THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

Volume 56 October 1977 Number 8

Anticausal Analysis of Feedback Amplifiers

By F. D. WALDHAUER

(Manuscript received February 11, 1977)

This paper discloses a technique for the direct analysis of linear active circuits, avoiding the solution of simultaneous equations. This is done by representing the circuit in such a way that the signal variables (currents and voltages) are determined sequentially: we only allow a signal variable to depend upon previously determined signal variables, not upon signal variables yet to be determined. Such a circuit is representable by a cascade signal flow graph, a graph containing no feedback loops. Not all circuits can be so represented, of course, but the number which can is expanded by the technique to be described to include most feedback amplifier configurations. This simplification in linear amplifier analysis allows us to trace a clear path from rough design approximations to exact computer analysis. The extension of the analysis to include the effect of nonsaturating nonlinearities is indicated but not developed here.

I. INTRODUCTION

Feedback regulators as human artifacts have been here for a long time. An early one (perhaps the first[1]) was a furnace temperature control invented by Cornelius Drebble (1572–1633) who used it in several versions including an incubator for chickens. The flyball governor may have originated with Huygens[2] in the seventeenth century, and was used for speed control of windmills by Thomas Mead and steam engines by James Watt, both in the early nineteenth century. In the same period, a much more diffuse feedback system was promulgated by Adam Smith in his

Wealth of Nations, which proposed that economic self-interest of individuals would automatically assure equilibrium of the economic system, without central control.[1]

Mathematical development of governors began with Maxwell,[3] who determined stability conditions for systems up to the third degree. He "hoped that the subject would obtain the attention of mathematicians." Routh, Lyapunov, Hurwitz, and others responded, extending the stability analysis to systems of higher degree.[4] Still the focus was on stability, that is, preventing the system from being useless. The engineer was pretty much on his own to make the system useful. Minorski was such an engineer. He developed an analysis for the *design* of a ship rudder servo in the early 1920s.[5]

In the same period, Black and Dickieson were working together on amplifiers for carrier transmission of telephone signals. Their *design* problem was to reduce nonlinearities in electronic amplifiers so that the several voice channels would not interfere with one another by modulation. Black's first entry in this area was his invention of feedforward,[6] a technique reinvented by many workers in the 1960s, and inspired, as Black relates, by "an approach to another problem, I don't remember what it was, in a lecture by Steinmetz." Black worked out the invention on the night of the lecture, and he and Dickieson got it working in six hours the next day.[7] The second invention was that of feedback,[8,9] which came out of the first invention in the sense that Black understood that it would do the same job of reducing nonlinearity. An appreciation of stability problems came later. Nyquist, with his paper "Regeneration Theory"[10] (unfortunately titled, according to Black), dealt with stability analysis or preventing oscillations—making sure that Black's invention would work. Dickieson has been quoted as saying about this theory of stability, "At last we knew what we were trying to achieve."[11] Bode later set down the theory of feedback amplifier design, which remains a landmark to this day.[12]

In the middle decades of this century, feedback amplifiers received much attention.[13,14] A recent library search turned up some 750 articles on the subject, indicating that the theory is hard to understand. By making the stability problem the central focus, and in solving it superbly well, Nyquist and Bode relegated the *design* problem to a position of lesser importance. What was the design problem? To reduce modulation products in frequency-division multiplex systems. What was the solution? To maximize the magnitude of the feedback signal, consistent with the stability constraint.

This paper questions the usefulness of feedback as a conceptual tool for design.[15] The physical connection of a portion of the output signal of an amplifier to the input is agreed to be a beneficial measure for many applications. The *analysis* of such a physical structure can be made

without recourse to the concept of feedback by a conceptual leap, one which has already been made by many engineers who design circuits. That leap is to reverse the direction of time, and think in terms of the input signal which is required to produce a given, preassigned output signal. The technique is old, having been applied to passive ladder circuits by many workers, although the origin is unclear.[16] Penfield refers to it as the "Guillemin trick,"[17] but many others of the era used it. Its application to active circuits has been much more limited, consisting of a few papers by the present author,[18] Beddoe,[19] and in control theory, by Rosenbrock.[20]

People working in computer-aided design have already rejected the concept of feedback in favor of general circuit analysis programs that calculate the performance of quite complex circuits by various matrix methods. These programs are most valuable in checking the performance of a circuit after it has been designed and before it is committed to production. They tend to be neutral with respect to circuit concepts, giving mostly correct answers as to how circuits, previously given to them by design engineers, will work. When a circuit doesn't work, the design engineer has difficulty tracing the source of the difficulty from the computer results. The CAD expert, on the other hand, complains that he is not brought into the design process early enough. The design, according to him, has been set in concrete. The problem is sometimes cast in terms of interpersonal relations, but I think that it is structural, in that there is a poor match between the intuitive thought process of the design engineer and the general analysis method of the CAD expert. The design method discussed in this paper should help to resolve this question, since it is at home as much on the computer as it is in the mind (potentially) of the designer.

The focus of this paper is on the *design* problem of feedback amplifiers. Sections II and III are tutorial, because the material is old, and may be unfamiliar to many who might like to understand the rest of the paper. Sections IV and V describe the new theory, and Sections VI and VII are concerned with applying it to familiar problems. Section VIII considers the stability question. While the substance of this paper is theoretical, it was derived from practical design experience with several amplifier configurations, the most recent of which is an operational amplifier with 1-GHz unity-gain bandwidth and 1 volt per nanosecond slew rate, to be reported upon later. The conceptual difficulties were discussed quite thoroughly in an in-hours course taught by the author at Bell Labs.

II. CAUSAL AND ANTICAUSAL ANALYSIS—SINGLE SIGNAL VARIABLES REPRESENTING CAUSE AND EFFECT

Two elementary examples will serve to define what is meant by feedback and its relationship to the choice of independent circuit vari-

ables. In Fig. 1a, a Thevenin source is connected to a load conductance: a first equation is written taking the cause, e_G, as the independent variable, giving rise to a loop gain or return ratio, $-G_L R_G$, and a return difference, $F = 1 + G_L R_G$, as shown in the signal flow graph.[21] A second equation is written taking the effect, v_o, as the independent variable, and the cause, e_G, as the dependent variable, in which case no loop gain appears, giving unity return difference. In Fig. 1b, the elementary feedback amplifier circuit of Black's feedback patent is shown with a similar set of causal and anticausal equations, showing again that loop gain does not appear under the anticausal formulation. Clearly, then, return ratio is a property of the mathematical description of the circuits, and not of the circuits themselves.

Feedback is seen to be associated with the departure from unity of the denominator in the circuit equation. Since circuit expressions are easier to evaluate (and think about) without denominators, a circuit description which avoids them is conceptually easier to deal with. In general circuit analysis, denominators (or return difference) cannot always be unity, of course, but in many active circuits it will be shown that they can be made to approximate it by appropriate choice of independent variables.

The word feedback is generally employed in a broader sense than Bode's strict definition of it as return difference. It connotes coupling from output to input of an active circuit, or portion of a circuit, and in this sense can exist, as in Figure 1b with its anticausal equation, without any loop gain. We shall employ the term feedback in this sense even though the description may include no return ratio.

H. Seidel, whose work on feedforward has been of substantial help to the author in clarifying amplifier input-output time relationships, has pointed out that the title of this paper might be interpreted (incorrectly) as describing a physical violation of the principle of causality. No such violation should be inferred. Rather, it is the *analysis* of the causal physical system, proceeding from output to input in a direction from effect to cause, which gives rise to the title of this paper.

III. TWO-PORT ANALYSIS USING THE TRANSMISSION MATRIX AND TRANSMISSION MATRIX SIGNAL FLOW GRAPHS

Much useful theory is based on single signal-variable analysis, including some introductory control theory and circuit analysis. For practical circuit work, however, we need to consider at least two signal variables in order to given an adequate, simple description of an amplifier made up from basic parts, such as transistors and passive devices. The simplest of such amplifiers will have an input port and an output port, and we are concerned with the current *and* voltage at each of these ports, four variables in total. The most general way to assign indepen-

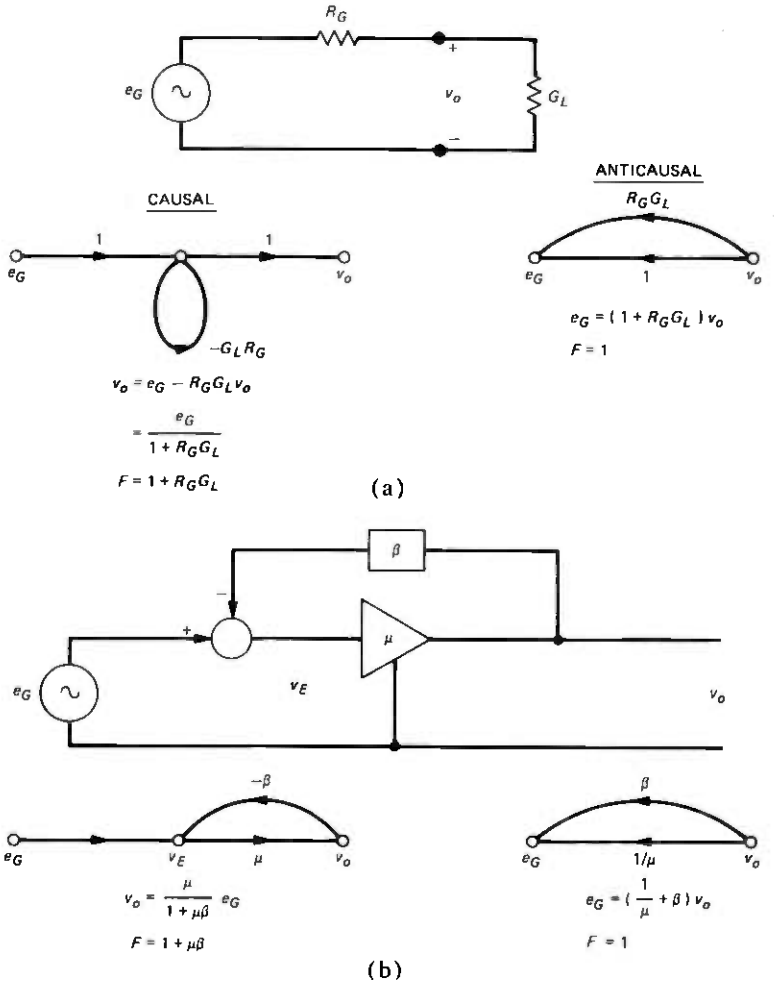


Fig. 1—One-dimensional analysis showing causal and anticausal functional dependencies.

dence and dependence to these four variables, necessary in most cases, is to choose two *independent* variables and two *dependent* variables. There are six possible assignments of port currents and voltages as independent and dependent: one is to choose the port voltages as the set of independent variables. The port currents, then, are the dependent variables, related to the port voltages by an admittance, or y matrix. The choice can profoundly affect the nature of the analysis of the amplifier. In what follows, we use five of the six choices as it suits the occasion, but the basic analysis is involved with the choice of the output current and voltage, the *output signal vector*, $u_o = v_o, i_i$, as the set of independent

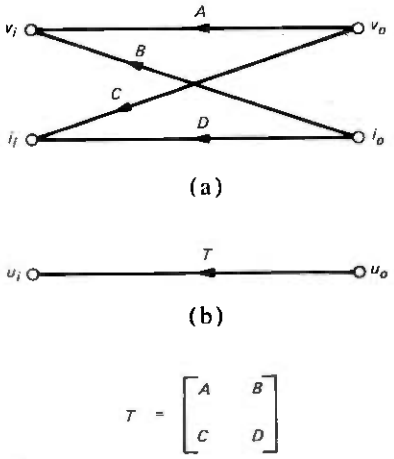


Fig. 2—Signal flow graph of an amplifier represented by its transmission parameters, and the equivalent transmission matrix signal-flow graph (TMSFG).

variables, and the *input signal vector* $u_i = v_i, i_i$, as the set of dependent variables. The dependent variables are related to the independent variables by the *transmission matrix*, or ABCD matrix:[22,23]

$$\begin{bmatrix} v_i \\ i_i \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} v_o \\ i_o \end{bmatrix} \tag{1}$$

or

$$u_i = Tu_o$$

These equations are shown in signal flow graph form in Fig. 2. In Fig. 2a, eq. (1) is represented by the usual signal flow graph, a graph of directed branches. For each branch, the tail originates at the independent circuit variable, and the nose points toward the dependent variable. The branch value multiplies the value of the independent variable at the tail, and adds the result to the dependent variable value at the nose.

Signal flow graphs are particularly useful in establishing and clarifying functional dependencies in circuits. They are not widely used in circuit analysis and design, however, because of their complexity, even in circuits of quite modest proportions.

In Fig. 2b, a simpler graph, a *transmission matrix signal flow graph* (TMSFG)* connects the output signal vector, u_o, to the input signal vector, u_i, through the matrix branch T. The TMSFG is simply a shorthand way of depicting the graph of Fig. 2a. While signal flow graphs having matrices for the branches were envisioned by Mason[24] and have been studied elsewhere,[25] the application to transmission matrices is new.

* A glossary of terms is given at the end of the paper.

In the transmission matrix signal flow graph, each graph node represents a signal vector consisting of a current and voltage at some point in a circuit, and each branch represents a transmission matrix. The correspondence between the graph and the circuit is direct, with the graph nodes having a direct counterpart in *vector nodes* of the circuit. A *circuit vector node* is defined as a node of the circuit with *only two connections* to it, allowing us to define uniquely the node voltage (to ground) and the node current, which together form the signal vector of the corresponding TMSFG node. The definitions of the transmission parameters are implicit in eq. (1):

$$
\begin{aligned}
A &= \frac{\partial v_i}{\partial v_o}, && \text{the reciprocal of } g_{21}\text{, the open circuit voltage gain} \\
B &= \frac{\partial v_i}{\partial i_o}, && \text{the negative reciprocal of } y_{21}\text{, the short circuit transadmittance} \\
C &= \frac{\partial i_i}{\partial v_o}, && \text{the reciprocal of } z_{21}\text{, the open circuit transimpedance} \\
D &= \frac{\partial i_i}{\partial i_o}, && \text{the negative reciprocal of } h_{21}\text{, the short circuit current gain}
\end{aligned}
\tag{2}
$$

Note that the *ABCD* parameters are all *reciprocals* or negative* reciprocals of familiar forward transfer or gain parameters.

Equation (1) can be written with the transmission matrix taken as a Jacobian matrix, making the equation suitable for analysis of an important class of nonlinear problems:

$$
\begin{bmatrix} dv_i \\ di_i \end{bmatrix} = \begin{bmatrix} \dfrac{\partial v_i}{\partial v_o} & \dfrac{\partial v_i}{\partial i_o} \\ \dfrac{\partial i_i}{\partial v_o} & \dfrac{\partial i_i}{\partial i_o} \end{bmatrix} \begin{bmatrix} dv_o \\ di_o \end{bmatrix} \tag{3}
$$

The partial derivatives can be expressed as nonlinear functions of the instantaneous output current and voltage, allowing us to find the input voltage and current as nonlinear functions of a preassigned output voltage and current. For a desired sinusoidal output, for example, we can find the input predistortion required to achieve that output. The study of transistor nonlinearities expressed in terms of the partials of eq. (3) is beyond the scope of this paper, and is mentioned here to indicate the

* The parameters which involve i_o are *negative* reciprocals because of differing sign conventions between the *ABCD* parameters, in which the positive direction of current is taken to be outward from the output port, and the *h*, *z*, *y*, and *g* parameters, in which the reverse is true.

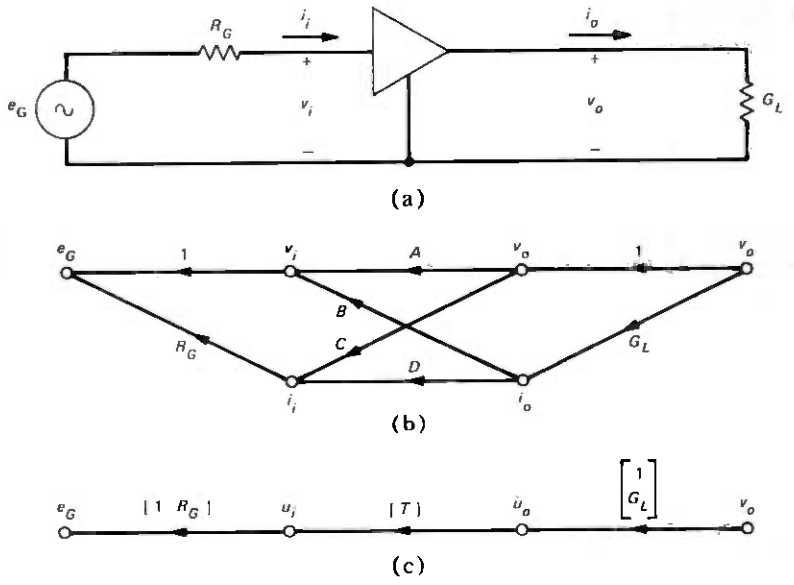


Fig. 3—Connection of a two-port amplifier to source and load, with appropriate signal-flow graphs.

direction of future efforts. For the remainder of this paper, we shall confine our attention to the small-signal case, where the partial derivatives are constants defined at a dc operating point, and are generally functions of frequency.

The calculation of the circuit properties of an amplifier connected as shown in Fig. 3a between a Thevenin source and load conductance is particularly simple if we retain the anticausal direction of analysis that finds the input for a given output. Thus, defining the *loss ratio, L,* as e_G/v_o, we simply add all of the paths from v_o to e_G in Fig. 3b, or, alternatively, perform the matrix multiplication indicated in Fig. 3c. Thus,

$$L = [1 \; R_G] \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} 1 \\ G_L \end{bmatrix} \tag{4}$$

$$= A + BG_L + R_G C + R_G D G_L \tag{5}$$

The graphs of Fig. 3 do not include any feedback loops (closed paths). Such a graph is termed a *cascade graph* and has the property that the graph gain (in this case representing the *loss ratio* since the graph source node corresponds to the circuit output) is the sum of all path products from the graph source node to the sink node, from v_o to e_G. With no feedback loops, no denominator appears in the expression for the loss ratio.

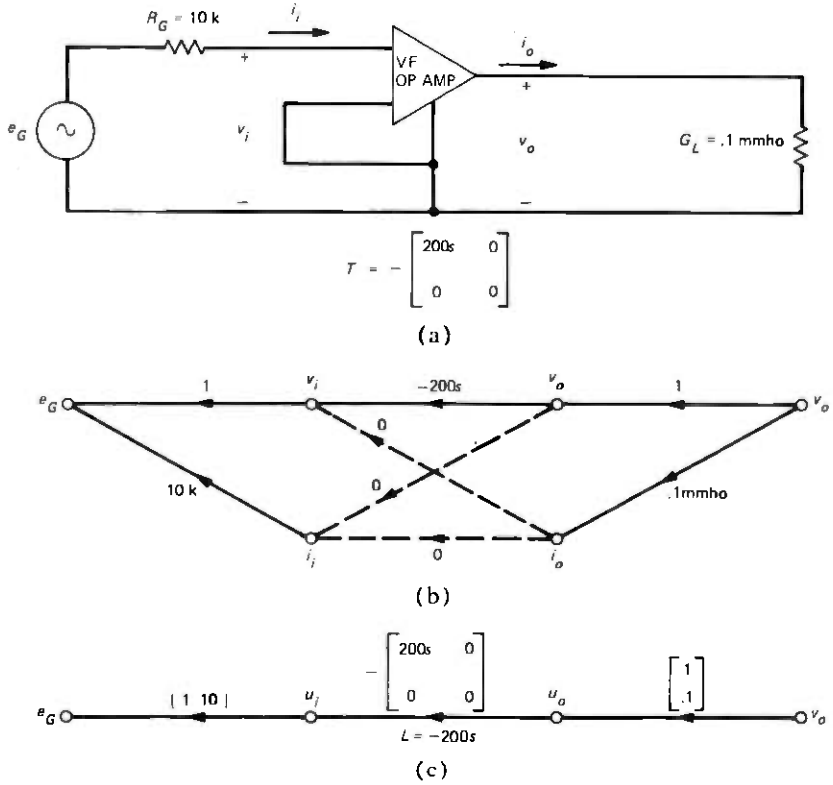


Fig. 4—Loss ratio calculation for a voice-frequency operational amplifier connected between source and load.

Two examples serve to illustrate the loss ratio calculation. In Fig. 4, a voice-frequency operational amplifier is connected between source and load, with the positive input grounded. The transmission matrix of this amplifier can be approximated over most of its frequency range by a single parameter: $A = -200s$, where s is the frequency variable in units of gigaradians per second.* (The low frequency value of input signal is over 100 dB down from the output, and is ignored.) Thus, the loss ratio is $-200s$, as shown in Fig. 4, and is seen not to depend upon the source or load immittances within the range of accuracy of the simple model.

In the signal flow graph of Fig. 4b, the zero value elements in the active path (operational amplifier) transmission matrix are represented by dotted lines; when these are ignored, only a single path exists between the v_o and e_G nodes. The TMSFG is shown in Fig. 4c.

* To save writing powers of 10, we shall adopt the following system of units throughout: the volt, milliampere, and nanosecond are taken as our fundamental units, leading to the derived units of kohms, mmhos, microhenries, picofarads, gigaradians per second (Gr/s), and GHz.

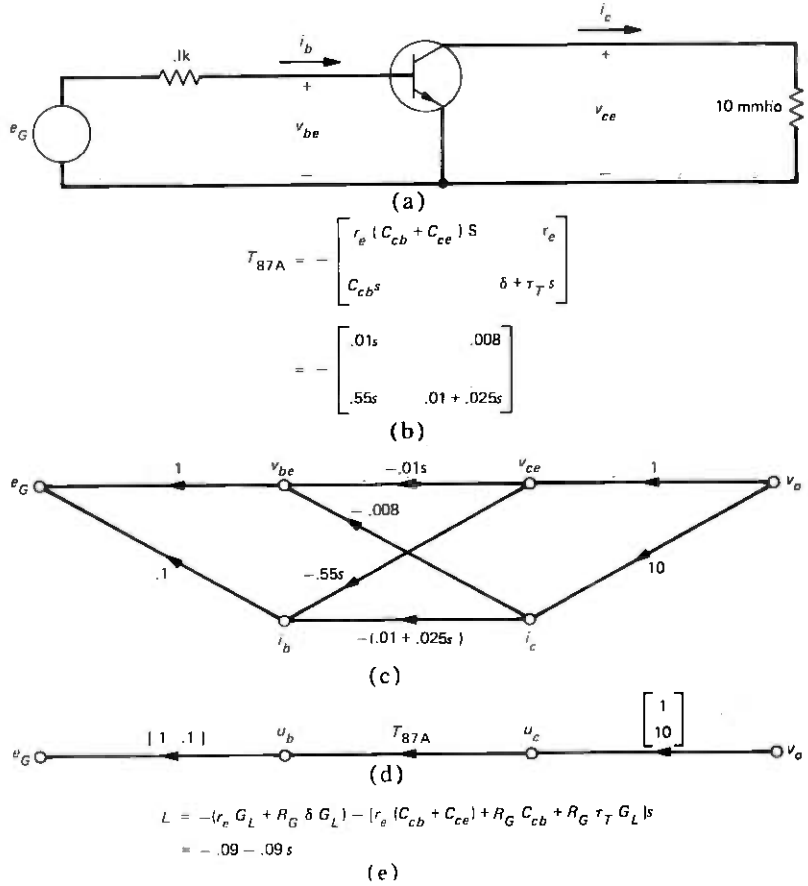


Fig. 5—Loss ratio calculation for a common emitter stage employing a type 87A transistor.

The second example, shown in Fig. 5, is a common emitter transistor stage using the Western Electric type 87A transistor. Accurate characterization of the transmission parameters of this transistor is under way; for purposes of the illustration, we approximate the transmission matrix as shown in Fig. 5b, accurate in magnitude to 1 GHz but somewhat deficient in phase. The transistor parameters are determined at a collector current of 5 mA, and a collector-to-emitter voltage of 3 volts, and are

r_e = emitter resistance	0.008 kohm
C_{CB} = collector-to-base capacitance	0.55 pF
$\delta = 1/h_{fe}$, the reciprocal current gain	0.01
$\tau_T = 1/2\pi f_T$, the current gain time constant	0.025 ns
C_{ce} = collector-to-emitter capacitance	0.7 pF

These parameter values yield the transmission matrix shown in Fig. 5b, whose elements are entered on the signal flow graph of Fig. 5c. The TMSFG is shown in Fig. 5d. The calculation of loss ratio, shown in Fig. 5e, can be made by adding all path products of the signal flow graph, or by performing the matrix operations indicated in the TMSFG. The loss ratio is seen to be a binomial in the frequency variable, with a cutoff frequency of 1 Gr/s, or 0.16 GHz.

This second example points to an advantage of the anticausal approach in determining circuit sensitivities. With feedback loops absent, the input signal is simply the sum of all paths from output to input of the signal flow graph, so that the sensitivity of the loss ratio to r_e, for example, is simply the proportion of the input contributed by r_e. At low frequencies, the r_e contribution is 0.08 out of a total input of 0.09, so that the sensitivity to r_e is 0.08/0.09, or 0.89.

The advantages of the anticausal approach for the simple circuits studied so far are implicit in the removal of feedback loops and therefore denominators from the transmission expressions. It remains to be shown that a method may be developed for retaining these advantages in more complicated circuits with more than mere ladder or cascade coupling.

IV. "FEEDBACK": THE EFFECT OF SPANNING NETWORKS

We define a *spanning network* as a two-port network which is connected between a pair of nonadjacent circuit vector nodes of a cascaded network. In the circuit of Fig. 6a, for example, the conductance G will be considered to be a spanning network around the transistor, and in Fig. 6b, the upper transistor will be considered to be a spanning network around the lower transistor. The choice at this point is arbitrary as to which is the spanning network and which is the cascade network. The consideration of active spanning networks is beyond our scope here, but in the case of the circuit of Fig. 6a, the reason for the choice will become clear. The conductance can be represented by its two-port dependent generator equivalent circuit as shown in the figure. Four separate effects are introduced by G, corresponding to the four elements of its y-parameter matrix:

$$y = \begin{bmatrix} G & -G \\ -G & G \end{bmatrix} \tag{6}$$

Clearly, y_{11} and y_{22} load the input and output circuits by shunt conductances equal to G. The generator $y_{12}v_o = -Gv_o$ augments the input *current* by an amount proportional to the output voltage upon which it depends. This y_{12} augmentation is usually the reason for connecting G to the circuit, and the other three effects (of y_{11}, y_{22}, and y_{21}) are side effects, usually deleterious. The fourth effect, introduced by the gen-

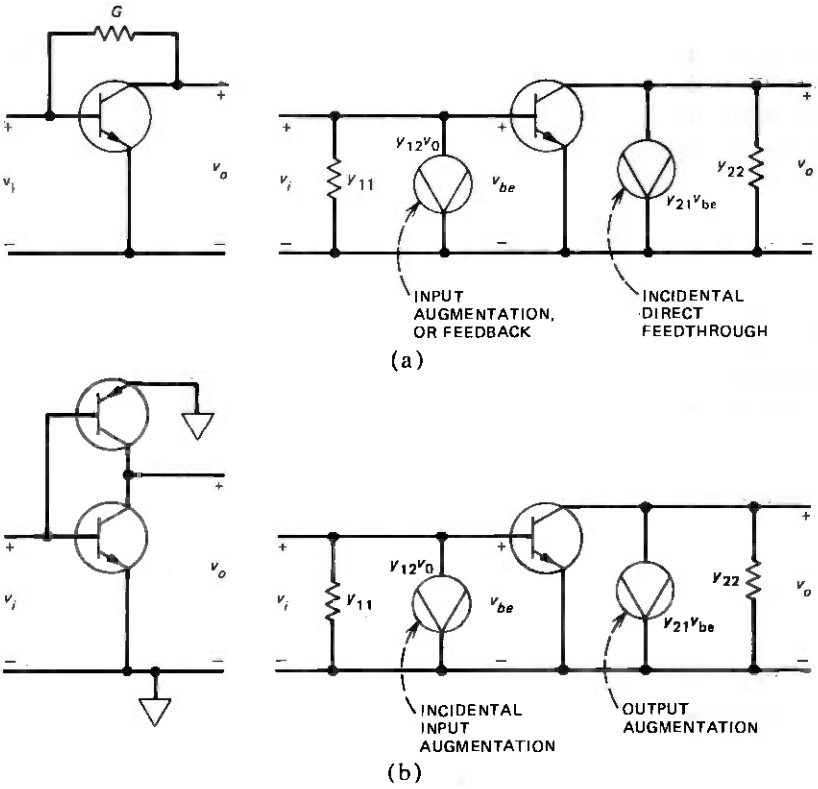


Fig. 6—Two types of spanning networks: (a) input signal augmentation, or feedback type, and (b) output signal augmentation, or feedforward type.

erator $y_{21}v_i$, is *direct feedthrough*. Where the transistor has high gain, e.g., where $v_i \ll v_o$, this effect is incidental, and can often be neglected.

The method of treating a spanning network in anticausal circuit analysis is (*i*) to *represent* its two-port characteristics by one of the four sets of network parameters whose dependent generator equivalent circuits and signal flow graphs are shown in Fig. 7, and (*ii*) to *decompose* its four network parameters into four separate transmission matrices, corresponding to the four effects of input and output circuit loading, input augmentation (or "feedback"), and output augmentation (direct feedthrough).

The four two-port representations of Fig. 7 correspond to the four well-known feedback configurations. The *h* parameters are chosen to represent a spanning network which provides series-input/parallel-output feedback, the main effect of which is to augment transmission parameter A by h_{12} of the spanning network. We shall term this A

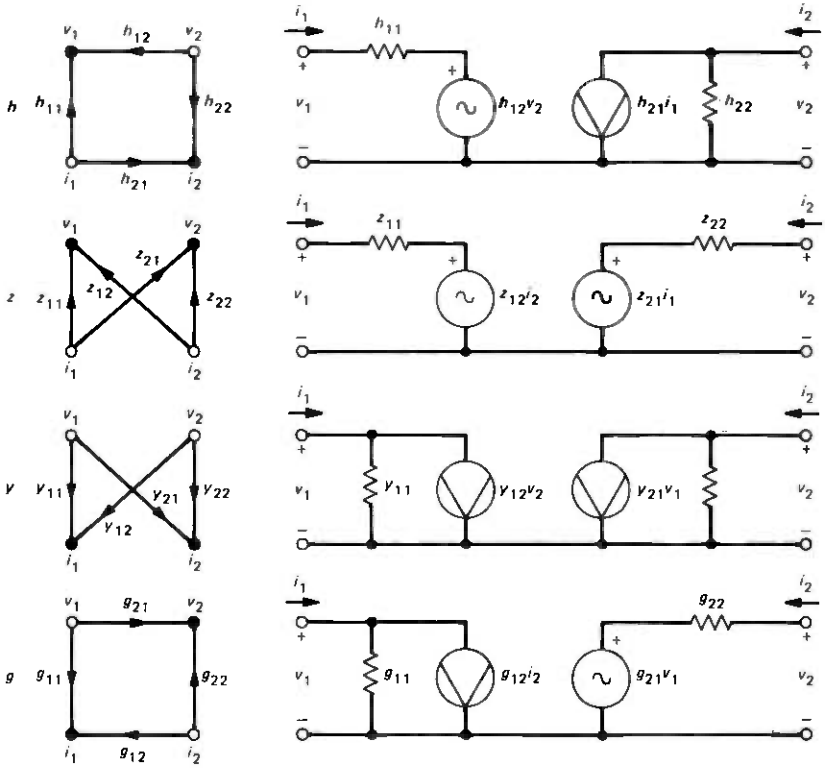


Fig. 7—Signal-flow graphs and equivalent circuits for two-ports corresponding to the h, z, y, and g parameter representations.

feedback. The z parameters augment B by z_{12}, and are appropriate for series-input/series-output, or B feedback. Similarly, y_{12} augments C, providing C feedback, and g_{12} augments D, giving D feedback. We shall consider all four types of feedback in the following section. In this section, we shall consider C feedback (parallel-input/parallel-output feedback) in detail.

A signal-flow graph for the circuit of Fig. 6a is shown in Fig. 8a. The four branches labeled A, B, C, and D represent the transmission parameters of the transistor. The other four (nonunity) branches represent the effect of the four y parameters of the spanning network. Three of these latter branches, corresponding to y_{11} and y_{22}, the input and output loading by the spanning network, and y_{21}, the direct feedthrough to the output through the spanning network, are shown as dashed lines to indicate their lesser importance.

In Fig. 8b, a TMSFG for this circuit is shown. The active path transmission matrix, T_a, includes the four transistor transmission parameters of Fig. 8a. The four branches of Fig. 8a which represent the spanning

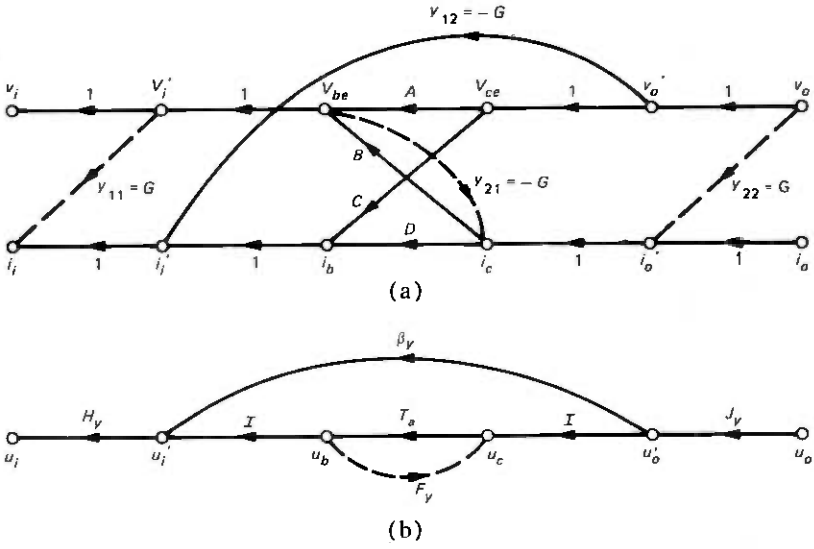


Fig. 8—Development of the signal-flow graph and the TMSFG for the circuit of Fig. 6a. (a) Signal-flow graph, with direct feedthrough and input and output feedback loading branches shown with dashed lines. (b) TMSFG.

network y parameters yield *four separate transmission matrices*, β_y, F_y, H_y, and J_y, defined as follows: β_y is the input augmentation or feedback matrix, given by

$$\beta_y = \begin{bmatrix} 0 & 0 \\ y_{12} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ -G & 0 \end{bmatrix} \tag{7}$$

F_y is the direct feedthrough, or feedforward matrix, given by

$$F_y = \begin{bmatrix} 0 & 0 \\ y_{21} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ -G & 0 \end{bmatrix} \tag{8}$$

H_y is the input loading matrix, given by

$$H_y = \begin{bmatrix} 1 & 0 \\ y_{11} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ G & 1 \end{bmatrix} \tag{9}$$

J_y is the output loading matrix, given by

$$J_y = \begin{bmatrix} 1 & 0 \\ y_{22} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ G & 1 \end{bmatrix} \tag{10}$$

In these four equations, the matrices containing G were obtained by substituting the y parameters of eq. (6) into the general expressions.

The transmission matrix for the C-feedback amplifier of Fig. 6a can be obtained by evaluating the graph gain of the TMSFG. As shown in Appendix A, this transmission matrix is

$$T = H_y[\beta_y + T_a(I - F_yT_a)^{-1}]J_y \tag{11}$$

The matrix F_yT_a will be called the *return ratio matrix*, and $(I - F_yT_a)^{-1}$ will be termed the *return difference matrix inverse*. These two matrices arise from the presence of a feedback loop in the signal flow graph and in the TMSFG, one which arises from the incidental direct feedthrough, or feedforward from input to output through the spanning network. Where the y_{21} branch can be ignored, F_y can be considered a null matrix, and the graphs become cascade graphs. The transmission matrix of eq. (11) can also be written

$$T = \beta_y + H_yT_a(I - F_yT_a)^{-1}J_y \tag{12}$$

since

$$H_y\beta_yJ_y = \beta_y$$

This equation states that the transmission matrix of the amplifier is the sum of the β_y matrix and the matrix of the active path, which itself is the transmission matrix of the transistor, modified by input and output loading and direct feedthrough. Our next step is to calculate the effect of these modifications of the active path transmission matrix.

To evaluate the effect of direct feedthrough, we begin by finding the return ratio matrix:

$$F_yT_a = \begin{bmatrix} 0 & 0 \\ y_{21}A & y_{21}B \end{bmatrix} \tag{13}$$

whereupon the return difference matrix inverse becomes

$$(I - F_yT_a)^{-1} = \frac{1}{1 - y_{21}B}\begin{bmatrix} 1 - y_{21}B & 0 \\ y_{21}A & 1 \end{bmatrix} \tag{14}$$

The effect of direct feedthrough is a small modification of the transmission parameters of the active path. Thus,

$$T_a' = T_a(I - F_yT_a)^{-1} = \frac{1}{1 - y_{21}B}\begin{bmatrix} A & B \\ C + y_{21}\Delta^t & D \end{bmatrix} \tag{15}$$

where $\Delta^t = AD - BC$ is the determinant of the transmission matrix of the transistor. Ordinarily, $|y_{21}B| \ll 1$ and $|y_{21}\Delta^t| \ll C$, so that the active path remains essentially unaffected by the direct feedthrough.

The loss ratio between a Thevenin source and a load conductance is found as in eq. 4:

$$L = [1 \quad R_G]T\begin{bmatrix} 1 \\ G_L \end{bmatrix} \tag{16}$$

Using eq. (12) for T, and substituting the matrix element values of eqs.

(7) to (10), we obtain

$$L = R_G y_{12} + \begin{bmatrix} 1 & R_G \end{bmatrix} \begin{bmatrix} 1 & 0 \\ y_{11} & 1 \end{bmatrix} T_a' \begin{bmatrix} 1 & 0 \\ y_{22} & 1 \end{bmatrix} \begin{bmatrix} 1 \\ G_L \end{bmatrix} \quad (17)$$

$$= R_G y_{12} + \begin{bmatrix} 1 + R_G y_{11} & R_G \end{bmatrix} T_a' \begin{bmatrix} 1 \\ G_L + y_{22} \end{bmatrix} \quad (18)$$

The term $1 + R_G y_{11}$ is a potentiometer term arising from the voltage divider action between the source resistance, R_G, and the spanning network input loading admittance, y_{11}, and the term $G_L + y_{22}$ represents the total output load admittance, including the spanning network output loading. Letting $P_G = 1 + R_G y_{11}$ and $G_L' = G_L + y_{22}$, we can write

$$L = R_G y_{12} + \frac{1}{1 - y_{21}B} [P_G A + P_G B G_L' + R_G C - R_G y_{21} \Delta^t + R_G D G_L'] \quad (19)$$

We can recapitulate the above development by identifying each term of this equation with the relevant spanning network effect, comparing it with eq. (5) for the loss ratio without the spanning network. The first term is the input augmentation, or "feedback," which is of course absent from eq. (2–5). This is the reciprocal of the familiar R_F/R_G gain approximation for this circuit, with $R_F = 1/G$. The remaining terms are divided by the return difference, $1 - y_{21}B$, which is ordinarily close to unity. The first term in the brackets, $P_G A$, is the same as that of eq. (5), except that it is magnified by the input loading factor, $P_G = 1 + y_{11}R_G$. The second term is magnified by this term as well as by the increased output loading provided by y_{22}. The third term is unchanged from eq (5). The fourth term is new: ordinarily very small, it constitutes a reverberation of the signal back and forth through the circuit. The fifth term in the brackets is the D term of eq. (5) magnified by the increased load conductance.

The main difference between the two equations is the feedback term, $R_G G$. The loading has a lesser effect but it is not normally negligible. The direct feedthrough effect is normally negligible.

As examples of the use of the above equations, consider the circuits of Figs. 4 and 5 with feedback conductances connected between input and output, as shown in Fig. 9. The loss ratio for the operational amplifier circuit consists of only two terms of eq. (19), since we have assumed that B, C, and D are zero. Hence, Δ^t is also zero, and

$$L = -R_G G + P_G A \quad (20)$$

Since $P_G = 1.1$ and $A = -200s$, we have

$$L = -(.1 + 220s) \quad (21)$$

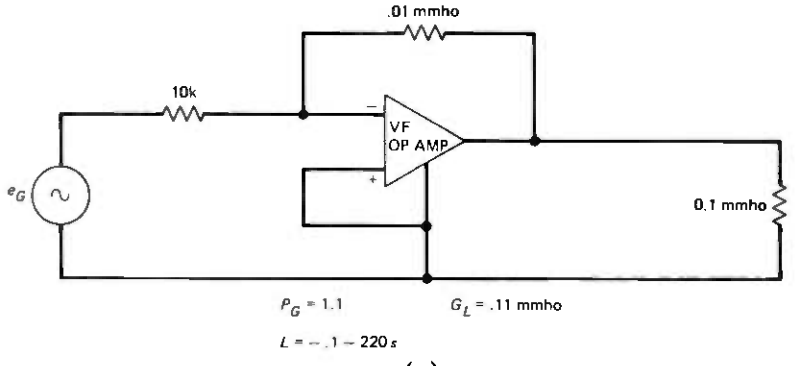


Fig. 9—Examples of loss ratio calculation with spanning network.

Direct feedthrough and output loading are of no concern by our assumption that B and D of the active path are zero. Feedback, or input augmentation, gives the low-frequency value of .1, and the value is A magnified by the input P_G term.

A more substantial example is provided in Fig. 9b, in which the common emitter stage of Fig. 5 is modified by connecting a 1 mmho conductance from input to output. With $R_G = .1$ k, we have $P_G = 1.1$, and with $G_L = 10$ mmho, we have $G_L' = 11$ mmho. From eq. (19), the loss ratio of the transistor stage is

$$L_{87A} = -.1 - \frac{1}{1 - 0.008}[.011s + .0968 + .055s + .00043s + .011 + .0275s] \quad (22)$$

in which the terms are in the order given in eq. (19). The value of Δ^t is taken as $.0043s$, ignoring the s^2 coefficient, since it affects the result only at frequencies higher than the range of approximation of the transistor model (1 GHz). Thus,

$$L_{87A} = -.209 - .0947s \quad (23)$$

When the direct feedthrough is ignored, the denominator in eq. (22) becomes unity, and the $R_G G \Delta^t$ term (.00043s) drops out, giving

$$L_{87A} \approx -.208 - .0935s \tag{24}$$

an approximation of which a circuit designer can be proud. When input and output loading are ignored, P_G becomes unity and G_L' reverts to G_L, 10 mmho. This approximation is rougher, but still valuable for circuit thinking:

$$L_{87A} \approx -.19 - .090s \tag{25}$$

A still rougher approximation is obtained by ignoring certain of the less important transistor parameters (for this case), such as δ and $C_{\dot{c}e}$. With these assumptions, the transmission matrix of the transistor becomes

$$T_{87A} \approx -\begin{bmatrix} .0044s & .008 \\ .55s & .025s \end{bmatrix} \tag{26}$$

and the loss ratio becomes

$$L_{87A} = -(.18 + .084s) \tag{27}$$

which is roughly 10 percent below the true value. A much rougher approximation is obtained by ignoring the contribution of the active path entirely, a good strategy where the loss ratio is controlled primarily by the feedback. For the case of the 87A, we would obtain $L_{87A} = .1$, a poor approximation, since r_e contributes to the low frequency loss ratio, and C_{cb} and τ_T provide most of the high-frequency loss. In the case of the op amp, this strategy accurately predicts the low-frequency loss ratio, but obviously cannot account for the increase in loss ratio at high frequencies. The significance of ignoring the active path contribution is that it defines the transmission matrix of the active path as the null matrix. This provides us with a convenient *reference condition* for a feedback circuit. Where Bode defined a reference condition for a feedback circuit as the circuit in which the "tube [active path] is dead," we stand the definition on its head, and take our reference condition as one in which the active path is very much alive—an ideal two-port amplifier—to be discussed in the next section. The approximation is widely used in operational amplifier applications such as active filter design.

The analysis of the C-feedback amplifier in this section shows that the essential character of the simple anticausal analysis of the circuits of Section II is retained when the y-parameter spanning network is added to the circuit. The cascade nature of the signal-flow graph is essentially retained because the loop gain of the inevitable feedback loops is below unity, and for usual feedback circuits, negligible. The sensitivities to circuit elements are easily evaluated. The low-frequency sensitivity of loss ratio to r_e in the 87A feedback circuit, for example, is seen

from eq. (22) to be .0976/.209 = .47, compared with the previously calculated value of .89 for the stage without the spanning network. The contribution of r_e to the input voltage has not been reduced—it actually has increased slightly because of the output loading by the spanning network and by the input P_G term—but the total generator voltage has been increased by the input signal augmentation of the spanning network, tending to swamp out the effect of r_e.

In this approach to active circuit analysis, the functional dependencies have been chosen in such a way that the increase in bandwidth and reduction in sensitivities usually ascribed to feedback are accounted for without the presence of denominators associated with feedback.

V. NOTES ON FEEDBACK THEORY

Equation (12) for the C-feedback stage neatly separates four essentials of a feedback amplifier. The two terms of the equation separate the feedback or spanning network and the active paths. The feedback network matrix, β_y, contains one nonzero element which augments the current at the amplifier input in proportion to the output voltage, as we have seen. The active path consists of four matrices, including T_a, the transmission matrix of the active path, H and J, the matrices representing the circuit loading by the spanning network at the amplifier input and output respectively, and $(I - F_y T_a)^{-1}$, the return difference matrix inverse representing direct feedthrough. This equation permits a clear definition of the β-matrix; by setting T_a equal to zero, that is, making T_a the null matrix, the second term in the brackets drops out, so that the transmission matrix of the amplifier becomes β_y. We shall define the *reference condition* for the amplifier by setting $T_a = [0]$. Thus, β_y is the transmission matrix of a C-feedback amplifier whose active path has been set in the reference condition. Later, this definition will be extended to A-, B-, and D-feedback amplifiers.

The concept of an amplifier whose input voltage and current are zero for all finite output signal vectors is a serviceable one which is fairly widely used in making rough calculations of gain of feedback circuits. Calling such an amplifier an ideal two-port amplifier* we can state the following.

Theorem: *An ideal two-port amplifier is an amplifier whose transmission matrix is the null matrix.*

Proof: From equation (5), we have

$$L = A + BG_L + R_G C + R_G D G_L = 0 \tag{28}$$

since $e_g/v_o = 0$ by definition. Since the terminations are arbitrary and

* To distinguish it from an ideal operational amplifier, which is a three-port.

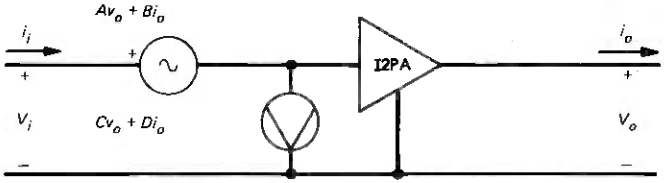


Fig. 10—Dependent generator equivalent circuit for a two-port represented by its transmission parameters.

nonzero, A, B, C, and D must be zero individually. Note that the input and output impedances are indeterminate, since

$$Z_{\text{in}} = \frac{A + BG_L}{C + DG_L} \tag{29}$$

and

$$Z_o = \frac{B + DR_G}{A + CR_G} \tag{30}$$

Impedances will be determined solely by externally applied spanning networks. An ideal two-port amplifier (I2PA) is a circuit element having no parameters to specify it (much like the nullator and norator),[26] and represents a limiting value for an active two-port. It is often useful in drawing equivalent circuits and in modeling; it can, for example, allow us to draw a dependent generator equivalent circuit for a two-port described by the transmission parameters, as shown in Fig. 10.

We can apply the concept of the I2PA to investigate the properties of various feedback configurations. With the active path of a feedback amplifier set in the reference condition, the resulting transmission matrix is simply the β matrix, without the complicating effect of a nonideal active path. In Fig. 11, the circuits of four *unitary feedback amplifiers* and their associated transmission matrices are shown. A *unitary feedback amplifier* is defined as one whose β matrix has but one nonzero element. Figures 11a and d employ permutative feedback—feedback obtained when the active device leads are permuted.[27] When the active path consists of a transistor, these are the common collector and common base stages, respectively. The transmission matrices shown are obtained by inspection, bearing in mind that the input current and voltage of the I2PA are zero. The circuits of Figures 11b and c are duals, with the transmission matrices likewise obtained by inspection.

We can obtain a good approximation to the actual transmission matrix of each of the four circuits of Fig. 11 with a nonideal active path by simply adding the transmission matrix of the active path, with due attention to the sign change introduced in Fig. 11a and d by the permutation of the device leads. This amounts to approximating the transmission matrix

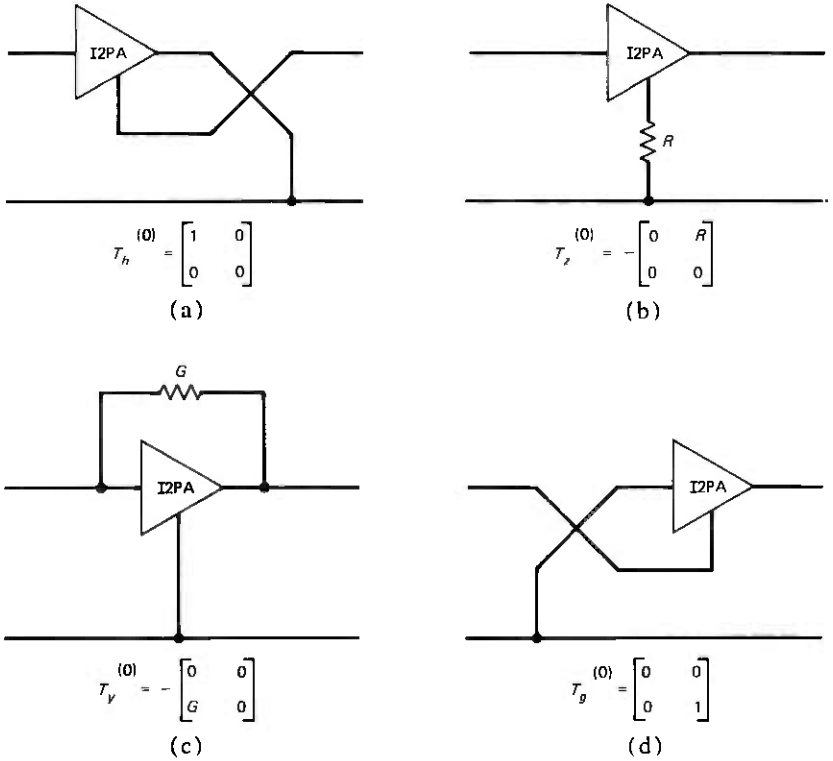


Fig. 11—Transmission matrices of four unitary feedback amplifiers whose active path is in the reference condition. (a) Common collector stage (example of A feedback); (b) emitter resistor feedback (example of B feedback); (c) collector-to-base feedback (example of C feedback); (d) common base stage (example of D feedback).

by the equation

$$T \approx \beta + T_a \tag{31}$$

In the case of the common collector stage, this amounts to approximating the transmission matrix as

$$T_{cc} \approx \begin{bmatrix} 1 - A & -B \\ -C & -D \end{bmatrix} \tag{32}$$

Since permutative feedback is lossless, there are no input and output loading terms, so that the approximation involves ignoring the direct feedthrough. The exact transmission matrix for this stage is derived in Appendix B, and is

$$T_{cc} = \frac{1}{1 - D} \begin{bmatrix} \theta & -B \\ -C & -D \end{bmatrix} \tag{33}$$

where $\theta = 1 - A - D + \Delta^t$, a quantity close to unity which will recur below. Where $D \ll 1$, that is, well below f_T, the approximation of eq. (32) is quite close.

In the case of B-feedback, for which the z-parameter description of the spanning network is appropriate, the approximation of eq. (31) gives

$$T_z \approx \begin{bmatrix} A & B - R \\ C & D \end{bmatrix} \tag{34}$$

The exact value, derived in Appendix C, is

$$T_z = \frac{1}{1 + CR} \begin{bmatrix} A + CR & B - R\theta \\ C & D + CR \end{bmatrix} \tag{35}$$

In comparing this expression with that of eq. (34) we note that the denominator and the multiplication of R by θ are due to direct feedthrough; the CR term added to A comes from input loading by the spanning network, and the CR term added to D comes from output loading by the spanning network. These modifications are small, but may become important at high frequencies, since C represents the (negative) admittance of the collector capacitance, a determining factor in high-frequency performance.

The transmission matrices for the four circuits of Fig. 11 with nonideal active paths are given in Table I. These expressions are intended for computer implementation, since they are complex, and their complexity arises from relatively small corrections on the approximations discussed here. The approximations can be used in the design process.

Table I also includes the transmission matrix of one nonunitary feedback amplifier, a *hybrid feedback amplifier*, incorporating both B- and C-feedback. As can be seen from the table, the matrix for the reference condition includes nonzero elements in all four positions of the matrix. The matrix was obtained by using the transmission matrix elements of T_z as a set of active path elements for the computation of T_y.

As noted above, a spanning network is represented by one of the four parameter sets of Fig. 7. Any one of these parameter sets contains four parameters, each of which generates a transmission matrix; these have been termed β, F, H, and J matrices corresponding to the four effects generated by the spanning network; input augmentation, direct feedthrough, and input and output loading, respectively. Each of the four types of unitary feedback can be represented by the same TMSFG, shown at the top of Table II. The rows of Table II define these four transmission matrices for each of the spanning network parameter sets of Fig. 7. Signs are a problem, since the sign conventions for two-port parameters are different for the parameter sets of Fig. 7 and for the

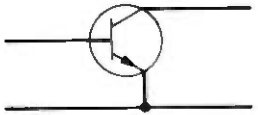
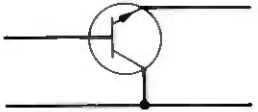
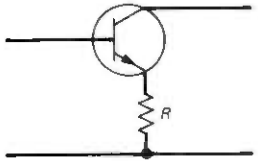
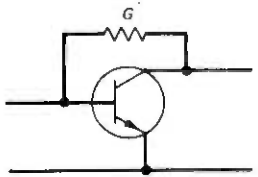
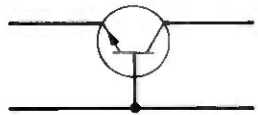
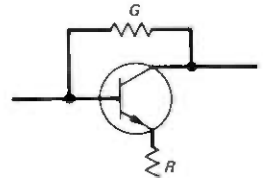
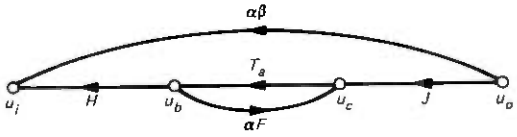
Circuit	Transmission matrix	Reference condition
	$T_a = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$
	$T_{cc} = \frac{1}{1-D}\begin{bmatrix} \theta & -B \\ -C & -D \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$
	$T_z = \frac{1}{1+CR}\begin{bmatrix} A+CR & B-R\theta \\ C & D+CR \end{bmatrix}$	$\begin{bmatrix} 0 & R \\ 0 & 0 \end{bmatrix}$
	$T_y = \frac{1}{1+BG}\begin{bmatrix} A+BG & B \\ C-G\theta & D+BG \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ G & 0 \end{bmatrix}$
	$T_{cb} = \frac{1}{1-A}\begin{bmatrix} -A & -B \\ -C & \theta \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$
	$T_{z/y} = \frac{1}{1+CR+BG-RG\theta}\begin{bmatrix} A+CR+BG-RG\theta & B-R\theta \\ C-G\theta & D+CR+BG-RG\theta \end{bmatrix}$ $\theta = 1-A-D+\Delta^t$	$\frac{1}{1-RG}\begin{bmatrix} RG & R \\ G & RG \end{bmatrix}$

Table II — Matrix element values for the four types of unitary feedback



TMSFG

Type of feedback	α	β	F	H	J
A	-1	$\begin{bmatrix} h_{12} & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & h_{21} \end{bmatrix}$	$\begin{bmatrix} 1 & h_{11} \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ h_{22} & 1 \end{bmatrix}$
B	-1	$\begin{bmatrix} 0 & z_{12} \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & z_{21} \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & z_{11} \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & z_{22} \\ 0 & 1 \end{bmatrix}$
C	1	$\begin{bmatrix} 0 & 0 \\ y_{12} & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ y_{21} & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ y_{11} & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ y_{22} & 1 \end{bmatrix}$
D	1	$\begin{bmatrix} 0 & 0 \\ 0 & g_{12} \end{bmatrix}$	$\begin{bmatrix} g_{21} & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ g_{11} & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & g_{22} \\ 0 & 1 \end{bmatrix}$

transmission parameters. This is accounted for in Table II by introducing the parameter α, which is -1 for the h and z parameter sets and $+1$ for the y and g parameter sets. The signs of the parameter values are all consistent with conventional practice.

The cascode stage of Fig. 12a illustrates a situation in which the common-base stage effectively removes or reduces certain active parameters of the common emitter stage. The transmission matrix of the cascode stage is

$$T_{\text{cascode}} = T_{a1}(\beta_D - T_{a2})$$

$$= T_{a1}\beta_D - T_{a1}T_{a2} \tag{36}$$

$$= \begin{bmatrix} 0 & B_1 \\ 0 & D_1 \end{bmatrix} - T_{a1}T_{a2} \tag{37}$$

The first matrix of eq. (37) is the transmission matrix of the common emitter stage with A_1 and C_1 removed, the primary effect of which is to remove C_1, the (negative) susceptance of the Miller capacitance. The second term is the negative of the cascaded pair of transistors in the common emitter configuration, a matrix whose elements are much smaller than those for a single stage up to frequencies at which the common emitter gain becomes small.

The process of sorting out the unique character of amplifier configurations is helpful for circuit design. Consider the cascades of unitary

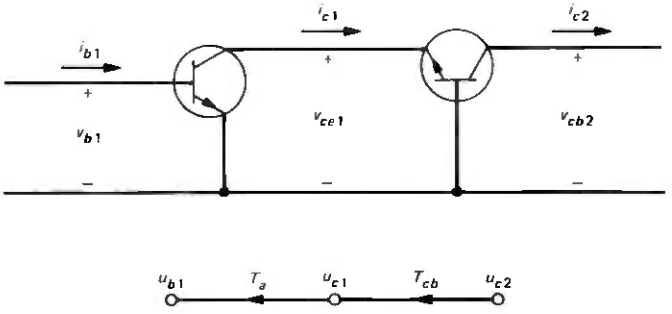
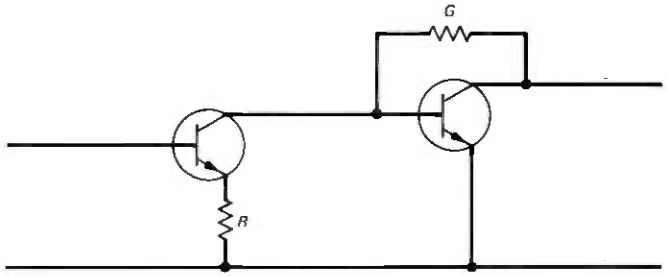


Fig. 12—Cascode stage.

feedback amplifiers of Fig. 13.[28] In Fig. 13a, the cascade of a B feedback amplifier with a C feedback amplifier has a transmission matrix given by $T_z T_y$ of Table I. When both transistors are placed in the reference condition, we observe that the elements of the β matrix of the combination are all zero except for the factor RG in the A position, so that the combination is itself a unitary A-feedback amplifier. Reversing the order of the stages, in Fig. 13b, gives a D-feedback amplifier. If we cascade two C-feedback stages (or two B-feedback stages) we find that the β matrix of the combination is null. In Fig. 13c, we note that the C feedback of the second stage augments the input current of that stage, but not the voltage. Since the β network of the first stage senses the input voltage of the second stage, which is zero in the reference condition, no overall feedback arises. The overall loss ratio increases as a result of the input augmentation of the individual stages; the increased input current of the second stage increases the contribution of B_1 and D_1 to the loss ratio, and the feedback around the first stage increases the effect of A_2 and B_2, but the β matrix for the combination is null.

VI. EQUIVALENT LADDER CIRCUITS FOR FEEDBACK AMPLIFIERS

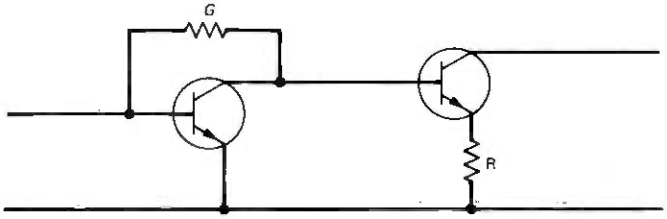
The circuit of any amplifier whose two-port characteristics are sought may be drawn as an *equivalent ladder circuit,* that is, a cascade of active and passive network elements, by the direct expedient of representing circuit couplings among nonadjacent nodes of the ladder by one or more of the dependent generator equivalent circuits of Fig. 7. This will be illustrated by deriving an equivalent ladder circuit for the A-feedback pair of Fig. 14, in which the output voltage of the second stage is divided down in a resistive divider and applied to the emitter of the first stage, where it augments the amplifier input voltage. Since the β network augments A of the active path transmission matrix, the β network is properly represented by its h parameters. The relationship of the h parameters to circuit elements R_E and R_F is given in Fig. 14b, which also defines a



$$T_{B\text{-}C} = T_z T_y$$

$$T_{B\text{-}C}^{(0)} = \begin{bmatrix} RG & 0 \\ 0 & 0 \end{bmatrix}$$

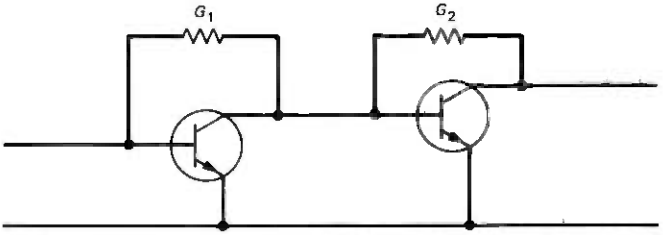
(a)



$$T_{C\text{-}B} = T_y T_z$$

$$T_{C\text{-}B}^{(0)} = \begin{bmatrix} 0 & 0 \\ 0 & RG \end{bmatrix}$$

(b)



$$T_{C\text{-}C} = T_y T_y$$

$$T_{C\text{-}C}^{(0)} = [\,0\,]$$

(c)

Fig. 13—Cascades of unitary feedback amplifier.

more convenient set of feedback parameters, R_A, G_A, and n_A, where R_A is the parallel combination of R_E and R_F, G_A is the conductance of the series combination of R_E and R_F, and n_A is $R_E/(R_E + R_F)$, which can be considered the turns ratio of the ideal transformer in the network shown. The application of the spanning network is shown in Fig. 14c. The

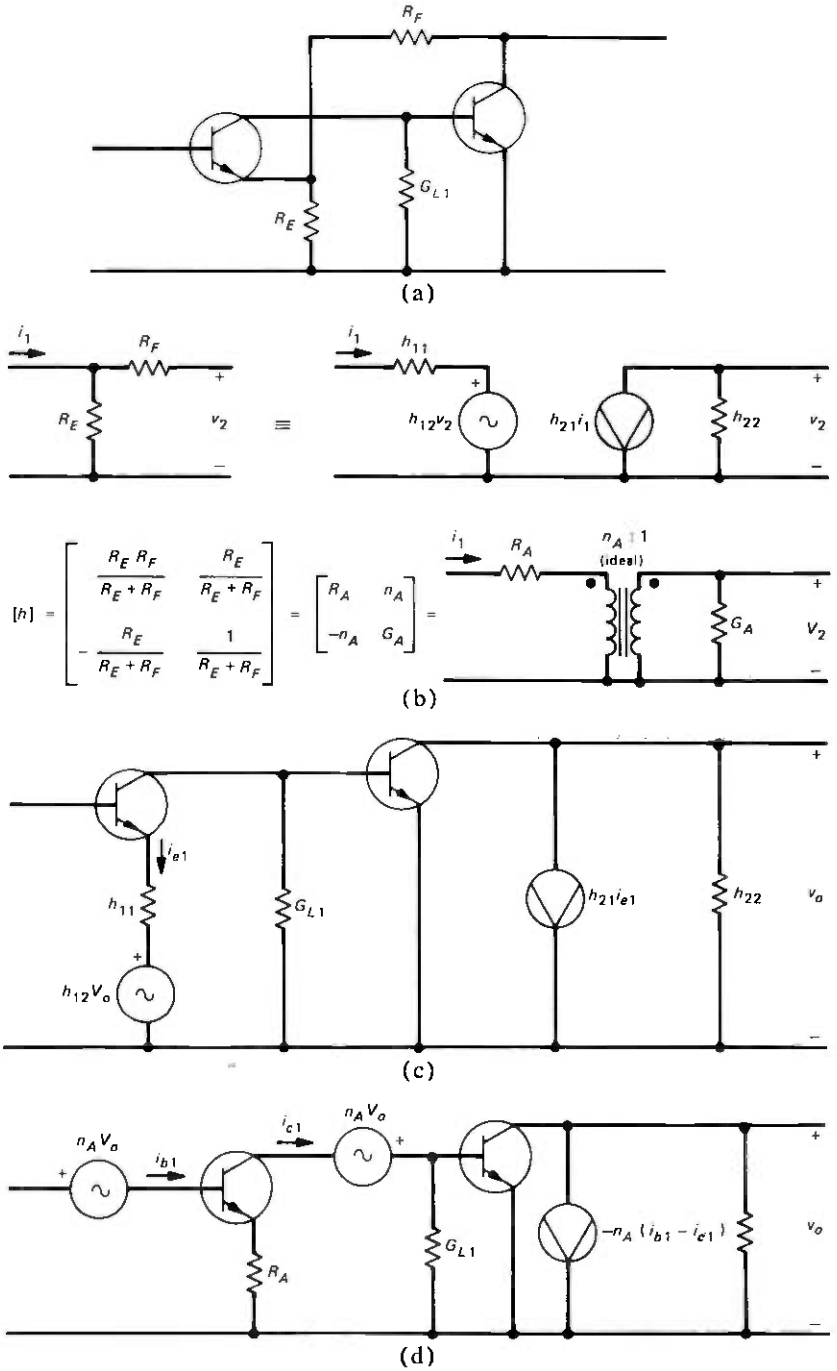


Fig. 14—The *A*-feedback pair. (a) Circuit; (b) analysis of spanning network; (c) application of dependent generator equivalent circuit of spanning network; (d) equivalent ladder network.

final equivalent circuit of Fig. 14d removes the $n_A v_o$ gererator from the emitter circuit of the first stage, replacing it by two generators, one in series with the base lead of the first stage, and one in series with the collector lead. This completes the transformation to the ladder configuration, except that we have left R_A in the emitter of the first stage. The reduction to a ladder of elementary active and passive devices would strictly require that the local B feedback of the first stage be represerted by a separate z-parameter spanning network. To save work, we shall take the single-stage circuits of Table I as elementary building blocks, so that the first-stage active path will be represented as T_z. Thus, there is no need to reanalyze the single-stage circuits of Table I each time they arise. In computer evaluation, the properties of T_z are derived from T_a in a subroutine. We note that when the transistors of the circuit of Fig. 14 are placed in the reference condition, the transmission matrix is

$$\beta_A = \begin{bmatrix} n_A & 0 \\ 0 & 0 \end{bmatrix} \tag{38}$$

so that the A-feedback pair is a unitary feedback amplifier.

The ladder equivalent circuit for the D-feedback amplifier of Fig. 15a is derived in an exactly analogous manner, and is shown in Fig. 15b. In this case, the g parameters of Fig. 7 are the appropriate set, since g_{12} relates the input current to the output current. When the transistors are placed in the reference condition, the transmission matrix of the circuit is

$$\beta_D = \begin{bmatrix} 0 & 0 \\ 0 & -n_D \end{bmatrix} \tag{39}$$

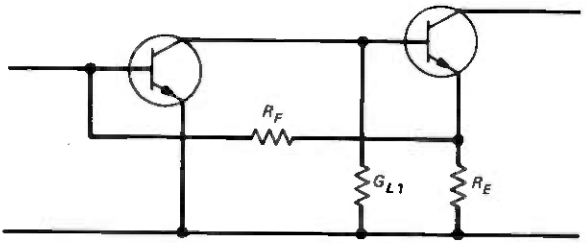
so that this, too, is a unitary feedback amplifier.

Simultaneous application of A- and D-feedback is shown in Fig. 15c, and the ladder equivalent circuit is shown in d. The circuit is an extension of the two unitary feedback circuits from which it is derived. When the transistors of this circuit are placed in the reference condition, the expression for the β matrix is complicated. It can be simplified by separating out the effects of G_D and G_A, which we would normally associate with the source and load immittances, respectively. The remaining matrix may be written by inspection, and is the middle matrix of:

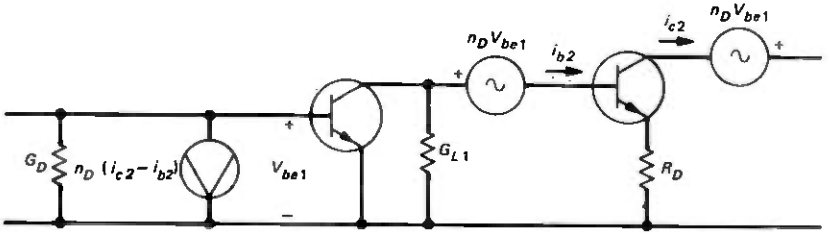
$$\beta_{A/D} = \begin{bmatrix} 1 & 0 \\ G_D & 1 \end{bmatrix} \begin{bmatrix} n_A & R_A G_{L1} R_D \\ 0 & n_D \end{bmatrix} \begin{bmatrix} 1 & 0 \\ G_A & 1 \end{bmatrix} \tag{40}$$

If G_{L1} is eliminated (by bootstrapping or by use of an active current source to provide dc for the first stage), the β matrix consists essentially of the two ratios, n_A to establish the voltage gain and n_D to establish the current gain.

