A} - \frac{A''^2}{A^2} - 2\,\frac{A'A'''}{A^2} + 2\,\frac{A'^2A''}{A^3}\,.$$

APPENDIX B

Fourier Transforms of $R_{\varphi'}{}^2(\tau)$, $R_{\varphi'\varphi''}{}^2(\tau)$, $R_{\varphi''}{}^2(\tau)$ and $R_{\varphi'}{}^3(\tau)$

The Fourier transforms of $R_{\varphi'}{}^2(\tau)$, $R_{\varphi'\varphi''}{}^2(\tau)$, $R_{\varphi''}{}^2(\tau)$ and $R_{\varphi'}{}^3(\tau)$ can be derived in a straightforward manner. However, considerable amount of algebra has been involved during the derivation. Without writing out the details here we merely present the final results.

$$\begin{aligned}\mathfrak{F}[R_{\varphi'}{}^2(\tau)] = P_0^2 f_b\{&D_1(f - 2f_b) + 2D_1(f) + D_1(f + 2f_b)\\ &- D_2(f - 2f_b) + 2D_2(f) - D_2(f + 2f_b)\\ &+ 2D_3(f - 2f_b) - 2D_3(f + 2f_b)\}\end{aligned}$$

$$\begin{aligned}\mathfrak{F}[R_{\varphi'}{}^3(\tau)] = P_0^3 f_b^2\{&3E_1(f - f_b) + E_1(f - 3f_b) + E_1(f + 3f_b)\\ &+ 3E_1(f + f_b) + 3E_4(f - f_b) - 3E_4(f - 3f_b)\\ &- 3E_4(f + 3f_b) + 3E_4(f + f_b) + 3E_3(f - f_b)\end{aligned}$$

$$
\begin{aligned}
&+ 3E_3(f - 3f_b) \\
&- 3E_3(f + 3f_b) - 3E_3(f + f_b) + 3E_2(f - f_b) \\
&- E_2(f - 3f_b) + E_2(f + 3f_b) - 3E_2(f + f_b)\}
\end{aligned}
$$

$$
\begin{aligned}
\mathfrak{F}[R_{\varphi'\varphi''}{}^2(\tau)] &= 4\pi^2 P_0^2 f_b^3 \{ J_1(f - 2f_b) + 2J_1(f) + J_1(f + 2f_b) - J_2(f - 2f_b) \\
&\quad + 2J_2(f) - J_2(f + 2f_b) - 2J_3(f - 2f_b) + 2J_3(f + 2f_b)\} \\
\mathfrak{F}[R_{\varphi''}{}^2(\tau)] &= 16\pi^4 P_0^2 f_b^5 \{K_1(f - 2f_b) + 2K_1(f) + K_1(f + 2f_b) \\
&\quad - K_2(f - 2f_b) + 2K_2(f) - K_2(f + 2f_b) + 2K_3(f - 2f_b) \\
&\quad - 2K_3(f + 2f_b)\},
\end{aligned}
$$

where

$$
\begin{aligned}
D_1(f) &= \sum_{n=2}^{6} (-1)^n \frac{d_{1n}}{2(2n-1)!} \left(\frac{f}{f_b}\right)^{2n-1} \operatorname{sgn} f \\
D_2(f) &= \sum_{n=1}^{7} (-1)^n \frac{d_{2n}}{2(2n-1)!} \left(\frac{f}{f_b}\right)^{2n-1} \operatorname{sgn} f \\
D_3(f) &= \sum_{n=1}^{6} (-1)^{n+1} \frac{d_{3n}}{2(2n)!} \left(\frac{f}{f_b}\right)^{2n} \operatorname{sgn} f \\
E_1(f) &= \sum_{n=3}^{9} (-1)^n \frac{e_{1n}}{2(2n-1)!} \left(\frac{f}{f_b}\right)^{2n-1} \operatorname{sgn} f \\
E_2(f) &= \sum_{n=1}^{10} (-1)^{n+1} \frac{e_{2n}}{2(2n)!} \left(\frac{f}{f_b}\right)^{2n} \operatorname{sgn} f \\
E_3(f) &= \sum_{n=2}^{9} (-1)^{n+1} \frac{e_{3n}}{2(2n)!} \left(\frac{f}{f_b}\right)^{2n} \operatorname{sgn} f \\
E_4(f) &= \sum_{n=2}^{10} (-1)^n \frac{e_{4n}}{2(2n-1)!} \left(\frac{f}{f_b}\right)^{2n-1} \operatorname{sgn} f \\
J_1(f) &= \sum_{n=1}^{7} (-1)^n \frac{j_{1n}}{2(2n-1)!} \left(\frac{f}{f_b}\right)^{2n-1} \operatorname{sgn} f \\
J_2(f) &= \sum_{n=2}^{8} (-1)^n \frac{j_{2n}}{2(2n-1)!} \left(\frac{f}{f_b}\right)^{2n-1} \operatorname{sgn} f \\
J_3(f) &= \sum_{n=1}^{7} (-1)^{n+1} \frac{j_{3n}}{2(2n)!} \left(\frac{f}{f_b}\right)^{2n} \operatorname{sgn} f \\
K_1(f) &= \sum_{n=2}^{8} (-1)^n \frac{k_{1n}}{2(2n-1)!} \left(\frac{f}{f_b}\right)^{2n-1} \operatorname{sgn} f
\end{aligned}
$$

$$K_2(f) = \sum_{n=1}^{9} (-1)^n \frac{k_{2n}}{2(2n-1)!} \left(\frac{f}{f_b}\right)^{2n-1} \operatorname{sgn} f$$

$$K_3(f) = \sum_{n=1}^{8} (-1)^{n+1} \frac{k_{3n}}{2(2n)!} \left(\frac{f}{f_b}\right)^{2n} \operatorname{sgn} f.$$

The coefficients in the above equations are given as

$$d_{12} = c_1^2, \quad d_{13} = 2c_1c_2, \quad d_{14} = c_2^2 + 2c_1c_3$$

$$d_{15} = 2c_2c_3, \quad d_{16} = c_3^2$$

$$d_{21} = c_4^2, \quad d_{22} = 2c_4c_5, \quad d_{23} = c_5^2 + 2c_4c_6$$

$$d_{24} = 2c_4c_7 + 2c_5c_6, \quad d_{25} = c_6^2 + 2c_5c_7$$

$$d_{26} = 2c_6c_7, \quad d_{27} = c_7^2$$

$$d_{31} = c_1c_4, \quad d_{32} = c_1c_5 + c_2c_4, \quad d_{33} = c_1c_6 + c_2c_5 + c_3c_4$$

$$d_{34} = c_1c_7 + c_2c_6 + c_3c_5, \quad d_{35} = c_2c_7 + c_3c_6, \quad d_{36} = c_3c_7$$

$$e_{13} = c_1^3, \quad e_{14} = 3c_1^2c_2, \quad e_{15} = 3c_1c_2^2 + 3c_1^2c_3$$

$$e_{16} = 6c_1c_2c_3 + c_2^3, \quad e_{17} = 3c_1c_3^2 + 3c_2^2c_3, \quad e_{18} = 3c_2c_3^2$$

$$e_{19} = c_3^3, \quad e_{21} = c_4^3, \quad e_{22} = 3c_4^2c_5$$

$$e_{23} = 3c_4c_5^2 + 3c_4^2c_6, \quad e_{24} = 3c_4^2c_7 + 6c_4c_5c_6 + c_5^3$$

$$e_{25} = 3c_4c_6^2 + 6c_4c_5c_7 + 3c_5^2c_6, \quad e_{26} = 6c_4c_6c_7 + 3c_5c_6^2 + 3c_5^2c_7$$

$$e_{27} = 3c_4c_7^2 + 6c_5c_6c_7 + c_6^3, \quad e_{28} = 3c_5c_7^2 + 3c_6^2c_7$$

$$e_{29} = 3c_6c_7^2, \quad e_{210} = c_7^3, \quad e_{32} = c_1^2c_4$$

$$e_{33} = 2c_1c_2c_4 + c_1^2c_5, \quad e_{34} = c_2^2c_4 + 2c_1c_3c_4 + 2c_1c_2c_5 + c_1^2c_6$$

$$e_{35} = 2c_2c_3c_4 + c_2^2c_5 + 2c_1c_3c_5 + 2c_1c_2c_6 + c_1^2c_7$$

$$e_{36} = c_3^2c_4 + 2c_2c_3c_5 + c_2^2c_6 + 2c_1c_3c_6 + 2c_1c_2c_7$$

$$e_{37} = c_3^2c_5 + 2c_2c_3c_6 + c_2^2c_7 + 2c_1c_3c_7$$

$$e_{38} = c_3^2c_6 + 2c_2c_3c_7, \quad e_{39} = c_3^2c_7, \quad e_{42} = c_1c_4^2$$

$$e_{43} = 2c_1c_4c_5 + c_2c_4^2, \quad e_{44} = c_1c_5^2 + 2c_1c_4c_6 + 2c_2c_4c_5 + c_3c_4^2$$

$$e_{45} = 2c_1c_4c_7 + 2c_1c_5c_6 + c_2c_5^2 + 2c_3c_4c_5 + 2c_2c_4c_6$$

$$e_{46} = c_1c_6^2 + 2c_1c_5c_7 + 2c_2c_4c_7 + 2c_2c_5c_6 + c_3c_5^2 + 2c_3c_4c_6$$

$$e_{47} = 2c_1c_6c_7 + c_2c_6^2 + 2c_2c_5c_7 + 2c_3c_4c_7 + 2c_3c_5c_6$$

$$
\begin{aligned}
e_{48} &= c_1c_7^2 + 2c_2c_6c_7 + c_3c_6^2 + 2c_3c_5c_7 \\
e_{49} &= c_2c_7^2 + 2c_3c_6c_7\,, \quad e_{410} = c_3c_7^2 \\
j_{11} &= c_4^2, \quad j_{12} = 2c_4(c_5 - 2c_1) \\
j^{13} &= (c_5 - 2c_1)^2 + 2c_4(c_6 - 4c_2) \\
j_{14} &= 2c_4(c_7 - 6c_3) + 2(c_5 - 2c_1)\,(c_6 - 4c_2) \\
j_{15} &= (c_6 - 4c_2)^2 + 2(c_5 - 2c_1)\,(c_7 - 6c_3) \\
j_{16} &= 2(c_6 - 4c_2)\,(c_7 - 6c_3), \quad j_{17} = (c_7 - 6c_3)^2 \\
j_{22} &= (c_1 + c_4)^2, \quad j_{23} = 2(c_1 + c_4)\,(c_2 + 3c_5) \\
j_{24} &= (c_2 + 3c_5)^2 + 2(c_1 + c_4)\,(c_3 + 5c_6) \\
j_{25} &= 14c_7(c_1 + c_4) + 2(c_2 + 3c_5)\,(c_3 + 5c_6) \\
j_{26} &= (c_3 + 5c_6)^2 + 14c_7(c_2 + 3c_5), \quad j_{27} = 14c_7(c_3 + 5c_6) \\
j_{28} &= 49c_7^2, \quad j_{31} = c_4(c_1 + c_4) \\
j_{32} &= c_4(c_2 + 3c_5) + (c_1 + c_4)\,(c_5 - 2c_1) \\
j_{33} &= c_4(c_3 + 5c_6) + (c_2 + 3c_5)\,(c_5 - 2c_1) + (c_1 + c_4)\,(c_6 - 4c_2) \\
j_{34} &= 7c_4c_7 + (c_5 - 2c_1)\,(c_3 + 5c_6) + (c_6 - 4c_2)\,(c_2 + 3c_5) \\
&\qquad + (c_7 - 6c_3)\,(c_1 + c_4) \\
j_{35} &= 7c_7(c_5 - 2c_1) + (c_3 + 5c_6)\,(c_6 - 4c_2) + (c_2 + 3c_5)\,(c_7 - 6c_3) \\
j_{36} &= 7c_7(c_6 - 4c_2) + (c_3 + 5c_6)\,(c_7 - 6c_3), \quad j_{37} = 7c_7(c_7 - 6c_3) \\
k_{12} &= (c_1 + 2c_4)^2, \quad k_{13} = -2(c_1 + 2c_4)\,(6c_1 - c_2 - 6c_5) \\
k_{14} &= (6c_1 - c_2 - 6c_5)^2 - 2(c_1 + 2c_4)\,(20c_2 - c_3 - 10c_6) \\
k_{15} &= 2(6c_1 - c_2 - 6c_5)\,(20c_2 - c_3 - 10c_6) - 2(c_1 + 2c_4)\,(42c_3 - 14c_7) \\
k_{16} &= (20c_2 - c_3 - 10c_6)^2 + 2(6c_1 - c_2 - 6c_5)\,(42c_3 - 14c_7) \\
k_{17} &= 2(20c_2 - c_3 - 10c_6)\,(42c_3 - 14c_7), \quad k_{18} = (42c_3 - 14c_7)^2 \\
k_{21} &= c_4^2, \quad k_{22} = -2c_4(4c_1 - c_5 + 2c_4) \\
k_{23} &= (4c_1 - c_5 + 2c_4)^2 - 2c_4(8c_2 + 12c_5 - c_6) \\
k_{24} &= 2(4c_1 - c_5 + 2c_4)\,(8c_2 + 12c_5 - c_6) - 2c_4(12c_3 + 30c_6 - c_7) \\
k_{25} &= (8c_2 + 12c_5 - c_6)^2 - 112c_4c_7 + 2(4c_1 - c_5 + 2c_4)\,(12c_3 + 30c_6 - c_7)
\end{aligned}
$$

$$k_{26} = 112c_7(4c_1 - c_5 + 2c_4) + 2(8c_2 + 12c_5 - c_6)(12c_3 + 30c_6 - c_7)$$

$$k_{27} = (12c_3 + 30c_6 - c_7)^2 + 112c_7(8c_2 + 12c_5 - c_6)$$

$$k_{28} = 112c_7(12c_3 + 30c_6 - c_7), \quad k_{29} = (56c_7)^2$$

$$k_{31} = c_4(c_1 + 2c_4)$$

$$k_{32} = -(c_1 + 2c_4)(4c_1 - c_5 + 2c_4) - c_4(6c_1 - c_2 - 6c_5)$$

$$\begin{aligned} k_{33} = {} & -(c_1 + 2c_4)(8c_2 + 12c_5 - c_6) + (6c_1 - c_2 - 6c_5)(4c_1 - c_5 + 2c_4) \\ & - c_4(20c_2 - c_3 - 10c_6) \end{aligned}$$

$$\begin{aligned} k_{34} = {} & -(c_1 + 2c_4)(12c_3 + 30c_6 - c_7) + (6c_1 - c_2 - 6c_5)(8c_2 + 12c_5 - c_6) \\ & + (20c_2 - c_3 - 10c_6)(4c_1 - c_5 + 2c_4) - c_4(42c_3 - 14c_7) \end{aligned}$$

$$\begin{aligned} k_{35} = {} & -56c_7(c_1 + 2c_4) + (6c_1 - c_2 - 6c_5)(12c_3 + 30c_6 - c_7) \\ & + (20c_2 - c_3 - 10c_6)(8c_2 + 12c_5 - c_6) \\ & + (42c_3 - 14c_7)(4c_1 - c_5 + 2c_4) \end{aligned}$$

$$\begin{aligned} k_{36} = {} & 56c_7(6c_1 - c_2 - 6c_5) + (20c_2 - c_3 - 10c_6)(12c_3 + 30c_6 - c_7) \\ & + (42c_3 - 14c_7)(8c_2 + 12c_5 - c_6) \end{aligned}$$

$$k_{37} = 56c_7(20c_2 - c_3 - 10c_6) + (42c_3 - 14c_7)(12c_3 + 30c_6 - c_7)$$

$$k_{38} = 56c_7(42c_3 - 14c_7),$$

where

$$c_1 = 2(a_2f_b^2 + 2a_4f_b^4 + 3a_6f_b^6)$$

$$c_2 = -24(a_4f_b^4 + 5a_6f_b^6), \quad c_3 = 720a_6f_b^6$$

$$c_4 = a_0 + a_2f_b^2 + a_4f_b^4 + a_6f_b^6$$

$$c_5 = -2(a_2f_b^2 + 6a_4f_b^4 + 15a_6f_b^6)$$

$$c_6 = 24(a_4f_b^4 + 15a_6f_b^6), \quad c_7 = -720a_6f_b^6.$$

REFERENCES

1. Carson, J. R. and Fry, T. C., Variable Frequency Electric Circuit Theory with Application to the Theory of Frequency-Modulation, B.S.T.J., *26,* October, 1937, pp. 513–540.
2. Van der Pol, B., The Fundamental Principles of Frequency Modulation, J. IEE (London), *93,* May, 1946, Part 3, pp. 153–158.
3. Stumpers, F. L. H. M., Distortion of Frequency Modulated Signals in Electrical Networks, Commun. News, *9,* April, 1948, pp. 82–92.
4. Gladwin, A. S., The Distortion of Frequency-Modulated Waves by Transmission Networks, Proc. IRE, *35,* December, 1947, pp. 1436–1445.

5. Medhurst, R. G., Harmonic Distortion of Frequency-Modulated Waves by Linear Networks, Proc. IEEE *101,* May, 1954, pp. 171–181.
6. Rice, S. O., Distortion Produced in a Noise-Modulated FM Signal by Nonlinear Attenuation and Phase Shift, B.S.T.J., *36,* July, 1957, pp. 879–890.
7. Medhurst, R. G., Explicit Form of FM Distortion Products With White-Noise Modulation, Proc. IEEE, *107,* January, 1960, Pt. C, pp. 120–126.
8. Magnusson, R. I., Intermodulation Noise in Linear FM Systems, Proc. IEEE *109,* March, 1962, Pt. C, pp. 32–45.
9. Baghdady, E. J., *Lectures on Communication System Theory*, McGraw-Hill Book Co., New York, 1961, Chapter 19.
10. Panter, P. F., *Modulation, Noise, and Spectral Analysis*, McGraw-Hill Book Co., New York, 1965, Chapter 9.
11. Newton, G. C., Gould, L. A., and Kaiser, J. F., *Analytical Design of Linear Feedback Controls*, John Wiley & Sons, Inc., New York, 1961, p. 120.
12. Laning, J. H. Jr. and Battin, R. H., *Random Processes in Automatic Control*, McGraw-Hill Book Co., New York, 1956, pp. 82–85.
13. Bodtmann, W. F., Private communication.

Bounds for Certain Multiprocessing Anomalies

By R. L. GRAHAM

(Manuscript received July 11, 1966)

It is known that in multiprocessing systems composed of many identical processing units operating in parallel, certain timing anomalies may occur; e.g., an increase in the number of processing units can cause an increase in the total length of time needed to process a fixed set of tasks. In this paper, precise bounds are derived for several anomalies of this type.

I. INTRODUCTION

In recent years there has been increased interest in the study of the potential advantages afforded by the use of a computer with many processors in parallel. While it is generally true that a set of tasks may be processed in less time by this type of multiprocessing, it has been pointed out that certain anomalies[1,2] may occur, even though the processors are used in a very "natural" way (e.g., it can happen that increasing the number of processors can *increase* the time required to complete a given set of tasks).

It is the purpose of this paper to derive precise bounds on the extent to which these anomalies can affect the time required to process a set of tasks, given certain rather natural rules for the operation of the multiprocessing system.

1.1 *Description of the System*

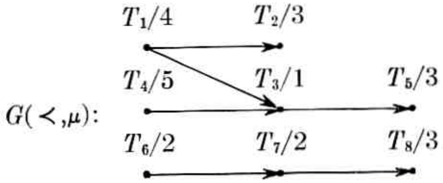
Let us suppose that we are given n identical processing units P_i, $1 \leqq i \leqq n$, and a set of tasks $T = \{T_1, \cdots, T_m\}$ to be processed by the P_i. We are also given a partial-order* $<$ on T and a function μ: $T \rightarrow [0, \infty)$. Once a processor P_i begins a task T_j, it works without interruption on T_j until completion of that task, taking altogether $\mu(T_j)$ units of time. It is also required that if $T_i < T_j$ then T_j cannot

* See Ref. 2.

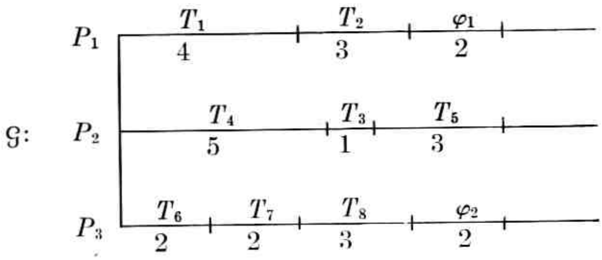
be started until T_i is completed. The P_i execute the T_j in the following way: We are given a linear ordering L: $(T_{k_1}, \cdots, T_{k_m})$ of T called a *task list* (or priority list). In general, at any time t a P_i completes a task, it immediately (and instantaneously) scans the list L (starting from the beginning) until it comes to the *first* task T_j which has not yet begun to be executed. If all the *predecessors* of T_j (i.e., those $T_i < T_j$) have been completed by time t then P_i begins working on T_j. Otherwise P_i proceeds to the *next* task $T_{j'}$, in L which has not yet begun to be executed, etc. If P_i proceeds through the entire list L without finding a task to execute then P_i becomes *idle* (we shall also say that P_i is working on an empty task). P_i remains idle until some other P_j *completes* a task at which time P_i (and of course P_j) immediately scans the list L as before for possible tasks to execute. If two processors P_i and P_j, $i < j$, simultaneously attempt to begin the same task T_k, it will be our convention to assign T_k to P_i, the processor with the smaller index. The processors all start scanning L at time $t = 0$ and proceed in the above-mentioned fashion until some time ω, the least time for which all the tasks have been completed.

It will be helpful here to consider several examples. We shall indicate the partial-order $<$ on T and the function μ by a *directed graph* $G(<,\mu)$. In $G(<,\mu)$, the vertices will correspond to the T_j and a directed edge from T_i to T_j will indicate that $T_i < T_j$. Each vertex of $G(<,\mu)$ will actually be labelled with the symbol $T_j/\mu(T_j)$, the $\mu(T_j)$ indicating the time necessary to execute T_j. The activity of each P_i is conveniently represented by a *timing diagram* $\mathcal{G}$ (also known as a Gantt diagram; see Ref. 1). $\mathcal{G}$ will consist of n horizontal half-lines (labelled by the P_i) in which each line is subdivided into segments* and labelled according to the state of the corresponding processor.

Example 1: $n = 3$, L: $(T_3, T_1, T_2, T_4, T_6, T_5, T_7, T_8)$

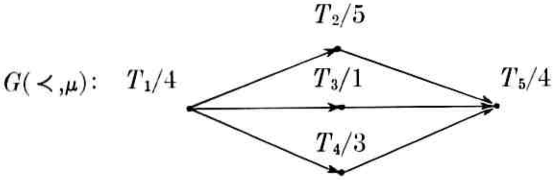


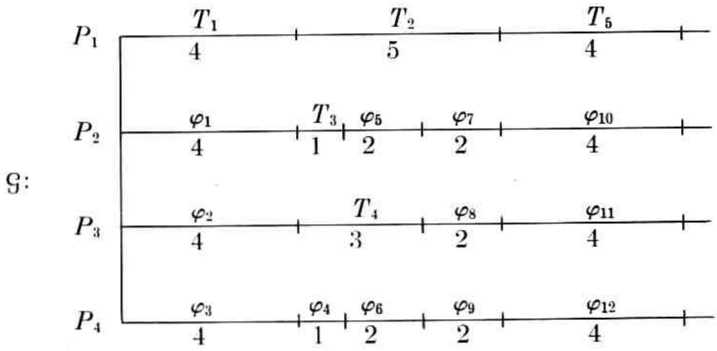
* We always consider the segments as being closed on the left and open on the right.



The symbol φ_i indicates a processor is idle (i.e., working on the empty task φ_i) but not all the T_j have been completed. The indexing of the φ_i is arbitrary. Thus, for $\mathcal{G}$ we have $\omega = 9$.

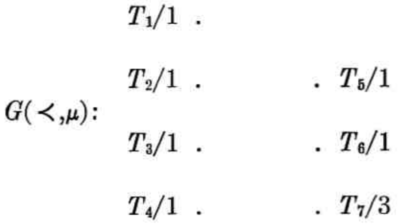
Example 2: $n = 4$, L: $(T_1, T_2, T_3, T_4, T_5)$

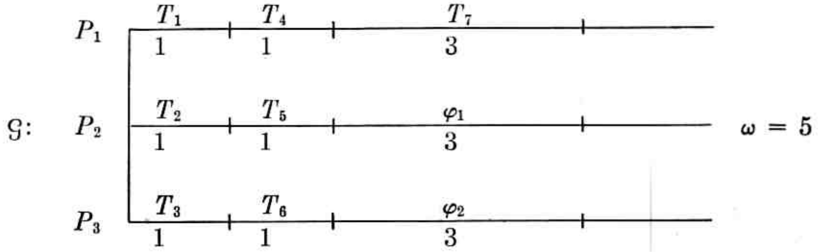




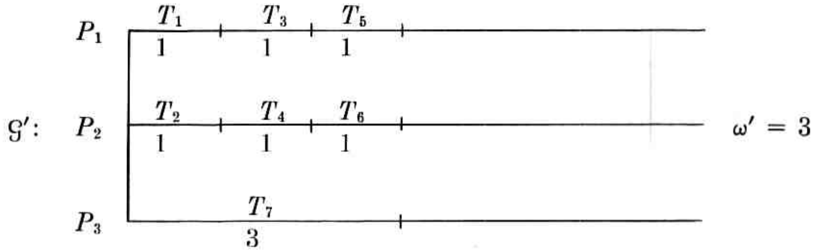
Here, $\omega = 13$. Note that in this example, ω is independent of L. We should also point out here that we are using the convention that whenever any T_j is completed, then *all* current empty tasks φ_i are also terminated. Processors still idle are then given "new" empty tasks to complete (e.g., P_4 in Example 2).

Example 3: $n = 3$, L: $(T_1, T_2, T_3, T_4, T_5, T_6, T_7)$





Suppose we use a different list L' given by L': $(T_1, T_2, T_7, T_3, T_4, T_5, T_6)$. We then have



Hence, by simply using a different list L', we have shortened ω by nearly a factor of two. The significance of this and similar examples will be brought out in the next section.

We see that, in general, ω is a function of the task list L, the "time" function μ, the partial-order $<$, and the number of processors n (in addition to the rules under which the P_i operate). In this note, we investigate the factor by which ω can increase if we simultaneously:

(*i*) Change* the task list L;

(*ii*) Decrease the function μ;

(*iii*) Relax the partial-order $<$;

(*iv*) Change the number of processors from n to n'.

While it might first be expected that (*ii*), (*iii*), or (*iv*) (with $n' > n$) would cause a decrease in ω, easy counterexamples† show that is not always the case. In the next section we obtain an upper bound on the factor by which ω can increase because of (*i*), (*ii*), (*iii*), and (*iv*) (cf. Theorem, p. 1571). This bound is just the expression $1 + n - 1/n'$. We also show that this bound is the *best possible* in the sense that it cannot be replaced by any smaller function of n and n'.

II. THE MAIN RESULTS

We begin this section by considering a special case of the general problem. We include this here in order to acquaint the reader with the basic ideas which will be used later. Suppose we are given a set of tasks $T = \{T_1, \cdots, T_m\}$ and a directed graph $G(<,\mu)$ giving a partial-order $<$ and a time function μ on T. We execute these tasks twice, each time using two identical processors P_1 and P_2. The first time the tasks are executed we use a task list L while the second time the tasks are executed we use another task list L'. Suppose the corresponding finishing times are ω and ω'. The question we consider now is this: How much can the ratio ω'/ω vary? This is answered by the following

Proposition: $\frac{2}{3} \leqq \frac{\omega'}{\omega} \leqq \frac{3}{2}$.

Proof: By the symmetry of ω and ω' it suffices to show that $\omega'/\omega \leqq \frac{3}{2}$. The basic idea we shall use is a simple one. Consider the timing diagram $\mathcal{G}$ obtained when the tasks are executed using the list L. We want to show that there is a *chain*‡ of tasks $T_{c_1} < T_{c_2} < \cdots < T_{c_r}$ which has the property that *whenever a processor is idle* (i.e., executing an empty task φ_i) *then the other processor is executing one of the* T_{c_k}.

* By "change" we mean "possibly change", etc.

† As far as the author is aware, these facts were first pointed out by Richards.[3]

‡ i.e., a linearly-ordered subset using the partial-order $<$.

To define the T_{c_k} we proceed as follows. First, let T_{j_1} be defined to be the task which has the latest finishing time in $\mathcal{G}$ (if there is more than one such task then we choose the task which is executed by the higher-indexed processor). Let φ_{t_1} be the empty task which has the *latest finishing time of all those empty tasks which finish at a time not later than the starting time of* T_{j_1}. By the construction of $\mathcal{G}$, there must be a task T_u which has the *same finishing time as* φ_{t_1}. Define T_{j_2} to be T_u. In general, suppose we have defined T_{j_k} for some $k \geqq 2$. To define $T_{j_{k+1}}$, let φ_{t_k} be the empty task which has the latest finishing time of all those empty tasks φ_i which finish at a time not later than the starting time of T_{j_k}. (If there are no such φ_i then we are done, i.e., $T_{j_{k+1}}$ is not defined.) By hypothesis, there must be a task T_v which has the same finishing time as $\varphi_{t_{k+1}}$ and which has a starting time not later than the starting time of $\varphi_{t_{k+1}}$. Define $T_{j_{k+1}}$ to be T_v. We continue this algorithm for as long as possible, say, until we have defined $T_{j_1}, \cdots, T_{j_r}$.

We first note that since no processor works on *one* empty task φ_i while the other processor works on *more than one task*, then *at any time a processor is executing an empty task*, φ_i, *the other processor is executing one of the* T_{j_k}. We next claim that $T_{j_{k+1}} \prec T_{j_k}$ for $1 \leqq k < r$. Suppose this is not the case. If t_o denotes the time at which a processor P_i started executing $\varphi_{t_{k+1}}$ then by the hypothesis concerning the operation of the processors, P_i should *not* have been idle (i.e., working on $\varphi_{t_{k+1}}$) since at least one task, namely T_{j_k}, was eligible to be executed at that time. Thus, the timing diagram $\mathcal{G}$ is not valid and we have a contradiction. Hence, we must have $T_{j_{k+1}} \prec T_{j_k}$ for $1 \leqq k < r$. By defining $T_{c_k} \equiv T_{j_{r+1-k}}$ for $1 \leqq k \leqq r$, the first assertion is proved. It follows at once that if we let $\mu(\varphi_i)$ denote the length of time a processor spends executing φ_i, then

$$\sum_{\varphi_i \in \mathcal{G}} \mu(\varphi_i) \leqq \sum_{k=1}^{r} \mu(T_{j_k}). \tag{1}$$

The proof of the proposition now follows directly. Let $T_{i_1} \prec T_{i_2} \prec \cdots \prec T_{i_s}$ be chosen (by the assertion just established) so that

$$\sum_{\varphi_i' \in \mathcal{G}'} \mu(\varphi_i') \leqq \sum_{k=1}^{s} \mu(T_{i_k}), \tag{2}$$

where the φ_i' are taken from $\mathcal{G}'$ (the timing diagram obtained when the list L' is used). Note that ω' can be written as:

$$\omega' = \tfrac{1}{2} \sum_{T_k \in T} \mu(T_k) + \sum_{\phi_i' \in \mathcal{G}'} \mu(\phi_i'). \tag{3}$$

From (2) and (3) we have

$$\omega' \leqq \tfrac{1}{2}\left(\sum_{T_k \in T} \mu(T_k) + \sum_{k=1}^{s} \mu(\phi_i')\right). \tag{4}$$

Since the following inequalities hold:

$$\omega \geqq \tfrac{1}{2} \sum_{T_k \in T} \mu(T_k) \tag{5}$$

$$\omega \geqq \sum_{k=1}^{s} \mu(T_{i_k}) \tag{6}$$

(where (6) follows from the fact that $T_{i_1} \prec T_{i_2} \prec \cdots \prec T_{i_s}$), then we have from (4), (5), and (6)

$$\omega' \leqq \tfrac{1}{2}(2\omega + \omega) = \frac{3\omega}{2}$$

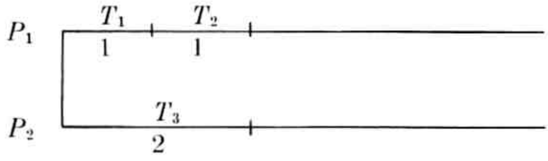
and the proposition follows.

The following example shows that the upper bound of $\frac{3}{2}$ cannot be replaced by any smaller value.

Example 4: $n = 2$, L: (T_1, T_3, T_2), L': (T_1, T_2, T_3)

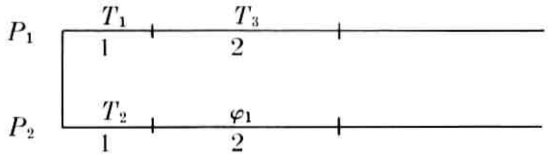
$G(\prec, \mu)$: $T_1/1$ $T_2/1$ $T_3/2$

$\mathcal{S}$:



$\omega = 2$

$\mathcal{S}'$:



$\omega' = 3$

Therefore, $\omega/\omega = \frac{3}{2}$ and the upper bound of the proposition is achieved.

Before stating the main theorem we introduce some notation. Let $T = \{T_1, \cdots, T_m\}$ be a set of tasks. Let $G = G(<,\mu)$ and $G' = G'(<',\mu')$ be two directed graphs for T with the partial-orders $<, <'$ and the time functions μ,μ'. We say that $G \leqq G'$ if:

(*i*) $\mu' \leqq \mu$, i.e., $\mu'(T_j) \leqq \mu(T_j)$ for all $T_j \in T$.

(*ii*) $<' \subseteq <$, i.e., $T_i <' T_j$ implies $T_i < T_j$ for all $T_i, T_j \in T$.

Finally, suppose we execute the tasks *twice*, one time using the graph G, a task list L and n processors, the other time using the graph G', a task list L' and n' processors. Let ω and ω' denote the respective finishing times. We then have the

Theorem: If $G' \leqq G$ *then*

$$\frac{\omega'}{\omega} \leqq 1 + \frac{n-1}{n'}.$$

Proof: By a slight modification of the argument used in the proposition, it follows that if φ_i', $1 \leqq i \leqq v$, denote the empty tasks of $\mathcal{G}'$ then there exists a chain of tasks $T_{i_1} <' T_{i_2} <' \cdots <' T_{i_s}$ of tasks in T with the property that *whenever a processor is idle then some other processor is executing one of the* T_{i_k}. From this we conclude

$$\sum_{\varphi_i' \in \mathcal{G}'} \mu'(\varphi_i') \leqq (n'-1) \sum_{k=1}^{s} \mu'(T_{i_k}). \tag{7}$$

As before we note that

$$\begin{aligned} \omega' &= \frac{1}{n'}\left(\sum_{T_j \in T} \mu'(T_j) + \sum_{\phi_i' \in \mathcal{G}'} \mu'(\phi_i')\right) \\ &\leqq \frac{1}{n'}\left(\sum_{T_j \in T} \mu'(T_j) + (n'-1) \sum_{k=1}^{s} \mu'(T_{i_k})\right) \end{aligned} \tag{8}$$

where the inequality follows by (7). Since

$$\omega \geqq \frac{1}{n} \sum_{T_j \in T} \mu(T_j) \geqq \frac{1}{n} \sum_{T_j \in T} \mu'(T_j) \tag{9}$$

and

$$\omega \geqq \sum_{k=1}^{s} \mu(T_{i_k}) \geqq \sum_{k=1}^{s} \mu'(T_{i_k}) \tag{10}$$

then by (8), (9), and (10) we conclude

$$\omega' \leqq \frac{1}{n'}\,(n\omega + (n' - 1)\omega).$$

Hence,

$$\frac{\omega'}{\omega} \leqq 1 + \frac{n-1}{n'}$$

and the theorem is proved.

To show that this bound is best possible, we give several examples, which show that the bound can be attained (to within ε) by varying any *one* of the four parameters L, μ, $<$, or n.

Example 5: L is varied.

$$n = n', \qquad \mu = \mu', \qquad < \; = \; <'.$$

$$L = (T_1, T_2, \cdots, T_{n-1}, T_{2n-1}, T_n, T_{n+1}, \cdots, T_{2n-2})$$

$$L' = (T_1, T_n, T_{n+1}, \cdots, T_{2n-2}, T_2, T_3, \cdots, T_{n-1}, T_{2n-1})$$

$. T_1/1$

$. T_2/1$

$\vdots$

$. T_{n-1}/1$

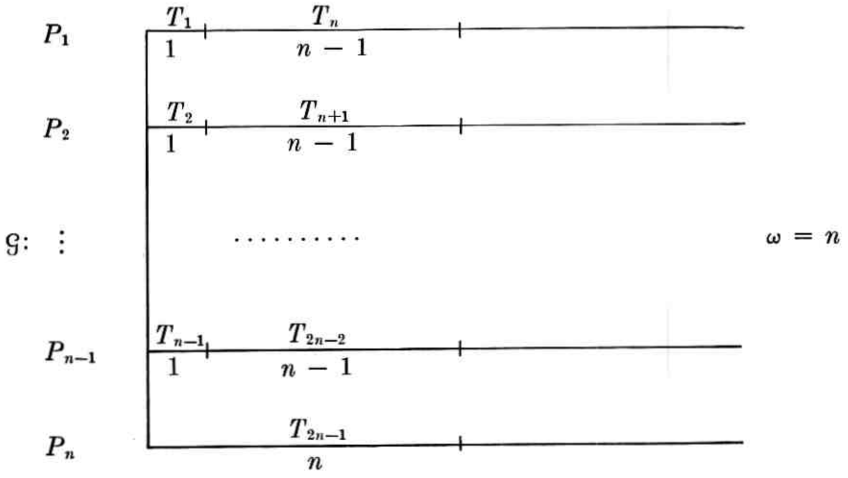
$G(<,\mu)$: $. T_n/n - 1$

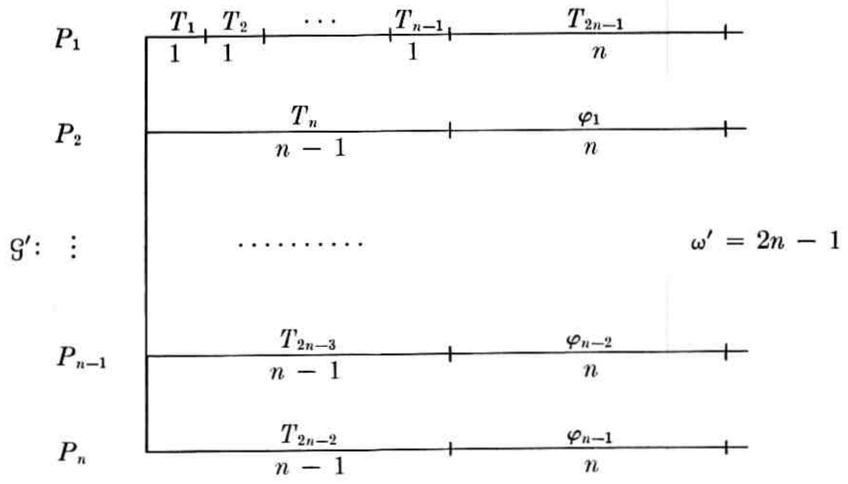
$. T_{n+1}/n - 1$

$\vdots$

$. T_{2n-2}/n - 1$

$. T_{2n-1}/n$


𝒢:
P_1
T_1
1
T_n
$n - 1$
P_2
T_2
1
T_{n+1}
$n - 1$
P_{n-1}
T_{n-1}
1
T_{2n-2}
$n - 1$
P_n
T_{2n-1}
n
$\omega = n$


𝒢′:
P_1
T_1
1
T_2
1
T_{n-1}
1
T_{2n-1}
n
P_2
T_n
$n - 1$
φ_1
n
P_{n-1}
T_{2n-3}
$n - 1$
φ_{n-2}
n
P_n
T_{2n-2}
$n - 1$
φ_{n-1}
n
$\omega' = 2n - 1$

Thus,

$$\frac{\omega'}{\omega} = 2 - \frac{1}{n},$$

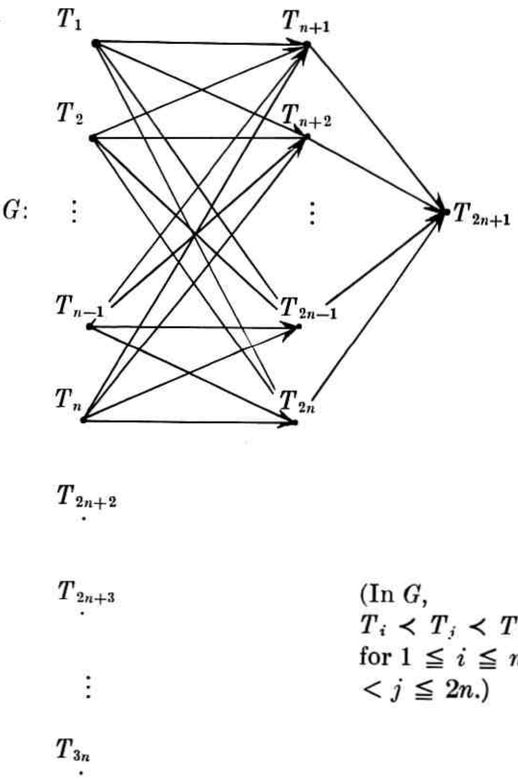
which is the value of $1 + (n - 1)/n'$ when $n = n'$.

Example 6: μ is decreased.

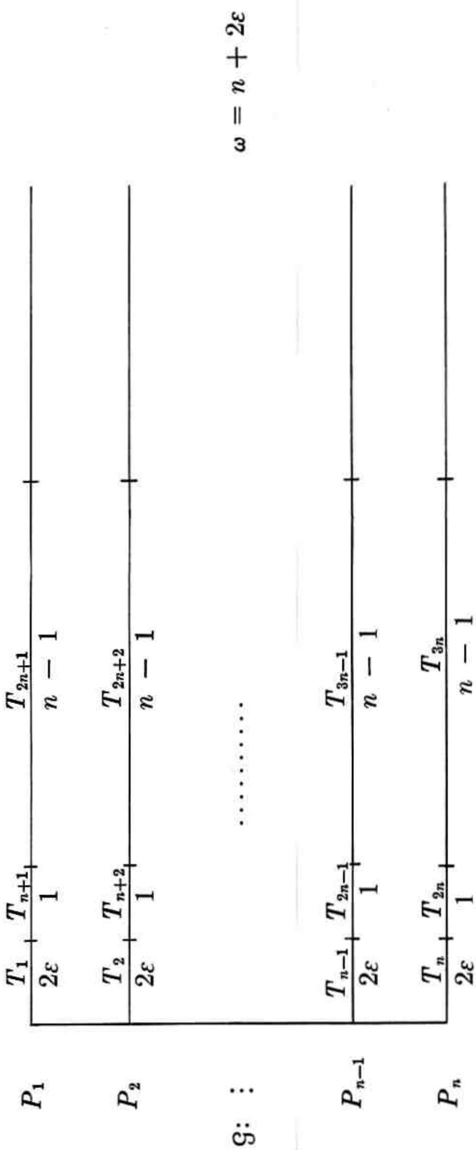
$$n = n', \qquad \mu \geqq \mu', \qquad \prec \; = \; \prec'$$

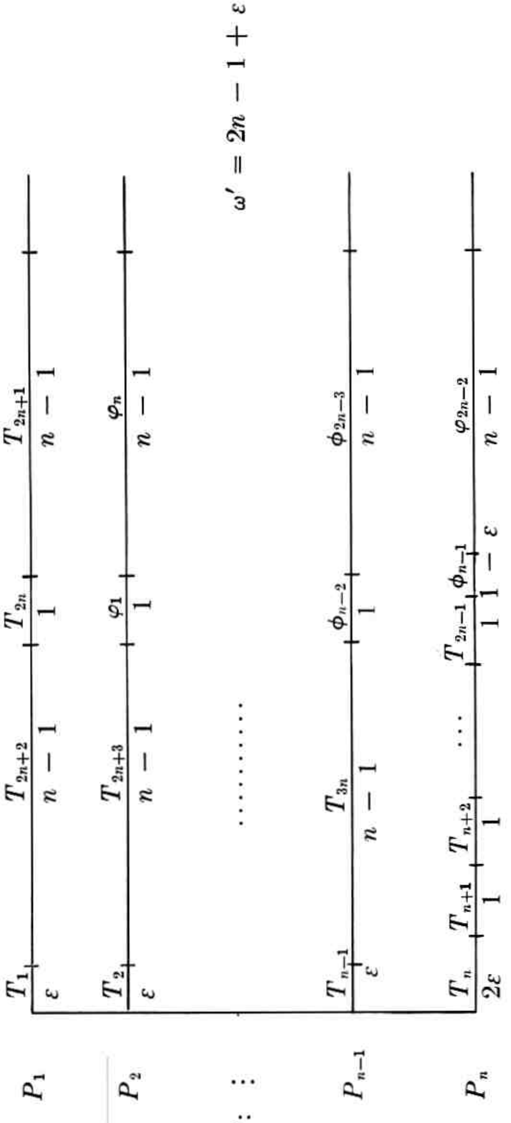
$$L = L': \quad (T_1, T_2, \cdots, T_{3n})$$

	$\mu(T_i)$	$\mu'(T_i)$
T_1	2ε	ε
T_2	2ε	ε
$\vdots$	$\vdots$	$\vdots$
T_{n-1}	2ε	ε
T_n	2ε	2ε
T_{n+1}	1	1
T_{n+2}	1	1
$\vdots$	$\vdots$	$\vdots$
T_{2n}	1	1
T_{2n+1}	$n - 1$	$n - 1$
T_{2n+2}	$n - 1$	$n - 1$
$\vdots$	$\vdots$	$\vdots$
T_{3n}	$n - 1$	$n - 1$



(In G, $T_i \prec T_j \prec T_{2n+1}$ for $1 \leqq i \leqq n < j \leqq 2n$.)


𝒢:
P_1
P_2
P_{n-1}
P_n
T_1
T_{n+1}
T_{2n+1}
T_2
T_{n+2}
T_{2n+2}
T_{n-1}
T_{2n-1}
T_{3n-1}
T_n
T_{2n}
T_{3n}
2ε
1
$n - 1$
$\omega = n + 2\varepsilon$


$\mathcal{S}'$:
P_1
T_1
ε
T_{2n+2}
$n-1$
T_{2n}
1
T_{2n+1}
$n-1$
P_2
T_2
ε
T_{2n+3}
$n-1$
φ_1
1
φ_n
$n-1$
$\vdots$
P_{n-1}
T_{n-1}
ε
T_{3n}
$n-1$
ϕ_{n-2}
1
ϕ_{2n-3}
$n-1$
P_n
T_n
2ε
T_{n+1}
1
T_{n+2}
1
$\cdots$
T_{2n-1}
1
ϕ_{n-1}
$1-\varepsilon$
φ_{2n-2}
$n-1$
$\omega' = 2n - 1 + \varepsilon$

Thus,

$$\frac{\omega'}{\omega} = \frac{2n - 1 + \varepsilon}{n + 2\varepsilon}$$

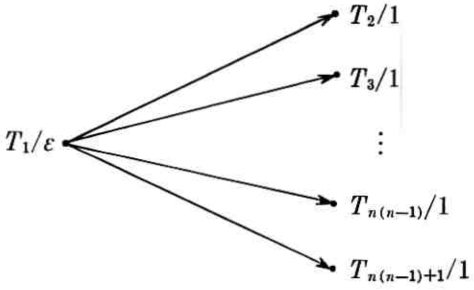
which is arbitrarily close to $2 - (1/n)$ for ε sufficiently small. We should note the interesting fact that $\omega' \geqq 2n - 1 + \varepsilon$ for *any* list L' which may be used.

Example 7: $\prec$ is relaxed.

$$n = n', \qquad \mu = \mu', \qquad \prec \supset \prec',$$

$$L = L': \quad (T_1, T_2, \cdots, T_{n(n-1)+2})$$

$G(\prec,\mu)$:



$.\,T_{n(n-1)+2}/n$

$G(\prec',\mu)$:

$.\,T_1/\varepsilon$

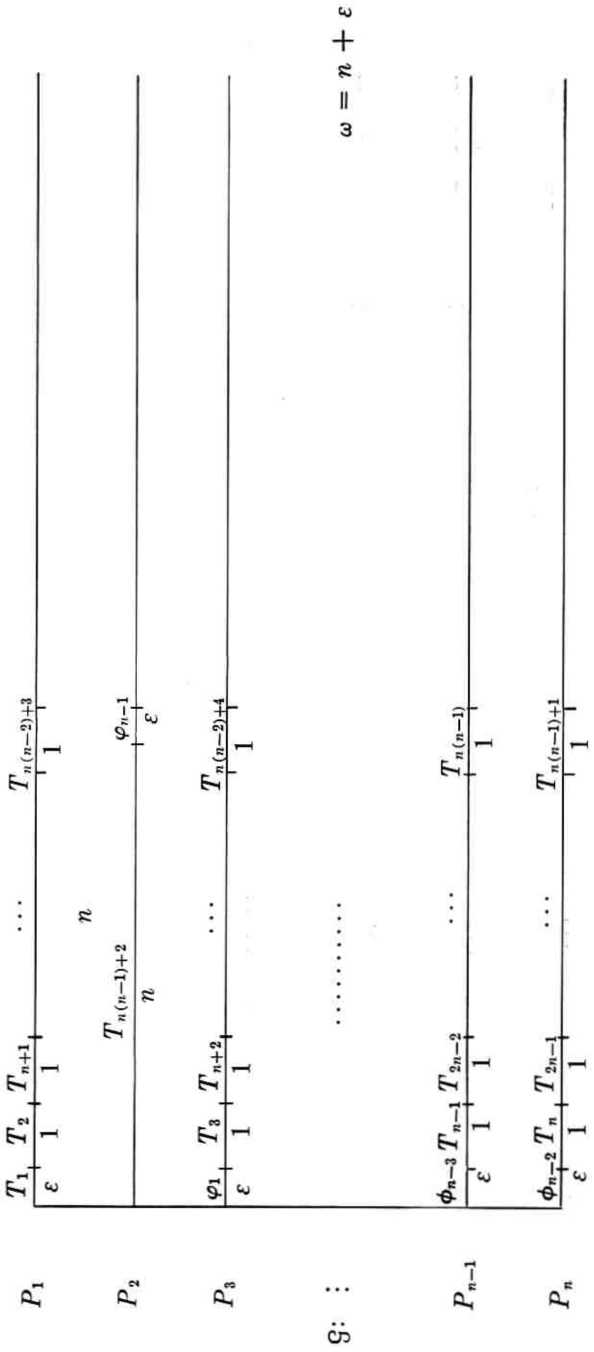
$.\,T_2/1$

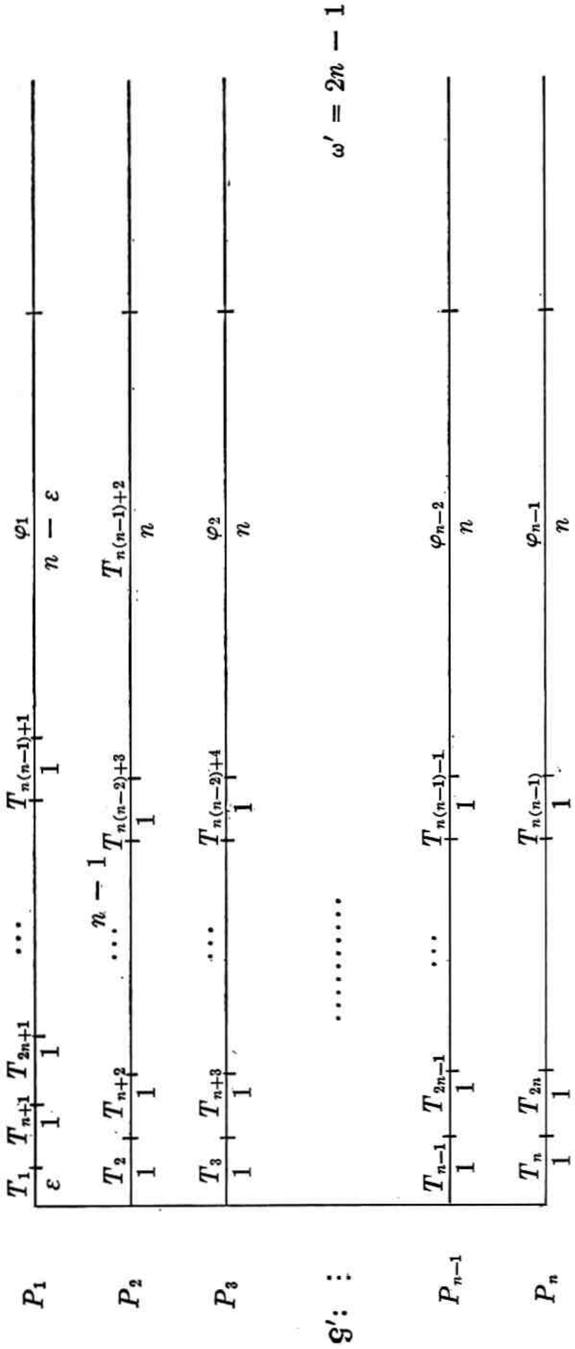
$\vdots$

$.\,T_{n(n-1)}/1$

$.\,T_{n(n-1)+1}/1$

$.\,T_{n(n-1)+2}/n$


$\mathcal{G}$:
P_1
P_2
P_3
$\vdots$
P_{n-1}
P_n
T_1
ε
T_2
1
T_{n+1}
1
$\cdots$
$T_{n(n-2)+3}$
1
$T_{n(n-1)+2}$
n
n
φ_{n-1}
ε
φ_1
ε
T_3
1
T_{n+2}
1
$\cdots$
$T_{n(n-2)+4}$
1
$\cdots\cdots\cdots$
ϕ_{n-3}
ε
T_{n-1}
1
T_{2n-2}
1
$\cdots$
$T_{n(n-1)}$
1
ϕ_{n-2}
ε
T_n
1
T_{2n-1}
1
$\cdots$
$T_{n(n-1)+1}$
1
$\omega = n + \varepsilon$


$\mathcal{S}'$:
$\omega' = 2n - 1$
P_1
P_2
P_3
P_{n-1}
P_n
T_1
ε
T_{n+1}
T_{2n+1}
$T_{n(n-1)+1}$
φ_1
$n - \varepsilon$
T_2
T_{n+2}
$n - 1$
$T_{n(n-2)+3}$
$T_{n(n-1)+2}$
T_3
T_{n+3}
$T_{n(n-2)+4}$
φ_2
T_{n-1}
T_{2n-1}
$T_{n(n-1)-1}$
φ_{n-2}
T_n
T_{2n}
$T_{n(n-1)}$
φ_{n-1}
1
n

Thus,

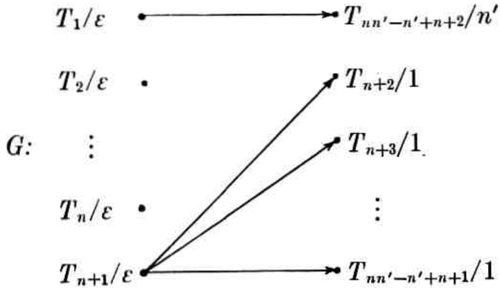
$$\frac{\omega'}{\omega} = \frac{2n - 1}{n + \varepsilon}$$

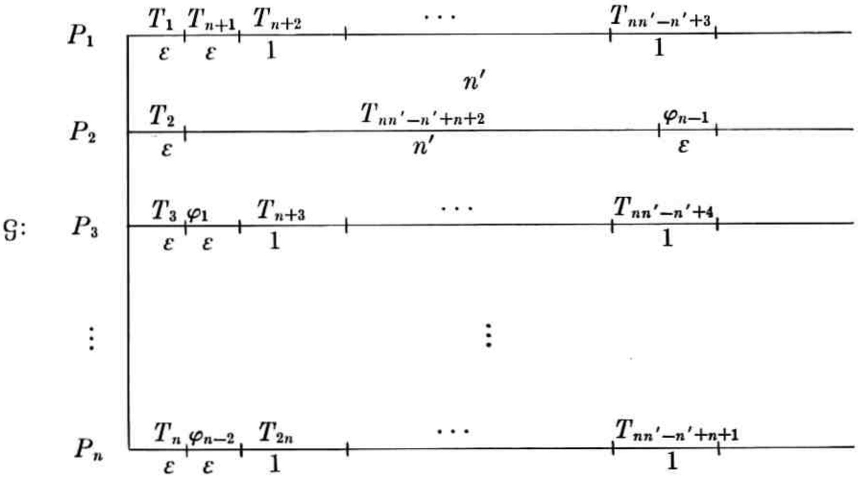
which is arbitrarily close to $2 - (1/n)$ for ε sufficiently small.

Example 8: n is varied.

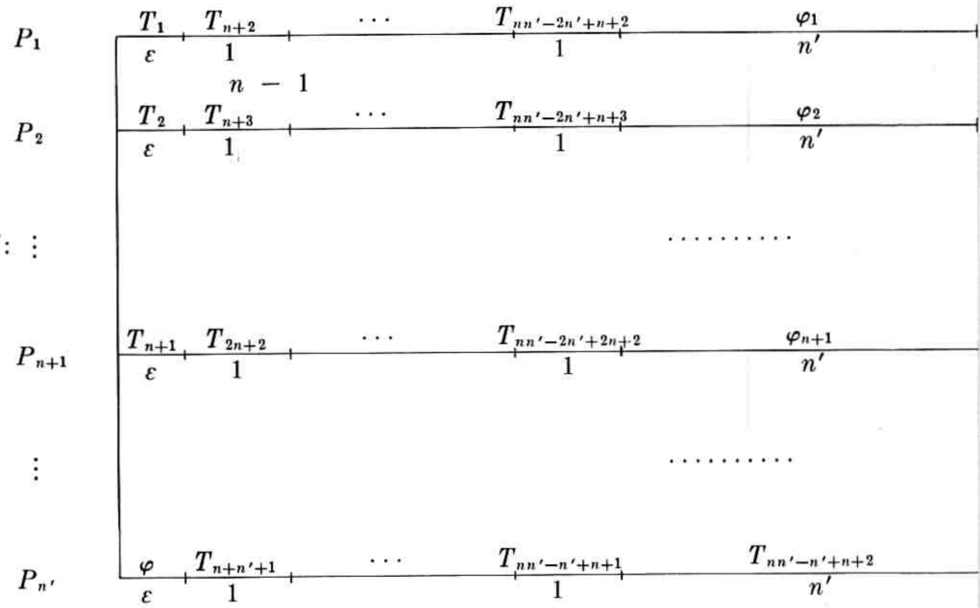
Case 1: $n < n', \qquad \mu = \mu', \qquad \prec \; = \; \prec'$

$$L = L': \quad (T_1, T_2, \cdots, T_{nn'-n'+n+2})$$




$\mathcal{G}$:
P_1
T_1 T_{n+1} T_{n+2} $\cdots$ $T_{nn'-n'+3}$
ε ε 1 1
n'
P_2
T_2 $T_{nn'-n'+n+2}$ φ_{n-1}
ε n' ε
P_3
T_3 φ_1 T_{n+3} $\cdots$ $T_{nn'-n'+4}$
ε ε 1 1
$\vdots$
$\vdots$
P_n
T_n φ_{n-2} T_{2n} $\cdots$ $T_{nn'-n'+n+1}$
ε ε 1 1

$$\omega = n' + 2\varepsilon$$



$$\omega' = n' + n - 1 + \varepsilon$$

Thus,

$$\frac{\omega'}{\omega} = \frac{n' + n - 1 + \varepsilon}{n' + 2\varepsilon}$$

which is arbitrarily close to $1 + (n - 1/n')$ for ε sufficiently small.

Case 2: $n > n'$. The construction in this case is similar to that of Case 1 and will not be presented.

We should note that in Example 8 we took $L = L'$. If it is of some consolation to a possibly battered intuition, it should be noted that if $n \leqq n'$, $\mu = \mu'$, and $\prec \; = \; \prec'$ then for any L which is chosen, it is possible to choose a suitable L' for which $\omega' \leqq \omega$.

III. CONCLUDING REMARKS

It should be pointed out here that we have not considered models of the multiprocessor system in which the priority list L is "dynamically formed" (as opposed to the *fixed* lists we have used thus far). For example, one seemingly quite reasonable way of doing this is as follows: At any time a processor is free, it immediately begins to execute the "ready" task (i.e., one which has all its predecessors completed) which currently heads the *longest chain* of unexecuted tasks (including itself). Suppose by following this algorithm in choosing tasks, we have a finishing time of ω^*. If we denote by ω_o the least possible finishing time (minimized over *all* lists), then we would like to assert something about the ratio ω^*/ω_o. It follows from what has been proved in this paper that $\omega^*/\omega_o \leqq 2 - (1/n)$, (where n is the number of processors) and we would hope that, in fact, we could show ω^*/ω_o is considerably closer to 1 than this. Unfortunately, this is not possible since it can be shown that the best possible bound on this ratio is given by

$$\frac{\omega^*}{\omega_o} \leqq 2 - \frac{2}{n+1}.$$

It is interesting to note, however, that in the case in which the partial-order $<$ on the tasks is *empty*, then this bound can be improved* to

$$\frac{\omega^*}{\omega_o} \leqq \frac{4}{3} - \frac{1}{3n},$$

which, again, is best possible.

In conclusion, one might ask just how "typical" the examples are for which ω'/ω_o is close to the upper bound $2 - (1/n)$. While very little work has been done on this aspect, empirical results (using computer simulation (see Ref. 1)) indicate that examples in which $\omega'/\omega_o \geqq 1.1$ are quite common.

IV. ACKNOWLEDGMENTS

I wish to acknowledge here the stimulating discussions I have had on this subject with S. Lin and with G. K. Manacher, who originally brought these questions to my attention.

REFERENCES

1. Manacher, G. K., The Production and Stabilization of Real-Time Task Schedules, Institute for Computer Research, Quarterly report, Univ. of Chicago, May, 1966.
2. Ochsner, B. P., Controlling a Multiprocessor System, Bell Laboratories Record, February, 1966
3. Richards, P., Parallel Programming, Report No. TD-B60-27, Tech. Operations Inc., August, 1960.
4. Heller, J., Sequencing Aspects of Multiprogramming, JACM, *8*, 1961, pp. 426–439.
5. Kelley, J. L., *General Topology*, Van Nostrand, Princeton, 1955.

* The proofs of this and the preceding result will appear in a later paper.

Phase and Amplitude Measurements of Coherent Optical Wavefronts

By JOSEPH T. RUSCIO

(Manuscript received March 24, 1966)

A phase-locked laser loop has been used as an amplitude and phase measuring device for coherent optical wavefronts. A relative phase resolution on the order of one degree and an amplitude resolution accurate to one dB or better were obtained. The system and measuring techniques used are described, and the results obtained are illustrated by several examples.