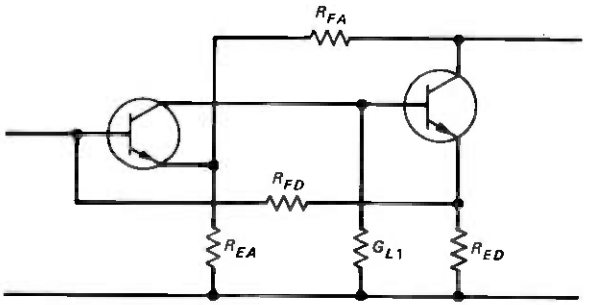
This configuration is another instance of a *hybrid feedback amplifier*



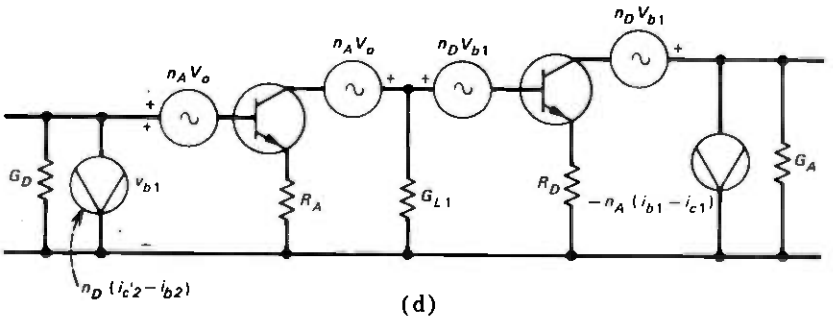
(a)



(b)



(c)



(d)

Fig. 15—Further development of equivalent ladder circuits. (a) A D-feedback amplifier and (b) its equivalent ladder circuit. (c) A hybrid A/D-feedback amplifier and (d) its equivalent circuit.

(the first was encountered as the last entry of Table I). Hybrid feedback can be used to provide a desired input and/or output impedance without incurring the power loss associated with build-out resistance or conductance.[29] At the amplifier output, such a build-out incurs a loss of power output capability, while at the input it increases noise. In the present instance, the two generators establish both the input current and voltage, and therefore the input impedance, and contribute only an incidental amount of noise (associated with the input loading of the spanning networks, which could be eliminated by making the spanning networks lossless, by use of transformers rather than voltage dividers). This leads to the surprising thermodynamic conclusion pointed out by Nyquist that such an amplifier cools down the source, since the source pumps noise power into this noiseless resistance, and receives no noise power in return.

At the beginning of this section, we stated that any amplifier for which the two-port characteristics are sought may be represented by a ladder network with (shunt) current generators and (series) voltage generators which are dependent upon voltages or currents at nonadjacent circuit nodes. Where parallel active paths are involved, a choice must be made as to which of the two paths is to be taken as the spanning network represented by its h, z, y, or g parameters. Where an active spanning network is involved, such as in feedforward circuits, it is advantageous to assign the role of spanning network in such a way that the loop gain arising from feedforward or direct feedthrough is minimized, since this most closely realizes a cascade graph representation. This procedure is beyond the scope of this paper, but will be treated in a subsequent publication.

When active spanning networks are admitted, it is clear that any active, linear network can be represented as a ladder in the sense defined here. It also appears to be true, in looking ahead to the analysis of active spanning networks, that the direct, intuitive understanding of active circuits which comes from the elimination or gross reduction of return ratio can be substantially retained when active spanning networks are used.

VII. WRITING THE TRANSMISSION MATRIX EQUATION FROM THE EQUIVALENT LADDER CIRCUIT

When the feedback amplifier circuit has been redrawn in equivalent ladder form with spanning networks represented by dependent generators, a TMSFG can be drawn directly from the circuit by inspection. As a mechanical aide, a set of circuit vector nodes are placed on the circuit between each element of the ladder. These become the graph nodes of the TMSFG. Branches connecting these nodes in sequence from circuit output to circuit input (from graph input to graph output) define the

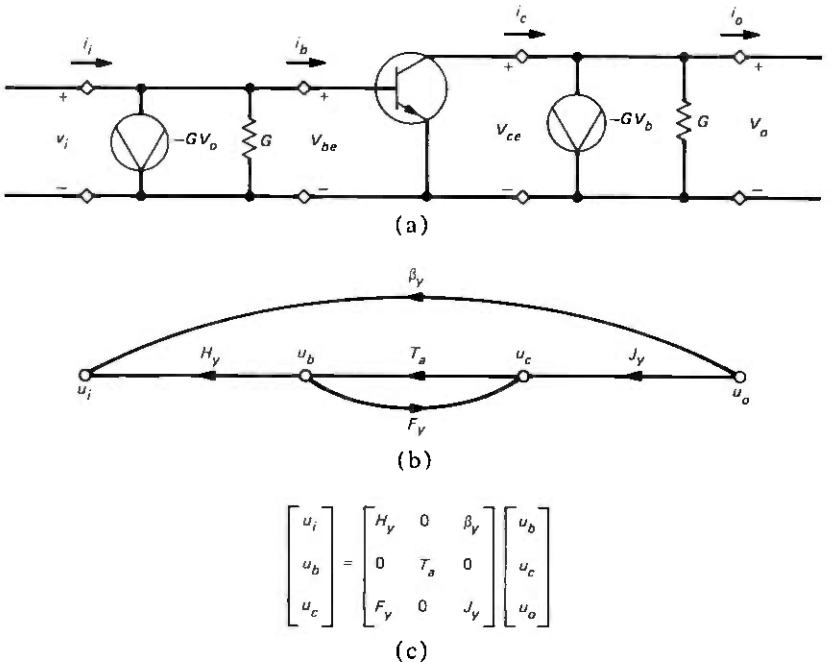


Fig. 16—(a) Equivalent ladder circuit, (b) TMSFG, and (c) transmission matrix array for a unitary C-feedback amplifier.

main transmission path. Each dependent generator will create a branch which spans one or more of these nodes: either a β branch in a direction from the circuit output toward the circuit input, or a direct feedthrough branch (an F branch) in a direction toward the circuit output. An example already examined in Section III is the C-feedback amplifier whose ladder circuit and TMSFG are shown in Fig. 16. The transmission matrix equation for the circuit is obtained as the transmission or graph gain of the TMSFG.

7.1 The transmission matrix array

Writing the transmission matrix equation of the amplifier is facilitated by putting the TMSFG into matrix form. Such a matrix form will be termed a *transmission matrix array* (TMA) which is itself a matrix relating the signal vectors at the nodes which receive signals to the signal vectors at the nodes which transmit them. The matrix elements are the branch values of the branches which connect these nodes in the TMSFG. In Fig. 16, for example, u_i, u_b, and u_c are nodes which receive signals; the signals at these nodes together form the *received signal vector*. Similarly, u_b, u_c, and u_o are nodes which transmit signals, which together form the *transmitted* signal vector. The matrix relating these two vectors

is the transmission matrix array, having a nonzero entry at any element position where a TMSFG branch transmits a signal from a component of the transmitted signal vector to a component of the received signal vector. Figure 16c shows the transmission matrix array for the circuit. It is evident that elements along the principle diagonal are matrices of the ladder network of cascaded circuit elements: elements above or to the right of the principle diagonal are β matrices, and elements below or to the left of the principle diagonal are F (direct feedthrough or feedforward) matrices.

Where all elements below the principle diagonal are zero or can be ignored, the transmission matrix equation relating u_i to u_o can be written by inspection. Thus, ignoring the direct feedthrough element in the TMA of Fig. 16, we can write

$$u_i \approx [\beta_y + H_y T_a J_y] u_o \tag{41}$$

To obtain the exact expression including direct feedthrough, the set of simultaneous equations represented by the TMA must be solved. Thus

$$u_c = F_y u_b + J_y u_o \tag{42}$$

$$= F_y T_a u_c + J_y u_o \tag{43}$$

so that

$$(I - F_y T_a) u_c = J_y u_o \tag{44}$$

and

$$u_c = (I - F_y T_a)^{-1} J_y u_o \tag{45}$$

In this equation, J_y is premultiplied by the return difference matrix inverse, which, for small values of the elements of the return ratio matrix, $F_y T_a$, is essentially the identity matrix. In any case, substitution of $(I - F_y T_a)^{-1} J_y$ for J_y in the TMA of Fig. 16c removes the direct feedthrough element from the TMA, allowing us to write the transmission matrix equation by inspection:

$$u_i = [\beta_y + H_y T_a (I - F_y T_a)^{-1} J_y] u_o \tag{46}$$

The TMA can be written directly from the circuit diagram, allowing us to dispense with the TMSFG, its graph equivalent. In the more complicated feedback amplifiers to be discussed below, the TMA gives a clearer picture of signal dependencies than does the TMSFG.

7.2 Examples: feedback pairs

Figure 17a gives the equivalent ladder circuit for the A-feedback amplifier of Fig. 14. In Fig. 17b, we review the process of drawing the

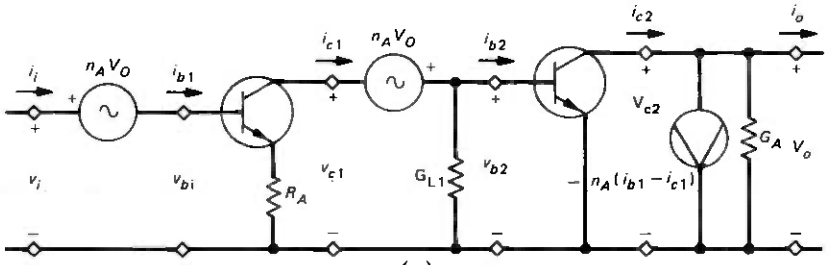


Fig. 17—The A-feedback pair. (a) Equivalent ladder circuit from Fig. 14d; (b) TMSFG; (c) TMA; (d) transmission matrix equation, ignoring direct feedthrough, written by inspection.

TMSFG. The input node, u_i, is identically the u_{b1} node except for the voltage-controlled voltage source, $n_A v_o$, so that u_i receives signals from two branches: the identity matrix branch from u_{b1} and the β_A branch from u_o. β_A is given by eq. (38). Next, u_{b1} is totally controlled by the transmission matrix of the transistor with R_A in the emitter, which we represent as T_z of Table I. Next, u_{c1} is equal to u_{b2} modified by the first-stage load conductance, G_{L1}, represented by the matrix

$$\mathbf{G}_{L1} = \begin{bmatrix} 1 & 0 \\ G_{L1} & 1 \end{bmatrix} \tag{47}$$

which appears as a TMSFG branch from u_{b2} to u_{c1}. In addition, the

voltage-controlled voltage source, $-n_A v_o$, adds branch $-\beta_A$ to u_{c1} from u_o. Node $u_{b2} = T_a u_{c2}$. Node u_{c2} has three inputs, $\mathbf{G}_A$ from u_o and the two F_A branches from u_{b1} and u_{c1} representing direct feedthrough of first-stage emitter current to the second collector. This completes the TMSFG.

The TMA can be constructed by exactly the same reasoning, and is shown in Fig. 17c. The columns of the TMA correspond to the transmitting nodes, and the rows to the receiving nodes. Where a node of a given row receives signal transmitted from a node of a given column, the branch transmission matrix is entered at the intersection, as shown.

The approximate transmission matrix equation, ignoring direct feedthrough, is written by inspection of the TMA as shown in Fig. 17d. This equation shows the overall input voltage augmentation by the first term, $\beta_A u_o$, and shows that the first stage incorporates local B-feedback, implicit in the transmission matrix, T_z. These are well-known characteristics of this feedback pair. What is less generally realized is that the second stage also incorporates local feedback, in this case local A-feedback, apparent from the additive $-\beta_A$ term in the parentheses, a term which is important to the high-frequency behavior of the circuit. (The input current from this cause alone is approximately $-C_1 n_A = n_A C_{cb1} s$.) The output loading of the feedback divider network is $G_A = 1/(R_E + R_F)$. If this is reduced by scaling R_E and R_F upward, the local feedback of the first stage is increased, since $R_A = R_E R_F/(R_E + R_F)$ is scaled up by the same factor, so that in the design of an A-feedback pair, a balance must be sought between these two effects.

The exact expression including the effect of direct feedthrough is useful as a final check, usually performed on the computer. It is obtained, as before, by solution of the simultaneous equations. We first reduce the TMA by direct substitution of the cascade portion of the TMA, that is, the portion containing no entries to the left of the principle diagonal. The TMA, thus condensed, is

$$\begin{bmatrix} u_i \\ u_{c2} \end{bmatrix} = \begin{bmatrix} T_z \mathbf{G}_{L1} T_a & (I - T_z)\beta_A \\ -F_A(I - T_z)\mathbf{G}_{L1} T_a & \mathbf{G}_A - F_A T_z \beta_A \end{bmatrix} \begin{bmatrix} u_{c2} \\ u_o \end{bmatrix} \tag{48}$$

Next, we remove the direct feedthrough term by removing the self-loop at node u_{c2}:

$$\begin{bmatrix} u_i \\ u_{c2} \end{bmatrix} = \begin{bmatrix} T_z \mathbf{G}_{L1} T_a & (I - T_z)\beta_A \\ 0 & M(\mathbf{G}_A - F_A T_z \beta_A) \end{bmatrix} \begin{bmatrix} u_{c2} \\ u_o \end{bmatrix} \tag{49}$$

Where M is the return difference matrix inverse, given by

$$M = [I + F_A(I - T_z)\mathbf{G}_{L1} T_a]^{-1} \tag{50}$$

The transmission matrix equation can now be written by inspection:

$$u_i = [(I - T_z)\beta_A + T_z \mathbf{G}_{L1} T_a M(\mathbf{G}_A - F_A T_z \beta_A)]u_o \qquad (51)$$

which reduces to the equation in Fig. 17d when $F_A = 0$.

The D-feedback pair whose equivalent ladder circuit is given in Fig. 15b can be analyzed by exactly similar means, with β_D given by eq. (39), G_D placed in shunt across the input terminals accounting for input circuit loading, and R_D in series with the emitter lead of the second transistor accounting for output loading by the spanning network. For this spanning network,

$$F_D = \begin{bmatrix} n_D & 0 \\ 0 & 0 \end{bmatrix} \qquad (52)$$

Writing the TMA and transmission matrix equation for this circuit is left as an exercise for the reader.

7.3 Hybrid feedback: the A/D hybrid feedback pair

Analysis of the A/D hybrid feedback pair demonstrates the utility of the transmission matrix array, as compared with the TMSFG. The TMA is essentially an incidence matrix of the TMSFG, presenting the same information in a better-ordered form. In Fig. 18a, the equivalent ladder circuit of Fig. 15d is repeated, and the TMSFG and TMA for this circuit are given in Fig. 18b and c. The TMSFG includes eight spanning branches; even if the four direct feedthrough branches are ignored, the tangle of β branches makes the writing of the graph gain (the transmission matrix of the circuit) hazardous. In the TMA, the role of each of these spanning branches is clarified, at least allowing us to write the approximate transmission matrix equation (in which the direct feedthrough branches are ignored) by inspection, proceeding row by row. Thus, the transmission matrix of the A/D feedback pair is written from the TMA as follows, in which the matrices G_D and G_A are first factored out, and the elements are considered row by row, starting from the right-hand end of the first row:

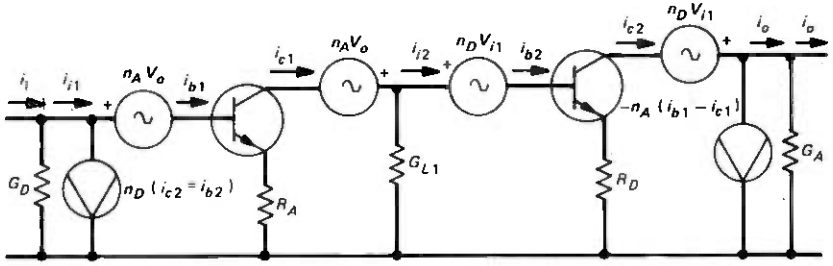
$$u_i \approx \mathbf{G}_d[\beta_A - \beta_D + \beta_D T_{z2} + T_{z1}(-\beta_A + \mathbf{G}_{L1} T_{z2})]\mathbf{G}_A u_o$$

or

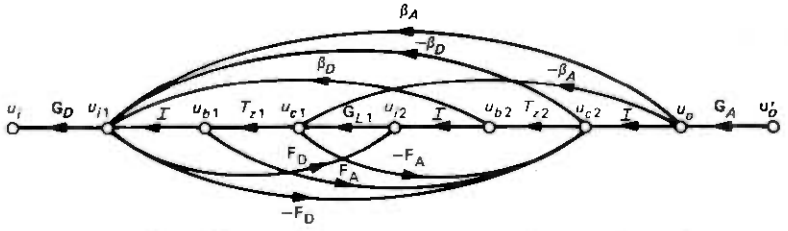
$$u_i \approx \mathbf{G}_D[(I - T_{z1})\beta_A - \beta_D(I - T_{z2}) + T_{z1}\mathbf{G}_{L1} T_{z2}]\mathbf{G}_A u_o \qquad (53)$$

The first term on the right in the brackets represents the A-feedback, the second term the D-feedback, and the third term the transmission matrix of the active path itself, modified by the series loading of the two spanning networks.

Equation (53) ignores the effects of the direct feedthrough branches, or the TMA entries below the principle diagonal, and is therefore approximate. The effect of the feedthrough branches is often to add excess



(a)



$$\beta_A = \begin{bmatrix} n_A & 0 \\ 0 & 0 \end{bmatrix} \quad \beta_D = \begin{bmatrix} 0 & 0 \\ 0 & -n_D \end{bmatrix} \quad F_A = \begin{bmatrix} 0 & 0 \\ 0 & -n_D \end{bmatrix} \quad F_D = \begin{bmatrix} n_D & 0 \\ 0 & 0 \end{bmatrix}$$

(b)

$$\begin{bmatrix} u_i \\ u_{i1} \\ u_{b1} \\ u_{c1} \\ u_{i2} \\ u_{b2} \\ u_{c2} \\ u_o \end{bmatrix} = \begin{bmatrix} G_D & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & \beta_D & -\beta_D & \beta_A & 0 \\ 0 & 0 & T_{z1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & G_{L1} & 0 & 0 & -\beta_A & 0 \\ F_D & 0 & 0 & 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & T_{z2} & 0 & 0 \\ -F_D & F_A & -F_A & 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & G_A \end{bmatrix} \begin{bmatrix} u_{i1} \\ u_{b1} \\ u_{c1} \\ u_{i2} \\ u_{b2} \\ u_{c2} \\ u_o \\ u_o' \end{bmatrix}$$

(c)

Fig. 18—The equivalent ladder circuit, TMSFG, and TMA for the *A/D* hybrid feedback pair.

phase to the active path, and is therefore of importance in investigating stability in the vicinity of the crossover frequency (the frequency at which the magnitudes of the contributions to the loss ratio from the β path and the active path are equal). The complete transmission matrix, including the effect of direct feedthrough brances, is obtained by direct

but tedious algebra from the TMA:

$$T_{A/D} = (I - T_{z1})\beta_A - \beta_D M_2(W + I - F_D\beta_A) + (T_{z1}\mathbf{G}_{L1} + \beta_D)[M_1F_D(I - T_{z1})\beta_A + M_1T_{z2}M_2(W + I - F_D\beta_A)] \quad (54)$$

where

$$M_1 = (I - F_DT_{z1}\mathbf{G}_{L1})^{-1}$$

$$M_2 = [I + [F_A(I - T_{z1}) - F_DT_{z1}]\mathbf{G}_{L1}M_1T_{z2}]^{-1}$$

and

$$W = [F_A(I - T_{z1}) + F_DT_{z1}][I - \mathbf{G}_{L1}M_1F_D(I - T_{z1})]\beta_A$$

A polynomial matrix manipulation computer program is of significant help in carrying out the indicated matrix operations. In such a program, currently in process of realization,* the elements of each of the matrices are put in a polynomial file, and the matrix operations indicated in eq. 54, for example, are carried out by simple commands. When the exact values of the transmission matrix elements have been found, the loss ratio and impedances can be found and automatically plotted.

While eq. (54) is far more complicated than (53), only two new matrices need be entered into file, namely F_D and F_A. Hence, the additional correction for the effects of direct feedthrough can be computed relatively easily, since most of the work involved in the computation is in entering the polynomial coefficients for the transistors and circuit elements into file. The instruction set for the computation consists essentially of the transmission matrix equation itself.

VIII. STABILITY

The above methods yield the transmission matrix of a feedback amplifier from which we derive a scalar measure of amplifier performance, such as loss ratio, in which we obtain the combined effect of the four matrix elements and the amplifier source and load immittances. For a linear, lumped-parameter circuit, the loss ratio will consist of a polynomial in the frequency variable, and may include a denominator polynomial, although this denominator often approximates unity. The condition for stability is that there shall be no roots of the (numerator) polynomial in the right-half plane of the complex frequency variable, since this would imply that, in the time domain, a growing exponential at the output could be supported with no input signal. The investigation of stability of distributed circuits, those containing transport delays, for example, is beyond our scope here, but these can be represented as

* By A. J. Osofsky

lumped systems by use of polynomial approximants for delay, such as the Padé approximants.[30] Hence, in the transmission matrix approach to amplifiers, stability is ascertained by direct investigation of the properties of what the conventional approach calls the closed-loop gain, or, in the present analysis, its reciprocal.

There are two aspects to the study of stability: investigation of the stability of a given amplifier, and design of an amplifier to be stable. A more restrictive form of the latter is to require the amplifier to have a prescribed transient response, since an amplifier which is merely stable may exhibit such damped oscillatory behavior as to be useless. The adjustment of the response of an amplifier to attain satisfactory transient response is termed *frequency compensation,* and involves adjustment of the coefficients of the loss ratio polynomial. In what follows, we shall study the case in which the loss ratio denominator is essentially unity.

Consider the loss ratio polynomial

$$L = \sum_{k=0}^{n} a_k s^k \tag{55}$$

We begin by normalizing the polynomial, to make the first and last terms unity, first by dividing throughout by a_0:

$$L = a_0 \left(1 + \sum_{k=1}^{n} \frac{a_k}{a_0} s^k\right) \tag{56}$$

Next, we change the frequency variable so that the coefficient of the highest-order term in the brackets is unity by letting

$$\frac{a_n}{a_0} s^n = p^n$$

or

$$s^k = \left(\frac{a_0}{a_n}\right)^{k/n} p^k \tag{57}$$

Thus,

$$L = a_0 \left(1 + \sum_{k=1}^{n-1} \frac{a_k}{{a_n}^{k/n} {a_0}^{1-k/n}} p^k + p^n\right) \tag{58}$$

The loss ratio is now in the desired form for investigation of stability and transient response. It may be written

$$L = a_0(1 + b_1 p + b_2 p^2 + \ldots + b_{n-1} p^{n-1} + p^n) \tag{59}$$

All information about the stability and transient response is contained in the values of the coefficients b_1 to b_{n-1}. In a cubic polynomial, for example, two coefficients, b_1 and b_2, determine the transient response.

Table III—Pascal-like triangles for normalized coefficients of the loss ratio polynomial

n	Multiple pole	Butterworth	Thomson
0	1	1	1
1	1 1	1 1	1 1
2	1 2 1	1 1.41 1	1 1.73 1
3	1 3 3 1	1 2 2 1	1 2.47 2.43 1
4	1 4 6 4 1	1 2.61 3.41 2.61 1	1 3.20 4.39 3.12 1

For a given set of polynomial coefficients, the roots are investigated to see if any lie in the right-half plane, in which case the amplifier is unstable. (For this determination, the normalization is unnecessary.) This is the only stability criterion necessary, since the design focuses on the performance of the amplifier, not a feedback loop. In traditional analysis, the focus was on the feedback loop and its analysis and design, so that an additional step, that of relating the closed-loop performance to the loop gain, had to be taken. The Nyquist criterion and its many later reinterpretations were worked out to ease this step. In the present method, these rather elaborate procedures are unnecessary. The loss ratio is found as the sum of the active-path and β-path contributions, and since the active path is usually expressible as a polynomial rather than a ratio of polynomials, the addition is simply made by adding the polynomial coefficients of the two paths. Denominators do arise. In the active path, these come from direct feedthrough and sometimes from frequency compensation networks which are used to adjust the polynomial coefficients to secure a prescribed transient response. In the β path, denominators arise when this path is used for equalization and filter applications. In these cases, we have no choice but to do the necessary multiplications to put both path polynomials over a common denominator.

In amplifier design where the denominators are incidental, prescribed transient response is obtained by designing the circuit such that the b coefficients of eq. (59) satisfy the performance criteria. Conversely, we may take the b coefficients as a performance specification for the amplifier. Examples of such criteria are given in Table III, which lists the b-values for an amplifier having either Butterworth or Thomson response characteristics.[31] Circuit methods for the adjustment of the b values to agree with a set of values such as those of Table III are beyond our scope in this paper, but a few comments are in order. The value of a_0 in eq. (59) is primarily established by the β path where the benefits of feedback (input augmentation) are to be obtained. The original reason for this was to reduce distortion introduced by the active devices, since the β path is linear and the active path is not, so that the β-path contribution was arranged to swamp out the smaller nonlinear contribution to the input signal. The second coefficient, $a_0 b_1$, as well as the remaining

frequency-dependent coefficients are ordinarily supplied by the active path, although it may be advantageous to have the β path supply and even dominate a_0b_1. In the case of C-feedback, for example, the a_0b_1 term is augmented by connecting a capacitor (a linear capacitor) between the output and the input in parallel with the β-path conductance. The a_0b_2 coefficient is adjusted upward by connecting capacitive feedback internal to the active path such that the capacitive current thus generated is multiplied by one of the active device matrix elements which is proportional to frequency, thereby augmenting the second-order coefficient, and so on through the set of b coefficients.

IX. DISCUSSION AND CONCLUSIONS

Viewed from both practical and theoretical standpoints, the process of analyzing, designing, and even thinking about active two-port circuits is simplified by taking an anticausal approach to the functional dependencies in the circuit. It does this because the importance of feedback or loop gain is greatly reduced, and with it denominators of the circuit expressions, which no longer depart greatly from unity.

The specific method described here for anticausal analysis of circuits is to base their transducer characteristics on the transmission matrix. This matrix puts cascades of two-ports into anticausal form directly, leaving the problem of how more remote circuit coupling is to be accommodated. In the method described here, such coupling is taken to be the property of spanning networks, which are described by the appropriate set of two-port parameters (h, z, y, or g). Each such spanning network parameter set yields four separate transmission matrices, each containing one of the four spanning network parameters, and each corresponding to one of four effects which are to be accounted for when the spanning network is applied to modify the amplifier characteristics. The 11 parameter and its associated transmission matrix corresponds to input circuit loading by the feedback network; the 12 parameter and its associated β matrix represents the input signal augmentation corresponding to the feedback signal of conventional analysis; the 21 parameter yields a transmission matrix which accounts for direct feedthrough of signals from circuit input to output through the spanning network; and the 22 parameter represents output circuit loading by the spanning network.

With all circuit element characteristics expressed as transmission matrices, it is desirable to be able to describe the whole circuit in these terms. The transmission matrix signal-flow graph, with its one-to-one correspondence between circuit vector nodes and graph nodes, provides a means for writing the transmission matrix equation of the whole circuit from the individual transmission matrices and their topological relationships. The transmission matrix array is a clearer way of showing the functional dependencies established by the transmission matrix sig-

nal-flow graph. From either of these two artifices, the transmission matrix equation for the whole may be written directly. This transmission matrix equation can be used for an initial look at a circuit to establish the basic properties of the configuration, by making suitable approximations such as that obtained by placing the transistors in their reference condition, through more accurate intermediate levels of approximation, by including the more important transistor parameters. Finally, an exact transmission matrix for the whole circuit may be derived, within the accuracy of the transistor and circuit element characterization available, traceable from initial approximation to final result.

Many problems remain, the most immediate of which is to complete the computational tools for linear analysis, and beyond that, the extension to quasilinear analysis and distortion, for which the present approach appears to offer substantial benefits. Active device characterization should be done in terms of anticausal functional dependencies: for linear analysis, we require transmission parameters of the active devices, an example of which is shown in Fig. 5. Nonlinear characterization of the partial derivatives of eq. (3) expressing the input signal vector as a function of the output, is needed. The noise of a two-port can be expressed as an equivalent input noise network including a series voltage generator and a shunt current generator.[32] It should be convenient to express not only the noise, but the predistortion, the dc input offsets, and the variation of the input signal vector due to transistor parameter variations, as an "input uncertainty network" consisting of a series voltage generator and a shunt current generator which in sum include all of these effects.

Beyond the two-port analysis discussed here, there are many instances where multiport analysis is needed. As a simple example, the operational amplifier with its positive and negative input leads can be considered a three-port (leaving out the power supply leads, which in most applications are at signal ground). The circuit partitioning resulting from the separation between the device supplier and user requires a three-port characterization, and with it, resolution of the question of functional dependencies of the six signal variables involved, comprising the two input signal vectors and the output signal vector.

It may be time to rid ourselves of the notion of feedback as a central concept in analysis of electronic amplifiers and other deterministic physical systems. As it applies to mercantile or social systems, where the reaction to a given event is barely predictable, the idea may still be of use, as for example, in the Club of Rome report.[33] Even in this area, anticausal analysis may supplant it. In project management, the PERT system, originally applied to the Polaris submarine, starts with the project goal and its projected date of completion, and works back to distinct events which must have happened to reach the goal.[34] In a sense,

the PERT chart is a TMSFG, without feedback loops. In comparison, the flow diagram of the world model (Fig. 26 of Ref. 33) is a feedback-loop-filled diagram inaccessible to human understanding. Were the projected goals introduced in that report used as flow graph *inputs*, the complex interrelationships among the variables might have been more readily understood. The mathematical description of feedback came out of the development of electronic amplifiers for carrier transmission, and has been widely adopted in other areas. The alternative suggested here might also find use in other areas.

X. ACKNOWLEDGMENTS

The work reported here is enough different from conventional amplifier theory that the author has had to rely on the experience and insights of many people. For the initial impulse to formalize an amplifier design method which the author has been using for some time, he is indebted to William McGee of Bell Northern Research, who invited him to give a paper in Toronto in 1973. For the kind of encouragement that only practical experimental realization can give, the author is indebted to Dan Wolaver and Walter Kruppa for their work on a UHF integrated operational amplifier, which was designed by use of the methods reported here. Special thanks are due Abraham Osofsky for developing a computer program (to be described in a future publication) which greatly enhances the usefulness of the method, allowing us to complete the traceable path from approximation to exact characterization.

The formalization of the method was advanced significantly by many class discussions during an in-hours course on amplifier design taught by the author in the fall of 1975; the efforts of the students of this course allowed the author to avoid many hazards of fuzzy thinking. For valuable discussions arising from this course, the author is indebted to J. A. Bellisio, W. I. H. Chen, D. L. Duttweiler, S. D. Personick, and J. M. Sipress. J. C. Candy and L. A. O'Neill offered valuable advice on the manuscript. Finally, the constant encouragement and insights of M. R. Aaron, who reviewed the key concepts as they evolved, are deeply appreciated.

GLOSSARY

Several terms have been introduced in this paper. For convenience, they are gathered here with brief definitions and the section number where they first appear.

Transmission matrix signal flow graph (III). A signal flow graph having signal vectors at circuit vector nodes for graph nodes and transmission matrices for branches.

Vector node or circuit vector node (III). A circuit node having only two

connections to it, allowing us to define uniquely the node voltage (to ground) and the node current, which together form the signal vector at the corresponding TMSFG node.

Loss ratio (III). The ratio of the generator voltage, e_G, to the output voltage, v_o, of a two-port circuit connected between a Thevenin source, e_G, R_G, and a load conductance, G_L.

Cascade graph (III). A signal flow graph or TMSFG having no feedback loops.

Spanning network (IV). A two-port network connected between two nonadjacent pairs of circuit vector nodes of a cascaded or ladder network. A spanning network is represented by one of the four sets of two-port parameters, h, z, y, or g.

Input signal augmentation or feedback (IV). The increase in input signal voltage or current (at constant output) due to the action of the spanning network; in particular, input augmentation is due to the 12 parameter (such as y_{12}) of the spanning network.

Direct feedthrough or *feedforward* (IV). The increase of change in the output signal voltage or current due to the action of the spanning network; in particular, direct feedthrough is due to the 21 parameter (such as y_{21}) of the spanning network.

Input circuit loading (IV). Shunt or series loading of a ladder network by the 11 parameter (such as y_{11} or z_{11}) of the spanning network.

Output circuit loading (IV). Shunt or series loading of a ladder network by the 22 parameter (such as y_{22} or z_{22}) of a spanning network.

β matrix (IV). A transmission matrix containing one nonzero element equal to the 12 parameter of a spanning network. Usually carries a subscript indicating which parameter set it is associated with, as in eq. (7).

F matrix (IV). A transmission matrix containing one nonzero element equal to the 21 element of the spanning network. Also called the direct feedthrough matrix. See eq. (8).

H matrix or *input loading matrix* (IV). A transmission matrix which is the sum of the identity matrix and a matrix having one nonzero element equal to the 11 parameter of the spanning network, as in eq. (9).

J matrix or *output loading matrix* (IV). A transmission matrix which is the sum of the identity matrix and a matrix having one nonzero element equal to the 22 parameter of the spanning network, as in eq. (10).

Return ratio matrix (IV). A transmission matrix equal to the loop gain of a feedback loop in a TMSFG.

Return difference matrix inverse (IV). A transmission matrix which postmultiplies the active path transmission matrix to account for the effect of a feedback loop.

Reference condition for a feedback circuit (V). A feedback circuit in which the active element(s) is (are) replaced by ideal two-port amplifier(s).

Ideal two-port amplifier (V). An amplifier whose transmission matrix is null.

Unitary feedback amplifier (V). An amplifier whose transmission matrix contains but one nonzero element when its active elements are placed in the reference condition.

Hybrid feedback amplifier (V). An amplifier whose transmission matrix contains more than one nonzero element when its active elements are placed in the reference condition.

Equivalent ladder circuit (VI). An equivalent circuit, drawn in ladder form, with remote couplings expressed by dependent generators.

Transmission matrix array (VII). An incidence matrix of the TMSFG which relates the received signals at a set of nodes to the transmitted signals at a set of nodes.

Received signal vector (VII). The set of signals at all nodes of a TMSFG which receive signals.

Transmitted signal vector (VII). The set of signals at all nodes of a TMSFG which transmit signals.

APPENDIX A

In what follows, we solve the simultaneous equations for the C-feedback amplifier of Fig. 6a. The TMSFG of Fig. 8b is repeated in Fig. 19a from which we can write (leaving out loading matrices H_y and J_y for the moment)

$$u_i = \beta_y u_o + u_b \tag{60}$$

$$u_b = T_a u_c \tag{61}$$

$$u_c = F_y u_b + u_o \tag{62}$$

Substituting (61) in (63), we form a self-loop at node u_c, as shown in Fig. 19b:

$$u_c = F_y T_a u_c + u_o \tag{63}$$

in which the matrix $F_y T_a$ will be called, following Bode's notation, the *return ratio matrix*. Solving for u_c, we have

$$u_c = (I - F_y T_a)^{-1} u_o \tag{64}$$

The matrix $I - F_y T_a$ corresponds to Bode's return difference, so that we term $(I - F_y T_a)^{-1}$ the *return difference matrix inverse*. Substituting (64) in (61), and thence in (60), we obtain

$$u_i = [\beta_y + T_a (I - F_y T_a)^{-1}] u_o \tag{65}$$

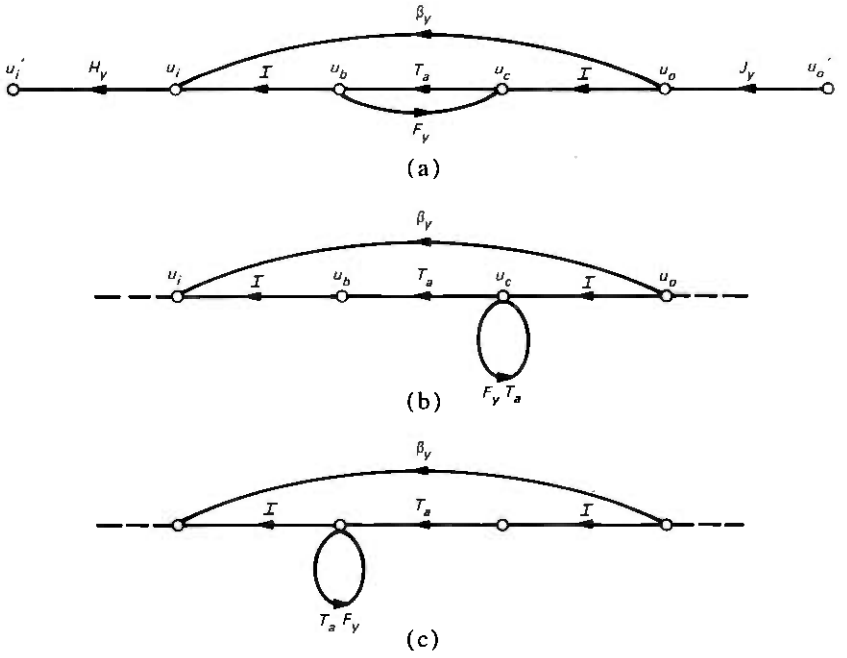


Fig. 19—Alternative TMSFG reductions of the C-feedback amplifier. (a) TMSFG from Fig. 8b; (b) creation of a self-loop at node u_c; (c) creation of a self-loop at node u_b. The preferred form is that of (b).

The transmission matrix for the stage is obtained by premultiplying by the input loading matrix, H_y, and postmultiplying by J_y:

$$T_y = H_y[\beta_y + T_a(I - F_yT_a)^{-1}]J_y \tag{66}$$

which is eq. (11) of the text.

Alternatively, we could solve the simultaneous equations by substituting (62) in (61), forming a self-loop at node u_b, as shown in Fig. 19c:

$$u_b = T_aF_yu_b + T_au_o \tag{67}$$

where T_aF_y is the new return ratio matrix. Solving for u_b, we also obtain a new return difference matrix inverse:

$$u_b = (I - T_aF_y)^{-1}T_au_o \tag{68}$$

from which we obtain

$$u_i = [\beta_y + (I - T_aF_y)^{-1}T_a]u_o \tag{69}$$

Clearly, $T_aF_y \neq F_yT_a$, since, as anyone knows who has tried to clean his glasses and blow his nose with the same tissue, the order in which the operation is carried out is important. It should not disturb the reader

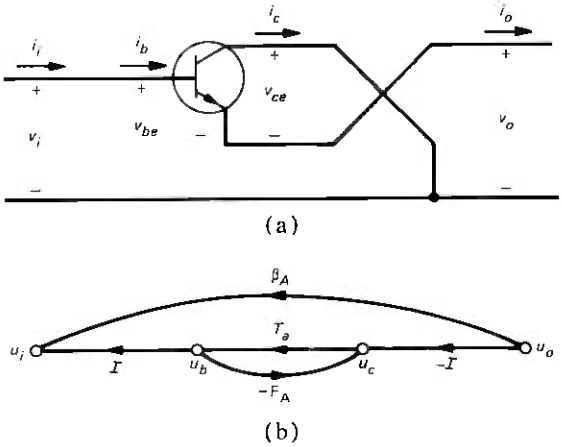


Fig. 20—Common collector stage. (a) Circuit and (b) TMSFG representing the transmission matrix equation derived in the text.

that the return ratio and return difference are dependent upon the way in which we solve the simultaneous equations, since these quantities are not invariants of the circuit, but depend entirely upon how we view the circuit.[15] On the other hand, comparing eqs. (65) and (69), $T_a(I - F_yT_a)^{-1} = (I - T_aF_y)^{-1}T_a$, so that the transmission matrix of the active path is invariant. For linear analysis, we are free to use either formulation. When we consider the extension to nonlinear analysis, however, eq (65) is preferred, since it preserves the actual signal level at u_c, the output node of the active device.* We therefore adopt a rule of procedure for solving simultaneous circuit equations: Always preserve the output node. This is done by placing the self loop at the output of a device exhibiting direct feedthrough, and allows a straightforward calculation of the waveform at the output of a nonlinear device. As a matter of practice, the node equation for a node nearer the input should be substituted into the node equation for a node nearer the output.

APPENDIX B

Common collector stage

The common collector stage is shown in Figure 20a. The circuit equations are written starting at the input:

$$v_i = v_b + v_o \tag{70}$$

$$i_i = i_b$$

* The advantage of the formulation of eq. (65) over that of eq. (69) was pointed out to the author by C. A. Desoer.

or

$$u_i = u_b + \beta_A u_o \tag{71}$$

where

$$\beta_A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \tag{72}$$

Next,

$$u_b = T_a u_c \tag{73}$$

and

$$v_{ce} = -v_o \tag{74}$$
$$i_c = -i_o + i_b$$

or

$$u_c = -u_o - F_A u_b \tag{75}$$

where

$$F_A = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \tag{76}$$

Substituting (73) into (75), and solving for u_c, we have

$$u_c = -(I + F_A T_a)^{-1} u_o \tag{77}$$

By successive substitution, we obtain the input vector as a function of the output:

$$u_i = [\beta_A - T_a(I + F_A T_a)^{-1}] u_o \tag{78}$$

To evaluate the matrix of the common collector stage, we obtain the return ratio matrix:

$$F_A T_a = -\begin{bmatrix} 0 & 0 \\ C & D \end{bmatrix} \tag{79}$$

The return difference matrix inverse is

$$(I + F_A T_a)^{-1} = \frac{1}{1-D}\begin{bmatrix} 1-D & 0 \\ C & 1 \end{bmatrix} \tag{80}$$

and the matrix for the stage is

$$T_{cc} = \frac{1}{1-D}\begin{bmatrix} 1-A-D+\Delta^t & -B \\ -C & -D \end{bmatrix} \tag{81}$$

as given in eq. (33). From eq. (78), we can draw a TMSFG for the stage as shown in Fig. 20b.

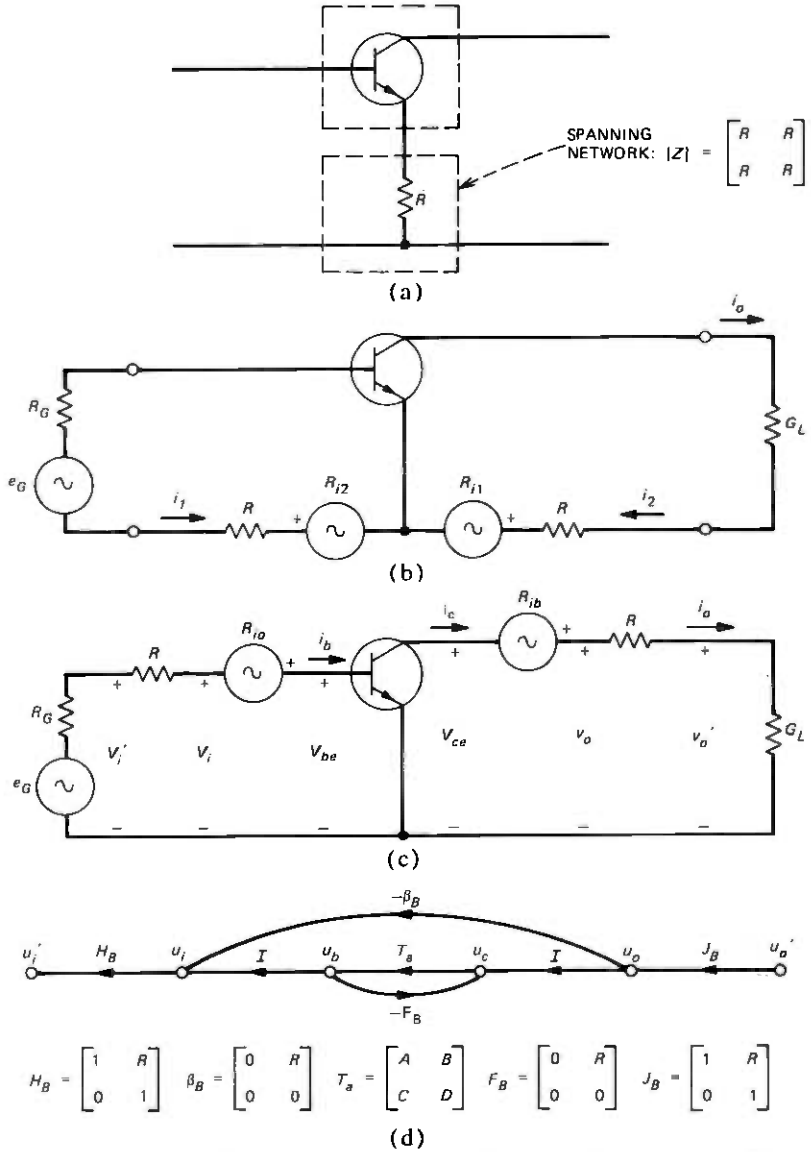


Fig. 21—Emitter resistor feedback analysis. (a) Circuit. (b) Circuit redrawn using dependent generator equivalent circuit of Fig. 7. (c) Redrawn circuit, interchanging position of series elements. (d) TMSFG with definitions of the matrices.

APPENDIX C

Emitter resistor feedback

A transistor with unitary B feedback is shown in Fig. 21. In (a), the circuit is divided into an active path and a resistive spanning network,

and in (b), the spanning network is represented by its dependent generator equivalent circuit from Fig. 7b. A common ground does not exist between the input and output loops, but the circuit as drawn in (b) is nevertheless a two-port, since $i_1 = -i_b$, and $i_2 = i_o$ with the current directions given in the figure. An equivalent representation is given in (c) of the figure, in which elements in series have had their circuit positions interchanged. The TMSFG for the circuit of (c) is given in (d), in which H_B and J_B represent the series input and output resistances, $-\beta_B$ represents the generator in series with the input lead, and F_B is the direct feedthrough supplied by the generator in series with the output lead. From the TMSFG, the transmission matrix equation is

$$T_z = H_B[-\beta_B + T_a(I + F_B T_a)^{-1}]J_B \tag{82}$$

With the element values of the matrices given in the figure, the return ratio matrix is

$$F_B T_a = \begin{bmatrix} CR & DR \\ 0 & 0 \end{bmatrix} \tag{83}$$

and the return difference matrix inverse is

$$(I + F_B T_a)^{-1} = \frac{1}{1 + CR}\begin{bmatrix} 1 & -DR \\ 0 & 1 + CR \end{bmatrix} \tag{84}$$

from which the active path matrix without loading becomes

$$T_a(I + F_B T_a)^{-1} = \frac{1}{1 + CR}\begin{bmatrix} A & B - R\Delta^t \\ C & D \end{bmatrix} \tag{85}$$

Adding the input augmentation from $z_{12} = R$, we have

$$-\beta_B + T_a(I + F_B T_a)^{-1} = \frac{1}{1 + CR}\begin{bmatrix} A & -R(1 + CR) + B - R\Delta^t \\ C & D \end{bmatrix} \tag{86}$$

Finally, we premultiply by H_B and postmultiply by J_B, and obtain

$$T_z = \frac{1}{1 + CR}\begin{bmatrix} A + CR & B - R(1 - A - D + \Delta^t) \\ C & D + CR \end{bmatrix} \tag{87}$$

as shown in eq. (35).

REFERENCES

1. O. Mayr, *The Origins of Feedback Control*, Cambridge, Mass.: The MIT Press, 1970.
2. H. Bateman, "The Control of an Elastic Fluid," Bull. Amer. Math. Soc., *51*, pp. 601–646. Also reprinted in Ref. 4.
3. J. C. Maxwell, "On Governors," Proc. Royal Soc. London, *16* (1868), pp. 270–283, reprinted in Ref. 4.
4. R. Bellman and R. Kalaba, ed., *Selected Papers on Mathematical Trends in Control Theory*, New York: Dover Publications, 1964.

5. N. Minorski, "Directional Stability of Automatically Steered Bodies," J. Am. Soc. Nav. Eng., *34* (1922), p. 280.
6. H. S. Black, "Translating System," U.S. Patent No. 1,686,792, issued December, 1937.
7. H. S. Black, private communication, January 21, 1977.
8. H. S. Black, "Stabilized Feed-back Amplifiers," B.S.T.J., *13*, (January 1934), p. 1.
9. H. S. Black, "Wave Translation System," U.S. Patent No. 2,102,671, filed April 1932, issued December 21, 1938.
10. H. Nyquist, "Regeneration Theory," B.S.T.J., *11* (1932), pp. 126–147.
11. P. C. Mabon, *Mission Communications*, Bell Telephone Laboratories, 1975, p. 45.
12. H. W. Bode, *Network Analysis and Feedback Amplifier Design*, Van Nostrand, 1945.
13. E. S. Kuh and R. A. Rohrer, *Theory of Linear Active Networks*, San Francisco: Holden-Day, Inc., 1967. See bibliography at the end of Chap. 11.
14. S. S. Haykin, *Active Network Theory*, Reading, Mass.: Addison-Wesley. See references at the end of Chap. 11.
15. F. D. Waldhauer, "Feedback—Conceptual or Physical?" Intl. Symp. on Circuit Theory, Toronto, 1973: IEEE Cat. No. 73CHO 76508CT, pp. 8–12.
16. S. J. Mason, "Feedback Theory—Some Further Properties of Signal Flow Graphs," Proc. IRE, *44*, No. 7, (July 1956), pp. 920–026.
17. G. J. Sussman and R. M. Stallman, "Heuristic Techniques in Computer-Aided Circuit Analysis," IEEE Trans. on Circuits and Systems, *CAS-22*, No. 11, November, 1975, pp. 857–865. See page 861 and Note 7. This reference has much to offer with respect to the concerns of the present paper.
18. F. D. Waldhauer, "Transistor Feedback Amplifiers," NEREM Record, 1963, Lewis B. Winner, New York.
19. B. Beddoe, "The Analysis of Feedback Amplifiers by Finding the Reciprocal of Gain," Electronic Engineering (G. B.), *36*, (February 1964), pp. 92–96.
20. H. H. Rosenbrook, "Design of Multivariable Control Systems Using the Inverse Nyquist Array," Proc. IEE, *116*, No. 11 (November 1969), pp. 1929–1935.
21. S. J. Mason and H. Zimmerman, *Electronic Circuits, Systems, and Signals*, New York: John Wiley and Sons, 1960.
22. L. Weinberg, *Network Analysis and Synthesis*, New York: McGraw-Hill, 1962.
23. L. P. Huelsman, *Circuits, Matrices, and Linear Vector Spaces*, New York: McGraw-Hill, 1963.
24. S. J. Mason, "Feedback Theory—Some Properties of Signal Flow Graphs," Proc. IRE, *41*, No. 9 (September 1953); pp. 1144–1157.
25. D. E. Riegle and P. M. Lin, "Matrix Signal Flow Graphs and an Optimum Topological Method for Evaluating their Gains," IEEE Trans., *CT-19*, No. 4 (September 1972), pp. 427–435.
26. H. J. Carlin, "Singular Network Elements," IEEE Trans., *CT-11*, No. 1 (March 1964), pp. 67–72.
27. S. J. Mason, "Power Gain in Feedback Amplifier," Trans. IRE, *CT-1*, No. 1, pp. 20–25.
28. E. M. Cherry and D. E. Hooper, "The Design of Wide-Band Transistor Feedback Amplifiers," Proc. IEE, *110*, No. 2 (February 1963), pp. 375–389.
29. T. J. Aprille, Jr., "Wideband Amplifier Design Using Major Multiloop Feedback Techniques," B.S.T.J. *54*, No. 7 (September 1975), p. 1253.
30. J. G. Truxal, "Automatic Feedback Control System Synthesis," New York: McGraw-Hill, 1955.
31. F. D. Waldhauer, "Analog Integrated Circuits of Large Bandwidth," 1963 IEEE Int. Convention Record, Part 2, pp. 200–207.
32. H. A. Haus and R. B. Adler, *Circuit Theory of Linear Noise Networks*, New York: John Wiley and Sons, 1959.
33. D. H. Meadows, D. L. Meadows, J. Randers, and W. W. Behrens, III, The Limits to Growth, second ed., New York: Universe Books, 1974.
34. Members of the Technical Staff of Bell Telephone Laboratories, *Physical Design of Electronic Systems*, Vol. IV, *Design Process*, In this book, see Chap. 4, "Project Scheduling," by P. W. McFadden.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 8, October 1977
Printed in U.S.A.

Vibrations of a Lithium Niobate Fiber

By LYNN O. WILSON

(Manuscript received March 8, 1977)

We discuss wave propagation along a crystalline piezoelectric fiber composed of lithium niobate or some other material in the trigonal 3m crystal class. The crystalline c axis is aligned with the fiber axis. We obtain an analytical description of all the vibrational modes. The method used is to make perturbation expansions about the modes of a hexagonal 6mm piezoelectric fiber, for which exact solutions are known.

I. INTRODUCTION

A single crystal of lithium niobate, grown in the form of a long fiber, has been considered for use as a low-loss acoustic delay line. Lithium niobate is of special interest because it is piezoelectric: it becomes electrically polarized when strained and, conversely, becomes strained when placed in an electric field. This piezoelectricity provides a means for electrically generating and detecting acoustic signals.

In this paper we study mathematically the vibrational properties of a $LiNbO_3$ crystal fiber, with the crystalline c axis aligned along the fiber axis. The problem is by no means simple. We illustrate this by giving a brief history of related problems for which exact solutions have been obtained. The elastic, or acoustic, wave equations for an infinitely long circularly cylindrical isotropic rod were solved exactly by Pochhammer[1] in 1876 and independently by Chree[2] in 1889. Even for an isotropic medium, exact solutions for a rod of finite length have not been obtained. It was not until 1965 that the next full exact solution was found. This was done by Mirsky,[3,4] who determined the vibrational modes of a circularly cylindrical rod consisting of a nonpiezoelectric medium which is transversely isotropic. Such a medium belongs to the hexagonal system of crystals; the crystalline c axis was aligned along the rod or fiber. Certain of the modes obtained by Mirsky, i.e., those which are azimuthally symmetric about the fiber axis, had also been obtained earlier.[5,6] Recently, the author and J. A. Morrison were able to solve the coupled acoustic and electromagnetic wave equations, in the customary quasi-static approximation, for piezoelectric transversely isotropic crystals

belonging to the hexagonal 6mm, 622, and 6 crystal classes.[7] Exact solutions were obtained for all the vibrational modes. The author is unaware of any other exact solutions, either for other crystals, or for other orientations of transversely isotropic crystals.

The difficulty lies in the acoustic wave equations which, for a general anisotropic medium, consist of three coupled wave equations for the three vector components of displacement. If piezoelectricity is added via the quasistatic approximation, for which the electric field is represented by the gradient of a potential, there are four coupled equations for four unknown functions. The boundary conditions may also involve all four functions coupled together. For the general anisotropic case, no method has been discovered to decouple the equations. For the specific crystals and orientations discussed above, it was possible to express the elastic displacements (and electric potential) in terms of three (or four) potential functions for which the wave equations decoupled.

Unfortunately, such a serendipitous situation does not exist for the lithium niobate fiber. It belongs to the trigonal 3m crystal class; we cannot expect to find an exact description of the vibrational modes. It will be possible, though, to find an approximate description by means of an infinite series perturbation expansion. We use a technique which is an extension of one used by the author to describe waves travelling along a sapphire fiber.[8] Sapphire is a nonpiezoelectric material belonging to the trigonal $\bar{3}$m crystal class. It is characterized by a stiffness matrix (used in the stress-strain relations) which has almost the same form as that for a transversely isotropic material. There is one additional stiffness coefficient. Since it turns out to be small in magnitude compared to the other stiffness coefficients, it is possible to describe the vibrational modes of a sapphire fiber (with the crystalline c axis aligned with the fiber axis) by means of perturbation expansions about the modes of a transversely isotropic fiber.

The situation for $LiNbO_3$ is similar, albeit somewhat more complicated. We will make an infinite series perturbation expansion about the known solutions for a hexagonal 6mm crystal. The same techniques, incidentally, can be used to describe vibrations of crystals in the trigonal 32 classes. We restrict ourselves to a discussion of trigonal 3m crystals only to keep the analysis from appearing extraordinarily complicated.

For the sapphire fiber, numerical results are available for the lowest-order torsional mode of vibration; they are presented in a paper by the author and M. A. Gatto.[9] A low-frequency asymptotic analysis for that mode was also performed by R. N. Thurston and the author.[10] Excellent numerical agreement between the results of the two independent theories provides a check on the rather complicated analyses involved and encourages us to extend the perturbation technique to a study of $LiNbO_3$.

In Section II we write down the basic equations of motion and boundary conditions. In Section III we apply the perturbation technique and introduce potential functions. In Section IV we solve the differential equations, and in Section V we sketch how to apply the boundary conditions.

Although it would be desirable to present numerical results as well, we shall not do so. Numerical results are not yet available for the unperturbed (hexagonal 6mm) problem. The computational effort required to describe quantitatively the vibrations of a lithium niobate fiber would be even greater than the considerable effort expended to present results for a sapphire fiber.

II. FORMULATION

Consider a single crystal of $LiNbO_3$ (or some other member of the trigonal 3m crystal class), grown in the form of a fiber of circular cross-section, with the crystallographic c axis along the fiber axis. We shall assume that the fiber is infinitely long and has radius R. We adopt a cylindrical coordinate system whose z axis coincides with the fiber axis.

In the quasistatic approximation, where the rotational part of the electric field is neglected, the basic differential equations are[11]

$$\nabla \cdot \mathbf{T} = \rho \frac{\partial^2 \mathbf{u}}{\partial t^2}, \tag{1}$$

$$\nabla \cdot \mathbf{D} = 0, \tag{2}$$

where $\mathbf{T}$ is the stress, $\mathbf{D}$ is the electric displacement, $\mathbf{u}$ is the elastic displacement, and ρ is the density. The properties of the specific crystal are introduced by means of the constitutive relations

$$\mathbf{T} = -\mathbf{e} \cdot \mathbf{E} + \mathbf{c}:\mathbf{S}, \tag{3}$$

$$\mathbf{D} = \boldsymbol{\epsilon} \cdot \mathbf{E} + \mathbf{e}:\mathbf{S}, \tag{4}$$

where

$$\mathbf{E} = -\nabla \Phi, \tag{5}$$

$$\mathbf{S} = \nabla_s \mathbf{u}. \tag{6}$$

Here $\mathbf{E}$ denotes the electric field, $\mathbf{S}$ the strain, and Φ the electric potential. The crystal is described by means of the elastic stiffness matrix $\mathbf{c}$, the piezoelectric stress matrix ϵ, and the dielectric permittivity at constant strain matrix $\boldsymbol{\epsilon}$. For a crystal in the trigonal 3m class, these matrices have the following forms in cylindrical coordinates:[12]

$$\mathbf{c} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14}C & c_{14}S & 0 \\ c_{12} & c_{11} & c_{13} & -c_{14}C & -c_{14}S & 0 \\ c_{13} & c_{13} & c_{33} & 0 & 0 & 0 \\ c_{14}C & -c_{14}C & 0 & c_{44} & 0 & -c_{14}S \\ c_{14}S & -c_{14}S & 0 & 0 & c_{44} & c_{14}C \\ 0 & 0 & 0 & -c_{14}S & c_{14}C & c_{66} \end{bmatrix}, \tag{7}$$

with

$$c_{66} = \tfrac{1}{2}(c_{11} - c_{12}), \tag{8}$$

$$C = \cos 3\theta, \quad S = \sin 3\theta. \tag{9}$$

$$\mathbf{e} = \begin{bmatrix} -e_{y2}S & e_{y2}S & 0 & 0 & e_{x5} & -e_{y2}C \\ -e_{y2}C & e_{y2}C & 0 & e_{x5} & 0 & e_{y2}S \\ e_{z1} & e_{z1} & e_{z3} & 0 & 0 & 0 \end{bmatrix} \tag{10}$$

$$\epsilon = \begin{bmatrix} \epsilon_{xx} & 0 & 0 \\ 0 & \epsilon_{xx} & 0 \\ 0 & 0 & \epsilon_{zz} \end{bmatrix}. \tag{11}$$

Let $\mathbf{n}$ denote a vector normal to the fiber surface, i.e., in the radial direction. For the three mechanical boundary conditions,[13] we shall specify either that the surface tractions vanish:

$$\mathbf{T} \cdot \mathbf{n} = 0 \text{ at } r = R \text{ (free surface)}, \tag{12}$$

or that there is no displacement at the surface:

$$\mathbf{u} = 0 \text{ at } r = R \text{ (clamped surface)}. \tag{13}$$

The free surface condition is the natural one to consider for an acoustic delay line; it is equally simple to show how to solve the problem for the clamped surface condition, so we include it, too.

For the electrical boundary condition,[13] we take either

$$\Phi = 0 \text{ at } r = R \text{ (short-circuit)}, \tag{14}$$

or

$$\mathbf{D} \cdot \mathbf{n} = 0 \text{ at } r = R \text{ (open-circuit)}. \tag{15}$$

The problem is to solve the four differential equations (1) and (2), in conjunction with eqs. (3) to (11), subject to four boundary conditions chosen from (12) to (15). Since we are concerned with waves travelling down the fiber, we assume the solution has an $\exp [i(\omega t - \beta z)]$ dependence, where ω is the angular frequency and β is the propagation constant; β will depend upon ω.

We begin by writing the differential equations and boundary conditions in dimensionless form. Let

$$\hat{c}_{ij} = c_{ij}/c, \quad c = \max_{i,j} |c_{ij}|,$$

$$\hat{e}_{ij} = e_{ij}/e, \quad e = \max_{i,j} |e_{ij}|,$$

$$\hat{\epsilon}_{ij} = \epsilon_{ij}/\epsilon, \quad \epsilon = \max(\epsilon_{xx}, \epsilon_{zz}). \tag{16}$$

Normalize $\mathbf{u}$ with respect to R, Φ with respct to Re/ϵ, β with respect to R^{-1}, and ω with respect to $(c/\rho)^{1/2}/R$. To simplify notation, we use the same symbols as we used for dimensional quantities, except for the hats on c_{ij}, e_{ij}, and ϵ_{ij}. Upon substituting eqs. (3) to (11) into (1) and (2), we can write the dimensionless differential equations in cylindrical coordinates as

$$\begin{aligned}
&\hat{c}_{11}\left(u_{rr} + \frac{1}{r}u_r - \frac{1}{r^2}u\right) + \hat{c}_{66}\frac{1}{r^2}u_{\theta\theta} + (\omega^2 - \beta^2\hat{c}_{44})u \\
&\quad - 2i\beta\hat{c}_{14}\cos 3\theta\,\frac{1}{r}u_\theta - 2i\beta\hat{c}_{14}\sin 3\theta\left(u_r - \frac{1}{r}u\right) \\
&\quad + (\hat{c}_{12} + \hat{c}_{66})\frac{1}{r}v_{r\theta} - (\hat{c}_{11} + \hat{c}_{66})\frac{1}{r^2}v_\theta \\
&\quad - 2i\beta\hat{c}_{14}\cos 3\theta\left(v_r - \frac{1}{r}v\right) + 2i\beta\hat{c}_{14}\sin 3\theta\,\frac{1}{r}v_\theta \\
&\quad - i\beta(\hat{c}_{13} + \hat{c}_{44})w_r + 2\hat{c}_{14}\cos 3\theta\left(\frac{1}{r}w_{r\theta} - \frac{1}{r^2}w_\theta\right) \\
&\quad + \hat{c}_{14}\sin 3\theta\left(w_{rr} - \frac{1}{r}w_r - \frac{1}{r^2}w_{\theta\theta}\right) \\
&\quad - i\beta\tau(\hat{e}_{x5} + \hat{e}_{z1})\Phi_r - 2\tau\hat{e}_{y2}\cos 3\theta\left(\frac{1}{r}\Phi_{r\theta} - \frac{1}{r^2}\Phi_\theta\right) \\
&\quad - \tau\hat{e}_{y2}\sin 3\theta\left(\Phi_{rr} - \frac{1}{r}\Phi_r - \frac{1}{r^2}\Phi_{\theta\theta}\right) = 0,
\end{aligned}$$

$$\begin{aligned}
&(\hat{c}_{12} + \hat{c}_{66})\frac{1}{r}u_{r\theta} + (\hat{c}_{11} + \hat{c}_{66})\frac{1}{r^2}u_\theta - 2i\beta\hat{c}_{14}\cos 3\theta\left(u_r - \frac{1}{r}u\right) \\
&\quad + 2i\beta\hat{c}_{14}\sin 3\theta\,\frac{1}{r}u_\theta + \hat{c}_{66}\left(v_{rr} + \frac{1}{r}v_r - \frac{1}{r^2}v\right) + \hat{c}_{11}\frac{1}{r^2}v_{\theta\theta} \\
&\quad + (\omega^2 - \beta^2\hat{c}_{44})v + 2i\beta\hat{c}_{14}\cos 3\theta\,\frac{1}{r}v_\theta + 2i\beta\hat{c}_{14}\sin 3\theta\left(v_r - \frac{1}{r}v\right) \\
&\quad - i\beta(\hat{c}_{13} + \hat{c}_{44})\frac{1}{r}w_\theta + \hat{c}_{14}\cos 3\theta\left(w_{rr} - \frac{1}{r}w_r - \frac{1}{r^2}w_{\theta\theta}\right)
\end{aligned}$$

$$
\begin{aligned}
&- 2\hat{c}_{14}\sin 3\theta\left(\frac{1}{r}w_{r\theta} - \frac{1}{r^2}w_\theta\right) - i\beta\tau(\hat{e}_{x5} + \hat{e}_{z1})\frac{1}{r}\Phi_\theta - \tau\hat{e}_{y2}\\
&\times \cos 3\theta\left(\Phi_{rr} - \frac{1}{r}\Phi_r - \frac{1}{r^2}\Phi_{\theta\theta}\right) + 2\tau\hat{e}_{y2}\sin 3\theta\left(\frac{1}{r}\Phi_{r\theta} - \frac{1}{r^2}\Phi_\theta\right) = 0,\\
&- i\beta(\hat{c}_{13} + \hat{c}_{44})\left(u_r + \frac{1}{r}u\right) + 2\hat{c}_{14}\cos 3\theta\left(\frac{1}{r}u_{r\theta} - \frac{2}{r^2}u_\theta\right)\\
&\qquad + \hat{c}_{14}\sin 3\theta\left(u_{rr} - \frac{1}{r^2}u_{\theta\theta} - \frac{3}{r}u_r + \frac{3}{r^2}u\right)\\
&- i\beta(\hat{c}_{13} + \hat{c}_{44})\frac{1}{r}v_\theta + \hat{c}_{14}\cos 3\theta\left(v_{rr} - \frac{1}{r^2}v_{\theta\theta} - \frac{3}{r}v_r + \frac{3}{r^2}v\right)\\
&- 2\hat{c}_{14}\sin 3\theta\left(\frac{1}{r}v_{r\theta} - \frac{2}{r^2}v_\theta\right) + \hat{c}_{44}\left(w_{rr} + \frac{1}{r}w_r + \frac{1}{r^2}w_{\theta\theta}\right)\\
&\qquad + (\omega^2 - \beta^2\hat{c}_{33})w + \tau\hat{e}_{x5}\left(\Phi_{rr} + \frac{1}{r}\Phi_r + \frac{1}{r^2}\Phi_{\theta\theta}\right) - \tau\beta^2\hat{e}_{z3}\Phi = 0,\\
&- i\beta(\hat{e}_{x5} + \hat{e}_{z1})\left(u_r + \frac{1}{r}u\right) - 2\hat{e}_{y2}\cos 3\theta\left(\frac{1}{r}u_{r\theta} - \frac{2}{r^2}u_\theta\right)\\
&\qquad - \hat{e}_{y2}\sin 3\theta\left(u_{rr} - \frac{1}{r^2}u_{\theta\theta} - \frac{3}{r}u_r + \frac{3}{r^2}u\right)\\
&- i\beta(\hat{e}_{x5} + \hat{e}_{z1})\frac{1}{r}v_\theta - \hat{e}_{y2}\cos 3\theta\left(v_{rr} - \frac{1}{r^2}v_{\theta\theta} - \frac{3}{r}v_r + \frac{3}{r^2}v\right)\\
&+ 2\hat{e}_{y2}\sin 3\theta\left(\frac{1}{r}v_{r\theta} - \frac{2}{r^2}v_\theta\right) + \hat{e}_{x5}\left(w_{rr} + \frac{1}{r}w_r + \frac{1}{r^2}w_{\theta\theta}\right)\\
&\qquad - \beta^2\hat{\epsilon}_{z3}w - \hat{\epsilon}_{xx}\left(\Phi_{rr} + \frac{1}{r}\Phi_r + \frac{1}{r^2}\Phi_{\theta\theta}\right) + \beta^2\hat{\epsilon}_{zz}\Phi = 0, \quad (17)
\end{aligned}
$$

where

$$\tau = \frac{e^2}{\epsilon c}, \quad (18)$$

and u, v, and w are the radial, azimuthal, and longitudinal components of the displacement vector $\mathbf{u}$.