I. INTRODUCTION

A laser phase-locked loop[1] consisting of two laser oscillators has been used to measure the relative phase and amplitude of the wavefront of a laser beam. A phase resolution on the order of one degree and an amplitude accuracy better than one dB have been obtained. This system was used to analyze the optical qualities of devices placed in the beam's path by measuring their effect on the wavefront. The system and techniques used along with the results obtained are described and illustrated in this paper.

II. DESCRIPTION OF THE MODIFIED PHASE-LOCKED LASER LOOP

The phase-locked system is shown in Fig. 1. It consists of controlled and uncontrolled optical oscillators which are single-frequency helium-neon lasers operated at 6328Å. Details of the oscillators' characteristics are shown in Figs. 2 and 3. The beam waists and spot sizes are defined and calculated in Appendix A. The two lasers used initially are shown in Fig. 2; however, tube replacements required a different combination of mirrors to maintain a single transverse mode, so the final measurements were made with the lasers shown in Fig. 3. Results are identified with the lasers used.

Prior to combining the two beams (Fig. 1) on the surface of the photomultiplier by means of a mirror and a beam splitter, the output beam of

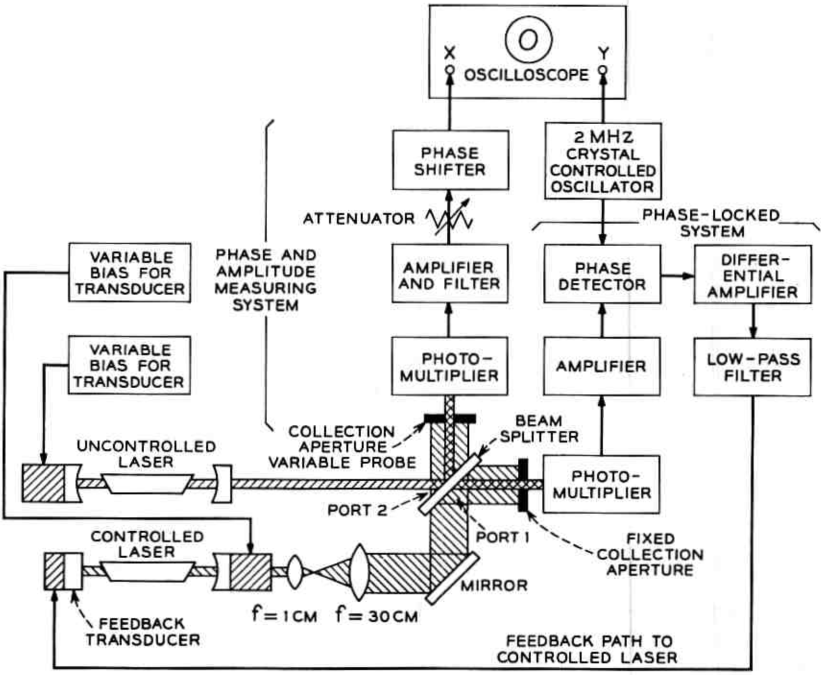


Fig. 1 — Phase-locked optical maser system.

the controlled laser, which will be referred to as the reference beam, is collimated by a telescope.

III. THEORY OF OPERATION

The beam splitter in Fig. 1 provides two outputs: Port 1 to phase-lock the system and Port 2 for making the phase and amplitude measurements. The photomultipliers are square law detectors. Thus, if the field at the photosensitive surface is

$$E = E_c \cos \omega_c t + E_u \cos \omega_u t$$

where

E_c is the controlled oscillator field amplitude,

E_u the uncontrolled oscillator field amplitude,

and

ω_c, ω_u the respective angular frequencies, then since E_c and E_u have the same polarization, the resulting photocurrent is proportional to

$$E^2 = \tfrac{1}{2}E_c^2 + E_cE_u \cos(\omega_c - \omega_u)t + \tfrac{1}{2}E_u^2 \tag{1}$$

which consists of a dc term $\frac{1}{2}(E_c^2 + E_u^2)$ plus the difference frequency term $[E_c E_u \cos(\omega_c - \omega_u)t]$.

In the original phase-lock loop,[1] the two lasers were locked at the same frequency with a dc error voltage proportional to their phase difference. When the loop locks, the controlled laser tracks the frequency of the uncontrolled laser in a manner such that the instantaneous phase error α remains smaller than 90° in absolute magnitude. A discussion of the phase relationships in the loop is given in Appendix B.

To improve the measurement of phase and amplitude, the laser oscillators were phase-locked at a fixed frequency difference of 2 MHz by using an additional phase detector with a 2-MHz crystal-controlled oscillator as a reference. When the lasers are tuned so that their difference frequency is 2 MHz, the 2-MHz output from the photomultiplier is amplified and applied to the phase detector. The phase detector output is a dc error voltage proportional to the phase difference between the 2-MHz beatnote and 2-MHz reference signal (Appendix B). This error voltage

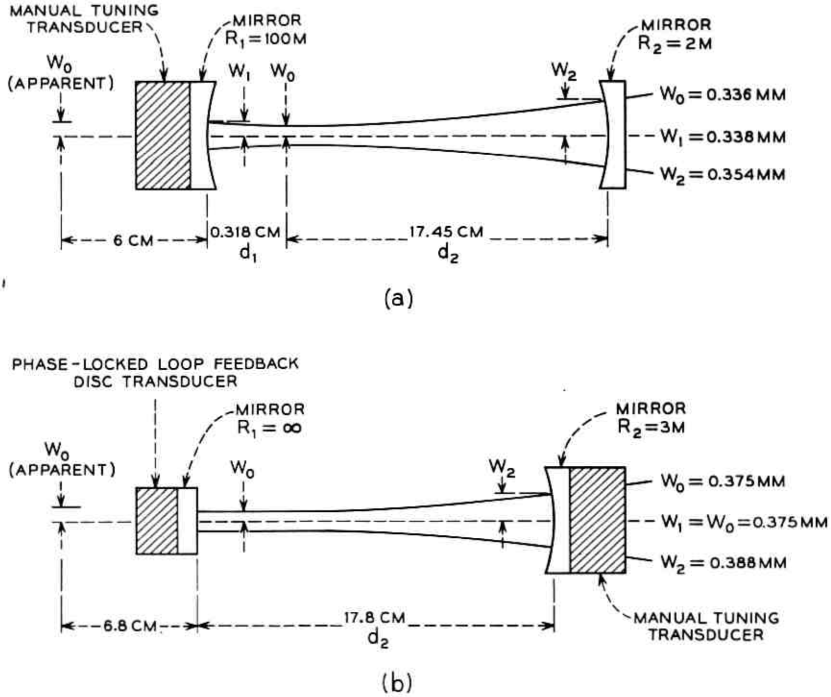


Fig. 2 — Initial lasers; (a) uncontrolled laser oscillator (signal), (b) controlled laser oscillator (reference).

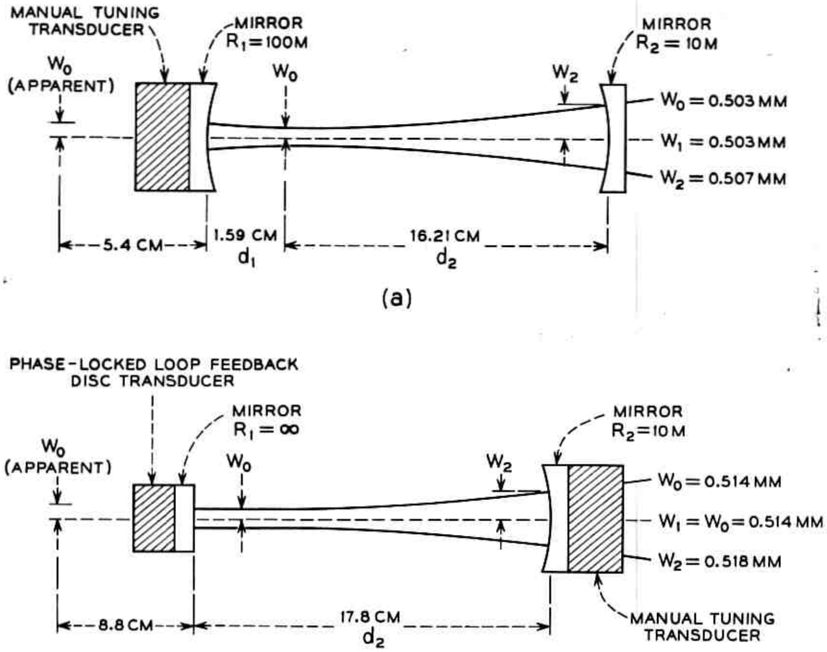


Fig. 3 — Replacement lasers; (a) uncontrolled laser oscillator (signal), (b) controlled laser oscillator (reference).

is fed back through a differential amplifier to a piezoelectric disc transducer. A mirror mounted on this transducer forms one end of the controlled laser cavity. There is an additional transducer-mounted mirror on the other end of the laser cavity; this is used for initial tuning (see Figs. 2(b) and 3(b)). The error voltage causes the cavity length and hence the frequency to change in a direction such that the phase error is decreased. When the loop locks, the controlled laser tracks the frequency of the uncontrolled laser in a manner such that the instantaneous phase error between the two 2-MHz signals remains less than 90°. The loop tracks over a frequency range of ±50 MHz, based on a feedback voltage of ±80 volts and a piezoelectric transducer having a sensitivity of 0.6 MHz/volt. This means that the phase difference between the reference signal and the beatnote signal remains less than 90° in absolute magnitude as long as the frequency of the uncontrolled laser does not vary more than ±50 MHz.

IV. TECHNIQUE OF MEASUREMENT

Port 2 of the beam splitter provides an output which is utilized for phase and amplitude measurements; this permits scanning the com-

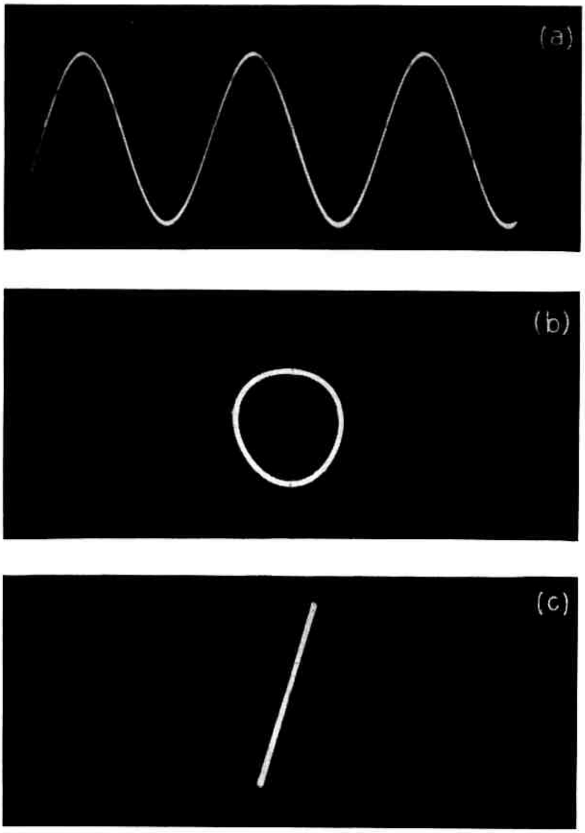


Fig. 4. — Beatnote (2 MHz) Lissajou; (a) 2-MHz Beatnote, (b) lissajou pattern indicating phase difference between two 2-MHz signals, (c) zero phase shift between two 2-MHz signals.

bined beams without interfering with the phase-locked loop. A circular collection aperture of a few mils diameter is used to scan the superimposed wavefronts, selecting the "point" area detected by the photomultiplier. The phase of the 2-MHz beatnote obtained from the photomultiplier is dependent on the position of the "point" area on the wavefronts. Frequency selective circuits, including a 2-MHz tuned circuit in the photomultiplier output and a crystal filter (2-MHz center frequency, 4-kHz bandwidth), assist in maintaining a signal-to-noise ratio that is better than 40 dB. The result is a well-defined 2-MHz signal (Fig. 4(a)), which with the 2-MHz reference can be used to produce Lissajou patterns, as in Figs. 4(b) and 4(c), on an oscilloscope. By this means relative phase measurements between the two beams are possible.

Distortion of the pattern in Fig. 4(b) is due to limitations in the horizontal amplifier of the oscilloscope. Measurement of the beatnote amplitude as a function of the probe position from the beam axis is used to determine the relative amplitude of the wavefronts. The techniques and theory used for both phase and amplitude measurements will be described.

V. PHASE MEASUREMENT

The phase measurement is based on the fact that each of the two spherical wavefronts of radii R_1 and R_2 can be expressed approximately as

$$E_1 = \exp\,(jkd^2/2R_1) = \exp\,(j\Phi_1),$$

where

$\Phi_1 = (kd^2/2R_1)$,

d is the distance from the beam axis, and

$k = 2\pi/\lambda$.

$E_2 = \exp\,(jkd^2/2R_2) + \gamma = \exp\,(j\Phi_2) + \gamma$,

where

$\Phi_2 = (kd^2/2R_2)$

and γ is the phase difference between the two wavefronts on the axis. Thus, the phase difference between the two wavefronts is given by

$$\Delta\Phi = \Phi_2 - \Phi_1 + \gamma = (kd^2/2)\,(1/R_2 - 1/R_1) + \gamma.$$

The telescope reduces the divergence of the reference beam so that its radius of curvature,* R_1, can be considered infinitely large, therefore,

$$\Phi_1 \approx 0$$

and

$$\Delta\Phi \approx (kd^2/2R_2) + \gamma.$$

From Fig. 5, it can be seen that if the phase shift as indicated by the Lissajou patten is adjusted to be zero at the center of the beam (by means of an auxiliary phase shifter), all measurements can be made relevative to this reference and the relative phase shift becomes

$$\Delta\Phi \approx (kd^2/2R_2).$$

* Calculations for the radii of curvature involved using the telescope are given in Appendix C.

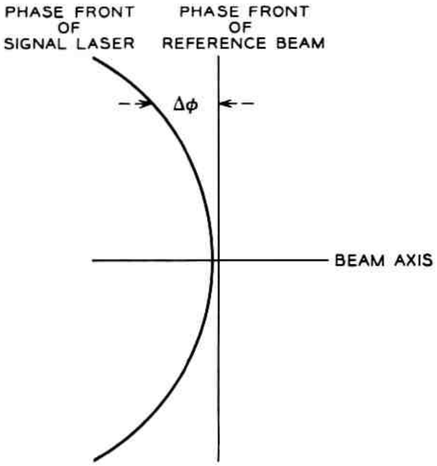


Fig. 5 — Phase relationship between spherical and planar wavefronts.

Measurement of ΔΦ as a function of distance (d) from the beam axis provides a means of determining R_2.

The experimental layout for measuring the relative phase between the two beams is shown in Fig. 1. The movable collection aperture positioned directly in front of the photomultiplier can be moved in 5-mil increments along the horizontal or vertical axis. Changes in phase with position is plotted as shown in Figs. 6(a), (b) and 7(a), (b).

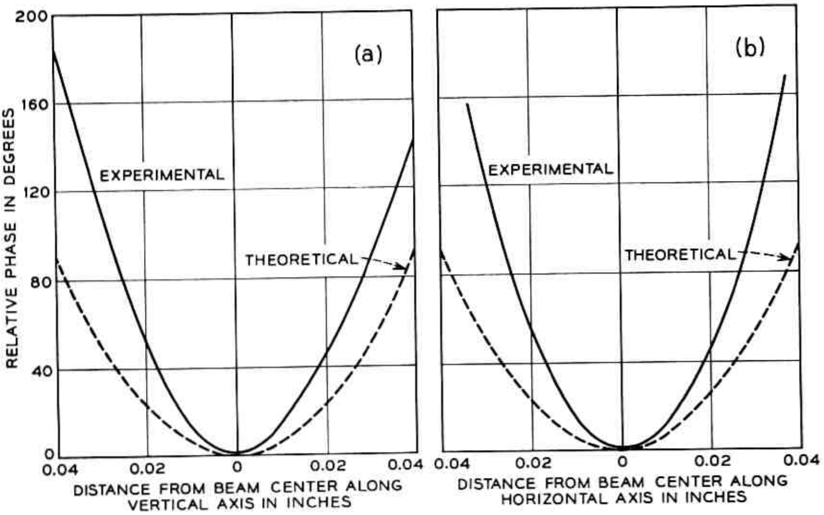


Fig. 6 — Optical wavefronts (laser in Fig. 2(a)).

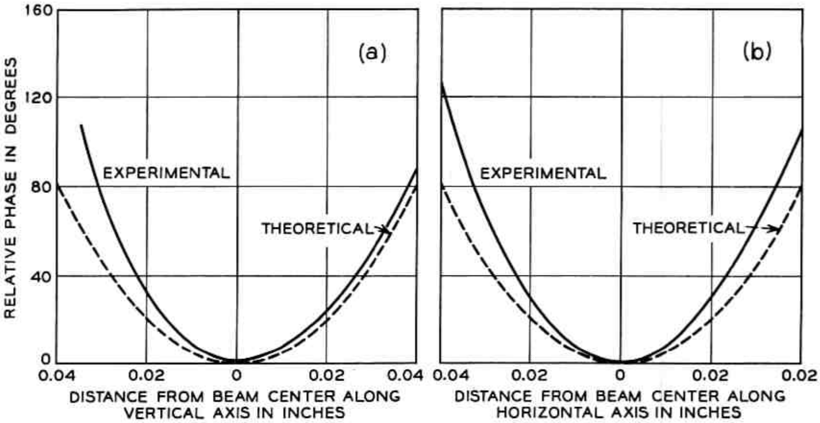


Fig. 7 — Optical wavefronts (laser in Fig. 3(a)).

The curves in Figs. 6(a) and (b) are for the laser shown in Fig. 2(a). Measurement of the optical wavefront was made at a distance of 3 meters from the apparent beam waist using a 0.015-inch diameter collection aperture. Similar data for the laser in Fig. 3(a) are shown in Figs. 7(a) and (b) in which case the collection aperture diameter was 0.009 inch. Theoretical curves indicate radii of curvature less than the measured values for all cases. The radius of curvature at a distance z from the apparent beam waist location is given by[2]

$$R = z\left[1 + \left(\frac{z_0}{z}\right)^2\right],$$

where $z_0 = \pi w_0^2/\lambda$, w_0 being the beam-waist radius, and $\lambda = 0.6328\mu$.

The disagreement between the theoretical and experimental values has not been resolved. This problem remains under consideration as work continues in this area.

VI. FIELD AMPLITUDE MEASUREMENTS

In a similar manner, observing the 2-MHz beatnote amplitude as a function of probe distance perpendicular to the beam axis provides a means of determining the field amplitude distribution of the combined beams. With the reference beam enlarged, in this case to 30 times its initial size, the amplitude of the reference beam over the distance scanned is relatively constant; therefore, the amplitude distribution measured is, in fact, the relative amplitude of the signal beam. The accuracy is governed by the variation in intensity of the reference beam over the area scanned as shown in Fig. 8. With a beam reference spot size of 1-inch diameter, scanning a distance of $0.2a$ (a = beam radius) from the beam axis introduces an error of 0.4 dB in the relative measurements.

Examples of the measured amplitude distribution as a function of distance from the center of the beam are shown in Figs. 9(a), (b) and 10(a), (b). In Fig. 9(b), which applies to the laser shown in Fig. 2(a), the theoretical Gaussian curve ($\exp(-r^2/a^2)$, a being the beam radius where the field amplitude falls to $1/e$, and r the distance from the beam axis) agrees quite closely with the measured values. Figs. 10(a) and (b) show amplitude distribution curves for the laser in Fig. 3(a); in these an unexplained lack of symmetry appears.

VII. DETERMINATION OF PROPERTIES OF OPTICAL ELEMENTS

In addition to measuring the signal laser's optical wavefront, it was also possible to determine the effects of putting a lens in the signal beam. Results of this experiment are described.

To facilitate measurement of lenses, the signal beam was also collimated so that now both beam wavefronts were planar. Under these conditions, placing a glass lens in the signal laser beam produced a phase-front at the collection aperture dependent on the focal length of the lens. The experimental arrangement is shown in Fig. 11 and the results of measurements on a 86.6-cm focal length lens are shown in Fig. 12. The measurements agree quite closely with the theoretical values.

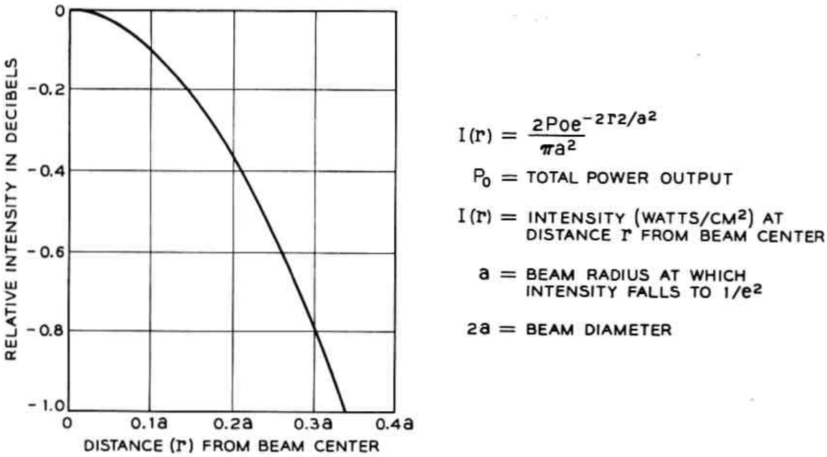


Fig. 8 — Intensity distribution of Gaussian curve.

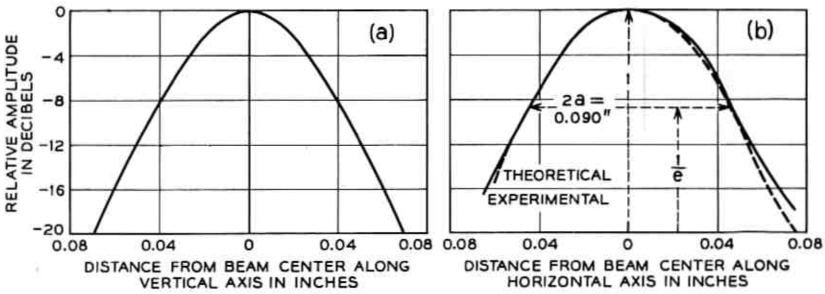


Fig. 9 — Optical wavefront—amplitude (laser in Fig. 2(a), 0.009-inch collection aperture).

VIII. POSSIBLE IMPROVEMENTS

In Fig. 11 it can be seen that the lens being tested is common to both the phase-lock loop branch and the phase and amplitude measuring system. To phase-lock the loop, the two beams must be made coincident at the photomultiplier. A fixed device, such as glass lens, can be inserted in the system, the beams aligned and the loop locked. However, if the item under test introduces random variations which displace the beams relative to each other, the phase-lock loop is affected and meaningful measurements are not possible. To eliminate this problem, the setup shown in Fig. 13 is preferable. Under these conditions, the phase-locked loop is independent of the component under test and is therefore, not affected by any instability introduced. This method may require lasers with greater output powers because of the additional beam splitters required.

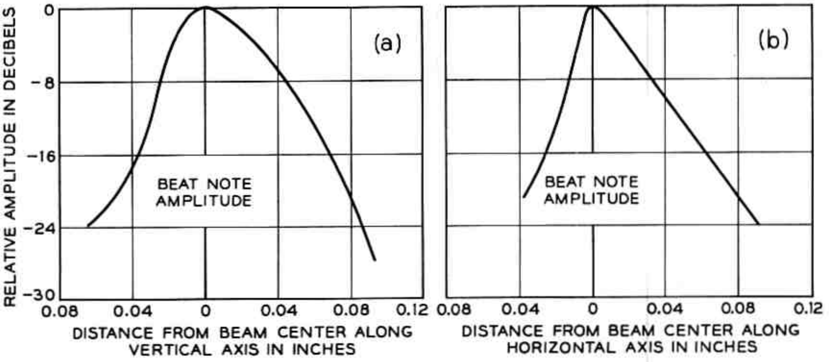


Fig. 10 — Optical wavefront—amplitude (laser in Fig. 3(a), 0.009-inch collection aperture).

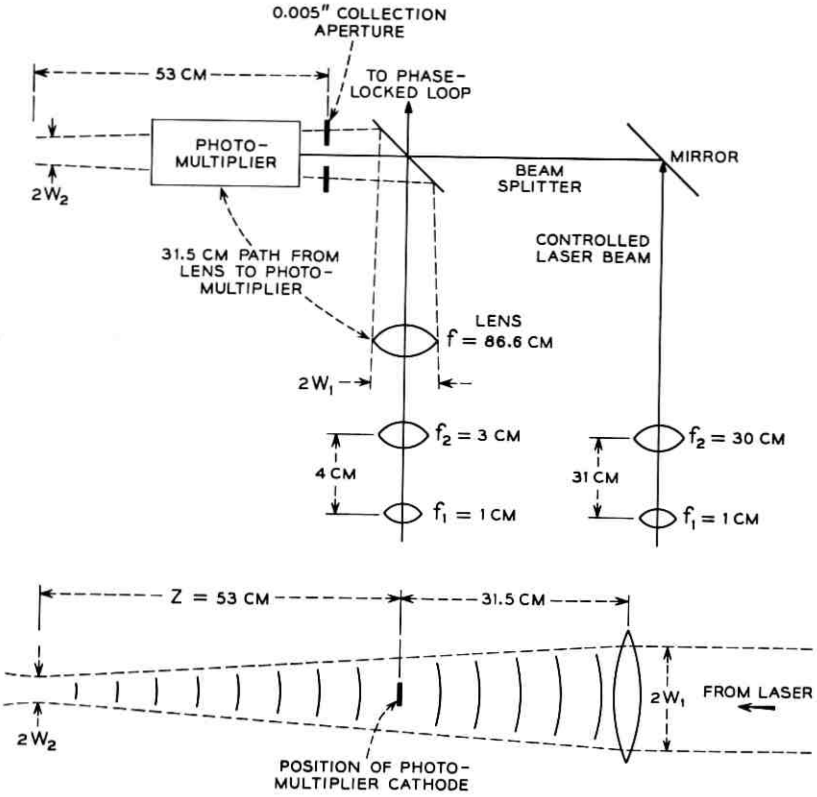


Fig. 11 — Experimental arrangement.

The system as it stands is sensitive to acoustic noises and for accurate phase measurements the laser must be maintained within a vault.[1] Enclosure in the vault permits phase-lock to be maintained for periods of 2 or 3 hours, with occasional tuning adjustments of the laser by means of the transducer-mounted mirror. Under these conditions the phase-lock is sufficiently stable to permit measurements without too much difficulty; however, it would be desirable to have portable lasers that could be used under less ideal conditions than a closed vault. Use of a transducer with a higher resonant frequency and additional gain in the feedback loop should increase the phase-lock stability.

IX. ACKNOWLEDGMENTS

Suggestions by L. H. Enloe were valuable, as were talks with J. L. Rodda, O. E. DeLange, A. F. Dietrich, T. Li, and W. H. Steier. C. P.

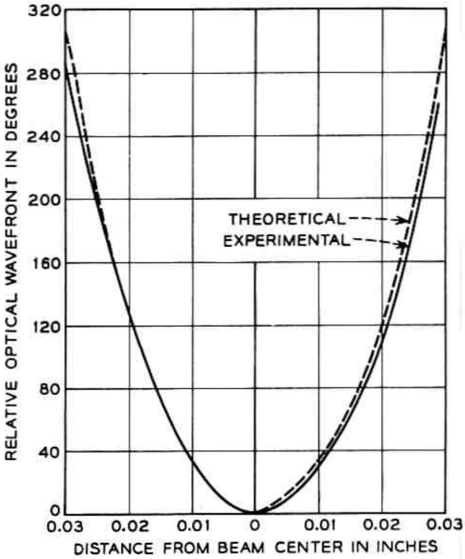


Fig. 12 — Phase front produced by a 86.6-cm focal length lens.

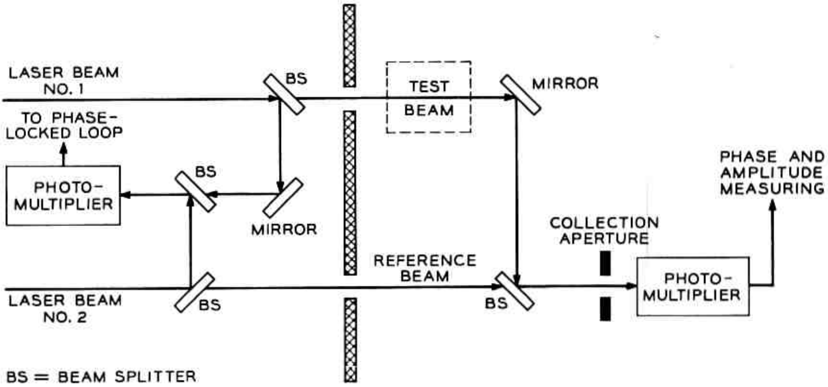


Fig. 13 — Proposed improved system.

Frazee's assistance on mechanical design was very helpful. The laser tubes were provided by E. I. Gordon of Bell Telephone Laboratories.

APPENDIX A

Calculation of Beam Waists and Spot Sizes

The following notations, some of which have already appeared, will apply in the following development:

w = spot radius, defined as the radius at which the field amplitude falls to $1/e$ of its maximum value on the z-axis.
w_0 = beam waist, which is the minimum spot radius.
w_1, w_2 = spot radii at their respective mirrors.
R_1, R_2 = radii of curvature of the two laser mirrors. One of the references[4] uses b_1 and b_2 as the notation for the radii of curvature of the mirrors.
d = separation of two laser mirrors.
d_1, d_2 = distances to mirrors as shown in Figs. 2 and 3.
λ = wavelength = 6328Å.

The beam waist w_0 is given by the following[3]:

$$w_0^2 = \frac{\lambda \sqrt{d(R_1 - d)(R_2 - d)(R_1 + R_2 - d)}}{\pi(R_1 + R_2 - 2d)}. \tag{2}$$

Output spot sizes were calculated using[4]

$$\left(\frac{w_1}{w_2}\right)^2 = \frac{R_1}{R_2} \cdot \frac{R_2 - d}{R_1 - d} \tag{3}$$

and

$$(w_1 w_2)^2 = \left(\frac{\lambda}{w}\right)^2 \frac{R_1 R_2 d}{R_1 + R_2 - d}. \tag{4}$$

Locations of the beam waists were obtained from[4]

$$d_1 = \frac{(dR_2 - d)}{R_1 + R_2 - 2d} \tag{5}$$

and

$$d_2 = \frac{(dR_1 = d)}{R_1 + R_2 - 2d}. \tag{6}$$

To compute the apparent beam waist location, it is necessary to first correct for the negative lens effect of the output mirror.[5] The output mirror acts like a negative lens transforming the phase front of the light wave emerging from the mirror. A mirror with a radius of curvature R and an index of refraction n transforms the phase front so that the radius of curvature is R/n. In this case (Fig. 2(a)), $R = 2$ m, $n = 1.46$ (quartz) so that the new radius of curvature

$$R' = R/n = 2/1.46 = 1.37 \text{ m}.$$

With this radius of curvature, using[5]

$$z = \frac{R'}{1 + \left(\frac{\lambda R'}{\pi w_2^2}\right)}, \tag{7}$$

the apparent beam waist appears to be at distance $z = 23.8$ cm from the output mirror; this places the apparent beam waist 6 cm outside the laser as shown in Fig. 2(a). Similar computations produce the apparent beam waist location for the other lasers as indicated in their respective figures. The radius of the apparent beam waist is obtained using the value of R_2/n rather than R_2 in (2).