In dimensionless form, the boundary conditions (12) to (15) are

Free surface:

$$
\begin{aligned}
&\hat{c}_{11}u_r + \hat{c}_{12}(u + v_\theta) - i\beta\hat{c}_{13}w - i\beta\tau\hat{e}_{z1}\Phi + \cos 3\theta[\hat{c}_{14}(-i\beta v + w_\theta)\\
&\qquad - \tau\hat{e}_{y2}\Phi_\theta] + \sin 3\theta[\hat{c}_{14}(-i\beta u + w_r) - \tau\hat{e}_{y2}\Phi_r] = 0,\\
&\hat{c}_{66}(u_\theta + v_r - v) + \cos 3\theta[\hat{c}_{14}(-i\beta u + w_r) - \tau\hat{e}_{y2}\Phi_r]\\
&\qquad - \sin 3\theta[\hat{c}_{14}(-i\beta v + w_\theta) - \tau\hat{e}_{y2}\Phi_\theta] = 0,
\end{aligned}
$$

$$\hat{c}_{44}(-i\beta u + w_r) + \tau\hat{e}_{x5}\Phi_r + \hat{c}_{14}\cos 3\theta(u_\theta + v_r - v) + \hat{c}_{14}\sin 3\theta(u_r - u - v_\theta) = 0 \text{ at } r = 1.$$

Clamped surface:

$$u = v = w = 0 \text{ at } r = 1.$$

Short-circuit:

$$\Phi = 0 \text{ at } r = 1.$$

Open-circuit:

$$-\hat{\epsilon}_{xx}\Phi_r + \hat{e}_{x5}(-i\beta u + w_r) - \hat{e}_{y2}\cos 3\theta(u_\theta + v_r - v) + \hat{e}_{y2}\sin 3\theta(u - u_r + v_\theta) = 0 \text{ at } r = 1. \quad (19)$$

III. PERTURBATIONS AND POTENTIALS

At any given frequency ω, we wish to solve the differential equations (17) and boundary conditions (19) for the elastic displacement components u, v, and w, and for the electric potential Φ; these are functions of r and θ. We also need to determine the propagation constant β. Unfortunately, we have been unable to obtain an exact solution. We shall find an approximate solution by combining two techniques which were applied successfully in earlier papers.[7,8] First, we observe that eqs. (17) and (19) have an exact solution if $\hat{c}_{14} = \hat{e}_{y2} = 0$.[7] In this case, the crystal is a member of the hexagonal 6mm class. We make an infinite series perturbation expansion about any modal solution to that problem. This results in systems of differential equations and boundary conditions for the perturbation contributions to the elastic displacement and electric potential. Second, we write these perturbation contributions in terms of certain potential functions. The differential equations then decouple. With the aid of the boundary conditions, the potential functions can be determined; perturbation contributions to the propagation constant can also be found.

The perturbation technique has been used to describe vibrations of a sapphire fiber.[8] The equations describing that crystal can be obtained from eqs. (17) and (19) by setting Φ and the components of the piezoelectric stress matrix e to zero.

The potential function technique used here is the same as the one used in obtaining an exact description of the vibrations of a fiber in the hexagonal 6mm class.[7]

For lithium niobate, we find from the definition (16) and the numerical values for the stiffness coefficients[12] that $\hat{c}_{14} \approx 3.6 \times 10^{-2}$. We will use $\hat{c}_{14}$ as a perturbation parameter. This is reasonable since it is small compared to one. Instead of treating $\hat{e}_{y2}$ as a separate perturbation parameter, we write it as a constant multiple of $\hat{c}_{14}$:

$$\hat{e}_{y2} = \xi \hat{c}_{14}. \tag{20}$$

For lithium niobate, it turns out that $\hat{e}_{y2} \approx 6.8 \times 10^{-1}$ and $\xi \approx 18$.[12] The perturbation scheme would work better if $\hat{e}_{y2}$ were smaller than this. It effectively is, in three out of four differential equations and in all but the open-circuit boundary condition, for it is then multiplied by the dimensionless constant $\tau \approx 1.4 \times 10^{-1}$. In the remaining differential equation and boundary condition, however, $\hat{e}_{y2}$ is not multiplied by a small constant in this fashion. How rapidly the perturbation series actually converges will have to be determined numerically.

We first make a perturbation expansion for the propagation constant:

$$\beta = \sum_{m=0}^{\infty} (\hat{c}_{14})^m \beta_m. \tag{21}$$

When we make a perturbation expansion for the elastic displacements and electric potential, it is convenient also to make a Fourier expansion in θ. Because of the three-fold symmetry of the crystal about the z axis, the Fourier expansion only needs to include multiples of 3θ, rather than θ. With Z used to represent u, v, w, or Φ, we assume that

$$Z(r, \theta) = \sum_{m=0}^{\infty} (\hat{c}_{14})^m \sum_{n=-\infty}^{\infty} e^{iN\theta} e^{i3n\theta} Z^{m,n}(r). \tag{22}$$

To begin the perturbation scheme, we choose (for $m = n = 0$) $u^{0,0}(r)e^{iN\theta}, \ldots, \Phi^{0,0}(r)e^{iN\theta}$, and β_0 to be a modal solution to the unperturbed problem, i.e., that for a hexagonal 6mm crystal. N can be any integer. It determines which type of modal solution is being considered. $N = 0$ corresponds to an azimuthally symmetric mode, $|N| = 1$ to a flexural mode, and $|N| > 1$ to a higher-order flexural mode. For $m = 0$ and $n \neq 0$, set $u^{0,n}, \ldots, \Phi^{0,n}$ to zero. The problem then is to determine $u^{m,n}(r)$, $\ldots, \Phi^{m,n}(r)$, and β_m for $m > 0$. We will see that the displacement and electric potential contributions vanish when $|n| > m$. The functions thus need only be determined in the "triangular" region $m = 0, 1, 2, 3, \ldots$ and $|n| \leq m$.

We next write the perturbation contributions to the elastic displacements and electric potential in terms of certain potential functions:

$$u^{m,n}(r) = \frac{d}{dr} \sum_{\ell=1}^{3} \psi_\ell^{m,n}(r) - \frac{s}{r} \psi_4^{m,n}(r),$$

$$v^{m,n}(r) = i \left[\frac{s}{r} \sum_{\ell=1}^{3} \psi_\ell^{m,n}(r) - \frac{d\psi_4^{m,n}(r)}{dr} \right],$$

$$w^{m,n}(r) = i \sum_{\ell=1}^{3} \mu_\ell \psi_\ell^{m,n}(r),$$

$$\Phi^{m,n}(r) = i \sum_{\ell=1}^{3} \eta_\ell \psi_\ell^{m,n}(r), \tag{23}$$

with

$$s = 3n + N. \tag{24}$$

The μ_ℓ and η_ℓ, $\ell = 1, 2, 3$, are constants (independent of m and n) which must be determined.

Substitute the potential functions defined in (23) into the perturbation expansions. Substitute these, in turn, into the differential equations (17). After considerable algebra, we find that the terms multiplied by $(\hat{c}_{14})^m e^{i(3n+N)\theta}$ yield the following system of differential equations for $\psi_1^{m,n}, \ldots, \psi_4^{m,n}$.

$$\frac{d}{dr} \sum_{\ell=1}^{3} \{\hat{c}_{11} \nabla_s^2 \psi_\ell^{m,n} + [(\omega^2 - \beta_0^2 \hat{c}_{44}) + \beta_0 \mu_\ell (\hat{c}_{13} + \hat{c}_{44}) + \beta_0 \eta_\ell \tau (\hat{e}_{x5} + \hat{e}_{z1})] \psi_\ell^{m,n}\}$$
$$- \frac{s}{r} [\hat{c}_{66} \nabla_s^2 \psi_4^{m,n} + (\omega^2 - \beta_0^2 \hat{c}_{44}) \psi_4^{m,n}] = F_1^{m,n}(r), \tag{25}$$

$$\frac{s}{r} \sum_{\ell=1}^{3} \{\hat{c}_{11} \nabla_s^2 \psi_\ell^{m,n} + [(\omega^2 - \beta_0^2 \hat{c}_{44}) + \beta_0 \mu_\ell (\hat{c}_{13} + \hat{c}_{44}) + \beta_0 \eta_\ell \tau (\hat{e}_{x5} + \hat{e}_{z1})] \psi_\ell^{m,n}\}$$
$$- \frac{d}{dr} [\hat{c}_{66} \nabla_s^2 \psi_4^{m,n} + (\omega^2 - \beta_0^2 \hat{c}_{44}) \psi_4^{m,n}] = F_2^{m,n}(r), \tag{26}$$

$$\sum_{\ell=1}^{3} [\mu_\ell \hat{c}_{44} + \eta_\ell \tau \hat{e}_{x5} - \beta_0 (\hat{c}_{13} + \hat{c}_{44})] \nabla_s^2 \psi_\ell^{m,n}$$
$$+ \sum_{\ell=1}^{3} [(\omega^2 - \beta_0^2 \hat{c}_{33}) \mu_\ell - \beta_0^2 \hat{e}_{z3} \tau \eta_\ell] \psi_\ell^{m,n} = F_3^{m,n}(r), \tag{27}$$

$$\sum_{\ell=1}^{3} [\hat{\epsilon}_{xx} \eta_\ell + \beta_0 (\hat{e}_{x5} + \hat{e}_{z1}) - \hat{e}_{x5} \mu_\ell] \nabla_s^2 \psi_\ell^{m,n}$$
$$+ \sum_{\ell=1}^{3} [-\beta_0^2 \hat{\epsilon}_{zz} \eta_\ell + \beta_0^2 \hat{e}_{z3} \mu_\ell] \psi_\ell^{m,n} = F_4^{m,n}(r), \tag{28}$$

with

$$\nabla_s^2 = \frac{d^2}{dr^2} + \frac{1}{r} \frac{d}{dr} - \frac{s^2}{r^2}. \tag{29}$$

The functions $F_j^{m,n}(r)$, $j = 1, \ldots, 4$ are written in the Appendix. They are written in terms of functions which have been determined in earlier stages of the iterative procedure. When $n = 0$, they also involve the

constant β_m, which must be found. The functions vanish when $m = 0$ or when $m > 0$ and $|n| > m$.

In a similar fashion, the perturbation procedure yields a system of boundary conditions.

Free surface:

$$\sum_{\ell=1}^{3}\left[\hat{c}_{11}\frac{d^2}{dr^2}+\hat{c}_{12}\left(\frac{d}{dr}-s^2\right)+\beta_0\hat{c}_{13}\mu_\ell+\beta_0\tau\hat{e}_{z1}\eta_\ell\right]\psi_\ell^{m,n}$$
$$-2\hat{c}_{66}s\left(\frac{d}{dr}-1\right)\psi_4^{m,n}=K_1^{m,n}\text{ at }r=1,$$

$$\hat{c}_{66}\left[\sum_{\ell=1}^{3}2s\left(\frac{d}{dr}-1\right)\psi_\ell^{m,n}-\left(\frac{d^2}{dr^2}-\frac{d}{dr}+s^2\right)\psi_4^{m,n}\right]=K_2^{m,n}\text{ at }r=1,$$

$$\sum_{\ell=1}^{3}\left[\hat{c}_{44}(\mu_\ell-\beta_0)+\eta_\ell\tau\hat{e}_{x5}\right]\frac{d}{dr}\psi_\ell^{m,n}+\beta_0\hat{c}_{44}s\psi_4^{m,n}=K_3^{m,n}\text{ at }r=1; \tag{30}$$

Clamped surface:

$$\sum_{\ell=1}^{3}\frac{d}{dr}\psi_\ell^{m,n}-s\psi_4^{m,n}=0\text{ at }r=1,$$

$$s\sum_{\ell=1}^{3}\psi_\ell^{m,n}-\frac{d\psi_4^{m,n}}{dr}=0\text{ at }r=1,$$

$$\sum_{\ell=1}^{3}\mu_\ell\psi_\ell^{m,n}=0\text{ at }r=1. \tag{31}$$

Short-circuit:

$$\sum_{\ell=1}^{3}\eta_\ell\psi_\ell^{m,n}=0\text{ at }r=1, \tag{32}$$

Open-circuit:

$$\sum_{\ell=1}^{3}\left[-\hat{\epsilon}_{xx}\eta_\ell+\hat{e}_{x5}(\mu_\ell-\beta_0)\right]\frac{d\psi_\ell^{m,n}}{dr}+\beta_0\hat{e}_{x5}s\psi_4^{m,n}=K_4^{m,n}\text{ at }r=1. \tag{33}$$

The constants $K_j^{m,n}$, $j = 1, \ldots, 4$ are written in the Appendix. Like the $F_j^{m,n}(r)$, they vanish when $m = 0$ and are known when $m > 0$; when $n = 0$, they also involve β_m.

IV. SOLUTION OF THE DIFFERENTIAL EQUATIONS

We now show how to decouple the differential equations (25) to (28) and solve them. First, let

$$H_1^{m,n}(r) = \sum_{\ell=1}^{3} \{\hat{c}_{11}\nabla_s^2\psi_\ell^{m,n} + [(\omega^2 - \beta_0^2\hat{c}_{44}) + \beta_0\mu_\ell(\hat{c}_{13} + \hat{c}_{44}) + \beta_0\eta_\ell\tau(\hat{e}_{x5} + \hat{e}_{z1})]\psi_\ell^{m,n}\}, \quad (34)$$

$$H_2^{m,n}(r) = \hat{c}_{66}\nabla_s^2\psi_4^{m,n} + (\omega^2 - \beta_0^2\hat{c}_{44})\psi_4^{m,n}. \quad (35)$$

Then by using (25), (26), and procedures similar to those exhibited in Ref. 8, we can show that, except in a certain special case to be discussed later,

$$H_1^{m,n}(r) = \tfrac{1}{2}r^s \int_0^r x^{-s}[F_1^{m,n}(x) + F_2^{m,n}(x)]\,dx + \tfrac{1}{2}r^{-s}\int_0^r x^s[F_1^{m,n}(x) - F_2^{m,n}(x)]\,dx, \quad (36)$$

$$H_2^{m,n}(r) = \tfrac{1}{2}r^s \int_0^r x^{-s}[F_1^{m,n}(x) + F_2^{m,n}(x)]\,dx - \tfrac{1}{2}r^{-s}\int_0^r x^s[F_1^{m,n}(x) - F_2^{m,n}(x)]\,dx, \quad (37)$$

Now consider eqs. (27), (28), and (34). They are equivalent to the three decoupled equations

$$\nabla_s^2\psi_\ell^{m,n} + p_\ell^2\psi_\ell^{m,n} = f_\ell^{m,n}, \ \ell = 1, 2, 3, \quad (38)$$

provided that

$$p_\ell^2 = [(\omega^2 - \beta_0^2\hat{c}_{44}) + \beta_0\mu_\ell(\hat{c}_{13} + \hat{c}_{44}) + \beta_0\eta_\ell\tau(\hat{e}_{x5} + \hat{e}_{z1})]/\hat{c}_{11}, \quad (39)$$

from (34),

$$-p_\ell^2[\mu_\ell\hat{c}_{44} + \eta_\ell\tau\hat{e}_{x5} - \beta_0(\hat{c}_{13} + \hat{c}_{44})] + [(\omega^2 - \beta_0^2\hat{c}_{33})\mu_\ell - \beta_0^2\hat{e}_{z3}\tau\eta_\ell] = 0, \quad (40)$$

from (27), and, from (28),

$$-p_\ell^2[\hat{\epsilon}_{xx}\eta_\ell + \beta_0(\hat{e}_{x5} + \hat{e}_{z1}) - \hat{e}_{x5}\mu_\ell] + [-\beta_0^2\hat{\epsilon}_{zz}\eta_\ell + \beta_0^2\hat{e}_{z3}\mu_\ell] = 0. \quad (41)$$

These imply that the p_ℓ^2 satisfy the cubic equation

$$(\hat{\epsilon}_{xx}p_\ell^2 + \beta_0^2\hat{\epsilon}_{zz})[(\hat{c}_{11}p_\ell^2 + \beta_0^2\hat{c}_{44} - \omega^2)(\hat{c}_{44}p_\ell^2 + \beta_0^2\hat{c}_{33} - \omega^2) - p_\ell^2\beta_0^2(\hat{c}_{13} + \hat{c}_{44})^2] + \tau(\hat{e}_{x5}p_\ell^2 + \beta_0^2\hat{e}_{z3})[(\hat{c}_{11}p_\ell^2 + \beta_0^2\hat{c}_{44} - \omega^2) \times (\hat{e}_{x5}p_\ell^2 + \beta_0^2\hat{e}_{z3}) - 2p_\ell^2\beta_0^2(\hat{c}_{13} + \hat{c}_{44})(\hat{e}_{x5} + \hat{e}_{z1})] + \tau p_\ell^2\beta_0^2(\hat{e}_{x5} + \hat{e}_{z1})^2(\hat{c}_{44}p_\ell^2 + \beta_0^2\hat{c}_{33} - \omega^2) = 0, \quad (42)$$

and that

$$\mu_\ell = \beta_0 \{(\hat{c}_{11}p_\ell^2 + \beta_0^2\hat{c}_{44} - \omega^2)(\hat{e}_{z3}\hat{\epsilon}_{xx} - \hat{e}_{x5}\hat{\epsilon}_{zz}) - p_\ell^2(\hat{e}_{x5} + \hat{e}_{z1})[\tau\hat{e}_{x5}(\hat{e}_{x5} + \hat{e}_{z1}) + \hat{\epsilon}_{xx}(\hat{c}_{13} + \hat{c}_{44})]\}$$
$$\times \{\beta_0^2(\hat{c}_{13} + \hat{c}_{44})(\hat{e}_{z3}\hat{\epsilon}_{xx} - \hat{e}_{x5}\hat{\epsilon}_{zz}) - \tau\hat{e}_{x5}(\hat{e}_{x5} + \hat{e}_{z1})(\hat{e}_{x5}p_\ell^2 + \beta_0^2\hat{e}_{z3}) - \hat{\epsilon}_{xx}(\hat{e}_{x5} + \hat{e}_{z1})(\hat{c}_{44}p_\ell^2 + \beta_0^2 c_{33} - \omega^2)\}^{-1} \quad (43)$$

and

$$\eta_\ell = [\hat{c}_{11}p_\ell^2 + \beta_0^2\hat{c}_{44} - \omega^2 - \mu_\ell\beta_0(\hat{c}_{13} + \hat{c}_{44})]/[\beta_0\tau(\hat{e}_{x5} + \hat{e}_{z1})]. \quad (44)$$

Furthermore, by eqs. (34), (27), and (28), the $f_\ell^{m,n}(r)$, $\ell = 1, 2, 3$, must satisfy

$$\hat{c}_{11}\sum_{\ell=1}^{3} f_\ell^{m,n} = H_1^{m,n}, \quad (45)$$

$$\sum_{\ell=1}^{3} [\mu_\ell\hat{c}_{44} + \eta_\ell\tau\hat{e}_{x5} - \beta_0(\hat{c}_{13} + \hat{c}_{44})]f_\ell^{m,n} = F_3^{m,n}, \quad (46)$$

$$\sum_{\ell=1}^{3} [\hat{\epsilon}_{xx}\eta_\ell + \beta_0(\hat{e}_{x5} + \hat{e}_{z1}) - \hat{e}_{x5}\mu_\ell]f_\ell^{m,n} = F_4^{m,n}. \quad (47)$$

These equations can be solved for the $f_\ell^{m,n}(r)$. By (35), eq. (38) also holds when $\ell = 4$, provided that

$$p_4^2 = (\omega^2 - \beta_0^2\hat{c}_{44})/\hat{c}_{66}, \quad (48)$$

$$f_4^{m,n} = H_2^{m,n}/\hat{c}_{66}. \quad (49)$$

The next step is to solve the uncoupled differential equations (38). The functions $f_j^{m,n}(r)$, $j = 1, \ldots, 4$, are either determined completely ($n \neq 0$) or else involve β_m in a known way ($n = 0$). Using the fact that $\psi_j^{m,n}$ is bounded at $r = 0$ to evaluate an integration constant, we have as a solution to (38)

$$\psi_j^{m,n}(r) = \left[A_j^{m,n} + \frac{\pi}{2}\int_r^1 xY_s(p_jx)f_j^{m,n}(x)\,dx\right]J_s(p_jr) + \frac{\pi}{2}\int_0^r xJ_s(p_jx)f_j^{m,n}(x)\,dx\,Y_s(p_jr) \text{ if } p_j^2 > 0,$$

$$\psi_j^{m,n}(r) = \left[A_j^{m,n} - \int_r^1 xK_s(q_jx)f_j^{m,n}(x)\,dx\right]I_s(q_jr) - \int_0^r xI_s(q_jx)f_j^{m,n}(x)\,dx\,K_s(q_jr) \text{ if } p_j^2 \equiv -q_j^2 < 0. \quad (50)$$

For any values of m and n, there are four constants $A_j^{m,n}$, $j = 1, \ldots, 4$, which remain to be evaluated. When $n = 0$, β_m must also be found. In the next section, we will show how to apply the boundary conditions to evaluate these constants.

It appears from (36), (37), (50), and (61) that a double integration must be performed computationally to obtain $H_1^{m,n}(r)$ and $H_2^{m,n}(r)$. Use of (38) and integration by parts, however, can reduce this to a single integration.

There is one special case for which the above analysis is not quite correct. By using arguments similar to those in Ref. 8, we can show that if $p_\ell^2 = 0$ for some ℓ, then

$$H_1^{m,n}(r) = C^{m,n} r^s + \tfrac{1}{2} r^s \int_1^r x^{-s}[F_1^{m,n}(x) + F_2^{m,n}(x)]\, dx + \tfrac{1}{2} r^{-s} \int_0^r x^s [F_1^{m,n}(x) - F_2^{m,n}(x)]\, dx,$$

$$H_2^{m,n}(r) = C^{m,n} r^s + \tfrac{1}{2} r^s \int_1^r x^{-s}[F_1^{m,n}(x) + F_2^{m,n}(x)]\, dx - \tfrac{1}{2} r^{-s} \int_0^r x^s [F_1^{m,n}(x) - F_2^{m,n}(x)]\, dx, \quad (51)$$

if $n \neq 0$. Here $C^{m,n}$ is a constant which remains to be determined. (When every p_ℓ^2 is nonzero, $C^{m,n}$ is arbitrary in the sense that changing it merely changes the constant by which the entire solution is multiplied.) Also, in this special case we have when $n = 0$,

$$H_1^{m,0}(r) = C_1^{m,n} - \int_r^1 F_1^{m,0}(x)\, dx,$$

$$H_2^{m,0}(r) = C_2^{m,n} - \int_r^1 F_2^{m,0}(x)\, dx, \quad (52)$$

where $C_1^{m,n}$ and $C_2^{m,n}$ must be determined. Now it can be shown that if the fiber is vibrating in the lowest-order torsional mode (with $v^{0,0}$ proportional to r and $u^{0,0} = w^{0,0} = \Phi^{0,0} = N = 0$), then $p_1^2 = p_4^2 = 0$. For this case, the solutions of (38) for $j = 1$ and 4 are

$$\psi_j^{m,n}(r) = \left[A_j^{m,n} - \frac{1}{2s} \int_r^1 x^{-s+1} f_j^{m,n}(x)\, dx \right] r^s - \frac{1}{2s} \int_0^r x^{s+1} f_j^{m,n}(x)\, dx r^{-s} \text{ if } n \neq 0, \quad (53)$$

$$\psi_j^{m,0}(r) = \int_r^1 x \ln x f_j^{m,0}(x)\, dx + \int_0^r x f_j^{m,0}(x)\, dx \ln r \text{ if } n = 0. \quad (54)$$

In eq. (54), an integration constant has been set to zero because it does not affect the final solution. It follows from eqs. (23), (43), (44), and (48), that when $p_1^2 = p_4^2 = 0$ and $n \neq 0$, the constants $A_1^{m,n}$ and $A_4^{m,n}$ appear in the displacements and electric potential only in the combination $A_1^{m,n} - A_4^{m,n}$. Thus for $n \neq 0$, the constants to be evaluated are $A_1^{m,n} - A_4^{m,n}$,

$A_2^{m,n}$, $A_3^{m,n}$, and $C^{m,n}$. When $n = 0$, we must find $A_2^{m,0}$, $A_3^{m,0}$, $C_1^{m,0}$, $C_2^{m,0}$, and β_m. In the next section, we show how to use the boundary conditions to determine these constants.

V. EVALUATION OF THE BOUNDARY CONDITIONS

For any pair (m, n), there are four boundary conditions, three of which are either eqs. (30) or (31), and the fourth of which is either (32) or (33); there are also four unknown constants to be found. When $n = 0$, β_m must be determined, too.

When the solutions $\beta_j^{m,n}$ to the differential equations (38) are substituted into the appropriate boundary conditions from eqs. (30) to (33), a system of equations results which can be written in matrix form as

$$\mathbf{J}^n \mathbf{A}^{m,n} = \mathbf{V}^{m,n}. \tag{55}$$

We will not write down here the specific components of these matrices and vectors, although it is straightforward to do so. The important things to know are the following: The 4×4 matrix $\mathbf{J}^n$ involves Bessel functions. It depends upon β_0, but is known once this is determined. The vector $\mathbf{A}^{m,n}$ consists of the four unknown constants to be determined. The vector $\mathbf{V}^{m,n}$ contains known constants: $K_j^{m,n}$, Bessel functions, integrals involving $f_j^{m,n}$. When $n = 0$, it also contains β_m linearly.

Incidentally, from a computational viewpoint, it is never necessary to differentiate the functions $\psi_j^{m,n}$ numerically, either for substitution into the boundary conditions or into the functions listed in the Appendix. Equation (38) can be used to eliminate all second derivatives of $\psi_j^{m,n}$ with respect to r. Differentiation with respect to r of the solutions (50) and the use of standard relations between Bessel functions and their derivatives result in analytical expressions for $d\psi_j^{m,n}/dr$.

The procedure for solving the differential equations and applying the boundary conditions is an iterative one. We start with $m = 0$. We choose a modal solution when $n = 0$ and set all $\psi_j^{0,n}$ to zero when $n \neq 0$. The $\psi_j^{0,0}$ satisfy eq. (38) with $f_j^{0,0} = 0$. The boundary conditions for this case are

$$\mathbf{J}^0 \mathbf{A}^{0,0} = 0. \tag{56}$$

From this we obtain for a nontrivial solution the frequency equation

$$\det \mathbf{J}^0 = 0, \tag{57}$$

which determines β_0 as a function of ω. This, of course, is the same as the dispersion relation for the hexagonal 6mm case about which we are perturbing.

As we iterate on m, we can see from the equations in the Appendix that $F_j^{m,n} = K_j^{m,n} = 0$ whenever $|n| > m$. It follows that $\mathbf{V}^{m,n} = 0$ in this case

and, since $\det \mathbf{J}^n \neq 0$, that $\mathbf{A}^{m,n} = 0$. Thus all $\psi_j^{m,n} = 0$ whenever $|n| > m$.

If $0 < |n| \leq m$, $\mathbf{V}^{m,n}$ is in general nonzero. Since $\det \mathbf{J}^n \neq 0$, we can immediately obtain $\mathbf{A}^{m,n}$ from eq. (55).

If $n = 0$, the analysis is slightly more complicated. It was explained in detail in Ref. 8, so we merely give the results here. To obtain β_m, replace any column of $\mathbf{J}^0$ by $\mathbf{V}^{m,n}$ and set the determinant of the resulting matrix to zero. The unknown vector $\mathbf{A}^{m,0}$ can be written as

$$\mathbf{A}^{m,0} = C_m \mathbf{A}^{0,0} + \mathbf{D}^{m,0}, \tag{58}$$

where $\mathbf{D}^{m,0}$ has *three* unknown components and $D_j^{m,0} = 0$ for some j for which $A_j^{0,0} \neq 0$. Then the equation

$$\mathbf{J}^0 \mathbf{D}^{m,0} = 0 \tag{59}$$

can be solved for $\mathbf{D}^{m,0}$. Furthermore, C_m is arbitrary in the sense that varying it varies the constant by which the full solution is multiplied. We set $C_m = 0$.

In this manner, the functions $\psi_j^{m,n}$ can be determined iteratively, starting with $m = 0$. For any given value of m, nontrivial results are obtained only when $|n| \leq m$. The perturbation contributions to the elastic displacements and electric potential are then found from eq. (23). The full solution is given by eqs. (21) and (22).

APPENDIX

Let

$$\begin{aligned} s &= 3n + N, \\ s_+ &= 3(n+1) + N, \\ s_- &= 3(n-1) + N, \end{aligned} \tag{60}$$

where n and N are integers. Then

$$\begin{aligned} F_1^{m,n}(r) + F_2^{m,n}(r) = \left(\frac{d}{dr} - \frac{s}{r}\right) \sum_{j=0}^{m-1} \Bigg\{ & \gamma_{m-j} \hat{c}_{44} \left[\sum_{\ell=1}^{3} \psi_\ell^{j\,n} + \psi_4^{j\,n} \right] \\ & - \beta_{m-j} \sum_{\ell=1}^{3} \left[(\hat{c}_{13} + \hat{c}_{44}) \mu_\ell + \tau (\hat{e}_{x5} + \hat{e}_{z1}) \eta_\ell \right] \psi_\ell^j \\ - \left[\nabla_{s_+}^2 - \frac{2(1-s_+)}{r} \left(\frac{d}{dr} + \frac{s_+}{r} \right) \right] \Bigg\{ & \sum_{j=0}^{m-1} 2\beta_{m-j-1} \left[\sum_{\ell=1}^{3} \psi_\ell^{j\,n+1} - \psi_4^{j\,n+1} \right] \\ & + \sum_{\ell=1}^{3} (\tau \xi \eta_\ell - \mu_\ell) \psi_\ell^{m-1,n+1} \Bigg\}, \end{aligned}$$

$$F_1^{m,n}(r) - F_2^{m,n}(r) = \left(\frac{d}{dr} + \frac{s}{r}\right) \sum_{j=0}^{m-1} \left\{ \gamma_{m-j} \hat{c}_{44} \left[\sum_{\ell=1}^{3} \psi_\ell^{j\,n} - \psi_4^{j\,n} \right] \right.$$

$$- \beta_{m-j} \sum_{\ell=1}^{3} \left[(\hat{c}_{13} + \hat{c}_{44}) \mu_\ell + \tau (\hat{e}_{x5} + \hat{e}_{z1}) \eta_\ell \right] \psi_\ell^j$$

$$+ \left[\nabla_{s-}^2 - \frac{2(1+s_-)}{r} \left(\frac{d}{dr} - \frac{s_-}{r} \right) \right] \left\{ \sum_{j=0}^{m-1} 2\beta_{m-j-1} \left[\sum_{\ell=1}^{3} \psi_\ell^{j\,n-1} + \psi_4^{j\,n-1} \right] \right.$$

$$\left. \left. + \sum_{\ell=1}^{3} (\tau \xi \eta_\ell - \mu_\ell) \psi_\ell^{m-1,n-1} \right\} \right],$$

$$F_3^{m,n}(r) = \frac{1}{2} \left\{ \left(\frac{d}{dr} - \frac{s_-}{r} \right) \nabla_{s-}^2 - \frac{2(2+s_-)}{r} \right.$$

$$\times \left[\nabla_{s-}^2 - \frac{2(1+s_-)}{r} \left(\frac{d}{dr} - \frac{s_-}{r} \right) \right] \left\} \left[\sum_{\ell=1}^{3} \psi_\ell^{m-1,n-1} + \psi_4^{m-1,n-1} \right] \right.$$

$$- \frac{1}{2} \left\{ \left(\frac{d}{dr} + \frac{s_+}{r} \right) \nabla_{s+}^2 - \frac{2(2-s_+)}{r} \left[\nabla_{s+}^2 - \frac{2(1-s_+)}{r} \left(\frac{d}{dr} + \frac{s_+}{r} \right) \right] \right\}$$

$$\times \left[\sum_{\ell=1}^{3} \psi_\ell^{m-1,n+1} - \psi_4^{m-1,n+1} \right] + \sum_{j=0}^{m-1} \left[\beta_{m-j} (\hat{c}_{13} + \hat{c}_{44}) \nabla_s^2 \sum_{\ell=1}^{3} \psi_\ell^{j\,n} \right.$$

$$\left. + \gamma_{m-j} \sum_{\ell=1}^{3} (\hat{c}_{33} \mu_\ell + \tau \hat{e}_{z3} \eta_\ell) \psi_\ell^{j\,n} \right],$$

$$F_4^{m,n}(r) = -\tfrac{1}{2}\xi \left\{ \left(\frac{d}{dr} - \frac{s_-}{r} \right) \nabla_{s-}^2 \frac{2(2+s_-)}{r} \right.$$

$$\times \left[\nabla_{s-}^2 - \frac{2(1+s_-)}{r} \left(\frac{d}{dr} - \frac{s_-}{r} \right) \right] \left\} \left[\sum_{\ell=1}^{3} \psi_\ell^{m-1,n-1} + \psi_4^{m-1,n-1} \right] \right.$$

$$+ \tfrac{1}{2}\xi \left\{ \left(\frac{d}{dr} + \frac{s_+}{r} \right) \nabla_{s+}^2 - \frac{2(2-s_+)}{r} \left[\nabla_{s+}^2 - \frac{2(1-s_+)}{r} \left(\frac{d}{dr} + \frac{s_+}{r} \right) \right] \right\}$$

$$\times \left[\sum_{\ell=1}^{3} \psi_\ell^{m-1,n+1} - \psi_4^{m-1,n+1} \right] + \sum_{j=0}^{m-1} \left[\beta_{m-j} (\hat{e}_{x5} + \hat{e}_{z1}) \nabla_s^2 \sum_{\ell=1}^{3} \psi_\ell^{j\,n} \right.$$

$$\left. + \gamma_{m-j} \sum_{\ell=1}^{3} (\hat{e}_{z3} \mu_\ell - \hat{\epsilon}_{zz} \eta_\ell) \psi_\ell^{j\,n} \right] \quad (61)$$

where

$$\gamma_m = \sum_{j=0}^{m} \beta_j \beta_{m-j}. \quad (62)$$

$$K_1^{m,n} = \frac{1}{2} \sum_{j=0}^{m-1} \beta_{m-1-j} \left\{ \left(\frac{d}{dr} - s_- \right) \left[\sum_{\ell=1}^{3} \psi_\ell^{j\,n-1} + \psi_4^{j\,n-1} \right] \right.$$

$$\left. - \left(\frac{d}{dr} + s_+ \right) \left[\sum_{\ell=1}^{3} \psi_\ell^{j\,n+1} - \psi_4^{j\,n+1} \right] \right\}$$

$$-\frac{1}{2}\sum_{\ell=1}^{3}(\mu_\ell-\tau\xi\eta_\ell)\left[\left(\frac{d}{dr}-s_-\right)\psi_\ell^{m-1,n-1}-\left(\frac{d}{dr}+s_+\right)\psi_\ell^{m-1,n+1}\right]$$

$$-\sum_{j=0}^{m-1}\beta_{m-j}\sum_{\ell=1}^{3}(\hat{c}_{13}\mu_\ell+\hat{e}_{z1}\tau\eta_\ell)\psi_\ell^{j,n}\ \text{at}\ r=1,$$

$$K_2^{m,n}=\frac{1}{2}\sum_{j=0}^{m-1}\beta_{m-1-j}\left\{\left(\frac{d}{dr}-s_-\right)\left[\sum_{\ell=1}^{3}\psi_\ell^{j,n-1}+\psi_4^{j,n-1}\right]\right.$$

$$\left.+\left(\frac{d}{dr}+s_+\right)\left[\sum_{\ell=1}^{3}\psi_\ell^{j,n+1}-\psi_4^{j,n+1}\right]\right\}$$

$$-\frac{1}{2}\sum_{\ell=1}^{3}(\mu_\ell-\tau\xi\eta_\ell)\left[\left(\frac{d}{dr}-s_-\right)\psi_\ell^{m-1,n-1}\right.$$

$$\left.+\left(\frac{d}{dr}+s_+\right)\psi_\ell^{m-1,n+1}\right]\ \text{at}\ r=1,$$

$$K_3^{m,n}=\hat{c}_{44}\sum_{j=0}^{m-1}\beta_{m-j}\left\{\sum_{\ell=1}^{3}\left[s\psi_\ell^{j,n}+\left(\frac{d}{dr}-s\right)\psi_\ell^{j,n}\right]-s\psi_4^{j,n}\right\}$$

$$+\frac{1}{2}\left[\nabla_{s_-}^2-2(1+s_-)\left(\frac{d}{dr}-s_-\right)\right]\left[\sum_{\ell=1}^{3}\psi_\ell^{m-1,n-1}+\psi_4^{m-1,n-1}\right]$$

$$-\frac{1}{2}\left[\nabla_{s_+}^2-2(1-s_+)\left(\frac{d}{dr}+s_+\right)\right]\left[\sum_{\ell=1}^{3}\psi_\ell^{m-1,n+1}-\psi_4^{m-1,n+1}\right]\ \text{at}\ r=1,$$

$$K_4^{m,n}=\hat{e}_{x5}\sum_{j=0}^{m-1}\beta_{m-j}\left\{\sum_{\ell=1}^{3}\left[s\psi_\ell^{j,n}+\left(\frac{d}{dr}-s\right)\psi_\ell^{j,n}\right]-s\psi_4^{j,n}\right\}$$

$$+\tfrac{1}{2}\xi\left[\nabla_{s_-}^2-2(1+s_-)\left(\frac{d}{dr}-s_-\right)\right]\left[\sum_{\ell=1}^{3}\psi_\ell^{m-1,n-1}+\psi_4^{m-1,n-1}\right]$$

$$-\tfrac{1}{2}\xi\left[\nabla_{s_+}^2-2(1-s_+)\left(\frac{d}{dr}+s_+\right)\right]\left[\sum_{\ell=1}^{3}\psi_\ell^{m-1,n+1}-\psi_4^{m-1,n+1}\right]$$

$$\text{at}\ r=1. \quad (63)$$

REFERENCES

1. L. Pochhammer, "Ueber die Fortpflanzungsgeschwindigkeiten kleiner Schwingungen in einem unbegrentzen isotropen Kreiscylinder," *J. reine angew. Math.*, *81* (1876), pp. 324–336.
2. C. Chree, "The equations of an isotropic elastic solid in polar and cylindrical coordinates, their solutions and applications," Trans. Cambridge Phil. Soc., *14* (1889), pp. 250–369.
3. I. Mirsky, "Wave propagation in transversely isotropic circular cylinders Part I: theory," J. Acoust. Soc. Am., *37* (1965), pp. 1016–1021.
4. I. Mirsky, "Wave propagation in transversely isotropic circular cylinders Part II: numerical results," J. Acoust. Soc. Am., *37* (1965), pp. 1022–1026.
5. R. W. Morse, "Compressional waves along an anisotropic circular cylinder having hexagonal symmetry," J. Acoust. Soc. Am., *26* (1954), pp. 1018–1021.
6. N. G. Einspruch and R. Truell, "Propagation of traveling waves in a circular cylinder having hexagonal elastic symmetry," J. Acoust. Soc. Am., *31* (1959), pp. 691–693.

7. L. O. Wilson and J. A. Morrison, "Wave propagation in piezoelectric rods of hexagonal crystal symmetry," Quart. J. Mech. Appl. Math. (1977).
8. L. O. Wilson, "Wave propagation along a sapphire rod," J. Acoust. Soc. Am., *61* (1977), pp. 995–1003.
9. L. O. Wilson and M. A. Gatto, "Torsional vibrations of a sapphire rod: a numerical description," J. Acoust. Soc. Am., *61* (1977), pp. 1004–1013.
10. R. N. Thurston and L. O. Wilson, "Torsional waves in a sapphire rod at low frequencies," IEEE Trans. Sonics and Ultrasonics, (1977), to be published.
11. B. A. Auld, *Acoustic Fields and Waves in Solids,* Wiley, New York, 1973, Vol. I, p. 299.
12. Reference 11, Appendix 2.
13. B. A. Auld, *Acoustic Fields and Waves in Solids,* Wiley, New York, 1973, Vol. II, p. 178.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 8, October 1977
Printed in U.S.A.

A Descent Algorithm for the Multihour Sizing of Traffic Networks

By W. B. ELSNER

(Manuscript received March 3, 1977)

Multihour engineering is a technique for designing trunk networks when the hours of peak traffic loads between various pairs of offices do not coincide. A new descent-type computational algorithm for the multihour engineering problem is derived. This algorithm obtains the unique solution to the minimization of the multihour cost function, which is strictly convex but only piecewise differentiable. The noninteger minimum-cost solution is subsequently rounded to the nearest allowable integer solution to give a realizable network. The new algorithm is applied to three numerical examples from the California network. The results are compared with the nonoptimal, nonunique solutions obtained with an earlier algorithm, and with the traditional single-hour solutions.

I. INTRODUCTION

This paper describes a numerical algorithm which obtains the unique, optimal noninteger solution to the multihour traffic network engineering problem. This solution is subsequently rounded to the nearest allowable integer solution to yield a unique, near-optimal realizable network.

As described in Ref. 1, multihour engineering is a procedure whereby a least-cost traffic network is engineered for more than one set of point-to-point loads, subject to the constraint that blocking on any last-choice trunk group not exceed a specified value. For networks which exhibit noncoincident traffic patterns,† the multihour engineering method has been shown to achieve significant capital-cost savings over the conventional single-hour engineering procedures.[1]

The results reported in Ref. 1 were based on an algorithm which optimizes the high-usage trunk group sizes‡ one at a time, in a fixed but

† Traffic loads between different pairs of offices are said to be noncoincident if their highest average values occur in different hours, or at the same hour but in different seasons.

‡ High-usage groups are direct groups which carry the majority of the load between those pairs of offices which have a large enough community of interest to warrant direct trunking.

arbitrary sequence, until no further cost reductions can be obtained. This algorithm is called a "coordinate-search" algorithm here. Such an algorithm has the undesirable property that it does not generally converge to a unique solution of the multihour problem. It can converge to any one of a family of suboptimal solutions, depending on the initialization of the algorithm, and on the specific order in which the calculations are performed.

The practical reasons for obtaining the optimal—and hence (as shown in Section III below) unique—solution to the multihour problem are as follows: First, the periodic re-engineering of the network in response to new load forecasts is facilitated. (In the Bell System, most networks must be re-engineered at least once each year.) A unique solution guarantees that changes in trunk-group sizes from one forecast period to the next reflect only changes in the loads. In contrast, the coordinate-search algorithm can produce changes in trunk-group sizes which are as much a function of nonuniqueness as they are of actual alterations in the loads. It is not possible to distinguish between these two effects, and thus use of the coordinate-search algorithm could lead to excessive rearrangements. Second, capital-cost savings with respect to the coordinate-search solutions can be realized in most cases.

The essential difficulty of the multihour engineering problem arises from the fact that the network cost function is not differentiable everywhere in its domain. The algorithm presented here, however, is assured of convergence to the minimum-cost noninteger solution by the convexity of the cost function and by the particular mechanism for executing the search process.

II. MULTIHOUR ENGINEERING—THE MODEL AND ITS COST FUNCTION

The model of the network considered in this paper is shown in Fig. 1. Traffic which is destined from the single originating office to one of n terminating offices is first offered to the appropriate one-way high-usage trunk group. If all the trunks in that group are busy, the traffic overflows and is offered to a final group which routes this parcel to a tandem switch, from where it is sent to its destination via a tandem-completing group. The final and tandem-completing groups are sized so that the blocking probability on each is 0.01 during its respective busy hour. The object of the engineering process is to determine the high-usage trunk group sizes which minimize the cost of the network subject to the blocking constraints on the alternate routes.

The cost of the network can be divided into four components, which are defined below:

(i) The direct-route cost: It is assumed that the cost of each high-usage trunk group is directly proportional to the number of trunks in the

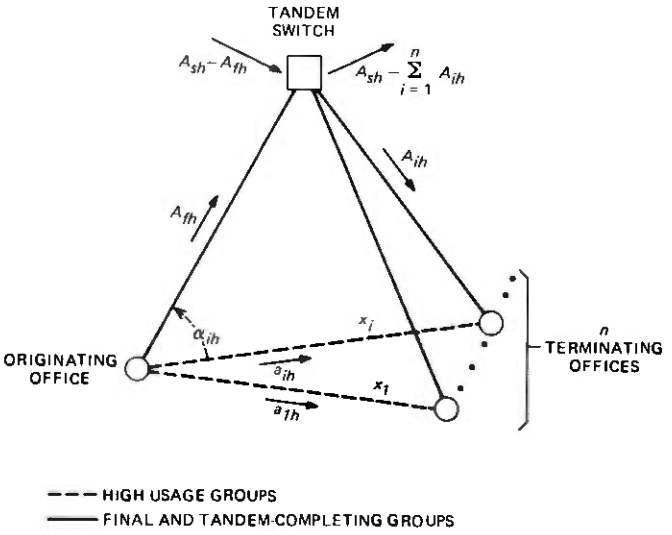


Fig. 1—Network model.

group.† If the cost per trunk of the ith high-usage group is c_{di}, and there are x_i trunks in this group, then the total cost for high-usage trunks in this network is given by

$$C_d = \sum_{i=1}^{n} c_{di} x_i. \tag{1}$$

In the theory which follows, as well as in the algorithm based on this theory, x_i is treated as a nonnegative real variable.

(*ii*) The final-route cost: Let $\mathcal{H} = \{1, 2, \ldots, H\}$ be the set of hours for which the network is to be engineered, and let $h \in \mathcal{H}$. Let a_{ih} be the load, in erlangs, offered to the ith high-usage group in hour h. Then the overflow from this group in hour h is given by

$$\alpha_{ih} = a_{ih} B(x_i, a_{ih}) \tag{2}$$

where $B(\cdot, a_{ih})$ is a strictly convex and continuously differentiable function of x_i which agrees with the Erlang loss function on the integers.‡ The overflow parcels from all the high-usage groups are combined and offered to the final trunk group. It is assumed that, in addition to the overflow traffic, the final group also has offered to it a first-routed load in hour h, designated by A_{fh}.

† Such an assumption is necessary, since the eventual realization of the network in terms of facilities is not known at the time that the groups are sized. Thus, average costs per trunk are used in approximating the eventual cost of each group.

‡ Such an interpolating function can always be constructed since $B(n-1, a) - B(n, a) > B(n, a) - B(n+1, a)$, $n = 1, 2, \ldots$.[2]

In sizing high-usage trunk groups, it is customary to approximate the number of trunks required in the final group by dividing the total load offered to this group by its so-called "marginal capacity." [1,3†] If the cost per trunk of the final group is c_f and its marginal capacity γ_f, then the cost of sizing the final to carry only its hour-h load is approximated by

$$C_{fh} = \frac{c_f}{\gamma_f}\left(A_{fh} + \sum_{i=1}^{n} \alpha_{ih}\right). \tag{3}$$

Since the final group must be engineered for its busiest hour, its approximate cost is

$$C_f = \max_{h \in \mathcal{H}} C_{fh}. \tag{4}$$

The actual sizing of the final group (which takes place only after all the high-usage groups have been sized) is done more precisely, of course.

(*iii*) The switching cost: It is assumed that the cost of switching is proportional to the load, with a unit-cost per erlang of c_s. Let A_{sh} denote the load switched by the tandem in hour h, excluding the overflows from the n high-usage groups. (The first-routed load on the final in hour h, A_{fh}, is included in A_{sh}.) Ignoring the blocking on the final group, the cost of switching only the hour-h load is

$$C_{sh} = c_s\left(A_{sh} + \sum_{i=1}^{n} \alpha_{ih}\right). \tag{5}$$

The cost of the tandem switch, when engineered for its busy hour, is then

$$C_s = \max_{h \in \mathcal{H}} C_{sh}. \tag{6}$$

(*iv*) The tandem-completing costs: The total load offered to the ith tandem-completing group in hour h is $A_{ih} + \alpha_{ih}$, where α_{ih} is the overflow from the ith high-usage group (neglecting the blocking on the final group and at the tandem) and A_{ih} is the remaining load destined to the ith terminating office via the tandem. As in the case of the final group, the size of the tandem-completing group is not computed exactly, but rather approximated by dividing its offered load by its marginal capacity. Let γ_{ti} be the marginal capacity of the ith tandem-completing group, and c_{ti} its cost per trunk. Then the cost of sizing this group to carry only its hour-h load is

† While the marginal-capacity assumption is not appropriate for determining actual trunk requirements on the final group, it is sufficiently accurate for the comparative purpose to which it is put here.

$$C_{tih} = \frac{c_{ti}}{\gamma_{ti}} (A_{ih} + \alpha_{ih}), \tag{7}$$

and sizing this group for its busy hour results in a cost

$$C_{ti} = \max_{h \in \mathcal{H}} C_{tih}. \tag{8}$$

The cost of providing trunks for all tandem-completing groups is thus

$$C_t = \sum_{i=1}^{n} C_{ti}. \tag{9}$$

The total cost of the network is simply the sum of these four components:

$$\begin{aligned} C(x) &= C_d + C_f + C_s + C_t \\ &= \sum_{i=1}^{n} c_{di} x_i + \frac{c_f}{\gamma_f} \max_{h \in \mathcal{H}} \left(A_{fh} + \sum_{i=1}^{n} \alpha_{ih} \right) \\ &\quad + c_s \max_{h \in \mathcal{H}} \left(A_{sh} + \sum_{i=1}^{n} \alpha_{ih} \right) \\ &\quad + \sum_{i=1}^{n} \frac{c_{ti}}{\gamma_{ti}} \max_{h \in \mathcal{H}} (A_{ih} + \alpha_{ih}), \end{aligned} \tag{10}$$

where $x = \{x_1, \ldots, x_n\}$ is the n-vector of high-usage group sizes. This function is called the "multihour cost function." Note that the final, switch, and tandem-completing costs may attain their maxima for different values of h, since each of these alternate-route components may be busy in a different hour.

The object of multihour engineering, then, is to minimize the cost function defined by eq. (10) with respect to the high-usage trunk group sizes, i.e., to determine $x = x^*$ such that

$$C(x^*) = \min_{x \in \mathcal{X}} C(x) \tag{11}$$

where the set $\mathcal{X}$ is defined by

$$\mathcal{X} = \{x: x_i \geq 0, \quad i = 1, \ldots, n\}. \tag{12}$$

From the point of view of the (continuous) multihour cost function, any x is "feasible" if $x \in \mathcal{X}$. Of course, an actual network is realizable only in whole trunks (or, in the presence of modular engineering rules,† in terms of whole modules of trunks). Thus, the noninteger solution x^*

† The uncertainty in load forecasts and the inherent modularity of some facilities have recently led to the engineering and administration of some networks in modules of trunks.

is subsequently rounded to an integer (or modular) solution, as discussed in Section IV.

III. MINIMIZATION OF THE MULTIHOUR COST FUNCTION

3.1 A reformulation

Equation (10) expresses the cost of the network as the sum of a linear term and $n + 2$ maxima of sets of nonlinear terms. For the analysis which follows, it is convenient to rewrite this cost function in terms of a single maximization operator.

Let $\mathcal{M}$ be a vector-valued index set with elements $\mu = (\mu_1, \mu_2, \ldots, \mu_{n+2})$. Each component of μ, in turn, is a member of the set $\mathcal{H} = \{1, 2, \ldots, H\}$, i.e., $\mathcal{M} = \{\mu = (\mu_1, \ldots, \mu_{n+2}): \mu_i \in \mathcal{H}\}$. Let $\{C_\mu(x): \mu \in \mathcal{M}\}$ be a new family of cost functions, called "elementary cost functions," which are defined by

$$C_\mu(x) \triangleq C_d(x) + C_{f\mu_{n+1}}(x) + C_{s\mu_{n+2}}(x) + \sum_{i=1}^{n} C_{ti\mu_i}(x). \tag{13}$$

In this equation, $C_{ti\mu_i}(x)$ is the cost of sizing the ith tandem completing group for its hour-μ_i load, $C_{f\mu_{n+1}}(x)$ is the cost of sizing the final for its hour-μ_{n+1} load, and $C_{s\mu_{n+2}}(x)$ is the cost of sizing the switch for its hour-μ_{n+2} load, as defined by eqs. (7), (3), and (5), respectively; the functional dependence upon the trunk-group-size vector x is explicitly indicated. It follows from eq. (10) that the multihour cost function is obtained from eq. (13) by maximizing each term on the right-hand side with respect to the appropriate component of μ:

$$C(x) = C_d(x) + \max_{\mu_{n+1} \in \mathcal{H}} C_{f\mu_{n+1}}(x) + \max_{\mu_{n+2} \in \mathcal{H}} C_{s\mu_{n+2}}(x) + \sum_{i=1}^{n} \max_{\mu_i \in \mathcal{H}} C_{ti\mu_i}(x). \tag{14}$$

This term-by-term maximization, however, is equivalent to maximizing $C_\mu(x)$ over all possible choices of μ:

$$C(x) = \max_{\mu \in \mathcal{M}} C_\mu(x). \tag{15}$$

In other words, the multihour cost function can be viewed as the pointwise maximum of the elementary cost functions defined by eq. (13). The multihour engineering problem now has the following form: Determine the vector $x^* \in \mathcal{X}$ with the property that

$$C(x^*) = \min_{x \in \mathcal{X}} \max_{\mu \in \mathrm{M}} C_\mu(x). \tag{16}$$

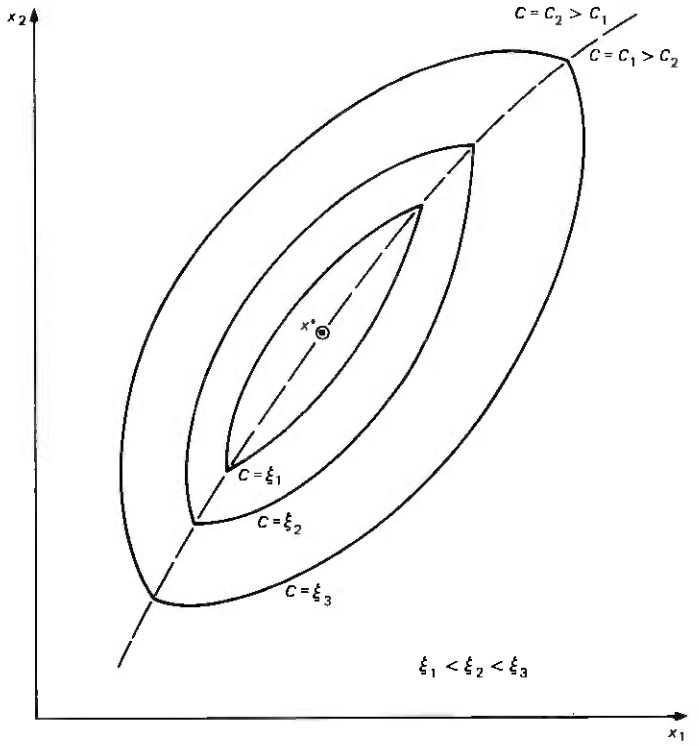


Fig. 2—Level curves of the multihour cost function.

3.2 Some properties of the multihour cost function

The multihour cost function has two fundamental properties on which the algorithm for finding its minimum is based: (*i*) it is a strictly convex function of trunk group sizes; (*ii*) it is only piecewise differentiable in these variables. Each member of the family of functions $C_\mu(x)$, $\mu \in \mathcal{M}$, is the sum of differentiable, strictly convex functions plus a linear term, and is therefore itself strictly convex and differentiable.[4] Since the multihour cost function is the pointwise maximum of a family of strictly convex functions, it is also strictly convex,[4] but not necessarily differentiable everywhere. In particular, if two or more elementary cost functions are maximal at some point (and hence their graphs intersect), the multihour cost function is generally not differentiable at that point.

Figure 2 illustrates the possible nondifferentiability of the multihour cost function for an example with two high-usage groups, and in which only two distinct elementary cost functions [denoted simply by $C_1(x)$ and $C_2(x)$] are maximal anywhere. The dashed curve separates the two regions in the $x_1 - x_2$ plane in which $C_1(x) > C_2(x)$ and $C_2(x) > C_1(x)$,

respectively. The solid lines are the level curves of $C(x)$, i.e., the loci of points for which $C(x) = \xi$, where ξ is a constant. The location of the minimum, x^*, is indicated by the circled point. Clearly, the multihour cost function is not differentiable anywhere along the dashed curve, where the graphs of the two elementary cost functions intersect.

This simple example also illustrates why a coordinate-search algorithm may fail to converge to the minimal solution. Figure 3 shows the same level curves as Fig. 2, together with three typical paths which a coordinate-search algorithm might follow: Path I (0-a-b) and Path II (0-c-d-e) start at the same initial point, but their orders of search are reversed. Path III (0′-f) starts with a different initial solution. Note that the three paths terminate at three different locations (b, e, and f), none of which is the minimal solution. In this example, the algorithm stops whenever it reaches a point x for which $C_1(x) = C_2(x)$ and at which no further decrease in the function $C(x)$ can be achieved by changing only one coordinate at a time.

3.3 A feasible search direction

The principle of the algorithm presented in this paper is to perform a sequence of searches through $\mathcal{X}$, in "feasible directions of descent." A feasible direction of descent is the direction of any vector $y(x)$ with the property that if $x \in \mathcal{X}$, there exists some $\lambda > 0$ such that $x + \lambda y \in \mathcal{X}$ and $C(x + \lambda y) < C(x)$. Whenever such a direction exists, a step size for the search is chosen to maximize the decrease of the multihour cost function in that direction, while maintaining the feasibility of the solution.

In order to determine a direction of descent, we use the concept of the one-sided directional derivative of $C(x)$ with respect to a vector $y \in \mathcal{Y}(x)$, where $\mathcal{Y}(x) = \{y \in R^n$: for $x \in \mathcal{X}$ and for some $\lambda > 0$, $x + \lambda y \in \mathcal{X}\}$. This derivative is denoted by $C'(x; y)$ and is defined as follows:

$$C'(x; y) \triangleq \lim_{\lambda \downarrow 0} \frac{C(x + \lambda y) - C(x)}{\lambda}. \tag{17}$$

$C'(x; y)$ is thus the rate of change of the function $C(x)$ in the direction of y, multiplied by $\|y\|$, where $\|\cdot\|$ is the Euclidean norm. If $C(x)$ is convex, $C'(x; y)$ exists and is a convex function of y for every x at which C is finite. If $C(x)$ is actually differentiable at x, then

$$C'(x; y) = \langle y, \nabla C(x) \rangle \tag{18}$$

where ∇ is the gradient operator and $\langle \cdot, \cdot \rangle$ denotes the scalar (or inner) product of two vectors.[4]

Substituting eq. (15) into the definition of the directional derivative

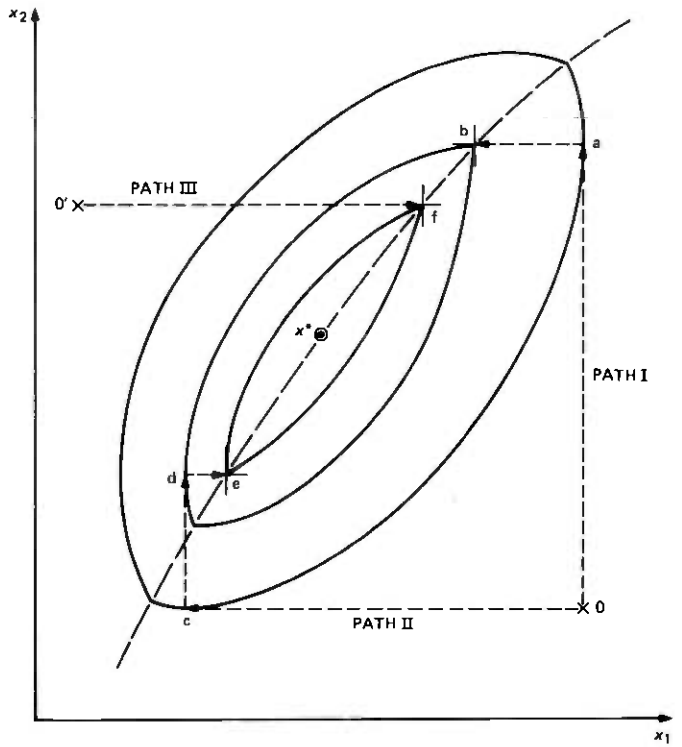


Fig. 3—Typical paths for a coordinate-search algorithm.

we have

$$C'(x;y) = \lim_{\lambda \downarrow 0} \frac{\max_{\mu \in \mathcal{M}} C_\mu(x + \lambda y) - C(x)}{\lambda}. \tag{19}$$

Let $\mathcal{I}(x) \subseteq \mathcal{M}$ be the set of indices of those elementary cost functions which are maximal at x:

$$\mathcal{I}(x) \triangleq \{\mu: C_\mu(x) = C(x)\}. \tag{20}$$

For each $\mu \in \mathcal{M}$, $C_\mu(x)$ is continuous for all $x \in \mathcal{X}$. Consequently, for each $x \in \mathcal{X}$ and for each $y \in \mathcal{Y}(x)$, there exists a $\lambda' > 0$ such that for all λ with $0 < \lambda < \lambda'$,

$$C(x + \lambda y) = \max_{\mu \in \mathcal{I}(x)} C_\mu(x + \lambda y). \tag{21}$$

In words, there exists a neighborhood of x in which no elementary cost function can be maximal which is not also maximal at x. Therefore, the maximization over $\mu \in \mathcal{M}$ in eq. (19) can be replaced by a maximization

over $\mu \in \mathcal{J} \equiv \mathcal{J}(x)$:

$$C'(x; y) = \lim_{\lambda \downarrow 0} \frac{\max_{\mu \in \mathcal{J}} C_\mu(x + \lambda y) - C(x)}{\lambda}. \tag{22}$$

Since $C_\mu(x) = C(x)$ for each $\mu \in \mathcal{J}$,

$$\begin{aligned} C'(x; y) &= \lim_{\lambda \downarrow 0} \max_{\mu \in \mathcal{J}} \left[\frac{C_\mu(x + \lambda y) - C_\mu(x)}{\lambda} \right] \\ &= \max_{\mu \in \mathcal{J}} \lim_{\lambda \downarrow 0} \left[\frac{C_\mu(x + \lambda y) - C_\mu(x)}{\lambda} \right] \\ &= \max_{\mu \in \mathcal{J}} C'_\mu(x; y), \end{aligned} \tag{23}$$

where $C'_\mu(x; y)$ is the directional derivative of $C_\mu(x)$ with respect to y. (The order of the lim and max operators can be interchanged since $\mathcal{J}$ is finite and C_μ is continuous for each $\mu \in \mathcal{J}$.) Since $C_\mu(x)$ is differentiable for each $\mu \in \mathcal{M}$,†

$$C'(x; y) = \max_{\mu \in \mathcal{J}} \langle y, \nabla C_\mu(x) \rangle. \tag{24}$$

Thus, the rate of change of $C(x)$ in the direction of y is negative (i.e., the direction of y is a feasible direction of descent) if and only if

$$\max_{\mu \in \mathcal{J}} \langle y, \nabla C_\mu(x) \rangle < 0, \quad y \in \mathcal{Y}(x) \tag{25}$$

or, equivalently,

$$\langle y, \nabla C_\mu(x) \rangle < 0, \quad \text{for all } \mu \in \mathcal{J}(x), y \in \mathcal{Y}(x). \tag{26}$$

A point at which no feasible direction of descent for $C(x)$ exists must be the location of the minimum of $C(x)$. In fact, a convex function $C(x)$ defined over a convex domain attains its global minimum at $x = x^*$ if and only if $C(x)$ is finite and

$$C'(x^*; y) \geq 0 \quad \text{for all } y \in \mathcal{Y}(x^*). \tag{27}$$

Furthermore, x^* is unique if $C(x)$ is strictly convex.[4] Since $0 \in \mathcal{Y}(x)$, eq. (27) is equivalent to

$$\min_{y \in \mathcal{Y}(x^*)} C'(x^*; y) = 0, \tag{28}$$

† If x is on the boundary of $\mathcal{X}$, $\nabla C_\mu(x)$ is defined as the limit of all sequences $\nabla C_\mu[x^{(1)}]$, $\nabla C_\mu[x^{(2)}], \ldots$, such that $x^{(i)} \in \mathcal{X}$ and $x^{(i)} \to x$.

or, with eq. (24) substituted,

$$\min_{y \in \mathcal{Y}(x^*)} \max_{\mu \in \mathcal{I}(x^*)} \langle y, \nabla C_\mu(x^*) \rangle = 0. \tag{29}$$

3.4 The descent algorithm

Inequality (25) gives the condition for y to be in a feasible direction of descent for $C(x)$. Such a y is generally not unique, and it is necessary to select a particular direction of descent at each iteration of the algorithm. A logical choice is the direction of steepest descent for $C(x)$, i.e., the vector y^* such that

$$C'(x; y^*) = \min_{y \in \mathcal{S} \cap \mathcal{Y}} C'(x; y) \tag{30}$$

where $\mathcal{S}$ is the unit sphere in R^n:

$$\mathcal{S} = \{y: \|y\| \leq 1\}. \tag{31}$$

A solution for y^* can be obtained in explicit form, as shown in the Appendix, provided $\mathcal{I}(x)$ contains either one or two elements, and $\mathcal{Y}(x) = R^n$ (i.e., x is not on the boundary of $\mathcal{X}$).

While it is possible, at least in principle, to solve for the steepest-descent vector in the general case (see the Appendix), the computation is cumbersome for three or more elements in $\mathcal{I}(x)$, or if boundary constraints are active. In this case it is more practical to choose a feasible search direction based on computational simplicity. For example, if $\mathcal{S}$ is chosen to be the set

$$\mathcal{S} = \{y: |y_i| \leq 1, \quad i = 1, \ldots, n\}, \tag{32}$$

the min-max problem expressed by eq. (30) can be converted into a linear program:

$$\begin{gathered} \min \sigma \\ \text{subject to} \\ \langle y, \nabla C_\mu(x) \rangle \leq \sigma \quad \text{for all } \mu \in \mathcal{I}(x) \\ |y_i| \leq 1, \quad i = 1, \ldots, n \\ y_i \geq 0 \quad \text{whenever } x_i = 0. \end{gathered} \tag{33}$$

This linear-programming problem is solved by standard methods. Although the vector y^* which solves that linear program may no longer be in the steepest-descent direction, the algorithm can still be shown to converge.†

† It can be shown that the algorithm will converge with y^* chosen according to eq. (30) as long as $\mathcal{S}$ is any convex, compact subset of R^n containing the origin in its interior.[5]

In principle, the descent algorithm consists of alternately computing a search direction according to eq. (30) and performing a one-parameter search to locate the minimum of $C(x)$ along that direction. (Each such combination is called an iteration.) While this procedure results in a sequence of feasible solutions with strictly decreasing costs, there is still no guarantee that this sequence will converge to the minimal solution. It is possible, as the result of a phenomenon known as "jamming" or "zig-zagging," for the sequence of solutions to converge to a point which does not satisfy eq. (29).[6]

The device used here to prevent jamming (and thus assure convergence to the minimal solution) is similar to that used by Zoutendijk.[6] This device consists of expanding the set $\mathcal{J}(x)$ to include all those elementary cost functions which are "nearly" maximal at x. Let $\epsilon > 0$ and define the new index set

$$\mathcal{J}_\epsilon(x) \triangleq \{\mu: C(x) - C_\mu(x) \leq \epsilon\}. \tag{34}$$

The direction-finding subproblem thus takes into account the directional derivatives of all those elementary cost functions which are within ϵ of being maximal at x. Similarly, the feasibility conditions are modified to prevent the algorithm from attempting to reduce any further those trunk-group sizes which are already "nearly" equal to zero. To this end, let $\delta > 0$ and define the set

$$\mathcal{Y}_\delta(x) \triangleq \{y \in \mathcal{Y}(x): \quad y_i \geq 0 \quad \text{whenever} \quad x_i \leq \delta\}. \tag{35}$$

For notational consistency, the original sets $\mathcal{J}(x)$ and $\mathcal{Y}(x)$ are henceforth denoted by $\mathcal{J}_0(x)$ and $\mathcal{Y}_0(x)$, respectively.

The original problem of determining a search direction y^* as expressed by eq. (30) is now replaced with the problem of finding $y = \hat{y}$ which solves the min-max problem

$$D(x) \triangleq \min_{y \in \mathcal{S} \cap \mathcal{Y}_\delta} \max_{\mu \in \mathcal{J}_\epsilon} \langle y, \nabla C_\mu(x) \rangle. \tag{36}$$

In this definition, the new symbol $D(x)$ replaces the symbol $C'(x; \hat{y})$, since the quantity which it denotes is no longer a directional derivative in the sense of eq. (17). Note, however, that if $\mathcal{J}_\epsilon(x) = \mathcal{J}_0(x)$ and $\mathcal{Y}_\delta(x) = \mathcal{Y}_0(x)$, then $\hat{y} = y^*$ and $D(x) = C'(x; y^*)$.

Since the inclusion of any additional necessary conditions may overly constrain the direction-finding subproblem, the values of ϵ and δ are reduced adaptively throughout the progress of the algorithm, so that $\epsilon \to 0$ and $\delta \to 0$. The use of this procedure also serves a computational purpose, in that ϵ and δ can be viewed as the tolerances within which elementary cost functions are deemed maximal and within which trunk-group sizes are considered to be zero, respectively.

The descent algorithm is now specified as follows:

Step 1 Let k be the iteration counter, and set $k = 0$.
Select an arbitrary initial solution $x^{(0)} \in \mathcal{X}$.
Select ϵ and δ.

Step 2 Compute $C[x^{(k)}]$.

Step 3 Determine the feasible search vector $y^{(k)} \equiv \hat{y}$ and compute $D[x^{(k)}]$.

Step 4 If $D[x^{(k)}] \leq -\epsilon$, go to Step 6.
If $-\epsilon < D[x^{(k)}] < 0$, go to Step 5.
If $D[x^{(k)}] = 0$, but $\mathcal{I}_\epsilon[x^{(k)}] \neq \mathcal{I}_0[x^{(k)}]$ or $\mathcal{Y}_\delta[x^{(k)}] \neq \mathcal{Y}_0[x^{(k)}]$, go to Step 5.
If $D[x^{(k)}] = 0$ and $\mathcal{I}_\epsilon[x^{(k)}] = \mathcal{I}_0[x^{(k)}]$ and $\mathcal{Y}_\delta[x^{(k)}] = \mathcal{Y}_0[x^{(k)}]$, stop. The solution $x^{(k)} = x^*$ has been found.

Step 5 Set $\epsilon = \epsilon/2$ and $\delta = \delta/2$; go to Step 3.

Step 6 Determine a scalar $\lambda^{(k)}$ such that

$$C[x^{(k)} + \lambda^{(k)} y^{(k)}] = \min_{\lambda \in \Lambda^{(k)}} C[x^{(k)} + \lambda y^{(k)}]$$

where

$$\Lambda^{(k)} = \{\lambda : x^{(k)} + \lambda y^{(k)} \in \mathcal{X}\}.$$

Set $x^{(k+1)} = x^{(k)} + \lambda^{(k)} y^{(k)}$, set $k = k + 1$, and go to Step 2.

The adaptive reduction of ϵ and δ is contained in Steps 4 and 5. Whenever $|D[x^{(k)}]|$ becomes sufficiently small (perhaps even zero), ϵ and δ are divided in half. If this reduction results in a decrease of the number of near-maximal elementary cost-functions or near-active boundary constraints, the direction-finding subproblem may be less constrained, and a new search direction may be found. If, on the other hand, the sets $\mathcal{I}_\epsilon(x)$ and $\mathcal{Y}_\delta(x)$ remain unaltered after ϵ and δ are divided in half, $D[x^{(k)}]$ remains unchanged as well, and the algorithm proceeds directly to Step 4.

It can be shown that this algorithm generates a sequence of solutions $\{x^{(k)}; k = 0, 1, 2, \ldots\}$ which is either finite, with its last term x^* satisfying

$$C'(x^*; y^*) = 0, \tag{37}$$

or infinite, with any accumulation point x^* satisfying eq. (37).[5†] It was shown earlier, however, that eq. (37) is a necessary and sufficient condition for x^* to be the unique, minimal noninteger solution to the multihour engineering problem.

† A practical stopping rule is suggested in Section 4.3.

IV. NUMERICAL RESULTS

4.1 The model

Three offices from the California network (Gardena, Compton, and Melrose) were engineered with the descent algorithm developed in Section III. For each office, the loads in two distinct hours (a morning hour and an evening hour) were considered. For the sake of simplicity, the loads A_{fh}, A_{sh}, and A_{ih}, $i = 1, \ldots, n$ (hereafter called "fixed loads," since they do not depend on the variables x_i) were assumed to be zero in both hours. All trunks were assumed to cost \$1000, and the switching cost was \$62/CCS.† The final and tandem-completing groups were assumed to have a common incremental capacity of 30 CCS/trunk. There were 37 high-usage groups in the Compton office, 43 in Gardena, and 35 in Melrose.

These three offices, together with the loads and costs, are the same as those used by M. Eisenberg;[1] they are used here again in order to allow direct comparisons with his results. While the simplifying assumptions in these cases certainly influence the numerical results, they preserve the essential properties of the multihour cost function and thus can be expected to demonstrate convergence characteristics similar to those which would occur in general.

4.2 Implementation of the descent algorithm

The assumption of zero fixed loads on the switch and on the tandem-completing groups allows the multihour cost function to be simplified considerably. Under this assumption, the busy hours on the tandem-completing groups are known *a priori*: the busy hour on the ith tandem-completing group is the same hour in which the load offered to the ith high-usage group is largest. Thus, only the final group and the switch have busy hours which may be functions of the high-usage group sizes. Furthermore, in the absence of fixed loads, the loads offered to the final group and to the switch are identical. Consequently, at most two of the elementary cost functions associated with each of these networks can ever be maximal.‡ For the sake of notational simplicity, these two functions are denoted by $C_1(x)$ and $C_2(x)$, respectively.

The feasible search vector for each iteration was chosen by specifying the set $\mathcal{S}$ to be the unit sphere. For experimental purposes, two distinct initial feasible solutions were used for each of the three offices:

† 1 erlang = 36 CCS (hundred call-seconds per hour).

‡ There are H^{n+2} elementary cost functions associated with a network with n high-usage groups which is engineered for H hours. Recent experience with more extensive data, including fixed loads, indicates that the busy hours of the tandem switch and of the tandem-completing groups are usually not affected by the sizes of the high-usage groups of the office which is being engineered.[7] Thus, one may need to consider only H elementary cost functions in a practical situation.

$$x_i^{(0)} = \max_{h \in \mathcal{H}} a_{ih}, \quad i = 1, \ldots, n \tag{38}$$

and

$$x_i^{(0)} = 0, \quad i = 1, \ldots, n. \tag{39}$$

At each iteration, the optimal step size $\lambda^{(k)}$ was determined by a simple parameter search: First, the location of the minimum of $C(x)$ along the search direction was bracketed between two points whose largest component-difference was 0.01 trunks. The minimal point was then computed more precisely, either by a quadratic approximation or by a linear interpolation, depending on whether $\mathcal{I}_\epsilon(x)$ contained one or two elements at that point. The Erlang-B function for noninteger trunk-group sizes was evaluated by an approximation due to Rapp,[8] and its partial derivatives were approximated by central differences with a step size of 0.01. The initial values for ϵ and δ were set equal to $10^{-3}C[x^{(0)}]$ and 0.1, respectively, and in all cases the algorithm was arbitrarily terminated after 25 iterations.

4.3 Convergence of the descent algorithm

The behavior of the algorithm was similar for all three of the offices and for both starting points. For each office, the final solutions obtained with each of the two starting points differed by less than 0.003 trunks on any high-usage group. Figures 4 to 7 summarize the convergence characteristics of the algorithm for the Gardena office, with the starting point given by eq. (38).

The cost of the network at each iteration, $C[x^{(k)}]$ (or simply $C^{(k)}$), as a function of the iteration number, k, is shown in Fig. 4. The cost decreased monotonically with k, and the rate of change became very small after the first few iterations (e.g., $C^{(4)} = 1.0003\, C^{(25)}$).

Figure 5 shows the magnitude of $D[x^{(k)}]$ (or simply $|D^{(k)}|$, and the value of ϵ, as functions of k. As this figure indicates, $|D^{(k)}|$, ϵ, and δ all approach zero as $k \to \infty$ (recall that $\epsilon < |D^{(k)}|$ for all $k \geq 0$, and that $\delta \sim \epsilon$). Thus, it is evident that the algorithm was converging to the minimal solution when it was terminated.

Unlike $C^{(k)}$, $|D^{(k)}|$ did not decrease monotonically. As the algorithm reduced ϵ, there was an occasional iteration ($k = 0, 2$, and 6) at which the solution point $x^{(k)}$ lay outside the region for which $|C_1^{(k)} - C_2^{(k)}| \leq \epsilon$. As a result, $\mathcal{I}_\epsilon[x^{(k)}]$ contained only one element at these iterations, and $\hat{y}$ was given by $-\nabla\{\max [C_1^{(k)}, C_2^{(k)}]\}$. (The dotted lines in Fig. 5 show the magnitudes of the gradients ∇C_1 and ∇C_2 as functions of k.) For the remaining iterations $\mathcal{I}_\epsilon[x^{(k)}]$ contained both indices, so that $\hat{y}$ and $D^{(k)}$ were computed via eqs. (61) to (63) in the Appendix. Note that since $\epsilon \to 0$ as $k \to \infty$ and $\mathcal{I}_\epsilon[x^{(k)}]$ contained both elements for all $k \geq 7$, the

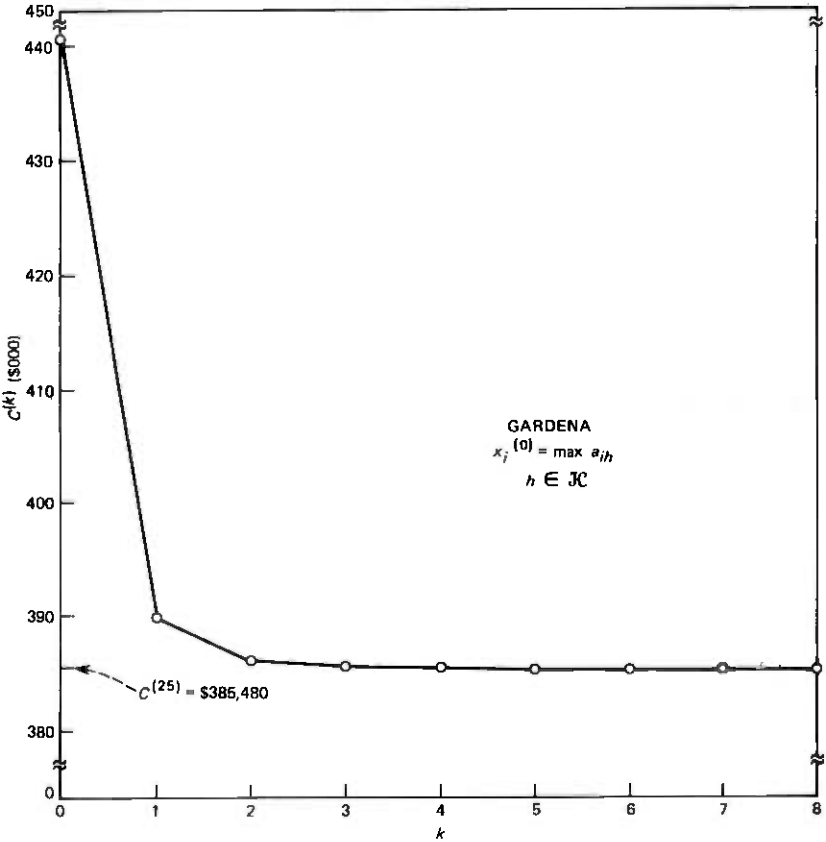


Fig. 4—Convergence of descent algorithm: network cost.

solution point x^* must be located on the intersection of the graphs of the two elementary cost functions.

The relative change in the solution from one iteration to the next, as measured by the Euclidean "distance" $\|x^{(k-1)} - x^{(k)}\|$, is shown in Fig. 6. Note that this quantity also tended to zero as k increased, although not monotonically. In particular, this plot shows that for each iteration (except the first one) at which $\mathcal{J}_\epsilon[x^{(k)}]$ contained only one element, the corresponding step size was small. Thus, the algorithm generated a sequence of solutions which tended to follow the intersection of the graphs of $C_1(x)$ and $C_2(x)$ toward the minimal solution x^*. Whenever the solution point moved too far from this intersection, a small step was taken to get back into the region defined by $|C_1(x) - C_2(x)| \leq \epsilon$, and the search along the intersection was resumed.

Figure 6 also suggests how the norm of the change in trunk-group sizes can be used as a measure for a practical stopping rule. Let Θ be a prese-

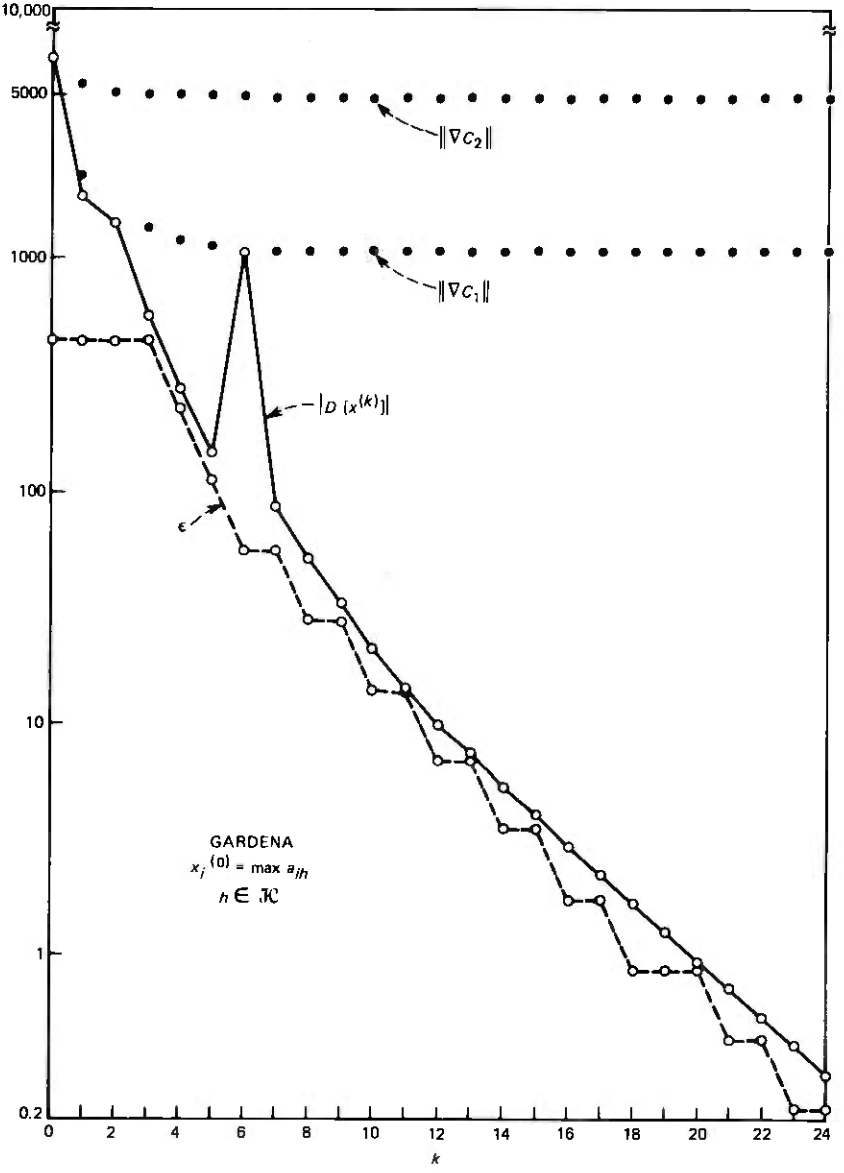


Fig. 5—Convergence of descent algorithm: $|D[x^{(k)}]|$ and ϵ.

lected threshold. Then the algorithm is deemed to have converged, and is terminated, if $k \geq 2$ and $\|x^{(k)} - x^{(k-1)}\| < \Theta$ for two consecutive iterations. The last solution $x^{(k)}$ is then an approximation to the exact solution x^*.

Figure 7 shows how the solution point $x^{(k)}$ converged. For this purpose,

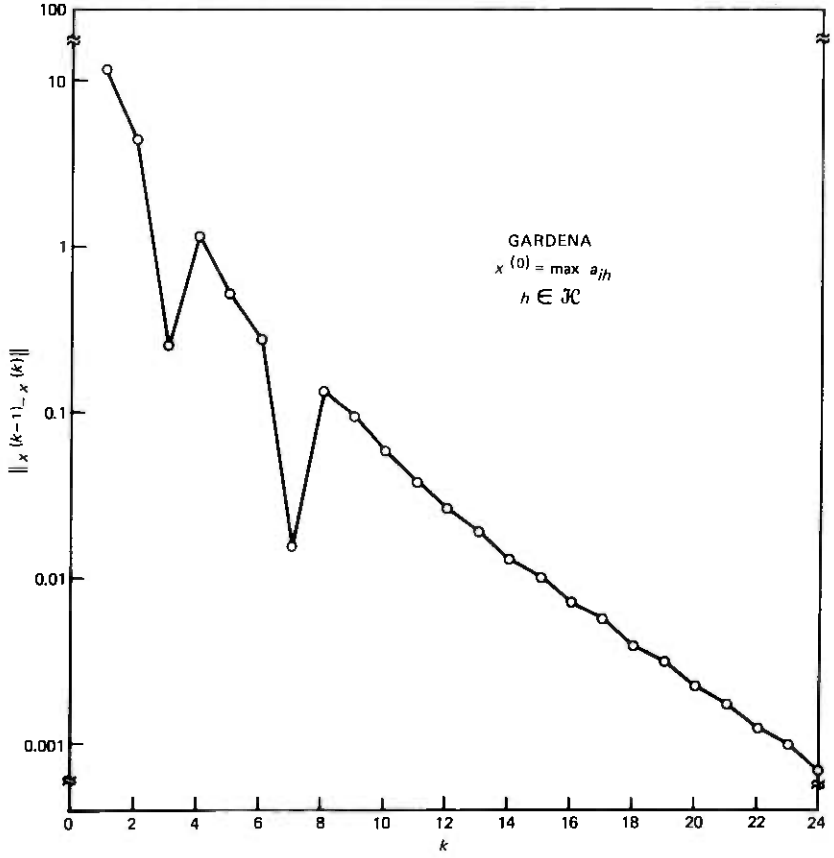


Fig. 6—Convergence of descent algorithm: change in trunk group sizes.

the Euclidean distance between $x^{(k)}$ and $x^{(25)}$, $\|x^{(k)} - x^{(25)}\|$, is plotted as a function of k. Note that $x^{(k)}$ reached a small neighborhood of $x^{(25)}$ in relatively few iterations (e.g., within one trunk after only four iterations, and within 0.1 trunks after ten iterations).

4.4 Further results

Table I shows the offered loads for the Gardena office (in CCS), and the trunk-group sizes (optimal, and rounded to the nearest integer) obtained by the descent algorithm. For the purpose of comparison, the following other sets of trunk-group sizes are included:

(*i*) For each high-usage group, the smallest and the largest number of trunks (in integers) selected from a set of solutions generated by the coordinate-search algorithm with 20 combinations of starting points and trunk-group orderings.

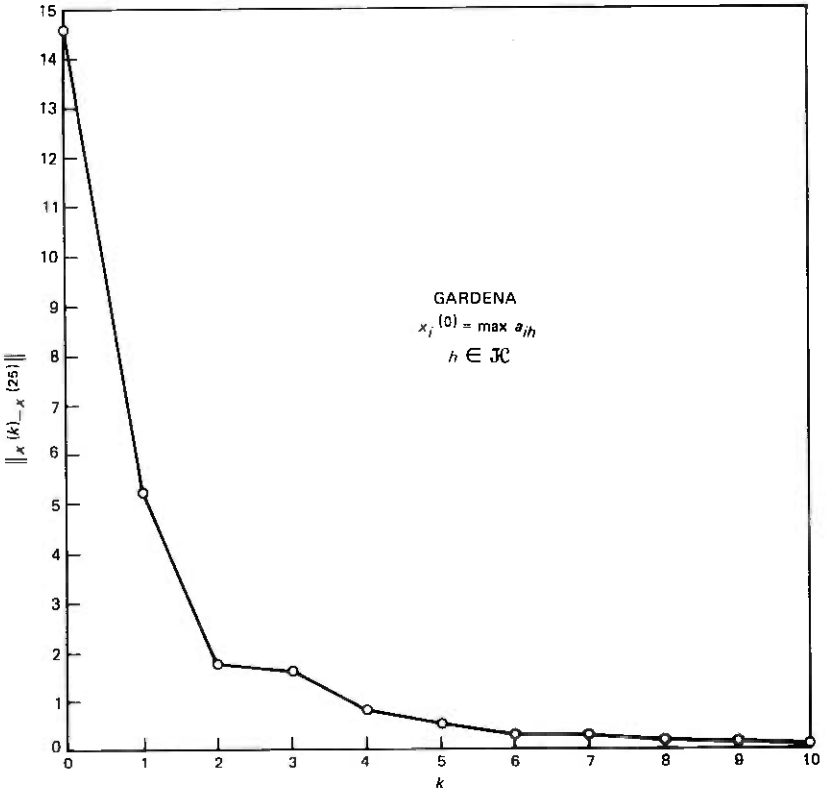


Fig. 7—Convergence of descent algorithm: trunk group sizes.

(*ii*) The number of trunks in each high-usage group (again in integers) as determined by the so-called "cluster-busy-hour" method.[1†]

This table illustrates the trunk-group-size variability when a coordinate-search algorithm is used, relative to the optimal—and hence unique—solution. Note that some groups varied by as many as eight trunks, while only six groups showed no variability. In view of the practical problems associated with such an uncertainty, as discussed in the Introduction, the need for obtaining a unique solution is clear.

Table II lists the costs[‡] of the optimal solutions for Compton, Gardena,

† Cluster-busy-hour engineering is a traditional method for sizing traffic networks, in which trunk-group sizes are chosen to minimize an elementary cost function whose alternate-route costs are all evaluated in a single, time-consistent hour. The hour selected is that hour in which the sum of the first-offered loads to all the high-usage groups which overflow to a common final group, plus the first-offered load to that final group, is largest.

‡ All costs are computed via eq. (3), and thus include the marginal-capacity approximation. The costs of comparable networks reported in Ref. 1 reflect the actual required sizes of the final groups as determined after the high-usage groups have been sized.

and Melrose. For the purpose of comparison, the table also lists the costs of the following other networks:

(*i*) A network with the optimal trunk-group sizes rounded to the nearest integers.

(*ii*) The two networks with the lowest and highest costs, respectively, selected from the set of 20 solutions generated by the coordinate-search algorithm.

(*iii*) The cluster-busy-hour network.

These results show that while some combinations of starting points and trunk-group orderings for the coordinate-search algorithm yielded solutions whose costs were only a fraction of a percent higher than the optimum, other combinations led to substantially higher costs (up to 5.5 percent in the case of Compton).

Since a network is realizable only in integer trunk-group sizes, the optimal (noninteger) solution must be rounded in some way. As indicated by the results in Table II, rounding of optimal trunk-group sizes to the nearest integers is an attractive alternative: It is simple; it yields an essentially unique solution; and, although it generally does not lead to the optimal integer solution, it yields networks whose costs are only slightly higher (from 0.2 to 0.44 percent in the three cases examined) than those of the optimal, noninteger solutions. (Subsequent studies[10] have shown that, among several practical approaches, rounding the optimal solution to the nearest integers, or to the nearest multiples of the module size, is indeed the policy most likely to minimize cost.)

The cluster-busy-hour networks are included to provide some perspective. It is evident that while the cost of any solution obtained by the coordinate-search algorithm is substantially lower than the cost of the cluster-busy-hour solution, additional nonneligibile capital savings may be obtained by computing the optimal solution via the descent algorithm.

A comparison of the rounded optimal solution with the cluster-busy-hour solution reveals that these two networks are not very different. (The average absolute difference in high-usage group sizes is only 0.8 trunks, while the average group size for the rounded optimal solution is 7.1 trunks.) The cost of the cluster-busy-hour network, however, is 11.7 percent higher than that of the integerized optimal solution. The sensitivity of the cost to relatively small changes in trunk-group sizes is a consequence of the "shape" of the multihour cost function. As Fig. 2 suggests, the contours of this function are long and narrow, and the slope is steep in directions normal to the intersection between the two elementary cost functions.

In contrast, $C(x)$ is much less sensitive to changes in x along the intersection of the graphs of $C_1(x)$ and $C_2(x)$. It is for this reason that the

Table I — High usage trunk group sizes—Gardena

Trunk Group	Offered loads (CCS)		Number of Trunks				
			Descent algorithm		Coordinate-search algorithm		Cluster-busy-hour
	Hour 1	Hour 2	Optimal	Rounded	Low	High	
1	60	140	4.42	4	4	6	3
2	119	9	5.25	5	2	6	6
3	82	20	3.78	4	2	4	4
4	305	76	11.97	12	10	12	12
5	30	0	1.47	1	1	2	2
6	59	7	2.81	3	1	3	3
7	102	56	4.64	5	5	5	5
8	256	161	10.32	10	9	11	11
9	366	230	14.10	14	12	15	15
10	469	310	17.57	18	14	18	18
11	115	115	5.37	5	5	5	5
12	144	34	6.20	6	3	7	7
13	206	335	10.81	11	10	13	9
14	310	650	18.58	19	16	22	13
15	284	319	12.14	12	12	13	12
16	93	152	5.43	5	5	6	4
17	17	24	1.08	1	1	1	1
18	74	325	8.92	9	6	13	4
19	102	158	5.74	6	5	7	5
20	137	322	9.36	9	8	13	6
21	222	247	9.78	10	10	10	9
22	252	390	12.58	13	7	15	11
23	445	194	16.73	17	12	17	17
24	176	86	7.41	7	6	8	8
25	83	29	3.83	4	2	4	4
26	98	21	4.43	4	4	5	5
27	158	74	6.75	7	6	7	7
28	124	36	5.44	5	5	6	6
29	54	25	2.64	3	2	3	3
30	38	1	1.86	2	1	2	2
31	31	17	1.60	2	1	2	2
32	140	46	6.06	6	6	6	6
33	96	30	4.35	4	3	5	5
34	122	62	5.40	5	4	6	6
35	163	57	6.92	7	5	7	7
36	163	72	6.93	7	5	7	7
37	296	238	11.84	12	11	12	12
38	33	28	1.77	2	2	2	2
39	240	3	9.70	10	6	10	10
40	136	7	5.90	6	4	6	6
41	54	4	2.59	3	2	3	3
42	52	35	2.61	3	2	3	3
43	206	9	8.48	8	3	9	9

Table II — Network costs

Office	Cost ($000)				
	Descent algorithm		Coordinate-search algorithm		Cluster-busy-hour
	Optimal solution	Integerized solution	Low	High	
Compton	402.9	403.7	406.8	425.0	488.6
Gardena	385.5	386.6	388.2	397.9	431.9
Melrose	271.7	272.9	273.7	276.0	305.3

largest cost difference between the 20 sample solutions generated by the coordinate-search algorithm is only 4.5 percent. (As suggested by Fig. 3, all termination points for the coordinate-search algorithm lie on this intersection in this particular example.)

The high sensitivity of network cost to trunk variations in some directions is not, of course, a consequence of engineering a network to a multihour (minimum-cost) solution. The cost function $C(x)$ as defined by eq. (10) represents the actual cost of the network. The multihour method is simply the one which recognizes this actual cost during the sizing process.

The imposition of modular engineering rules tends to diminish the capital savings of multihour engineering over single-hour engineering, by blurring some of the fine structure of the networks. A substantial portion of the savings can still be realized, however, as long as the module size is not large relative to the average group size. For example, in the three networks studied here, where the average group size is 7.4, rounding the high-usage group sizes to the nearest multiples of six trunks resulted in a reduction of the original savings by approximately one-fourth.

V. SUMMARY

The cost of a traffic network which gives a minimum specified grade of service on the last-choice routes in more than one hour can be approximated by a strictly convex, although possibly nondifferentiable, function of the high-usage trunk-group sizes. A descent algorithm, which can be shown to converge to the unique, noninteger minimum-cost network, has been developed. The noninteger solution is subsequently rounded to the nearest allowable integer (or modular) solution to yield a realizable network. The main advantage of this algorithm relative to the coordinate-search method is that the uniqueness of the solution prevents unnecessary, expensive rearrangements from being undertaken as traffic loads change with time. A secondary advantage is a small possible additional saving in network capital cost.

The results obtained from applying the descent algorithm to three numerical examples (and to others not discussed here) demonstrate that even after only a few iterations a sufficiently high degree of precision can be achieved to ensure the reproducibility of the results and hence the stability which motivated the design of the algorithm.

VI. ACKNOWLEDGMENTS

The author wishes to acknowledge helpful discussions with E. J. Messerli and R. Saigal.

APPENDIX

The steepest-descent direction for the multihour cost function

Consider the problem of finding a vector $y^* \in R^n$ with the property that

$$C'(x, y^*) = \min_{\|y\| \le 1} C'(x; y)$$

$$= \min_{\|y\| \le 1} \max_{\mu \in \mathcal{J}(x)} \langle y, \nabla C_\mu(x) \rangle \tag{40}$$

where $C'(\cdot\,;\cdot)$, $C_\mu(x)$ and $\mathcal{J}(x)$ have all been defined in Section III. Without loss of generality, let $\mathcal{J}(x) = \{1, \ldots, m\}$ and let Z be the convex hull of the set of all the inner products $\langle y, \nabla C_j(x) \rangle$, $j = 1, \ldots, m$:

$$Z = \left\{ z : z = \sum_{j=1}^{m} \lambda_j \langle y, \nabla C_j(x) \rangle, \sum_{j=1}^{m} \lambda_j = 1, \quad \lambda_j \ge 0 \right\}. \tag{41}$$

In other words, Z is the shortest closed segment of the real line which contains all the inner products $\langle y, \nabla C_j(x) \rangle$, $j = 1, \ldots, m$. Consequently, the maximal inner product is also the maximal element in Z:

$$\max_{j \in \mathcal{J}(x)} \langle y, \nabla C_j(x) \rangle = \max_{z \in Z} z. \tag{42}$$

Since the inner product is a linear operator, the elements z defined by eq. (41) can be rewritten as

$$z = \left\langle y, \sum_{j=1}^{m} \lambda_j \nabla C_j(x) \right\rangle, \sum_{j=1}^{m} \lambda_j = 1, \quad \lambda_j \ge 0. \tag{43}$$

Now define a new set, $\mathcal{G}$, as the convex hull of the gradients of the elementary cost functions:

$$\mathcal{G} = \left\{ g : g = \sum_{j=1}^{m} \lambda_j \nabla C_j(x), \sum_{j=1}^{m} \lambda_j = 1, \quad \lambda_j \ge 0 \right\}. \tag{44}$$

The relationship between the sets $\mathcal{G}$ and Z is evidently

$$Z = \{ z : z = \langle y, g \rangle, g \in \mathcal{G} \} \tag{45}$$

and thus

$$\max_{z \in Z} = \max_{g \in \mathcal{G}} \langle y, g \rangle \tag{46}$$

Combining eqs. (40), (42), and (46) yields

$$C'(x; y^*) = \min_{\|y\| \le 1} \max_{g \in \mathcal{G}} \langle y, g \rangle. \tag{47}$$

Since the two constraint sets on the right-hand side of eq. (47) are convex and compact, the minimization and maximization operations can be interchanged:[9]

$$C'(x; y^*) = \max_{g \in \mathcal{G}} \min_{\|y\| \le 1} \langle y, g \rangle. \tag{48}$$

For any $g \in \mathcal{G}$ with $g \neq 0$,

$$\begin{aligned} \min_{\|y\| \le 1} \langle y, g \rangle &= \left\langle -\frac{g}{\|g\|}, g \right\rangle \\ &= -\|g\|. \end{aligned} \tag{49}$$

If $g = 0$, then

$$\min_{\|y\| \le 1} \langle y, g \rangle = 0. \tag{50}$$

Equation (48) is then equivalent to

$$\begin{aligned} C'(x; y^*) &= \max_{g \in \mathcal{G}} (-\|g\|) \\ &= -\min_{g \in \mathcal{G}} \|g\| \end{aligned} \tag{51}$$

Let g^* be the vector with minimum norm in $\mathcal{G}$, i.e.,

$$\|g^*\| = \min_{g \in \mathcal{G}} \|g\|. \tag{52}$$

The desired result is now given by

$$C'(x; y^*) = -\|g^*\| \tag{53}$$

$$y^* = \begin{cases} 0 & , g^* = 0 \\ -\dfrac{g^*}{\|g^*\|} & , g^* \neq 0. \end{cases} \tag{54}$$

The set $\mathcal{G}$ and its elements g are called the "subdifferential" and the "subgradients," respectively, of the convex function $C(x)$.[4] The steepest-descent vector y^* is then in the negative direction of the minimum-norm subgradient of $C(x)$. Note, incidentally, that $C'(x^*; y^*) = 0$—the necessary and sufficient condition for $C(x^*)$ to be the minimum—is equivalent to the condition that $0 \in \mathcal{G}$ at x^*.

Explicit solutions for y^* can now be found for the cases $m = 1$ and $m = 2$, as follows:

(*i*) $m = 1$:

In this case we simply have

$$g = \nabla C_1(x) = \nabla C(x) = g^*. \tag{55}$$

(*ii*) $m = 2$:

The subdifferential is now given by

$$\mathcal{G} = \{g: g = \beta \nabla C_1(x) + (1 - \beta)\nabla C_2(x), \quad 0 \le \beta \le 1\}. \tag{56}$$

If $0 \in \mathcal{G}$, then $y^* = g^* = 0$. Suppose now that $0 \notin \mathcal{G}$. The unconstrained minimum of $\|g\|$ is found by setting its derivative (with respect to β) equal to zero:

$$\frac{d}{d\beta} \|g\| = 0. \tag{57}$$

Since

$$\|g\|^2 = \langle g, g \rangle, \tag{58}$$

the relationship

$$2\|g\| \frac{d}{d\beta} \|g\| = \frac{d}{d\beta} \langle g, g \rangle \tag{59}$$

is obtained. Thus, since $\|g\| \neq 0$, eq. (57) is equivalent to

$$\frac{d}{d\beta} \langle g, g \rangle = 0. \tag{60}$$

Let $\beta = \hat{\beta}$ satisfy eq. (60). Expanding the inner product, taking the derivative, and solving for $\hat{\beta}$ yields

$$\hat{\beta} = \frac{\|\nabla C_2\| - \langle \nabla C_1, \nabla C_2 \rangle}{\|\nabla C_1\|^2 + \|\nabla C_2\|^2 - 2\langle \nabla C_1, \nabla C_2 \rangle}. \tag{61}$$

However, since β is constrained by $0 \le \beta \le 1$, the minimum-norm subgradient g^* is given by

$$g^* = \beta^* \nabla C_1 + (1 - \beta^*)\nabla C_2 \tag{62}$$

where

$$\beta^* = \begin{cases} 0, & \hat{\beta} < 0 \\ \hat{\beta}, & 0 \le \hat{\beta} \le 1 \\ 1, & \hat{\beta} > 1 \end{cases} \tag{63}$$

REFERENCES

1. M. Eisenberg, "Engineering Traffic Networks for More Than One Busy Hour," B.S.T.J., *56*, No. 1 (January 1977), pp. 1–20.

2. E. J. Messerli, "Proof of a Convexity Property of the Erlang-B Formula," B.S.T.J., *51*, No. 4 (April 1972), pp. 951–953.
3. C. J. Truitt, "Traffic Engineering Techniques for Determining Trunk Requirements in Alternate Routing Trunk Networks," B.S.T.J., *39*, No. 2 (March 1954), pp. 277–302.
4. R. T. Rockafellar, *Convex Analysis*, Princeton, New Jersey: Princeton University Press, 1970.
5. M. D. Cannon, C. D. Cullum, Jr., and E. Polak, *Theory of Optimal Control and Mathematical Programming*, New York: McGraw-Hill, 1970.
6. G. Zoutendijk, *Methods of Feasible Directions*, Amsterdam: Elsevier, 1960.
7. M. Eisenberg, "Multihour Engineering in Alternate-Route Networks," Eighth International Teletraffic Conference, Melbourne, Australia, November 10–17, 1976.
8. Y. Rapp, "Planning of Junction Network in a Multiexchange Area," (Part I), Ericsson Technics, *20*, No. 1 (1964), pp. 77–130.
9. D. G. Luenberger, *Optimization by Vector Space Methods*, New York: Wiley, 1969, p. 208.
10. W. B. Elsner, unpublished work.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 8, October 1977
Printed in U.S.A.

On Blocking Probabilities for Switching Networks

By F. R. K. CHUNG and F. K. HWANG

(Manuscript submitted March 30, 1977)

We study the blocking probabilities of multistage switching networks through their linear graphs using Lee's model. We give results which allow us to compare the blocking probabilities of various classes of linear graphs. In particular, we derive techniques for deciding when the blocking probability of one linear graph does not exceed the blocking probability of another linear graph under all possible traffic loads. This allows us to compare the blocking performances of corresponding switching networks containing these linear graphs. Our results apply not only to series-parallel linear graphs, but also to the more general "spider-web" linear graphs, which have recently attracted substantial interest in the theory of switching networks.

I. INTRODUCTION

A network N consists of a set of switches, a set of links, and two sets of terminals denoted by I and Ω, and called, respectively, the set of input terminals and output terminals. The union of all paths that can be used to connect one call between an input terminal u and an output terminal v is called the *linear graph* determined by u and v, and is denoted by $G(u,v)$. (A linear graph is also called a channel graph.[10]) Let P^* be the union of all paths connecting input terminals to output terminals. A *state* of N is a subset S of P^* such that no two paths in S have a common link. For a given state S, a link is *busy* if it is on a path in S. Otherwise it is *idle*.

Many existing switching networks consist of several stages. We say that N is an *n-stage network* if the set of switches of N can be partitioned into n sets, called *stages*, and links exist only between a switch in stage i and a switch in stage $i + 1$, for $1 \leq i \leq n - 1$. All input terminals are connected to switches in the first stage and all output terminals are connected to switches in the last stage.

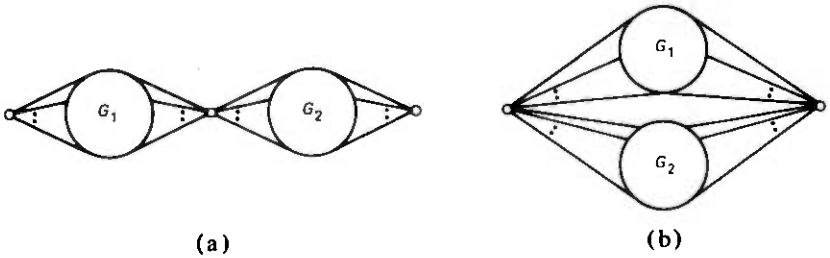


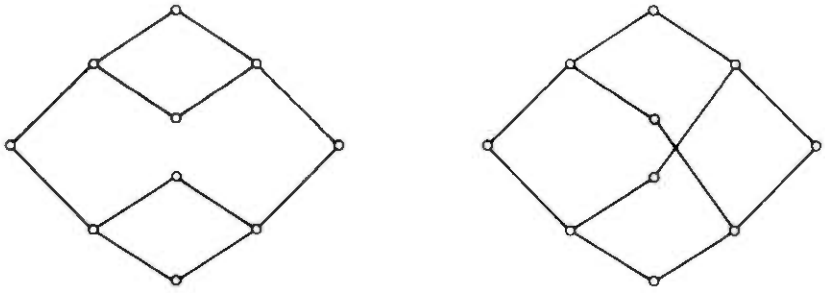
Fig. 1—(a) Series combination. (b) Parallel combination.

In order to simplify the analysis of the switching networks under consideration we will employ Lee's model in Ref. 8. We will also use Lee's independence assumption, namely, that the probabilities of being busy for links in successive stages are independent. Thus, we will assume all links between stage i and stage $i + 1$ have some probability p_i of being busy and some probability $q_i = 1 - p_i$ of being idle, for any i, $1 \leq i \leq k - 1$. Let $P(u,v)$, $u \in I$, $v \in \Omega$, denote the probability that there does not exist a path connecting u and v which consists of idle links. $P(u,v)$ is called the blocking probability for u and v. Note that because of the independence assumption, $P(u,v)$ actually only depends on the linear graph $G(u,v)$ between u and v. Furthermore, we will assume all switches in the same stage are of the same size (i.e., for any switch in stage i, there are r_i inlet lines and r_i' outlet lines).

A network is said to be *balanced* if all the linear graphs $G(u,v)$, $u \in I$, $v \in \Omega$, are isomorphic.[4] It is said to be *partially balanced* if there are only relatively few nonisomorphic linear graphs. We can then compare the blocking probabilities of two partially balanced switching networks by comparing the blocking probabilities of the corresponding linear graphs.

A linear graph is said to be a *series-parallel* linear graph if it is either a series combination or a parallel combination of two series-parallel linear graphs of smaller sizes (see Fig. 1a,b). A linear graph is said to be a *spider-web* linear graph if it is not series-parallel. In Fig. 2 we give examples of a series-parallel linear graph (Fig. 2a) and a spider-web linear graph (Fig. 2b). A linear graph $G(u,v)$ is said to be a *multilink* linear graph if there exist two switches in $G(u,v)$ which are connected by more than one link. Any linear graph which is not a multilink graph is said to be a *simple-link* linear graph.

In this paper, we present several general methods for comparing blocking probabilities of various classes of switching networks. These methods generalize and improve previous results in this area.[2,7] These results can be applied not only to series-parallel linear graphs but also to more general spider-web linear graphs. We also consider the general case in which two switches can be connected by more than one link.



(a) (b)

Fig. 2—(a) A series-parallel linear graph. (b) A spider-web linear graph.

II. LINEAR GRAPHS IN THREE-STAGE NETWORKS

We denote an n-stage network by the following:

(i) The switch set

$$\bigcup_{i=1}^{n} \{s(i,j): 1 \le j \le t_i\}$$

where the stage i consists of t_i switches which are labeled by $s(i,j)$, $j = 1,2,\ldots,t_i$;

(ii) The link set

$$\bigcup_{i=1}^{n} \{L(i,j,k): 1 \le j \le t_i,\ 1 \le k \le t_{i+1}\}$$

where $L(i,j,k)$ denotes the set of links connecting $s(i,j)$ and $s(i+1,k)$;

(iii) I and Ω, the input and output terminals, respectively. We note that for fixed i we have

$$\sum_{k=1}^{t_{i-1}} \ell(i-1,k,j) = \sum_{k=1}^{t_{i-1}} \ell(i-1,k,j') = r_i$$

$$\sum_{k=1}^{t_{i+1}} \ell(i,j,k) = \sum_{k=1}^{t_{i+1}} \ell(i,j',k) = r_i'$$

for any j,j', $1 \le j, j' \le t_i$, where $\ell(i,j,k)$ denotes the cardinality of $L(i,j,k)$.

An n-stage linear graph $G(u,v)$ can then be characterized by the following:

(i) The switch set is

$$\bigcup_{i=1}^{n} s_i'$$

where s_i' is a subset of the switch set in stage i and $s_1' = \{u\}$, $s_n' = \{v\}$. (We

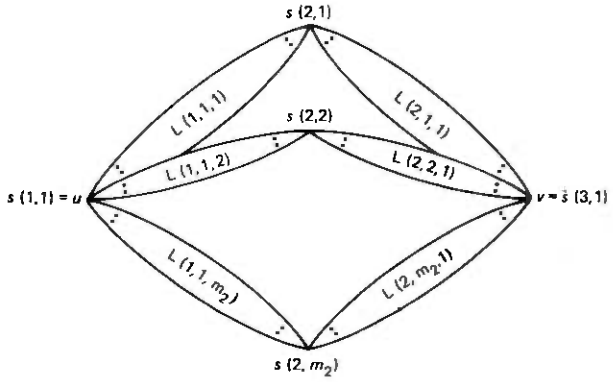


Fig. 3—Three-stage linear graphs.

relabel switches if necessary so that $s_i' = \{s(i,j): 1 \le j \le m_i\}$ for some $m_i \le t_i$, $m_1 = m_n = 1$);

(*ii*) The link set is $\{L(i,j,k): s(i,j)$ and $s(i+1,k)$ are in the switch set of $G(u,v)\}$.

Let $G'(u',v')$ be an n-stage linear graph with the set of switches

$$\bigcup_{i=1}^{n} \{s'(i,j): 1 \le j \le m_i'\}$$

and the set of links $\{L'(i,j,k)\}$. We say $G(u,v)$ and $G'(u',v')$ are *isomorphic* if the following conditions are satisfied.

(*i*) $m_i = m_i'$ for $1 \le i \le n$;

(*ii*) The set of switches in each stage can be properly relabeled such that the following holds:

$$\ell(i,j,k) = \ell'(i,j,k).$$

Now, we consider a three-stage linear graph as shown in Fig. 3 (where switches in middle stages are labeled $s(2,1), \ldots, s(2,m_2)$).

Theorem 1: Let $G(u,v)$ be the linear graph with the set of switches

$$\bigcup_{i=1}^{3} \{s(i,j): 1 \le j \le m_i\}$$

and the set of links $\{L(i,j,k)\}$.

Let $G'(u',v')$ be the linear graph with the set of switches

$$\bigcup_{i=1}^{3} \{s'(i,j): I \le j \le m_i'\}$$

and the set of links $\{L'(i,j,k)\}$. Moreover, suppose $G(u,v)$ and $G'(u',v')$ satisfy the following conditions (see Fig. 4a,b):

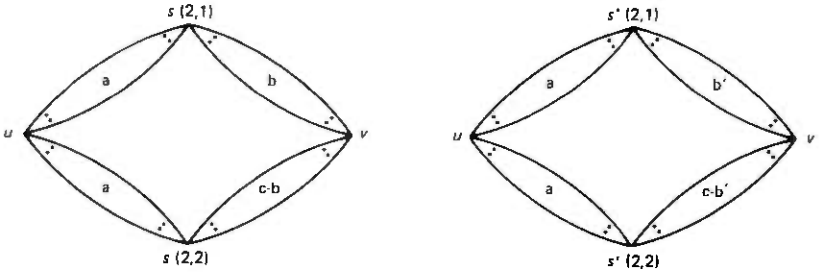


Fig. 4—Graphs for Theorem 1.

(*i*) $m_2 = m_2' = 2$
(*ii*) $\ell(1,1,1) = \ell(1,1,2) = \ell'(1,1,1) = \ell'(1,1,2)$
(*iii*) $\ell(2,1,1) + \ell(2,2,1) = \ell'(2,1,1) + \ell'(2,2,1)$
(*iv*) $|\ell(2,1,1) - \ell(2,2,1)| \leq |\ell'(2,1,1) - \ell'(2,2,1)|$

where $\ell(i,j,k)$, $\ell'(i,j,k)$ denote the cardinalities of $L(i,j,k)$, $L'(i,j,k)$, respectively. Then we have $P(u,v) \leq P(u',v')$.

Proof: Let p_i denote the probability of a link being busy between stage i and stage $i+1$, $i = 1,2$. Let

$$a = \ell(1,1,1) = \ell(1,1,2) = \ell'(1,1,1) = \ell'(1,1,2)$$

and

$$c = \ell(2,1,1) + \ell(2,2,1) = \ell'(2,1,1) + \ell'(2,2,1).$$

We may assume without loss of generality that

$$b = \ell(2,1,1) \leq \ell(2,2,1),$$

$$b' = \ell'(2,1,1) \leq \ell'(2,2,1).$$

It is easy to verify that $b' \leq b \leq c/2$. Define the function $f(x)$ as follows:

$$f(x) = [1 - (1 - p_1^a)(1 - p_2^x)][1 - (1 - p_1^a)(1 - p_2^{c-x})]$$

We note that $P(u,v) = f(b)$ and $P(u',v') = f(b')$. Furthermore, f attains its minimum at $x = c/2$ and f is a convex function. Thus we have

$$f(b) \leq f(b')$$

and

$$P(u,v) \leq P(u',v').$$

We note that the number of paths connecting u and v in $G(u,v)$ is ac, which is also equal to the number of paths connecting u' and v' in $G'(u',v')$.

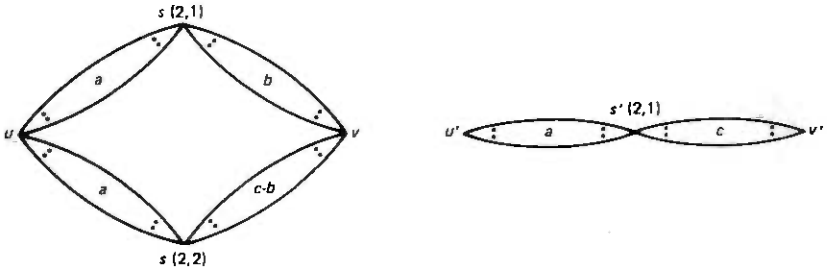


Fig. 5—Graphs for Theorem 2.

The following theorem can be viewed as a special case of Theorem 1. Because it is very useful in comparing linear graphs, we will state it here.

Theorem 2: Let $G(u,v)$ be the linear graph with the set of switches

$$\bigcup_{i=1}^{3} \{s(i,j): 1 \le j \le m_i\}$$

and the set of links $\{L(i,j,k)\}$, and let $G'(u',v')$ be the linear graph with the set of switches

$$\bigcup_{i=1}^{3} \{s'(i,j): 1 \le j \le m_i'\}$$

and the set of links $\{L'(i,j,k)\}$.

Suppose $G(u,v)$ and $G'(u',v')$ satisfy the following conditions (see Fig. 5a,b):

(*i*) $m_2 = 2, m_2' = 1,$
(*ii*) $\ell(1,1,1) = \ell(1,1,2) = \ell'(1,1,1),$
(*iii*) $\ell(2,1,1) + \ell(2,2,1) = \ell'(2,1,1).$

where $\ell(i,j,k)$, $\ell'(i,j,k)$ denote the cardinalities of $L(i,j,k)$, $L'(i,j,k)$, respectively.
Then we have

$$P(u,v) \le P(u',v')$$

Theorem 2 can be proved by taking $b' = 0$ in Theorem 1.

In the following corollary, we give a short proof for the main theorem in Ref. 4, which asserts that a multilink linear graph can always be replaced by a simple-link linear graph having smaller blocking probability whereas the total numbers of paths in the two linear graphs are the same.

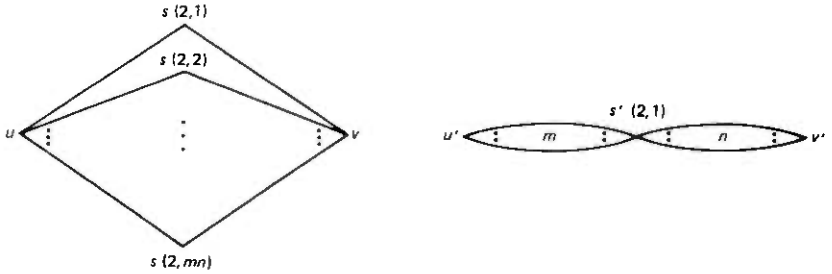


Fig. 6—(a) A single-link linear graph. (b) A multilink linear graph.

Corollary: Let $G'(u,v)$ be a three-stage linear graph with the set of switches $\{u,v\} \cup \{s(2,i): i = 1, \ldots, mn\}$ and $\ell(1,1,i) = \ell(2,i,1) = 1$ for $1 \leq i \leq m$ (see Fig. 6a). Let $G(u',v')$ be a three-stage linear graph with the set of switches $\{u,v,s(2,1)\}$ and satisfying $\ell(1,1,1) = m$, $\ell(2,1,1) = n$, (see Fig. 6b). Then we have

$$P(u,v) \leq P(u',v').$$

Proof: We let $G''(u'',v'')$ have the set of switches $\{u'',v''\} \cup \{s''(2,i): 1 \leq i \leq m\}$ and satisfying $\ell''(1,1,i) = 1$ for $1 \leq i \leq m$, $\ell''(2,i,1) = n$ for $1 \leq i \leq m$ (see Fig. 7).

By using Theorem 2 (repeatedly), we have

$$P(u'',v'') \leq P(u',v').$$

Similarly, we have

$$P(u,v) \leq P(u'',v'').$$

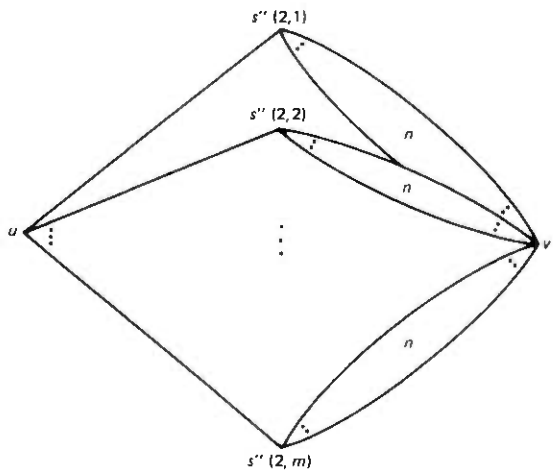


Fig. 7—An intermediate linear graph.

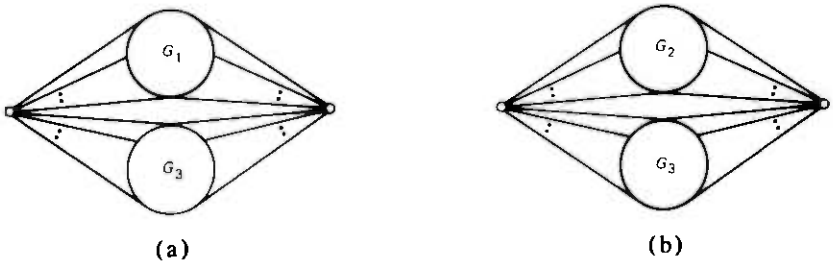


Fig. 8—Parallel combinations for n-stage linear graphs.

Thus, we have

$$P(u,v) \leq P(u',v')$$

and the corollary is proved.

III. LINEAR GRAPHS IN MULTISTAGE NETWORKS

In Section II, we presented several methods to compare blocking probabilities of small linear graphs. In fact, large linear graphs can be compared in very much the same way. The following two theorems show how to extend these methods to multistage linear graphs with a comparatively large set of switches.

Theorem 3: Let $G_1(u_1,v_1)$, $G_2(u_2,v_2)$, $G_3(u_3,v_3)$ be three n-stage linear graphs. We suppose the blocking probability $P(u_1,v_1)$ is smaller than or equal to the blocking probability $P(u_2,v_2)$. Let $G(u,v)$ be an n-stage linear graph obtained by a parallel combination of $G_1(u_1,v_1)$ and $G_3(u_3,v_3)$ (see Fig. 8a). Let $G'(u',v')$ be an n-stage linear graph obtained by a parallel combination of $G_2(u_2,v_2)$ and $G_3(u_3,v_3)$ (see Fig. 8b). Then we have

$$P(u,v) \leq P(u',v').$$

Similarly, if $\overline{G}(u_*,v_*)$ is a $(2n-1)$-stage linear graph obtained by a series combination of $G_1(u_1,v_1)$ and $G_3(u_3,v_3)$ and $\overline{G}'(u'_*,v'_*)$ is a $(2n-1)$-stage linear graph obtained by a series combination of $G_2(u_2,v_2)$ and $G_3(u_3,v_3)$, then we have

$$P(u_*,v_*) \leq P(u'_*,v'_*).$$

Proof: It is easy to see that

$$P(u,v) = P(u_1,v_1)P(u_3,v_3)$$

$$P(u_*,v_*) = 1 - [1 - P(u_1,v_1)]\,[1 - P(u_3,v_3)],$$

and

$$P(u',v') = P(u_2,v_2)P(u_3,v_3)$$

$$P(u'_*,v'_*) = 1 - [1 - P(u_2,v_2)]\,[1 - P(u_3,v_3)].$$

Thus we have

$$P(u,v) \le P(u',v'),\ P(u_*,v_*) \le P(u'_*,v'_*).$$

The following theorem is a generalized version of Theorem 2. Theorem 1 and Corollary 1 can be generalized similarly but will not be stated here.

Theorem 4: Let $G(u,v)$ be an n-stage linear graph with the set of switches

$$\bigcup_{i=1}^{n} \{s(i,j): 1 \le j \le m_i\}$$

and the set of links $\{L(i,j,k)\}$. Let $G'(u',v')$ be an n-stage linear graph with the set of switches

$$\bigcup_{i=1}^{n} \{s'(i,j): 1 \le j \le m'_j\}$$

and the set of links $\{L'(i,j,k)\}$.

Suppose $G(u,v)$ and $G'(u',v')$ satisfy the following conditions.

(*i*) $m_i = m'_i$ for any $i \ne w$, $1 \le i \le n$ (for a fixed w).

(*ii*) There exist k_1,k_2,k_3 such that the linear graph $G(u,v) - \{s(w,k_1),s(w,k_2)\}$ is isomorphic to the linear graph $G'(u',v') - \{s'(w,k_3)\}$.

(*iii*) $s(w,k_1)$, $s(w,k_2)$ and $s'(w,k_3)$ are connected to other switches so that the following conditions hold:

$$\ell(w-1,j,k_1) = \ell(w-1,j,k_2) = \ell'(w-1,j,k_3) \text{ for } 1 \le j \le m_{w-1},$$

$$\ell(w,k_1,k) + \ell(w,k_2,k) = \ell'(w,k_2,k) \text{ for } 1 < k < m_{w+1}.$$

where $\ell(i,j,k)$, $\ell'(i,j,k)$ denote the cardinalities of $L(i,j,k)$, $L'(i,j,k)$, respectively.

Then we have

$$P(u,v) \le P(u',v').$$

We note that (*iii*) could be replaced by (*iii'*) because of symmetry:

(*iii'*) $\ell(w-1,j,k_1) + \ell(w-1,j,k_2) = \ell'(w-1,j,k_3)$ for $1 \le j \le m_{w-1}$, $\ell(w,k_1,k) = \ell(w,k_2,k) = \ell'(w,k_3,k)$ for $1 \le k \le m_{w+1}$.

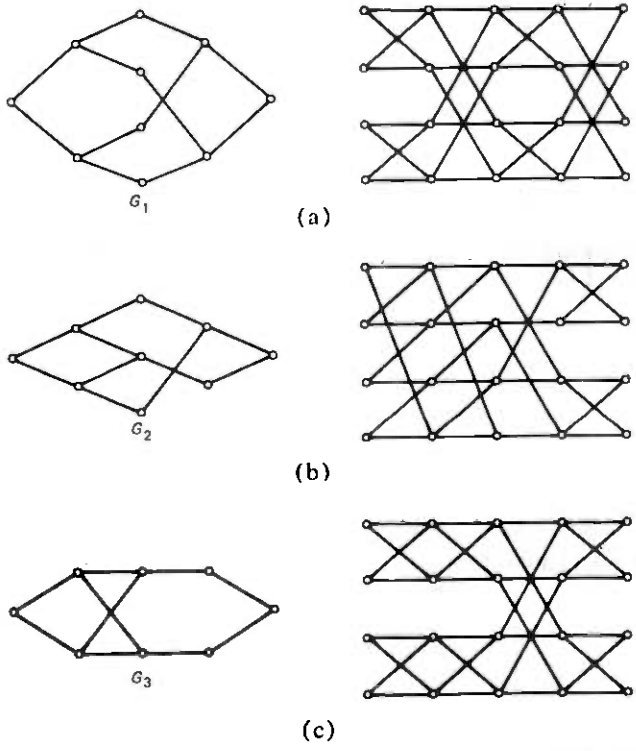


Fig. 9—Examples of linear graphs and corresponding balanced switching networks.

Proof: We may assume $n \geq 4$ because of Theorem 2. Thus, we may assume without loss of generality that $w \neq n-1$. Therefore $m_i = m_i'$ for $i \neq w$, and in particular, $m_{n-1} = m_{n-1}'$. Let A be a subset of $\{j: 1 \leq j \leq m_{n-1}\}$. Let $G_A(u,v_A)$ be an $(n-1)$-stage linear graph which can be viewed as the union of all paths in G which connect u and a switch $s(n-1,j)$, where $j \in A$ and all switches in A have been identified. (It can be viewed that all switches in A are condensed into one switch.) In other words, $G(u,v_A)$ has the set of switches $\{v_A = s_A\ (n-1,1)\} \cup \{s_A(i,j): i \neq n-1$ and $s(i,j)$ is on a path which passes through a switch $s(n-1,j)$ where $j \in A\}$. G_A has the set of links $\{L_A(i,j,k)\}$ where

$$\ell_A(n-2,j,1) = \sum_{k \in A} \ell(n-2,j,k)$$

and $\ell_A(i,j,k) = \ell(i,j,k)$ for $i \neq n-2$. Let $G_A'(u',v_A')$ be the linear graph similarly obtained from G' by identifying all switches in A. By the induction assumption, we have

$$P(u,v_A) \leq P(u',v_A').$$

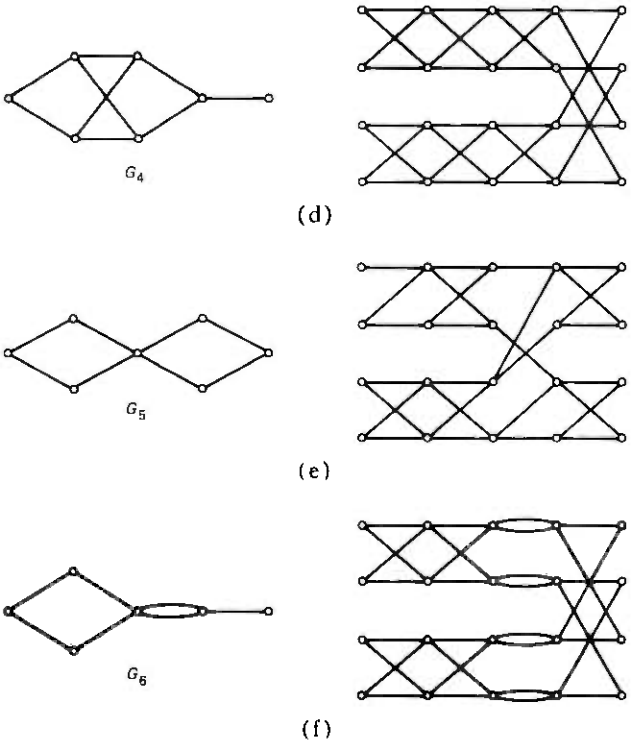


Fig. 9 (*continued*)

Moreover, $P(u,v)$ can be written as follows:

$$P(u,v) = \sum_A p_{n-1}^{|A|} (1 - p_{n-1})^{m_{n-1}-|A|} P(u,v_A)$$

where A ranges over all subsets of $\{j: 1 \le j \le m_{n-1}\}$.

Since $P(u',v')$ has the similar expression

$$P(u',v') = \sum_A p_{n-1}^{|A|} (1 - p_{n-1})^{m_{n-1}-|A|} P(u',v'_A),$$

then we have

$$P(u,v) \le P(u',v')$$

In Ref. 2, the present authors consider a special class of linear graphs $G(u,v)$ with $m_{n-i} = m_i$, n odd and m_i dividing m_{i+1} for $i = 1,2,\ldots,[n/2]$. It can be easily seen that the linear graphs in the class can be compared by using Theorem 4.

In Fig. 9a to f, we give several examples of linear graphs together with their corresponding balanced switching networks.

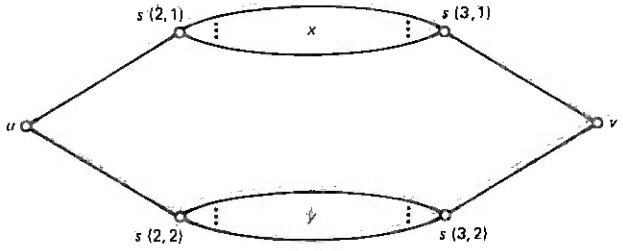


Fig. 10—A four-stage linear graph.

Let P_{G_i} denote the blocking probability for the balanced network N_i with linear graph G_i. It is easy to verify that $P_{G_1} \leq P_{G_2}$ by taking $w = 3, k_1 = 1, k_2 = 3, k_3 = 2$. Similarly, it is easy to see that

$$P_{G_1} \leq P_{G_2} \leq P_{G_3} \leq P_{G_4} \leq P_{G_6},$$

and

$$P_{G_3} \leq P_{G_5} \leq P_{G_6}.$$

We note that the numbers of crosspoints in N_i, $i = 1, \ldots, 6$, are the same. Thus we know that the switching network N_1 is "better" than the switching network N_2 and so forth.

IV. SERIES-PARALLEL LINEAR GRAPHS

In this section, we consider series-parallel linear graphs. Series-parallel linear graphs are sometimes preferred to spider-web linear graphs[6] because of the conditions for implementation and control. The following two theorems treat the blocking probabilities of series-parallel linear graphs.

Theorem 5: We consider the following four-stage linear graph $G_{x,y}$ (see Fig. 10).

(*i*) $m_2 = m_3 = 2$

(*ii*) $\ell(1,1,1) = \ell(1,1,2), \ell(2,1,2) = \ell(2,2,1) = 0, \ell(3,1,1) = \ell(3,2,1)$,

(*iii*) $\ell(2,1,1) = x, \ell(2,2,2) = y$.

If there are integers a and b with $x + y = a + b, x \leq a \leq b \leq y$, then we have

$$P_{G_{ab}} \leq P_{G_{xy}}$$

The proof of Theorem 5 is quite similar to the proof of Theorem 1—by setting $f(x) = [1 - (1 - p_1)(1 - p_2^x)(1 - p_3)][1 - (1 - p_1)(1 - p_2^{\varepsilon - x})(1 - p_3)]$—and is omitted.

Remark: The above theorem can be extended to multistage linear graphs by replacing each link by a linear graph under the condition that all links

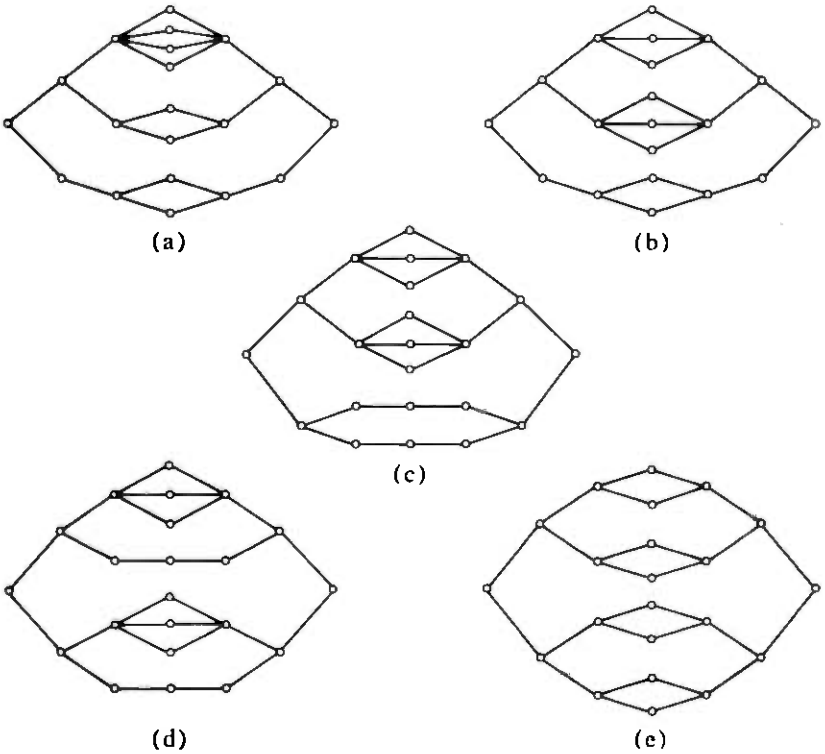


Fig. 11—Examples.

between stage i and stage $i + 1$ are replaced by copies of a linear graph or by linear graphs with the same blocking probabilities.

In Fig. 11, some examples are illustrated. The linear graph in Fig. 11b has a smaller blocking probability than the linear graph in Fig. 11a by Theorems 3 and 5. The linear graph in Fig. 11c has a smaller blocking probability than the linear graph in Fig. 11b by Theorems 3 and 4.

Theorem 6: We consider the following linear graph G_{xyzw} (see Fig. 12):

(*i*) $m_i = m_j = 2$.

(*ii*) u and $s(i,1)$ are connected by a linear graph N_1. u and $s(i,2)$ are connected by a linear graph N_2. N_1 and N_2 have the same number of stages and $P_{N_1} = P_{N_2}$.

(*iii*) v and $s(j,1)$ are connected by a linear graph N_3. v and $s(j,2)$ are connected by a linear graph N_4. N_3 and N_4 have the same number of stages and $P_{N_3} = P_{N_4}$.

(*iv*) There exist $(j - i + 1)$-stage linear graphs G_1, G_2 such that $s(i,1)$

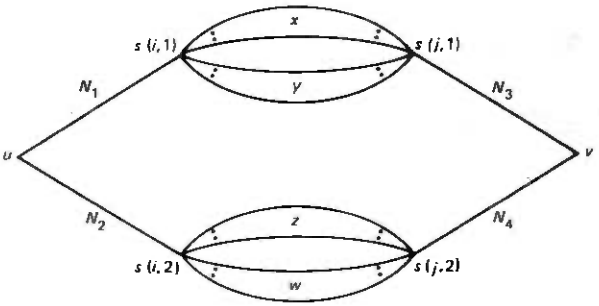


Fig. 12—Linear graph for Theorem 6.

and $s(j,1)$ are connected by x copies of G_1 and y copies of G_2 and $s(i,2)$ and $s(j,2)$ are connected by z copies of G_1 and w copies of G_2.

Suppose $x + y = z + w = c$ and $x + z = d$ for some constants c and d. We also suppose $x' + y' = z' + w' = c, x' + z' = d$ where $x' \leq x \leq z \leq z'$. Then we have

$$P_{G_{xyzw}} \leq P_{G_{x'y'z'w'}}.$$

Proof: Let $\alpha = (1 - P_{N_1})(1 - P_{N_3})$.

Define the following function $f(x)$:

$$f(x) = [1 - \alpha(1 - P_{G_1}^{x} P_{G_2}^{c-x})][1 - \alpha(1 - P_{G_1}^{d-x} P_{G_2}^{c-d+x})].$$

It is easy to see that $P_{G_{xyzw}} = f(x)$, $P_{G_{x'y'z'w'}} = f(x')$. Now,

$$\frac{df}{dx}(x) = \alpha(1-\alpha)(\log P_{G_1} - \log P_{G_2})(P_{G_1}^{x} P_{G_2}^{c-x} - P_{G_1}^{d-x} P_{G_2}^{c-d+x})$$

$$= \alpha(1-\alpha)(\log P_{G_1} - \log P_{G_2}) P_{G_1}^{x} P_{G_2}^{c-x}(1 - P_{G_1}^{d-2x} P_{G_2}^{2x-d}).$$

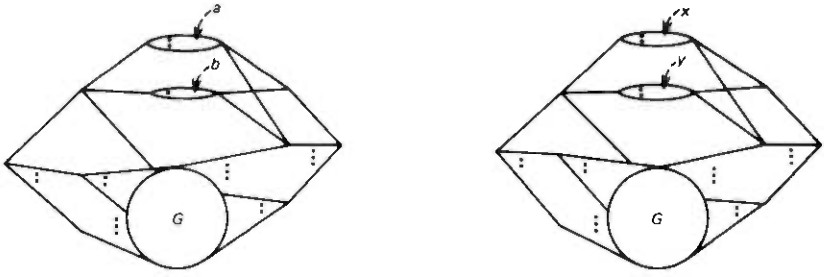
If $P_{G_1} = P_{G_2}$, we have $f(x) = f(x')$. If $P_{G_1} \neq P_{G_2}$, $f(x)$ attains its minimum at $x = d/2$. Since $f(x)$ is convex, then

$$f(x) \leq f(x') \text{ for } x' \leq x \leq \frac{d}{2}$$

Thus we have

$$P_{G_{xyzw}} \leq P_{G_{x'y'z'w'}}.$$

Theorem 5 and Theorem 6 essentially say that the more regular (i.e., evenly distributed) the linear graph, the better it is. Of course, all these results are based on the Lee model and the related independence assumption. In some existing networks, irregular linear graphs might sometimes be desirable because of the preference schemes in routing.[1]



(a) (b)

Fig. 13—Linear graphs for Theorem 7.

In Fig. 11, the linear graph in 11d has a smaller blocking probability than the linear graph in 11c by Theorem 6. By Theorem 5, the linear graph in 11e has the smallest blocking probability. We note that 11e is the most regular linear graph in Fig. 11.

Theorems 5 and 6 can be generalized to a class of spider-web linear graphs. We will state the generalized version of Theorem 5.

Theorem 7: Let $\overline{G}_{ab}$ and $\overline{G}_{xy}$ be two n-stage linear graphs satisfying the following properties (see Fig. 13).