APPENDIX B

Phase Relationships of Phase-Locked Loop

It has been shown when the field at the photomultiplier is $E = E_c \cos \omega_c t + E_u \cos \omega_u t$ that the difference frequency term is $E_c E_u \cos (\omega_c - \omega_u)t$ where ω_c and ω_u are the respective angular frequencies of the controlled and uncontrolled laser beams. To determine the phase relationships in the dc system, the field amplitudes are omitted and the angular frequencies and their phases are expressed as

$$\cos [(\omega_c t + \varphi_1) - (\omega_u t + \varphi_2)] = \cos [(\omega_c - \omega_u)t + \varphi_1 - \varphi_2].$$

Let

$$(\omega_c - \omega_u) = \Delta\omega \quad \text{and} \quad \varphi_1 - \varphi_2 = \Delta\varphi$$

then the error signal from the photomultiplier is

$$\cos (\Delta w \cdot t + \Delta\varphi).$$

When the system is phase-locked, the frequency difference $\Delta w \cdot t$ is equal to zero, therefore,

$$\cos (\Delta w \cdot t + \Delta\varphi) = \cos \Delta\varphi.$$

This, in turn, can be written as $\sin (90° - \Delta\varphi)$; the error signal for the phase-lock laser loop. At phase-lock, this error voltage approaches zero and $\Delta\varphi$, the phase difference, is equal to $90° \pm \alpha$, where α is the instantaneous phase error of the loop. The output of the phase detector is proportional to $\sin \alpha$ and since α is small, $\sin \alpha \doteq \alpha$, the dc error voltage is proportional to the phase difference.

When the system was modified to permit the use of a 2-MHz intermediate frequency, an additional phase detector was utilized to develop an error signal based on the output difference frequency term of the photomultiplier and a 2-MHz reference signal. Computations similar to the above show that the output error voltage of the IF phase detector is also proportional to the phase difference between the 2-MHz beat frequency from the photomultiplier and the reference frequency of 2-MHz.

APPENDIX C

Beam Transformation Using A Telescope

The location of the output beam waist after passing through a telescope consisting of two lenses with focal lengths f_1 and f_2, spaced at a distance $d = f_1 + f_2 \pm d$, is as follows,[2] where it is assumed the telescope is adjusted so that the misadjustment Δd is approximately equal to zero.

$$S_2 = -\left[(S_1 - f_1)\frac{f_2^2}{f_1^2} + f_2\right]$$

where S_1 is the distance from the input beam waist to the first lens and S_2 is the distance to the output beam waist from the output lens. Since we know S_1, substitution of the values for the reference telescope [$f_1 = 1$ cm and $f_2 = 30$ cm] gives us a value of 270 m for S_2.

The radius of curvature of the wavefront is[2]

$$Rz = z\left[1 + \left(\frac{z_0}{z}\right)^2\right],$$

where $z_0 = \pi w_0^2/\lambda$ and z is the distance from the output beam waist to the photomultiplier [in this case $z = S_2 - 3$ m $= 267$ m]. Thus, R_z is determined to be approximately 2000 m which for our purpose is considered to be a planar phase front. Since there is a good possibility that the lens arrangement is not well adjusted, it is wise to observe the output beam of the telescope over an extended distance to insure that there is no noticeable divergence or convergence of the beam diameter.

REFERENCES

1. Enloe, L. H. and Rodda, J. L., Laser Phase-Locked Loop, Proc. IEEE, Feb., 1965, *53*, p. 165.
2. Kogelnik, H., Imaging of Optical Modes—Resonators with Internal Lenses, B.S.T.J., *44*, March, 1965, pp. 455–494.
3. Kogelnik, H., Modes in Optical Resonators, in *Advances in Lasers*, edited by A. K. Levine, Dekker Publishers, New York (to be printed).
4. Boyd, G. D. and Kogelnik, H., Generalized Confocal Resonator Theory, B.S.T.J., *41*, July, 1962, pp. 1347–1369.
5. Fork, R. L., Herriott, D. R., and Kogelnik, H., A Scanning Spherical Mirror Interferometer for Spectral Analysis of Laser Radiation, Appl. Opt., *3*, December, 1964, pp. 1471–1484.

State of the Art in GaP Electroluminescent Junctions*

By M. GERSHENZON

(Manuscript received June 10, 1966)

Quantum efficiencies and brightness values for green and particularly for red light emission from currently available GaP p-n junctions in forward bias at room temperature are sufficiently high to merit consideration in electroluminescence applications where the human eye is the detector.

I. INTRODUCTION

Although the recombination radiation from forward-biased GaP p-n junctions could be used for the same applications as the emission from lower band-gap materials (e.g., in photon-coupled circuitry), the GaP emission occurs mainly in the visible portion of the spectrum and is thus more appropriate for applications where the human eye is the detector. To obtain emission in the visible from a forward-biased p-n junction, one needs a semiconductor with a band gap greater than 1.8 eV. The II–VI compounds that meet this requirement cannot be made into simple p-n junctions (although some of their alloys can). Hence, only GaP, BP and the various polytypes of SiC are considered. (These are all indirect gap semiconductors so that stimulated emission is not normally expected.) Of these three, GaP (band gap 2.26 eV) is characterized by the simplest materials technology.

II. RADIATIVE RECOMBINATION MECHANISMS

Fig. 1 shows a typical room temperature forward-bias emission spectrum from a diode prepared by Zn diffusion into an n-type crystal containing Te and O. Two emission bands appear in the visible, separated both spectrally and spatially. A weak green band is generated

* Presented at the Seminar on Electroluminescence and Semiconductor Lasers sponsored by the New York Section IEEE at Stevens Institute of Technology, May 11, 1966.

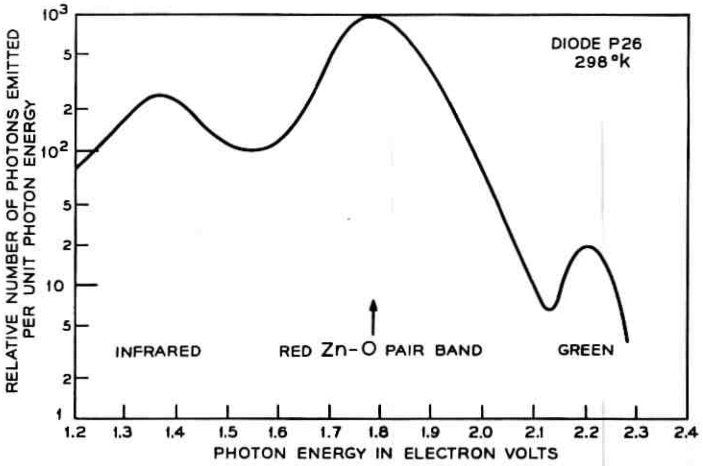


Fig. 1 — Emission spectrum from a forward biased Zn-diffused diode at room temperature.

close to the junction proper, while a much stronger red band seems to originate on the p-side of the junction. Infrared emission seen in Fig. 1 will not be discussed.

2.1 *The Red Emission*

Mostly by comparison with photo-luminescence, it has been shown that the red band is due to donor-acceptor pair recombination involving shallow Zn acceptors and deep O donors.[1] External photoluminescence quantum efficiencies of up to 11 percent have been reported at room temperature in p-type samples.[2] Zn-O pair band recombination in a p-n junction is sketched in Fig. 2. On the p-side of the junction the Zn level ($N_A \approx 2 \times 10^{18}$ cm^{-3}) is about half full of holes in thermal equilibrium. Injected minority carrier electrons are captured efficiently by the ionized compensating O donors and, because the O donor is relatively deep, the electrons remain trapped, with little thermal ionization back to the conduction band, until they recombine radiatively with holes on the Zn acceptors. This situation is identical to the Zn-O pair emission in photoluminescence and should lead to high efficiencies. On the n-side of the junction, the O donors are always filled with electrons. Injected minority carrier holes may be captured by the empty Zn acceptors, but, because the Zn acceptor level is quite shallow, they are thermally released back to the valence band, from where they may find other means to recombine. In the space-charge

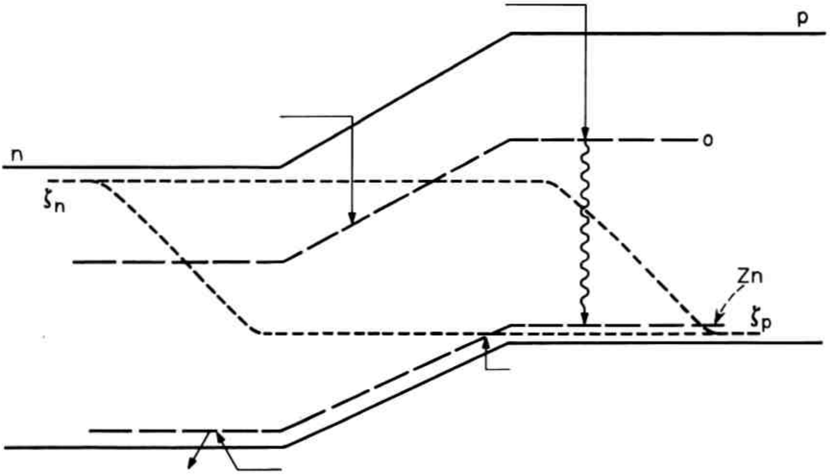


Fig. 2 — The Zn-O pair band mechanism in a forward-biased p-n junction.

layer, the O donors are below the electron quasi-Fermi level from the n-side to deep into the depletion layer, and these donor states can be populated by electrons. However, the Zn acceptors lie above the hole quasi-Fermi level only very close to the p-side. It is only these acceptors that contain trapped holes. Hence, there is no region in the space-charge layer that contains both trapped electrons and trapped holes, and therefore, the Zn-O pair band should not be an efficient recombination mechanism in the depletion layer. Thus, we expect the red Zn-O band to originate predominantly from the p-side, beyond the space-charge layer.

2.2 *The Green Emission*

Among the many types of recombination leading to photoluminescence near the band edge at low temperatures there are (*i*) pair transitions involving a shallow donor and a shallow acceptor (e.g., Te and Zn),[3, 4] and (*ii*) the "A" line and its phonon replicas due to exciton recombination at an N atom substituting isoelectronically for a P atom.[5,6] The green emission at low temperatures observed from junctions prepared from such material can be identified as due to these transitions by simply comparing electroluminescence and photoluminescence spectra.[7] As the temperature of the diodes is increased, the pair band becomes weaker and the "A" line grows at first but then also diminishes in intensity and, at the same time, it broadens and merges with its phonon replicas. Above approximately 200°K, only a broad green emission band remains. It is, therefore, not clear whether the room temperature green band is due to isoelectronic N traps, or to shallow pairs, or to some new mechanism. The possibility of simple band-to-band recombination can be eliminated because the observed efficiency is several orders of magnitude greater than the efficiency (on the n-side, on the p-side, and in the space-charge layer) calculated using the band-to-band rate constant derived from a detailed-balance analysis of the absorption edge.

III. INJECTION — RECOMBINATION KINETICS

3.1 *Dominant Current*

The current-voltage characteristics of Zn-diffused diodes can be explained quantitatively by assuming that there are several current generating mechanisms, each of which dominates in a different range of forward bias.[1] These mechanisms are summarized in Table I. We assume that the current J can always be written as exp qV/nkT, where n will depend upon bias. At the lowest applied bias, surface leakage predominates and the effective n (at room temperature) is about four. In the next bias range the dominant current is due to recombination at deep levels in the space-charge layer. Here $n = 2$, but with increasing bias, preexponential terms (W is the junction width and V_D the built-in potential) cause the effective value of n to decrease. In the next bias range, not observed in all diodes, recombination at a shallow level in the space-charge layer dominates. Although n is nominally equal to unity here, again pre-exponential terms perturb its value somewhat. Here the effective n lies between one and two and slowly decreases toward unity with increasing bias. Thus, in the space-charge regime, n starts at two, and approaches one at high bias. Finally, at the highest biases, simple injection beyond the depletion layer dominates with $n = 1$. (Conductivity modulation which should set in at even higher biases has so far not been observed.)

3.2 *Red Emission*

We have already noted that the red Zn-O emission seems to originate from the p-side of the junction. From near-field spatial distributions on a surface cleaved perpendicular to the junction plane it is evident that the green emission, at least at high biases, is centered at the junction itself, as defined by observations of the junction electro-

Table I — Dependence of Current (J) and of Zn-O Pair Band Emission (L) upon Bias (V) and Their Comparisons

Dominant Current, J $J \alpha \exp \frac{qV}{nkT}$				
Surface Leakage	Space Charge Recombination		Diffusion Current on n-Side	
	Deep Levels	Shallow Levels	Linear Range	Conductivity Modulation
$J\alpha \exp \beta V$	$J\alpha \frac{W}{V_D - V} \exp \frac{qV}{2kT}$		$J\alpha \exp \frac{qV}{kT}$	
$\beta \approx q/4kT$ $n \approx 4$	$n \leqq 2$	$n \to 1$	$n = 1$	$n = 2$

$L\alpha J^4$	$L\alpha J^2$	$L\alpha J^1$	$L\alpha J^{\frac{1}{2}}$

$m = 1$ $L \alpha \exp \frac{qV}{kT}$	$m = 2$ $L \alpha \exp \frac{qV}{2kT}$	$m = \frac{n}{n+1}$	$m = 2$
Linear Range	Diffusion	Diffusion and Drift	Conductivity Modulation
	Saturation		
Light Emission from p-Side, L $L \alpha \exp \frac{qV}{mkT}$			

optic effect.[1] However, the red emission is not centered on the junction but lies on the p-side. At high biases the emission closest to the junction saturates and the emission volume simultaneously expands deeper into the p-side. This observation is inconsistent with n-side or space-charge layer recombination.[1] Thus, the red emission is generated on the p-side beyond the space-charge layer, as expected from Fig. 2. The spatial motion at high biases is due to the saturation of the recombination centers on the p-side.[8] The injected carriers, therefore, must travel beyond the normal diffusion length in order to recombine. We again assume that we can write the red light intensity L in the form exp qV/mkT. At low bias, with simple injection into the p-side and recombination at Zn-O pairs, $m = 1$. However, in the saturation range at high bias, with minority carrier transport limited by diffusion only, $m = 2$.[8]

3.3 *Green Emission*

It is an experimental result that m is always equal to unity for the green emission, independent of the bias.

3.4 *Light versus Current*

In Table I the *J-V* and *L-V* data (for the red band) are combined to show the dependence of light intensity upon current. At low bias, where surface leakage predominates, the light emission varies as $\approx J^4$. In the space-charge regime the relationship is quadratic, but it approaches linearity at high bias. At the highest biases, with saturation on the p-side, and with the current due to injection beyond the space-charge layer (hence, into the n-side), the relationship becomes sublinear. Thus, the quantum efficiency of the Zn-O red band increases rapidly at first, then slowly levels off and finally decreases at the highest biases, thus exhibiting a maximum in the linear range. For the green emission $m = 1$ always. Hence, the quantum efficiency rises rapidly, then slowly levels off and remains constant up to the highest biases measured.

IV. DIODE STRUCTURES

The various types of GaP p-n junctions that have been reported in the literature are summarized in Table II. The circled structures were prepared expressly to exhibit the Zn-O red band. A typical in-diffused diode is made by diffusing Zn into an n-type crystal containing Te and O. A typical out-diffused diode is made by heating a p-type crystal doped with Zn, Te, and O, so that some Zn diffuses out, leaving an n-layer near the surface. Grown junctions may be prepared by floating-zone,[9,10] by vapor phase epitaxy[10,11] or by solution-growth epitaxy.[12,13] For example, in the latter case one can grow (by "tipping") an n-type Te-doped layer from solution onto a p-type seed containing Zn + O.[12] A typical alloyed diode is made by alloying a Sn ball onto a p-type sample containing Zn + O.[14,15,16] Finally, surface structures may be prepared by evaporating a metal on a cold p-type substrate containing Zn, Te, and O.[16]

The diffused structures and the grown junctions are simple p-n

TABLE II — GaP DIODES (CIRCLED) DESIGNED TO EXHIBIT THE ZN-O RED PAIR EMISSION

GaP Junctions								
Diffused			Grown		Alloyed		Surface	
In-Diffused		Out-Diffused	Method	Doping	Substrate	Alloy	Substrate	Film
Substrate	Diffusant	Substrate						
n	Zn	p	Float-zone	Mg-S	p	Sn	p	Au
(Te + O)	Cd	(Zn + Te + O)		Cd-S	(Zn + O)	Ag-Te	(Zn + Te + O)	Sn
		(Cd + Te + O)	Vapor	Mg-S	n	In-Zn		Ag paste
p	Si			Cd-S		Au-Zn		
			Solution	(Zn + O)-Te		Ag-Zn		
p-n		p-n	p-n		p-n Surface-barrier, tunneling		Tunneling	

junctions, where injection is due to thermal activation over the normal junction barrier.[1] In the surface diodes injection arises from tunneling through a thin surface layer.[16] Three injection mechanisms can occur in parallel in the alloyed diodes.[16] In the regions where Sn alloying produces an n-type regrowth layer on the p-type substrate, a simple p-n junction is formed. In regions where no n-type regrowth layer is produced, the metal is in intimate contact with the p-type substrate. This is a surface barrier junction which at forward bias can only extract majority carrier holes. Since it cannot inject minority carriers it results in an excess nonradiative current component. In regions where a thin layer of insulator (perhaps an oxide) separates the metal from the substrate, it is also possible to inject minority carriers by tunneling.

V. RADIATIVE EFFICIENCIES

5.1 *Quantum Efficiency*

Table III summarizes the maximum reported external quantum efficiencies of the red Zn-O band at room temperature in the five classes of diodes described previously. Note that while the highest measured efficiency, 1.5 percent, corresponds to an alloyed diode,[17] the maximum efficiencies observed in the other four classes are all within less than a factor of ten of this value. The table also lists some "average" efficiencies,[12,14] which is the range obtainable with high yields with present technology. The highest quantum efficiency reported for the green emission is 0.015 percent,[18] or 100 times less than the corresponding figure for the red. Since the external quantum efficiencies for spontaneous emission in GaAs diodes at room temperature are usually one to five percent, it is obvious that the red emission from GaP, only slightly less efficient, might be useful in applications where spontane-

TABLE III — EXTERNAL QUANTUM EFFICIENCIES OF Zn-O PAIR BAND IN GaP DIODES AT ROOM TEMPERATURE

		Maximum (%)	Average (%)	Source
In-diffused	Zn/Te + O	0.2		BTL
Out-diffused	Zn + Te + O	0.7		BTL
Solution-grown	Te/Zn + O	0.75	0.3–0.5	IBM
Alloyed	Sn/Zn + O	1.5		Philips
			0.01–0.1	SERL
Surface	Au/Zn + Te + O	0.4		BTL

ous GaAs emitters are considered, as in optoelectronic devices. However, the significant distinction is that the GaP emission lies in the visible range.

5.2 *Luminous Efficiency*

By integrating the product of the emission spectrum of the Zn-O red band and the visual acuity curve, it is found that one watt of Zn-O red light is equivalent to 20 Lumens as far as the eye is concerned. The GaP green emission corresponds to about 650 Lumens/watt. (For comparison, the emission from GaP_xAs_{1-x}, where x corresponds to the maximum P concentration before the band structure becomes indirect, is equivalent to approximately 100 Lumens/watt.) Consider a typical diode, available with current technology, as summarized in Table IV. It may operate at 20 mA with a dc bias of 2 volts emitting red light with an external quantum efficiency of 0.5 percent. (Since the energy of the emitted photon is 1.77 eV, the power efficiency is only slightly less than the quantum efficiency.) With a junction area of 10^{-3} cm^2 the current density is 20 amps/cm^2, which is close to the maximum in quantum efficiency. Table IV also notes the output in normal power units as well as in luminous units. By assuming that the light leaves the diode from only one surface in the active junction area of 10^{-3} cm^2, the predicted brightness is 3600 foot-Lamberts. (Although present measurements are about a factor of ten lower, the discrepancy might be decreased by using large ratios of active junction area to inactive surfaces, or by using special geometries or index-of-refraction-matching glasses to increase the light output from a given region.) The maximum reported efficiency for the green emission is only 0.015 percent but the luminous equivalent of the

TABLE IV — TYPICAL LUMINOUS EFFICIENCY FOR THE Zn-O RED BAND (20 LUMENS/WATT)

Typical diode:	20 mA 2 V 0.5% external quantum efficiency 10^{-3} cm^2 area
Output:	2×10^{-4} watts 4×10^{-3} Lumens 3×10^{-4} candles
Brightness:	4 lamberts 3600 foot-Lamberts 1200 candles/square foot

green band is more than 30 times greater than that for the red band. Thus, the brightness currently available from the *best* green diodes are approximately equal to that available from current *average* red diodes at biases where the red efficiency is a maximum. (At higher biases the green emission of course will increase more rapidly than the red emission.) For comparison, the brightness of the green emission from a ZnS:Cu electroluminescent cell is about 1 foot-Lambert (at 60 Hz, and up to 10 foot-Lamberts at higher excitation frequencies, but with significant deterioration during aging). Thus, with present technology, the red emission from GaP diodes corresponds to much higher brightness values than for standard ZnS EL cells, and this occurs in the bias range corresponding to maximum efficiencies — 0.3 to 0.5 percent quantum efficiency. Similarly, the brightness of the GaP green emission is also higher than that available from ZnS panels.

5.3 *Efficiency Outlook*

Since the quantum efficiency of the Zn-O red band is as high as 11 percent in photoluminescence of p-type samples, it might be possible to obtain similar electroluminescence efficiencies from p-side injection in junctions. At low to moderate biases the dominant competing mechanism is due to space-charge recombination at deep levels. Thus, a reduction of this current component could increase the red emission efficiency. Since the room temperature green emission mechanism has not been established, similar predictions for the green band cannot be made. Finally, it is noted that a number of other deep pair combinations exhibit donor-acceptor pair recombination in the orange and red in photoluminescence at room temperature with efficiencies of several percent.[19,20] These may eventually provide useful recombination centers in GaP diodes.

VI. SUMMARY

Currently available p-n junctions in GaP emit in the red with an external quantum efficiency (roughly equal to a power efficiency) which exhibits a maximum (with bias) of 0.1 to 0.5 percent. The brightness at this maximum is within a factor of 10 of 3600 foot-Lamberts, far greater than that from a normal ZnS EL cell. The *best* green emitting diodes available correspond to a similar brightness value, but the efficiency does not drop with increasing bias. Such diodes should possess the normal advantages of semiconductor devices: low dc operating bias, small size, probably cheap to manufacture and hope-fully little deterioration with aging. Special diode geometries or the use of index-of-refraction-matching glasses might be used to increase the external quantum efficiency (although the red band falls in a region of low internal absorption, the green band, near the band edge, does not) or to focus the emitted light, thereby increasing the apparent brightness.

REFERENCES

1. Gershenzon, M., Logan, R. A., and Nelson, D. F., Phys. Rev., *149*, 1966, p. 580.
2. Gershenzon, M. and Mikulyak, R. M., Appl. Phys. Letters, *8*, 1966, p. 245.
3. Thomas, D. G., Gershenzon, M., and Trumbore, F. A., Phys. Rev., *133*, 1964, p. A269.
4. Trumbore, F. A. and Thomas, D. G., Phys. Rev., *137*, 1965, p. A1030.
5. Thomas, D. G., Gershenzon, M., and Hopfield, J. J., Phys. Rev., *131*, 1963, p. 2397.
6. Thomas, D. G., Hopfield, J. J., and Frosch, C. J., Phys. Rev. Letters, *15*, 1965, p. 857.
7. Gershenzon, M., Mikulyak, R. M., Logan, R. A., and Foy, P. W., Sol. St. Elect., *7*, 1964, p. 113.
8. Nelson, D. F., Phys. Rev., *149*, 1966, p. 574.
9. Frosch, C. J. and Derick, L., J. Electrochem. Soc., *108*, 1961, p. 251.
10. Gershenzon, M. and Mikulyak, R. M., Sol. St. Elect., *5*, 1962, p. 313.
11. Frosch, C. J., J. Electrochem. Soc., *111*, 1964, p. 180.
12. Lorenz, M. R. and Pilkuhn, M., J. Appl. Phys., in press.
13. Trumbore, F. A. and Logan, R. A., to be published.
14. Allen, J. W., Moncaster, M. E., and Starkiewicz, J., Sol. St. Elect., *6*, 1963, p. 95.
15. Grimmeiss, H. G. and Koelmans, H., Phil. Res. Rpts., *15*, 1960, p. 290; Phys. Rev., *123*, 1961, p. 1939; Grimmeiss, H. G., Rabenau, A., and Koelmans, H., J. Appl. Phys., *32*, 1961, p. 2123.
16. Logan, R. A., Gershenzon, M., Trumbore, F. A., and White, H. G., Appl. Phys. Letters, *6*, 1965, p. 113.
17. Grimmeiss, H. G. and Scholz, H., Phys. Letters, *8*, 1964, p. 233.
18. Pilkuhn, M. H. and Foster, L. M., IBM J., *10*, 1966, p. 122.
19. Gershenzon, M., Trumbore, F. A., Mikulyak, R. M., and Kowalchik, M., J. Appl. Phys., *37*, 1966, p. 486.
20. Gershenzon, M., Trumbore, F. A., Mikulyak, R. M., and Kowalchik, M., to be published.

Schottky Barrier Photodiodes with Antireflection Coating

By M. V. SCHNEIDER

(Manuscript received July 19, 1966)

Schottky barrier diodes can be used for fast and efficient photodetectors if the incident light is coupled into the depletion layer of the diode and if electron-hole pairs are created by the internal photoelectric effect in the depletion layer. Fast response of the diode is achieved by designing a Schottky barrier with a small RC product. High quantum efficiency is obtained by coupling the light through a thin metal layer into the depletion region of the diode and by using an antireflection coating on the metal layer for matching the incident light beam.

Schottky barrier photodiodes have been made with thin semitransparent gold layers on n-type epitaxial silicon and with zinc sulfide as an antireflection coating. A net quantum efficiency of 70 percent has been achieved at the He-Ne laser wavelength of 6328 Å. The pulse response of packaged diodes with 0.5-nanosecond wide pulses shows a symmetrical pulse shape with only small distortion due to carrier diffusion and reactance in the completed package.

The diode structure is suitable for detector arrays. It is also useful for optical time domain reflectometry. The technique of coupling light through metal layers can be extended to other optical devices which require efficient transfer of radiation into a semiconductor through conducting electrodes.

I. DEFINITION OF THE SCHOTTKY BARRIER PHOTODIODE

A Schottky barrier is a rectifying metal-to-semiconductor contact with certain properties which have been originally described by Schottky.[1,2] The main feature of the Schottky barrier is that it has the properties of an ideal step junction and that only majority carriers are involved in the rectification process. Schottky barriers have been used for various devices in the microwave region. A few examples are the Au-n-type GaAs and the Au-n-type Si varactors described by Kahng and D'Asaro[3,4] and the honeycomb millimeter diode by Irvin and Young.[5] The Schottky barrier has not been used to any great extent for optical devices because of the difficult problem of coupling optical radiation through the metal contact into the semiconductor. The purpose of this paper is to present a solution to this problem and to describe the properties of a completed diode which will be defined as a Schottky barrier photodiode.

The Schottky barrier photodiode is a rectifying metal-to-semiconductor contact in which electron-hole pairs are created in the semiconductor by the internal photoelectric effect under incident illumination. The separation of the pairs is accomplished by the built-in electric field in the barrier or by an externally applied field across the barrier. The separation of the carriers leads to a photocurrent in the external circuit which may be amplified and detected. Internal amplification by avalanche multiplication cannot be achieved in a Schottky barrier photodiode because of nonuniform field intensities at the boundary of the metal-semiconductor interface.

II. STRUCTURE OF OPTICAL JUNCTION DETECTORS

Photodetectors with a high frequency response consist usually of a semiconductor p-n junction or a semiconductor p-i-n structure. A schematic drawing of such a detector is shown in Fig. 1. The incident radiation is absorbed in the intrinsic layer which is sandwiched between a p and an n-layer. Electron-hole pairs are created by the internal photoelectric effect and are separated by an applied electric field across the junction. Metal contacts are required on both sides of the structure in order to apply the electric field and to collect the carriers. The contacts on top of the p-layer shown in Fig. 1 are made in the form of stripes in order to transmit the incident radiation between adjacent stripes into the p-i-n region.

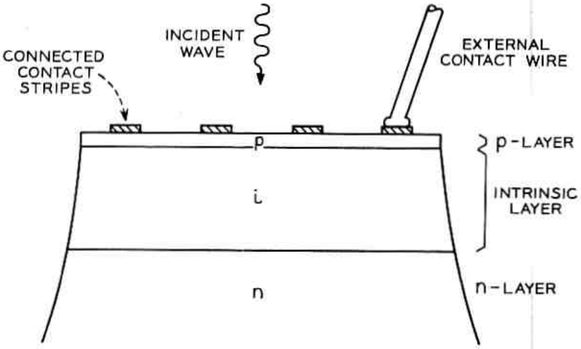


Fig. 1 — p-i-n photodiode with contact stripes on *p*-layer.

High quantum efficiency and fast response are achieved by proper choice of the semiconductor material and the physical dimensions of the layers including the contact stripes. Design criteria have been discussed by Anderson,[6] Lucovsky and Emmons,[7] and Riesz.[8] Internal multiplication with uniform and microplasma-free junctions with a guard ring has been achieved by Anderson, McMullin, D'Asaro and Goetzberger[9] and by Melchior and Lynch.[10]

A different approach is necessary for the case of a Schottky barrier photodiode shown in Fig. 2. A semitransparent metal is deposited on the surface of a semiconductor in order to create a surface barrier. The light is matched into the barrier by an antireflection coating which is deposited on the semitransparent metal. A thick metal dot or a metal ring with a contact wire serves as an external contact for applying the dc back bias and for collecting carriers. The semiconductor material and the applied back bias are chosen in such a way that most of the carriers are created within the depletion layer. The net quantum efficiency of the device is determined mainly by the transmission loss in the metal film and by the quality of the antireflection coating. The response time is determined by the transit time of the carriers through the depletion layer and the RC product of the diode. Design criteria for achieving a small RC product will be discussed later in this paper.

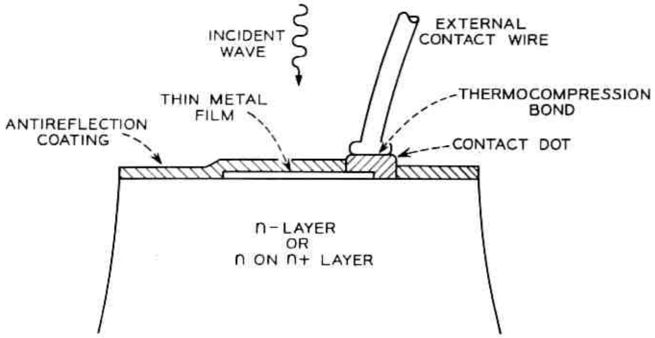


Fig. 2 — Schottky barrier photodiode with antireflection coating on metal film.

Coupling of the incident light beam into the Schottky barrier can also be achieved by other means. Fig. 3 is a cross-sectional view of the Sharpless photodiode.[11,12] A point contact is formed on epitaxial material and the light is focused into the Schottky barrier through an etched dome in the semiconductor. An antireflection coating is not necessarily required because the semiconductor surface does not present a serious optical mismatch to the incident wave.

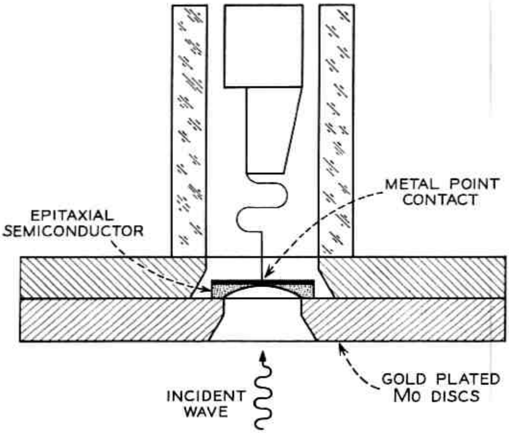


Fig. 3 — Point contact Schottky barrier photodiode with etched dome in epitaxial semiconductor.