(*i*) There exists k, $2 \leq k \leq n-2$, such that $\overline{G}_{ab} - \{s(k,1),s(k,2),s(k+1,1)s(k+2,2)\}$ is isomorphic to $\overline{G}_{xy} - \{s'(k,1),s'(k,2),s'(k+1,1),s'(k+2,2)\}$, where $\{s(i,j)\}$, $\{s'(i,j)\}$ are the sets of switches of $\overline{G}_{ab}, \overline{G}_{xy}$, respectively.

(*ii*) $\ell(k-1,i,1) = \ell(k-1,i,2) = \ell'(k-1,i,1) = \ell'(k-1,i,2)$ for $1 \leq i \leq m_{k-1}$, and $\ell(k+1,1,j) = \ell(k+1,2,j) = \ell'(k+1,1,j) = \ell'(k+1,2,j)$ for $1 \leq j \leq m_{k+1}$, where $\ell(i,j,k)$ and $\ell'(i,j,k)$ are the cardinalities of links of $\overline{G}_{ab}$, $\overline{G}_{xy}$, respectively.

(*iii*)

$$\ell(k,1,j) = \begin{cases} a & \text{if } j = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\ell(k,2,j) = \begin{cases} b & \text{if } j = 2 \\ 0 & \text{otherwise} \end{cases}$$

Similarly,

$$\ell'(k,1,j) = \begin{cases} x & \text{if } j = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\ell'(k,2,j) = \begin{cases} y & \text{if } j = 2 \\ 0 & \text{otherwise} \end{cases}$$

(*iv*)

$$x + y = a + b,\ x \leq a \leq b \leq y.$$

Then we have

$$P_{\overline{G}_{ab}} \leq P_{\overline{G}_{xy}}.$$

Proof: The proof is by induction on the number of stages. Suppose $n = 4$. Following the notation in Theorem 5, we note that $\overline{G}_{xy}$ is the parallel combination of G_{xy} and G. Thus by Theorem 5, we have

$$P_{\overline{G}_{ab}} = P_{G_{ab}} P_G \leq P_{G_{xy}} P_G = P_{\overline{G}_{xy}}.$$

For $n > 4$, we apply the same reduction scheme which is used in the proof of Theorem 4. The theorem is then proved by mathematical induction.

V. CONCLUDING REMARKS

Lee[8] first proposed the concept of a linear graph in connection with his study of the blocking probabilities of switching networks. Since then his model has been widely used. However, a systematic study of linear graphs is still far from complete. There are some results in extending Lee's method[5,9] or for studying the blocking probabilities for certain classes of series-parallel linear graphs[2]. Takagi[10,11] has defined a class of spider-web linear graphs and finds the optimal one in that class. Some of his results have been obtained earlier by Le Gall[3]. Van Bosse[12,13] extends results in Refs. 3, 10, and 11 in the sense that the occupancy distribution for links can be arbitrary. In this paper, several new methods for analyzing blocking probabilities of certain classes of switching networks are presented. We hope it will lead to more research in this direction.

REFERENCES

1. D. Bazlen, G. Kampe, and A. Lotze, "On the influence of hunting mode and link wiring on the loss of link systems, 7th ITC, Stockholm, 232/1–12 (1973).
2. F. R. K. Chung and F. K. Hwang, "A problem of blocking probabilities in connecting network," Networks, 7 (1977), pp. 185–192.
3. P. Le Gall, "Étude du blocage dans les systèmes de commutation téléphonique, Ann. Télécommun., *11*, No. 9 (1956).
4. F. K. Hwang, "Balanced networks," 1976 International Conference on Communications, Vol. 1, 7–13 to 7–16.
5. F. K. Hwang and A. M. Odlyzko, "A Probability Inequality and Its Application to Switching Networks," B.S.T.J., *56*, No. 5 (May-June 1977), pp. 821–826.
6. J. G. Kappel, personal communication.
7. R. S. Krupp, "Analysis of Toll Switching Networks," B.S.T.J., *55*, (1976), 843–856.
8. C. Y. Lee, "Analysis of Switching Networks," B.S.T.J., *34* (1955) 1287–1315.
9. N. J. Pippenger, "The complexity theory of switching networks," MIT Ph.D. thesis, 1973.
10. K. Takagi, "Design of multi-stage link systems by means of optimum channel graphs," Electronics and Communications in Japan, *51A*, No. 4 (1968), 37–46.
11. K. Takagi, "Optimum channel graph of link system," Electronics and Communications in Japan, *54A*, No. 8 (1971), 1–10.
12. J. G. Van Bosse, "On an inequality for the congestions in switching networks," IEEE Trans. on Communications, *COM-22* (1974), 1675–1677.
13. J. G. Van Bosse, "A generalization of Takagi's results on optimum link graphs," 8th International Telegraphic Congress (1976), 513-1 to 513-7.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 8, October 1977
Printed in U.S.A.

Coupled Surface-Acoustic-Wave Resonators

By P. S. CROSS and R. V. SCHMIDT

(Manuscript submitted March 18, 1977)

Coupled Surface-Acoustic-Wave (SAW) grating resonators are investigated analytically with a transmission-matrix technique, and the measured frequency responses at ~145 MHz of devices on YZ-$LiNbO_3$ with Ti-diffused gratings are compared with the theoretical results. Coupled-mode theory is applied to derive the two-by-two transmission matrix relating the acoustic wave amplitudes at the input and output of a surface wave grating. Using the transmission matrices, the external transmission through a SAW resonator is found by matrix multiplication. Some fundamental aspects of resonator passband synthesis are introduced by considering the transmission through several acoustically cascaded resonators. Resonator filters where the transducers couple directly to the resonant cavities are treated by developing a description of the transducer that is compatible with the transmission matrix of the grating. The analysis technique is then applied to the familiar two-port resonator-filter. Next, coupled resonator-pairs with a transducer in each cavity are considered in detail for: (i) collinear acoustic coupling, (ii) multistrip coupling, and (iii) transducer coupling. Experimental results are presented for each configuration considered and good agreement with the analytical description is found in each case.

I. INTRODUCTION

Surface-acoustic-wave resonators are now well established as one-pole, narrowband filters in the frequency range 30 to 1000 MHz.[1,2] Recent work[3–10] has shown that multipole filters can be formed by coupling several resonators. In general, multipole filter responses can be synthesized by using one or more of the three established coupling mechanisms: (*i*) collinear acoustic coupling,[3–5] (*ii*) acoustic directional coupling (multistrip coupler),[6,7] or (*iii*) electrical coupling using transducers.[8–10]

Examples of two-pole resonator filters using the three types of cavity-coupling mechanisms are presented in Fig. 1. In each configuration,

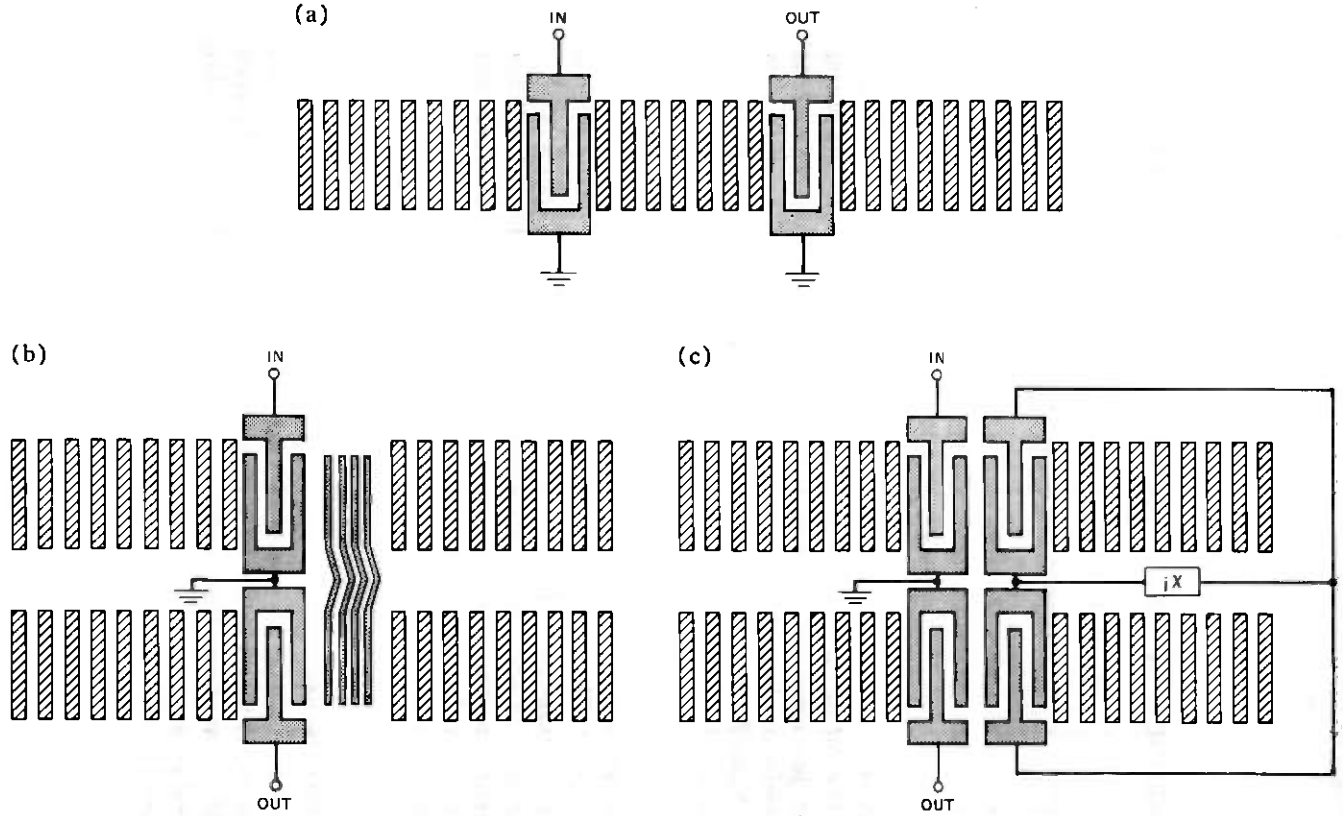


Fig. 1—Diagram of two-pole, coupled resonator filters using (a) collinear acoustic coupling, (b) multistrip coupling, and (c) transducer coupling with an external shunt susceptance.

there are two resonant cavities with a transducer in each cavity for coupling to the external circuitry.

In the collinear cascade structure the central grating, common to both cavities, is the coupling element. The strength of the central grating determines how much power can "leak" from one cavity to the other.

When either multistrip or transducer coupling is employed, each cavity has a distinct set of gratings, and the resonators are conveniently arranged in parallel with the acoustic power flowing in two separate "tracks." Coupling with the multistrip is effected by simply extending the electrodes of the coupler into both cavities. The degree of multistrip coupling is determined by the length or, equivalently, the number of electrodes of the coupler.

In the transducer-coupling configuration, a second transducer is placed in each cavity. The cavities are then coupled by connecting the transducers together, either directly or through an external electrical network. The external network provides a means for adjusting both the strength and phase of the cavity coupling.

In order to design a filter using coupled grating resonators it is necessary to be able to relate the frequency response of the filter to the parameters describing the gratings, transducers, and coupling elements. We present here a general technique for obtaining the frequency response of coupled resonators. In addition, the technique yields closed-form expressions for the insertion loss, out-of-band response and the near-in-band shape which aid in filter design.

The approach taken in this paper is to first develop the transmission matrix of a uniform grating and use it to analyze the external transmission response of a single resonator. Next, the properties of coupled resonators are introduced by studying the external transmission response of acoustically cascaded resonators.

We then present a description of the interdigital transducer which is compatible with the transmission matrix description of the gratings. With this description one can calculate the transmission response of any resonator structure which includes internal transducers.

The technique is applied to the familiar two-port resonator-filter. Then coupled-resonator pairs are treated in detail for each of the three cavity-coupling mechanisms. Experimental results at ~145 MHz are presented for each configuration considered, and the good agreement with theory that is found in each case substantiates the analytical models.

II. GRATING TRANSMISSION MATRIX

In this section, the transmission matrix of a surface-wave grating is derived. A transmission matrix relates the forward and backward traveling-wave-amplitudes at the left side of an element to those on the right

side. It is therefore useful to establish a compact notation by introducing the vector

$$\mathbf{W}_i = \begin{bmatrix} w_i^+ \\ w_i^- \end{bmatrix} \tag{1}$$

which represents the complex amplitudes of the forward-, w_i^+, and backward-, w_i^-, traveling waves at the right-hand reference plane of the ith element of a filter structure. The amplitudes have dimensions of $\sqrt{\text{Power}}$. Thus, the transmission characteristics of the ith element of the structure are described by the matrix equation

$$\mathbf{W}_{i-1} = \mathcal{M}_i \mathbf{W}_i \tag{2}$$

where $\mathcal{M}_i$ is the 2×2 transmission matrix of the ith element.

The transmission matrix of a grating is derived using a plane-wave, coupled-mode analysis which was originally applied to thick holograms[11] and subsequently to distributed feedback lasers[12,13] and acoustic grating reflectors.[3] The grating to be analyzed is taken to have constant period Λ, and to extend from $x = -L$ to $x = 0$. Near the Bragg frequency, only the fundamental Fourier component of the grating perturbation provides phase-matching between the forward- and backward-traveling waves. Thus, in the analysis, a lossless grating is mathematically modeled by a sinusoidal velocity perturbation given by

$$v(x) = v_0 - \frac{\Delta v}{2} \cos(Kx) \tag{3}$$

where $K = 2\pi/\Lambda$. Furthermore, we assume that the surface wave propagation can be represented by the scalar wave equation

$$\frac{d^2\Psi}{dx^2} + \frac{\omega^2}{v^2(x)}\Psi = 0 \tag{4}$$

where ω is the surface-wave radian frequency. The scalar Ψ represents the quasistatic electric potential at the surface of the piezoelectric crystal associated with the surface wave. The general solution[11] of eq. (4) is

$$\Psi(x) = w^+(x) + w^-(x) \tag{5a}$$

where

$$w^+(x) = \psi^+(x)e^{-j\beta_0 x} \tag{5b}$$

$$w^-(x) = \psi^-(x)e^{+j\beta_0 x} \tag{5c}$$

are respectively the forward and backward wave amplitudes in the grating and $\beta_0 = \pi/\Lambda$ is the propagation constant of the surface wave at the Bragg frequency $\omega_0 = \pi v_0/\Lambda$. By appropriately combining eqs. (3) through (5) and dropping higher harmonic terms one obtains the coupled

wave equations

$$-\frac{d\psi^+}{dx} - j\delta\psi^+ = j\frac{\beta}{\beta_0}\kappa\psi^- \tag{6a}$$

$$\frac{d\psi^-}{dx} - j\delta\psi^- = j\frac{\beta}{\beta_0}\kappa\psi^+ \tag{6b}$$

where $\beta = \omega/v_0$, $\kappa = (\beta/4)\cdot\Delta v/v_0$ is the grating coupling coefficient and

$$\delta = \frac{\beta^2 - \beta_0^2}{2\beta_0} \tag{7}$$

is a measure of the frequency deviation from the Bragg frequency. For high-Q resonators, we are particularly interested in a limited frequency range such that $\beta/\beta_0 \approx 1$ and the coupled wave equations can be simplified by setting $\beta/\beta_0 = 1$ and letting $\delta = (\omega - \omega_0)/v_0$. In the remainder of this paper we use this narrowband approximation. The exact forms of (6) must be used if responses over large bandwidths are required.

Solving eqs. (3) through (7) for the wave amplitudes at $x = -L$ in terms of the wave amplitudes at $x = 0$ yields the following transmission relation

$$\mathbf{W}(-L) = \mathcal{G}\mathbf{W}(0) \tag{8a}$$

where the transmission matrix $\mathcal{G}$ for a grating an integral number of periods long is given by

$$\mathcal{G} = (-1)^{N_g}\frac{\cosh(\sigma L)}{\sqrt{1-\Delta^2}} \times \begin{bmatrix} \sqrt{1-\Delta^2} + j\Delta\tanh(\sigma L) & j\tanh(\sigma L) \\ -j\tanh(\sigma L) & \sqrt{1-\Delta^2} - j\Delta\tanh(\sigma L) \end{bmatrix} \tag{8b}$$

where

$$N_g = L/\Lambda, \tag{8c}$$

$$\sigma = \sqrt{\kappa^2 - \delta^2} = \kappa\sqrt{1-\Delta^2} \tag{8d}$$

and

$$\Delta = \delta/\kappa \tag{8e}$$

is the normalized frequency deviation.

The reflection coefficient, Γ, at the plane $x = -L$ for a wave incident from the left is

$$\Gamma(\Delta) = \frac{w^-(-L)}{w^+(-L)} = \frac{-j}{\sqrt{1-\Delta^2}\coth(\sigma L) + j\Delta} \tag{9}$$

and at the Bragg frequency, $\Delta = 0$,

$$\Gamma(0) = -j\rho \tag{10}$$

where $\rho = \tanh(\kappa L)$.

The grating transmission matrix and reflection coefficient have been derived by postulating a sinusoidal velocity perturbation grating. The final expressions are, however, in terms of a coupling coefficient, κ, which describes the strength of the perturbation that forms the grating. By appropriately identifying the coupling coefficient of other grating types (such as surface corrugations), the grating transmission matrix (8b) describes the behavior of surface-wave gratings formed with any perturbation mechanism.

Equations (9) and (10) provide a means for experimentally determining the coupling coefficient for a particular physical grating. It has been found[14] that κ can be obtained by either measuring the reflectivity at $\delta = 0$ and using (10) or by measuring the fractional bandwidth $\Delta\omega/\omega_0$ between reflection zeros and calculating κ from the expression

$$\kappa = \frac{\pi}{2\Lambda}\sqrt{\left(\frac{\Delta\omega}{\omega_0}\right)^2 - \left(\frac{2\Lambda}{L}\right)^2} \tag{11}$$

obtained from (9). The first method is most suitable for weakly reflecting gratings while the second method works best on highly reflective gratings.

For the specific case of shallow-groove gratings, one can use, in addition to the above techniques, the results of Li *et al.*[15,16] to determine the coupling coefficient which gives $\kappa = h/3\Lambda^2$ for corrugations of depth h. The various second-order effects associated with stored reactive energy have been neglected here for simplicity.

The phase of the reflection coefficient and the off-diagonal terms of the transmission coefficient depend on the choice of grating reference planes. In Appendix A, the question of specifying reference planes is treated in detail, and it is shown that reference planes can be found for any grating such that the transmission matrix in (8b) is applicable.

If the ith element of the structure is a transmission line extending from $x = -L_i$ to $x = 0$, it is described by the familiar transmission equation

$$\mathbf{W}_{i-1} = \Phi_i \mathbf{W}_i \tag{12}$$

where

$$\Phi_i = \begin{bmatrix} e^{j\beta L_i} & 0 \\ 0 & e^{-j\beta L_i} \end{bmatrix}. \tag{13}$$

Thus far, the surface-wave gratings have been treated as lossless. However, in many circumstances, small grating losses have a significant

influence on the grating filter transmission response. In Appendix A, the transmission matrix of a lossy grating with a distributed attenuation coefficient, α, is given in eq. (82). This matrix is unnecessarily complicated when only frequencies near the Bragg frequency, $|\Delta| \ll 1$, are considered. An approximate transmission matrix for a lossy grating can be considered when $\alpha/\kappa \ll 1$ and $\Delta \ll 1$ by decomposing the lossy matrix (82) at $\Delta = 0$ as follows:

$$\mathcal{G} \simeq \mathcal{A}\mathcal{T}\mathcal{A} \tag{14a}$$

where

$$\mathcal{A} = \begin{bmatrix} \exp\left(\frac{\alpha\rho}{2\kappa}\right) & 0 \\ 0 & \exp\left(-\frac{\alpha\rho}{2\kappa}\right) \end{bmatrix} \tag{14b}$$

$$\mathcal{T} = \left(1 + \frac{\Delta^2}{2}\right)\cosh(\kappa L)\begin{bmatrix} 1 - \frac{\Delta^2}{2} + j\Delta\rho & j\rho \\ -j\rho & 1 - \frac{\Delta^2}{2} - j\Delta\rho \end{bmatrix} \tag{14c}$$

where, as before $\rho = \tanh(\kappa L)$ and $\Delta = \delta/\kappa$. This decomposition is equivalent to placing a lumped, frequency-independent loss[17] at each side of the grating. The matrix $\mathcal{T}$ is the lossless grating transmission matrix (8b) simplified for the condition $|\Delta| \ll 1$ and N_g even. The decomposition of $\mathcal{G}$ in (14) has two advantages. First, other loss mechanisms (such as bulk radiation loss) that are localized in nature can be mathematically included as a component of α. And, second, the important frequency dependence of $\mathcal{G}$ is all contained in $\mathcal{T}$ so that the simplified matrix $\mathcal{T}$ can be used to obtain closed-form expressions for the resonant passband shape of a given structure.

III. TRANSMISSION RESPONSE OF CASCADED GRATING STRUCTURES

The transmission matrices derived in Section II provide the means to calculate the properties of cascaded structures of gratings and transmission lines. As an example of the application of the transmission matrices, we first consider a grating resonator as illustrated in Fig. 2a. The resonator consists of two identical gratings each of length L, which are separated by a quarter-wave transmission line. The wave amplitudes $\mathbf{W}_0$ and $\mathbf{W}_3$, at the left and right reference planes respectively, are related by the matrix equation

$$\mathbf{W}_0 = \mathcal{G}_1\Phi_2\mathcal{G}_3\mathbf{W}_3 \equiv \mathcal{R}\mathbf{W}_3 \tag{15}$$

where $\mathcal{G}_1 = \mathcal{G}_3$ are the transmission matrices of the first and third elements (gratings) and Φ_2 is the transmission matrix of the second element

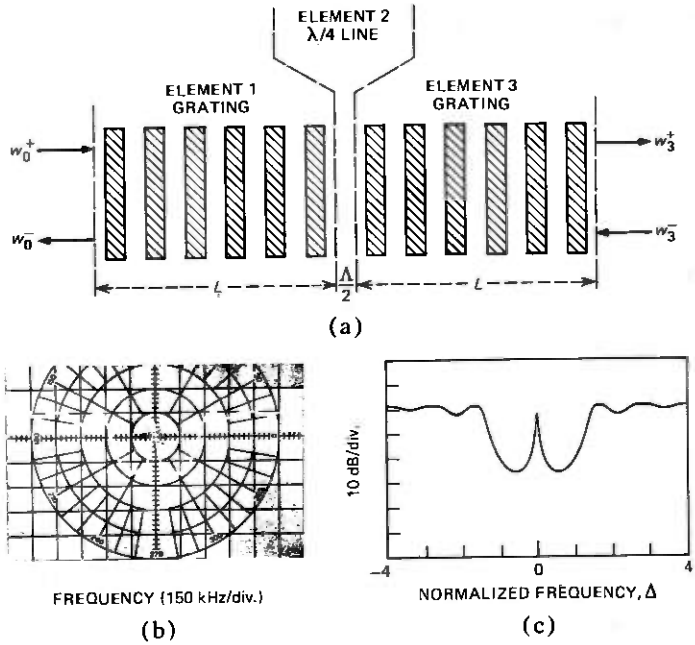


Fig. 2—(a) Diagram of a grating resonator in the external transmission configuration. (b) The transmission spectrum at ~145.5 MHz for a resonator on YZ-$LiNbO_3$ using Ti-diffused gratings with $\Lambda = 12$ μm and L = 6.48 mm. (c) The calculated transmission spectrum for the device in (b) using $\kappa = 3.74\ cm^{-1}$ and $\alpha = 0.036\ cm^{-1}$.

(in this case a quarter-wave line). The matrix $\mathcal{R}$ is the transmission matrix of the resonator.

In the laboratory, the external power transmission, $|w_3^+/w_0^+|^2$, through the structure of Fig. 2a is the most conveniently measured quantity. A typical experimental transmission spectrum for a resonator formed by Ti-diffused gratings[14] in YZ-$LiNbO_3$ is shown in Fig. 2b. Far off resonance, where the gratings are transparent, the transmission is near unity. Inside the grating stopband, the gratings are highly reflective and there is a deep transmission minimum. Near resonance, there is once again near-unity transmission.

The theoretical transmission response can be obtained by applying the boundary condition $w_3^- = 0$ to eq. (15). The external power transmission is then

$$\left|\frac{w_3^+}{w_0^+}\right|^2 = \frac{1}{R_{11}R_{11}^*} \tag{16}$$

where R_{11} is the 11 element of the $\mathcal{R}$ matrix of (15). In Fig. 2c, the calculated spectrum is given for the structure of Fig. 2b where α and κ are

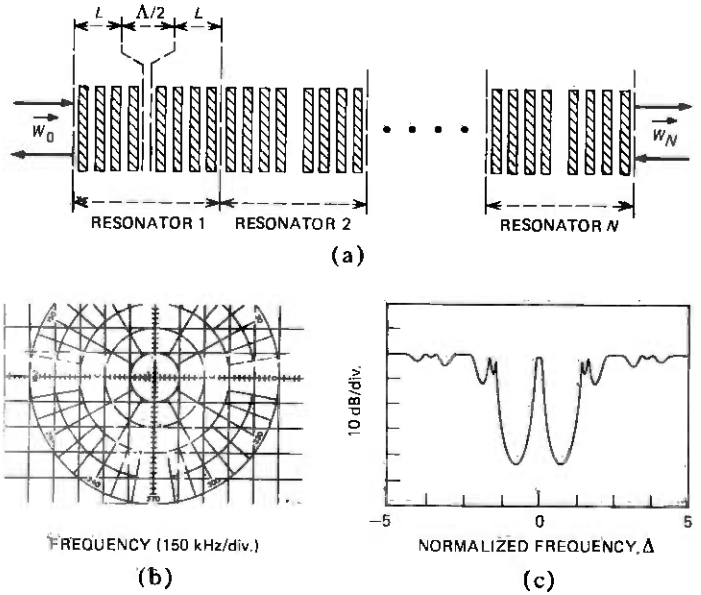


Fig. 3—(a) Diagram of a cascade of N identical resonators. (b) The transmission spectrum at 145.5 MHz of a cascade of three identical resonators with $\Lambda = 12\ \mu m$ and $L = 3.84$ mm. (c) The calculated transmission spectrum for the device in (b) using $\kappa = 3.55\ cm^{-1}$ and $\alpha = 0.027\ cm^{-1}$.

chosen to fit the insertion loss and stopband width. The complete grating transmission matrix of eq. (82) in Appendix A is used in the calculation.

In many cases, only the frequency response near resonance is of interest and the external power transmission can be found using the approximate grating transmission matrix (14a). Under the conditions $|\delta|$, $\alpha \ll \kappa$ and $2 \cosh(\kappa L) \approx \exp(\kappa L)$, eq. (16) simplifies to

$$\left|\frac{w_3^+}{w_0^+}\right|^2 = \frac{1}{1 + \frac{\alpha}{\kappa}\exp(2\kappa L) + \frac{\delta^2 + \alpha^2}{4\kappa^2}\exp(4\kappa L)} \tag{17}$$

From (17), an analytical expression for the unloaded resonator quality factor, Q_u, can be obtained and is given by

$$\frac{1}{Q_u} = \frac{1}{Q_r} + \frac{1}{Q_g} \tag{18}$$

where

$$Q_r = \frac{\pi}{\kappa\Lambda}\sinh^2(\kappa L) \simeq \frac{\pi}{4\kappa\Lambda}\exp(2\kappa L) \tag{19}$$

is the Q associated with radiation loss from the ends of the gratings and

$$Q_g = \frac{\pi}{2\alpha\Lambda} \tag{20}$$

is the Q associated with the distributed internal grating loss (material losses, surface imperfections, and diffraction).

The distributed internal grating loss can be determined from the resonant transmission loss through the resonator. From eq. (17) the resonant transmission is

$$\left|\frac{w_3^+}{w_0^+}\right|^2 = \frac{1}{\left[1 + \frac{1}{2}\frac{\alpha}{\kappa}\exp(2\kappa L)\right]^2} \tag{21}$$

from which α can be determined.

The transmission matrix analysis technique is easily extended to more complicated structures such as those that are encountered in multipole filter-synthesis applications. The simple case of a collinear cascade of identical resonators shown in Fig. 3a provides an illustrative example since such a cascade has been shown to have a near-resonance transmission response described by a Chebyshev polynominal.[3,4,18] The transmission matrix of a cascade of N lossless, identical resonators is given by $(\mathcal{R})^N$ where $\mathcal{R}$ is the transmission matrix of a single resonator. If the lossless grating transmission matrix (8b) is used, the following expression for $\mathcal{R}$ is found:

$$\mathcal{R} = \mathcal{G}\Phi\mathcal{G} = \begin{bmatrix} \frac{2\Delta}{\sqrt{1-\Delta^2}}SC - j\left[1 - 2\frac{\Delta^2}{1-\Delta^2}S^2\right] & -j\frac{2\Delta}{1-\Delta^2}S^2 \\ j\frac{2\Delta}{1-\Delta^2}S^2 & \frac{2\Delta}{\sqrt{1-\Delta^2}}SC + j\left[1 - 2\frac{\Delta^2}{1-\Delta^2}S^2\right] \end{bmatrix} \tag{22}$$

where $S = \sinh(\sigma L)$ and $C = \cosh(\sigma L)$. Equation (22) is applicable over the region of validity of the coupled-mode approximation, $|\Delta| \ll \pi/\kappa\Lambda$.

Using the results of Storch[19] to evaluate $(\mathcal{R})^N$, one can obtain the following expression for the transmission response through the cascade:

$$\left|\frac{w_N^+}{w_0^+}\right|^2 = \frac{1}{\left[1 + 2\frac{\Delta}{1+\Delta^2}S^2U_N(\xi)\right]^2} \tag{23}$$

where

$$\xi = \frac{2\Delta}{\sqrt{1-\Delta^2}} CS$$

and U_N is the Chebyshev polynominal of the second kind of Nth order. Near the resonant frequency the response is simplified to

$$\left|\frac{w_N^+}{w_0^+}\right|^2 \approx \frac{1}{1+\Omega^2 U_N^2(\Omega)} \tag{24}$$

where $\Omega = 2Q_r(\omega - \omega_0)/\omega_0$ and Q_r is the radiation Q of a single resonator. In Fig. 3b the experimental transmission response of a cascade of three coupled resonators is presented, and in Fig. 3c, the theoretical frequency response calculated using the lossy grating matrix (82) is given. The theoretical description again provides an excellent fit to the data.

The comparisons made in this section between the experimental and theoretical transmission spectra of cascaded grating structures provide a quantitative verification of the analytical model and approximations presented in Section II. In particular, over the frequency range used in the measurements ($\Delta\beta/\beta \lesssim 1$ percent), the excellent agreement between the calculated and experimental responses justifies both the use of the coupled mode equations and the narrow-band ($\beta/\beta_0 \simeq 1$) simplification. It should also be noted that the loss coefficient required to theoretically fit the data is only about twice the surface wave propagation loss of $LiNbO_3$. Thus, the titanium diffusion process[14] produces a low-loss surface perturbation that is ideal for high-Q resonators.

IV. INTRACAVITY TRANSDUCERS AND THE TWO-PORT RESONATOR

In the preceding sections, coupled-mode theory has been applied to derive a transmission matrix description of SAW gratings and resonators. The resonators become useful bandpass filters with low out-of-band transmission, when the transducers are placed inside the cavity.[20–22] In Fig. 4 an interdigital transducer (IDT) is depicted schematically and the various physical quantities associated with the IDT are indicated. The quantities $w_i^{\pm}$ and $w_{i-1}^{\pm}$ are the local amplitudes of the various acoustic waves as previously defined, and a_i and b_i are the amplitudes of the electrical waves incident and emanating from the transducer, respectively.

The terminal amplitudes at the transducer can be related by a dimensionless matrix $\mathcal{T}$, such that

$$\begin{pmatrix} w_{i-1}^+ \\ w_{i-1}^- \\ b_i \end{pmatrix} = \mathcal{T} \begin{pmatrix} w_i^+ \\ w_i^- \\ a_i \end{pmatrix} \tag{25}$$

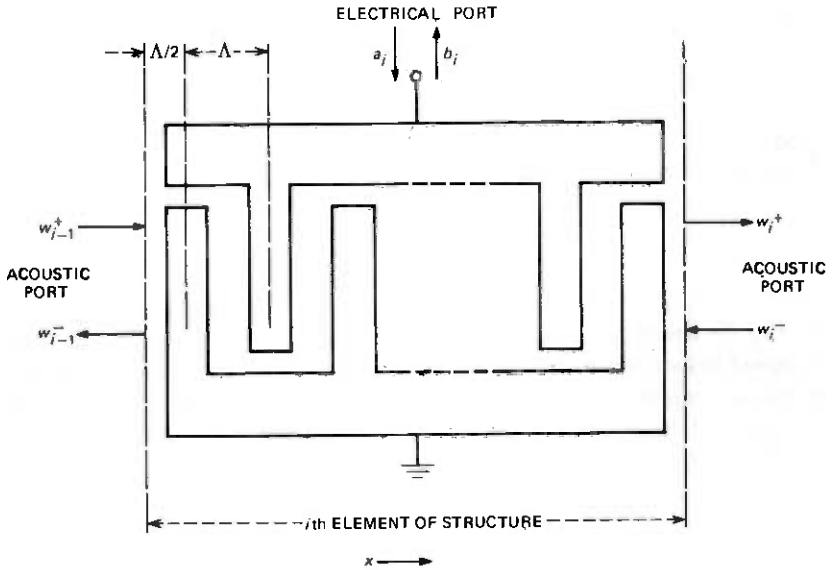


Fig. 4—Diagram of an interdigital transducer.

where $\mathcal{T}$ is given by

$$\mathcal{T} = \begin{pmatrix} t_{11} & t_{12} & t_{13} \\ -t_{12} & t_{22} & t_{23} \\ st_{13} & -st_{23} & t_{33} \end{pmatrix} \tag{26}$$

and s is a symmetry parameter expressing whether the transducer has an even ($s = 1$) or odd ($s = -1$) number of electrodes.

The transducer description of eq (25) has the useful property that the acoustic amplitudes are expressed in transmission matrix form. As a result, (25) is conveniently decomposed into two equations:

(*i*) The acoustic amplitudes at the transducer reference planes are related by

$$\mathbf{W}_{i-1} = t_i \mathbf{W}_i + a_i \tau_i \tag{27}$$

where t_i is the transmission matrix

$$t_i = \begin{pmatrix} t_{11} & t_{12} \\ -t_{12} & t_{22} \end{pmatrix}_i \tag{28}$$

and τ_i is the input coupling vector

$$\tau_i = \begin{pmatrix} t_{13} \\ t_{23} \end{pmatrix}_i \tag{29}$$

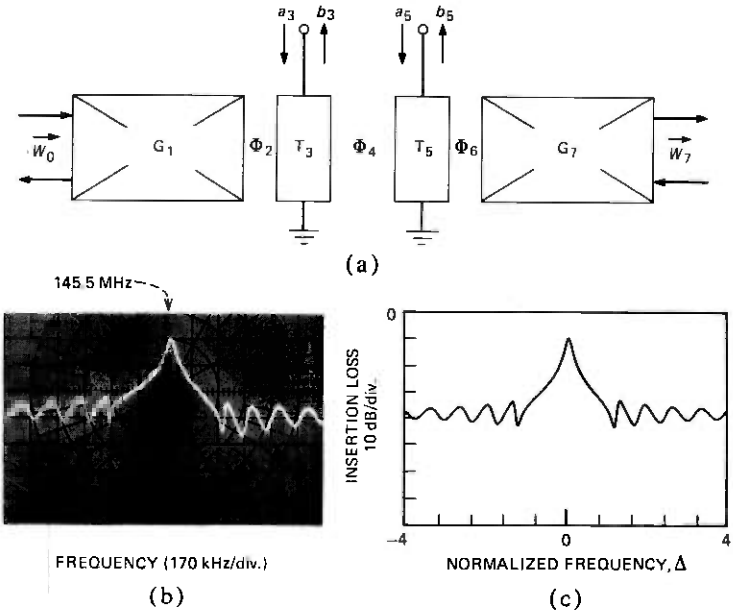


Fig. 5—(a) Diagram of a two-port resonator. (b) The electrical transmission spectrum for a two-port resonator on YZ-$LiNbO_3$ with gratings 9.6 mm long and 12 μm period, optimally placed transducers with $N_t = 5$, $Z_e = 50\ \Omega$, and an acoustic aperture of 50 wavelengths. (c) The calculated spectrum for the device in (b) using $\kappa = 4.5\ cm^{-1}$, $\alpha/\kappa = 0.01$, $R_g = 11\ \Omega$, $\epsilon = 0.04$, and $\phi_4 = 9.98\ \pi$ on resonance.

(*ii*) The electrical signal leaving the transducer is expressed by

$$b_i = \tau_i' \cdot \mathbf{W}_i + a_i(t_{33})_i \tag{30}$$

where τ_i' is an output coupling vector

$$\tau_i' = s \begin{pmatrix} t_{13} \\ -t_{23} \end{pmatrix}_i \tag{31}$$

The symbol · in (30) indicates the scalar (dot) product.

As shown in Appendix B, eqs. (27) and (30) allow the analysis of resonators and coupled-resonators to be reduced to a simple, matrix-multiplication algorithm.

The elements of the matrix $\mathcal{T}$ are evaluated by using an appropriate transducer model.[23,24] The accuracy of the matrix elements depends on the degree of sophistication of the model used. For example, the Mason equivalent circuit model first used for interdigital transducers by Smith[23] *et al.* has proven very useful in practice. The complete matrix $\mathcal{T}$ based on the Smith-Mason model is given in Appendix B.

In many resonator applications, however, only a first-order analysis is required. Thus, by neglecting the static transducer capacitance and

the frequency dependence of the propagation phase-shift through the transducer, $\mathcal{T}$ is given by

$$(\mathcal{T})_{\text{first-order}} \approx s \begin{bmatrix} 1+g+g_s & -g-g_s & s\sqrt{2g} \\ g+g_s & 1-g-g_s & s\sqrt{2g} \\ \sqrt{2g} & -\sqrt{2g} & s \end{bmatrix} \tag{32}$$

where

$$g = G_r Z_e \tag{33}$$

$$g_s = G_r R_s \tag{34}$$

and G_r, Z_e, and R_s are the transducer radiation conductance, load resistance, and series electrode resistance, respectively. The first-order matrix in eq. (32) is sufficient for calculating the near-resonance properties of many SAW resonators, but the more complete matrix in eq. (84) is required for wideband descriptions.

As a first application of the transducer matrix in (32) and of the matrix-multiplication algorithm in Appendix B, consider the two-port resonator in Fig. 5a. Ideally, the transducers are optimally-placed[25] ($\phi_2 = \phi_6 = \pi/4$), and the cavity is resonant at $\Delta = 0$ ($\phi_4 = m\pi$). Thus, from eq. (32) and eqs. (96)–(103), the electrical power-transmission factor P_{53} of the optimal two-port is given by

$$P_{53} = \left|\frac{b_5}{a_3}\right|^2 = \left|\frac{2g}{2g + 2g_s + \dfrac{1-r}{1+r}}\right|^2 \tag{35}$$

where, $r = j\Gamma$, Γ is the frequency-dependent reflection coefficient of each grating (G_1 is assumed to be identical to G_7), and g and g_s are given in eqs. (33) and (34), respectively.

The total loading on the cavity can be separated into two components: (*i*) the power coupled to external circuit and (*ii*) the power lost in the filter structure.

Thus, eq. (35) can be written in the more intuitively recognizable form

$$P_{53} = \left|\frac{\mu_C}{\mu_C + \mu_L}\right|^2 \tag{36}$$

where

$$\mu_C = 8g \tag{37}$$

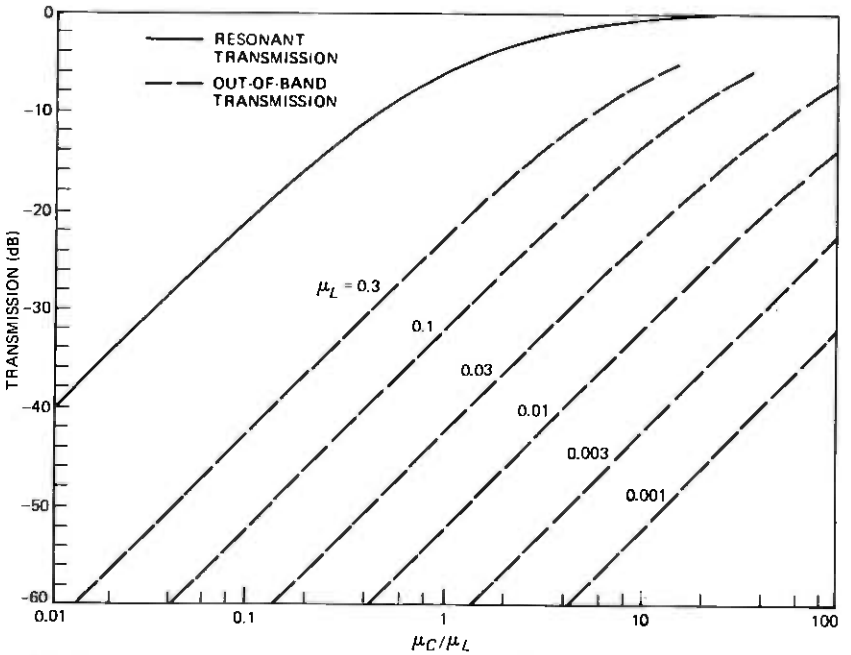


Fig. 6—Nomogram giving the resonant and out-of-band transmission for two-port, surface-acoustic-wave grating resonators and matched grating resonator pairs. The resonant transmission is determined by the ratio of the transducer cavity loading μ_C to the cavity loading μ_L due to all other mechanisms. The out-of-band transmission is only a function of μ_C. The dashed curves are contours of out-of-band transmission for constant cavity loss. The resonant and out-of-band transmission can be found from μ_C and μ_L or vice-versa. The nomogram is directly applicable to single resonators and matched collinearly coupled resonator pairs. To use the nomogram with matched multistrip-coupled pairs, multiply the ordinate by $4\nu_m^2|\Gamma|^2$ (see Section VII) and for matched transducer-coupled cavities, multiply the ordinate by $(\nu_t/4)^2$ (see Section VIII).

is the single-transit, fractional power coupling to the external circuit and

$$\mu_L = 8g_s + 4(1 - r)/(1 + r) \tag{38}$$

is the single-transit, fractional power loss due to all other mechanisms (ohmic loss, bulk scattering, intrinsic propagation losses, and transmission through the gratings). Note that in the optimal resonator described here, the transducers are spaced an integral number of half wavelengths apart so that coherent interactions take place that allow μ_C to be greater than 1 for strong-coupling transducers.

On resonance ($\Delta = 0$), for highly reflective gratings [$\exp(2\kappa L) \gg 1$], eq. (38) becomes

$$\mu_L \approx 8g_s + 2\epsilon + 2\alpha/\kappa + 4\exp(-2\kappa L) \tag{39}$$

where ϵ is a localized[17] excess loss that accounts for mode-conversion losses. By dividing the numerator and denominator in eq. (36) by μ_L^2, the

resonant power transmission through a resonator is described by the single parameter μ_C/μ_L.

Figure 6 is a nomogram for finding the resonant and out-of-band transmission of grating resonators. The solid curve is the resonant insertion loss versus μ_C/μ_L. Plotted with dashed curves is the out-of-band transmission with the cavity loss μ_L as a parameter. Using the nomogram, the resonant and out-of-band transmission can be found knowing μ_C and μ_L or vice versa. The nomogram is also applicable to coupled resonators as described in the caption to Fig. 6 and in Sections VI, VII, and VIII. Coldren and Rosenberg[6,17] have used similar diagrams for the resonant insertion loss of single and multistrip-coupled resonators as a function of coupling and loss parameters.

Equation (35) can also be used to find the loaded electrical Q, Q_{Le_1}, of a single-cavity, two-port resonator. For $\exp(2\kappa L) \gg 1$, it is found that

$$Q_{Le_1} = \frac{\pi}{\kappa\Lambda} \frac{1}{(\mu_C + \mu_L)} \tag{40}$$

where $\pi/\kappa\Lambda$ is the single-transit cavity phase-shift.

The algorithm used to derive eq. (35) provides a flexible tool for interpreting experimental device performance, since a large number of electrical, mechanical, and geometrical properties are explicitly contained in the analysis. For example, consider the transmission response in Fig. 5b of a two-port resonator with Ti-diffused gratings on YZ-$LiNbO_3$. The resonant insertion loss is 10 dB, and from eq. (36) or Fig. 6,

$$\frac{\mu_C}{\mu_L} = 0.46 \tag{41}$$

Next, the transducers each have five electrodes 50 wavelengths long, and, from eqs. (33), (93), and (94),

$$\mu_C = 0.052 \tag{42}$$

for $Z_e = 50\Omega$. From (41) and (42), it is found that

$$\mu_L = 0.112 \tag{43}$$

The transmission minima on each side of the resonance occur near the first reflection zeroes of the gratings. Thus, eq. (11) can be used to estimate κ, with the result

$$\kappa \approx 4.3 \text{ cm}^{-1} \tag{44}$$

The gratings are each 0.96 cm long (800 Λ), and from eq. (44),

$$e^{-2\kappa L} = 0.00026 \tag{45}$$

From external transmission measurements on resonators (see Sections II and III), the loss α/κ associated with diffused gratings is found to be $\sim$0.01. Thus, from (39)–(45), the remaining loss is probably associated with the transducers and is given by

$$8g_s + 2\epsilon = \mu_L - \frac{2\alpha}{\kappa} - 4e^{-2\kappa L} = 0.092 \tag{46}$$

The electrode resistance ($R_s = 11\ \Omega$) is calculated from the metal thickness (2700 Å of aluminum),

$$g_s = 0.0014 \tag{47}$$

and, finally, from eqs. (46) and (47)

$$\epsilon = 0.040 \tag{48}$$

The 4 percent excess loss ϵ is probably due to bulk mode conversion by the transducer electrodes. Both loss mechanisms associated with the transducers (series resistance and bulk mode conversion) should be less significant on low-coupling materials such as ST-quartz due to the increased transducer length.

In order to complete the description of the resonator in Fig. 5a, the phase-shifts ϕ_2, ϕ_4, and ϕ_6 must be specified. It is observed in practice that the velocity of propagation is very sensitive to surface perturbations (piezoelectric-loading, mass-loading, and reactive energy storage). As a result, the separation between the gratings must be empirically adjusted to compensate for the velocity variations in the structure. For the device of Fig. 5b, the appropriate empirical values are $\phi_2 = \phi_4 = \pi/4$ and $\phi_6 = 9.98\ \pi$ on resonance.

The parameters estimated in (41)–(48) have been used with the algorithm in Appendix B to calculate the complete transmission spectrum shown in Fig. 5c.

V. COUPLED GRATING-RESONATORS—GENERAL CONSIDERATIONS

Multipole filters are formed by coupling together two or more cavities. The general configuration for a cascade-coupled multipole resonator-filter is shown in Fig. 7. Acoustic energy is launched by the transducer in the input cavity, propagates through the coupling structure C_5, and is detected by the transducer in the output cavity. The coupling structure C_5 consists in general of some combination of gratings, phase shifts, transducers, and multistrip couplers. The overall filter response is determined by the properties of C_5 as well as the properties of the input and output cavities.

In order to better understand the various elements that can be used in the coupling structure C_5, two-pole resonators formed by acoustic collinear coupling, multistrip coupling, and transducer coupling are

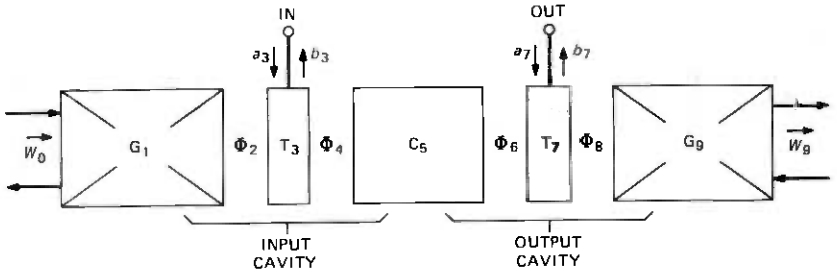


Fig. 7—Diagram of the configuration for cascade-coupled SAW grating resonators with an arbitrary coupling element, C_5.

discussed individually in the next three sections. It is shown for each coupling mechanism that the important, near-resonance properties of the coupling structure are expressed by the matrix $\hat{\mathcal{C}}$

$$\hat{\mathcal{C}} = \frac{1}{\nu}\begin{bmatrix} e^{j2\delta L_{\text{eff}}} & j\sqrt{1-\nu^2} \\ -j\sqrt{1-\nu^2} & e^{-j2\delta L_{\text{eff}}} \end{bmatrix} \tag{49}$$

where ν is a real parameter ≤ 1, and L_{eff} is the effective contribution to the cavity length by the coupling structure.

The parameter ν is the magnitude of the amplitude transmission coefficient through the coupling structure and is a measure of the degree of coupling between the cavities. The quantity $\exp(j2\delta L_{\text{eff}})$ is a propagation phase factor that accounts for the phase shift through the coupling structure.

The degree of coupling between the cavities (specified by ν) largely determines the transmission characteristics of the resonator pair. For example, using the method outlined in Appendix B, the resonant transmission of a pair of cavities is found to be

$$\left|\frac{b_7}{a_3}\right|^2_{\delta=0} = \frac{1}{4}\left|\frac{\nu\mu_C}{1+\left(\frac{\mu_C+\mu_L}{4}\right)^2 - \sqrt{1-\nu^2}\left[1-\left(\frac{\mu_C+\mu_L}{4}\right)^2\right]}\right|^2 \tag{50}$$

where the quantity $(\mu_C + \mu_L)$ is the single-transit, fractional power loading on the combined resonator pair. Equivalently, $(\mu_C + \mu_L)$ can be interpreted as the *round-trip* power loading on each cavity.

By differentiating eq. (50) with respect to ν, it is found that maximum, resonant transmission is obtained when the coupling structure "matches" [26] the two cavities according to

$$\nu_{\text{opt}} = \frac{1}{2}\frac{\mu_C+\mu_L}{1+\left(\frac{\mu_C+\mu_L}{4}\right)^2} \tag{51}$$

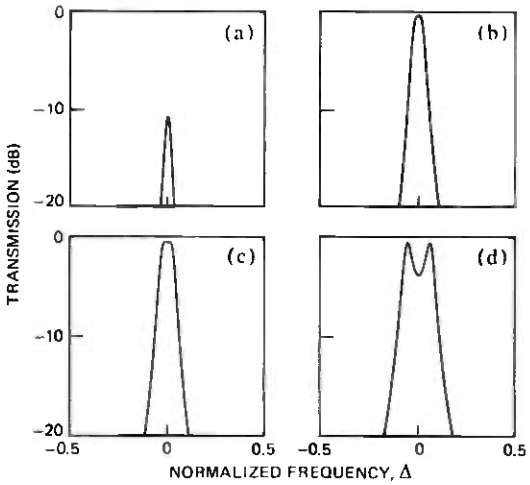


Fig. 8—Near-resonance transmission spectra for lossless, coupled resonator pairs that are: (a) undercoupled, (b) matched, (c) slightly overcoupled, and (d) overcoupled for maximum 3-dB bandwidth.

Qualitatively, the cavities are matched when the loading on each cavity due to the coupling structure is equal to the loading due to all other mechanisms.

The importance of matching the individual resonators in a coupled structure is illustrated in Fig. 8. If the parameter ν is too small, the cavities are *undercoupled* and there is a large resonant insertion loss as in Fig. 8a. When $\nu = \nu_{opt}$ from eq. (51), the cavities are *matched* and minimum insertion loss is obtained as shown in Fig. 8b. As ν is increased slightly beyond ν_{opt}, the peak flattens and broadens as in Figure 8c. For still larger values of ν the cavities become *overcoupled* and the resonance splits into two peaks as in Fig. 8d where the dip between peaks is 3 dB. Thus, the degree of cavity-coupling, ν, is a central parameter in determining the passband shape and insertion loss.

The matched condition (51) has a further interesting consequence. When the frequency dependence of the transfer function is included in (50), it can be shown for matched cavities that

$$\left|\frac{b_7}{a_3}\right|^2 \approx \left|\frac{\mu_C}{\mu_C + \mu_L}\right|^2_{\delta=0} \left[\frac{1}{1 + \Omega^2 U_2^2(\Omega)}\right] \tag{52}$$

where $U_2 = 2\Omega$ is the second Chebyshev polynomial of the second kind. The parameter Ω is a normalized frequency

$$\Omega = 2\frac{\Delta\omega}{\omega_0} Q_{Le_2} \tag{53}$$

where Q_{Le_2} is the loaded electrical Q of each cavity in the coupled pair.

The Chebyshev-polynomial form in eq. (52) is the same as the form obtained for a coupled pair of identical resonators in the external transmission configuration [see eq. (24)]. Although it is not rigorously proven here, eq. (52) indicates that the passband shapes that can be obtained in external transmission can also be obtained with intracavity transducers. Thus, the procedure for synthesizing resonant passbands can be simplified by first investigating the passband in the external transmission configuration, and subsequently including the transducers.

VI. COLLINEAR ACOUSTICALLY COUPLED RESONATORS

An acoustically coupled resonator pair is formed by inserting a section of grating between the input and output transducers as shown in Fig. 9a. Comparing Figs. 7 and 9a, the coupling structure in Fig. 9a is simply a section of grating. From eq. (14a), the near-resonance coupling matrix is given by

$$\mathcal{C} \approx \mathcal{A}\mathcal{F}\mathcal{A} \tag{54}$$

The loss matrix $\mathcal{A}$ in eq. (54) has the same form as that given in eq. (14b), but any excess loss due to the transducers must be included.

The near-resonance behavior of a highly reflective grating, described by the matrix $\mathcal{F}$, is approximately equal to the coupling matrix $\hat{\mathcal{C}}$ in eq. (49) when the identification

$$\nu_g = \operatorname{sech}(\kappa L_5) \tag{55}$$

is made and L_{eff} is the effective penetration depth into the grating, $1/2\kappa$. The quantity ν_g is the coupling parameter for collinear acoustic coupling and L_5 is the total length of the coupling grating.

Including dissipative loss, the matching condition (51) specialized to collinear acoustic coupling, is given by

$$(e^{-\kappa L_5})_{\text{opt}} = \frac{1}{4}(\mu_C + \mu_{Lg}) \tag{56}$$

where $\mu_C = 8g$ is the transducer loading on the resonator pair, and μ_{Lg} is the effective loading on each cavity due to all other mechanisms,

$$\mu_{Lg} = 8g_s + 4\epsilon_g + 4\frac{\alpha}{\kappa} + 4e^{-2\kappa L_1} \tag{57}$$

In deriving eq. (57), it is assumed that the outer gratings, G_1 and G_9, are identical.

Comparing eqs. (39) and (57), the expression for μ_{Lg} is similar to that for μ_L (for a single cavity) with the exceptions: (*i*) the grating-loss contribution is twice as large ($4\alpha/\kappa$ versus $2\alpha/\kappa$) because there are four effective reflection planes instead of two, and (*ii*) the excess loss ϵ_g is in

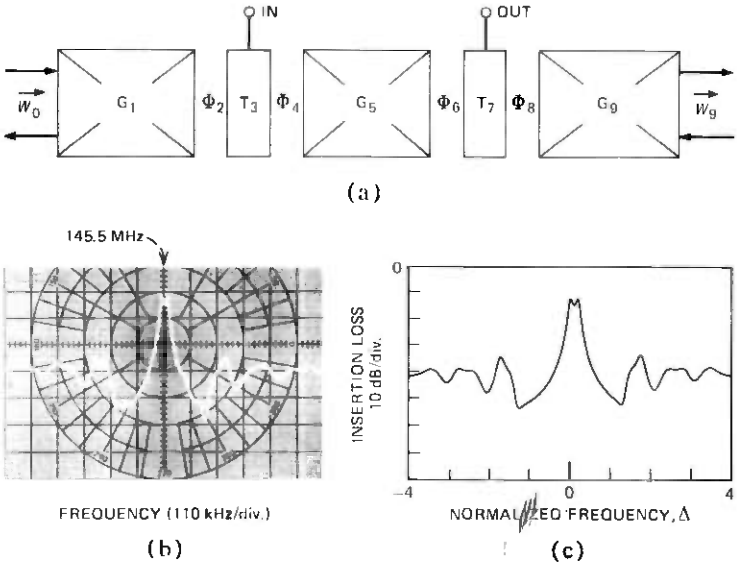


Fig. 9—(a) Diagram of an acoustically cascaded resonator-filter pair. (b) The electrical transmission spectrum in a 50 Ω system for an acoustically cascaded resonator-filter pair on YZ-$LiNbO_3$ with $L_1 = L_9 = 9.60$ mm, $L_5 = 7.296$ mm, $\Lambda = 12$ μm, $N_t = 5$ and an acoustic aperture of 50 wavelengths. (c) The calculated spectrum for the device in (b) using $\kappa = 3.3$ cm^{-1}, $\alpha/\kappa = 0.01$, $R_s = 12$ Ω, $\epsilon_g = 0.018$ and $\phi_2 = \phi_4 = \phi_6 = \phi_8 = 0.234\ \pi$ on resonance.

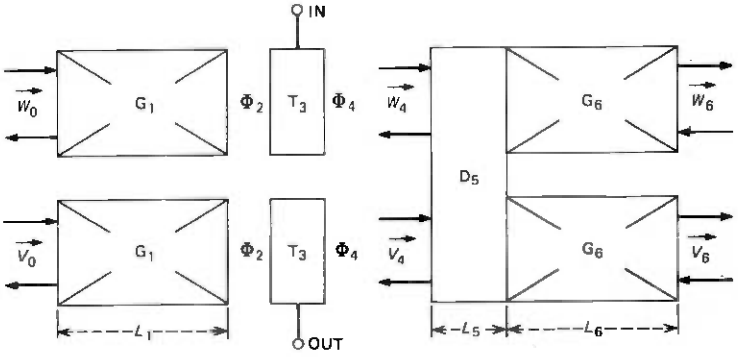
general different from the excess loss ϵ for a single cavity. In fact, the origins of the excess loss can be investigated, by comparing measured values of ϵ_g and ϵ. For example, if the excess loss is predominantly caused by the gratings, $\epsilon_g \simeq \epsilon$. If, however, the excess loss is transducer-associated, $\epsilon_g \approx \epsilon/2$ since there is only one transducer in each cavity in an acoustically coupled pair.

As an aid in design and data interpretation, the nomogram in Fig. 6 is directly applicable to matched acoustically coupled cavities when μ_{Lg} is substituted for μ_L.

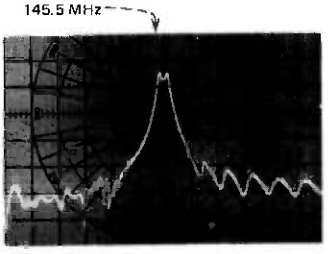
The transmission spectrum of an acoustically coupled resonator pair is shown in Fig. 9b. The transducers and outer gratings are identical to those used in the single-cavity resonator in Fig. 5b. The experimental parameters have been estimated as described in the previous section, and the calculated response is shown in Fig. 9c. It is interesting to note that the value $\epsilon_g = 0.018 \approx \epsilon/2$ is found, providing further evidence that the excess loss is transducer associated on $LiNbO_3$.

VII. MULTISTRIP-COUPLED RESONATORS

Grating resonators can also be coupled using a directional (multistrip) coupler[6,7] as shown in Fig. 10a. A detailed analysis of the multistrip-coupled resonator pair from a scattering-matrix point of view has been

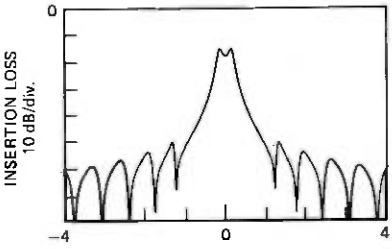


(a)



FREQUENCY (175 kHz/div.)

(b)



NORMALIZED FREQUENCY, Δ

(c)

Fig. 10—(a) Diagram of a multistrip-coupled resonator-filter pair. (b) The electrical transmission spectrum in a 50 Ω system for a multistrip (5 μm strips, $L_6 = 90$ μm) coupled resonator-filter-pair on YZ-$LiNbO_3$ with $L_1 = L_6 = 9.60$ mm, $\Lambda = 12$ μm, $N_t = 5$ and an acoustic aperture in each track of 50 wavelengths. (c) The calculated transmission spectrum of the device in (b) with $q = 0.163$, $\kappa = 4.3$ cm^{-1}, $\alpha/\kappa = 0.01$, $R_s = 10$ Ω, $\epsilon_m = 0.047$, $\phi_2 = 0.25$ π and $\phi_4 = 9.89$ π on resonance.

given by Rosenberg and Coldren.[6] In this section we derive the coupling matrix ($\mathcal{C}_5$ in Fig. 7) for multistrip-coupled cavities.

The overall structure consists of two resonators in parallel connected by an ideal, directional coupler[27] described by the fourth-order vector equation

$$\begin{pmatrix} \mathbf{W}_4 \\ \mathbf{V}_4 \end{pmatrix} = \begin{pmatrix} \mathcal{P} & \mathcal{Q} \\ \mathcal{Q} & \mathcal{P} \end{pmatrix} \begin{pmatrix} \mathbf{W}_5 \\ \mathbf{V}_5 \end{pmatrix} \tag{58}$$

where

$$\mathcal{P} = \begin{pmatrix} p & 0 \\ 0 & p \end{pmatrix} \tag{59}$$

$$\mathcal{Q} = \begin{pmatrix} -jq & 0 \\ 0 & jq \end{pmatrix} \tag{60}$$

and

$$p^2 + q^2 = 1 \tag{61}$$

For simplicity, the frequency dependence of the propagation phase shifts through the multistrip coupler is ignored in eqs. (58)–(60).

Comparing Figs. 7 and 10a, the coupling element is the multistrip coupler in combination with the gratings G_6. The transmission between W_4 in the upper track and V_4 in the lower track can be treated as a two-port cascade element. Thus, the 2×2 matrix $\mathcal{D}$ satisfying

$$\begin{pmatrix} w_4^+ \\ w_4^- \end{pmatrix} = \mathcal{D} \begin{pmatrix} v_4^- \\ v_4^+ \end{pmatrix} \tag{62}$$

becomes the coupling matrix for multistrip-coupled cavities.

To solve for $\mathcal{D}$, the appropriate acoustic boundary conditions are

$$w_6^- = v_6^- = 0 \tag{63}$$

and the resulting matrix is

$$\mathcal{D} = \frac{j}{2pq} \begin{pmatrix} -\dfrac{1}{\Gamma_6} & (p^2 - q^2) \\ -(p^2 - q^2) & \Gamma_6 \end{pmatrix} \tag{64}$$

where Γ_6 is the reflection coefficient of the gratings, G_6.

Near resonance ($|\Delta| \ll 1$), Γ_6 can be expanded as in eq. (14), and for $\exp(\kappa L_6) \gg 1$, eq. (64) becomes

$$\mathcal{D} = \frac{1}{\nu_m} \mathcal{A} \begin{bmatrix} e^{j\Delta} & j\sqrt{1 - \nu_m^2} \\ -j\sqrt{1 - \nu_m^2} & e^{-j\Delta} \end{bmatrix} \mathcal{A} \tag{65}$$

where ν_m is the coupling parameter for multistrip-coupled cavities

$$\nu_m = 2q\sqrt{1 - q^2} \tag{66}$$

and $L_{\text{eff}} \approx 1/2\kappa$ since the length of the multistrip coupler is neglected. The loss matrix $\mathcal{A}$ in eq. (65) has the same form as that given for a single grating in (14b), but any excess loss due to the multistrip coupler must now be included. Thus, the matching condition for multistrip-coupled cavities becomes

$$q_{\text{opt}} = \frac{\mu_C + \mu_{Lm}}{\sqrt{16 + (\mu_C + \mu_{Lm})^2}} \tag{67}$$

where μ_{Lm} is the single-transit power loss of the resonator pair (excluding transducer coupling):

$$\mu_{Lm} = 8g_s + 4\epsilon_m + 4\frac{\alpha}{\kappa} + 4e^{-2\kappa L_6} \tag{68}$$

The excess loss ϵ_m now includes additional losses suffered due to the multistrip coupler.

As pointed out by Rosenberg,[6] far away from resonance ($|\Delta| \gg 1$) the multistrip-coupled structure has low out-of-band transmission, since the path connecting input and output requires a reflection from a grating. Quantitatively, from eq. (64).

$$\left|\frac{v_4^-}{w_4^+}\right|^2 = 4\nu_m^2 |\Gamma_6(\Delta)|^2 \tag{69}$$

As indicated by eq. (69), the effective cavity-coupling is directly proportional to Γ_6. Thus, the out-of-band transmission of a multistrip-coupled pair is low and can be suppressed to arbitrarily small values by using sidelobe-free apodized gratings.[28]

The nomogram in Fig. 6 can be used for matched, multistrip-coupled cavities when μ_{Lm} is substituted for μ_L, and the ordinate for out-of-band transmission is multiplied by $4\nu_m^2|\Gamma_6|^2$.

In Fig. 10b is shown the experimental transmission spectrum of a multistrip-coupled device, and in Fig. 10c is shown the spectrum for the same device calculated using (64) and the parameters given in the caption. The high resonant insertion loss (15 dB) is due to the large cavity perturbations ($\epsilon_m = 0.047$) caused both by the transducers and multistrip coupler. The distortion in the sidelobe response is due to slight nonuniformities in the gratings and direct capacitive coupling between the input and output transducers (RF feedthrough).

VIII. TRANSDUCER-COUPLED RESONATORS

The general scheme for using transducers to electrically couple two resonators is depicted in Fig. 11a. The coupling structure is topologically similar to the multistrip-coupled case, but with the important advantage that an electrical coupling network can be inserted between the resonators if desired. In general, both passive and active electrical circuit components can be employed so that passband shaping and amplification/attenuation can be performed in the coupling network. Thus, the electrically coupled configuration offers more design flexibility than either the acoustic cascade or directionally coupled configurations.

To gain an insight into the performance of electrically coupled resonators, we examine the important case[9,10] where the coupling network is simply a shunt susceptance $j\chi$. The coupling structure (transducers T_5 in combination with gratings G_7 and shunt susceptance $j\chi$) is described by the electrical coupling matrix $\mathscr{E}$ satisfying

$$\begin{pmatrix} w_4^+ \\ w_4^- \end{pmatrix} = \mathscr{E} \begin{pmatrix} v_4^- \\ v_4^+ \end{pmatrix} \tag{70}$$

Using the acoustic boundary conditions (as for the multistrip-coupled structure)

$$w_7^- = v_7^- = 0 \tag{71}$$

and assuming the two coupling transducers are identical, with N_t electrodes, the matrix $\mathcal{E}$ is found to be

$$\mathcal{E} = \frac{-2j(-1)^{N_t}Q_t}{(1+r)^2}\begin{bmatrix} 1 - j\left(\dfrac{1+r}{Q_t}\right) & r + j\left(\dfrac{1-r^2}{2Q_t}\right) \\ -r - j\left(\dfrac{1-r^2}{2Q_t}\right) & -r^2 + j\left(\dfrac{r(1+r)}{Q_t}\right) \end{bmatrix} \tag{72}$$

where $r = j\Gamma_7$, Γ_7 is the reflection coefficient of gratings G_7, and Q_t is the effective radiation Q of the cavity-coupling transducers:

$$Q_t = (\omega C_T + \chi/2)/G_r \tag{73}$$

The quantities C_T and G_r are the transducer static capacitance and radiation conductance, respectively. For clarity of exposition, in deriving eq. (72), the transducer length is assumed small compared to energy penetration depth in the gratings ($\theta_t \approx 0$) and the series resistance is neglected ($R_s = 0$). In eq. (72), the loss due to series resistance in the cavity-coupling transducers T_5 can be mathematically included in the grating loss coefficient as done for the losses in the multistrip coupler in Section VII.

For the electrical coupling structure, the phase shifts $\phi_6 = \pi/4$ must be included between the coupling transducers and the gratings G_7 in order to obtain optimum coupling of the transducers to the cavity standing-wave-pattern. Further, as noted by Matthaei *et al.*,[10] the coupling transducers introduce a small phase shift due to the finite value of Q_t. Thus, expanding (72) for $|\Delta| \ll 1$ and $\exp(\kappa L_7) \gg Q_t$, the matrix $\mathcal{E}$ is given by

$$\mathcal{E} = -(-1)^{N_t}\frac{1}{\nu_t}\mathcal{A}\begin{bmatrix} e^{j(\Delta+\phi_{ex})} & j\sqrt{1-\nu_t^2} \\ -j\sqrt{1-\nu_t^2} & e^{-j(\Delta+\phi_{ex})} \end{bmatrix}\mathcal{A} \tag{74}$$

where the matrices $\mathcal{A}$ account for all dissipative cavity losses due to gratings and transducers T_5, and the constant excess phase shift ϕ_{ex} is given by

$$\phi_{ex} = \pi/2 - 2/Q_t \tag{75}$$

The quantity ν_t is the cavity-coupling parameter for transducer coupling,

$$\nu_t = \frac{2Q_t}{Q_t^2 + 1} \tag{76}$$

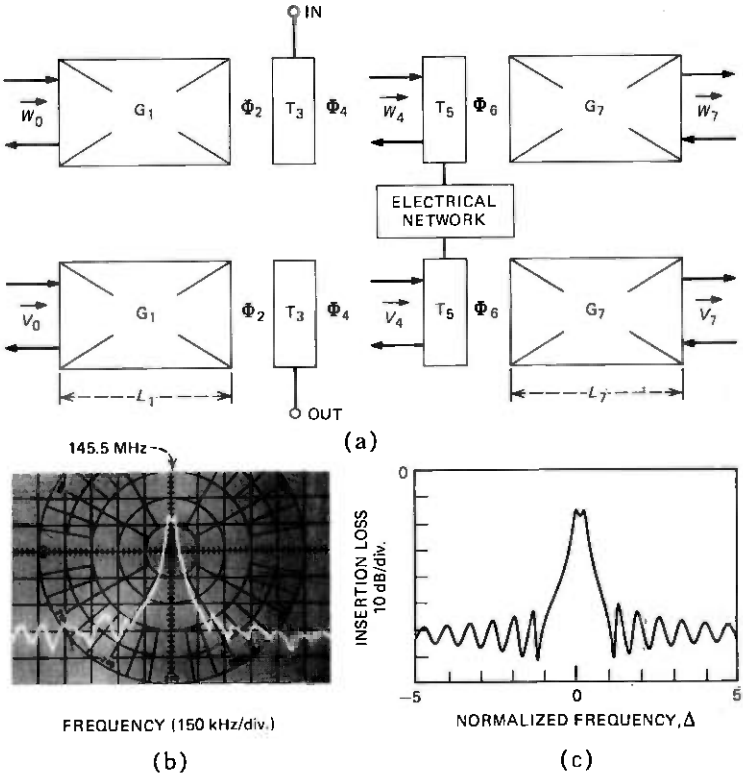


Fig. 11—(a) Diagram of a transducer-coupled resonator-filter pair. (b) The electrical transmission spectrum in a 50 Ω system of an electrically coupled, resonator-filter pair on YZ-$LiNbO_3$ with optimally placed transducers with $N_t = 5$, $L_1 = L_7 = 9.60$ mm, $\Lambda = 12$ μm and an acoustic aperture in each track of 50 wavelengths. (c) The calculated spectrum of the device in (b) with $Q_t = 6.69$, $\kappa = 4.0$ cm^{-1}, $\alpha/\kappa = 0.01$, $R_s = 11$ Ω, $\epsilon_t = 0.047$, and $\phi_4 = 10\pi$ on resonance.

Here again, L_{eff} in (49) is given by the penetration depth $(1/2\kappa)$ into the grating since the transducer length has been ignored. The matching condition for transducer-coupled cavities becomes

$$(Q_t)_{\text{opt}} = \frac{4}{\mu_C + \mu_{Lt}} \tag{77}$$

where μ_{Lt} is the single-transit power loss of the resonator pair (excluding loading by the external circuit),

$$\mu_{Lt} = 8g_s + 4\epsilon_t + 4\frac{\alpha}{\kappa} + 4e^{-2\kappa L_7} \tag{78}$$

The excess loss ϵ_t accounts for all additional losses due to the cavity-coupling transducers as well as the excess loss from the input-output

transducers, and the term $8g_s$ is due to the input-output transducers T_3.

Outside the grating stop-band ($|\Delta| \gg 1$), the power transmission $|v_4^-/w_4^+|^2$ through the coupling elements tends to the limit

$$\left|\frac{v_4^-}{w_4^+}\right|^2 \approx \left(\frac{\nu_t}{4}\right)^2 \tag{79}$$

The out-of-band transmission of the transducer-coupled configuration is therefore lower than with collinear acoustic coupling but is still higher than the out-of-band level for multistrip-coupled resonators [see (69)].

The resonator nomograph in Fig. 6 can be used for matched, transducer-coupled resonators when μ_{Lt} is substituted for μ_L, and the ordinate for out-of-band transmission is multiplied by $(\nu_t/4)^2$.

In Fig. 11b, the experimental transmission spectrum of a pair of transducer-coupled cavities is shown, and the theoretical response of the same device calculated using (72) and the parameters given in the caption is shown in Fig. 11c. The excess loss ϵ_t is about the same as ϵ for a single-cavity resonator, as would be expected. As for the multistrip-coupled pair, the distortion in the sidelobe response is caused by grating nonuniformities and RF feedthrough.

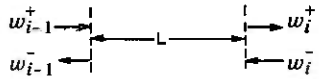
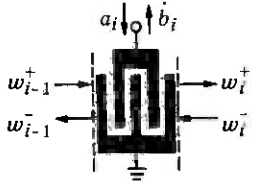
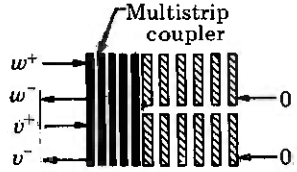
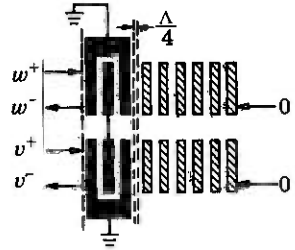
IX. SUMMARY AND CONCLUSIONS

The major results derived in this paper are summarized in Table I. Gratings and small pieces of transmission line are the fundamental elements for SAW resonators. Using coupled-mode theory, gratings and transmission lines are described by 2×2 transmission matrices. Resonators and combinations of resonators can be analyzed simply by multiplying together a sequence of transmission matrices. A matrix-multiplication algorithm is also presented for analyzing bandpass filters with intracavity transducers.

To form multipole filters, several resonators can be coupled together using one or more of the three mechanisms: (*i*) collinear acoustic coupling, (*ii*) multistrip coupling, or (*iii*) transducer coupling. Near the resonant frequency all three mechanisms are mathematically equivalent and can be used interchangeably in passband synthesis applications. Far off the resonant frequency, the three mechanisms have quite different sidelobe suppression characteristics.

The essential properties of the three coupling mechanisms are illustrated in Fig. 12. The calculated transmission spectra for three different coupled resonator pairs are shown. In each case, the cavities are of identical length and are coupled to the same degree (same value of ν) with only the cavity-coupling mechanism being changed from case to case. All three spectra have nearly the same passband shape, but the electri-

Table I—Summary of resonator elements

Element	Schematic diagram	Simplified, near-resonance description	Relevant equations
Grating		$\begin{bmatrix} w_{i-1}^+ \\ w_{i-1}^- \end{bmatrix} = \frac{1}{\nu_g} \begin{bmatrix} e^{j\Delta} & j\sqrt{1-\nu_g^2} \\ -j\sqrt{1-\nu_g^2} & e^{-j\Delta} \end{bmatrix} \begin{bmatrix} w_i^+ \\ w_i^- \end{bmatrix}$ $\nu_g = \operatorname{sech}(\kappa L)$	(8), (14) (82)
Transmission line		$\begin{bmatrix} w_{i-1}^+ \\ w_{i-1}^- \end{bmatrix} = \begin{bmatrix} e^{j\beta L} & 0 \\ 0 & e^{-j\beta L} \end{bmatrix} \begin{bmatrix} w_i^+ \\ w_i^- \end{bmatrix}$	(13)
Transducer		$\begin{bmatrix} w_{i-1}^+ \\ w_{i-1}^- \\ b_i \end{bmatrix} = (-1)^{N_t} \begin{bmatrix} 1+g & -g & (-1)^{N_t}\sqrt{2g} \\ g & 1-g & (-1)^{N_t}\sqrt{2g} \\ \sqrt{2g} & -\sqrt{2g} & (-1)^{N_t} \end{bmatrix} \begin{bmatrix} w_i^+ \\ w_i^- \\ a_i \end{bmatrix}$	(25), (26), (32), (84)
Multistrip cavity coupler		$\begin{bmatrix} w^+ \\ w^- \end{bmatrix} = \frac{1}{\nu_m} \begin{bmatrix} e^{j\Delta} & j\sqrt{1-\nu_m^2} \\ -j\sqrt{1-\nu_m^2} & e^{-j\Delta} \end{bmatrix} \begin{bmatrix} v^- \\ v^+ \end{bmatrix}$ $\nu_m = 2q$	(64), (65)
Transducer cavity coupler		$\begin{bmatrix} w^+ \\ w^- \end{bmatrix} = \frac{1}{\nu_t} \begin{bmatrix} e^{j(\Delta+\phi_{ex})} & j\sqrt{1-\nu_t^2} \\ -j\sqrt{1-\nu_t^2} & e^{-j(\Delta+\phi_{ex})} \end{bmatrix} \begin{bmatrix} v^- \\ v^+ \end{bmatrix}$ $\nu_t = 2/Q_t$ $\phi_{ex} = \pi/2 - \nu_t$	(72), (74), (75), (76)

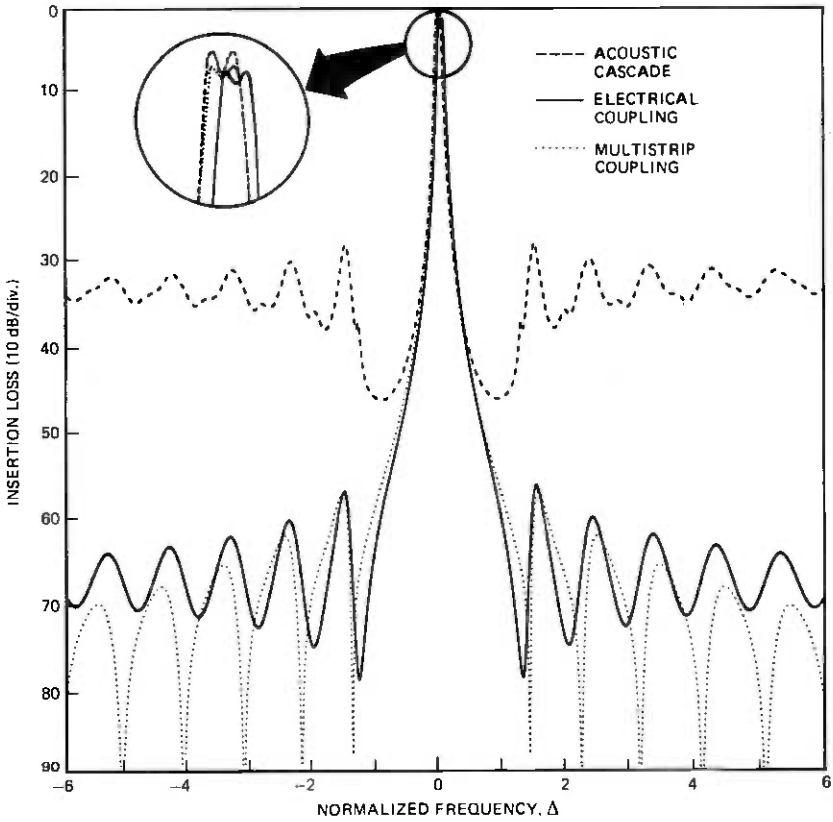


Fig. 12—The calculated transmission spectra of three equivalent resonator-filter pairs on YZ-$LiNbO_3$, each using a different cavity-coupling mechanism. In each case, the devices are assumed lossless and the outer gratings are 800 periods long with $\Lambda = 12$ μm and $\kappa = 3.27$ cm^{-1}. The transducers have $N_t = 5$ with an acoustic aperture of 100 wavelengths. The degree of cavity coupling is the same in each case with $v_g = v_m = v_t = 0.077$.

cally coupled pair resonates at a higher frequency than the others due to the phase shift introduced by the cavity-coupling transducers. The multistrip and electrically coupled cavities have a slightly greater resonant insertion loss than the acoustic cascade because some energy is lost through the end gratings G_6 in Fig. 10 and G_7 in Fig. 11. The sidelobe levels are highest for the acoustic cascade and progressively lower for transducer and then multistrip coupling.

For the synthesis of multipole filters each coupling mechanism has unique advantages so that a combination of two or more coupling mechanisms will probably be optimal. The acoustic cascade is particularly easy to design because coupling between cavities can be accomplished without disturbing the intrinsic cavity properties. That is, there

are no velocity perturbations, ohmic losses, or spurious reflections introduced into the cavity by the coupling structure.

Transducer coupling allows the flexibility of using an external electrical network in addition to the additional sidelobe suppression mentioned above. The external network can be used to contribute to passband shaping and as a convenient means for post-fabrication trimming of device performance.

Finally, the multistrip coupler offers the lowest sidelobe levels and the technological advantage that no critical alignment of the coupler within the cavity is required (as is the case for transducers).

Beginning with the gross properties of the various coupling mechanisms discussed above and emphasized in Fig. 12, the simple matrices given in Table I can be used to obtain first-order results for a wide variety of filter configurations. More precise results can then be obtained using the exact expressions given earlier in the text. Thus, the analytical techniques presented in this paper should provide a sound basis for developing a synthesis procedure for multipole SAW resonator filters.

ACKNOWLEDGMENTS

The authors wish to acknowledge the valuable technical assistance of L. L. Buhl in device fabrication; of R. H. Bosworth in photomask layout; and of M. J. Madden in the numerical computations. The authors also benefited from numerous discussions with G. D. Boyd, L. A. Coldren, and R. L. Rosenberg which stimulated several aspects of this work.

APPENDIX A—TRANSMISSION MATRIX FOR LOSSY GRATINGS

In this appendix a general grating transmission matrix is derived which includes a propagation attenuation and allows for an arbitrary choice of reference planes.

As in Section II, the grating extends from $x = -L$ to $x = 0$. The velocity perturbation is now generalized to allow an arbitrary phase shift, θ, of the grating with respect to the x axis:

$$v(x) = v_0 - \frac{\Delta v}{2} \cos (Kx + \theta) \tag{80}$$

The scalar wave equation is modified to

$$\frac{d^2\Psi}{dx^2} + \left(\frac{\omega^2}{v^2(x)} - j\frac{2\omega\alpha}{v(x)}\right)\Psi = 0 \tag{81}$$

which includes a propagation attenuation coefficient, α. The grating transmission matrix is found in the manner described in Section II. For

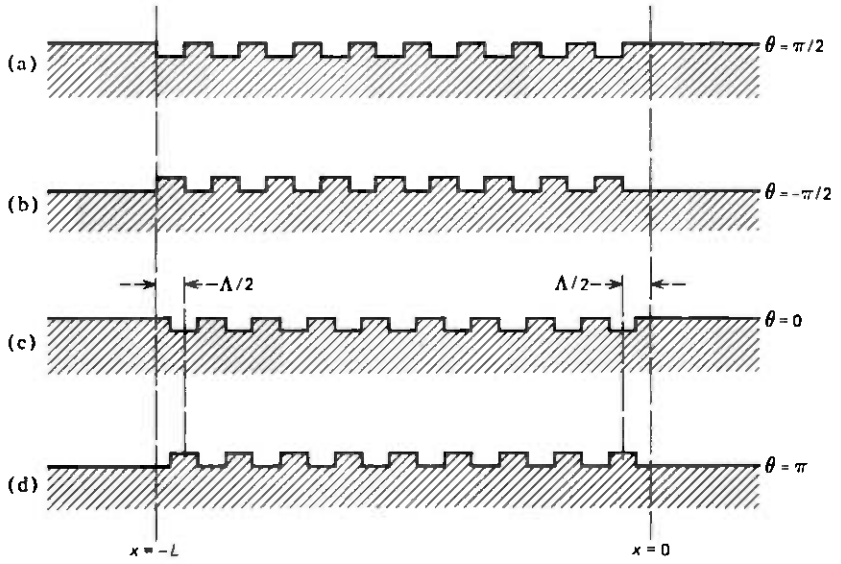


Fig. 13—Location of reference planes and phase angles, θ, for YZ-$LiNbO_3$ and ST quartz surface-deformation gratings: (a) step-down grating with reference plane at the first down-step, (b) step-up grating with reference plane at the first up-step, (c) step-down grating with symmetrically placed reference planes; (d) step-up grating with symmetrically placed reference planes.

the narrowband approximation, $\beta/\beta_0 \approx 1$, the transmission matrix becomes

$$\mathcal{G} = \frac{\kappa}{\sigma}\cosh(\sigma L) \times \begin{bmatrix} \left[\frac{\sigma}{\kappa} + j\left(\frac{\delta - j\alpha}{\kappa}\right)\tanh(\sigma L)\right] e^{j\beta_0 L} & je^{-j\theta}\tanh(\sigma L)e^{j\beta_0 L} \\ -je^{j\theta}\tanh(\sigma L)e^{-j\beta_0 L} & \left[\frac{\sigma}{\kappa} - j\left(\frac{\delta - j\alpha}{\kappa}\right)\tanh(\sigma L)\right] e^{-j\beta_0 L} \end{bmatrix} \quad (82)$$

where

$$\sigma = [\kappa^2 - (\delta - j\alpha)^2]^{1/2}$$

This matrix reduces to eq. (8b) when α and θ are set equal to zero.

It is shown in Section II that the magnitude of the grating reflection coefficient provides a means of determining the coupling coefficient. Similarly, the phase of the reflection coefficient specifies the parameter

θ for a particular choice of reference planes. For a lossless grating an integral number of periods long, the reflection coefficient at the Bragg frequency is

$$\Gamma(0) = -je^{+j\theta} \tanh(\kappa L) \tag{83}$$

Thus, when the reference planes of a grating are spaced by an integral number of periods, one need only measure the phase of the reflection coefficient at the Bragg frequency in order to determine θ. For example, consider surface corrugation gratings of the step-down and step-up type as shown in Fig. 13. The experimentally observed optimum transducer placement has shown for both YZ-$LiNbO_3$[29,30] and ST-quartz[17,30] that the electric potential, Ψ, is a maximum at the edge of a step-down grating, and a minimum at the step-up grating edge. Accordingly, for reference planes shown in Fig. 13a, $\theta = +\pi/2$ for a step-down grating and in Fig. 13b, $\theta = -\pi/2$ for a step-up grating. Similarly, for any type of grating and choice of reference plane, θ can be determined from knowledge of the optimum transducer[25] location which gives the position of the potential maximum. For the case of step-down gratings, $\theta = 0$ corresponds to the symmetrical choice of reference planes as shown in Fig. 13c. A symmetrical choice of reference planes for a step-up grating is as shown in Fig. 13d, which requires $\theta = \pi$. In this paper we assume the reference planes have been chosen such that $\theta = 0$ for mathematical simplicity.

APPENDIX II—TRANSDUCER TRANSMISSION MATRIX AND RESONATOR-ANALYSIS ALGORITHM

The transmission matrix $\mathcal{T}$ of an IDT can be found by manipulating the well-known admittance matrix[23,24] based on a Mason equivalent-circuit model. Using the results of Smith *et al.*,[23] and including an effective series electrode resistance, R_s, $\mathcal{T}$ is given by

$$\mathcal{T} = s\begin{pmatrix} (1+t_0)e^{j\theta_t} & -t_0 & st_{13} \\ t_0 & (1-t_0)e^{-j\theta_t} & st_{13}e^{-j\theta_t} \\ t_{13} & -t_{13}e^{-j\theta_t} & st_{33} \end{pmatrix} \tag{84}$$

where

$$t_0 = \frac{G_r(R_s + Z_e)}{1 + j\theta_e} \tag{85}$$

$$t_{13} = \frac{\sqrt{2G_rZ_e}}{1 + j\theta_e} e^{j\theta_t/2} \tag{86}$$

$$t_{33} = 1 - \frac{2j\theta_c}{1 + j\theta_e} \tag{87}$$

$$s = (-1)^{N_t} \tag{88}$$

N_t = number of electrodes in the transducer

$$\theta_t = N_t \Lambda \delta \tag{89}$$

G_r = transducer radiation conductance

$$\theta_c = \omega C_T (R_s + Z_e) \tag{90}$$

$$\theta_e = (\omega C_T + B_r)(R_s + Z_e) \tag{91}$$

$$C_T = (N_t - 1)C_s/2 \tag{92}$$

B_r = transducer radiation susceptance

C_s = static capacitance/electrode pair

For uniform transducers,[23,31]

$$G_r \approx 2G_0(N_t - 1)^2 \left[\frac{\sin\left(\frac{\theta_t}{2}\right)}{\frac{\theta_t}{2}} \right]^2 \tag{93}$$

$$G_0 = k_c^2 C_s \omega / 2\pi \tag{94}$$

k_c^2 = electromechanical coupling constant

$$B_r \simeq 4G_0(N_t - 1)^2 \frac{\sin(\theta_t) - \theta_t}{\theta_t^2} \tag{95}$$

Using the transducer description in eq. (84), we develop an algorithm for analyzing coupled resonators with intracavity transducers. Consider the general cascaded-resonator structure in Fig. 7. The input signal is applied to transducer T_3 which is separated by phase shift Φ_2 from grating G_1. The output is taken from transducer T_7 which is separated by phase-shift Φ_8 from grating G_9. The element C_5 is a generalized coupling element that can be composed of gratings, transducers, phase shifts, and multistrip couplers. The coupling element C_5 is described by the 2×2 transmission matrix $\mathscr{C}_5$. Specific examples of the matrix $\mathscr{C}_5$ are given in the main text for: (*i*) a single-cavity, two-port resonator, (*ii*) acoustically cascaded resonators, (*iii*) multistrip-coupled resonators, and (*iv*) electrically coupled resonators.

From eq. (27), the acoustic amplitudes associated with transducer T_3 can be expressed

$$\mathbf{W}_2 = t_3 \mathbf{W}_3 + a_3 \tau_3 \tag{96}$$

Vector equation (96) is actually two equations with four unknowns $w_2^{\pm}$, $w_3^{\pm}$. Two further equations are obtained from the boundary conditions

expressing the fact that there are no acoustic waves externally incident on the resonator

$$w_0^+ = w_9^- = 0 \tag{97}$$

Next, the boundary conditions can be referred to the reference planes of transducer T_3:

$$\mathbf{W}_0 = \mathcal{G}_1\Phi_2\mathbf{W}_2 \tag{98}$$

$$\mathbf{W}_3 = \mathcal{C}_5\Phi_6 t_7\Phi_8\mathcal{G}_9\mathbf{W}_9 \tag{99}$$

where it is assumed transducer T_7 is connected to a matched load (i.e., $a_7 = 0$).

Combining eqs. (96), (98), and (99), the outward propagating acoustic waves w_9^+ and w_0^- are specified in terms of the electrical input, a_3,

$$\begin{pmatrix} 0 \\ w_0^- \end{pmatrix} = \mathcal{M}\begin{pmatrix} w_9^+ \\ 0 \end{pmatrix} + a_3\mathcal{G}_1\Phi_2\tau_3 \tag{100}$$

where $\mathcal{M}$ is the overall acoustic transmission matrix

$$\mathcal{M} = \mathcal{G}_1\Phi_2 t_3\Phi_4\mathcal{C}_5\Phi_6 t_7\Phi_8\mathcal{G}_9 \tag{101}$$

The vector $\mathbf{W}_7$ is next found from $\mathbf{W}_9$,

$$\mathbf{W}_7 = \Phi_8\mathcal{G}_9\mathbf{W}_9 \tag{102}$$

Finally, from eq. (30) the electrical output amplitude b_7 is given by

$$b_7 = \tau_7' \cdot \mathbf{W}_7 \tag{103}$$

The analysis leading up to eq. (103) is essentially a derivation of a general algorithm for finding the two-port, electrical-transmission characteristics of a grating resonator with an arbitrary coupling element C_5. The algorithm can therefore be applied to single-cavity resonators as well as more complex, multipole structures.

The analysis can be further simplified by considering transducer T_3 in combination with grating G_1 as an "input" coupler described by the matrix $\mathcal{C}^{IN}$

$$\begin{pmatrix} a_3 \\ b_3 \end{pmatrix} = \mathcal{C}^{IN}\begin{pmatrix} w_3^+ \\ w_3^- \end{pmatrix} \tag{104}$$

Similarly, transducer T_7 and grating G_9 form an "output" coupler described by $\mathcal{C}^{OUT}$

$$\begin{pmatrix} w_6^+ \\ w_6^- \end{pmatrix} = \mathcal{C}^{OUT}\begin{pmatrix} b_7 \\ a_7 \end{pmatrix} \tag{105}$$

The overall electrical transfer function is then found from

$$\begin{pmatrix} a_3 \\ b_3 \end{pmatrix} = \mathcal{C}^{IN}\Phi_4\mathcal{C}_5\Phi_6\mathcal{C}^{OUT}\begin{pmatrix} b_7 \\ a_7 \end{pmatrix} \tag{106}$$

For optimal transducer placement and, for simplicity, neglecting R_s and θ_t, $\mathcal{C}^{IN}$ and $\mathcal{C}^{OUT}$ are given by

$$\mathcal{C}^{IN} = \frac{-(-1)^{N_t}}{\sqrt{2g}} \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \tag{107}$$

and

$$\mathcal{C}^{OUT} = \frac{1}{\sqrt{2g}} \begin{bmatrix} c_{11} & -c_{21} \\ -c_{12} & c_{22} \end{bmatrix} \tag{108}$$

where

$$c_{11} = g + \frac{1 + j\theta_e}{1 + r} \tag{109}$$

$$c_{12} = -g + \frac{(1 + j\theta_e)r}{1 + r} \tag{110}$$

$$c_{21} = -g + \frac{1 - j\theta_e}{1 + r} \tag{111}$$

$$c_{22} = g + \frac{(1 - j\theta_e)r}{1 + r} \tag{112}$$

$$g = G_r Z_e$$

$$r = j\Gamma$$

and Γ is the reflection coefficient of the appropriate grating (G_1 or G_9).

REFERENCES

1. D. T. Bell and R. C. M. Li, "Surface-Acoustic-Wave Resonators," Proc. IEEE, *64* (May 1976), pp. 711–721.
2. C. S. Hartmann and R. C. Rosenfeld, U.S. Patent No. 3,886,504, May 1975.
3. H. A. Haus and R. V. Schmidt, "Transmission Response of Cascaded Gratings," IEEE Trans. Son. Ultrason., *SU-24* (March 1977), pp. 94–101.
4. P. S. Cross, R. V. Schmidt, and H. A. Haus, "Acoustically Cascaded SAW Resonator Filters," 1976 IEEE Ultrason. Symp. Proc., September 1976, pp. 277–280.
5. E. J. Staples, Proc. 30th Annual Symposium on Frequency Control, U. S. Army Electronics Command, Fort Monmouth, New Jersey, June 1976, pp. 322–327.
6. R. L. Rosenberg and L. A. Coldren, "Reflection-Dependent Coupling Between Grating Resonators," 1976 IEEE Ultrason Symp. Proc., September 1976, pp. 281–286.
7. M. Redwood, R. B. Topolevsky, R. F. Mitchell, and J. S. Palfreeman, "Coupled-Resonator Acoustic-Surface-Wave Filter," Electron. Lett., *11*, No. 12 (June 12, 1975), pp. 253–254.
8. P. S. Cross, R. S. Smith, and W. H. Haydl, "Electrically Cascaded Surface-Acoustic-Wave Resonator Filter," Electron. Lett., *11*, No. 11 (May 29, 1975), pp. 244–245.
9. W. R. Shreve, "Surface-Wave Two-Port Resonator Equivalent Circuit," 1975 IEEE Ultrasonics Symp. Proc., September 1975, pp. 295–298.
10. G. L. Matthaei, B. P. O'Shaughnessy, and F. Barman, "Relations for Analysis and Design of Surface Wave Resonators," IEEE Trans. Son. Ultrason. *SU-23*, March 1976, pp. 99–107.
11. H. Kogelnik, "Coupled Wave Theory for Thick Hologram Gratings," B.S.T.J., *48*, No. 9 (November 1969), pp. 2909–2947.

12. H. Kogelnik and C. V. Shank, "Coupled-Wave Theory of Distributed Feedback Lasers," J. Appl. Phys., *43*, No. 5 (May 1972), pp. 2327–2335.
13. H. A. Haus and C. V. Shank, "Antisymmetric Taper of Distributed Feedback Lasers," IEEE J. Quantum Electron., *QE-12*, No. 9 (September 1976), pp. 532–538.
14. R. V. Schmidt, "Acoustic Surface Wave Velocity Perturbations in $LiNbO_3$ by Diffusion of Metals," Appl. Phys. Lett., *27*, No. 1 (July 1, 1975), pp. 8–10.
15. R. C. M. Li, R. C. Williamson, D. C. Flanders, and J. A. Alusow, "On the Performance and Limitations of the Surface-Wave Resonator Using Grooved Reflectors," 1974 IEEE Ultrason. Symp. Proc., November 1974, pp. 257–262.
16. R. C. M. Li, J. A. Alusow, and R. C. Williamson, "Experimental Exploration of the Limits of Achievable Q of the Grooved Surface Wave Resonator," 1975 IEEE Ultrason. Symp. Proc., September 1975, pp. 279–283.
17. L. A. Coldren and R. L. Rosenberg, "Scattering Matrix Approach to SAW Resonators," 1976 IEEE Ultrason. Symp. Proc., September, 1976, pp. 266–271, and L. A. Coldren, "Characteristics of Surface Acoustic Wave Resonators Obtained from Cavity Analysis," IEEE Trans. Son. Ultrason, May 1977, pp. 212–217.
18. R. C. M. Li, J. A. Alusow, and R. C. Williamson, "Surface-Wave Resonators Using Grooved Reflectors," Proc. 29th Annual Freq. Control Symp., U.S. Army Electronics Command, Ft. Monmouth, New Jersey, May, 1975, pp. 167–176.
19. L. Storch, "The Transmission Matrix of N Alike Cascaded Networks," AIEE Trans. (Communications and Electronics), *73*, January 1955, pp. 616–618.
20. W. H. Haydl and P. S. Cross, "Fine Tuning of Surface-Acoustic-Wave Resonator Filters with Metallization Thickness," Electron. Lett., *11*, No. 12 (June 12, 1975), pp. 252–253.
21. E. J. Staples, J. S. Schoenwald, R. C. Rosenfeld, and C. S. Hartmann, "UHF Surface Acoustic Wave Resonators," 1974 IEEE Ultrason. Symp. Proc., November 1974, pp. 245–252.
22. G. L. Matthaei, F. Barman, E. B. Savage, and B. O. O'Shaughnessy, "A Study of the Properties and Potential Application of Acoustic-Surface-Wave Resonators," 1975 IEEE Ultrason. Symp. Proc., September 1975, pp. 284–289.
23. W. R. Smith, H. M. Gerard, J. H. Collins, T. M. Reeder, and H. J. Shaw, "Analysis of Interdigital Surface Wave Transducers by Use of an Equivalent Circuit Model," IEEE Trans. Microw. Theory Tech., *MTT-17*, November, 1969, pp. 850–873.
24. K. M. Lakin and T. R. Joseph, "Surface Wave Resonators," 1975 IEEE Ultrason. Symp. Proc., September 1975, pp. 269–278.
25. P. S. Cross, "Properties of Reflective Arrays for Surface Acoustic Resonators," IEEE Trans. Son. Ultrason. *SU-23*, No. 4 (July 1976), pp. 255–262.
26. A. Ashkin, G. D. Boyd, and J. M. Dziedzic, "Resonant Optical Second Harmonic Generation and Mixing," IEEE Jour. Quantum Electron., *QE-2*, No. 6 (June 1966), pp. 109–124.
27. R. N. Ghose, *Microwave Circuit Theory and Analysis*, New York: McGraw-Hill, 1963, Chap. 10.
28. P. S. Cross and H. Kogelnik, "Sidelobe Suppression in Corrugated-Waveguide Filters," Opt. Lett., *1*, No. 1 (July 1977), pp. 43–45.
29. R. V. Schmidt and P. S. Cross—unpublished work.
30. W. H. Haydl, B. Dischler, and P. Hiesinger, "Multimode SAW Resonators—A Method to Study the Optimum Resonator Design," Proc. 1976 IEEE Ultrason. Symp., September 1976, pp. 287–296.
31. G. L. Matthaei, D. Y. Wong, and B. P. O'Shaughnessy, "Simplifications for the Analysis of Interdigital Surface-Wave Devices," IEEE Trans. Son. Ultrason., *SU-22*, No. 2 (March 1975), pp. 105–114.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 8, October 1977
Printed in U.S.A.

A Quasioptical Feed System for Radioastronomical Observations at Millimeter Wavelengths

By P. F. GOLDSMITH

(Manuscript received March 1, 1977)

We describe a quasioptical feed system for use with a 7-meter Cassegrain antenna at millimeter wavelengths. This system is designed to take full advantage of low noise, broadband mixer receivers and will be used for radioastronomical observations at frequencies between 60 GHz and 140 GHz. Two offset parabolic mirrors couple the radiation from the $f/D = 5.7$ antenna into the receiver feedhorn. A Fabry-Perot resonator operating at oblique incidence is used to inject the local oscillator energy into the signal path and to suppress response at the image frequency. The loss of the Fabry-Perot diplexer is 0.25 dB for the signal, while the coupling loss between the mixer waveguide flange and the main lobe of the antenna pattern should be ≤ 1 dB.

I. INTRODUCTION

For optimal use of an antenna for radio astronomy at millimeter wavelengths, the feed system should provide a number of functions and must satisfy a variety of stringent performance criteria. These include

(*i*) Low loss for the signal over an instantaneous bandwidth of ≥ 500 MHz.

(*ii*) A well-controlled antenna illumination pattern which should remain unchanged over as large a range of frequencies as possible.

(*iii*) A provision for making accurate absolute calibrations of the receiver gain and atmospheric attenuation—both of these require suppression of the image frequency response in systems incorporating mixers.

(*iv*) A facility for antenna beam switching at a rapid rate to minimize the sky-noise contribution to receiver noise.

(*v*) Since mixers are currently the dominant type of receiver at

frequencies between 60 GHz and 300 GHz, it would be advantageous to include local oscillator injection as part of the feed system if this can be done with low loss.

The present feed system has been designed to satisfy all of the preceeding requirements. In Section II we describe the feed system optics and analyze the measurements of system performance. In Section III we discuss various aspects of the Fabry-Perot diplexer including bandwidth, image rejection, local oscillator noise suppression, and loss for the signal and for the local oscillator. In Section IV we discuss the calibration system.

II. FEED SYSTEM OPTICS

2.1 Antenna

This feed system is designed to operate with the recently completed Bell Laboratories millimeter antenna located at Holmdel, N.J. The antenna is an offset Cassegrain with a diameter of 7 meters and a f/D ratio of 5.7. The overall surface accuracy is approximately 0.01 cm rms, allowing operation with a moderately high beam efficiency at frequencies as high as 300 GHz. The main advantage of the offset Cassegrain design is that there is zero aperture blockage, and a very low reflection coefficient and low sidelobe levels can be achieved.[1]

2.2 Gaussian beam theory

We shall discuss the feed system optics in terms of the propagation of a single gaussian mode. As discussed by Arnaud,[2] a gaussian beam propagating in free space has a power distribution perpendicular to the direction of propagation (taken to be the z axis) of the form

$$\frac{P(r)}{P(o)} = e^{-[r/\xi(z)]^2} \tag{1}$$

The beam half-width (radius) ξ depends on z, the distance along the axis of propagation, as

$$\xi^2(z) = \xi_o^2 + \left(\frac{z}{k_o \xi_o}\right)^2 \tag{2}$$

where ξ_o is the minimum beam half-width (beam waist radius), taken to be located at $z = 0$, and $k_o = 2\pi/\lambda$. The asymptotic angle of beam half-width growth is seen from eq (2) to be

$$\theta_\xi = 1/k_o \xi_o \tag{3}$$

Equations (1) to (3) apply to gaussian beams of infinite transverse extent. In any practical system the beam will be truncated at some level,

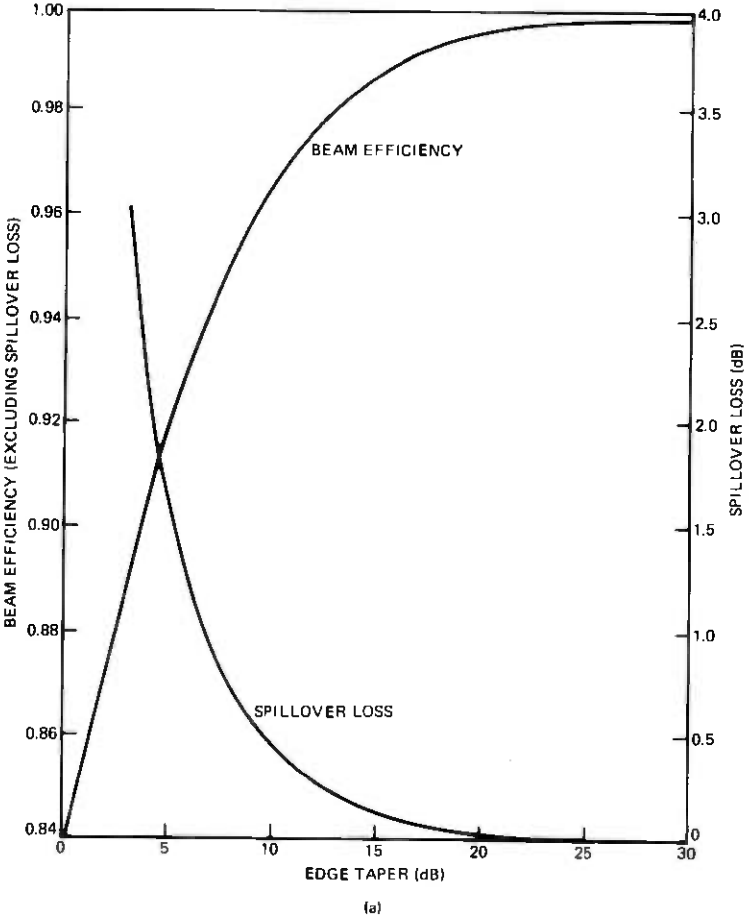


Fig. 1—(a) Beam efficiency and spillover loss for an unblocked, ideal antenna with gaussian aperture illumination, as a function of the edge taper. The edge taper is defined as the power density at the center of the antenna divided by the power density at the edge of the antenna. (b—next page) Beamwidth (full width at half maximum) for the same conditions. The radius of the antenna is a, and $k_0 = 2\pi/\lambda$.

which will produce sidelobes. In considering at what level the beam at the main reflector should be truncated, we have to balance consideration of spillover loss, sidelobe levels, and beam efficiency[3] against those of beamwidth and on-axis gain. Figure 1a shows the spillover loss and beam efficiency while Fig. 1b shows the beamwidth as a function of edge taper for an antenna with a gaussian aperture illumination pattern. The edge taper is defined as the power density at the center of the antenna divided by the power density at the edge. We have chosen an edge taper T_M close to 14 dB as being a satisfactory compromise. All other optical elements in the feed system truncate the beam at a much lower level (at least 23

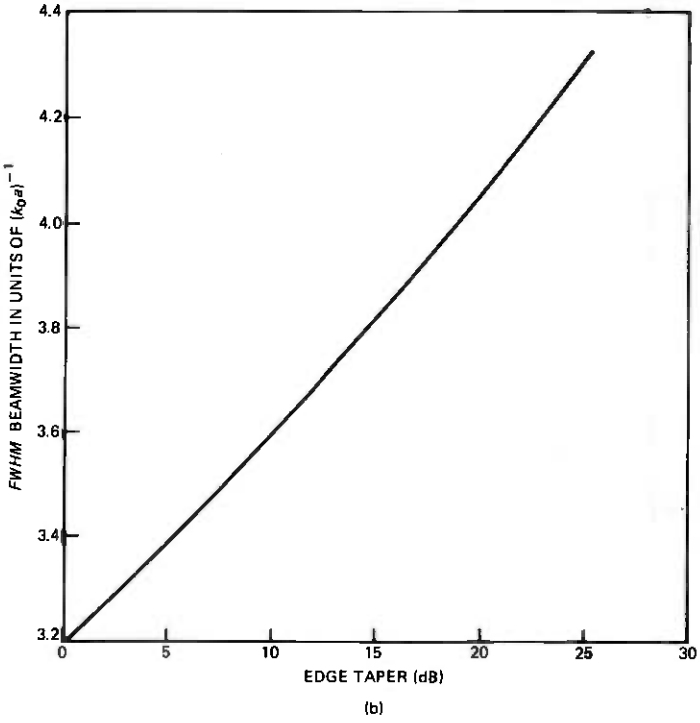


Fig. 1. (*continued from previous page*)

dB below the on-axis power level). We will thus ignore the effects of beam truncation within the feed system.

The edge taper at the main reflector is related to ξ_A, the antenna illumination beam half-width, by the formula

$$\xi_A = a\sqrt{\frac{10}{T \ln 10}} \tag{4}$$

where a is the main reflector radius (350 cm for this antenna) and T is the edge taper in decibels. We find that $\xi_A = 195$ cm for $T = 14$ dB. Since ξ_A is much larger than ξ_o, eq. (2) reduces to

$$\xi_A \approx z_A \theta_\xi = \frac{f}{k_o \xi_o} \tag{5}$$

where f is the focal length of the antenna (3955 cm). The resulting value for ξ_o at 100 GHz is 0.97 cm.

2.3 Feed system components

The large f/D ratio and resulting large beam waist size of the antenna makes coupling to the antenna beam waist directly with a feedhorn

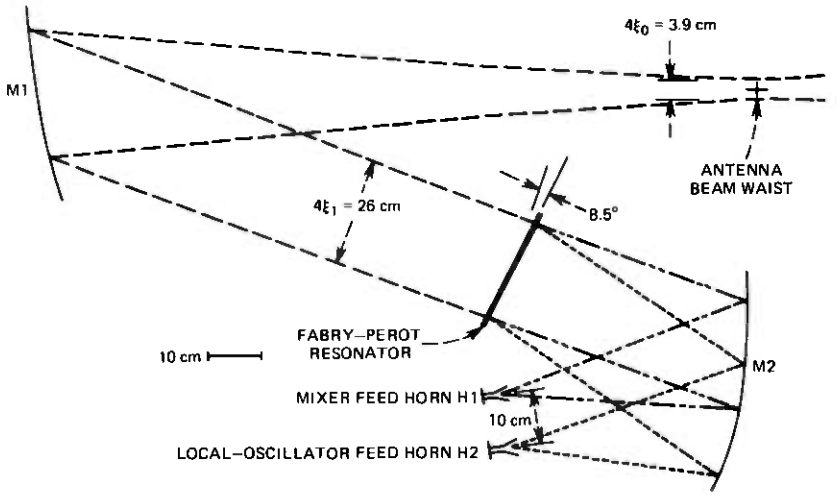


Fig. 2—Feed system optics. M1 and M2 are offset paraboloids. The diplexing action of the Fabry-Perot resonator (tilted 8.5 degrees from normal incidence) is shown schematically.

undesirable, especially for cryogenic receivers. Horn-lens arrangements were investigated but the losses involved were felt to be a significant disadvantage, especially when operation over very large bandwidths is required. In view of these facts, and also because of the desirability of an even larger beam waist size required for low loss in the Fabry-Perot diplexer (Section III), a feed system using metal mirrors is preferable. The arrangement of the feed system components is shown in Fig. 2. The overall size of the feed system is dictated by the beam waist size and the desire to minimize the number of mirrors involved.

Mirrors M1 and M2 are offset paraboloids; the offset angle for M1 is 20 degrees and the focal length is 136 cm. For M2 the offset angle is 30 degrees for the signal beam and the focal length is 44 cm. Offset antennas of this type have been shown to have excellent beam patterns.[4] The mirrors used in this work were cut on a numerically controlled milling machine; the peak deviation from the desired surface contour is approximately 0.05 mm.

The beam from the antenna expands until it reaches M1; at this point the beam half-width, denoted ξ_1, is 6.5 cm and is essentially frequency-independent. The distance from the beamwaist to M1 is equal to the focal length of the mirror so that in the geometrical optics limit the resulting beam would be collimated. The diffraction effects in the beam between M1 and M2 are small; in actuality a second beam waist is created in the large beam at a distance equal to the focal length from M1. Ideally, the separation between M1 and M2 would be equal to the sum of their focal lengths (180 cm) but a calculation[5] of the mismatch due to the

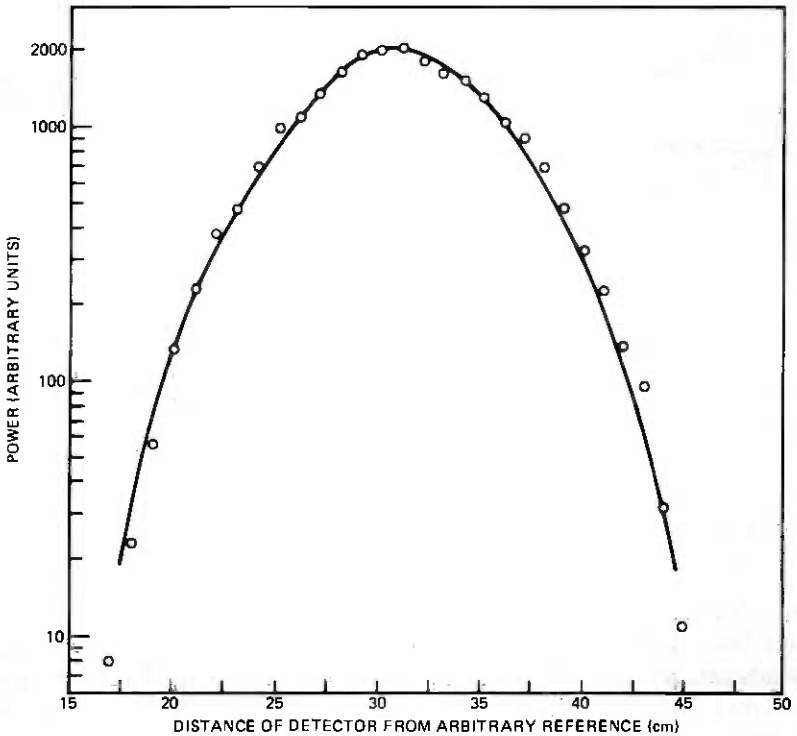


Fig. 3—Profile of beam in region between M1 and M2. The axis of the scan is perpendicular to the plane of the components in Fig. 2, and passes through the axis of the beam. Also shown is a gaussian beam with a beam half-width equal to 6.5 cm.

distance being only 140 cm indicates that this is an insignificant effect.

The difficulty in measuring the power distribution in the beam at the antenna beam waist can be overcome by utilizing the properties of a gaussian beam focused by lenses or mirrors; the beam half-width in the focal plane on one side of a converging lens with focal length f will be related to the beam-waist radius on the other side by[6]

$$\xi_{fp} = \frac{f}{k_o \xi_o} \tag{6}$$

In Fig. 3 we show a profile of the beam in the collimated region measured with a small-aperture (0.4 cm × 0.6 cm) horn and square-law detector. This measurement, which is well-fitted by a guassian with $\xi_1 = 6.5$ cm, together with eq. (6) confirms that the beam-waist size at 100 GHz is 1.0 cm, very close to the design value.

A signal passing through the Fabry-Perot resonator is focused by M2 into the feed horn attached to the mixer, located at the beam waist of

M2. The beam-waist radius at the feed horn is 0.32 cm at 100 GHz. The utilization of the Fabry-Perot with a diplexing angle of 8.5 degrees and M2 focal length equal to 44 cm requires that the dimension of M2 in the plane of the paper in Fig. 2 be approximately twice as large as would be required for focusing the signal beam alone.

The feedhorn for the receiver, which is the same design as that for the local oscillator, is a corrugated horn[7] with a beamwidth between −17 dB power points of 29 degrees. This type of feedhorn allows waveguide-bandwidth (90 to 140 GHz for the initial version) operation with high efficiency and very low sidelobes. For system tests performed at frequencies near 100 GHz we have, however, used relatively narrowband dual-mode horns[8,9] constructed in a manner similar to those described in Ref. 4. The power patterns are very similar to those of the corrugated horns, although with a beamwidth approximately 10 percent larger. All feed system characteristics refer to those measured with the dual-mode horns, but these should differ only in minor ways from those obtained with the corrugated horns.

2.4 Measurements of feed system efficiency

As discussed in the previous section, measurements of the power distribution in the collimated region indicate that the feed system will produce the correct taper in the illumination of the main antenna. In order to mesure the efficiency of the feed system, a separate collector was placed at the beam waist of M1, corresponding to the antenna beam waist. This collector, consisting of an ellipsoidal reflector and dual-mode feed horn, was independently measured to have a gaussian angular response pattern corresponding to a beam-waist size of 0.99 cm. A 100-GHz klystron with ~50 dB attenuation was used as a signal source. By interchanging a square law detector between the signal-source flange and the collector output flange (with the signal source connected to the feed system mixer flange), we determined the loss of power between the signal source and the collector output flange to be 1.1 dB. It should be noted that if part of this loss is due to the mode produced by the feed system not coupling to that accepted by the collector, this will not necessarily lower the efficiency when used with the antenna, but will only result in an illumination function slightly different from that anticipated. Thus the loss measured in this manner is an upper limit to the loss when used with an antenna. While the losses of the individual elements cannot easily be measured separately, the symmetry of the system suggests that half of the measured loss is due to the collector, and half is in the feed system, with a resulting feed system loss of 0.5 dB.

In Table I we summarize the salient characteristics of the feed system.

Table I — Feed system characteristics at 100 GHz

Characteristic	Value
ξ_1, collimated beam-waist radius to $1/e$ power point	6.5 cm
ξ_0, beam-waist radius to $1/e$ power point at antenna beam waist	0.97 cm
T_M, edge taper at main reflector	14.1 dB
θ_{FWHM}, full angular beamwidth to half-power points	1'.8
First sidelobe level relative to on-axis gain	−30 dB
ϵ_F, feed system loss (mixer waveguide flange to antenna beam waist)	0.5 dB
Spillover loss	0.14 dB
ϵ_M, beam efficiency	0.95

III. QUASIOPTICAL DIPLEXER

3.1 Introduction

The limited local-oscillator output power available at shorter millimeter wavelengths and the difficulty of fabricating low-loss waveguide diplexers[10] are incentives to seek an alternative to injection cavities and directional filters made in waveguide that are currently available. The use of a Fabry-Perot resonator as a diplexer is not new,[11,12] but the realization of a very low loss device to combine two signals differing in frequency by ~5 percent puts stringent restrictions upon the design of the resonator. There are a variety of configurations in which a Fabry-Perot resonator be used as a diplexer, e.g., with the signal in transmission or in reflection. A desirable characteristic of an ideal diplexer would be the ability to transmit power at the frequency of either one or both mixer sidebands. Single-sideband operation is important for accurate calibrations at millimeter wavelengths because the atmospheric attenuation in certain regions of the spectrum is a rapidly varying function of frequency.[13,14,15] Thus, although data analysis procedures have been developed which attempt to circumvent this problem,[16] the fact remains that an accurate determination of atmospheric extinction for spectral line work requires measurement of the attenuation in the sideband in which the line of interest is located. Also, the gain of a mixer receiver may well be different in the two sidebands, especially with the relatively high IF frequencies (4 to 5 GHz) that are now in use. For these reasons, systems have previously been devised which incorporate a Fabry-Perot resonator which either can be inserted in the optical path to measure the gain and attenuation in the two sidebands individually[17] or is permanently placed in front of the feed horn and which, at the expense of a small loss (~0.4 dB), suppresses the mixer response to the unwanted sideband.[18] In order to minimize the number of resonant elements and consequent adjustments required when changing frequencies, we decided

Table II — Characteristics of Fabry-Perot resonator at 100 GHz

T*	Image rejection ratio (dB)	0.5-dB bandwidth (MHz)	1-dB bandwidth (MHz)
0.10	26	220	320
0.15	22	350	500
0.20	19	540	800
0.25	17	620	890
0.30	15	760	1100
0.40	12	1090	1600
0.50	9.5	1500	2200

* T is the transmission of a single mirror.

to use the Fabry-Perot resonator in transmission for the signal (the local oscillator is reflected by the resonator, thus providing the diplexing action). This design allows us either to operate in a double-sideband mode with the two sidebands being transmitted in successive orders (for continuum work) or in a single-sideband mode (desirable for spectral line observations). Only one adjustment is required to set the diplexer for operation at a particular frequency, which proves to be a significant advantage in use.

3.2 Fabry-Perot resonator theory

The analysis of the propagation in a noninfinite Fabry-Perot resonator has been treated by Arnaud et al.[11] Since we will be dealing with a strongly tapered beam, it is sufficient to use the standard formulas for a plane wave in a resonator of infinite transverse dimension to calculate the response. Neglecting absorption in the mirrors, we find[19] that the fraction of the incident power transmitted by the resonator is given by

$$\tau = \frac{1}{1 + \dfrac{4(1-T)}{T^2} \sin^2 (k_o d \cos \theta)} \tag{7}$$

where d is the distance between the mirrors, θ is the angle from normal incidence of the radiation, T is the power transmission of a single mirror, and we have set the phase of the reflection coefficient equal to π which causes no loss of generality. In this limit we see that the peak transmission (for $k_o d \cos \theta = n\pi$, n being the order of operation) is equal to unity. The peak-to-valley ratio, or contrast factor, which will in our case be the image rejection ratio, and the 0.5-dB and 1-dB bandwidths for a resonator operating at 100 GHz are given in Table II as a function of T, which is assumed to be frequency-independent. It also has been assumed that the free spectral range of the resonator is approximately equal to 4 ν_{IF}; this is not a severe restriction since the transmission is only weakly dependent on frequency near the transmission minimum. There is a

tradeoff between bandwidth and image reflection, as expected for a simple resonator. This restriction could be eased by using a multiple-mirror resonator, but only at the expense of easy tunability. Efficient utilization of the bandwidth of available IF amplifiers (~600 MHz) indicates that T should not be less than 0.2; the resulting image rejection ratio of 19 dB is certainly adequate to assure proper calibration accuracy. It should be pointed out, however, that this ratio is not so high that the leakage of very strong lines from the opposite sideband in a high-sensitivity spectrogram can be entirely ruled out.

The Fabry-Perot diplexer exhibits quite high directivity for local oscillator injection. Power coming from the local oscillator feed horn that directly leaks though the Fabry-Perot resonator does not end up in the beam waist area at all, and is caught by a sheet of absorbing material. Only local oscillator power which is reflected from the Fabry-Perot, then reflected from the mixer feed horn, and which is finally transmitted by the resonator, can reach the calibration area; the level of this radiation should be at least 17 dB below that of the local oscillator power reaching the mixer.

The loss in a Fabry-Perot resonator operated at oblique incidence is primarily due to a walk-off effect in the finite-sized beam.[11] In this reference, the peak fractional transmission τ through a resonator (assumed to be much larger than the beam size) consisting of two mirrors of transmission T, spacing d, inclined at an angle θ to a gaussian beam of beamwaist radius ξ_o, is given by

$$\tau = 1 - G^2$$

where

$$G = \frac{2d \sin \theta}{\xi_o T} \tag{8}$$

For operation with $\nu_{IF} = 5$ GHz and $\nu_{SIGNAL} = 100$ GHz, obtaining the best image rejection ratio requires that the resonator be operated in fifth order so that $d = 5\lambda/2 = 0.75$ cm. The exact spacing will be determined by the resonance condition for the signal frequency; the condition $4\nu_{IF} = \nu_{SIGNAL}/5$ will be satisfied only approximately, but d will be close to the value given above. For $T = 0.2$ we find for small angles $\tau = 1 - (7.5\ \theta/\xi_o)^2$.

A lower limit on θ of $\sim 4/k_o\xi_o$ is found[11] from the condition that the beams be separable at the −17-dB level when the diffraction of each is considered. Thus the maximum transmission is (again for $T = 0.2$, $d = 0.75$ cm)

$$\tau_{MIN\theta} = 1 - \left(\frac{30}{k_o \xi_o{}^2}\right)^2 \tag{9}$$

As seen from eq. (8) the insertion loss, defined in decibels as $10 \log_{10} \tau^{-1}$, can, in theory, be made as low as desired, at the expense of enlarging the beam-waist radius. The beam waist required for low loss even in the optimum situation [eq. (9)] is moderately large; at $\nu = 100$ GHz and for the above conditions, $\xi_o = 2.6$ cm is required to achieve an insertion loss of 0.2 dB (the beam diameter will be at least $4\xi_o$). The most straightforward geometry (see, for example, Ref. 11) then results in a very large distance between the Fabry-Perot and the inputs for the signal and local oscillator; on the order of 1 meter for the above conditions. For this reason, and due to the simplicity of having the one mirror (M2) serve as collector for both the mixer and the local oscillator, the geometry of Fig. 2 was adopted. With a room temperature mixer, it would not be difficult to achieve a diplexing angle close to the theoretical minimum for a given loss, since the diameter of a dual-mode or corrugated feed is approximately 5 times the beam-waist diameter of the beam it launches. With a cryogenic receiver the minimum diplexing angle is set by the size of the dewar containing the mixer; we have used $\theta = 8.5$ degree (0.148 radian). To obtain an insertion loss of 0.15 dB the required beam-waist radius is approximately 6 cm; this number sets the size of the various mirrors and the focal length of M1, as well as the size of the Fabry-Perot resonator. The Fabry-Perot is shown in Fig. 4. In principle, one could utilize the minimum diplexing angle required for a given loss and collect the two spatially separated beams by mirrors which would refocus the beams wherever desired (i.e., into a dewar). This approach was not adopted because of alignment difficulties associated with the additional mirrors involved.

3.3 Measurements

3.3.1 Fabry-Perot mirrors

Each Fabry-Perot mirror consists of a photoetched copper mesh stretched on a metal support ring; the latter is similar to those described by Wannier et al.[18] The theory of one-dimensional wire grids[20] indicates that for the wires parallel to the electric field the grid behaves as a shunt inductance. We expect that a grid with square apertures will behave as a polarization-independent reflector as long as the angle of inclination of the incident beam is small.[21] For these grids with period $p = 1.07$ mm, strip widths $s = 0.29$ mm, and grid thickness $t = 0.08$ mm, one expects the relatively large value of t/s to decrease the equivalent inductance and thus decrease the transmission, compared to that of an infinitely thin grid.[20] The measured transmission at an incidence angle of 8.5 degrees is 0.19 ± 0.02 (at $\nu = 100$ GHz) compared to a transmission of 0.13 predicted theoretically; for an infinitely thin grid with the same aperture parameters, the theoretical transmission is 0.30.

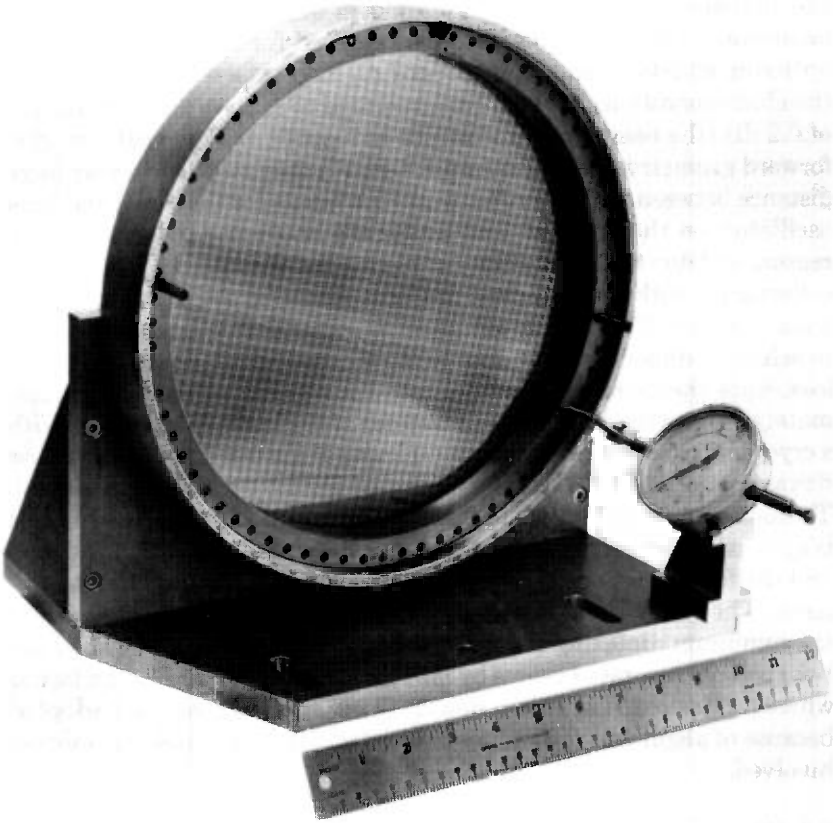


Fig. 4—The Fabry-Perot resonator. The dial indicator on the right is used to monitor the mirror separation.

3.3.2 Fabry-Perot resonator

Examples of the frequency response of the Fabry-Perot resonator are shown in Fig. 5. These curves were obtained by sweeping a Siemens RWO 110B BWO connected to the mixer horn flange and monitoring the output from the collector located at the beam waist of M1. A measurement system consisting of a digitizer, log amplifier, and 1024 channel memory (Pacific Measurements model 1038) was used to first record the output without the Fabry-Perot. We then used this to correct the output measured with the Fabry-Perot in place for frequency-dependent variations in the oscillator output. The following parameters are obtained from these scans:

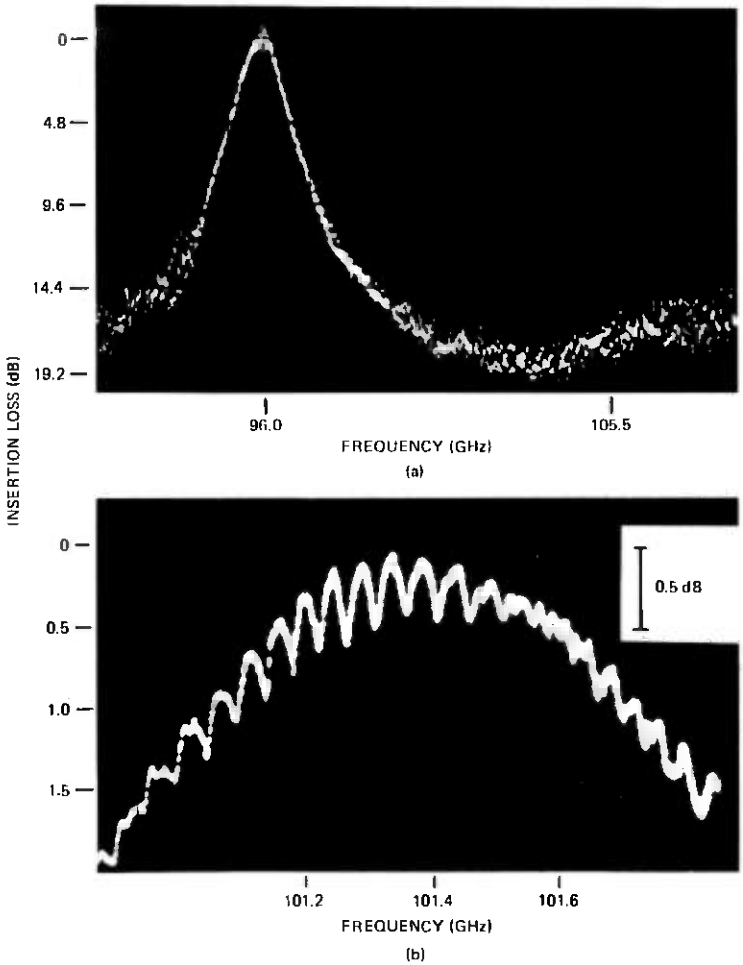


Fig. 5—(a) Transmission of the Fabry-Perot resonator as a function of frequency. The nominal value of 5 dB per vertical division was determined to be 4.5 dB from measurements with a precision attenuator. (b) Response near the transmission maximum, for a different mirror separation. Each vertical division corresponds to 0.5 dB. The ripple pattern is characteristic of the separation between the transmitter and receiver feed horns used in making the measurement.

Image rejection ratio = 19 dB
0.5-dB bandwidth = 510 MHz
1-dB bandwidth = 780 MHz
Minimum insertion loss = 0.25 dB (10)

This last number is obtained by averaging over the ripple pattern in the central 250 MHz of the response pattern. The results presented here,

when compared to those given in Table II, indicate that the image rejection ratio measurement is consistent with a mirror transmission of 0.2, while the bandwidth measurements imply a transmission of about 0.21. The minimum resonator loss predicted by a mirror transmission of 0.2, $\theta = 8.5$ degrees, $\xi_o = 6.5$ cm, and $d = 0.75$ cm is 0.13 dB. If we allow for a loss of 0.12 dB from ohmic dissipation and/or other losses in the resonator, all of these measured characteristics are consistent within the errors with the expected resonator performance assuming a mirror transmission of 0.2.

3.3.3 Local oscillator loss

From the response curve of the Fabry-Perot (Fig. 5a), we see that the fraction of the local oscillator power leaking through the resonator will be only a few percent. If, for the moment, we consider the local oscillator injection process in reverse, we see that the mixer feed horn would produce essentially a plane wave heading towards M2, after reflection from the Fabry-Perot. In this case, the M2-local oscillator feedhorn combination should be considered as an off-axis offset parabolic antenna. The diplexing angle $\theta = 8.5$ degrees requires that the local oscillator feedhorn be 17 degrees or 24 half-power beamwidths off-axis. For a symmetric antenna with the same f/D ratio, the loss in gain would be less than 0.4 dB.[22] For an offset antenna, the theoretical loss is approximately 4 dB.[23] The measured loss for transmission between the flange of the local oscillator feed horn and that of the mixer feed horn is 2.7 dB. This is somewhat better than that achieved with a waveguide directional filter,[24] and far superior to results obtained with waveguide injection cavities.[25] If the diplexing angle were reduced by only a factor of two, the theoretical loss would be less than 1 dB.

3.3.4 Local-oscillator noise suppression

The Fabry-Perot diplexer as used here provides only 3 dB suppression of local oscillator noise since noise power at the image frequency is coupled into the mixer essentially as efficiently as power at the nominal local oscillator frequency. At an IF frequency of 5 GHz, a 3-dB filtering of the local-oscillator noise from a 100-GHz reflex klystron is sufficient to reduce the local oscillator noise to the equivalent of a 20 K input signal as measured with a single-ended mixer.[24] This is consistent with our measurements, in which we were unable to measure any increase in the diode noise temperature[26] using the Fabry-Perot diplexer, compared to using a high-Q injection cavity, with equal bias voltages and diode currents with the local oscillator on. In any case, local oscillator noise can easily be further reduced by a simple bandpass filter installed in the local oscillator waveguide.

3.3.5 Mixer performance

It is difficult to accurately measure the effect of the quasioptical diplexer on mixer performance, since most mixers when used with an injection cavity or directional filter are sensitive to signals in both sidebands, while with the Fabry-Perot resonator in its usual configuration the mixer in the quasioptical diplexer is sensitive to only one sideband. If we assume that the mixer is equally sensitive in the two sidebands, a comparison can be made. A room-temperature mixer with a transistor IF amplifier, when used with the quasioptical diplexer, was found to have an SSB noise temperature 0.7 dB better than that implied by a double-sideband measurement using an injection cavity diplexer. This same injection cavity was measured to have an insertion loss of 0.74 dB for the signal at 100 GHz while the quasioptical diplexer insertion loss is ~0.25 dB. The difference in noise temperatures is seen to be larger than the difference in diplexer losses, a fact which probably reflects the uncertainty in the relative response in the mixer sidebands. We do conclude, however, that the very low insertion loss for the quasioptical diplexer will probably be reflected in lower system noise temperatures.

3.4 Discussion

The Fabry-Perot diplexer described here exhibits low loss for the signal and for the local oscillator. The metal mesh mirrors actually had a lower transmission (0.2) than was expected (0.25) due to the larger thickness-to-aperture-size ratio compared to lower-frequency grids. Examination of Table II indicates that a mirror transmission of 0.27 might be optimum; this would lower the theoretical loss by a factor of 2. A more elaborate optical system would allow a diplexing angle at least 2 times smaller than that used, which would lower the loss by a factor of 4, or else would allow the beam and resonator diameters to be halved for the same loss. Thus it is seen that this technique has not been pushed to its limit in terms of low loss or compactness.

The use of the Fabry-Perot as a diplexer is also feasible in the submillimeter region. The techniques for making the mirrors are available and have been used to make resonators, operating at wavelengths between 80 μ and 600 μ.[19,27] If the ratio of the IF frequency to signal frequency is held constant, the order of operation of the resonator will remain fixed and the mirror separation will be proportional to the signal wavelength. Then, to obtain a given loss [eq. (8)], the beam size will also be proportional to the wavelength. If, on the other hand, a fixed IF frequency is used, the beam size required to obtain a given loss will be independent of the wavelength.

This quasioptical diplexer is also well suited to dual-polarization applications. The properties of the Fabry-Perot resonator are essentially

polarization independent. Thus, if the polarization angle of the local oscillator feedhorn is rotated 45 degrees to that of the mixer feedhorn, equal amounts of local-oscillator power would be detected in the two polarizations at the mixer feed horn. Either a dual-polarization feed horn or two feed horns with orthogonal polarizations fed by a wire-grid polarization splitter could be utilized.

IV. CALIBRATION SYSTEM

The calibration system shown in Fig. 6 is designed to provide a convenient method of measuring the receiver gain and atmospheric attenuation, and to allow various modes of observation. Each of these functions will be briefly discussed.

4.1 Receiver calibration

Not shown in Fig. 6 is a load consisting of truncated pyramids of Eccosorb* VHP-2 absorber which can be inserted into the beam that has passed from M1 through the rotary chopper. This provides a load at near ambient temperature. A cold load at liquid nitrogen temperatures has been constructed from pyramids of Eccosorb VHP-2 absorber in a dewar of liquid nitrogen. The index of refraction of liquid nitrogen is 1.4 at low frequencies[28] and should not be significantly higher at millimeter wavelengths. The resulting power reflection coefficient is 0.03. The power reflected by the absorber at the bottom of the dewar filled with nitrogen is measured to be approximately 20 dB below that reflected from a metal plate at the bottom of an empty dewar. We thus conclude that cold load is likely to be a moderately good calibration standard; its stability and emissivity have not been measured. By rotating the chopper (with the movable mirror out of the beam) a temperature difference of approximately 210 K is produced. It is possible that for very low noise receivers, this change in total power produced may exceed the limit allowable for good detector linearity. In this event, a calibrated, computer-controlled attenuator will be switched in synchronism with the chopper to keep the total power more nearly constant.

4.2 Measurement of atmospheric attenuation

This function is accomplished by chopping between the sky and either the ambient temperature or the cold load. The choice of reference depends on the sky temperature; the maximum temperature difference of ~100 K will probably be small enough to ensure good detection linearity. The atmospheric attenuation is then computed from an assumed

* Registered trademark of Emerson Cuming Inc., Canton, Mass.