Another way to build a Schottky barrier photodiode is shown in Fig. 4. A thick metal coating is applied to the semiconducting material. An array of slots or holes are etched into the metal. The holes are close to resonance at the wavelength of the incident radiation, e.g., they are approximately a quarter wavelength wide and are spaced approximately a half wavelength apart. The thickness of the metal has to be much smaller than a wavelength because the excited mode in the hole or the slot is under cutoff. The remaining reactive part of the surface impedance

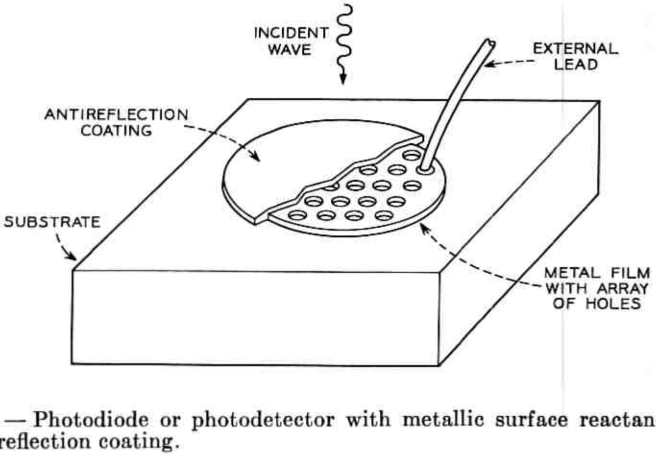


Fig. 4 — Photodiode or photodetector with metallic surface reactance sheet and antireflection coating.

of this structure is compensated by a suitable antireflection coating. This coating is only required for improving the final match of the device because a reactance sheet can be designed with a high return loss without any further matching elements.

Photodiodes of the type shown in Fig. 2 have been made on n-type epitaxial silicon for maximum response at the 6328 Å line of a He-Ne gas laser. Gold has been used for formation of the Schottky barrier and zinc sulfide for the antireflection coating. The results are discussed in the following sections of this paper. Various technological improvements in the technique of fabricating microarrays will have to be achieved in order to fabricate the photodiode shown in Fig. 4.

III. TRANSMISSION OF LIGHT THROUGH METAL FILMS

Metal films are characterized by high reflectivity and low transmission in the visible range of the spectrum. The transmission through the film can be increased by reducing its thickness. The reflection can be decreased by a dielectric film acting as a quarter wave transformer on top of the metal film. These two simple steps make it possible to transmit light into an optical device which requires metal electrodes.

Optical constants of thin metal films are listed by Schopper,[13] Heavens,[14] and Mayer.[15] The physical theory and measurements are described by Parker Givens[16] and by Abelès.[17] A marked dependence of the optical constants on film thickness is observed. Other parameters of importance include the technique used in the deposition process, the substrate temperature, deposition rate and surface properties of the substrate. A typical example of steps taken in substrate preparation, purity of materials and pressures observed in the vacuum chamber is described by Bennett and Ashley[18] and a review on nucleation and film growth as a function of various parameters is given in a paper by Behrndt.[19] What complicates matters for device applications is the fact that the films may not be continuous and that the index of refraction depends on the angle of incidence as described by Hall.[20]

These difficulties do not prevent the fabrication of an antireflecting metal-semiconductor surface. Fairly consistent results can be achieved with gold evaporated from tungsten or molybdenum boats under high vacuum or ultra-high vacuum conditions. The optical surface impedance of the structure can be measured and from this one can determine a unique dielectric constant and a unique thickness which will allow optical matching at a specified wavelength. The steps in this procedure are similar to matching microwave networks by using the Smith Chart.

The reflectance and the transmittance which one can expect from a thin gold film at λ = 6328 Å are shown in Fig. 5. Reflectance, transmittance and loss are plotted as a function of film thickness for an unsupported Au film with an index of refraction $N = n - j \cdot k = 0.30 - j \cdot 3.0$. The only assumptions used in this plot are that one deals with normal incidence and that this particular index of refraction is independent of the film thickness. The index $N = 0.30 - j \cdot 3.0$ is an approximate value for bulk gold obtained from measurements described by Parker Givens.[16] Other data for gold deposited under various conditions and listed by Schopper[13] cover an approximate range of $n = 0.30 \pm 0.10$ and $k = 3.0 \pm 1.0$ at wavelengths in the range from 6000 Å to 6600 Å.

The exact thickness of a thin metal film is usually of secondary importance for many device applications. What one needs to know for devices described in this paper is reflectance, transmittance, and loss for a specified surface resistance (sheet resistance) of the film. The surface resistance of the film limits the frequency response of the device because it will contribute to the resistive part of the device. The relationship between the surface resistance and the RC product will be discussed later.

Fig. 6 is a plot of reflectance, transmittance and loss measured for Au films at λ = 6328 A for surface resistances in the 3 to 6 ohm/square range. The Au films are deposited on fused quartz slides under the following conditions:

(*i*) The substrates are cleaned ultrasonically in successive baths

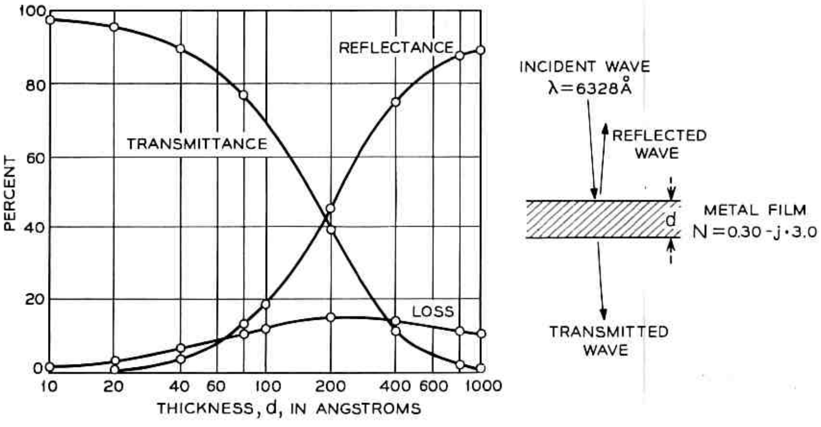


Fig. 5 — Transmittance, reflectance, and loss for thin metal film with $N = 0.30 - j \cdot 3.0$.

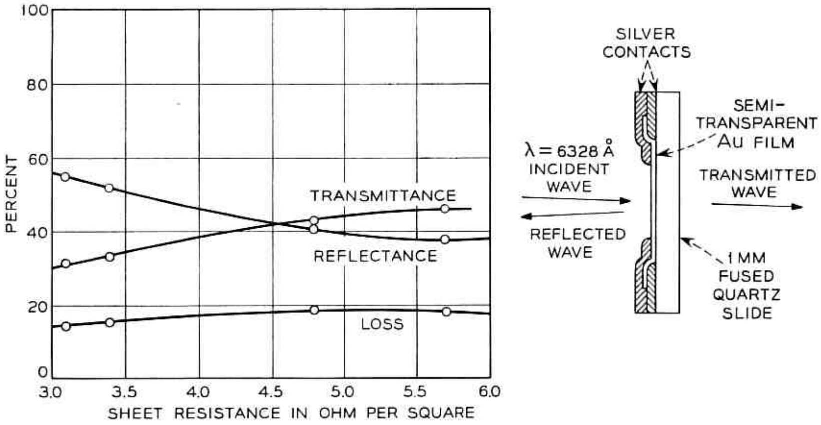


Fig. 6 — Transmittance, reflectance, and loss of evaporated gold films on fused quartz substrates.

of a detergent, distilled water and alcohol. They are dried with tank nitrogen and vapor degreased in isopropyl alcohol in the apparatus described by Holland.[21]

(*ii*) The substrate is transferred into a VE-400 (Vacuum Electronics, Inc.) vacuum system which is pumped down to a pressure of $2 \cdot 10^{-7}$ torr. Due to the location of the ionization gauge, which is between the diffusion pump and the liquid nitrogen cold trap, the pressure in the bell jar is an order of magnitude higher.

(*iii*) Gold is evaporated from a tungsten coil located 6 inches from the substrate with estimated deposition rates in the range of 5 to 10 Å/sec. The quartz substrate is not heated.

(*iv*) The dc resistance of the film is continuously monitored during evaporation with two silver contacts shown in Fig. 6. Additional silver contacts are applied immediately after the gold evaporation in order to sandwich the gold layer between two layers of silver. This method insures minimum contact resistance between the silver and the gold. All three layers (Ag, Au, Ag) are applied consecutively without opening the vacuum system.

The thickness of the films is measured with a multiple beam interferometer; e.g., the film with a 5.7 ohm/square sheet resistance has a thickness of 180 ± 15 Å. This particular sheet resistance is approximately 4.5 times higher than that which one would obtain from the resistivity ρ of bulk gold for a thickness of 180 Å ($\rho = 2.44 \times 10^{-6}$ ohm cm at room temperature). One of the reasons for this discrepancy is the fact that conduction in thin films depends upon the scattering from the film boundaries. This means that bulk resistivities cannot be achieved for thin films. Another effect of importance is that the film may be discontinuous; that means the film consists of a number of islands with partial bridging between adjacent islands as described by Chopra[22] and Francombe and Sato.[23] The optical properties of such a film can be characterized by a complex index of refraction provided that the average distance between neighbouring islands is a small fraction of the optical wavelength.

The reflectance and transmittance curves shown in Fig. 5 are calculated for a metal film which is not supported by any substrate. Fig. 7 is a similar plot for a film supported by a substrate with an index of refraction $N = 3.75$. This particular index corresponds approximately to silicon with $N = 3.75 - j \cdot 0.18$ at $\lambda = 6328$ Å. Comparison with Fig. 5 shows that the reflectance for a specified film thickness is higher. The loss in the metal film is lower because of the increased reflectance. Reflectance, transmittance, and loss are the same for very thick films as shown by the calculated points for $d = 1000$ Å.

IV. OPTICAL MATCHING OF A METAL-SEMICONDUCTOR CONTACT

A metal-semiconductor contact can be optically matched at a specified wavelength if the reflection coefficient or the surface impedance is known for that particular wavelength.

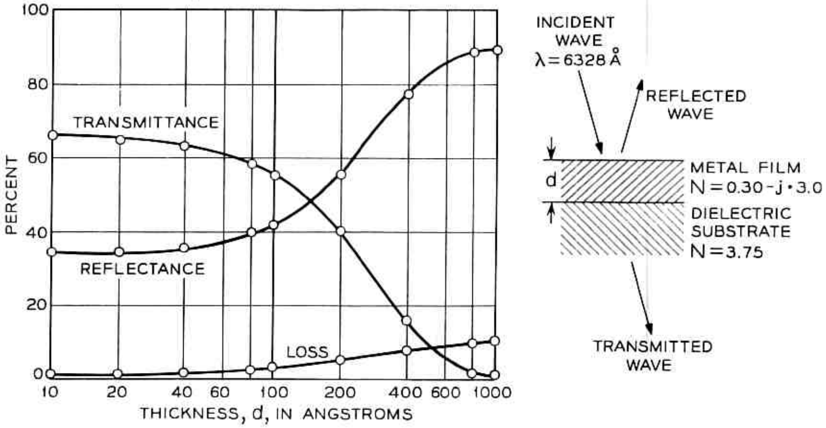


Fig. 7 — Transmittance, reflectance, and loss of metal film on dielectric substrate.

The reflection coefficient r_1 at the interface of two media in Fig. 8 is given by

$$r_1 = \frac{g_1 - g_0}{g_1 + g_0}. \tag{1}$$

The quantities g_k ($k = 0, 1$) are generalized impedances or admittances (immittances) of the two media and μ_k ($k = 0, 1$) is the permeability of the medium. The immittance is directly related to the index of refraction of that particular medium. The transmission coefficient t_1, is

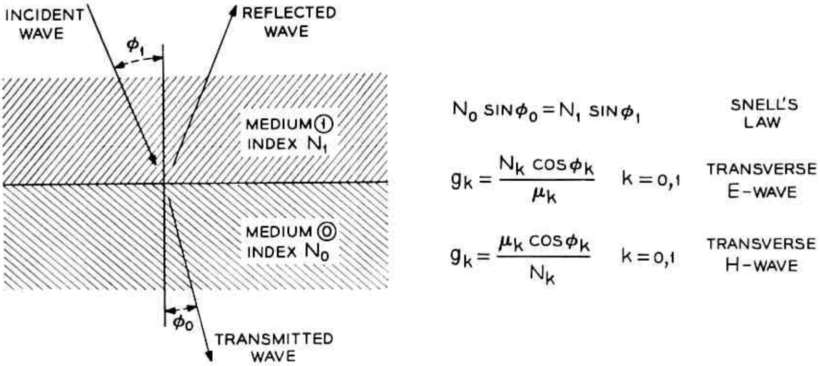


Fig. 8 — Plane wave reflected and transmitted from plain boundary.

$$t_1 = \frac{2g_1}{g_1 + g_0}. \tag{2}$$

Equations (1) and (2) are exactly identical with the equations used for computing the voltage or current reflection coefficient and the transmission coefficient for two adjacent RF transmission lines at different impedance levels.

A sequence of plane parallel films can be treated by applying (1) and (2) with a recursion formula which takes into account the phase shift between two adjacent media. The exact procedure is derived by Wolter.[24] The result with a sequence of three media for the reflection coefficient r_2 and the transmission coefficient t_2 is

$$r_2 = \frac{(g_2 - g_1)(g_1 + g_0)\exp(\rho_1 d_1) + (g_2 + g_1)(g_1 - g_0)\exp(-\rho_1 d_1)}{(g_2 + g_1)(g_1 + g_0)\exp(\rho_1 d_1) + (g_2 - g_1)(g_1 - g_0)\exp(-\rho_1 d_1)} \tag{3}$$

$$t_2 = \frac{4g_1g_2}{(g_2 + g_1)(g_1 + g_0) \exp(\rho_1 d_1) + (g_2 - g_1)(g_1 - g_0) \exp(-\rho_1 d_1)} \tag{4}$$

with the notation shown in Fig. 9. The quantities g_k are again the immitances of the media. The exponential term $\exp(\pm\rho_1 d_1)$ represents the phase shift between the two adjacent boundaries shown in Fig. 9.

Equations (3) and (4) are valid if the impedance or admittance of the center medium is complex; e.g., if it is a metal. A plane wave launched in medium 2 will excite a hybrid wave in medium 1; that means planes of equal phase and of equal amplitude will not coincide unless one deals with normal incidence. A wave with parallel planes for equal phase and equal amplitude can be propagated in an absorbing medium. Such a wave, however, cannot be excited by a plane wave coupled into the absorbing medium through a plane boundary at an oblique angle. This property leads to an index of refraction which is a function of the angle of incidence. Further details may be found in the original work by Fry.[25,26]

The reflection coefficient r_2 in (3) should be made as small as possible for building devices with a high transmission into the substrate. This cannot be achieved for a metal-semiconductor structure. It is possible, however, to deposit an additional film on the metal and to compensate the complex reflection coefficient by proper choice of the index of refraction and the thickness of this antireflection coating. The surface impedance on top of medium 1 in Fig. 9 which represents the metal is

$$G_0 = U - j \cdot V = \frac{1 - r_2}{1 + r_2} \tag{5}$$

with U being the real part and jV the complex part of the surface impedance. G_0 can be matched to an impedance G_2 by a dielectric layer with a proper thickness D and a proper impedance G_1 if

$$G_1^2 = G_2 \left\{ \frac{V^2}{U - G_2} + U \right\} \tag{6}$$

$$D = \frac{\lambda}{2\pi n} \arctan \left(\frac{G_1}{G_2} \cdot \frac{U - G_2}{V} \right). \tag{7}$$

The notation is shown in Fig. 10. The quantity n is the index of refraction of the dielectric material. For an interface with a real surface impedance $G_0 = U$, one obtains with $V = 0$ from (6) and (7)

$$G_1 = \sqrt{G_0 G_2} \tag{8}$$

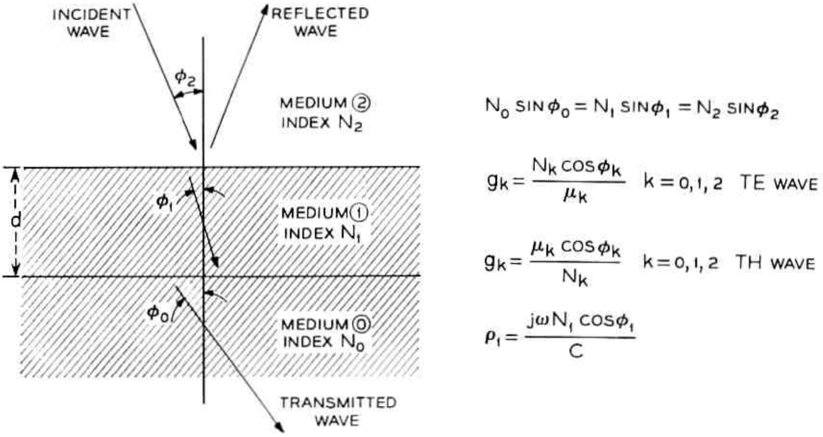

INCIDENT WAVE
REFLECTED WAVE
ϕ_2
MEDIUM ②
INDEX N_2
d
ϕ_1
MEDIUM ①
INDEX N_1
ϕ_0
MEDIUM ⓪
INDEX N_0
TRANSMITTED WAVE
$N_0 \sin\phi_0 = N_1 \sin\phi_1 = N_2 \sin\phi_2$
$g_k = \frac{N_k \cos\phi_k}{\mu_k}$ $k = 0, 1, 2$ TE WAVE
$g_k = \frac{\mu_k \cos\phi_k}{N_k}$ $k = 0, 1, 2$ TH WAVE
$\rho_1 = \frac{j\omega N_1 \cos\phi_1}{C}$

Fig. 9 — Reflection and transmission for 3 media.

$$D = \frac{1}{n} \cdot \frac{\lambda}{4}. \tag{9}$$

This is the well-known relationship for a quarter-wave transformer connecting microwave transmission lines at different impedance levels. The wavelength λ is the vacuum wavelength.

A practical example is treated in Fig. 11. Reflection coefficients for a Au-Si structure are plotted in the complex plane for a gold layer with a thickness of 100 Å and 200 Å. The example refers to normal incidence at a wavelength of $\lambda = 6328$ Å. It is assumed that the index of refraction is $N_1 = 0.28 - j \cdot 3.01$ for gold and $N_0 = 3.72 - j \cdot 0.18$ for silicon. The index of refraction and the thickness of the antireflection coating are listed in Table I.

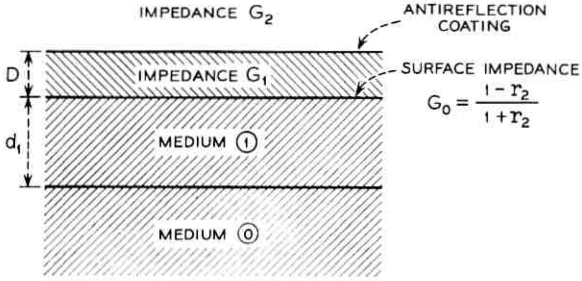


Fig. 10 — Impedance matching with antireflection coating on medium No. 1.

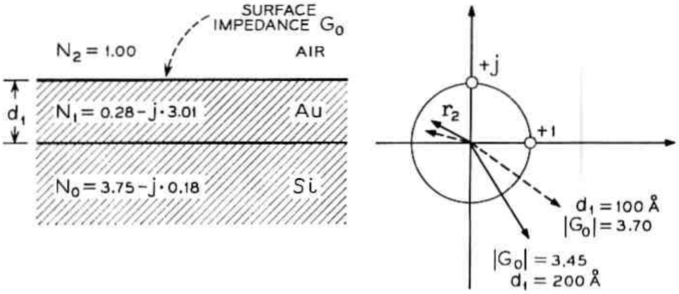


Fig. 11 — Surface impedance of gold-silicon Schottky barrier at $\lambda = 6328$ A.

The loss in the gold film and the reflectance and the transmittance of the complete structure is shown in Fig. 12 and Fig. 13. All three quantities are plotted as a function of the thickness of the gold film for two *fixed* antireflection coatings with $n = 2.30$, $D = 500$ Å and $n = 3.30$, $D = 240$ Å. Minimum reflectance is achieved as predicted in Table I. All curves are obtained by applying (3) and (4) with a recursion formula for one additional layer.

The conclusion from the results of Figs. 12 and 13 is that transmission with low loss into the silicon substrate is feasible.

The reflectance achieved for three evaporation processes with zinc sulfide deposited on a Au-Si surface barrier is shown in Fig. 14. Gold layers with sheet resistances in the range of 6 ohm/square are first evaporated on epitaxial silicon. Zinc sulfide is evaporated on the gold layer. The return loss at $\lambda = 6328$ Å is continuously measured with an optical reflectometer and a He-Ne laser as a signal source. The reflectometer is similar to the one described by Perry.[27] The measured return loss is calibrated in dB. The evaporation is continued after reaching the first minimum in one case in order to show the periodicity of the process. One concludes that an improvement of 8 dB to 9 dB in return loss is possible with a single layer of zinc sulfide. The return loss without the matching layer is 3 dB. The total return loss is therefore, 11 dB to 12 dB.

Table I

Thickness of Au Film	Index *n* of Coating	Thickness of Coating in Å	Thickness of Coating in Terms of Phase Angle
100 A	2.28	510 A	66°
200 Å	3.33	242 Å	46°

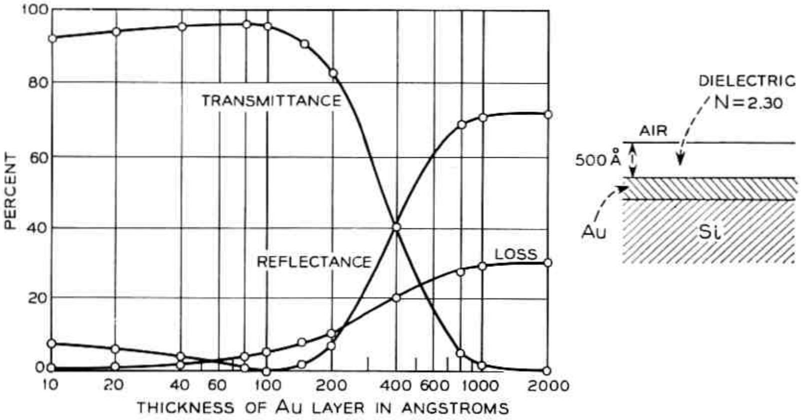


Fig. 12 — Reflectance, transmittance, and loss from Au-Si surface barrier with 500 Å thick ZnS antireflection coating.

It is difficult to measure the transmittance or the loss in the metal for a ZnS-Au-Si structure. Some indication may be obtained from the measurement of the net quantum efficiency of the device if all the carriers can be collected and if there is no internal multiplication. Another direct method is to use the fact that the losses in the metal will increase its temperature and change its resistance. The resistance changes could be simulated by obtaining the same increase with a dc current flowing in

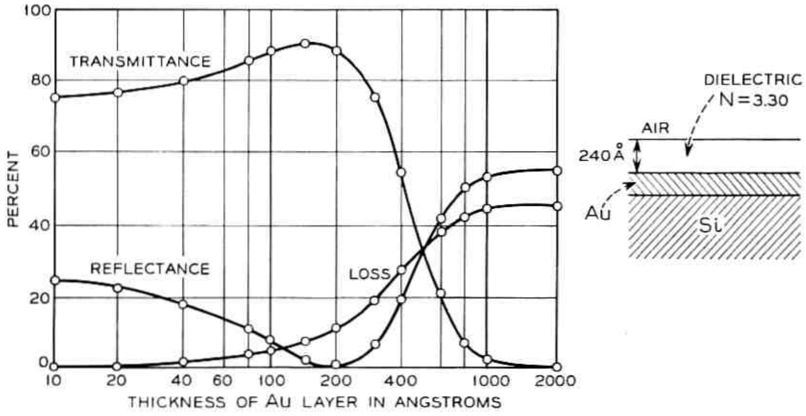


Fig. 13 — Reflectance, transmittance, and loss from Au-Si surface barrier with 240 Å thick ZnS antireflection coating.

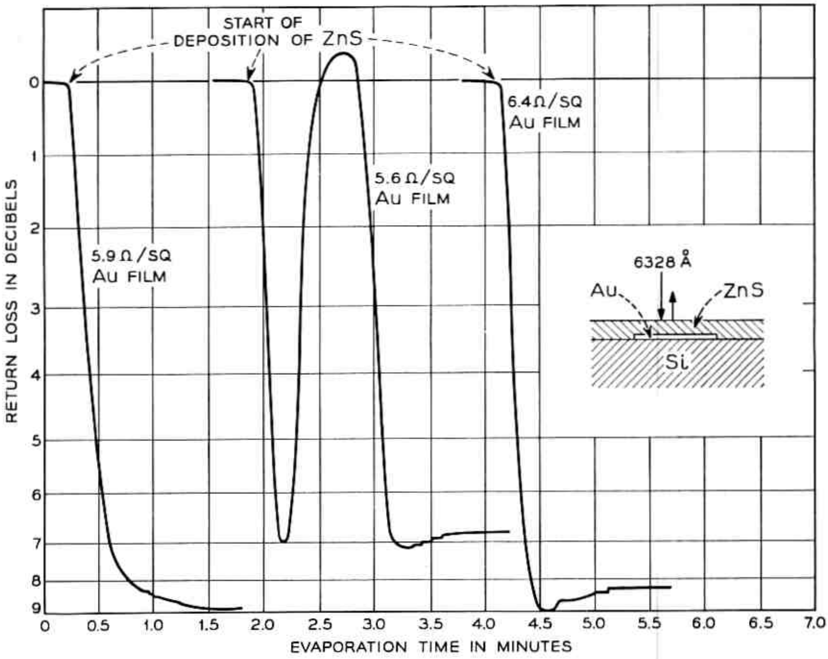


Fig. 14 — Return loss from Si-Au surface barrier during deposition of antireflection ZnS layer.

the Au film. The same procedure is used for power detector calibrations in the microwave frequency range.

The transmittance through Au-ZnS has been measured for a slightly modified case shown in Fig. 15 using a 1-mm thick quartz slide as a substrate. A 5-ohm/square Au film is evaporated on the quartz. The reflectance is 40 percent and the transmittance is 43 percent at 6328 Å as shown in Fig. 6. Zinc sulfide is then evaporated on the Au. The return loss and the transmission are continuously measured with a double reflectometer at $\lambda = 6328$ Å mounted inside the vacuum system. The double reflectometer records transmitted and reflected power simultaneously as shown by the coincidence of maxima and minima on the time scale in Fig. 15. The calibration in dB is obtained with a set of standard optical attenuators. The results from this experiment are

(*i*) The transmittance is improved by 2.5 dB. This means that 76 percent of the incident light is transmitted.

(*ii*) The reflectance is decreased by 10 dB which means that the reflectance is reduced from 40 to 4 percent.

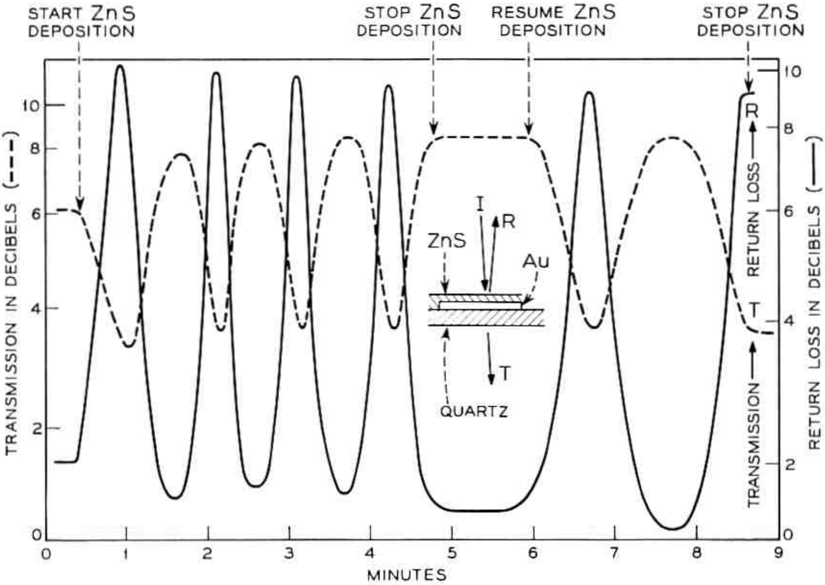


Fig. 15 — Return loss and transmission of gold film on quartz substrate during deposition of ZnS antireflection coating.

(*iii*) The process is periodic which means the losses in the ZnS are small.

(*iv*) The evaporation process can be interrupted at any time and resumed later without changing the periodicity of the process and the levels of the minima and the maxima.

V. DESIGN OF THE SCHOTTKY BARRIER PHOTODIODE

5.1 *Optical Absorption and Carrier Generation in the Schottky Barrier*

The absorption coefficient α of the semiconductor and the width of the depletion layer w are important parameters for designing a Schottky barrier photodiode. The reason for this is as follows. The photocurrent through the depletion layer consists of two contributions. One is due to the carriers created within the layer, the other is due to carriers generated in the adjacent bulk material which diffuse later into the junction. Minority carriers which enter the junction by diffusion will be swept across the junction by the applied external field. This diffusion current may lead to delay distortion if the incident wave is pulsed or rf modulated. The diffusion current is small if most of the optical power is absorbed within the depletion layer. This requires

$$w \gg \frac{1}{\alpha}. \tag{10}$$

The upper limit for w is determined by the transit time which can be tolerated for the carriers.