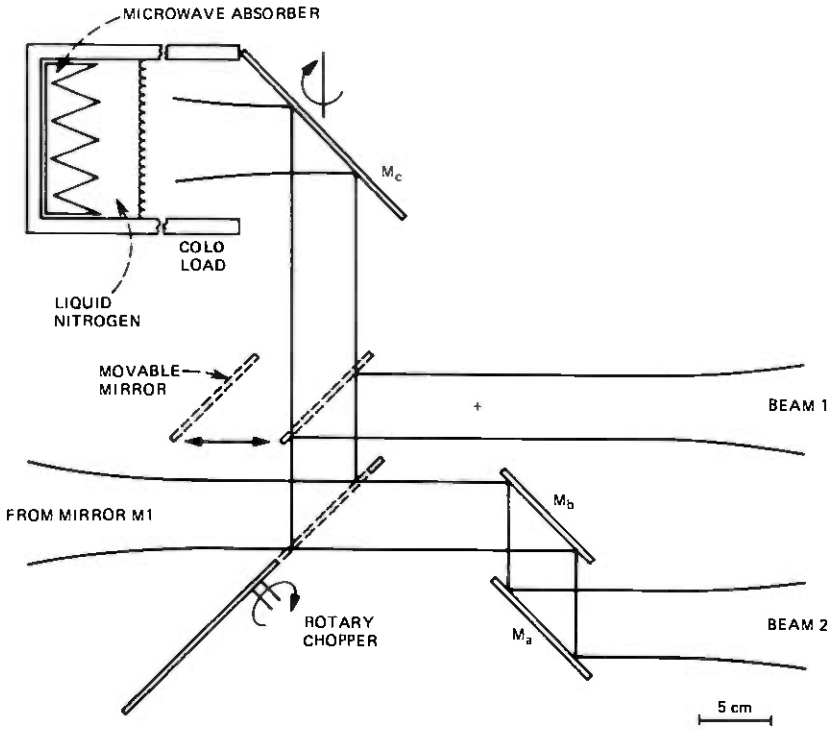


Fig. 6—The calibration system. The cross indicates the location of the antenna beam waist, while the lines shown approximate the −17 dB contours of the power distribution. The view presented is with the antenna pointing at zenith; at other elevation angles the cold load mirror M_c pivots about the axis indicated to keep the surface of the liquid nitrogen parallel to the horizon and perpendicular to the incident beam. Not shown is an ambient-temperature absorber that can be inserted between the chopper and M_b; its position, as well as that of the rotary chopper and movable mirror, is under computer control.

physical temperature (or temperature distribution) for the absorbing gas.

4.3 Beam switching

For observation of moderately small sources this technique is advantageous in that fluctuations in atmospheric emission will cancel if the chopping rate is sufficiently high and the scale size of the inhomogeneities is larger than the beam separation.[29] The separation between the two beams is 13′. This large value will be useful astronomically, but if the separation proves too large for effective noise cancellation, it can easily be reduced to about 6′. The uncertainty in the power spectrum of atmospheric fluctuations has led us to make the chopper speed variable between 2 Hz and 50 Hz. Observational experience will be required to determine the optimum chopping speed at different wavelengths under different atmospheric conditions.

V. SUMMARY

We have designed and tested a feed system for use with millimeter radio-astronomical receivers on a 7-meter Cassegrain antenna. We have measured that power incident on the mixer waveguide flange is transmitted to the antenna beam waist in the desired mode with a loss less than 1.1 dB and probably close to 0.5 dB. The antenna beam efficiency should be 0.95. The feed system incorporates a Fabry-Perot diplexer which has an insertion loss of 0.25 dB (transmission = 0.94) for a signal at 100 GHz and a loss of 2.7 dB for the local oscillator with a frequency differing by 5 GHz. A calibration system incorporates an ambient temperature load and a liquid nitrogen load, and a rotary chopper to switch between the two, between either one and the sky, or between two beams separated by 13′ on the sky.

The low loss and versatility of quasioptical techniques at millimeter wavelengths are expected to prove advantageous in obtaining well-calibrated high-sensitivity astronomical data.

ACKNOWLEDGMENT

I wish to thank J. Arnaud, T. S. Chu, A. A. M. Saleh, and R. W. Wilson for devoting considerable time to many helpful discussions about various aspects of this work. R. A. Linke supplied the mixer used in the tests, and also valuable information about mixer operation. Several coworkers at Crawford Hill generously made the results of their work available before publication. In particular, C. Dragone provided data on offset cassegrain antennas, M. J. Gans supplied data on truncated gaussian beams, and J. T. Ruscio made available the results of his measurements on mesh transmission. F. A. Pelow supervised making the metal mesh mirrors, R. A. Semplak supplied the collector used in measuring the feed-system efficiency, and W. Legg tuned and measured the patterns of the two feed horns. Thanks are also extended to R. L. Plambeck for a variety of helpful suggestions and to A. A. Penzias for encouragement to begin this project.

REFERENCES

1. C. Dragone and D. C. Hogg, "The Radiation Pattern and Impedance of Offset and Symmetrical Near-Field Cassegrainian and Gregorian Antennas," IEEE Trans. Ant. Propag., *AP-22*, May 1974, pp. 472–475.
2. J. A. Arnaud, *Beam and Fiber Optics*, New York: Academic Press, 1976, pp. 50–64.
3. J. D. Kraus, *Radio Astronomy*, New York: McGraw-Hill, 1966, pp. 154–159.
4. M. J. Gans and R. A. Semplak, "Some Far-Field Studies of an Offset Launcher," B.S.T.J, *54*, No. 7 (September 1975), pp. 1319–1340.
5. J. A. Arnaud, op. cit., pp. 74–79.
6. J. A. Arnaud, op. cit., pp. 65–67.
7. A. J. Simmons and A. F. Kay, "The Scalar Feed—A High Performance Feed for Large Parabolid Reflectors" in *Design and Construction of Large Steerable Aerials*, IEE Conf. Pub. 21, 1966, pp. 213–217.
8. P. D. Potter, "A New Horn Antenna With Suppressed Sidelobes and Equal Beamwidths," Microw. J., *6*, June 1963, pp. 71–78.

9. R. H. Turrin, "Dual Mode Small Aperture Antennas," IEEE Trans. Ant. Propag., *AP-15*, March 1967, pp. 307–308.
10. G. T. Wrixon, "Low-Noise Diodes and Mixers for the 1–2mm Wavelength Region," IEEE Trans. Microw. Theory Tech., *MTT-22*, December 1974, pp. 1159–1165.
11. J. A. Arnaud, A. A. M. Saleh, and J. T. Ruscio, "Walk-Off Effects in Fabry-Perot Diplexers," IEEE Trans. Microw. Theory Tech., *MTT-22*, May 1974, pp. 486–493.
12. J. A. Arnaud and F. A. Pelow, "Resonant Grid Quasi-Optical Diplexers," B.S.T.J., *54*, No. 2 (February 1975), pp. 263–282.
13. P. W. Rosenkranz, "Shape of the 5mm Oxygen Band in the Atmosphere," IEEE Trans. Ant. Propag., *AP-23*, July 1975, pp. 498–506.
14. C. J. Gibbins, A. C. Gordon-Smith, and D. L. Croom, "Atmospheric Emission and Attenuation in the Region 85-118 GHz," in *Conference on Propagation of Radio Waves at Frequencies Above 10 GHz*, IEE Conf. Pub. 98, 1973, pp. 132–140.
15. F. T. Ulaby, "Absorption in the 220 GHz Atmospheric Window," IEEE Trans. Ant. Propag., *AP-21*, pp. 266–269, March 1973.
16. J. H. Davis and P. Vandenbout, "Intensity Calibration of the Interstellar Carbon Monoxide Line at λ2.6 mm," Astrophys. Lett., *15*, September 1973, pp. 43–47.
17. R. L. Plambeck, D. R. W. Williams, and P. F. Goldsmith, "Comparison of $J = 2 \rightarrow 1$ and $J = 1 \rightarrow 0$ Spectra of CO in Molecular Clouds," Ap. J. (Letters), *213*, April 1, 1977, pp. L41–45.
18. P. G. Wannier, J. A. Arnaud, F. A. Pelow, and A. A. M. Saleh, "Quasioptical Band-Rejection Filter at 100 GHz," Rev. Sci. Instrum., *47*, January 1976, pp. 56–58.
19. R. Ulrich, K. F. Renk, and L. Genzel, "Tunable Submillimeter Interferometers of the Fabry-Perot Type," IEEE Trans. Microw. Theory Tech., *MTT-11*, September 1963, pp. 363–371.
20. N. Marcuvitz, *Waveguide Handbook*, New York: McGraw-Hill, 1951, pp. 280–289.
21. J. A. Arnaud and F. A. Pelow, opt. cit., pp. 262–264.
22. J. Ruze, "Lateral-Feed Displacement in Paraboloid," IEEE Trans. Ant. Propag., *AP-13*, September 1965, pp. 660–665.
23. C. Dragone, private communication.
24. B. D. Moore and J. R. Cogdell, "A Millimeter Wave Directional Filter Cavity," IEEE Trans. Microw. Theory Tech., *MTT-24*, November 1976, pp. 843–847.
25. R. A. Linke, private communication.
26. S. Weinreb and A. R. Kerr, "Cryogenic Cooling of Mixers for Millimeter and Centimeter Wavelengths," IEEE J. Solid State Circuits, *SC-8*, February 1973, pp. 58–63.
27. D. Brandshaft, R. A. McLaren, and M. W. Werner, "Spectroscopy of the Orion Nebula From 80 to 135 Microns," Ap. J., *199*, July 1975, pp. L115–L117.
28. R. C. Weast, ed., *Handbook of Chemistry and Physics*, Cleveland: CRC Press, 1975, pp. E55–E56.
29. J. W. M. Baars, *Dual Beam Parabolic Antennae in Radio Astronomy*, Groningen: Wolters-Noordhoff, 1970, pp. 59–116.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 8, October 1977
Printed in U.S.A.

An Evaluation of Two Simple Methods for Detecting Tones over Telephone Lines

By D. K. CHRISTOPHER, L. R. RABINER, P. SCHWEITZER, and D. E. BOCK

(Manuscript received March 24, 1977)

An important practical application of signal processing theory is the problem of complex tone detection. Within the telephone plant there often arises a need for a simple, yet efficient, method for detecting the various tones which are used in telephone communication. Two such methods are discussed in this paper. One method uses measurements of the short-time signal energy and makes the decision as to whether or not there is a particular tone present on the line based on the periodicity of the envelope of the signal. This method has application in determining if the energy on the line is periodic or aperiodic where sample examination time is not limited. The second method uses measurements of the short-time zero crossings of the signal. A parallel processing scheme is used to determine if a particular tone is present based on the detailed statistical properties of each of the tones. This method has application in determining if specific frequencies are present, especially when the examination time of the sample is limited. Using a large number of dialed-up connections, both systems were evaluated as to accuracy and speed. Results are presented which show the properties of the two tone detection methods.

I. INTRODUCTION

The need to reliably detect tones arises in a number of systems which are in use within the telephone plant. A wide variety of methods have been proposed for solving this problem including digital filtering, spectral analysis etc.[1–4] In this paper we discuss a particular problem in tone detection and show some characteristics of two simple systems designed to solve this problem.

Figure 1 shows a pictorial description of a simple telephone call in which the calling party initiates a call and the call is switched through a central office. When the called party picks up the telephone a dc path

Fig. 1—Pictorial description of a simple telephone call.

is completed to the central office to indicate that the connection has been made, and that billing of the calling party can begin. The indication is formally called answer supervision. On a certain percentage of calls, anomalies occur such that calls are not completed in the normal manner. A recent survey of a total of 3 million unanswered calls indicated that 90 seconds after the calling party initiated the call, no answer supervision had been received on 61,000 (approximately 2 percent) of the calls. There are several possibilities which account for these seemingly long unanswered calls. These include:

(*i*) Persistent callers—i.e., the calling party is waiting 90 seconds (15 ring cycles) for the called party to answer the telephone. Another possibility is that the line is busy, and the calling party remains on the line in spite of hearing 90 seconds of busy tone. Yet another possibility is that the call was improperly routed and that the calling party is listening to a fast busy or reorder signal.

(*ii*) Announcement service—e.g., the calling party dialed an inoperative number and is listening to an announcement concerning the called telephone. Generally such announcements are short and will not last 90 seconds, but it is not impossible for this to occur.

(*iii*) Defective telephone—i.e., the telephone of the called party is defective or the circuitry which generates the answer supervision signal is not working.

(*iv*) Network irregularities—i.e., the equipment making up the telephone network interconnecting telephone offices may be defective or inoperative and fail to relay the proper answer supervisory signals.

It is important to be able to isolate cases (*iii*) and (*iv*) in order to take appropriate action. Since direct methods for determining whether a telephone is defective or whether a network irregularity exists are not easily implemented, an indirect approach is indicated. The approach studied here was one in which the signal on the line was processed to determine whether or not tones were on the line [case (*i*)], and if not, whether speech could be detected [cases (*ii*)–(*iv*)].

Although sophisticated approaches to tone detection are applicable, such methods are unnecessarily complex (and expensive) for the problem of detecting the three tones described above. Thus two relatively simple, and yet very different, methods for detecting tones were studied. The first method was developed on the assumption that the frequencies of

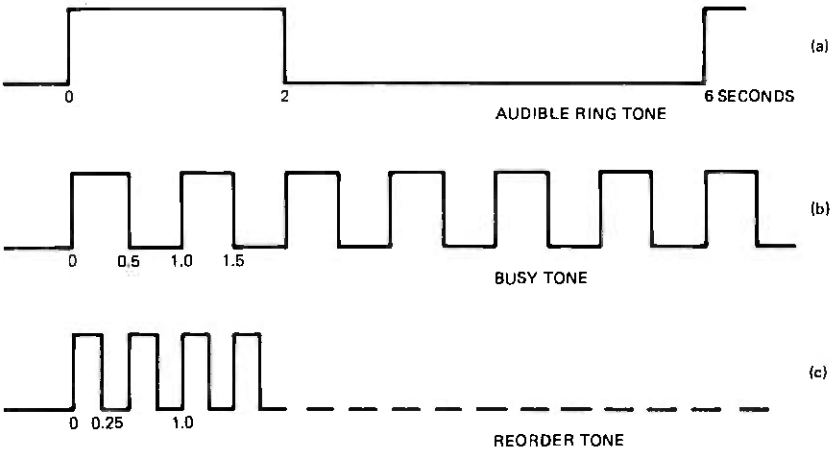


Fig. 2—Temporal characteristics of the tones.

the individual tones would vary greatly from one central office to another and that gathering adequate statistics on these frequencies was impractical. Therefore long-term properties of the tones were used in the detection algorithm. This first approach, known herein as the energy system, was developed by F. T. Boesch and R. E. Thomas[4] to comply with the above constraint. The second method uses measurements of the zero crossing rate of the signal, and used a parallel processing scheme to determine if a particular tone is present based on the assumption that detailed statistical properties were available for each of the tones. This method was developed in the Acoustics Research Department. Descriptions of each of these two methods are given in Sections III and IV. In the next section we discuss the properties of the individual tones for which the two systems were designed. Finally, in Section V we give the results of simulation experiments with both tone-detection methods.

II. PROPERTIES OF THE TONES

The three tones which had to be detected were: (*i*) audible ring tone, (*ii*) line busy tone, (*iii*) reorder tone. Figure 2 shows a sketch of the nominal temporal pattern of these three tones. The audible ring tone is on for 2 seconds and off for 4 seconds. The line busy tone is on for 0.5 second, and off for 0.5 second. The reorder (fast busy) tone is somewhat variable in its temporal pattern. Generally its overall period is 0.5 second, with an on period of from 0.2 to 0.3 second, and an off period from 0.3 to 0.2 second. Nominally the on and off periods are 0.25 second.

Within one period the spectral properties of these tones can vary a fair amount. This is due both to the variability in the mechanical equipment which produces the tones (e.g., motor generators), and to the different

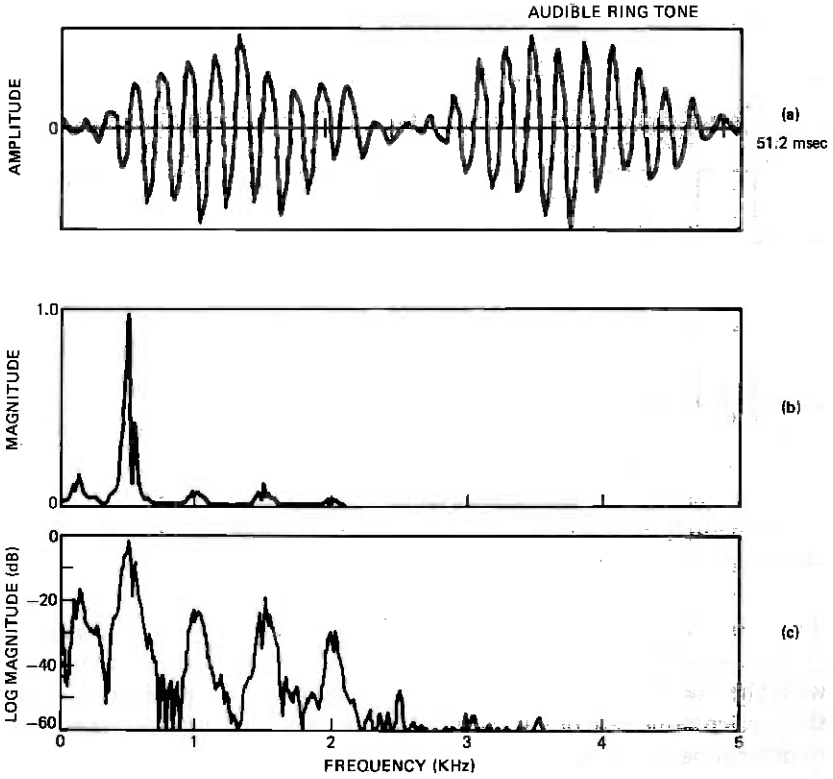


Fig. 3—Spectral characteristics of audible ring tone.

standards in use within the telephone plant. However, the dominant amount of spectral energy in these tones is concentrated in a region around 500 Hz. By way of example, Figs. 3 to 5 show plots of typical 51.2 msec sections of audible ring (Fig. 3), busy tone (Fig. 4), and reorder tone (Fig. 5), along with linear and log magnitude spectra for these tones. A Hamming window was used on the data to minimize the effects of the endpoints of the signal on the resulting spectrum. It can readily be seen from these figures that these tones have a reasonably complex structure both in time and in frequency.

Preliminary analysis also uncovered some prominent temporal properties of these tones with which the tone-detection systems would have to deal. For the audible ring tone a substantial transient generally was found at the beginning and end of each on cycle. Figure 6 shows an example of such a transient occurring at the beginning of a cycle. Such transients are distinctly audible as clicks. Also for each of the tones a substantial amount of variability in the duration of the on-off cycles was also observed—even within consecutive cycles of the same tone. Figure

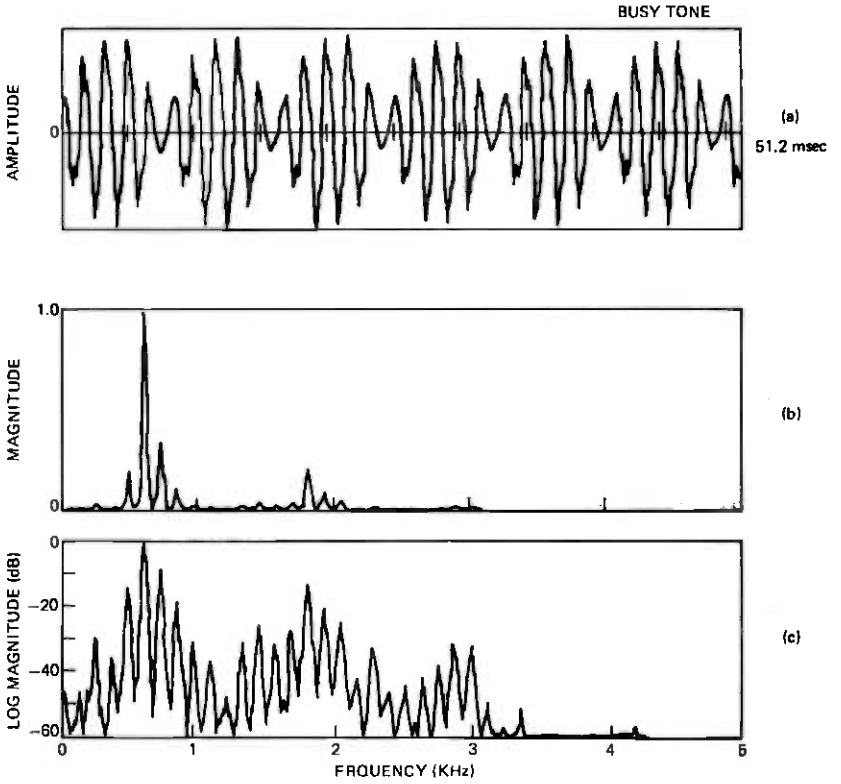


Fig. 4—Spectral characteristics of busy tone.

7 illustrates this effect during 2 cycles of a busy tone. The first on cycle lasts about 0.44 second, whereas the second on cycle lasts about 0.47 second. Such variability was not uncommon for the tones which were studied.

III. ENERGY-BASED SYSTEM FOR TONE DETECTION

Figure 8 shows a block diagram of the simulation of the energy-based system for tone detection. The input signal $x(t)$ is first band-pass filtered to the range 300–3200 Hz, and then sampled at a 10-kHz rate. An energy contour of $x(n)$ is computed using a 501-point (50-msec) Hamming window to give

$$E(n) = \sum_{m=0}^{N-1} [x(n-m)w(m)]^2 \tag{1}$$

where

$$N = 501$$

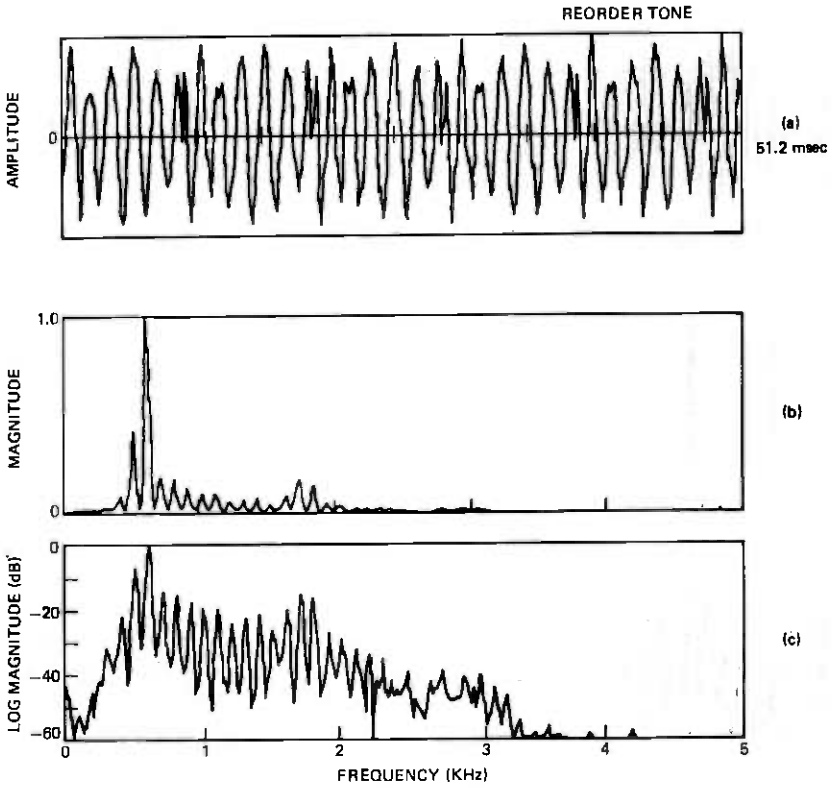


Fig. 5—Spectral characteristics of reorder tone.

and

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \tag{2}$$

The energy contour is resampled at a rate of 20 times/sec.* A noise threshold is computed by finding the minimum energy of the signal, and setting the threshold 12 dB above this level. Based on the energy threshold, the signal $E(n)$ (at the 20-Hz rate) is infinite-peak-clipped to give a binary signal $b(n)$, which is of the form

$$\begin{aligned} b(n) &= 1 \quad \text{if } E(n) \geq T \\ &= 0 \quad \text{if } E(n) < T \end{aligned} \tag{3}$$

(To eliminate the effects of spurious transient on the line, a delay of 150 msec, i.e., 3 samples, is built into the infinite peak clipper for off-on

* In the implementation $E(n)$ is computed only 20 times/second.

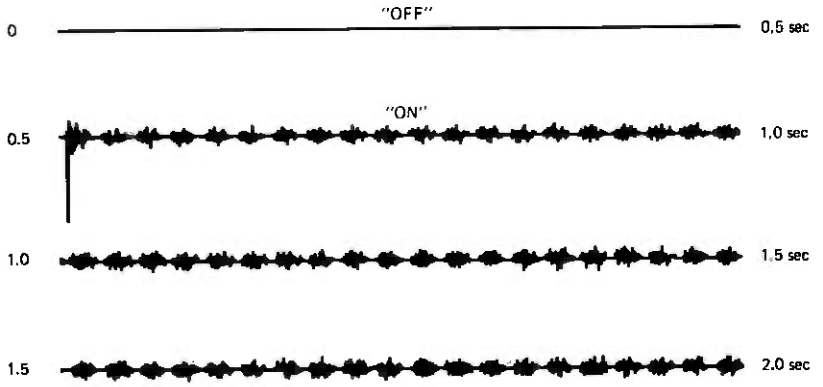


Fig. 6—Example showing transient at beginning of audible ring tone.

transitions to guarantee that $E(n)$ stays above T for this period. If $E(n)$ falls below T during this period, then $b(n)$ stays at 0. Similarly a delay of 50 msec, i.e., 1 sample, is built into on-off transitions. Thus, $E(n)$ must fall below T for 2 consecutive samples for $b(n)$ to be set to 0).

Following clipping, the signal $b(n)$ is blocked into runs. A run is defined as a sequence of 1's followed by a sequence of 0's. The detection system processes $x(n)$ until a total of 5 runs is obtained, or until 40 seconds of data are processed, whichever occurs first. The duration of each run, $r(j)$, $j = 1,2,\ldots 5$, is measured, and the average run length, RL, is determined as

$$\mathrm{RL} = \frac{1}{5}\sum_{j=1}^{5} r(j) \tag{4}$$

The signal, $b(n)$, is then comb-filtered using a fixed comb of delay RL samples, giving

$$c(n) = b(n) - b(n - RL) \tag{5}$$

The signal $c(n)$ is of the form

$$\begin{aligned} c(n) &= +1 \quad \text{if } b(n) = 1,\ b(n - RL) = 0 \\ &= 0 \quad \text{if } b(n) = 0,\ b(n - RL) = 0 \\ &\qquad \text{if } b(n) = 1,\ b(n - RL) = 1 \\ &= -1 \quad \text{if } b(n) = 0,\ b(n - RL) = 1 \end{aligned} \tag{6}$$

The absolute value of $c(n)$ is then accumulated, and the result is normalized by dividing by the number of samples of $b(n)$ which went into the computation, giving

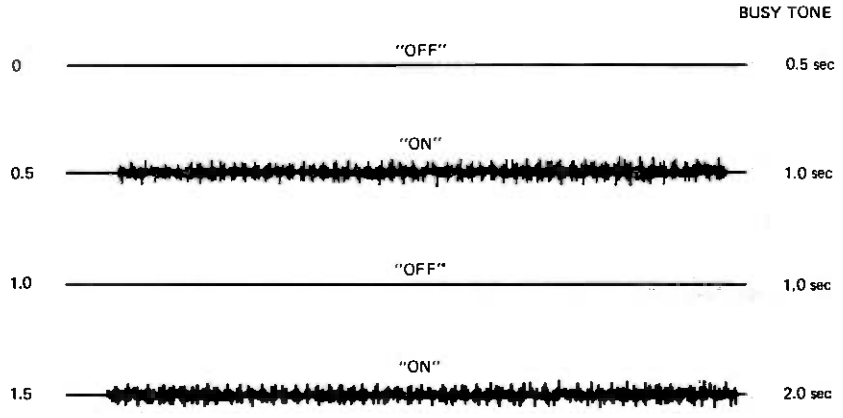


Fig. 7—Example showing variation in on-off cycle of busy tone.

$$D_N = \frac{\sum_{n=0}^{N-1} |c(n)|}{N} \tag{7}$$

Clearly D_N is a normalized measure of the lack of periodicity of the signal, since

$$0 \leq D_N \leq 1 \tag{8}$$

and $D_N \to 0$ if the signal is perfectly periodic and of period RL.

The final tone detection is based on the values of RL and D_N. If D_N is sufficiently small (indicating the signal is periodic) then the signal is classified as a tone of period RL samples. The tone whose period is closest to RL samples is chosen as the correct tone. If D_N is sufficiently large (indicating a lack of signal periodicity) then the signal is classified as speech (or silence). Based on experimentation with the system the thresholds chosen for D_N and the corresponding decision rules are

if $D_N < 0.15$	signal is periodic (tone) of period R_L samples
$D_N > 0.30$	signal is aperiodic (speech or silence)
$0.30 > D_N > 0.15$	signal is undefined

The undefined region accounts for tones on a very noisy-line, or speech conversations with a fairly periodic rhythm of talking.

Figure 9 shows some typical energy contours $[E(n)]$ for 15-second recordings of an audible ring tone (Fig. 9a), a reorder tone (Fig. 9b), a busy tone (Fig. 9c), and a weather announcement (Fig. 9d). It can be seen that by appropriate placement of the silence energy threshold, the resulting binary signal $b(n)$ will be periodic. However, a variation of over

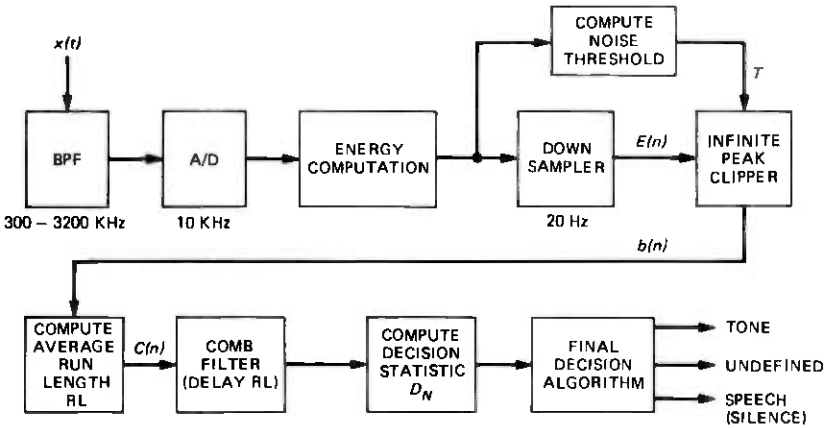


Fig. 8—Block diagram of energy system for tone detection.

20 dB in the level of $E(n)$ is obtained between the audible ring and the reorder signal. Thus careful choice of the silence threshold is extremely important to the proper functioning of this system. The transients present in the audible ring signal are also clearly seen in Fig. 9a, as noted previously.

IV. ZERO-CROSSING-BASED SYSTEM FOR TONE DETECTION

Figure 10 shows a block diagram of the zero crossing system which was used for tone detection. This system is organized as a parallel processing system with an individual detector for each tone. Speech (or silence) is indicated by the absence of a detected tone for an 8-second interval. The operation of this system is as follows. The input signal, $x(n)$, is sampled at a 10-kHz rate and then fed into three parallel tone detectors. The output of the tone detector is zero until a tone is detected at which point the output becomes one until the detector is reset by the decision logic. The decision logic to choose the tone is very simple. The output of each tone detector is monitored at a fixed rate until either one of the lines indicates a tone, or until 8 seconds have passed.* If one of the tone lines indicates a tone the logic decides which tone was detected and resets the tone detectors. After 8 seconds without a tone indication, the logic classifies the signal as speech (or silence).

Figure 11 shows a block diagram of the individual tone detectors. For each tone detector there are two parameters which define the range of the level crossing parameter for a particular tone. A level crossing for the signal $x(n)$ occurs at $n = n_0$ when

* The choice of a maximum interval of 8 seconds is explained later.

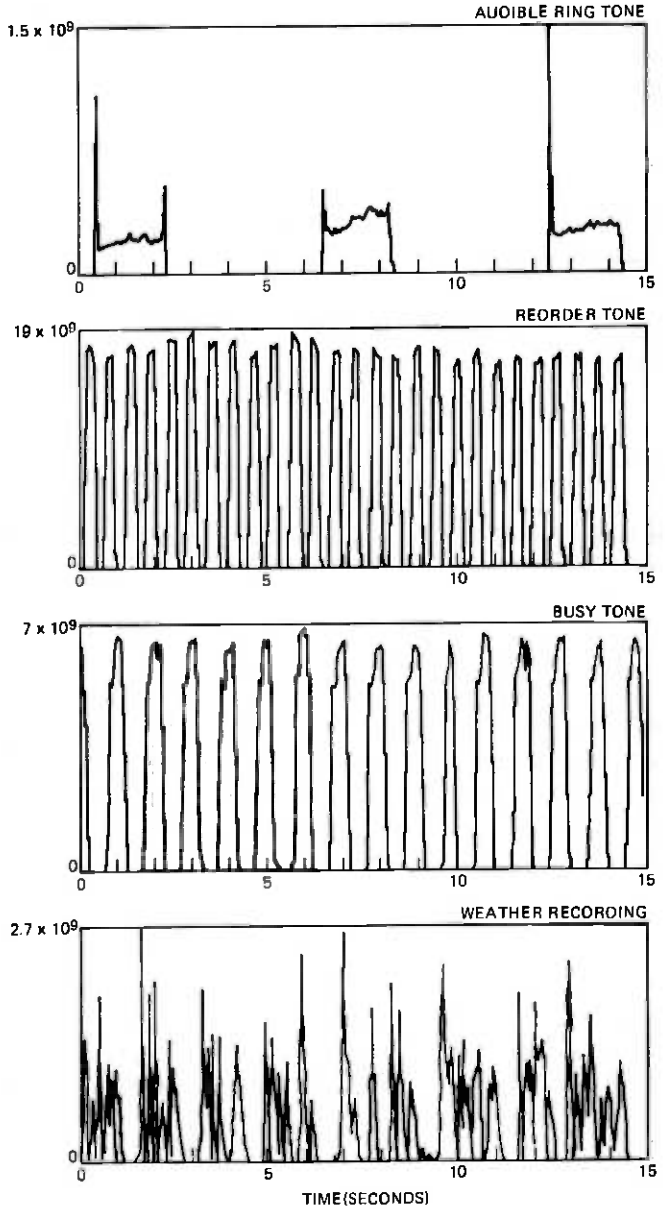


Fig. 9—Typical energy contours of three tones and a weather announcement.

$$x(n_0) \geq T \tag{9}$$

and

$$x(n) < -T \tag{10}$$

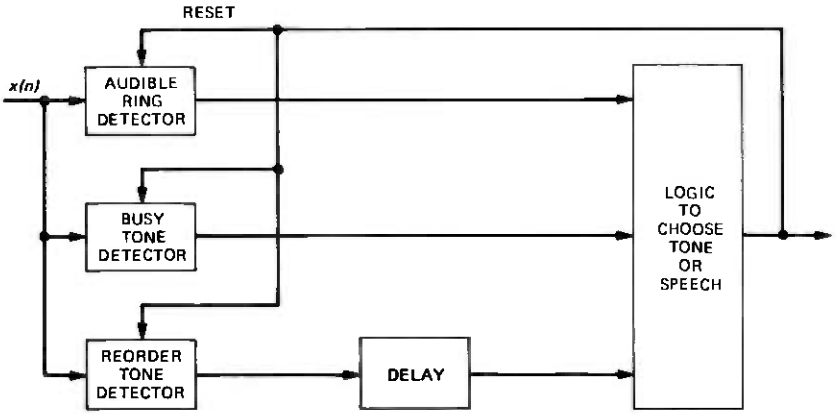


Fig. 10—Block diagram of the zero crossing system for tone detection.

for some value of $n = n_1 < n_0$ where n_1 is greater than the value at which the previous level crossing occurred—i.e., the signal must have been below the level $(-T)$ and risen above the level T for a level crossing to have occurred. The parameter used for tone detection was the number of level crossings of the signal within a specified duration D. Call this parameter $L_x(T,D)$. For each D second duration the quantity $L_x(T,D)$ is computed and compared to the range parameters R_L and R_H which define an expected range for $L_x(T,D)$ for the jth tone. For the jth tone to be detected the parameter $L_x(T,D)$ must satisfy the relation

$$R_L \leq L_x(T,D) \leq R_H \tag{11}$$

for three consecutive D second intervals. This sequence is called a triple

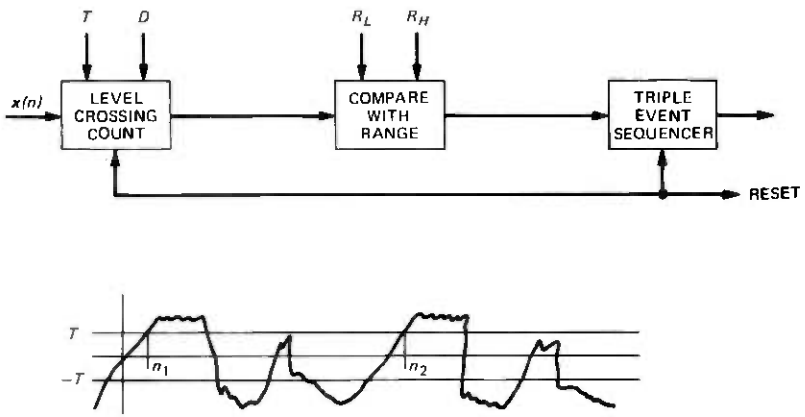


Fig. 11—Block diagram of the individual tone detectors.

Table I — Values for three tones

Signal	D	T	R_L	R_H
Reorder tone	60 msec	200	34	37
Busy tone	120 msec	200	67	73
Audible ring tone 1	480 msec	100	190	245
Audible ring tone 2	480 msec	100	220	270

event—an event being whenever eq. (11) is satisfied. The reason a triple event sequencer is used is because the major concentration of energy for the tones being detected is in the range 200–1000 Hz, and this in the region in which the major concentration of speech energy also occurs. Thus it is quite possible for eq. (3) to be satisfied by a speech signal, as well as by the correct tone. To minimize the possibility of a speech signal being detected by the tone logic, the triple event sequencer was used.

Table I gives the values of T,D, R_L and R_H for the three tones which were used in this investigation. For each tone the quantity D was chosen to be approximately ¼ of the on cycle of the tone to guarantee that at least 3 complete cycles of D seconds would occur within the on period of the tone, independent of the phase of the initial D second region. (Recall that the D second intervals are asynchronous with the tone—thus an interval may contain a transition from on to off or vice versa.) In addition, for both software and hardware convenience, the values of D are all multiples of 60 msec.

It is seen in Table 1 that two sets of parameters were used for the audible ring tones. This was due to the bimodal distribution of $L_x(T,D)$ which was measured when representative tones were recorded in different areas. In practice, an additional tone detector is used for such cases.

The reason for the delay in the reorder tone detector of Fig. 10 should now be clear. Since the frequency characteristics of the busy tone and the reorder tone were almost identical, a busy tone on the line would cause the reorder tone detector to detect the tone before the busy tone detector. Thus, a delay of 240 msec was used to allow the busy tone detector a chance to detect a busy tone prior to classifying the tone as a reorder tone.

The purpose of the reset signal following each signal classification is to clear the level crossing count, and to clear the triple event sequencer, so that the overall tone detector can be switched to a new line in order to classify a new signal.

Finally the reason for waiting 8 seconds until classifying the signal as speech (or silence) is that in the worst case the tone detector might be switched into an audible ring tone just past the beginning of an on cycle. Thus, the detector would not indicate the audible ring tone during this first on cycle, and would have to wait for the second on cycle for detecting

Table II — Results of tests on audible ring tone

(a) Accuracy of detection
Signal—Audible Ring Tone

Connection	Energy System	Zero Crossing System
1	100/100	100/100
2	100/100	100/100
3	100/100	100/100
4	100/100	100/100
5	100/100	100/100
6	100/100	100/100
Total	600/600	600/600
Percentage Accuracy	100%	100%

(b) Speed of detection

Connection	Energy System	Zero Crossing System
1	33.70 sec	4.574 sec
2	33.83	4.291
3	31.70	4.142
4	33.77	4.569
5	33.22	3.912
6	33.32	4.032
Total	189.54 sec	25.518 sec

Relative Time = $\frac{189.54}{25.518} = 7.43$

the tone. This would require a total of 2 seconds for the first on cycle, plus 6 seconds for the second off-on cycle, or a total of 8 seconds before audible ring can be eliminated.

V. EXPERIMENTAL EVALUATIONS

An extensive experimental evaluation of these two methods for tone detection was carried out over standard dialed-up telephone lines. For each type of signal (i.e., the three tones, and speech) a number of different connections was tested. For each connection a total of 60 seconds of the signal was recorded. A total of 100 trials were made on each signal with each trial beginning at a randomly selected point in the recording.

Results of these evaluation tests are given in Tables II to V for audible ring, busy, reorder, and speech respectively. Each table shows the accuracy for each system, as well as the average length of signal required to make the decision. Thus, for the audible ring tone (Table II), six different telephone numbers were used giving a total of 600 trials. Both systems detected audible ring 100 percent reliably. However, the energy-based system required about 7.4 times the amount of signal required by the zero crossing system—i.e., the average time to detect audible ring tone was 4.4 seconds for the zero crossing system, and 31.9 seconds for the energy system. (It should be kept in mind that the two tone-detection

Table III — Results of tests on busy tone

(a) Accuracy of detection

Signal—Busy Tone

Connection	Energy System	Zero Crossing System
1	100/100	79/100 (21 Ro)
2	100/100	77/100 (23 Ro)
3	99/99	82/100 (18 Ro)
4	100/100	91/100 (9 Ro)
5	100/100	79/100 (21 Ro)
6	100/100	84/100 (16 Ro)
7	100/100	85/100 (15 Ro)
Total	699/699	577/700 (123 Ro)
Percentage Accuracy	100%	82.4% → 100%

(b) Speed of detection

Connection	Energy System	Zero Crossing System
1	5.77 sec	0.596 sec
2	5.81	0.592
3	5.78	0.666
4	5.65	0.643
5	5.84	0.560
6	5.74	0.614
7	6.09	0.678
Total	40.48 sec	4.349 sec

Relative time = $\frac{40.48}{4.349} = 9.31$

systems were designed with different constraints, as discussed previously.)

For the busy tone (Table III) a total of seven different connections, or 700 trials, were used. The energy-based system detected 699 out of 699 correctly,* whereas the zero crossing system detected only 577 out of 700 correctly. Of the 123 errors, all were classified as reorder signal. This is in fact no real error since it occurs whenever the signal starting point is past the beginning of an on cycle, but before the middle of the on cycle—a fairly common occurrence. In such cases the busy tone is indistinguishable from a reorder tone. Thus the zero crossing system was effectively 100 percent accurate.

In terms of processing time the zero crossing system took, on average, 0.62 second to detect busy tone, whereas the energy system took about 5.8 seconds. Thus the zero crossing system was a factor of 9.3 times faster than the energy system.

Table IV shows the results for the reorder tone tests. A total of five connections were tested resulting in 500 individual trials. The energy system accurately detected 493 of 500, or 98.6 percent of the trials. The 7 errors involved classifying the signal as speech. These errors were

* On one trial the random starting point was too close to the end of the recording so no decision was made.

Table IV — Results of tests on reorder tone

(a) Accuracy of detection
Signal—Reorder Tone

Connection	Energy System	Zero Crossing System
1	100/100	100/100
2	100/100	100/100
3	100/100	100/100
4	93/100	100/100
5	100/100	87/100
Total	493/500	487/500
Percentage Accuracy	98.6%	97.4%

(b) Speed of detection

Connection	Energy System	Zero Crossing System
1	2.98 sec	0.576 sec
2	3.00	0.572
3	3.18	0.572
4	3.35	0.583
5	2.99	1.526
Total	15.40 sec	3.829 sec

Time Ratio = $\frac{15.4}{3.829} = 4.02$

eliminated by raising the noise level threshold to a value 18 dB above the noise level (instead of 12 dB as was normally the case). For the zero crossing system the accuracy was 487 of 500 or 97.45 percent of the trials. Of the 13 errors, 2 were classified as audible ring tone, and 11 were classified as speech. All these errors were corrected by increasing the search duration to include a second cycle of the tone.

In terms of speed the energy system required, on average, 3.08 seconds to detect the reorder tone, whereas the zero crossing system required about 0.76 second. Thus, the zero crossing system was a factor of 4 times faster than the energy system for this tone.

For testing the two systems on speech signals, two classes of signals were used. One class was of the announcement type—i.e., weather, news service, recorded phone messages, etc. The other class was a set of conversions. Table V shows the results on these two sets of signals. For recording and announcements the energy system detected 228 of 272 trials, or 83.8 percent of the trials. (In 128 cases the random starting point of the message was too close to the end of the message for a decision to have been made). All 44 errors were cases when the threshold fell in the undefined region. For conversational speech the energy system detected 398 of 398 trials, for 100 percent accuracy.

For the zero crossing system on recorded speech an accuracy of 364 of 400 trials, or 96 percent was obtained, and for conversations an accuracy of 390 of 400 trials or 97.5 percent was obtained. For recorded announcements 15 of the 16 errors were audible ring tone, and in one case

Table V — Results of tests on speech

(a) Accuracy of detection
Signal-Speech Recordings and Conversation

1. Recordings

Connection	Energy System	Zero Crossing System
1–Chicago Weather	25/43–18UD	84/100
2–NYC Weather	100/100	100/100
3–Viking News	3/29–26UD	100/100
4–Recorded Phone Trouble	100/100	100/100
Total	228/272–44UD	384/400
Percentage Accuracy	83.8%	96.0%

2. Conversation

Connection	Energy System	Zero Crossing System
1–2 Females	100/100	93/100
2–1 Female–1 Male	100/100	99/100
3–1 Female–1 Male	98/98	98/100
4–1 Female–1 Male	100/100	100/100
Total	398/398	390/400
Percentage Accuracy	100%	97.5%

(b) Speed of detection

1. Recordings

Connection	Energy System	Zero Crossing System
1	25.42 sec	7.347 sec
2	12.89	8.040
3	38.04	8.040
4	6.44	8.040
Total	82.79 sec	31.467 sec

$$\text{Time Ratio} = \frac{82.79}{31.467} = 2.63$$

2. Conversation

Connection	Energy System	Zero Crossing System
1	7.21 sec	7.798 sec
2	9.10	7.974
3	6.82	7.890
4	12.11	8.040
Total	35.24 sec	31.702 sec

$$\text{Time Ratio} = \frac{35.24}{31.702} = 1.11$$

it was busy tone. For conversations 7 of the 10 errors were audible ring tone, with 2 reorder, and 1 busy tone.

For recorded announcements the energy system required, on average, 20.7 seconds, whereas the zero crossing required 7.9 seconds. Thus the zero crossing system was about 2.6 times faster. For conversations, however, the average detection time was 8.9 seconds for the energy system, and 7.9 seconds for the zero crossing system. Therefore the zero crossing system was only about 1.1 times faster.

The only other comparison which was made between the two tone-detection systems was in terms of ease of implementation. The zero crossing system appears to be less costly to implement than the energy

system since it requires only a simple threshold device, a counter, and some simple logic, whereas the energy system requires arithmetic computations to compute the average run length, and the distance measure D_N.

VI. SUMMARY

We have presented two relatively simple methods for detecting tones based on temporal and spectral properties of the signals. The main requirement on the systems was that they be as accurate as possible in classifying tones or speech within the constraints of each individual system. Computer simulations of both systems were performed to compare and contrast the two systems. In terms of tone detection, both systems were essentially 100 percent reliable; however, the zero crossing system was from 4 to 10 times faster than the energy system. For recorded announcements the zero crossing system was more accurate than the energy system; however, for conversations the energy system was better by a small percentage. The processing time for the zero crossing system was always less than for the energy system, although the differences for speech were much less than for tones. Finally, in terms of implementation, it was argued that although both systems are relatively easy to implement, the zero crossing system is somewhat less costly than the energy system.

In summary, this study showed that the constraint of not using detailed tone frequency statistics in the detection process for the energy system led to an algorithm which was considerably slower than the zero crossing method, but equally accurate.

REFERENCES

1. L. B. Jackson, J. F. Kaiser, and H. S. McDonald, "An Approach to the Implementation of Digital Filters," IEEE Trans. Audio and Electroacoust., *AU-16*, No. 3 (September 1968), pp. 413–421.
2. J. N. Denenberg, "Spectral Moment Estimators: A New Approach to Tone Detection," B.S.T.J., 55, No. 2 (February 1976), pp. 143–155.
3. J. J. Dubnowski, J. C. French, and L. R. Rabiner, "Tone Detection For Automatic Control of Audio Tape Drives," IEEE Trans. on Acoustics, Speech, and Signal Proc., *ASSP-24*, No. 3 (June 1976), pp. 212–215.
4. F. T. Boesch and R. E. Thomas, "A Detection Scheme for Almost Periodic Functions," unpublished work.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 8, October 1977
Printed in U.S.A.

Some Experiments in Adaptive and Predictive Hadamard Transform Coding of Pictures

By A. N. NETRAVALI, B. PRASADA, and F. W. MOUNTS

(Manuscript received April 15, 1977)

This paper describes some experiments in adaptive and predictive Hadamard transform coding of still pictures using a small transform block (2 × 2 × 2). Predictive coding of the transform coefficients is discussed using certain combinations of coefficients of the present as well as previously transmitted blocks as predictors. Two separate adaptive quantization techniques are considered. The first technique relates to PCM quantization, in which a uniform PCM quantizer with a different number of quantization levels is used, depending upon the spatial activity within the block. The second technique alters the quantizer of a predictive transform coder based on a weighted sum of already transmitted coefficients of the present and previous blocks. Finally, we give a comparison of three coding techniques: (i) adaptive predictive transform coding, (ii) nonadaptive transform coding, and for comparison, (iii) nonadaptive predictive coding in the picture element domain using a two-dimensional prediction.

I. INTRODUCTION

In a recent paper[1] we considered Hadamard transform coding of still pictures using a small three-dimensional block (a 2 × 2 × 2 array of picture elements). There we described the design of optimum quantizers for the Hadamard transform coefficients based on psychovisual criteria in the transform domain. Starting with subjective tests to evaluate the visibility of quantization noise, we then developed a design procedure to minimize the "mean-square subjective distortion" (MSSD) due to quantization noise. We compared the performance of the resulting quantizers with the widely used Max-type[2] quantizers (i.e., quantizers which minimize the mean-square quantization error) and demonstrated our quantizers to be better in terms of picture quality and entropy of the quantizer output, for a given number of levels.

The present paper, which consists of three parts, extends the previous

work by considering techniques for adaptive and predictive coding of the transform coefficients, based on both statistical and psychovisual criteria. In the first part, we develop prediction algorithms for predictive coding of the coefficients. Although the small block size that we use ensures that the quantization noise can be placed in those parts of the picture where it is least visible, thereby permitting coarser quantization and thus achieving a higher coding efficiency, it does not exploit the statistical correlation between adjacent blocks. To overcome this, we consider predictive coding for the coefficients. We predict the value of a coefficient using a linear combination of already-transmitted values of other coefficients of both the present and the previous block. The prediction error is then quantized and transmitted. A reverse operation is performed at the receiver to reconstruct the picture elements. Our predictors are not limited to small block sizes. We show how they can be extended to larger spatial blocks as well as to spatiotemporal blocks.

The second part of this paper is concerned with two separate techniques for adaptive quantization, one useful in PCM quantization and the other in predictive quantization of the coefficients. These adaptive quantizers change to match the fidelity requirements of a viewer in different parts of the picture, as measured by subjective tests. We illustrate our methodology only for the coding of the first Hadamard coefficient. In PCM quantization, we use coefficients within a block representing a measure of spatial detail, to determine when to change the number of levels of the quantizer. Based on a theoretical analysis, we obtain a formula to change the number of quantizer levels and demonstrate the usefulness of this formula on a hardware system. In predictive quantization, the quantizers are switched on the basis of a weighted sum of the coefficients of the present and previous blocks. In areas of low spatial detail, a fine quantizer optimized for that area is used, whereas in areas of high spatial detail, a coarse quantizer is used which is optimized for such an area. The advantage of adapting the quantizer is evaluated by measuring the entropy for a given picture quality.

The third part of the paper deals with some comparisons between the techniques discussed in the first two parts and in our previous paper[1]. These comparisons are based on the picture quality versus entropy tradeoffs. They show that adaptive-predictive transform coding requires about 1.84 bits/pel for an excellent picture quality; and this represents, for the same picture quality, a decrease of almost 1.3 bits/pel over the bit rate obtainable by two-dimensional predictive coding in the picture element (pel) domain.

1.2 Relationship with some previous work

The combining of transform coding with predictive coding has been

discussed by many authors. Reudink[3] investigated simple DPCM coding of Hadamard transform coefficients which are obtained from a transform of 4, 8, or 16 pels along a scan line of video. Habibi[4] generalized the combining of transform and predictive coding. He considered several different transforms using one-dimensional blocks or small two-dimensional blocks and found that such a hybrid coding system performed better, in terms of signal-to-noise ratio for a given bit rate, than either the transform coding or the predictive coding system separately. Ishii[5] has considered a similar coding scheme using Hadamard transform coding, whereas Heller[6] and Roese et al.[7] have extended this concept to interframe coding.

Our contribution here is twofold. First, we develop methods for prediction that use coefficients from the present as well as previous blocks. Second, we develop a technique to quantize coefficients taking into account subjective effects of the quantization noise.

Adaptive coding of transform coefficients has been discussed by many authors.[8] Simple techniques of threshold sampling, which transmit only those coefficients whose magnitudes exceed a certain threshold, have been in existence for some time. Tasto and Wintz[9] have used local statistical properties of pictures to divide the picture into a number of segments and have chosen the coding strategy suited for each subpicture. Their division of pictures does depend on spatial activity, although not explicitly. It should be noted that their "best" quantizers were from those encountered in their trial-and-error procedure. Gimlett[10] has proposed a definition of an "activity index" using a weighted sum of absolute values of the transform coefficients and assigned more bits for coding those subpictures having a higher "activity index". This does not take advantage of the observer's reduced sensitivity for reproducing areas of higher activity.

Our adaptive quantization techniques divide the picture on the basis of subjective noise visibility and then design the quantizer for each segment. This is done using the data from the subjective tests in which noise visibility is related to certain measures of spatial detail.

II. PREDICTORS FOR TRANSFORM COEFFICIENTS

The objective of this section is to show that, for predictive coding of the transform coefficients, predictions better than the corresponding coefficients from the previous block can be made. In general, a predictor can utilize the information contained in the corresponding coefficient as well as other coefficients of the previous block. It can also utilize information contained in the other coefficients of the same block that may be available to the receiver when reconstructing the coefficient which is being differentially encoded.

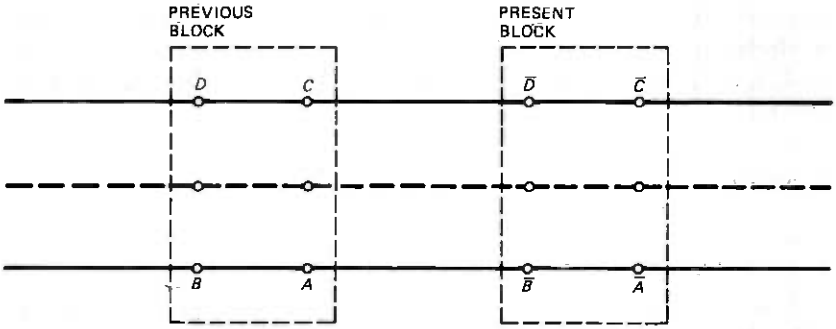


Fig. 1—Pel locations of two successive Hadamard transform blocks.

To illustrate the technique, we take a specific example of a 2×2 block of pels and develop a predictor for H_1, the first transform coefficient. The configuration of the block is shown in Fig. 1, where A is the current pel, B is the previous pel in the same line, C is the pel corresponding to A in the previous line in the same field, and D is the previous element with respect to C. After Hadamard transformation, the four coefficients are defined as follows:

$$\begin{aligned} H_1 &= A + B + C + D \\ H_2 &= A + B - C - D \\ H_3 &= A - B - C + D \\ H_4 &= A - B + C - D \end{aligned} \tag{1}$$

Now consider two horizontally consecutive blocks (as in Fig. 1), one having pels A, B, C, D giving rise to coefficients H_1, H_2, H_3, H_4; and the other having pels $\overline{A}, \overline{B}, \overline{C}, \overline{D}$ giving rise to coefficients $\overline{H}_1, \overline{H}_2, \overline{H}_3, \overline{H}_4$. Then the prediction for $\overline{H}_1$ is taken to be

$$(H_{1Q} + H_{4Q} + \overline{H}_{4Q}) \tag{2}$$

where subscript Q denotes the quantized values available both at the transmitter and the receiver. The prediction error is evaluated, quantized and transmitted.

The prediction error in the absence of quantization will be

$$\begin{aligned} \Delta\hat{H}_1 &= \overline{H}_1 - H_1 - H_4 - \overline{H}_4 \\ &= (\overline{D} - C) + (\overline{B} - \mathrm{A}) + (\overline{D} - \mathrm{C}) + (\overline{B} - A) \end{aligned} \tag{3}$$

From Fig. 1, $(\overline{D} - C)$ and $(\overline{B} - A)$ are the element differences and are, in general, small. Thus, the problem of transmitting H_1 is converted to the problem of transmitting a sum of certain element differences. The prediction error using the previous block H_1 as the prediction of $\overline{H}_1$ is

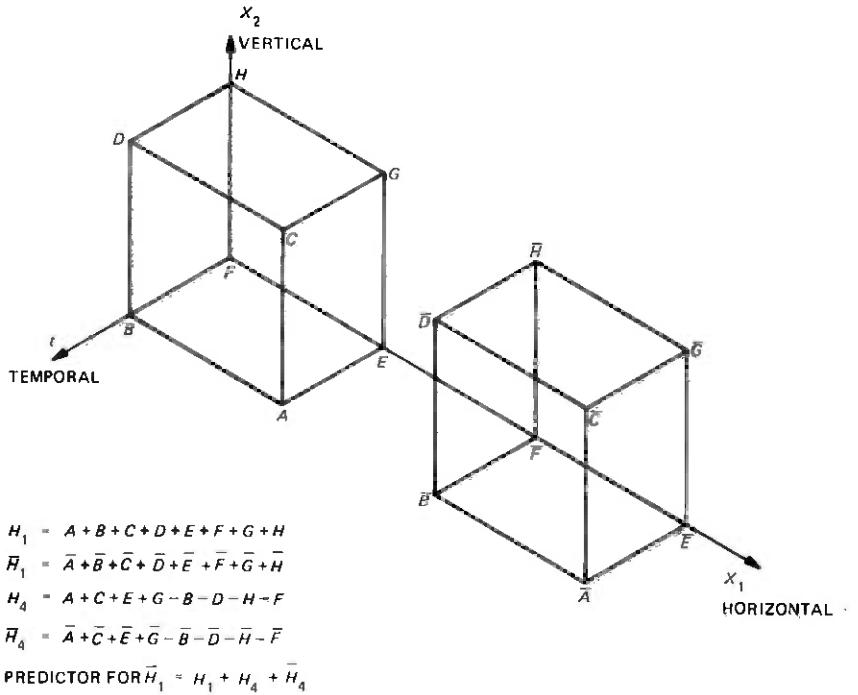


Fig. 2—Predictor for $\overline{H}_1$ of a spatiotemporal transform block.

given by

$$\Delta H_1 = \overline{H}_1 - H_1$$
$$= (\overline{D} - C) + (\overline{B} - A) + (\overline{C} - D) + (\overline{A} - B) \qquad (4)$$

Comparing the prediction errors [eqs. (3) and (4)], we see that the first two terms of the right-hand side are identical. However, the next two terms in eq. (4) will in general have higher values due to larger spatial separation, and therefore the predictor shown in eq. (2) will have a lower average error.

The example described above can be extended to more general cases than a spatial block of 2×2. Thus, better prediction of $\overline{H}_1$ is possible in the case of spatiotemporal blocks as well as larger spatial blocks. As an illustration, we show in Fig. 2 a case with a $2 \times 2 \times 2$ "spatiotemporal" block. The definition of the predictor is shown in the same figure. Notice that the prediction error can again be written as a summation of certain spatially adjacent element differences and this reduces the entropy of the prediction error.

The above procedure can be used for constructing better predictors for other transform coefficients. As an example for the blocks in Fig. 1,

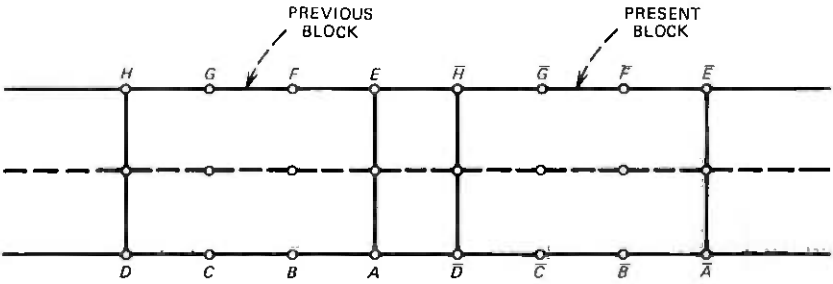


Fig. 3—Pel locations of two successive large transform blocks.

the predictor for $\overline{H}_2$ is taken to be

$$(H_2 + H_3 + \overline{H}_3) \tag{5}$$

It is easy to see that the prediction error is given by

$$2\{\overline{B} + C - (A + \overline{D})\} \tag{6}$$

which is two times H_3 of the intermediate block with pels $\{\overline{B}, A, \overline{D}, C\}$. Since H_3 generally has very small value, the prediction error will again have lower entropy as compared to the entropy of $(\overline{H}_2 - H_2)$. We note in passing that, instead of using a horizontally adjacent block as above, coefficients from the vertically adjacent block can also be used to construct predictors (e.g., the prediction for $\overline{H}_4$ can be $(H_4^V + H_3^V + \overline{H}_3)$, where superscript V denotes coefficients from the vertically adjacent block).

As an extension of our predictor to larger blocks, consider a block of 8 pels as shown in Fig. 3. Hadamard transformation of the pels from "previous block" gives us

$$\begin{bmatrix} H_1' \\ H_2' \\ H_3' \\ H_4' \\ H_5' \\ H_6' \\ H_7' \\ H_8' \end{bmatrix} = \begin{vmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \end{vmatrix} \begin{bmatrix} A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \end{bmatrix} \tag{7}$$

Hadamard transformation of the pels from the "present block" which generates $\overline{H}_1, \ldots, \overline{H}_8$ are similarly defined. The prediction error by using H_1 as a prediction of $\overline{H}_1$ is given by

$$\begin{aligned} \overline{H}_1 - H_1 = [&(\overline{D} - A) + (\overline{C} - B) + (\overline{H} - E) + (\overline{G} - F) \\ &+ (\overline{B} - C) + (\overline{A} - D) + (\overline{F} - G) + (\overline{E} - H)] \end{aligned} \tag{8}$$

We note that the last four terms of the right-hand side are often larger because of the wider separation of picture elements. A better predictor

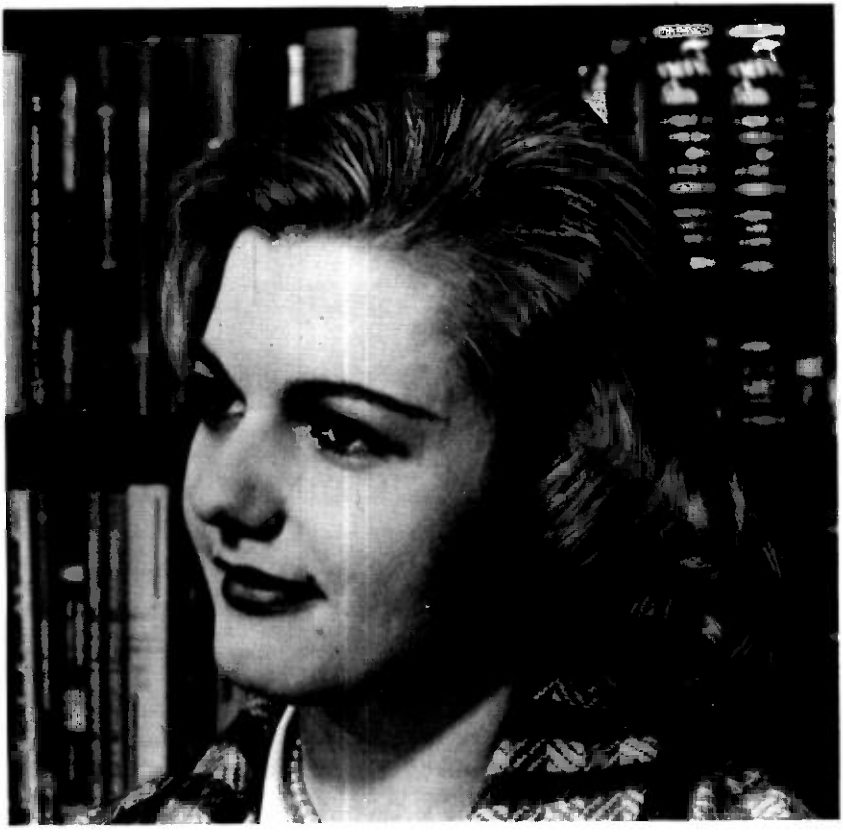


Fig. 4—Original picture used for subjective tests.

for $\overline{H}_1$ is taken to be $H_1 + H_4 + \overline{H}_4$ and then the prediction error would be

$$= [(\overline{D} - A) + (\overline{C} - B) + (\overline{H} - E) + (\overline{G} - F) + (\overline{D} - A) + (\overline{C} - B) + (\overline{H} - E) + (\overline{G} - F)] \quad (9)$$

The first four terms of eqs. (8) and (9) are equal; but comparing the last four terms, we see that in general the right-hand side of eq. (9) would be smaller than that of eq. (8).

We evaluated the performance of the new prediction scheme by hardware simulation using a $2 \times 2 \times 2$ block. We considered a still picture; and, therefore, except for frame-to-frame noise, this is equivalent to considering a 2×2 block. The picture entitled "Library Girl" shown in Fig. 4 was used for all the experiments discussed in this paper. We calculated the entropy of the unquantized prediction error for both predictors. The entropy of prediction error of eq. (3) was 4.99 bits/block,

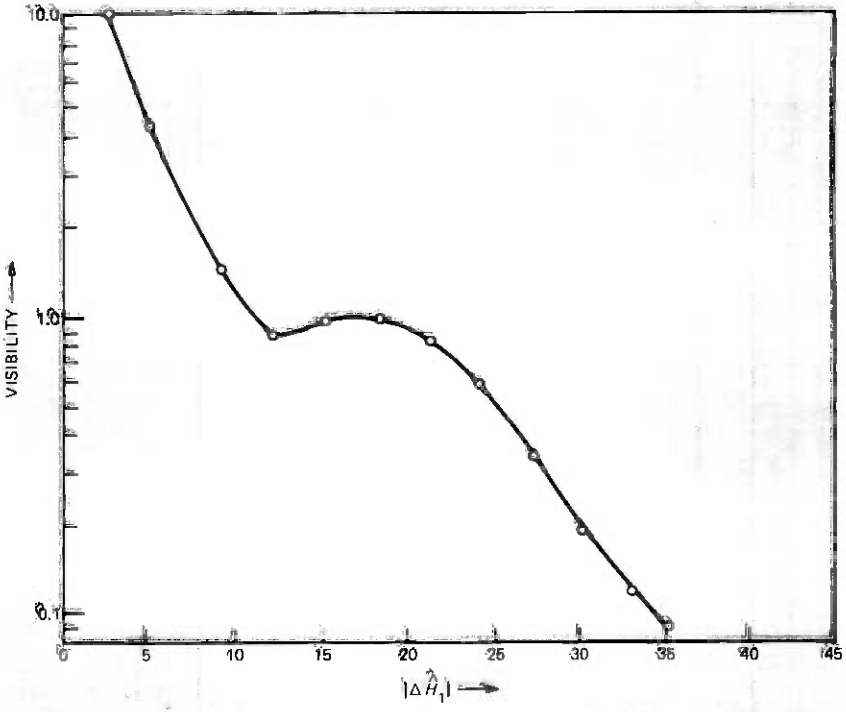


Fig. 5—Visibility function for noise in H_1. Prediction error $|\Delta \hat{H}_1|$ is used as a control variable.

and of eq. (4) was 6.35 bits/block, clearly showing that the new predictor is better for the picture used. Quantizers were then designed for each of these predictors. They were optimized by performing subjective experiments to measure the visibility of quantization noise as a function of the unquantized prediction error; and then, following a procedure analogous to that in Ref. 1, the mean-square subjective distortion due to quantization noise was minimized. We note that the prediction error $\Delta \hat{H}_1$ with respect to which the noise visibility is determined in these experiments is a better choice than ΔH_1 since $\Delta \hat{H}_1$ is the sum of element differences spatially closer to block $\{\overline{A}, \overline{B}, \overline{C}, \overline{D}, \overline{E}, \overline{F}, \overline{G}, \overline{H}\}$, where the quantization error appears. This provides a better spatial masking of the quantization noise.

The visibility function* $f_{H_1}(\cdot)$ for noise in H_1 as a function of $\Delta \hat{H}_1$ is shown in Fig. 5. Quantizers were obtained for a different number of quantization levels (N), and their performance was observed by the authors. For $N = 21$, a fairly good picture was obtained with an entropy of 3.13 bits/block. There was a noticeable (not objectionable) noise

* The method of obtaining visibility function is described in Ref. 1.

pattern with some structure in the gray regions. For $N = 23$, very good picture quality was obtained. A noise pattern was slightly visible in the gray regions of the picture, and the entropy was 3.17 bits/block. For $N = 25$, a near perfect picture was obtained with a very slight amount of noise in the gray regions of the picture. The entropy was 3.19 bits/block.

During these evaluations, the other coefficients H_2, H_3 and H_4 were unquantized. Even though the picture was stationary, the experiment was done in real time, so that the effects of camera noise which changes from frame to frame were included.

It is interesting to compare these observations with results[1] obtained using H_1 as the predictor for $\overline{H}_1$. In that case, a near perfect picture was obtained with $N = 36$ and an entropy of 4.25 bits/block. This shows that the new predictor gave about 25 percent lower entropy than the previous block coefficient predictor.