The absorption coefficient α of Ge and Si as a function of wavelength is given in Fig. 16. The width of the depletion layer in a uniformly doped material with a carrier concentration N under an applied external voltage V is given by Kahng[3,28] as

$$w = \sqrt{\frac{2\varepsilon(V_D + V)}{qN}} \tag{11}$$

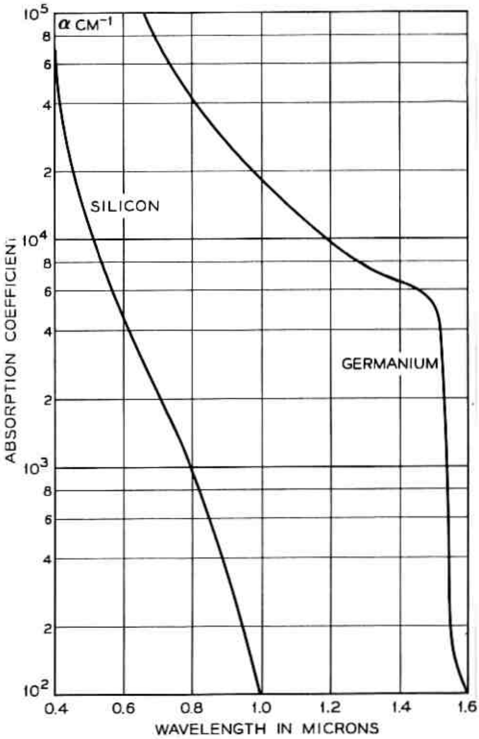


Fig. 16 — Optical absorbtion coefficient of silicon and germanium at 300° K.

where V_D is the diffusion potential, ε the dielectric constant and q the electron charge. The diffusion potential is the potential difference of the conduction band level between its value at the surface and its value inside the bulk material. Diffusion potentials of various metal-semiconductor combinations can be obtained from data supplied by Cowley and Sze.[29] The diffusion potential in a Schottky barrier photodiode is usually much smaller than the applied back bias V because of the requirement $w \gg 1/\alpha$. With $q = 1.60 \times 10^{-19}$ coulomb and $\varepsilon = 8.85 \times 10^{-14}\, \varepsilon_r$ farad/cm one obtains

$$w = 1.05 \times 10^7 \sqrt{\frac{\varepsilon_r (V_D + V)}{N}} \text{ micron.} \tag{12}$$

N is the doping level in carriers/cc and w is measured in μm or micron (1 μm $= 10^{-4}$ cm). The ratio of capacitance C to junction area A is

$$\frac{C}{A} = \frac{\varepsilon}{w} = \sqrt{\frac{\varepsilon q N}{2(V_D + V)}}\,. \tag{13}$$

The width of the depletion layer for Ge and Si as a function of the total voltage $V_D + V$ is shown in Figs. 17 and 18. The breakdown limit

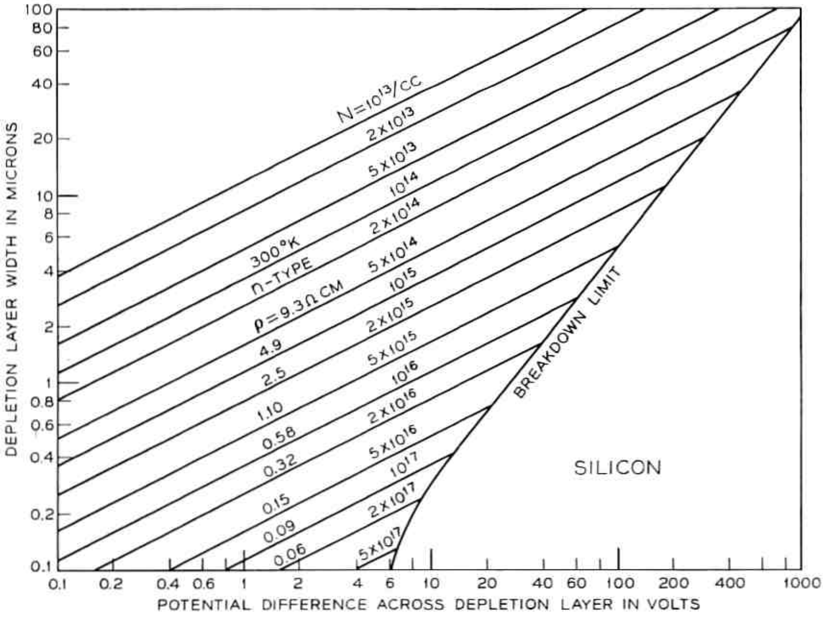


Fig. 17 — Depletion layer width in n-type silicon as a function of potential difference $V + V_D$ across depletion layer.

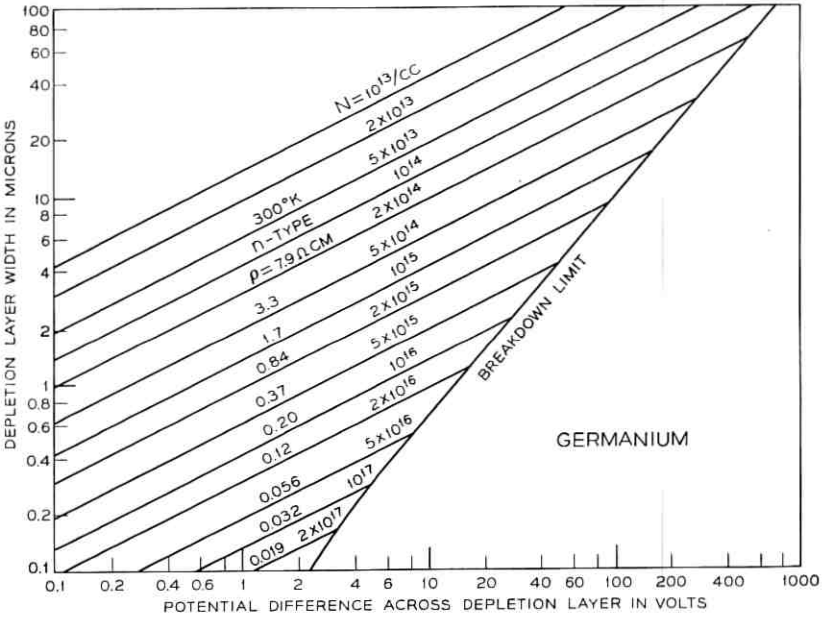


Fig. 18 — Depletion layer width in n-type germanium as a function of potential difference $V + V_D$ across depletion layer.

refers to an abrupt junction in the bulk. It is desirable that the depletion width satisfy (10). Moreover, for any particular application, the thickness w should be no thinner than is necessary to achieve a cutoff frequency which is twice the maximum operating frequency since the maximum available power from a photodiode depends inversely upon the square of the diode capacitance. It is not always possible to satisfy these requirements because of transit time considerations or because of material properties. The drift current for a specified depletion layer width w is

$$J_{\text{Drift}} = q\varphi(1 - e^{-\alpha w}) \tag{13}$$

where φ is the incident photon flux at the front of the depletion layer and q the electron change. One obtains e.g., for $w = 2/\alpha$ a value of 0.86 for the reduction factor $1 - \exp(-\alpha w)$. The total current will be larger because 14 percent of the radiation will be absorbed beyond the depletion layer in the bulk material and create diffusion current. The exact amount of diffusion current under static condition may be found in a paper by Gärtner.[30]

5.2 *Frequency Limitations of the Photodiode*

The frequency limitations for Schottky barrier photodiodes are determined by the transit time of the carriers through the depletion layer, the sheet resistance of the metal film, the resistance of the bulk material and the capacitance of the junction.

The transit time for a carrier depends on the type of carrier and the location of its origin within the depletion layer. The carriers may reach saturation velocities for sufficient high field intensities; e.g., 10^7 cm/sec for electrons in silicon. The holes will move at lower velocities. One has to remember, however, that holes are created predominantly in the high field region in the vicinity of the metal and will travel only a short distance to the metal electrode. Electrons will have to travel over a much longer distance and through a region of low field intensities as shown in Fig. 19. The electron transit time τ_{el} will thus be the predominant factor. This transit time has been calculated by B. C. DeLoach[31] by assuming that

(*i*) electrons reach the saturation velocity v_s at the maximum field in the junction, and

(*ii*) the transit of the carrier through the junction is completed when it reaches the field $E_o = kT$, that means when it joins the free carriers with an average energy kT (0.026 eV at 300°K) to the right of the swept space charge.

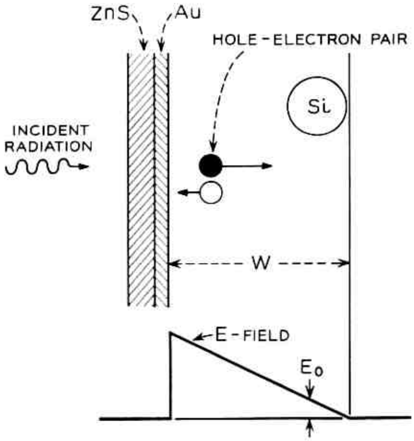


Fig. 19 — Pair creation under incident illumination with electric field intensity E in depletion layer.

The result for the electron transit time τ_{el} is

$$\tau_{el} = \frac{w}{v_s}. \tag{14}$$

Typical values which may be achieved in a silicon surface barrier are $w = 5$ microns and $v_s = 10^7$ cm/sec. This leads to an electron transit time $\tau_{el} = 5.10^{-11}$ sec. The corresponding cutoff frequency is $f_c = 1/\tau_{el} =$ 20 GHz.

The frequency response of a photodiode with a capacitance C per unit area and a sheet resistance R_s has been calculated by Lucovsky and Emmons.[32] The cutoff frequency depends on the geometry of the diode and in particular on the location of the ohmic contact. Three types of contacts shown in Fig. 20 have been discussed by the authors. The 3-dB cutoff frequencies for the short circuited diodes are

$$\omega_c = \frac{3}{R_s C b^2} \qquad \text{linear contact} \tag{15}$$

$$\omega_c = \frac{8}{R_s C a^2} \qquad \text{ring contact} \tag{16}$$

$$\omega_c = \frac{4}{R_s C \cdot F(a,b)} \qquad \text{dot contact} \tag{17}$$

with

$$F(a,b) = \frac{a^2}{2} - \frac{3b^2}{2} + \frac{2b^4}{b^2 - a^2} \log \frac{b}{a}. \tag{18}$$

The dimensions a,b are defined in Fig. 20. The formulas have been derived for a p-n photodiode with the p-layer as the conducting layer. They remain fully valid if the p-layer is replaced by a thin metal film with a sheet resistance of R_s ohm/square. The cutoff frequency for the

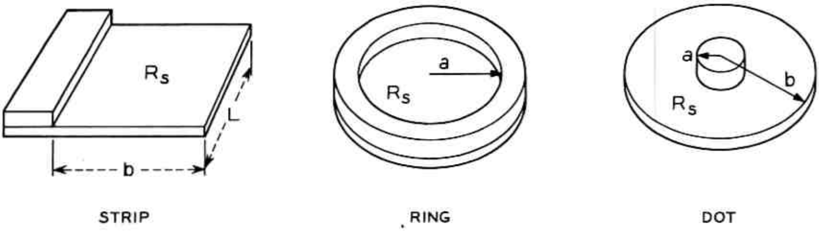


Fig. 20 — Contact shapes for ohmic contacts on thin metal film of Schottky barrier photodiode.

linear contact shown in Fig. 20 is independent of the dimension L because identical diodes connected in parallel will display the same frequency response if they are operated under short circuit conditions.

All three types of contacts can be used for Schottky barrier photodiodes. Ring contacts will give the highest cutoff frequencies for a specified diode area and given material properties. The fabrication of dot contacts is simpler because the ring contact has to be deposited by masking off the center area of the diode for the deposition of the contact ring. Dot contacts can be evaporated through an ordinary metal mask with an array of holes. This makes dot contacts particularly useful for photodiode arrays or for the fabrication of a large number of photodiodes on a single wafer which can be sliced up later. It is convenient to set the contact dot off center in order to facilitate the attachment of an external connection without interfering with the incident light beam. This is shown in Fig. 21. The large dots are semitransparent gold films on Si with a diameter of 10 mils and a sheet resistance of 5 to 7 ohm/square. The small contact dots have a diameter of 3 mils. The capacitance of each diode for a substrate material with a resistivity of 2.7 ohm cm at a back bias of 60 volt is 0.9 to 1.0 pF. The cutoff frequency cannot be calculated from (17) for the dot contact because the contact in Fig. 21 is off centered. A good approximation is obtained by applying (15) for the linear contact with b being the diameter of the semitransparent gold film. The cutoff frequency obtained for this particular diode at the specified back bias of 60 volt is $f_c = \omega_c/2\pi = 22$ GHz. This cutoff is of the same order as the cutoff frequency obtained from transit time considerations.

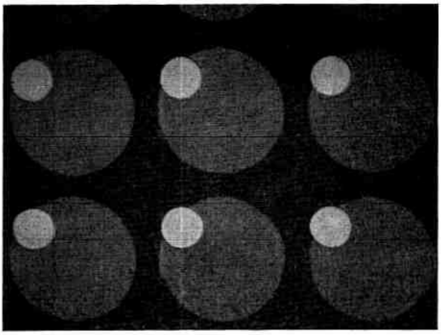


Fig. 21 — Array of Schottky barrier photodiodes on epitaxial silicon wafer before deposition of antireflecting coating. The diameter of the large semitransparent gold dots is 0.25 mm, the diameter of the small gold contact dots is 0.075 mm.

VI. FABRICATION AND PACKAGING OF Si-Au-ZnS PHOTODIODES

The combination of materials for Schottky barrier photodiodes depends on the frequency range of the incident radiation and the metallurgical properties of the metal-semiconductor system. Silicon and germanium are both suitable for the visible range of the spectrum. Stable Schottky barriers can be formed with a number of other metals, e.g., Ag, Al, Pt, and Ni. The eutectic temperatures of the various metal-semiconductor combinations determine the maximum device temperature. The eutectic temperature of Au-Si is 370°C. The choice of the matching coating is governed by the optical surface impedance of the metal-semiconductor combination. A total optical return loss of 12 dB can be achieved with ZnS which has a dielectric constant $\varepsilon_r = 2.3$. A higher return loss may be desirable; however, the stability of ZnS and the good adherence to the Au is an advantage compared to other dielectric materials.

The surface preparation of the semiconductor substrate is relatively simple. Epitaxial silicon wafers n on n^+ with alloyed gold antimony back contacts are rubbed with a clean cotton swab under methanol, boiled in distilled water, etched in HF, washed in distilled deionized water, washed in methanol and finally dried with nitrogen. The wafer is covered with a molybdenum mask with an array of 10-mil diameter holes. Only $\frac{2}{3}$ of the wafer are covered by this mask. The remaining $\frac{1}{3}$ of the wafer is later used for test purposes of the optical return loss during the evaporation of the antireflection coating. The unit is transferred into the vacuum system which is pumped down to a pressure of $2\text{–}3 \times 10^{-7}$ torr measured at its pumping port. Gold is evaporated from a tungsten coil. The sheet resistance of the gold is measured on a 1-mm quartz slide which is located adjacent to the silicon wafer. The evaporation is discontinued when the sheet resistance measured on the quartz slide is in the range of 5 to 7 ohm/square. Separate measurements have shown that the sheet resistance of the Au on the Si wafer is also in the 5 to 7 ohm/square range.

A second deposition of Au through a molybdenum mask with an array of 3-mil holes is made on the wafer. The 3-mil Au dots are needed later for contacting purposes. A photograph of the semitransparent Au dots and the contacting 3-mil dots is shown in Fig. 21. The second mask is removed after the Au evaporation and the unit is mounted in a vacuum system which is equipped with an optical reflectometer at $\lambda = 6328$ Å. Zinc sulfide is evaporated on the wafer. The reflectance from the test area on the wafer is measured continuously and the evaporation is stopped at the first maximum of the return loss. Typical results obtained from the reflectometer recording are shown in Fig. 14.

The step of depositing 3-mil Au contacts is repeated in order to facilitate thermocompression bonding to the contact area. The second mask is mounted in exact registry with the first evaporation of contact dots. This means that a ZnS layer is sandwiched between two identical contact dots. This layer is shorted out after completion of the thermocompression bonding process of a 1-mil Au wire to the contact dot. Fig. 22 is a photograph of the wafer surface after completion of all evaporation processes. The wafer surfaces shown in Figs. 21 and 22 are both illuminated from a standard tungsten lamp for obtaining the photographs.

A cross sectional view of the detector packaged into a modified type N connector body is shown in Fig. 23 and a photograph of the completed structure in Fig. 24. A metallized quartz washer is used for electrical separation of the diode terminals. A bypass capacitor provides an RF short between the outer conductor of the connector body and the terminal which is connected to the metal side of the Schottky barrier. Various parts of the package are identified in the figure caption.

VII. MEASUREMENT OF PULSE RESPONSE AND NET QUANTUM EFFICIENCY

The pulse response of packaged Schottky barrier photodiodes has been examined by phase locking the TEM_{00q} modes of a 6328 Å He-Ne gas laser with an internal phase modulator. The laser output consists of pulses with a half width of approximately 0.5 nanoseconds separated by 11 nanoseconds. The average optical power is 0.3–0.4 milliwatt.

A typical pulse response obtained with a completed photodiode dis-

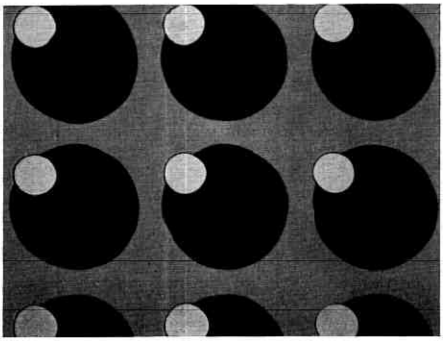


Fig. 22 — Array of Schottky barrier photodiodes on epitaxial silicon wafer after deposition of the antireflecting coating. The dot dimensions are the same as in Fig. 21.

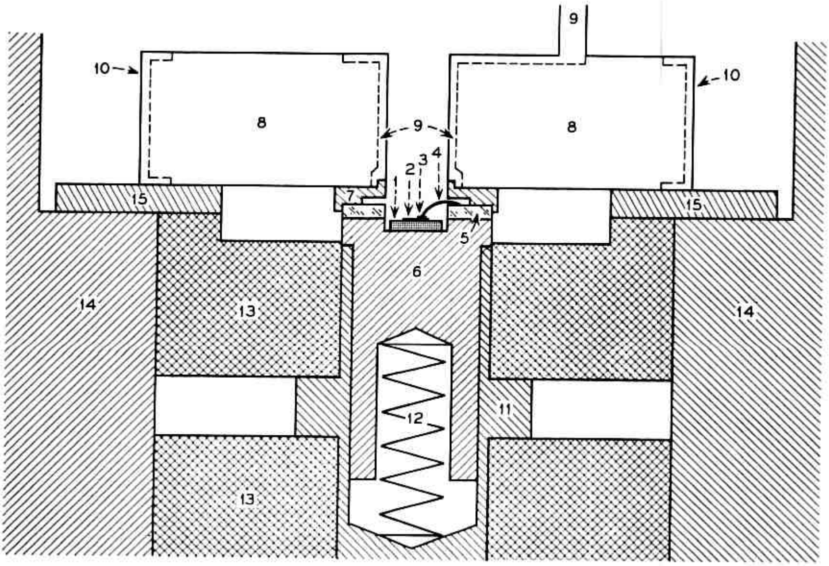


Fig. 23 — Cross-sectional view of diode package. (1. Silicon wafer. 2. Thin gold film. 3. Thermocompression bond on gold contact. 4. Contact wire to quartz washer. 5. Quartz washer, top and bottom are metallized. 6. Brass pin. 7. Brass adapter ring. 8. Bypass button capacitor 1000 pF. 9. Inside metal connection of button capacitor and external lead for applying dc bias to photodiode. 10. External metal connection of button capacitor. 11. Brass pin forming part of center conductor of the connector. 12. Steel spring. 13. Teflon spacers. 14. Connector body. 15. Steel washer.)

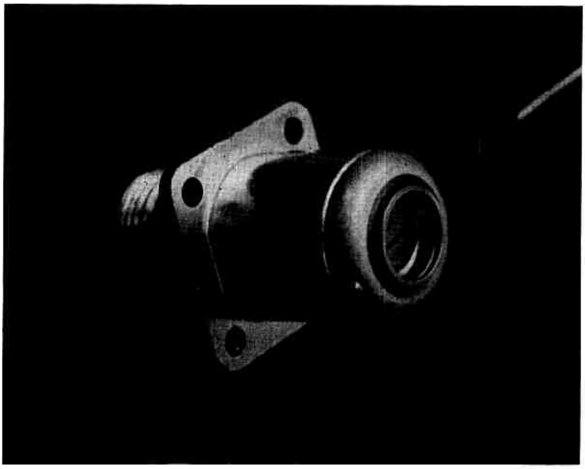


Fig. 24 — Completed diode package showing connector body and external lead wire for dc bias.

played on a Tektronix sampling oscilloscope Type 661 with a rise time of 0.1 nsec is shown in Fig. 25. The diode is made on 2.7-ohm cm epitaxial silicon with the 10-mil diameter dots shown in Figs. 21 and 22. The half width of the pulses is 0.45 nsec at a back bias of 50 volts. The net quantum efficiency measured with an Eppley thermocouple #4952 as a reference is 70 percent. This efficiency is obtained by graphical integration of the pulse shape shown in Fig. 25 and by assuming that the diode acts like an ideal current source into the 50-ohm broadband load of the sampling oscilloscope.

A close inspection of the pulse shape shows that the leading edge is slightly steeper than the trailing edge. The distortion in the trailing edge is due to diffusion current and to case capacitance in the package. The influence of the diffusion current can be examined by observing the pulse shape for various back bias conditions. Fig. 26 is the pulse response of the same diode at a back bias of 0, 4, 15, and 50 volt. A diffusion tail is clearly visible at a back bias of 0 volt and 4 volt. The diffusion tail is depressed at higher back bias because more carriers are created within the depletion layer.

An important property required for many practical applications is a uniform response of the photodiode over the entire area of the junction.

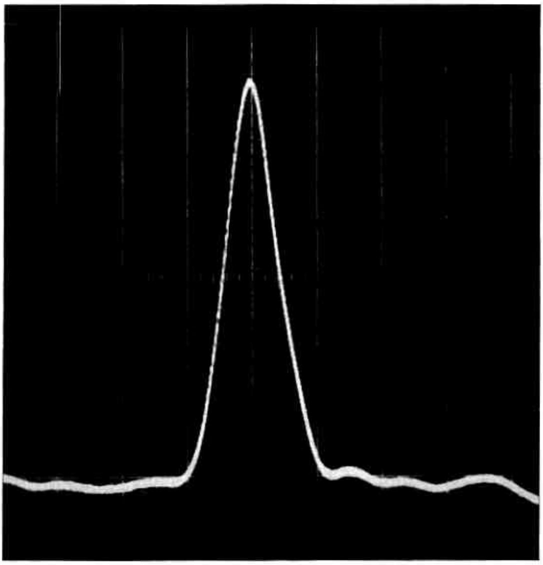


Fig. 25 — Pulse response of packaged diode at 50-volt back bias into 50-ohm load obtained from phase locked modes of He-Ne gas laser. Horizontal scale 0.5 nsec/cm, vertical scale 20 mvolt/cm.

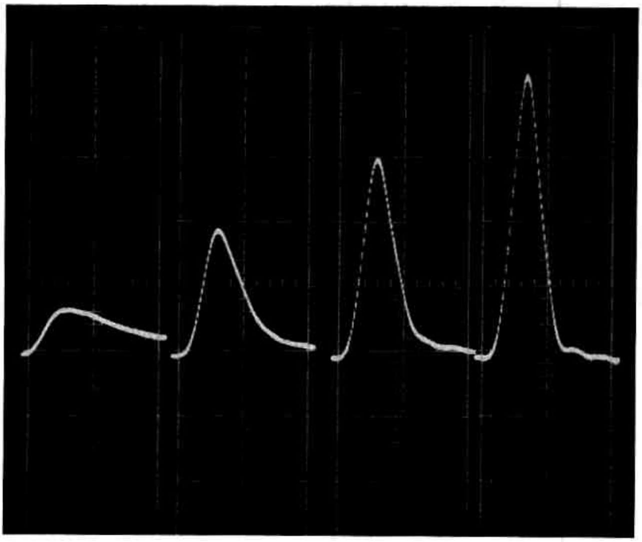


Fig. 26 — Pulse response of Schottky barrier photodiode at 0-volt, 4-volt, 15-volt, and 50-volt back bias. Horizontal scale 1 nsec/cm, vertical scale 25 mvolt/cm.

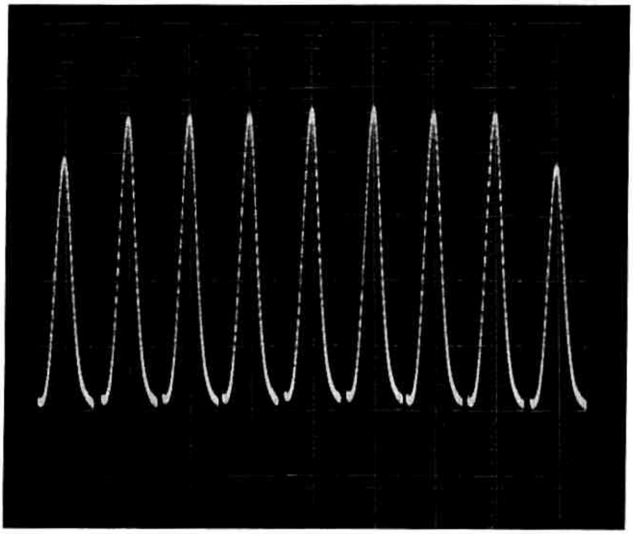


Fig. 27 — Pulse response for 9 points on the same photodiode obtained by linear scanning of a focused laser beam over diode area by 0.025-mm increments. Horizontal scale 2 nsec/cm, vertical scale 25 mvolt/cm.

Fig. 27 shows the pulse response of a Schottky barrier photodiode at various locations of the diode. A laser beam is focused on the front surface of the diode and is scanned across the diode. The pulse response is measured on an axis at discrete points which are spaced 1 mil apart. The peak variation is less than 3 percent over a total distance of 7 mils. The reduced pulse response in the vicinity of the boundaries is due to the fact that there is a small thickness change of the antireflection coating close to the boundary. This change of thickness is due to different sticking coefficients and different surface mobilities of the zinc sulfide on gold and on silicon during the evaporation process. One observes therefore a reduced amplitude response with no degradation of the pulse shape.

VIII. CONCLUSIONS

Schottky barrier photodiodes can be used for fast and efficient optical detectors. The high efficiency is obtained because radiation can be coupled through thin metal films with relatively low loss at optical frequencies. The small reflectance of the diode is achieved by proper choice of the matching layer. A diode with a fast response is obtained by designing junctions with a small RC product. The problem is similar to building high cutoff Schottky barriers for microwave and millimeter wave circuits. Additional limitations are due to transit time effects which are common to all solid-state radiation detectors based on carrier generation.

IX. ACKNOWLEDGMENTS

The author would like to thank E. I. Gordon and D. L. Perry for providing gas lasers and optical components for monitoring thin film deposition. He also thanks R. F. Trambarulo for the use of Figs. 17 and 18 and R. E. Neeld for micromechanical precision work. Provision of pulsed lasers based on mode phase locking and built by O. E. DeLange, A. F. Dietrich, and F. R. Nash is also gratefully acknowledged.

REFERENCES

1. Schottky, W., Semiconductor Theory of Blocking Layer and Point Contact Rectifier, Z. Physik, *113*, 1939, pp. 367–414.
2. Schottky, W. and Spenke, E., Zur quantitativen Durchführung der Raumladungs- und Randschichttheorie der Kristallgleichrichter, Wiss. Veröff. Siemens-Werken, *18*, 1939, pp. 225–291.
3. Kahng, D., Au-n-Type GaAs Schottky Barrier and Its Varactor Application, B.S.T.J., *43*, January, 1964, pp. 215–224.
4. Kahng, D. and D'Asaro, L. A., Gold-Epitaxial Silicon High-Frequency Diodes, B.S.T.J., *43*, January, 1964, pp. 225–232.

5. Irvin, J. C. and Young, D. T., Millimeter Frequency Conversion using Au-n-Type GaAs Schottky Barrier Epitaxial Diodes with a Novel Contacting Technique, Proc. IEEE, *53*, December, 1965, pp. 2130–2131.
6. Anderson, L. K., Photodiode Detection, Proc. Symp. Optical Masers, Microwave Research Institute Symposia Series, Polytechnic Institute of Brooklyn, *8*, April, 1963, pp. 549–566.
7. Lucovsky, G. and Emmons, R. B., High Frequency Photodiodes, Appl. Opt., *4*, June, 1965, pp. 697–702.
8. Riesz, R. P., High Speed Semiconductor Photodiodes, Rev. Sci. Instr., *38*, September, 1962, pp. 994–998.
9. Anderson, L. K., McMullin, P. G., D'Asaro, L. A., and Goetzberger, A., Microwave Photodiodes Exhibiting Microplasma-free Carrier Multiplication, Appl. Phys. Letters, *6*, February, 1965, pp. 62–64.
10. Melchior, H. and Lynch, W. T., Signal and Noise Response of High Speed Germanium Avalanche Photodiodes, IEEE Trans. Electron Devices, to be published.
11. Sharpless, W. M., Cartridge-Type Point-Contact Photodiode, Proc. IEEE, *52*, February, 1964, pp. 207–208.
12. Di Domenico, M., Sharpless, W. M.,and McNicol, J. J., High Speed Photodetection in Germanium and Silicon Cartridge-Type Point-Contact Photodiodes, Appl. Opt., *4*, June, 1965, pp. 677–682.
13. Schopper, H., *Landolt-Börnstein Zahlenwerte aus Physik, Chemie und Technik*, Springer-Verlag, Göttingen, *2*, Part 8, 1962, pp. 1–42.
14. Heavens, O. S., *Reports on Progress in Physics*, Institute of Physics and the Physical Society, London, *23*, 1960, pp. 1–65.
15. Mayer, H., *Physik Dünner Schichten*, Wissenschaftliche Verlagsgesellschaft, Stuttgart, Part 1, 1950, pp. 293–307.
16. Parker Givens, M., *Solid State Physics*, ed. by F. Seitz and D. Turnbull, Academic Press, Inc., New York, *6*, 1958, pp. 313–352.
17. Abelès, F., *Physics of Thin Films*, ed. by G. Hass and R. E. Thun, Academic Press, Inc., New York, *3*, 1965.
18. Bennett, J. M. and Ashley, E. J., Infrared Reflectance and Emittance of Silver and Gold Evaporated in Ultrahigh Vacuum, Appl. Opt. *4*, February, 1965, pp. 221–224.
19. Behrndt, K. H., *Deposition Processes for Thin Films*, *Techniques in Metals Research*, ed. by R. F. Bunshah, *1*, Interscience Publishers, New York, 1967.
20. Hall, A. C., Experimental Determination of the Optical Constants of Metals, J. Opt. Soc. Am., *55*, August, 1965, pp. 911–915.
21. Holland, L., *The Properties of Glass Surfaces*, John Wiley and Sons, Inc., New York, 1964, pp. 300–301.
22. Chopra, K. L., Influence of Electric Field on the Growth of Thin Metal Films, J. Appl. Phys., *37*, May, 1966, pp. 2249–2254.
23. Francombe, M. H. and Sato, H., *Single-Crystal Films*, The MacMillan Company, New York, 1964.
24. Wolter, H., *Optics of Thin Films*, *Encyclopedia of Physics*, *24*, Springer Verlag, Göttingen, 1956, pp. 461–473.
25. Fry, T. C., Plane Waves of Light, J. Opt. Soc. Am., *15*, September, 1927, pp. 137–161.
26. Fry, T. C., Plane Waves of Light, J. Opt. Soc. Am., *16*, January, 1928, pp. 1–25.
27. Perry, D. L., Low Loss Multilayer Dielectric Mirrors, Appl. Opt., *4*, August, 1965, pp. 987–995.
28. Kahng, D., Conduction Properties of the Au-n-Type-Si Schottky Barrier, Solid State Elec., *6*, May, 1963, pp. 281–295.
29. Cowley, A. M. and Sze, S. M., Surface States and Barrier Height of Metal-Semiconductor Systems, J. Appl. Phys., *36*, October, 1965, pp. 3212–3220.
30. Gärtner, W. W., Depletion Layer Photoeffects in Semiconductors, Phys. Rev., *116*, October 1, 1959, pp. 84–87.
31. DeLoach, B. C., personal communication.
32. Lucovsky, G. and Emmons, R. B., Lateral Effects in High-Speed Photodiodes, IEEE Trans. Electron Devices, *ED-12*, January, 1965, pp. 5–12.

Topology of Thin Film RC Circuits

By F. W. SINDEN

(Manuscript received August 31, 1966)

Integrated RC circuits can be made by depositing exceedingly thin metallic and dielectric films in suitable patterns on an insulating substrate. Resistors are strips of conductor; capacitors are patches on which conducting, dielectric, and conducting layers are superimposed. Since conductors can cross at capacitor patches, RC networks need not be strictly planar to be realizable in thin film.

Determining which RC circuits are realizable poses new problems in topology which are remarkably simple to state but are as yet unsolved. The results reported here are fragmentary, but they do cover some cases of small order that may be of practical interest.

I. INTRODUCTION

Integrated RC circuits can be made by depositing exceedingly thin metallic and dielectric films in suitable patterns on an insulating substrate. A resistor is made by depositing a long, narrow strip of conductor (usually in a zag-zag for compactness); a capacitor is made by superimposing conducting, dielectric, and conducting layers. Because the dielectric is thin, the capacitance per unit area is high. Fig. 1 shows a typical thin film pattern.

Ordinarily printed circuits are strictly planar; crossovers are made only by leading one of the conductors entirely out of the plane of the circuit. In the thin film technique, however, conductors can be separated by thin insulating layers within the plane of the circuit. Thus, crossovers can be permitted provided a nonzero capacitance between the crossing conductors is acceptable. If an RC circuit can be laid out so that conductors cross only if the circuit requires a nonzero capacitance between them, we will say the circuit is realizable in thin film or just *realizable.*

An example of a realizable nonplanar circuit is shown in Fig. 2. In this case, the schematic thin film layout brings out intrinsic symmetries not displayed by the circuit diagram.

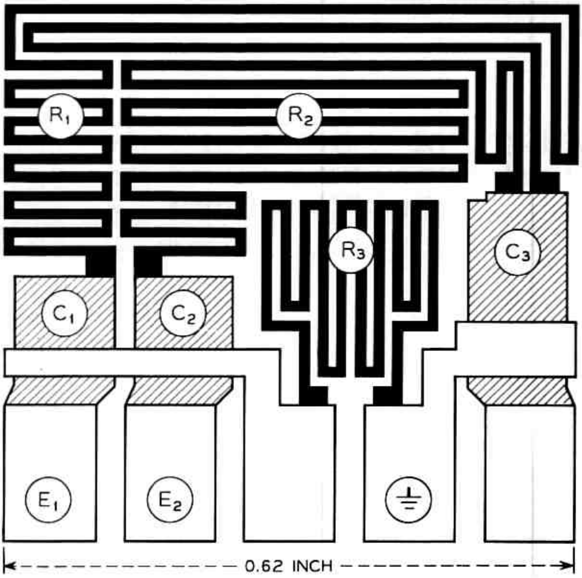


Fig. 1 — Thin film layout for a notch filter (courtesy W. H. Orr). Black region is bottom conductor; shaded region is dielectric; white region is top conductor.

Finding feasible layouts, or even determining when they exist, leads to unsolved problems in topology. The results presented here give answers only in special cases. Moreover, these results concern only the topological side of the problem; *electrical* equivalences are not taken into account. It is assumed that the network is given topologically and that

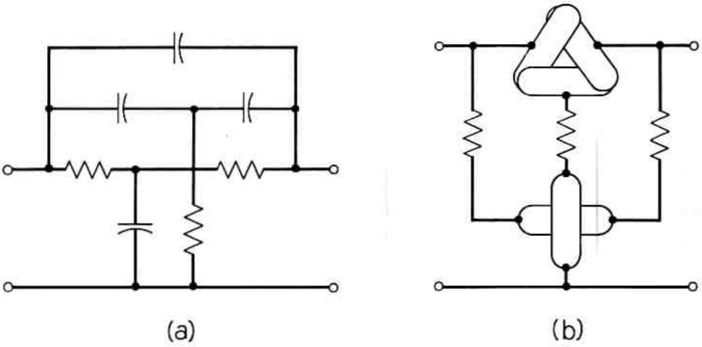


Fig. 2 — (a) Nonplanar circuit ("twin-tee", Ref. 3, p. 309); (b) schematic thin film layout for the circuit in (a).

terminals to the outside are located in given fixed positions on the periphery of the board.

II. SEPARATION OF THE RESISTIVE AND CAPACITIVE PARTS

Given an RC network N, let R_N be the purely resistive network obtained by replacing every capacitor by a direct connection. Clearly N is not realizable in thin film unless R_N is. R_N is realizable only if its graph (a vertex for each conductor, an edge for each resistor) is planar under the restrictions imposed by the locations of the terminals to the outside (see Fig. 3). This observation provides a first check: if R_N is not planar, there is no need to proceed further.

Each vertex in the graph of R_N replaces a purely capacitive network. In Fig. 3, for example, the vertex V in R_N replaces the network shown in Fig. 4.

One way to construct a realization of N is to construct realizations for the individual vertex-networks, and then to fit these into the planar layout of R_N. Since the layout of R_N may not be unique (there may be more than one ordering of edges about a vertex) the conditions on the vertex-networks may not be unique.

Another approach, discussed briefly in the final section, is to modify algorithms for purely capacitive networks to take account of resistors. In either case, one needs to study the purely capacitive networks first.

III. PURE C NETWORKS

A pure C network is a set of zero-resistance conductors $c_1, \cdots, c_r$ some pairs of which are connected by capacitors. The problem of finding a feasible layout for such a network is the following:

For each conductor c_i find a connected region R_i in the plane such that

(*i*) R_i and R_j have common points if and only if c_i and c_j are connected by a capacitor, and

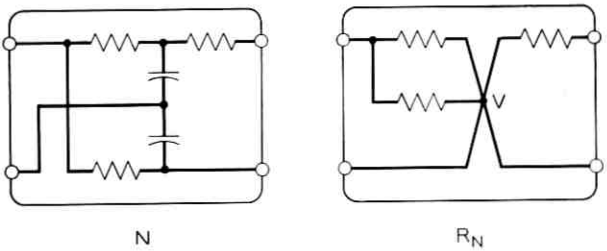


Fig. 3 — Nonplanar RC network N and reduced purely resistive network R_N.

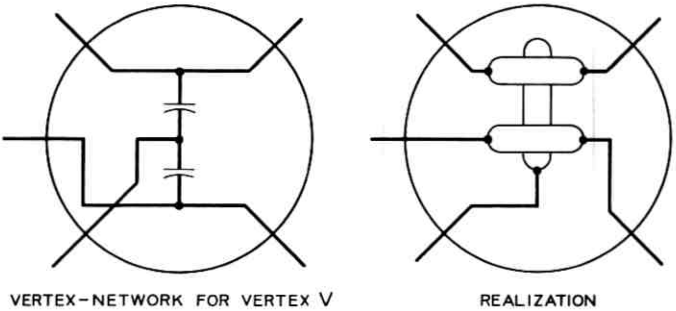


Fig. 4 — Capacitive network for vertex V of Fig. 3 and realization of this network.

(*ii*) no point belongs to more than two regions.

Condition (*ii*) says that no more than two conductors (separated by dielectric) may be superimposed. If, contrary to condition (*ii*), conducting and dielectric layers can be stacked up indefinitely, then *every* connected C network has a feasible layout. (The network is *connected* if any conductor can be reached from any other through a sequence of capacitors.) This is not quite immediately obvious; a proof is given in Appendix A.1.

Indefinite stacking offers other advantages as well.[1] Unfortunately it also presents technical difficulties. To date most thin film circuits have been limited to two conducting layers.

It does not change the problem to replace the connected regions R_i by curves C_i of finite length, since a connected region can be nearly filled by a curve of finite length, and a curve of finite length can be approximated by a narrow region. When convenient, the curves can have branches, although this is not necessary, since a branch can be approximated by letting the curve double back. In some cases, a pair of curves, whether branched or not, have to cross more than once (examples later). Such multiple crossings will be permitted on the assumption that a capacitance, if need be, can be distributed over several crossovers. Sometimes the curves are more convenient and sometimes the regions. I will use both.

In addition to satisfying conditions (*i*) and (*ii*) the regions (or curves) may have to satisfy constraints associated with the terminals to the outside. More specifically, $R_1, \cdots, R_r$ may be required to lie within a given region R and certain of the R_i may be required to contain specified points P_i on the boundary of R. I will consider mainly the two extreme cases where (a) there are no such terminal constraints and (b) every region R_i satisfies a terminal constraint.

IV. UNCONSTRAINED CASE

The problem is simply stated: It is specified which pairs of a set of curves (or connected regions) in the plane cross and which pairs do not. When are such specifications consistent?

To get a feeling for the problem, the reader may wish to try the examples in Fig. 5.

The crossings are conveniently specified by means of a graph G. Associate a vertex with each curve, and let two vertices be joined by an edge if and only if the corresponding curves are required to cross. If a set of curves satisfying the crossing specifications exists, we will say that the graph G is *realizable*.

If G is planar, then it is realizable. In a planar representation of G one has merely to replace each vertex v_i by a star-shaped region R whose points extend out along the edges emanating from v_i far enough to overlap the points of neighboring regions.

The converse is not true; some nonplanar graphs are realizable. For instance, any complete graph (nonplanar if the order is greater than four) is realizable, for in this case every curve C_i crosses every other. (Let the C_i be straight lines in general position; i.e., no two parallel, no three through a point.)

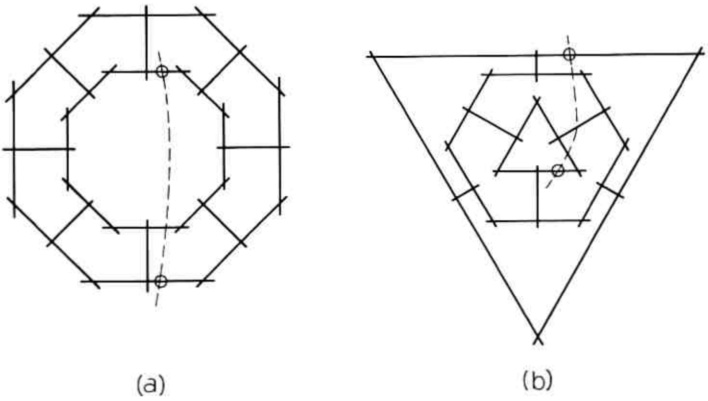


Fig. 5 — Examples of unconstrained case. With the exception of the dashed curve, a pair of curves must cross if and only if they cross in the figure. The dashed curve must make only the encircled crossings. One of these examples has a solution; the other does not. Answers are given in Appendix A.2.

Although nonrealizable is different from nonplanar there is a class of nonrealizable graphs that is related to nonplanar graphs. If G is nonplanar, then the graph G^* obtained by inserting a new vertex into each edge of G is nonrealizable (see Fig. 6). If G^* were realizable, one could construct a planar representation of G as follows. In a realization of G^* let each of the curves C_i corresponding to an original vertex of G shrink to a point in such a way that no new crossings are generated. This is always possible. Since by assumption the remaining curves (corresponding to edges of G) do not cross each other, the resulting figure is a planar representation of G.

A theorem of Kuratowski[2] states that any nonplanar graph can be reduced to one of two minimal nonplanar graphs G_1 or G_2 (Fig. 7) by (*i*) deleting edges and (*ii*) combining adjacent vertices.†

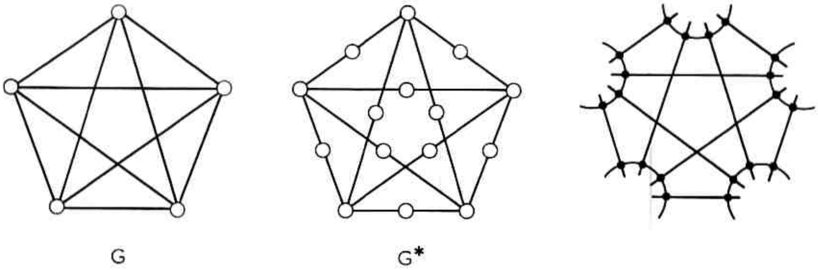


Fig. 6 — G is nonplanar; G^* is nonrealizable. On the right is a nonrealization of G^*; crossings marked with dots are required, no others are permitted.

The two operations (*i*) and (*ii*) clearly preserve planarity. Operation (*ii*) also preserves realizability, but (*i*) does not. (If it did, all graphs would be realizable, since any graph can be constructed by deleting edges of a complete graph, which is realizable.) To preserve realizability it is necessary to replace (*i*) by the weaker operation (*i'*): deleting *vertices* (together with attached edges). To see that (*i'*) and (*ii*) do indeed preserve realizability one has only to interpret them as operations on the curves C_i.

Using operations (*i'*) and (*ii*) and Kuratowski's theorem we can identify a class of nonrealizable graphs as follows.

Let G_1^* and G_2^* be the graphs obtained by inserting a new vertex

† G_1 is the graph involved in the familiar problem of connecting three utilities (e.g., the gas, water, and electric plants) to three houses without crossing lines. Since G_1 is nonplanar there is no solution. In Fig. 7 vertices 1, 3, and 5 can be taken as the utilities and 2, 4, and 6 as the houses.

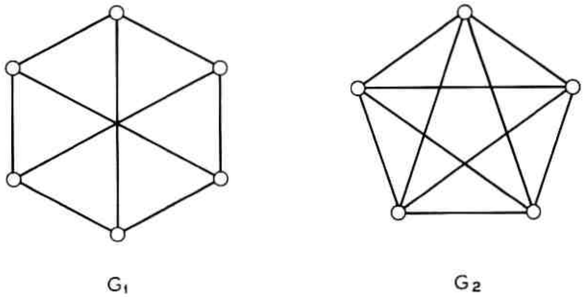


Fig. 7 — Kuratowski graphs.

into each edge of the Kuratowski graphs G_1 and G_2. A graph is nonrealizable if it can be reduced to G_1^* or G_2^* by application of (*i'*) and (*ii*). G_1^* and G_2^* are themselves irreducible. In Appendix A.2 one of the examples in Fig. 5 is shown to be reducible to G_1^*, hence nonrealizable.

The analogue of Kuratowski's theorem which would say that every nonrealizable graph can be reduced to G_1^* or G_2^* is false. An example of a nonrealizable graph that cannot be so reduced is given in Appendix A.3.

V. CONSTRAINED CASE

In addition to satisfying the conditions (*i*) and (*ii*) in Section III, the curves C_i (or the regions R_i) will now be required to lie within a simply-connected region R (which we shall take to be a disk) and each C_i will be required to contain a specified point P_i on the boundary of R. (This covers the case where a single conductor is required to join two or more separate terminals. One has only to require that the corresponding curves cross each other; their union represents the conductor.)

Before proceeding further, the reader may wish to try the examples in Fig. 8.

In passing, we observe that any constrained problem can be imbedded in an unconstrained problem. The constraints can be simulated by means of a ring structure containing $2r$ curves, where r is the number of curves in the constrained problem. This is proved in connection with the example discussed in Appendix A.3. Unfortunately, this observation is of little use in the absence of more information about the unconstrained case.

We will regard the vertices $v_1, \cdots, v_r$ of graph G as residing at the terminal points $P_1, \cdots, P_r$. We will often make use of the complement

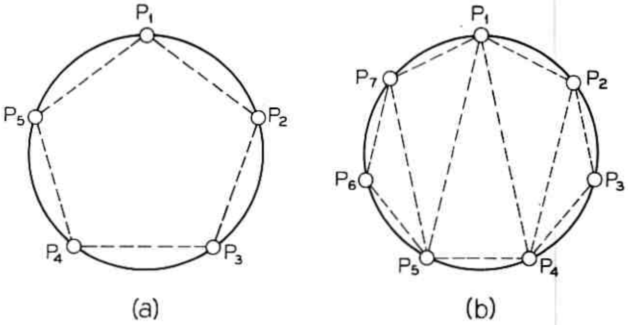


Fig. 8 — Examples of constrained case. Curve C_i must contain point P_i and lie otherwise within the circle. The dashed lines show the edges *not* in G, i.e., if P_i and P_j are connected by a dashed line then curves C_i and C_j may *not* cross; otherwise C_i and C_j *must* cross. One example has a solution, the other does not. Answers in Appendix A.4.

$\bar{G}$ of G, where $\bar{G}$ consists of all edges *not* in G. Edges in G will be shown as solid lines, edges in $\bar{G}$ as dashed lines.

A subset of vertex points $P_{i_1}, \cdots, P_{i_n}$ such that $i_1 < i_2 < \cdots < i_n$ will be called a *cycle* if all the pairs

$$(P_{i_1}, P_{i_2}),(P_{i_2}, P_{i_3}), \cdots, (P_{i_n}, P_{i_1})$$

are joined by edges. A cycle will be called empty if no other pairs are joined by edges. We will be primarily concerned with empty cycles in the complementary graph $\bar{G}$. (See Fig. 9)

Theorem 1: A necessary condition for a constrained graph G to be realizable is that $\bar{G}$ contain no empty cycles of order four or more.

Proof: (*i*) If G is an empty cycle of order four, then G is not realizable. This is easily verified by inspection. If, therefore, $\bar{G}$ *contains* an empty cycle of order four, then G is not realizable.

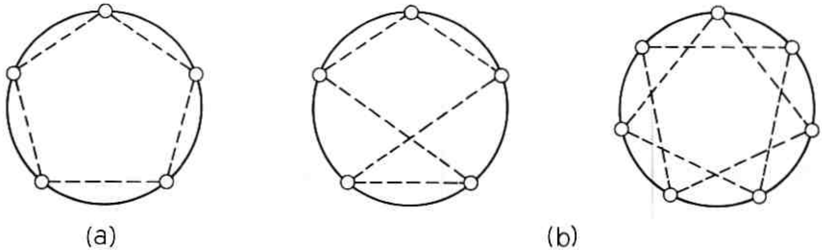


Fig. 9 — (a) Empty cycle in $\bar{G}$, (b) non-cycles. Dashed edges belong to $\bar{G}$; edges not shown belong to G.

(*ii*) Suppose the theorem is known to be true for cycles of order $4, \cdots, m - 1$ and suppose, contrary to the theorem, that $\bar{G}$ contains an empty cycle of order m and that G is realizable. The realization of G can be generated in the following way: let curve C_1 grow continuously out of point P_1 until it reaches its full length, then let curve C_2 grow out of point P_2 until *it* reaches *its* full length, and so on until all curves are complete.

Let $\bar{G}(t)$ be the corresponding complementary graph at time t. At the beginning, $\bar{G}(t)$ is the complete graph (no crossings); as the crossings are generated one by one, edges are deleted from $\bar{G}(t)$. At some stage the postulated empty cycle of order m, which is contained in the final form of $\bar{G}$, must have just one internal edge left. But this last internal edge forms two empty cycles inside the final cycle, at least one of which must be of order four or more (since $m > 4$) and less than order m. Therefore, by the induction hypothesis, there can be no realization at this intermediate stage. Contradiction.

For some time it appeared to me that the empty cycle condition was not only necessary for the realizability of a constrained graph, but sufficient as well. Recently, though, I found a counterexample of order eight. This example is discussed in Appendix A.5.

Following are a number of results that help to identify and construct special classes of realizable constrained graphs. Taken together these seem to cover most cases of small order.

If no two edges of G cross, then clearly G is realizable. Less obvious is a similar result for $\bar{G}$:

Theorem 2: A sufficient condition for a constrained graph G to be realizable is that $\bar{G}$ contain no empty cycles of order four or more and that no two edges of $\bar{G}$ cross.

An example of such a $\bar{G}$ is the triangulated polygon of Fig. 8(b). This example, typical of the genre, has a complicated solution with unavoidable multiple crossings.

Theorem 2 is proved in Appendix B. A more general result, also proved in Appendix B, is the following:

Theorem 3: (i) If (P_1, P_k) is an edge of G that crosses no other edges of G, and if the subgraphs G' with vertices $P_1, P_2, \cdots, P_k$, and G'' with vertices $P_k, \cdots, P_r, P_1$ are both realizable, then G is realizable.

(ii) If (P_1, P_k) is an edge of $\bar{G}$ that crosses no other edges of $\bar{G}$, and if subgraphs G' with vertices $P_1, P_2, \cdots, P_k$, and G'' with vertices $P_k, \cdots, P_r, P_1$ are both realizable, then G is realizable.

The following two theorems describe circumstances under which a new curve C_{r+1} can be added to an existing solution. In many cases the entire solution can be generated by adding curves one at a time.

Theorem 4: Let G be a constrained graph with vertices $P_1, \cdots, P_r, P_{r+1}$. *G is realizable if (i) the subgraph of G with vertices* $P_1, \cdots, P_r$ *is realizable, and (ii) there do not exist three vertices* P_i, P_j, P_k, $i < j < k < r + 1$ *such that* $P_{r+1}P_i$ *and* $P_{r+1}P_k$ *are edges of* $\bar{G}$ *and* P_iP_k *and* $P_{r+1}P_j$ *are edges of G. (See Fig. 10.)*

Though cumbersome to state, this theorem is usually easy to apply. The following special cases are often useful by themselves. Let S be the set of vertices joined to P_{r+1} by edges of $\bar{G}$. Special case 1: the vertices of S are an adjacent string. Special case 2: every pair of vertices in S is joined by an edge of $\bar{G}$. Special case 2, for instance can solve examples like 8(b) in which $\bar{G}$ is a triangulated polygon. One has only to add new vertices one at a time in such a way that each additional vertex forms one new triangle in $\bar{G}$. The set S always has just two members.

Theorem 4 is proved in Appendix B. Though somewhat involved when worked out in detail, the idea of the proof is simple. In the situation of Fig. 10 the curves C_i and C_k (emanating from P_i and P_k) form a barrier which C_{r+1} cannot cross. This does not necessarily prevent C_{r+1} from intersecting C_j, for it is possible that C_j could cross the barrier. If, however, the barrier is not there, then C_{r+1} can reach C_j on its own without C_j's help. If there are no barriers of the Fig. 10 type, then C_{r+1} can reach all of the curves it is supposed to cross no matter how these may have been drawn. Thus, the new curve C_{r+1} can be added without disturbing the old ones.

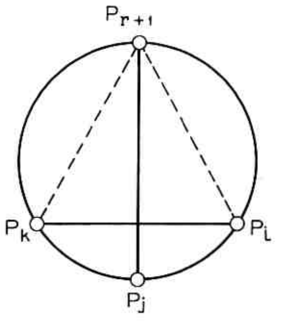


Fig. 10 — Configuration forbidden by hypothesis of Theorem 4. Dashed lines show edges of $\bar{G}$; solid lines show edges of G.

The next theorem concerns an operation which I will call an *adjacent interchange*. Given the circle R with the peripheral points $P_1, \cdots, P_r$, let R' be a slightly smaller circle concentric to R with corresponding peripheral points $P_1', \cdots, P_r'$. Let the primed points have the same order as the unprimed points except for one adjacent pair P_k', P_{k+1}', which is interchanged. The points $P_1, \cdots, P_r$, can be joined, respectively, to $P_1', \cdots, P_r'$ by curves $C_1 \cdots, C_r$ in such a way that only C_k and C_{k+1} cross. (See Fig. 11.)

If the operation is repeated by means of a new circle R'' inside R', then the curves C_i are extended inward and one new crossing is generated. A sequence of such operations can be specified by giving the pair of currently adjacent points that is to be interchanged.

Theorem 5 states the conditions under which all of the intersection requirements of a curve can be satisfied by a sequence of adjacent interchanges. These conditions involve cycles in $\bar{G}$ (not necessarily empty) as defined just before Theorem 1. Note that the order of vertices in a cycle of $\bar{G}$ is invariant under adjacent interchanges.

We will say that a member P_i of a cycle in $\bar{G}$ is *active* if it is joined to some other member of the cycle by an edge of G.

Theorem 5: The intersection requirements of a curve C_i can be satisfied entirely by a sequence of adjacent interchanges if and only if P_i is not an active member of any cycle in $\bar{G}$.

Theorems 4 and 5 tend to be complementary; where one fails, the other often works. Fig. 8(b) is an example where Theorem 5 fails (every vertex is an active member of several cycles) and Theorem 4 works.

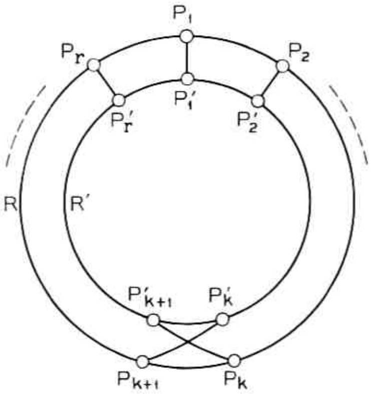


Fig. 11 — An adjacent interchange.

An example of the opposite kind is shown in Fig. 12. In this example Theorem 4 fails (every vertex has the forbidden configuration) but Theorem 5 works. The whole realization can be constructed by adjacent interchanges.

A realizable example to which neither Theorem 4 nor Theorem 5 applies is given in Appendix A.6. This is the smallest such example I have found (twelve vertices), but I doubt that it is really minimal.

VI. ORDER OF CROSSINGS

It is possible to obtain directly from the graph G information about the order in which crossings must occur along a given curve C_i. This information is contained in configurations I will call *empty chains.*

An *empty chain* is a subset of vertex points $P_{i_1}, P_{i_2}, \cdots, P_{i_n}$ in cyclic order such that the pairs (P_{i_1}, P_{i_2}), (P_{i_2}, P_{i_3}), $\cdots$, $(P_{i_{n-1}}, P_{i_n})$ are joined by edges of $\bar{G}$ and all other pairs are joined by edges of G.

An empty chain is just an empty cycle with a gap in it. Since the empty cycle is nonrealizable, it is not surprising to find that the realization of the empty chain, though not quite unique, is tightly determined. (See Fig. 13.)

Theorem 6: Let $P_1, \cdots, P_n$ be the vertices of an empty chain. Along curve C_k the first crossings with $C_1, \cdots, C_{k-2}$ must occur in that order; the first crossings with $C_{k+2}, \cdots, C_n$ must occur in reverse order.

The proof is given in Appendix B.

Every empty chain of length four or more yields ordering information. If, for instance, P_1, P_2, P_5, P_7 is an empty chain, then C_1 must cross C_7 before it crosses C_5 and C_7 must cross C_1 before C_2. Since most

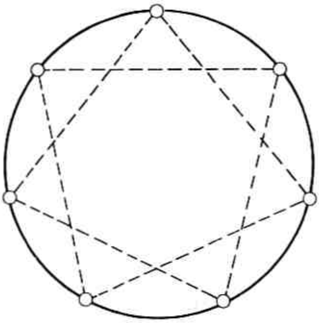


Fig. 12 — No vertex is an active member of any cycle in $\bar{G}$, therefore, a realization exists.

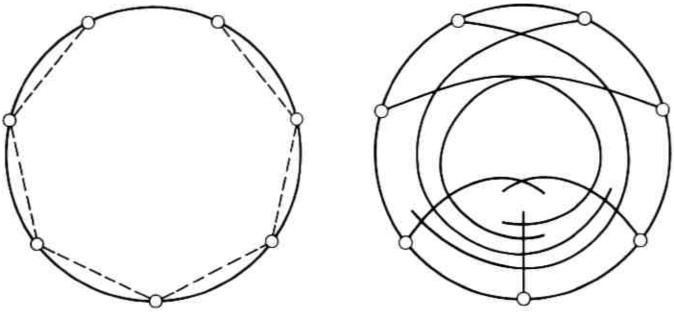


Fig. 13 — Realization of the empty chain of order seven.

examples of interest contain several such empty chains, this theorem is very generally applicable. The example of Fig. 8(b), for instance, contains six empty chains of order four and one of order five, which together give complete information about first crossings.

Searching for empty chains is tedious to do by hand, but could easily be done by machine.

A weakness of Theorem 6, evident in the example of Fig. 8(b), is that it says nothing about multiple crossings. It is clear in many examples that multiple crossings are determined by G. A way of extracting this information would be very useful.

VII. CONSTRUCTION OF SOLUTIONS — SUMMARY

The preceding results are not strong enough to define a guaranteed procedure for constructing realizations of constrained graphs. They do, however, seem to work in most cases of small order. To apply them one can proceed as follows:

(*i*) Look for empty cycles in $\bar{G}$ of order four or more. If any exist, G cannot be realized (Theorem 1).

(*ii*) Look for edges of G (or $\bar{G}$) that do not cross other edges of G (or $\bar{G}$). Such edges, if internal, permit the graph to be broken into two independent parts (Theorem 3).

(*iii*) Look for vertices that are free of the configuration shown in Fig. 10. Such vertices can be temporarily deleted, since the corresponding curves can be drawn in after the remaining curves have been drawn (Theorem 4).

(*iv*) Look for vertices that are not active members of any cycle in $\bar{G}$. These are typically on tree-like branches of $\bar{G}$. The corresponding curves can be constructed either at the beginning or the end by means of adjacent interchanges (Theorem 5).

(*v*) Find all the empty chains of order four or more and write down all the ordering relations they imply. Try to locate each crossing on both of its curves. This cannot always be done uniquely.

In a systematic procedure one could combine 1, 4, and 5 since these all involve chains and cycles.

Chains and cycles in $\bar{G}$ seem to be important in this problem; they certainly yield much information. But apparently they are not enough. To set up necessary and sufficient conditions for realizability, some other element is needed.

VIII. LOOSE ENDS

So far we have considered only completely constrained and completely unconstrained graphs, corresponding to networks where none or all of the conductors are connected to outside terminals. In general, of course, one wants the intermediate case where only some of the conductors are connected to outside terminals. This remains to be studied.

The preceding results can be used to construct realizations for the pure C networks represented by the nodes of the resistive network R_N. (See Section II.) Alternatively, one can generalize the pure C problem as follows to take account of resistors *a priori.*

The graph G can be replaced by its associated matrix A, where $a_{ij} = X$ (for "crossing") if conductors C_i and C_j are connected through a capacitor (or a short circuit) and $a_{ij} = 0$ (for "no crossing") if C_i and C_j are not so connected. To take account of resistors, we let $a_{ij} = T$ if C_i is connected to C_j through a resistor but not through a capacitor. This will mean topologically that C_i and C_j must touch without crossing.

T and X can be defined more precisely as follows. Consider instead of the curves C_i the regions R_i. We can assume that the $R_\imath$ are simply connected. If $a_{ij} = T$, then the part of R_i's boundary that lies inside R_j must be connected (i.e., a single piece). If $a_{ij} = X$ then the part of R_i's boundary inside R_j may (but need not) consist of several pieces.

A. J. Goldstein has observed that in constructing an algorithm, the regions R_i have advantages over the curves C_i. (The ends of the curves have an unnecessarily special character.) He suggests that an algorithm might be constructed that would keep track of all of the pieces of the boundaries of the R_i and take, so far as possible, only steps that are topologically mandatory. Such an algorithm could easily take account of both T and X connections. This idea has not been worked out in detail and we do not know how often one would be forced to take an arbitrary step that might be wrong.

APPENDIX A

Examples and Answers

A.1 If indefinite stacking of conducting and dielectric layers is permitted, then any connected G is realizable regardless of the positions of the outside terminals. A universal realization can be constructed as follows.

Since G is connected, there is a path in G that contains every vertex at least once. In their order along this path let the vertices be $v_1, \cdots, v_n$. Over a disk D, stack n layers of conductor separated by layers of dielectric. Associate the conductors with the vertices of G according to their order along the path. This is permissible since the conductors have nonzero capacitances only with their neighbors in the stack. These capacitances correspond to the edges in the path. An extension of any conductor can be brought out of the stack radially in any direction. Thus, any pair of conductors required to have a nonzero capacitance can be brought out together and superimposed in an arbitrarily long radial strip. Similarly, any conductor can be brought out in the appropriate direction to connect to an outside terminal.

Although this construction shows the existence of a topological realization, it would hardly do as a practical layout in every case, even if indefinite stacking were permitted. Some of the metrical difficulties can be overcome by substituting an annulus for the disk D, but even so, this construction should be regarded as an existence proof, not as a practical solution.

A.2 *Answers to the Examples in Fig. 5.*

The example (a) of Fig. 5, constructed by R. L. Graham, was the first nonrealizable example found. It turns out to be of the type discussed in the text. Its graph is shown in Fig. 14(a). By deleting vertices and combining adjacent vertices it can be reduced to the graph shown in Fig. 14(b), which is a Kuratowski graph with a vertex inserted into each edge. Therefore, the example is nonrealizable. (See discussion subsequent to Fig. 5.)

Example (b) of Fig. 5 has the solution shown in Fig. 15.

A.3 Fig. 16(a) shows a nonrealizable graph which does not contain either of the augmented Kuratowski graphs G_1* or G_2*. The outer ring

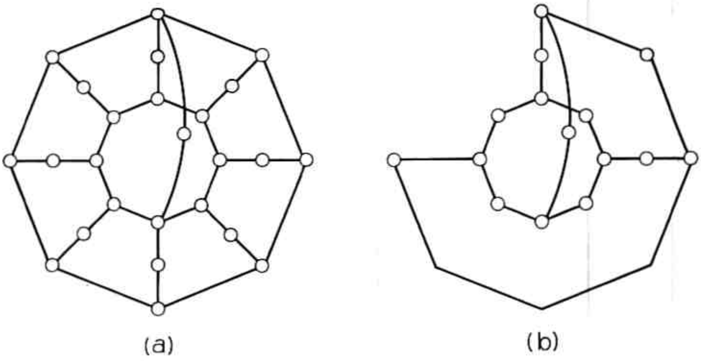


Fig. 14 — (a) Graph for example (a) of Fig. 5. (b) Reduced graph G_1^*.

(B and C vertices) simulates terminal constraints; the inner part (A vertices) is a constrained graph (empty cycle of order 5) that is known to be nonrealizable.

Proof: (*i*) The graph G of Fig. 16(a) cannot be reduced to G_1^* or G_2^*. The operations (*i′*) and (*ii*) always reduce the number of vertices. But G already has the same number of vertices (fifteen) as G_1^* and G_2^*. (*ii*) G is nonrealizable. Suppose a realization exists. In this realization let $\bar{C}$ be the union of C-curves (Fig. 16(b)). No A-curve intersects $\bar{C}$. Therefore all A-curves must lie in the same mesh of $\bar{C}$. Call the interior of this mesh R. R is (or may be) partitioned into subregions by segments of B-curves. We will show that all *intersections* between pairs of A-curves lie within the same subregion of R.

The A-curves may be indexed so that in the cycle $A_1, A_2, \cdots, A_5, A_1$ each curve intersects only its neighbors. Let I be an intersection between

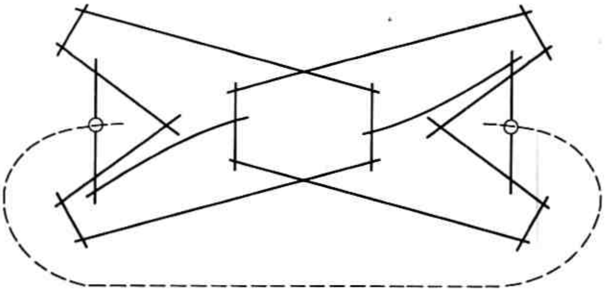


Fig. 15 — Solution to example (b) of Fig. 5. Both triangles can be drawn outside the hexagon.

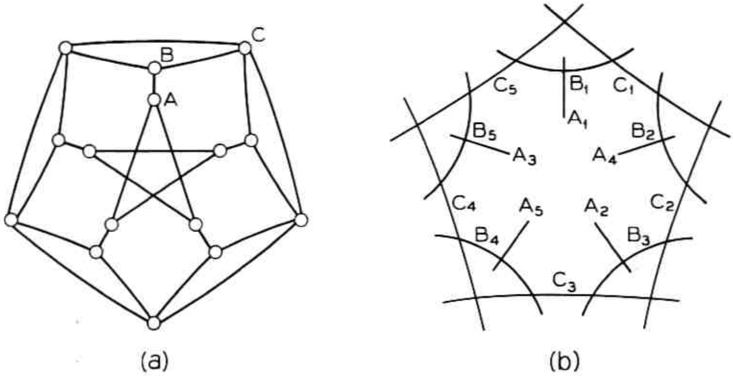


Fig. 16 — Nonrealizable graph which does not contain either of the augmented Kuratowski graphs G_1^* or G_2^* and a partial realization.

A_i and $A_{i+1(\mathrm{mod}\ 5)}$ and let J be an intersection between A_j and $A_{j+1(\mathrm{mod}\ 5)}$. There exist two distinct paths along A-curves joining I and J. One path P_1 traverses segments of $A_{i+1}, A_{i+2}, \cdots, A_j$ and the other path P_2 traverses $A_i, A_{i-1}, \cdots, A_{j+1}$ (indices mod 5). (In case $i = j$, P_1 traverses A_{i+1} and P_2 traverses A_i.) The sets of A-curves represented in the two paths are disjoint. Since each B-curve can cross only one A-curve and cannot cross any other B-curve it is not possible for a continuous boundary made up of B-curves to cross both P_1 and P_2. Therefore, I and J cannot belong to different subregions.

Let R^* be the subregion to which all A-intersections belong. The boundary G of R^* is made up of segments of B and C curves. (Every B-curve is represented since every A-curve must intersect its corresponding B-curve and could leave R^* only at a point belonging to this curve.) If $b_1, b_2, \cdots, b_5$ are any points on G belonging to $B_1, B_2, \cdots, B_5$, respectively (indexed according to the BC cycle), then the points $b_1, \cdots, b_5$ must lie in cyclic order around G. If not, it is possible to find a subset of four out of cyclic order, say b_1, b_3, b_2, b_4. But b_1 is joined to b_2 by a path lying within B_1, C_1, B_2, and b_3 is joined to b_4 by a path lying within B_3, C_3, B_4. These paths cannot cross, yet must be outside R^*. This is not possible under the postulated ordering b_1, b_3, b_2, b_4. All other noncyclic orderings can be similarly ruled out.

The points at which the A-curves join G must, therefore, lie in the order determined by the BC cycle and all intersections between A-curves must lie within R^*. But these are the conditions of a constrained case known (Theorem 1) to be nonrealizable.

A.4 *Answers to Examples in Fig. 8*

Example (a) of Fig. 8 has no solution (empty cycle); example (b) has the solution shown in Fig. 17.

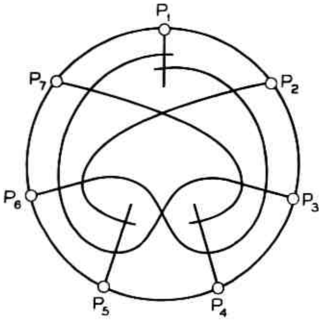


Fig. 17 — Solution to example in Fig. 8 (b). Curves C_3 and C_6 cross thrice. Multiple crossings are unavoidable in this example.

A.5 Counterexample to the conjecture that all constrained graphs free of empty cycles of order four or more are realizable. Fig. 18 shows the graphs G and $\tilde{G}$ for this example and a near-realization in which only one required crossing does not occur.

The lack of empty cycles of order four or more can be verified by inspection; the nonrealizability can be shown as follows.

Consider curves 4 and 8, which do not cross. Since curve 2 crosses both of these, there exists a path from vertex 4 to vertex 8 traversing curves 4, 2, and 8. In case of multiple crossings, there may be more than one such path. We will assume that the path is chosen so that the segment of curve 2 contained in it has no crossings with curves 4 and 8 except at its endpoints. Since this path is to serve as a barrier, we will denote it by B_2.

There exists a similar path traversing curves 4, 6, and 8. We will call this one B_6.

Since curves 2 and 6 do not cross, the barriers B_2 and B_6 can have no points in common except along a single segment of curve 4 and a single segment of curve 8. Thus, the barriers must be related to each other in one of two ways shown in Fig. 19.

Curve 1 cannot cross B_2 and curve 5 cannot cross B_6. Thus, the barriers cannot be oriented as in case (a) of Fig. 19, for if they were curve 1 could not cross curve 5. By a similar argument, case (b) is eliminated by curves 3 and 7. Thus, neither case can occur; the example is nonrealizable.

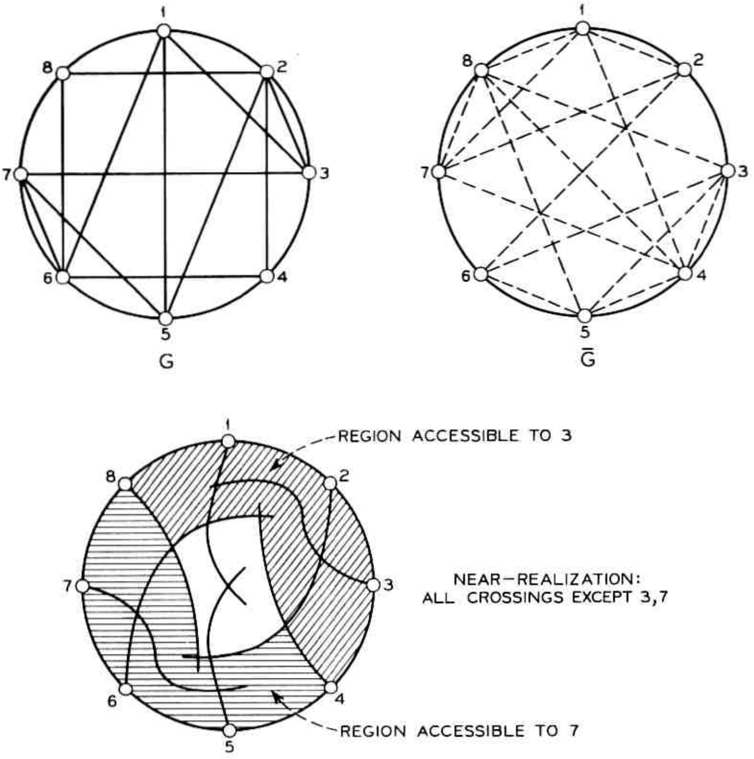


Fig. 18 — Counterexample. $\bar{G}$ contains no empty cycles of order four or more, yet G is not realizable.

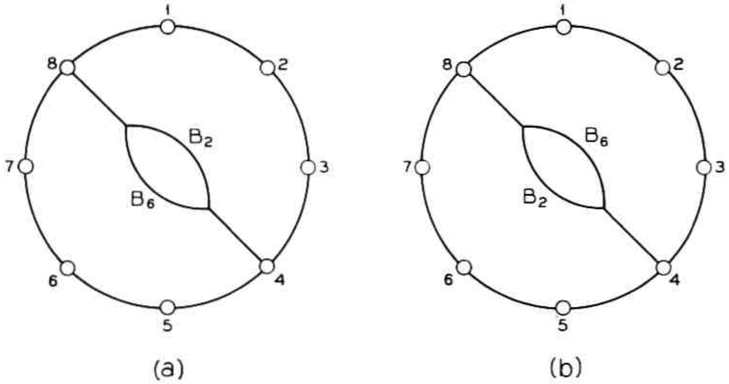


Fig. 19 — Proof of counterexample.

A.6 Fig. 20 shows the realization for an example to which neither Theorem 4 nor Theorem 5 applies. The ordering information supplied by Theorem 6 is very complete in this case. Only the order of curves 5 and 6 along curve 2 (and the symmetric counterparts) is unspecified. Indeed this could not be specified since either order is feasible. The order 6, 5 however, requires multiple crossings. The realization without multiple crossings is unique.

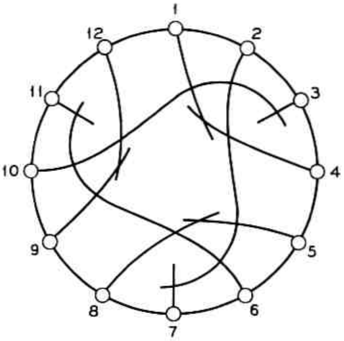


Fig. 20 — Realization for an example to which neither Theorem 4 nor 5 applies.

APPENDIX B

Proofs of Theorems

Theorem 2: A sufficient condition for a constrained graph G to be realizable is that $\bar{G}$ contain no empty cycles of order four or more and that no two edges of $\bar{G}$ cross.

Proof: The following proof depends on Theorems 3 and 5 whose proofs are independent.

The theorem is certainly true if G has three or fewer vertices. Suppose it is known to be true if G has m or fewer vertices. Consider a graph G with $m + 1$ vertices. If G satisfies the hypotheses of the theorem, then either $\bar{G}$ is an empty chain (see discussion preceding Theorem 6) or else $\bar{G}$ has an internal edge. If G is an empty chain, then by Theorem 5 it is realizable. If $\bar{G}$ has an internal edge then this edge separates $\bar{G}$ into two parts as defined in Theorem 3. Each of these parts has m or fewer vertices and is free of crossing edges and empty cycles of order four or more (by hypothesis), hence by the induction assumption is realizable. By Theorem 3, G is realizable.

Theorem 3: If (P_1, P_k) is an edge of G ($\bar{G}$) that crosses no other edge of G ($\bar{G}$), and if the subgraphs G' with verticles $P_1, P_2, \cdots, P_k$ and G'' with vertices $P_k, \cdots, P_r, P_1$ are both realizable, then G is realizable.

Proof: The method of proof is proof by picture (Fig. 21). *Case (a):* (P_1, P_k) is an edge of G crossing no other edges of G. None of the curves $C_2, \cdots, C_{k-1}$ crosses any of the curves $C_{k+1}, \cdots, C_r$. Therefore, except for C_1 and C_k the realizations of G' and G'' can be confined to separate parts of the disk R. C_1 and C_k can participate in both parts. (See Fig. 21(a).) *Case b:* (P_1, P_k) is an edge of $\bar{G}$ crossing no other edges of $\bar{G}$. Every one of the curves $C_2, \cdots, C_{k-1}$ crosses every one of the curves $C_{k+1}, \cdots, C_r$. The realizations of G' and G'' can be confined to the regions labelled with these letters in Fig. 21(b). The peripheral terminals for these realizations can be connected to the terminals on the periphery of the disk as shown in the figure. The connections to G' can cross G'''s region since this can only generate allowable crossings. The required crossings between curves of G' and curves of G'' occur in the center of the figure.

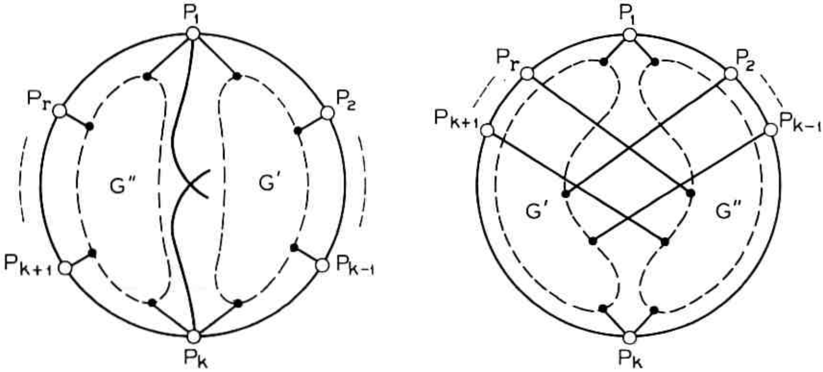


Fig. 21 — Proof of Theorem 3.

Theorem 4: Let G be a constrained graph with vertices $P_1, \cdots, P_r, P_{r+1}$. G is realizable if (i) the subgraph of G with vertices $P_1, \cdots, P_r$ is realizable, and (ii) there do not exist three vertices $P_i, P_j, P_k, i < j < k < r + 1$ such that $P_{r+1}P_i$ and $P_{r+1}P_k$ are edges of $\bar{G}$ and P_iP_k and $P_{r+1}P_j$ are edges of G. (See Fig. 10.)

Proof: We suppose that a realization for the subgraph with vertices $P_1, \cdots, P_r$ is at hand. It will be convenient to think of this realization as made up of regions $R_1, \cdots, R_r$ instead of curves. To simplify the notation later on we will designate the disk R to which the realization is confined by the indexed name R_0. We may assume that R_i intersects the boundary of R_0 only in the vertex P_i.

Let R^* be that connected piece of R_0 which contains P_{r+1} but is exterior to the regions that R_{r+1} may not intersect. R^* is the set of points that can be reached by R_{r+1}. We will show that the boundary of R^* contains all the vertices corresponding to regions R_{r+1} must intersect, i.e., all vertices joined to P_{r+1} by edges of G.

The boundary of R^* can be partitioned into a sequence of segments $S_1, \cdots, S_n$ where S_i belongs to the boundary of $R_{k(i)}$ and $k(i) \neq k(i+1)$. The segments S_1 and S_n adjacent to P_{r+1} belong to R_0, hence $k(1) = k(n) = 0$. If $k(i) = 0$, $1 < i < n$, then S_i is that segment of the boundary of the disk R_0 which runs from $P_{k(i-1)}$ to $P_{k(i+1)}$. (The end cases $i = 1$ and $i = n$ can be included by defining $k(0) = k(n+1) = r + 1$.)

Now suppose $P_{r+1}\, P_j$ is an edge of G (i.e., R_{r+1} must intersect R_j). We will show that P_j belongs to the boundary of R^*.

If $j > k(i)$, $i = 1, \cdots, n$, then P_j belongs to S_n, hence to the boundary of R^*. If not, let i be the first index such that $j < k(i+1)$. It is not possible that $k(i) = j$ because R_j as a region that intersects R_{r+1} is not involved in the boundary of R^*. It is also not possible that $0 < k(i) < j$ for this would violate hypothesis (*ii*). ($R_{k(i)}$ intersects $R_{k(i+1)}$ because segments of their boundaries are adjacent.) Therefore, $k(i) = 0$. Hence, S_i runs from $P_{k(i-1)}$ to $P_{k(i+1)}$. Since $k(i-1) < j < k(i+1)$, S_i must contain P_j. Therefore, P_j is on the boundary of R^*, which was to be proved.

Theorem 5: The intersection requirements of a curve C_i can be satisfied entirely by a sequence of adjacent interchanges if and only if P_i is not an active member of any cycle in $\bar{G}$.

Proof: If: A *chain* is a sequence of vertices $P_{i_1}, P_{i_2}, \cdots, P_{i_n}$ in cyclic order such that (P_{i_1}, P_{i_2}), (P_{i_2}, P_{i_3}), $\cdots$, $(P_{i_{n-1}}, P_{i_n})$ are edges of $\bar{G}$. For the duration of this proof a chain must have at least three vertices.

Let the vertices be numbered in clockwise order and suppose P_1 is not an active member of any cycle in $\bar{G}$. Let S be the set of vertices joined to P_1 by edges of G. We will show that by a sequence of adjacent interchanges the members of S can be moved around the circle and finally interchanged with P_1.

S can be divided into three subsets:

(*i*) The clockwise set S_c : $P_k \varepsilon S_c$ if P_1 is joined to P_k by a chain whose intermediate members have indices between 1 and k.

(*ii*) The counterclockwise set S_{cc} : $P_k \varepsilon S_{cc}$ if P_1 is joined to P_k by a chain whose intermediate members have indices greater than k.

(*iii*) The rest S_R .

S_c and S_{cc} must be disjoint because otherwise P_1 would be an active member of a cycle. Let P_i be that member of S_c with highest index. P_i can be interchanged with all vertices with higher indices. Thus, it can be moved clockwise around the circle past P_1 . With P_i out of the way, the member of S_c with next highest index can also be moved clockwise past P_1 . The process can continue until all members of S_c have been interchanged with P_1 . Similarly, the members of S_{cc} can be moved counterclockwise past P_1 . The members of S_R can be moved either way. Hence, every member of S can be interchanged with P_1 , which was to be proved.

Only if: Suppose P_1 is an active member of a cycle. Then it is joined to another member P_k by an edge of G. P_1 cannot be brought adjacent to P_k because the order of vertices in a cycle is invariant under adjacent interchanges. Hence, P_1 cannot be interchanged with P_k .

Theorem 6: Let $P_1, \cdots, P_n$ *be the vertices of an empty chain. Along curve* C_k *the first crossings with* $C_1, \cdots, C_{k-2}$ *must occur in that order; the first crossings with* $C_{k+2}, \cdots, C_n$ *must occur in reverse order.*

Proof: It is only necessary to prove the first part of the statement (concerning $C_1, \cdots, C_{k-2}$) since the second part follows from the first by symmetry. The first part is trivially true if $k \leqq 3$. We assume then that $k > 3$.

The region bounded by C_{k-1} and C_{k-3} encloses C_{k-2} . (See Fig. 22.) Since C_k cannot cross C_{k-1} it must cross C_{k-3} before it can cross C_{k-2} .

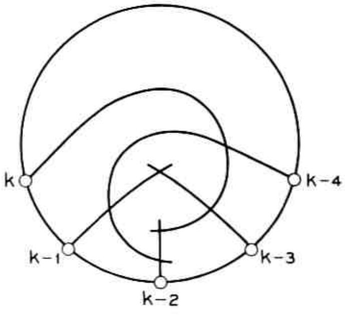


Fig. 22 — Proof of Theorem 6.

If $k > 4$, then there is a curve C_{k-4}. The region bounded by C_{k-2} and C_{k-4} encloses C_{k-3}. Therefore, C_k must cross either C_{k-2} or C_{k-4} before it can cross C_{k-3}. But by the previous argument it cannot cross C_{k-2} before C_{k-3}. Therefore, it must cross C_{k-4} before C_{k-3}. Since this argument can be iterated indefinitely, the theorem holds for arbitrary k.

REFERENCES

1. Feldman, D., private communication.
2. Berge, C., *Theory of Graphs and its Applications*, John Wiley & Sons, New York, 1962.
3. Balbanian, N., *Network Synthesis*, Prentice-Hall, 1958.

Contributors To This Issue

Václav E. Beneš, A.B., 1950, Harvard College; M.A. and 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. He is the author of General Stochastic Process in the Theory of Queues (Addison-Wesley, 1963), and of Mathematical Theory of Connecting Networks and Telephone Traffic (Academic Press, 1965). Member, American Mathematical Society, Association for Symbolic Logic, Institute of Mathematical Statistics, SIAM, Mind Association, Phi Beta Kappa.

James L. Flanagan, B.S., 1948, Mississippi State University; S.M., 1950 and Sc.D., 1955, Massachusetts Institute of Technology; faculty, Mississippi State University, 1950–52; Rockefeller Foundation Fellow, 1952–53; Air Force Cambridge Research Center, 1954–57; Bell Telephone Laboratories, 1957—. Mr. Flanagan has specialized in signal coding and speech communication over narrow bandwidths, including studies of acoustical, physiological, and psychophysical phenomena related to speech and hearing. Since 1961, he has headed the Speech and Auditory Research Department. Fellow, Acoustical Society of America; member, IEEE, Sigma Xi, Tau Beta Pi, Committee on Hearing and Bioacoustics of the National Academy of Sciences.

M. Gershenzon, B.S., 1949, City College of New York, A.M., 1953, Ph.D., 1957, Columbia University; Bell Telephone Laboratories, 1957—. Mr. Gershenzon has worked on radiative recombination in semiconductors and on physical and chemical properties of gallium phosphide. Currently, he is on leave of absence at the Department of Electrical Engineering and Materials Science, University of Southern California, Los Angeles, California.

Roger M. Golden, B.S.E.E., 1954, M.S.E.E., 1955, Ph.D., 1959, California Institute of Technology; Fulbright student Technical Institute at Eindhoven, Netherlands, 1959–60; Bell Telephone Laboratories, 1960–66. While at Bell Laboratories, Mr. Golden has worked on speech bandwidth compression systems and speech analysis-synthesis systems for telephone communications. He has developed new techniques for the simulation of signal processing systems on digital computers. In November of this year, Mr. Golden will return to California where he will join the research and engineering staff of Autonetics, Division of North American Aviation. Member, Acoustical Society of America, IEEE, Association for Computing Machinery, Sigma Xi, Tau Beta Pi.

Ronald L. Graham, B.S., 1958, University of Alaska; M.A., Ph.D., 1962, University of California (Berkeley); Bell Telephone Laboratories, 1962—. Mr. Graham has been engaged in research into a variety of problems arising in coding theory, graph theory and combinatorial geometry. Member, American Mathematical Society, Mathematical Association of America, Sigma Xi.

Ming-Lei Liou, B.S.E.E., 1956, National Taiwan University; M.S.E.E., 1961, Drexel Institute of Technology; Ph.D., 1964, Stanford University; Bell Telephone Laboratories, 1963—. Mr. Liou has been working on various kinds of system analyses and characterization. His current interests lie in the area of computational analysis of linear and nonlinear systems. He is presently Supervisor, Transmission Studies Group in the Transmission Technology Laboratory. Member, Eta Kappa Nu, Sigma Xi, IEEE.

James E. Mazo, B.S., 1958, Massachusetts Institute of Technology; M.S., 1960, and Ph.D., 1963, Syracuse University; Research Associate, University of Indiana, 1963–64; Bell Telephone Laboratories, 1964—. At Indiana University, Mr. Mazo was engaged in work on quantum scattering theory. At present, he is engaged in theoretical analysis of data systems. Member, American Physical Society, IEEE, Sigma XI.

Joseph T. Ruscio, B.S., 1962, Monmouth College; Bell Telephone Laboratories, 1957—. Mr. Ruscio has been engaged in the Echo and *Telstar*® satellite communication projects, light modulation, and other phases of laser technology. He is currently involved in light propagation experiments at the Crawford Hill Laboratory.

J. Salz, B.S.E.E., 1955, M.S.E., 1956, Ph.D., 1961, University of Florida; the Martin Company, 1958–60; Bell Telephone Laboratories, 1961—. Mr. Salz first worked on the remote line concentrators for the electronic switching system. He has since engaged in theoretical studies of data transmission systems. Member, IEEE; associate member, Sigma Xi.

IRWIN W. SANDBERG, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—. Mr. Sandberg has been concerned with analysis of military systems, particularly radar systems, with synthesis and analysis of active and time-varying networks, and with studies of the signal-theoretic properties of nonlinear systems. Member, IEEE, SIAM, Eta Kappa Nu, Sigma Xi, Tau Beta Pi.

M. V. SCHNEIDER, M.S., 1956, and Ph.D., 1959, Swiss Federal Institute of Technology, Zurich, Switzerland; Bell Telephone Laboratories, 1962—. Mr. Schneider has been engaged in experimental work on microwave solid-state devices and optical detectors. Member, Optical Society of America, IEEE, American Vacuum Society.

FRANK W. SINDEN, B.S., 1948, University of Chicago; Dr. Sci. Math., Swiss Federal Institute of Technology, 1954; Bell Telephone Laboratories, 1956—. Mr. Sinden has worked mainly in operations research and mathematical programming. Member, AMS, MAA; Editor, SIAM Review.

JURGEN H. W. UNGER, Dipl. Eng., 1951, Technische Hochschule (Institute of Technology) Karlsruhe, Germany; Bell Telephone Laboratories, 1962—. Mr. Unger has been engaged in the development of guidance and control systems for spacecraft, and in research on the effects of random errors in control systems. His present work involves atmospheric effects in microwave, radar, and optical systems. Member, IEEE, American Institute of Aeronautics and Astronautics.

B.S.T.J. BRIEFS

Realizability Conditions for the Impedance Function of the Lossless Tapered Transmission Line

By P. L. ZADOR

(Manuscript received August 2, 1966)

In the study of tapered transmission lines or accoustical horns, an unsolved problem of great practical interest is the determination of the taper function (inductance or capacitance per unit length as a function of distance; it is assumed that the product of these quantities is unity) for the structure which will possess a prescribed driving point impedance function. For the case where the structure may be modeled by a cascade of sections of *uniform* transmission line segments, physical realizability conditions and a synthesis procedure have been given by B. K. Kinariwala.[1] For the case of continuous taper no results of a general nature are known.

In this note, we shall give an almost complete characterization of driving point impedances for structures possessing once continuously differentiable taper functions. Although the proof of the realizability theorem will not be given here, the author wants to point out that the sufficiency is, in fact, proved by a construction of the taper function. However, this construction is too unwieldly to be of practical use.

The mathematical formulation of the problem is as follows.

Suppose that for all complex s $y(x,s)$, $0 \leqq x \leqq l$ is the solution of the Horn equation

$$(c(x)y'(x,s))' = s^2c(x)y(x,s) \tag{1}$$

satisfying the boundary condition

$$y(0,s) = -i, \qquad y'(0,s) = \frac{is}{c(0)}^{*}.$$

If by driving point impedance we mean the function

$$Z(s) = -\frac{s}{c(l)}\,\frac{y(l,s)}{y'(l,s)},$$

then the following theorem is true.

Necessity: If $c(x)$ is a positive real function continuously differentiable on $0 \leqq x \leqq l$ then

* Unit terminating resistance at zero is assumed.

(*i*) $Z(s)$ is a positive real function.

(*ii*) There exist entire functions of the exponential type* l, $N_i(s)$, $D_i(s)$, $i = 1, 2$, such that

$$(a)\ Z(s) = \frac{N_1(s) + N_2(s)}{D_1(s) + D_2(s)},$$

$$(b)\ N_1D_1 - N_2D_2 \equiv e^{2ls},$$

$$(c)\ N_1(s) = N_1(-s),\ D_1(-s) = D_1(s)$$
$$N_2(s) = -N_2(-s),\ D_2(s) = -D_2(-s).$$

(*iii*) If for real ω

$$f(\omega) = \operatorname{Re} e^{-2il\omega} Z(i\omega) \quad \text{or} \quad \operatorname{Re} e^{-2il\omega} \frac{1}{Z(i\omega)}$$

then the function $f(\omega)$ has an asymptotic expansion† at $\pm\infty$ of the following kind

$$f(\omega) \approx 1 + \frac{a}{\omega^2} + \frac{b}{\omega^4} + \frac{c}{\omega^6} + 0\left(\frac{1}{\omega^7}\right)$$

(The constants of course may be different).

Sufficiency: In order that a complex function $Z(s)$ be the driving point impedance of the differential equation (1) for some continuously differentiable positive taper function $C(x)$ it is sufficient that

(i') $Z(s)$ be positive real,

(ii') there exist complex functions $N_i(s)$, $D_i(s)$, $i = 1, 2$ of the exponential type satisfying (2) (*a*), (*b*), (*c*), and

(iii') the function $f(\omega)$ defined in (3) have asymptotic expansion at infinity

$$f(\omega) \approx 1 + \frac{a}{\omega^2} + \frac{b}{\omega^4} + \frac{c}{\omega^6} + \frac{d}{\omega^8} + 0\left(\frac{1}{\omega^9}\right).$$

Remarks: (*i*) It is conjectured that the existence of the two asymptotic expansions are not independent, that is (i'), (ii'), and one expansion may be sufficient.

(*ii*) A similar result is probably valid for infinite transmission lines.

* The function $h(s)$ is called exponential type l if $e^{-lr}M(r)$ remains bounded for all $r > 0$ but for any $l' < l$ $e^{-l'r}M(r)$ grows to infinity. Here $M(r) = \max_{|s| \leqq r} |h(s)|$.

† This means that $\lim_{\omega=\pm\infty} \omega^2 (f(\omega) - 1) = a$,

$$\lim_{\omega=\pm\infty} \omega^4 \left(f(\omega) - 1 - \frac{a}{\omega^2}\right) = b, \text{ etc.}$$

Substituting the words "functions of order unity"[2] for "functions of the exponential type l" should yield the correct theorem.

(*iii*) As a last conjecture we offer the following. Let $f(\omega)$ be a positive even function possessing the properties

(*a*) $f(\omega)$ has an asymptotic expansion at infinity as in (*iii'*).

(*b*) The Fourier transform of $1/f(\omega) - 1$ vanishes outside the interval $(-2l, 2l)$.

Then there exists a unique taper function such that if $Z(s)$ is the impedance function of the associated differential equation (1) then

$$\operatorname{Re} Z(i\omega)e^{-2i\omega l} = \frac{1}{f(\omega)}.$$

REFERENCES

1. Kinariwala, B. K., Theory of Cascaded Structures: Lossless Transmission Lines, B.S.T.J., *45*, April, 1966, pp. 631–650.
2. Titchmarsh, *Theory of complex functions*, Oxford Univ. Press, 1937.