III. ADAPTIVE QUANTIZATION OF THE FIRST COEFFICIENT

In this part we discuss two separate techniques for the adaptive quantization of the first transform coefficient H_1. The first technique is applicable to PCM quantization, and the second to DPCM quantization of the coefficients.

The general approach is to identify measures of spatial luminance activity in terms of certain transform coefficients and then to obtain relations between noise visibility and these measures by subjective experiments. The visibility function is used for the categorization of blocks into subpictures of approximately equal visibility for a given quantity of noise. Separate quantizers are used for each category. We will now describe the application of this general approach for the quantization of H_1.

3.1 Adaptive PCM quantization of H_1

In general, a picture may be categorized into several regions depending on spatial detail. H_1 can be specified with different accuracy in each of these regions without degrading the picture quality as seen by a human viewer. The magnitude of either H_2 or H_4 or both is large in the busy regions of the picture and, hence, is taken as an indication of picture busyness. Since H_2 and H_4 are available to the receiver prior to decoding of H_1, there is no need to transmit information regarding the adaptation of coding of H_1 explicitly to the receiver.

3.1.1 Design of adaptive PCM Quantizer for H_1 as a function of $|H_2|$, $|H_4|$

Let

$$x = \max(|H_2|, |H_4|)$$

$f(x)$ = visibility function for noise in H_1 obtained as a function of x

$p(x)$ = probability density of x, measured for picture of Fig. 4.

We carry out our derivation for a uniform quantizer with "ℓ_1" levels used for quantizing H_1 from all blocks where $0 \leq x \leq x_1$, x_1 being a positive number, and "ℓ_2" levels used for all other cases. Assuming that the quantization noise is proportional to $1/(\ell_i)^\gamma$, $i = 1, 2$, for a positive constant γ, we can express the visible distortion (D) due to the quantization noise as†

$$D = \frac{1}{\ell_1^\gamma} \int_0^{x_1} f(x)\, dx + \frac{1}{\ell_2^\gamma} \int_{x_1}^{\infty} f(x)\, dx \tag{10}$$

Assuming no variable-length coding, the average number of bits required for such quantization is†

$$B = \log \ell_1 \int_0^{x_1} p(x)\, dx + \log \ell_2 \int_{x_1}^{\infty} p(x)\, dx \tag{11}$$

Using calculus of variations, we solve the problem of minimizing D, for a given B, with respect to ℓ_1, ℓ_2, and x_1. It is seen that the optimum ℓ_1 and ℓ_2, defined as ℓ_1^*, ℓ_2^*, are given by

$$\ell_1^* \propto \sqrt[1/\gamma]{\frac{\gamma \int_0^{x_1} f(x)dx}{\int_0^{x_1} p(x)dx}} \tag{12a}$$

$$\ell_2^* \propto \sqrt[1/\gamma]{\frac{\gamma \int_{x_1}^{\infty} f(x)dx}{\int_{x_1}^{\infty} p(x)dx}} \tag{12b}$$

Also the optimum x_1^* is given by

$$\frac{f(x_1^*)}{p(x_1^*)} \propto \frac{\log (\ell_2^*/\ell_1^*)}{(\ell_1^*)^{-\gamma} - (\ell_2^*)^{-\gamma}} \qquad (\ell_1^* \neq \ell_2^*) \tag{12c}$$

As shown in the next section, we simulated a system with adaptive quantization to check the above equations.

3.1.2 Experimental investigation and results

An experiment was performed to obtain a value of γ and to verify the result of Section 3.1.1. First, the visibility function was obtained by subjective testing. Figure 6 shows the visibility function $f(x)$ obtained with x as the control function.

† Except for a proportionality constant.

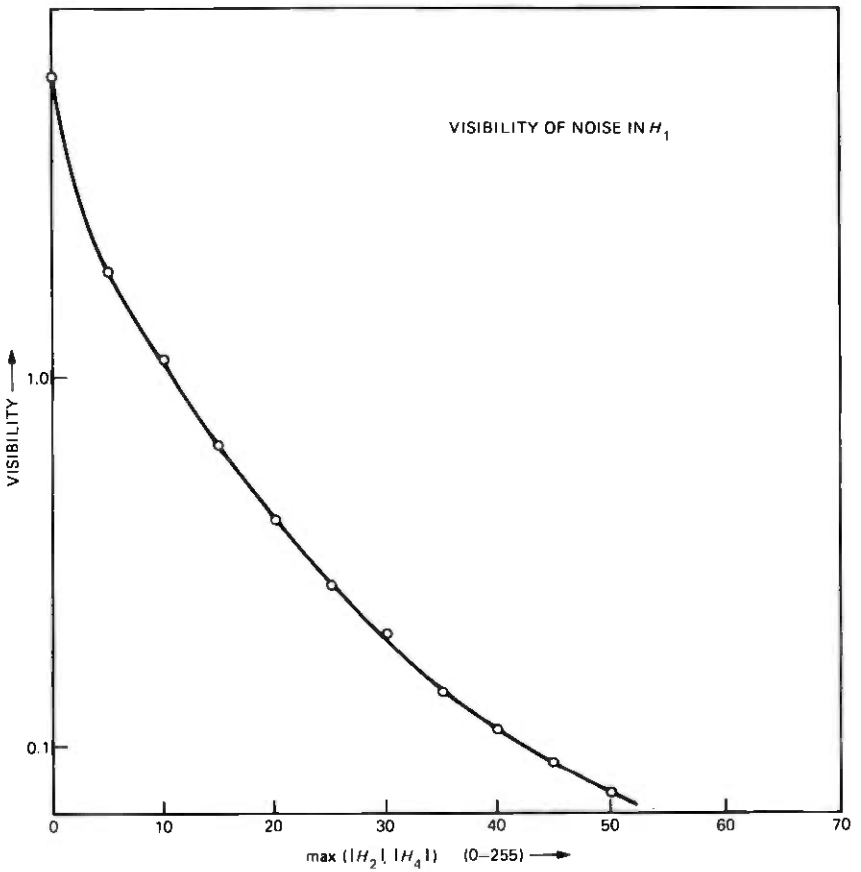


Fig. 6—Visibility function for noise in H_1. Max ($|H_2|$,$|H_4|$) is used as a control variable.

In the experiment, two quantizers, Q_A and Q_B, were used to quantize H_1. For a block, the function $x = \max(|H_2|,|H_4|)$ was determined† and the value compared with a threshold to decide whether quantizer Q_A or Q_B should be used. The block diagram of the experimental setup is shown in Fig. 7. Condition I refers to nonadaptive quantization of H_1 by a uniform quantizer. We considered two cases: for case 1 the uniform quantizer uses 128 levels, and for case 2, 64 levels. Condition II pertains to quantization of H_1 by either quantizer Q_A (for $x \leq T$) or quantizer Q_B (for $x > T$). In an A–B test, two subjects compared pictures corresponding to conditions I and II and adjusted the threshold T to the smallest value at which the pictures appeared to be of the same quality.

† Effect of quantization of H_2 and H_4 was neglected.

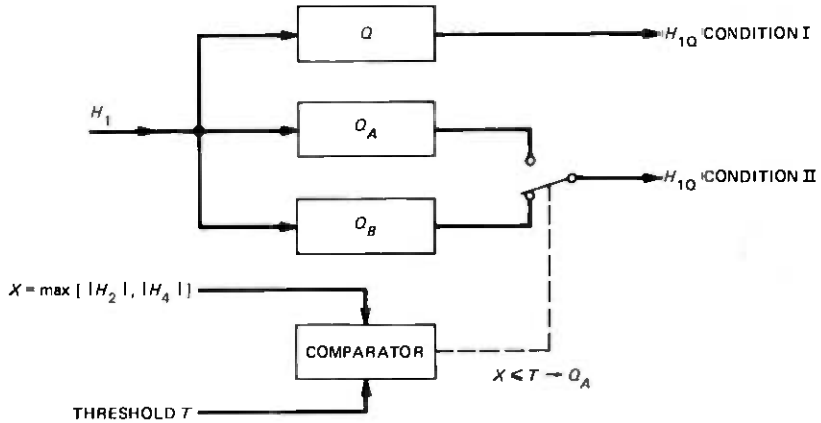


Fig. 7—Experimental setup for quantizer optimization. H_1 is quantized using a PCM quantizer Q and the resulting picture compared with the picture obtained by using PCM quantizers Q_A or Q_B. The choice of Q_A or Q_B depends on whether max $(|H_2|, |H_4|)$ is $< T$ or $\geq T$, respectively.

The results of the test are shown in Table I. Q_A was a quantizer with the same number of levels as the uniform quantizer used for condition I (7 or 6 bits per H_1 sample, as the case may be). Q_B had a smaller number of levels. By changing the threshold T, the percentage of blocks which were coded by Q_A and Q_B were varied. The table gives N_A, the number of coefficients coded by the quantizer Q_A and N_B, the number of coefficients coded by the Q_B. The entropies of the output signals of the quantizers Q_A and Q_B are denoted by E_A and E_B, and the overall entropy is given by E. The entropy of H_1 for condition I with 64-level quantization was 5.66 bits/block, and with 128-level quantization it was 6.64 bits/block.

The table also shows the advantages of adaptation. It is seen that to get the same quality as a picture with 7-bit quantization of H_1, the combination of 7 and 6 bits for Q_A and Q_B, respectively, results in lower entropy than combinations 7 and 5 or 7 and 4 bits. Using the combination of 7 and 6 bits, the saving in entropy is of the order of 15 percent over the nonadaptive quantization.

In order to judge the usefulness of eq. (12), we took values of ℓ_1^*/ℓ_2^* and x_1^* obtained from the above experiments and found that an approximate value of 2 for γ gave a good fit to all the different cases. The precise value of x_1^* which could be obtained from eq. (12c) was checked by evaluating the proportionality constant (between the left-hand side and the right-hand side of the equation) for different cases and was found to vary by about 14 percent. This allows us to conclude that our experimental results are within reasonable agreement of the optimality conditions of eq. (12).

Table I — Results of adaptive PCM quantization of H_1

Test no.	Subject	Bits for quantizer Q_A	Bits for quantizer Q_B	Bits for quantizer Q	Threshold T	Number of blocks in A (average)	Number of blocks in B (average)	Conditional entropy of blocks from Q_A	Conditional entropy of blocks from Q_B	Overall entropy of bits/block
1	I	7	4	7	42	9286	521	6.665	2.922	6.467
	II				47					
2	I	7	5	7	19	6903	2902	6.688	4.230	5.960
	II				23					
3	I	7	6	7	5	2482	7363	6.503	5.518	5.770
	II				5					
4	I	6	4	6	35	8816	990	5.671	3.048	5.340
	II				35					
5	I	6	5	6	12	5729	4066	5.699	4.348	5.140
	II				6					

Entropy of H_1 with 6-bit quantization = 5.657 bits/block
Entropy of H_1 with 7-bit quantization = 6.642 bits/block

3.2 Adaptive DPCM Quantization of H_1

In this section we describe our experiments in adaptive predictive coding of H_1 using a $(2 \times 2 \times 2)$ block. This is done by switching the quantizer in the predictive coder "loop" as a function of a measure of spatial detail. We define the spatial detail S as

$$S = \max [|\overline{H}_4|, \alpha|\overline{H}_2|, \beta|H_4|, \delta|H_2|] \tag{13}$$

This is used as a measure of spatial detail for the transform block consisting of elements $\{\overline{A}, \overline{B}, \overline{C}, \overline{D}, \overline{E}, \overline{F}, \overline{G}, \overline{H}\}$ of Fig. 2. Weight α is used to compensate for the wider separation between the lines due to interlace. We took α to be equal to 1/2. Weight β, which was taken to be 1/2, compensates for the spatial separation between the blocks consisting of $\{A, B, C, D, E, F, G, H\}$ and $\{\overline{A}, \overline{B}, \overline{C}, \overline{D}, \overline{E}, \overline{F}, \overline{G}, \overline{H}\}$. Weight δ was taken to be 1/4 and compensated for the spatial separation as well as effects of interlace.

Using this measure of spatial detail, we performed subjective tests to determine the visibility of noise in $\overline{H}_1$ as a function of S. The visibility function from these tests was used to divide the picture into subpictures. This is done by making a two-step approximation (i.e., piecewise constant approximation with two pieces) to the visibility function. The threshold T, corresponding to the point of separation of the two pieces of approximation, is used to divide the picture. Thus, if $S \leq T$, the block consisting of $\{\overline{A}, \overline{B}, \overline{C}, \overline{D}, \overline{E}, \overline{F}, \overline{G}, \overline{H}\}$ belongs to subpicture I, otherwise it belongs to subpicture II. Each subpicture contains blocks wherein the visibility of a unit of quantization noise is approximately equal.

We performed subjective experiments to determine the characteristics of the quantizer for each subpicture. We used the new predictor for $\overline{H}_1$, as described in Section II. The conditional visibility function, i.e., the visibility function for noise in $\overline{H}_1$ for all blocks belonging to subpicture I, is obtained by adding noise to $\overline{H}_1$ as a function of the unquantized prediction error $(\overline{H}_1 - H_1 - \overline{H}_4 - H_4)$, whenever the spatial detail for the block is less than T. This visibility function is shown in Fig. 8. The quantizer for the prediction error of $\overline{H}_1$ from blocks in subpicture I is obtained by minimizing the mean-square subjective quantization error, using the visibility function as the weighting function. The quantization characteristics for $\overline{H}_1$ of subpicture II are obtained similarly.

We used the quantizers obtained by the above procedure in the real-time system. The picture of Figure 4 was quantized using a 15-level quantizer for subpicture I and a 21-level quantizer for subpicture II. The entropy of the quantized output was 2.41 bits/block for the first transform coefficient. The picture produced by such a quantization was fairly good, although the quantization noise was certainly visible (but not objectionable). The quality of this picture was approximately the same

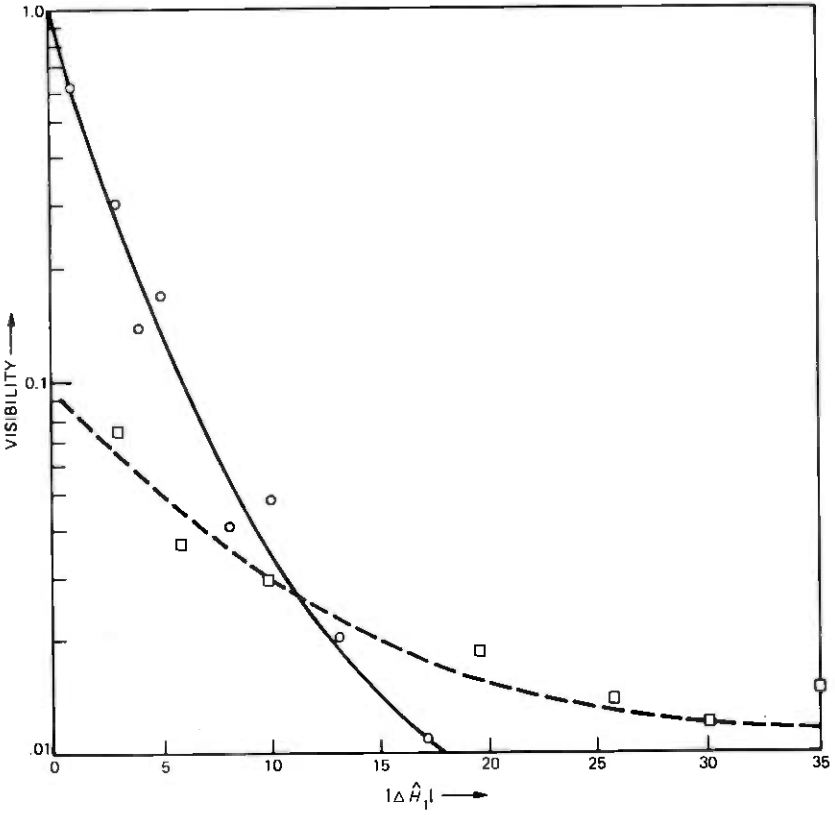


Fig. 8—Conditional visibility functions for noise in H_1. Prediction error $|\Delta \hat{H}_1|$ is used as a control variable. Segment I circles (—) is for the quiet area and segment II (squares) is for the busy area.

as the quality using a nonadaptive 21-level quantizer. Thus the saving in entropy using adaptive quantization was 0.22 bits/block for H_1, which was about 7 percent. We also did adaptive quantization to produce almost perfect picture quality. This required a 17-level quantizer for subpicture I and a 25-level quantizer for subpicture II. The picture quality for this case was equivalent to that produced by 25-level nonadaptive quantization; the advantage of adaptation is about 0.18 bits/block, which amounts to about 6 percent.

IV. SOME COMPARISONS BETWEEN PREDICTIVE CODING AND PREDICTIVE TRANSFORM CODING

In this part, we give a comparison of some of the techniques discussed in our previous paper[1] and the first two parts of this paper. This comparison, done on our real-time system, is limited to the performance in terms of entropy for a given picture quality.

We simulated, for the purpose of comparison, two predictive coding systems in the pel domain. One used the previous element prediction, and the other used a two-dimensional prediction (the predictor for picture element A of Fig. 1 was $B + C - D$). For each of these cases, we optimized the quantizer characteristics by doing subjective experiments in which the visibility of noise was determined by adding noise to the picture element being coded as a function of its prediction error. Pictures of different quality were produced by quantizers having a different number of levels (N). In the case of the previous element predictor, for $N = 23$, a near perfect picture was obtained. There was very slight noise in low-brightness regions. The entropy was 3.57 bits/pel. For $N = 16$, the picture quality obtained was good; however, a slight amount of slope overload and edge busyness was observed. In the low-brightness area the picture was more noisy than for $N = 23$. The entropy was 3.20 bits/pel. For a 13-level quantizer, noise was observed in low-brightness levels. Slope overload and busyness were observed on the edges. The picture was acceptable but impairments were certainly visible. The entropy was 3.02 bits/pel.

Using the two-dimensional predictor and a 16-level optimized quantizer, a near perfect picture was obtained. There was slight slope overload observed in the corner of the mouth of the picture in Fig. 4. The entropy was 3.12 bits/pel. For $N = 13$, a very good picture was obtained except for the slight slope overload in regions of large changes. The entropy in this case was 2.82 bits/pel.

We recall from our earlier work[1] that nonadaptive transform coding (in which the first coefficient is coded using predictive coding techniques with the previous block coefficient as the predictor and the other coefficients are PCM encoded) is capable of generating an excellent picture quality with 2.17 bits/pel. Thus there is almost a 0.95 bit/pel advantage by using transform coding over DPCM with a two-dimensional predictor, and a 1.4 bits/pel advantage over DPCM with a previous element predictor. This advantage is increased by using our new predictor for coding of H_1 and adapting the quantizer. An excellent picture would then be obtained with 1.80 bits/pel. It should be noted, however, that the DPCM techniques which we used for comparison are rather simple, and they can be made more sophisticated to decrease the bit rate significantly.[11] Also, the advantage of transform coding in localizing the transmission error to within a block is lost by doing predictive coding of the coefficients or by adapting the coding using coefficients from many surrounding blocks.

V. SUMMARY AND CONCLUSIONS

We have described techniques for adaptive and predictive coding of Hadamard transform coefficients. We have shown how predictors for

transform coefficients could be designed to reduce the bit rates. For the picture we used, the advantage of using our predictor for H_1 appears to be about 25 percent in terms of entropy reductions over the conventional predictor using the corresponding coefficients from a previous block. We demonstrate this by simulating a predictive coder for coding of H_1. Adaptive quantization, in which a coarse quantizer is used for areas of pictures with larger spatial detail and a fine quantizer is used for relatively flat areas of the picture, was demonstrated by PCM quantization of H_1, as well as by predictive quantization of H_1. We showed that adaptation reduces the bit rate by about 5 to 15 percent without changing the picture quality. We attempted a comparison of the predictive coding in the pel and transform domain. Here, on the basis of picture quality and bit-rate considerations only, we found that using a $2 \times 2 \times 2$ block for transform coding allows a lower bit rate by about 1.8 bits/pel over simple DPCM techniques using the previous element predictor and 1.3 bits/pel over DPCM with a two-dimensional predictor. This comparison does not consider complexity of the encoding schemes. It should be noted that throughout this paper our emphasis has been on investigation of certain techniques rather than a description of a complete coding system; several aspects (e.g., channel errors) which are important to a coding system have not been discussed.

VI. ACKNOWLEDGMENTS

We would like to thank K. Walsh for his help in various phases of this work and the members of the Electronics and Computer Systems Research Laboratory who were test subjects.

REFERENCES

1. F. W. Mounts, A. N. Netravali, and B. Prasada, "Design of Quantizers for Real-Time Hadamard Transform Coding of Pictures," B.S.T.J., *56*, No. 1 (January 1977) pp. 21–48.
2. J. Max, "Quantizing for Minimum Distortion," *IEEE Trans. on Information Theory, IT-6* (March 1960), pp. 7–12.
3. D. O. Reudink, unpublished work, 1971.
4. A. Habibi, "Hybrid Coding of Pictorial Data," *IEEE Trans. on Communications, COM-22*, No. 5 (May 1974), pp. 614–624.
5. M. Ishii, "Picture Bandwidth Compression by DPCM in the Hadamard Transform Domain," *Fujitsu Scientific and Technical Journal*, September 1974, pp. 51–65.
6. J. A. Heller, "A Real Time Hadamard Transform Video Compression System Using Frame-to-Frame Differencing," *NTC-74*, San Diego, 1974.
7. J. A. Roese, A. Habibi, W. K. Pratt, and G. S. Robinson, "Interframe Transform Coding and Predictive Coding Methods," *ICC-75*, San Francisco, 1975.
8. P. A. Wintz, "Transform Picture Coding," *Proc., IEEE* July 1972, pp. 809–820.
9. M. Tasto and P. A. Wintz, "Picture Bandwidth Compression by Adaptive Block Quantization," Technical Report TR-EE-70-14, July 1970, Purdue University, Lafayette, Indiana.
10. J. Gimlett, "Use of Activity Classes in Adaptive Transform Image Coding," *IEEE Trans. on Communications*, July 1975, pp. 785–786.
11. A. N. Netravali and B. Prasada, "Adaptive Quantization of Picture Signals Using Spatial Masking," *Proc. IEEE*, April 1977, pp. 536–548.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 8, October 1977
Printed in U.S.A.

A Scanning Spot-Beam Satellite System

By D. O. REUDINK and Y. S. YEH

(Manuscript received July 20, 1977)

We propose a satellite with a high gain, movable spot beam to communicate with individual earth stations time-sharing a single channel in the TDMA (Time-Division Multiple Access) mode. It is estimated that this approach could readily save some 20 dB in the link budget while still providing full U.S. coverage. When this 20 dB is apportioned with the objectives of reducing the earth-station antenna size, increasing the satellite capacity, and reducing transmitter power, the effects are dramatic. This technique can be combined with a fixed-spot beam system serving major traffic areas. This combination can provide both full area coverage as well as multiple reuse of the frequency band. A TDMA burst organization is proposed, and estimates of burst lengths, beam switching intervals, and buffer storage size are made for a 100-earth-station network operating on a 600 Mb/s channel. A phased array antenna with each element irradiating the entire U.S. is employed to form the movable spot-beam. This provides an attractive solution even though a closed-loop beam-forming algorithm may be required. It appears feasible to construct such an antenna with nearly 50-dB gain capable of forming a spot beam toward any position within the continental United States with a switching time of a few nanoseconds.

I. INTRODUCTION

The current approaches to domestic-satellite systems divide along the lines of area-coverage and spot-beam concepts. Each system has its merits as well as disadvantages. A spot-beam satellite system[1,2] allows high antenna gain and several reuses of the allocated frequency spectrum. In Ref. 1, a 12/14-GHz system with 11 frequency reuses was described which could provide reliable service at digital rates of 600 Mb/s with 30 watts peak transmitter power, employing a satellite antenna having 47-dB gain in each spot-beam. The disadvantage of such a system stems from the fact that each spot-beam covers only a small area. To avoid cochannel interference, a dead space between any two adjacent beams much larger than the beam coverage area (e.g., 3-dB contour) is

required.[3,4] Also, there are regions needing service which do not have enough traffic to justify a dedicated spot-beam.

Area coverage satellites, such as used by AT&T/GTE, Western Union, or RCA use broad antenna beams covering the whole United States. They are capable of providing service everywhere within the continental U.S.A. but lack channel capacity because the alloted spectrum can be reused at most once by polarization reuse. A more significant disadvantage, however, is the power penalty associated with the gain of an area-coverage antenna. The 3-dB contour gain of a U.S. coverage antenna is only 27 dB, and there is little that can be done to improve it further. To obtain the same SNR as the previously mentioned spot-beam antenna system, the required RF power to transmit at a 600-Mb/s data rate would be 3 kW. Equivalently, one could use a 10 times larger diameter earth station antenna than used by a spot-beam system. Since neither alternative is practical, the link SNR must be compromised by approximately 10 dB. As a result the rain outage at 12 GHz might be expected to increase by an order of magnitude.[5] Even with a 10-dB sacrifice in margin, an additional 10 dB must be obtained through a combination of higher satellite transmitter power and larger earth stations. This is the unfortunate price one must pay to use a wide-area-coverage antenna.

In this paper, we discuss a new concept which achieves area coverage using a rapidly scanned spot-beam. The beam is steered so that all parts of the country can be covered, but at different times, which works perfectly with a time-division multiple access (TDMA) configuration. Because only one ground station accesses the satellite at a time, a spot-beam toward that ground station is all that is needed and spreading energy over the entire United States is not necessary. To achieve total service, it is necessary to scan both the transmit and receive beams, coordinating their movements in accordance with the pair-wise traffic demands of the system. Each station is assigned a time slot where it transmits bursts of information to other stations. It is envisioned that the antenna gains would be of the order of 50 dB so that approximately 1 percent of the U.S. is illuminated at any one time. Thus, 100 beam directions will provide complete U.S. coverage. Once a particular scanning sequence is set up, it would be repeated at a frame period of perhaps a few milliseconds.

Let us examine the potential advantage of such a scanning spot-beam system. At 12/14 GHz, with polarization reuse, an area coverage satellite has enough bandwidth (1 GHz total) to support about a 1.2-Gb/s data rate. Under the conditions of Ref. 1, which assumed a large margin to minimize outage due to rain, the area coverage system would require 6 kW of RF power. Even using the 10 dB less margin, the required weight in solar cells is so large polarization reuse cannot be employed if the most popular of today's launch vehicles is used.* In a scanned-beam system only 30 watts of RF power are needed for a 600 Mb/s transmit beam.

Since the weight required for electrical power generation by solar cells scales linearly with power, the scanning spot-beam concept potentially offers a 100-fold decrease in the weight required for electrical power compared with an area coverage system operating with the same signal-to-noise ratio. Thus, it appears that the scanned beam offers significant satellite weight savings, provided the scanning system can be realized without an exorbitant cost in weight.

The weight savings may be utilized in a number of ways. An important first option might be to choose a smaller, cheaper booster to launch the payload. A second option might be to increase the satellite transmitter power, consequently allowing smaller earth station attennas, or third, to attempt to increase the satellite capacity. A technique which readily lends itself to increased capacity is to combine a fixed spot-beam and scanning spot-beam system. By letting the scanning beams (one transmit and one receive) occupy one polarization they can be dedicated to serving the low-traffic areas, while cross-polarized fixed spot-beams would be concentrated on the major metropolitan areas. These spot beams would be spatially separated far enough from one another to allow complete reuse of the frequency spectrum; as many as 10 simultaneous reuses of the frequency band may be possible.

Let us see how the advantages of this technique might effect an overall system. In Table I below we have selected typical values of some key elements of the earth-space link for both an area coverage and a scanning spot beam system. Because of reciprocity the 20 dB advantage for the scanning spot beam is enjoyed on both the up-link and down-link. Obviously, there are many tradeoffs to be considered among the various link parameters, even for an area coverage satellite system; in the examples given below the major consideration was to provide a system capable of serving many low-cost earth stations. Employing several fixed beams together with the scanning spot-beam system, a 10-fold increase in capacity is possible, and moreover, earth-station antenna size can be significantly reduced while providing the same signal-to-noise ratio as a conventional area coverage system.

For the remainder of this paper we will concentrate only on the scanning spot-beam portion of the system and defer consideration of combinations with fixed spot-beams to a later publication. In the next section we shall describe the system concepts and in particular the burst formats and timing organization. In Section III, the formation of rapidly

* The high-power Japanese Broadcast Satellite[6] generates dc power at 0.23 lb/W. Allowing a 40 percent overall transmission efficiency, 7.5 kW (or 1725 lb) are needed for an average coverage satellite with high rain margins. As a comparison, the Thor-Delta 3914 rocket provides about 400 lbs payload for the communication and power supply packages. The Atlas-Centaur provides about 800 lbs. Significantly higher payloads will be possible with the advent of the Space Transportation System.

Table I — Example differences of key elements in 12/14-GHz satellite systems when one has a 20-dB advantage in the link budget

	Area coverage system	Fixed and scanning spot-beam system	dB difference
Up-link:			
Earth station antenna (meters)	6	2.25	8.5
Earth station transmitter (watts)	500	35	11.5
			20
Down-link:			
Earth station antenna (meters)	6	2.25	8.5
Receiver noise temperature (kelvin)	200	280	1.5
Satellite transmitter power (watts/500 MHz)	300	30	
Total transmitted power (watts)	300	300	
Capacity (Mb/sec)	600	6000	10.0
			20

scanned beams is discussed. The array design and its performance is examined in Section IV.

II. TDMA BURST ORGANIZATION

There may be hundreds of ground stations in a scanned spot-beam system. For example, with 100 ground stations in the system, the number of possible distinct links is 4950 pairs. Of course, at any particular time the total number of connected links may be far less than this, and the number of channels required between various pairs of earth stations would be by no means equal. We shall discuss one possible organization format that provides the connections among the ground stations in the following paragraphs.

To illustrate a possible organization of such a system let us refer to Fig. 1. Shown here in the time domain are time-interleaved bursts from 100 ground stations which are repeated at a frame length T. Each burst occupies a time length τ_k and consists of preambles as well as data streams for all other earth stations as illustrated by the burst τ_2 in Fig. 1. The preamble enables carrier and timing recovery on the satellite. At the satellite, the digital bursts are detected and remodulated onto a carrier and are sent down to the ground stations via the scanned spot-beams as shown in the time-sequence plot of Fig. 2. Consider burst τ_1, which consists of many subbursts intended for different ground stations. The scanned spot-beam has to be formed and moved fast enough at the sub-burst rate to illuminate all the ground stations in the duration of the burst length τ_1. Each ground station only receives the intended message; the time domain sequence of the received sequence of subbursts is shown in Fig. 2b. Again, each subburst should carry a preamble to facilitate carrier and timing recovery at the ground station.

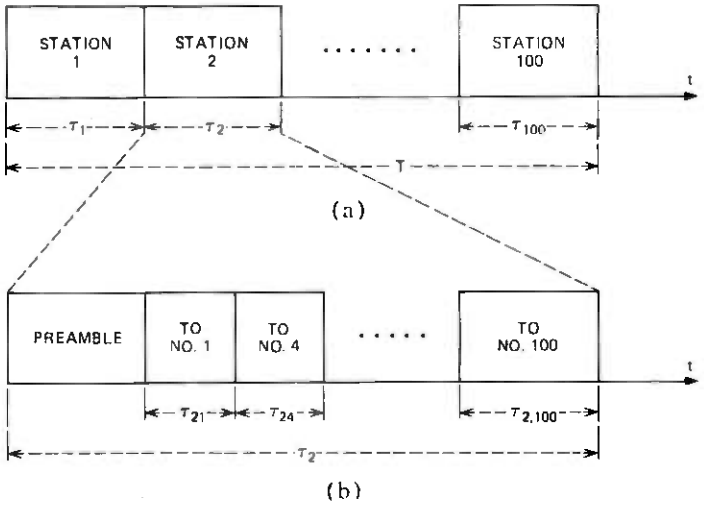


Fig. 1—Uplink frame and burst format. Frame length T, burst length τ_k, Subburst Length τ_{ij}.

With 500-MHz bandwidth available, it is reasonable to assume a bit rate of 600 Mb/s or 300 Mbauds/s using four-phase PSK modulation. Assuming 32 kb/s per channel, the total capacity is 18,800 circuits or 9400 two-way circuits. Allowing the simultaneous participation of 100 ground stations and that each station might communicate with 10 other stations, each burst would then average 94 circuits and each subburst carries only 9.4 circuits. In fact, it is quite possible that some subbursts may carry

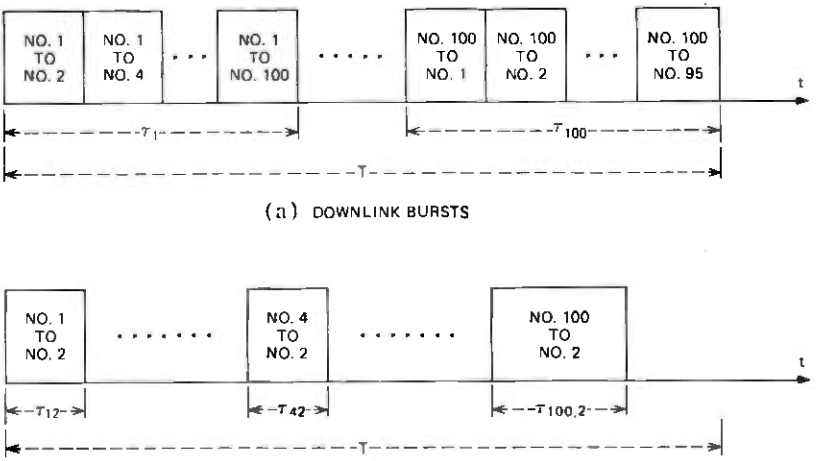


Fig. 2—Downlink burst formats.

only one or two circuits at a time. For a frame length of 125 μsec, i.e., 8-kHz sampling rate, a subburst carrying one voice circuit consists of only 4 bits. This is far less than the preamble requirement and results not only in inefficiency but also in an unrealistically high switching rate of the spot-beam. However, by buffering at ground stations, the frame length may be lengthened by a factor of 100 to, say, 12.5 msec. This added round-trip delay of 50 msec is still small compared to the 480 msec round-trip delay over the satellite path and should not cause significant echo degradation. In this way, each subburst contains a minimum of 400 bits and the necessary preamble 20 to 40 bauds[7] becomes a small penalty, even in the case of single channel subbursts. The required switching time of the spot beams should be achieved in the order of a few bauds, e.g., 10 ns.

The number of bits in a frame is simply 600 Mb/s times 12.5 ms = 7.5×10^6 bits. A station using 1 percent of the capacity of the channel would need to buffer only 150 kb for both up- and down-link transmission. Since 16k bits of memory are available on integrated circuits chips today, the buffer requirement can be readily satisfied with minimal cost and effort.

III. BEAM-FORMING NETWORKS

There are many ways to form rapidly scanned beams.[8] The simplest approach for satellite application is to use a parabolic reflector with multiple feeds as shown in Fig. 3. In Fig. 3a, a 5 × 5 feed horn array is shown at the focal plane of an offset paraboloid. Each feed horn, if singly excited, would produce a main lobe which coincides with the intended coverage area on the ground. Figure 3b illustrates the far-field pattern of two adjacent beams, e.g., beam No. 1 and 2. However, there are significant drawbacks with this approach in that the beam switching must be performed at high power level because all the power is fed into a single horn. As a result, the switching speed and/or drive power presents serious design problems. Furthermore, to produce full area coverage, the adjacent beams overlap at the 3-dB points. This requires an undersized feed horn and thus antenna gain suffers because of spillover loss. Significant cross-coupling loss into the adjacent feed horns further reduces the reflector antenna gain. One possible alternative is to form a beam by simultaneously feeding the center horn and the adjoining horns with reduced magnitude.[9] This reduces the spillover and cross-coupling loss but most of the power is still handled by the center horn. Furthermore, the feed network becomes extremely complicated.

A more attractive approach is to form an array of high-gain elements. For example, employing element patterns covering the United States with 27-dB edge gain, only 100 elements are needed to produce a 47-dB gain beam-forming array. A typical radiated power requirement of 30

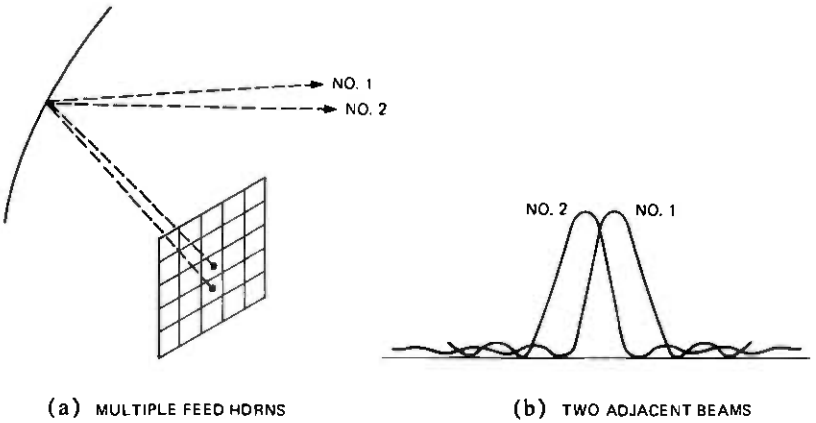


Fig. 3—Beam forming by multiple feeds.

watts is distributed among the 100 elements resulting in 0.3 watts per element. This should permit the use of solid-state microwave power devices such as GaAs FET amplifiers at each element as the final power stage. This may allow a weight reduction and will increase reliability because failure of elements merely reduces the radiated power. Beam forming is easily achievable by microstrip phase shifters which can change state within a few nanoseconds. By placing the phase shifters before the GaAs FET amplifiers at the low-power points, rapid beam switching can be controlled easily with high-speed logic.

IV. ARRAY DESIGN

Phased arrays have some characteristics different from reflector antennas that affect their performance. When a phased array is scanned off-axis there is a difference in path length between the array edge and its center. This limits its useful bandwidth. Also, it is most convenient to form a beam using discrete phase steps, and using steps which are too coarse will reduce the array gain. Another source of gain degradation arises when elements fail. Finally, component phase drift may make it impossible to form beams in an open-loop manner.

To treat the above topics in a quantitative manner is beyond the scope of this brief paper, and they will be published at another time. We have calculated, however, that a 120λ aperture phased array scanning ± 3 degrees would satisfy the bandwidth requirements of 500 MHz at 12-GHz carrier frequency with little degradation. Such an antenna would serve the continental United States from geosynchronous orbit.

Since we would envision that the phase shifter settings for all the possible beam-pointing angles would be stored in a digital memory, the values for the individual phase shifts necessarily become quantized. It

is interesting to note that very coarse approximations to the precise phase settings result in very little gain degradation. For example, quantizing the phase into one of four quadrants (± 45 degrees) results in an expected value of on-axis gain decrease of less than 1 dB for a 100-element array. Other considerations such as sidelobe performance and high assurance of little gain loss for any scan angle will probably dictate phase shifters quantized to either ±22.5 or ±11.25 degrees. This results in a storage requirement of 3 or 4 bits per element per beam position. Thus, on-board the satellite 30 to 40 thousand bits of memory are required for a 100-element array to scan to 100 positions. Since upwards of 16,000 bits are readily available on a single memory chip, this is a very modest requirement.

Let us consider briefly the array gain degradation due to failure of elements. Denote the gain of an N-element array by N. Let each element radiate unity power, so that the EIRP in the main beam direction is $P_o = N^2$. If M elements fail, the array gain reduces to N-M and the EIRP becomes $(N\text{-}M)^2$. We are assuming, of course, that there is no mutual coupling between elements, which is reasonable because the aperture size of the elements is large. Thus, the EIRP for the case of failed elements becomes

$$P = P_o \left(\frac{N\text{-}M}{N}\right)^2$$

It is interesting to note the above equation is identical to the failure performance of cascaded hybrid power combiners. If 10 percent of the elements fail, 19 percent of the radiated power would be lost compared with a perfectly functioning phased array.

In both the case of failed elements and discrete phase-shift settings, the sidelobe performance may be adversely affected. This does not pose a significant problem in these considerations because only one spot-beam is contemplated. However, for a more sophisticated system which would have two or more cochannel spot-beams, the questions of sidelobe performance and mutual interference would have to be seriously addressed.

For a synchronous satellite located at 98° E longitude (mid-U.S.A.), the continental U.S.A., when viewed from the satellite, spans about 6 degrees in the east-west direction and 3 degrees in the north-south direction as shown in Fig. 4. We want to concentrate the array radiated power into the main beam so that for a given gain, the spot-beam will cover the largest area. This would reduce the number of spot-beams required for total U.S.A. coverage. The above requirement implies arrays with closely packed elements. Thus the use of random arrays or other forms of thinned arrays is ruled out of our considerations and a periodic array is more appropriate.

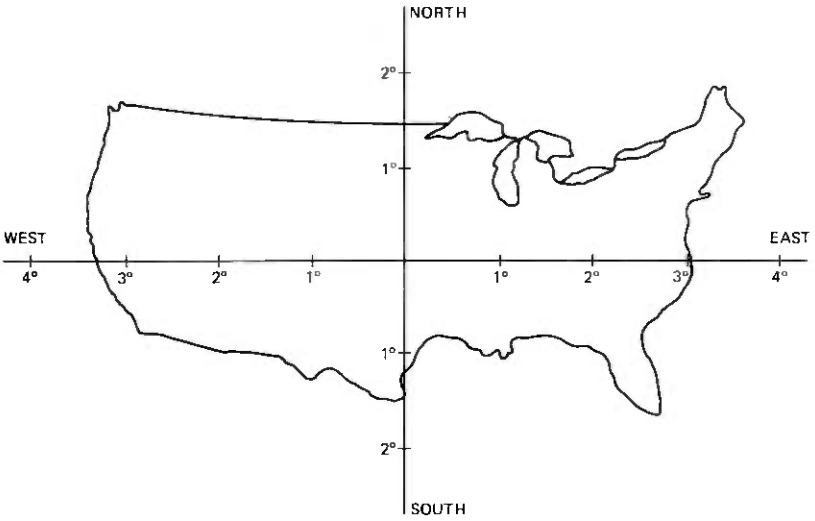


Fig. 4—The United States as viewed by a synchronous satellite at 98° E pointed at 98° E and 36° N.

To reduce the number of array elements we use high gain elements for the individual radiators. This invariably leads to grating lobes. To avoid grating lobes falling on the continental United States while the array is scanning across the desired coverage area, the grating lobes should be at least 3 degrees apart in N-S direction and more than 7 degrees in E-W direction. Recall that a grating lobe angle, ψ, is related to element spacing, d, by $\psi = \lambda/d$ radians, the maximum allowable spacings are:

$$d_{\text{E-W}} = \frac{\lambda}{\pi}\frac{180}{7} = 8.2\lambda$$

$$d_{\text{N-S}} = \frac{\lambda}{\pi}\frac{180}{3} = 19.1\lambda$$

The array is shown in Fig. 5 with element spacings prescribed by the above equations. The elements are rectangular in shape with dimensions of 8.2λ and 19.1λ. A pyramidal horn antenna would be one simple method of realizing the array element. A more attractive antenna design would be one which accommodates spot-beams on one polarization and the scanning beams on the other. This work will be published later by other authors.

The 3-dB beamwidths expressed in radians of these elements in the two principal planes are approximately

$$\theta_{3\text{dB}} = 1.2\ \lambda/d$$

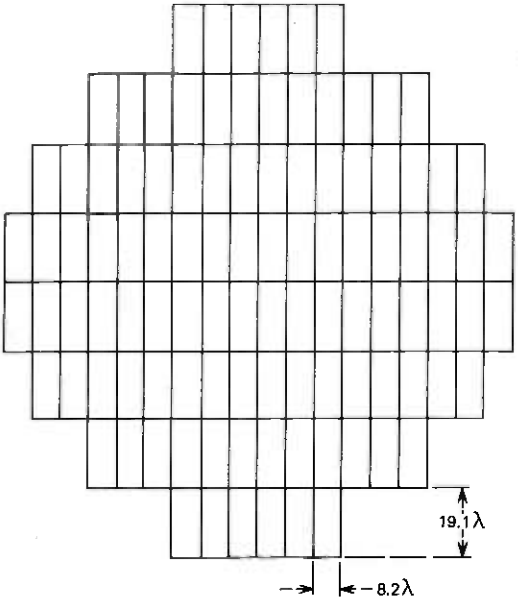


Fig. 5—A sample 104-element array.

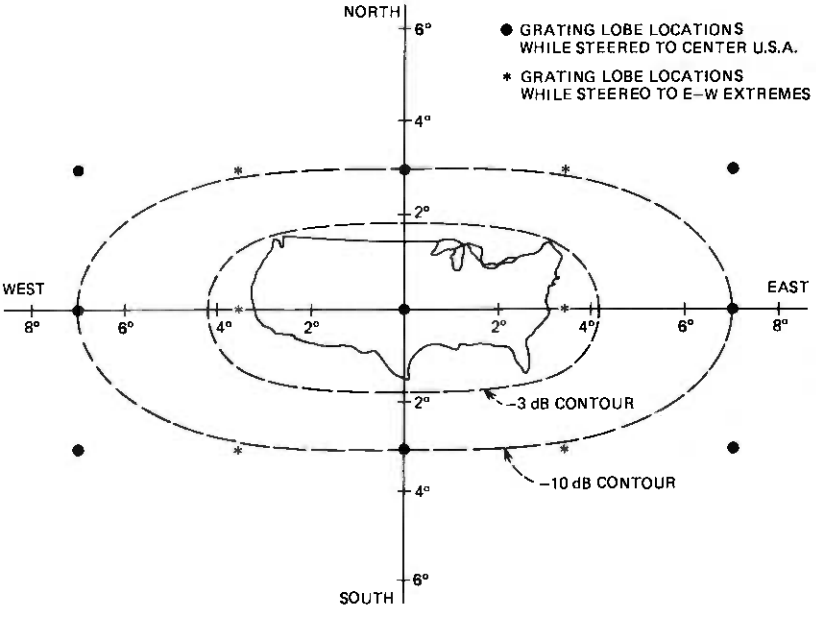


Fig. 6—Element pattern and grating lobes locations.

For the dimensions given, $\theta_{3dB(E\text{-}W)} = 8.4$ degrees and $\theta_{3dB(N\text{-}S)} = 3.6$ degrees. Assuming 50 percent aperture efficiency the gain of this high-gain element would be

$$G = 10 \log \left(0.5 \frac{4\pi A}{\lambda^2}\right) = 29.9 \text{ dB}$$

In Fig. 6, the equal intensity contour of the individual radiator, extrapolated from Silver,[10] is plotted over a map of the U.S.A. Also shown are the grating lobe locations when the array beam is pointed toward mid-U.S.A. For off-center scanning beams, the grating lobe locations are shifted according to the angles scanned. It can be seen that grating lobes will be outside of the United States for all cases as expected. The array gain when all elements are equally excited is 50 dB at the center of the United States and will drop to 47 dB when scanned to the 3-dB edge of the element pattern.

The array shown is basically circular in shape; this has reduced sidelobes compared with a rectangular array. Further control of the sidelobes is possible by a minor amount of spatial tapering of the intensity of the array excitation, but of course gain will be sacrificed. A more detailed study is needed to determine the optimal design.

V. CONCLUSIONS

The best approach for digital communications among multiple earth stations with varying traffic requirements appears to be Time-Division Multiple Access (TDMA). In an area coverage concept all earth stations time-share a single up-link channel; a single antenna broadcasts all messages on a common down-link channel and each station selects only those messages intended for it. In this paper we propose using a movable spot-beam to radiate to each earth station consistent with the TDMA approach. With a reasonable-size aperture antenna it is estimated for the equivalent SNR of an area coverage antenna that approximately 20 dB can be saved in the link budget. This savings can be advantageously applied to reduce the satellite transmitter power, increase its capacity, and significantly reduce the size of the earth station antennas.

A TDMA burst organization is proposed, and estimates of burst lengths, beam-switching intervals, and buffer storage size are made for a 100-earth-station network operating on a 600-Mb/s channel; all requirements for operating such a system appear feasible and within the state of today's art. Two approaches for forming rapidly scanning spot-beams were discussed. One approach used a single reflector with a multiple feed-horn array; the other employs a phased array with each element radiating the entire U.S. An equally spaced array of rectangular elements arranged inside a circle appears to provide an attractive solution. Using this approach an antenna capable of forming spot-beams with nearly 50-dB gain

in any direction within ± 3 degrees of center within 10 ns appears feasible.

VI. ACKNOWLEDGMENT

The authors wish to express their thanks to N. Amitay, M. J. Gans, and C. Dragone for many helpful discussions on antennas and arrays; also we thank A. A. Penzias for aiding in clarifying many concepts in this manuscript and for his enthusiastic support of this project.

REFERENCES

1. D. O. Reudink, "A Digital 11/14 GHz Multibeam Switched Satellite System," AIAA/CASI 6th Communication Satellite Systems Conference, Montreal, April 5–8, 1976.
2. D. Jarett, "A Baseline Domestic Communications Satellite System for the 1980's," AIAA/CASI 6th Communication Satellite Systems Conference, Montreal, April 5–8, 1976.
3. E. A. Ohm, "A Proposed Multi-Beam Microwave Antenna for Earth Stations and Satellites," B.S.T.J., *53*, No. 8 (October 1974), pp 1657–1665.
4. D. O. Reudink, Y. S. Yeh, and A. S. Acampora, "Spectral Reuse in 12 GHz Satellite Communication Systems," International Communications Conference Record, *3*, No. 75-CH1209-6 CSCB, pp 27.5-32–37.5-35, June 12–15, 1977.
5. A. J. Rustako, Jr., "The 12 GHz Crawford Hill Receiving System for Use with the CTS Satellite Beacon," to be published.
6. "A High-Powered Satellite for Communication Applications," General Electric Co. brochure PIB-A-86(8-75)-5M.
7. P. P. Nuspl, K. E. Brown, W. Steenaart, and B. Ghicopoulas, "Synchronization Methods for TDMA," Proc. IEEE, 65, No. 3 (March 1977), pp. 434–444.
8. A. A. Oliver and G. H. Knittel, *Phased Array Antennas*, Dealham, Mass: ARTECH House, 1972.
9. A. R. Dion, "Variable-Coverage Communication Antenna for LES-7," Communication Satellites for the '70s, Progress in Astronautics and Aeronautics, *25*, Alpine Press, 1971, pp 255–276.
10. S. Silver, *Microwave Antenna Theory Design*, New York: Dover Pub., 1965, p. 451.

Contributors to This Issue

Donald Bock, Union County Technical Institute, 1963; Bell Laboratories, 1963—. Associate member of technical staff in the Acoustics Research Department. Mr. Bock has worked on the design of computer interfaces for speech coding and voice response systems.

Deborah K. Christopher, B.S. (E.E.C.S.), 1976, Massachusetts Institute of Technology; summer research associate in the Acoustics Research Department in 1975 and 1976. Currently a graduate student in the Electrical Engineering Department of the Massachusetts Institute of Technology. Member, Eta Kappa Nu, Sigma Xi.

Fan R. K. Chung, B.S., 1970, National Taiwan University; Ph.D., 1974, University of Pennsylvania; Bell Laboratories, 1974–. Mrs. Chung's current interests include combinatorics, graph theory, and the analysis of algorithms. She is presently investigating various problems in the theory of switching networks.

Peter S. Cross, B.S.E.E., 1968, California Institute of Technology; M.S., 1969, Ph.D., 1974, acting assistant professor, 1974, University of California, Berkeley; member of technical staff, Institut für Angewandte Festkörperphysik der Fraunhofer—Gesellschaft, Freiburg, W. Germany, 1974–1975; Bell Laboratories, 1975–. At the University of California, Mr. Cross engaged in studies of the optical properties of semiconductors with emphasis on the 6 to 12 μm region. At the Institut in Freiburg, he studied the basic properties of surface acoustic wave resonators. He is currently in the Coherent Optics Research Department working on optical and acoustical guided-wave devices.

B. R. Eichenbaum, B. S. 1963, City College of New York; M. S. 1965, Ph.D. 1969, New York University; Bell Laboratories, 1972—. Mr. Eichenbaum has worked on a variety of optical fiber development problems in the areas of coating materials, coating application techniques, ribbon fabrication, and fiber mechanics. He is currently developing field stripping procedures. Member, OSA.

Wolfgang B. Elsner, B.S. (E.E.), 1965, M.S. (E.E.), 1967, Ph.D., 1974, Clarkson College of Technology; NASA, Goddard Space Flight Center, Greenbelt, Md., 1966–1967; Instructor at Clarkson College of Technology, Potsdam, N.Y., 1967–1973; Bell Laboratories, 1973—. Since joining Bell Laboratories, Mr. Elsner has been concerned with various aspects of trunk-network engineering. Member, IEEE, Eta Kappa Nu, Sigma Xi.

Paul F. Goldsmith, B.A. (physics), University of California, Berkeley 1969; Ph.D. (physics), University of California, Berkeley, 1975; Bell Laboratories, 1975–. His graduate work was concerned primarily with building a receiver system to study the spectral line emission from the $J = 2 \rightarrow 1$ transition of CO ($\nu = 230$ GHz) and using it to analyze the structure of interstellar molecular clouds. He also worked on analyzing the excitation of rotational transitions of molecules such as CO under interstellar conditions. At Bell Laboratories, he has worked on the radioastronomical feed system for the Crawford Hill millimeter-wave antenna and also on several astrophysical subjects including isotopic abundance variations, and the excitation of molecules by electrons in interstellar clouds. Member, American Astronomical Society, IEEE, and Sigma Xi.

Frank K. Hwang, B.A., 1960, National Taiwan University; M.B.A., City University of New York; Ph.D. (Statistics), 1968, North Carolina State University; Bell Laboratories, 1967—. Mr. Hwang spent the fall of 1970 visiting the Department of Mathematics of National Tsing-Hua University. He has been engaged in research in statistics, computing science, discrete mathematics, and switching networks.

F. W. Mounts, E. E., 1953; M.S., 1956, University of Cincinnati; Bell Telephone Laboratories, 1956—. Mr. Mounts has been concerned with research in efficient methods of encoding pictorial information for digital television systems. Member, Eta Kappa Nu; Senior Member, IEEE.

Arun N. Netravali, B. Tech. (Honors), 1967, Indian Institute of Technology, Bombay, India; M.S., 1969 and Ph.D. (E.E.), 1970, Rice University; Optimal Data Corporation, Huntsville, Alabama, 1970–1972; Bell Laboratories, 1972—. Mr. Netravali has worked on various aspects of signal processing. Member, Tau Beta Pi, Sigma Xi.

Birendra Prasada, B.S., 1953, M.S., 1955, Banaras University; Ph.D., 1960, University of London; Central Electronics Engineering Research Institute, Pilani, India, and Defence Science Laboratory, Delhi, India, 1961–1963; Massachusetts Institute of Technology, 1965–1966; Indian Institute of Technology, 1968–1972; Bell Laboratories, 1963–1965, 1973–1976; Bell Northern Research, 1976—. Mr. Prasada's main research and teaching interests are in the areas of visual communications, systems engineering, systems design, and human communication. He has worked as an industrial consultant in India and the United States. Member, 1963, Senior Member 1976, IEEE.

Lawrence R. Rabiner, S.B., S.M, 1964, Ph.D., 1976, Massachusetts Institute of Technology. Bell Laboratories, 1962—. He worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently he is engaged in research on speech communications and digital signal processing techniques. Member, Eta Kappa Nu, Sigma Xi, Tau Beta Pi; fellow, Acoustical Society of America, IEEE.

Douglas O. Reudink, B.A. (physics), 1961, Linfield College; Ph.D., 1965 (mathematics), Oregon State University; Bell Laboratories, 1964—. Mr. Reudink has been engaged in electronics systems research; from 1965–1972 his research efforts were concerned with mobile communications. In 1972, he was appointed Head, Satellite System Research Department. He is responsible for fundamental studies related to the use of satellites in future communication systems.

M. R. Santana, B.S.E.E., 1970, University of Hartford; M.S.E.E., 1971, Georgia Institute of Technology; Bell Laboratories, 1970—. Mr. Santana has been continuously involved in cable design and development in the Loop Transmission Division. At present he is involved in optical fiber cable design, analysis, and testing. Member, IEEE, Kappa Mu.

Ronald V. Schmidt, B.S., 1966, Ph.D. 1970, University of California, Berkeley; postdoctoral research assistant, University College, London, 1970–71; Bell Laboratories, 1971–. Mr. Schmidt is presently engaged in research on guided optical wave and acoustic surface wave devices.

Philip F. Schweitzer, B.S.E.E., 1968, Newark College of Engineering; M.S.E.E., 1972, Monmouth College; Bell Laboratories, 1972—. Mr.

Schweitzer, a member of the Billing and Local Switching Systems Department, is engaged in the formulation of a system philosophy for mechanizing the message accounting billing process in the 1980s. Previously he worked on developing techniques for improving revenue recovery through the use of computerized support systems.

Fred D. Waldhauer, B.E.E., 1948, Cornell; M.S.E.E., 1960, Columbia; RCA, 1948–1955; Bell Laboratories, 1956—. At Bell Laboratories, Mr. Waldhauer has worked on the T1 digital transmission system and analog/digital converters for television signals. He later supervised the development of high-speed PCM repeaters for digital coaxial transmission, including final development of repeaters for the T4M 274-Mb/s system. He is currently engaged in applications of fiber optics to telecommunications. Fellow, IEEE; registered professional engineer.

Lynn O. Wilson, A.B. (physics), 1965, Oberlin College; Ph.D. (applied mathematics), 1970, University of Wisconsin; Bell Laboratories, 1970—. Ms. Wilson has pursued research in various areas of applied mathematics. She has worked on problems concerning *PICTUREPHONE*® service demand, electromagnetic theory, dielectric waveguides, elastic waves, crystalline vibrations, and the growth of semiconductor crystals. Member, Sigma Xi, American Physical Society, SIAM.

Y. S. Yeh, B.S.E.E., 1961, National Taiwan University; M.S., 1964, and Ph.D., 1966, University of California at Berkeley; Chinese Navy 1961–1962; Harvard University, 1966–1967; Bell Laboratories, 1967—. Mr. Yeh was engaged in mobile radio research from 1967–1972. Since that time he has worked on communication satellite systems.

Papers by Bell Laboratories Authors

BIOLOGY

Linear Electric Field Effects in Electron Paramagnetic Resonance for Two Bis Imidazole Heme Complexes, Model Compounds for B and H Hemichromes of Hemoglobin. W. B. Mims, J. Peisach, Biochem. J., *16* (1977), pp. 2795–2799.

CHEMISTRY

Sub-Picosecond Relaxation of Large Organic Molecules in Solution. C. V. Shank, E. P. Ippen, and O. Teschke, Chem. Phys. Lett., *45*, No. 2 (January 15, 1977), pp. 291–294.

Comment on Host Nuclear Resonance in a Spin Glass: CuMn. D. A. Levitt, R. E. Walstedt, Phys. Rev. Lett., *38*, No. 4 (January 24, 1977), pp. 178–181.

Evidence for Crystalline Electric Field and Spin-Orbit Splittings for Co Cimpurities in Au. R. Dupress, R. E. Walstedt, W. W. Warren, Jr., Phys. Rev. Lett., *38*, No. 11 (March 14, 1977), pp. 612–615.

^{23}Na Nuclear Relaxation in Naβ-Alumina: Barrier Height Distributions and the Diffusion Process. R. E. Walstedt, R. R. Dupree, J. P. Remeika, A. Rodrigues, Phys. Rev. B, *15*, No. 7 (April 1, 1977), pp. 3442–3454.

Scanning Electron Microscopy of Liquid Mercury-Solid Metal Interactions. J. E. Bennett, Microstructural Sci., ed. by J. D. Braun, H. W. Arrowsmith, J. L. McCall, Elsevier North-Holland, Inc., *5* (1977), pp. 395–402.

^{77}Se Nuclear Magnetic Resonance in $TiSe_2$. R. Dupree, W. Warren, F. J. DiSalvo, Bull. Amer. Phys. Soc., *22* (March 1977), p. 281.

COMPUTING

Algorithms and Computational Complexity. Alfred V. Aho, Acta Crystallgr., *33* (1977), pp. 5–12.

Code Generation for Expressions with Common Subexpressions. A. V. Aho, S. C. Johnson, J. D. Ullman, J. Ass. Comput. Mach., *24*, No. 1 (January 1977), pp. 146–160.

MATERIALS SCIENCE

Correlation Between Leaking Glass/Metal Seals and Wire Defects in Dry Reed Contacts. M. R. Pinnel, J. E. Bennett, F. E. Bader, J. Test. Eval., *5* (January 1977), pp. 30–42.

Corrosion of Solder-Coated TiPdAu Thin Film Conductors in a Moist Chlorine Atmosphere. F. N. Fuss, C. T. Hartwig, J. M. Morabito, Thin Solid Films, *43* (1977), pp. 189–213.

Effect of Surface Stress on the Natural Frequency of Thin Crystals. M. E. Gurtin, X. Markenscoff, R. N. Thurston, Appl. Phys. Lett., 29 (1976), pp. 529–530.

Effects of Hydrostatic Pressure on the Magnetic Properties of Disordered Monosilicide $Fe_xCo_{1-x}Si$ Alloys. J. Beille, D. Bloch, V. Jaccarino, J. H. Wernick, G. K. Wertheim, J. Phys. (Paris), *38* (March 1977), pp. 339–343.

Electrical Characterization of Vapor-Phase-Epitaxially Grown Large Area n-AlAs/p-GaAs Heterojunctions. W. D. Johnston, IEEE Trans. Electron. Dev., *ED-24* (February 1977), pp. 135–138.

The Electron Stimulated Interaction of H_2O with a Nickel Surface. H. G. Tompkins, Surf. Sci., *62* (January 1977), pp. 293–302.

Empirical Electron Backscatter Model for Thin Resist Films on a Substrate. R. D. Heidenreich, J. Appl. Phys., *48* (April 1977), pp. 1418–1425.

Epitaxial Growth of High T_c Superconducting Nb_3Ge on Nb_3Ir. A. H. Dayem, T. H. Geballe, R. B. Zubeck, A. B. Hallak, G. W. Hull, Appl. Phys. Lett., *30*, No. 10 (May 15, 1977), pp. 541–543.

Growth and Properties of Liquid-Phase Epitaxial $GaAs_{1-x}Sb_x$. R. E. Nahory, M. A. Pollack, J. C. DeWinter, K. M. Williams, J. Appl. Phys., *48* (April 1977), pp. 1607–1614.

Low Noise and High Power GaAs Microwave Field-Effect-Transistors Prepared by Molecular Beam Epitaxy (MBE). A. Y. Cho, J. V. DiLorenzo, B. S. Hewitt, W. C. Niehaus, W. O. Schlosser, C. Radice, J. Appl. Phys., *48* (January 1977), pp. 346–349.

Low Temperature Heat Capacity of Alkali and Silver β-Aluminas. D. B. McWhan, C. M. Varma, F. L. S. Hsu, J. P. Remeika, Phys. Rev. Lett., *B15* (January 15, 1977), pp. 553–560.

Nuclear Modulation of the Electron Spin Echo Envelope in Glassy Materials. W. B. Mims, J. Peisach, J. L. Davis, J. Chem. Phys., *66*, No. 12 (June 15, 1977), pp. 5536–5550.

Preparation of CdS/InP Solar Cells by Chemical Vapor Deposition of CdS. M. Bettini, K. J. Bachmann, J. L. Shay, S. Wagner, J. Appl. Phys., *48* (April 1977), pp. 1603–1606.

Solution Decomposition Behavior of Chemically Modified Poly(vinyl chloride). I. M. Plitz, R. A. Willingham, W. H. Starnes, Jr., Macromolecules, *10* (March/April 1977), pp. 499–500.

Spherulitic Crystallization in Blends of Poly(vinylidene fluoride) and Poly(methyl methacrylate). T. T. Wang, T. Nishi, Macromolecules, *10* (March 1977), pp. 421–425.

Transistors with Boron Bases Predeposited by Ion Implantation and Annealed in Various Oxygen Ambients. T. E. Seidel, R. S. Payne, R. A. Moline, W. R. Costello, J. C. C. Tsai, K. R. Gardner, IEEE Trans. Electron. Dev., *ED-24*, No. 6 (June 1977), pp. 717–722.

Water Tree Growth in Polyethylene Under DC Voltage Stress. E. A. Franke, J. R. Stauffer, E. Czekaj, IEEE Trans. Elec. Insul., *EI-12* (June 1977), pp. 218–223.

GENERAL MATHEMATICS AND STATISTICS

Solution of "Solvable Model of a Spin Glass." D. J. Thouless, P. W. Anderson, R. G. Palmer, Phil. Mag., *35* (1977), pp. 593.

The Spanning Subgraphs of Eulerian Graphs. F. Boesch, C. Suffel, R. Tindell, J. Graph. Theory., *1*, No. 1 (April 1977), pp. 79–84.

ELECTRICAL AND ELECTRONIC ENGINEERING

Experimental Fiber-Optic Transmission System for Interoffice Trunks. T. L. Maione, D. D. Sell, IEEE Trans. Commun., *COM-25*, No. 5 (May 1977), pp. 517–523.

A Model for Advanced Reservations for Intercity Visual Conferencing Services. Hanan Luss, Oper. Res. Quart., *28* (May 1977), pp. 275–284.

A Multi-Microphone Signal Processing Technique to Remove Room Reverberation from Speech Signals. J. Allen, D. Berkley, J. Blauert, J. Acoust. Soc. Amer. (October 1977).

Optimal Reception of Digital Data Over the Gaussian Channel with Unknown Delay and Phase Jitter. D. D. Falconer, J. Salz, IEEE Trans. Inform. Theory, *IT-23*, No. 1 (January 1977), pp. 117–126.

PHYSICS

Determination of the Valence Band Structure of InSe by Angle-Resolved Photoemission Using Synchrotron Radiation. P. K. Larsen, S. Chiang, N. V. Smith, Phys. Rev. Lett., *B15* (March 15, 1977), pp. 3200–3210.

Discovery of Optical Emission in Radio Lobes of Double Radio Galaxies. W. C. Saslaw, J. A. Tyson, P. Crane, Astron. Astrophys., *59* (July 15, 1977), pp. L15–L16.

Evidence for Crystalline Electric Field and Spin-Orbit Splittings for Co Impurities in Au. R. Dupree, R. E. Walstedt, W. W. Warren, Jr., Phys. Rev. Lett., *38* (March 14, 1977), pp. 612–615.

Frequency-Frequency Correlations of Ocean Ambient Noise Levels. R. H. Nichols, C. E. Sayer, J. Acoust. Soc. Amer., *61* (May 1977), pp. 1188–1190.

Intensity Calibration of Millimeter-Wave Spectrometers. P. S. Henry, Astrophys. Lett., *18* (February 1977), pp. 75–78.

Miniature Plane Mirror Analyzer Suitable for Angle-Resolved Photoelectron Spectroscopy. N. V. Smith, P. K. Larsen, M. M. Traum, Rev. Sci. Instrum., *48* (April 1977), pp. 454–459.

A New Opto-Acoustic Cell with Improved Performance. C. K. N. Patel, R. J. Kerl, Appl. Phys. Lett., *30*, No. 11 (June 1, 1977), pp. 578–579.

NMR in Liquid Selenium to High Temperature and Pressure. W. Waren, R. Dupree, Bull. Amer. Phys. Soc., *22* (March 1977), pp. 385.

Nonlinear Optical Response of Metal-Barrier-Metal Junctions. B. Fan, S. M. Faris, T. K. Gustafson, T. J. Bridges, Appl. Phys. Lett., *30*, No. 4 (February 15, 1977), pp. 177–179.

A Study of the Optical Emission from an RF Plasma during Semiconductor Etching. W. R. Harshbarger, R. A. Porter, T. A. Miller, P. Norton, Appl. Spectrosc., *31* (May/June 1977), pp. 201–207.

Synchrotron Radiation Photoemission Spectroscopy of III–VI Compounds. G. Margaritondo, J. E. Rowe, S. B. Christman, Phys. Rev. B, *15* (April 15, 1977), pp. 3844–3854.

Textures and NMR in Superfluid ^{3}He(B). H. Smith, W. F. Brinkman, S. Engelsberg, Phys. Rev. Lett., *15*, No. 1 (January 1, 1977), pp. 199–213.

A Time Dispersion Tuned Fiber Raman Oscillator. R. H. Stolen, C. Lin, R. K. Jain, Appl. Phys. Lett., *30*, No. 7 (April 1, 1977), pp. 340–342.

SYSTEMS ENGINEERING AND OPERATIONAL RESEARCH

Inspection Policies for a System Which is Inoperative During Inspection Periods. H. Luss, AIIE Trans., *9* (June 1977), pp. 189–194.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 8, October 1977
Printed in U.S.A.

B.S.T.J. BRIEFS

Use of Variable-Quality Coding and Time-Interval Modification in Packet Transmission of Speech

By S. A. WEBBER, C. J. HARRIS, and J. L. FLANAGAN

(Manuscript received December 14, 1976)

Speech transmission by switched digital packets offers several opportunities for increasing the utilization of transmission capacity. We comment here upon a combination of variable-quality coding and time-interval modification that can efficiently load a transmission facility and accommodate fluctuating demands on it.

Consider, typically, that a conventional voice switch detects speech energy bursts and demarks each as a packet. A time stamp is given to each packet, and the interburst silences are discarded. Each packet is digitally encoded with a quality that reflects service demands being made on the transmission facility at the moment. Coding bit rate and time-stamp are written in the header data for each packet, along with necessary supervisory information, such as destination and source addresses. Successive packets are assembled in a transmit buffer and are transmitted when capacity is available. Figure 1 illustrates the process.

At the receiver, arriving packets are accepted into a receive buffer. The receiver decodes each packet (in accordance with the header bit rate), reassembles the packets in temporal order (according to the time-stamp), and reinserts the silent intervals, not necessarily exactly as in the original, but with a variation that is perceptually acceptable.

Relevant design questions include: (*i*) how much saving in transmission capacity can be achieved by discarding the silent intervals, (*ii*) what range of signal quality is acceptable in digitally coding the packets, (*iii*) what latitude is perceptually acceptable in reconstructing the speech silent intervals, (*iv*) what total round-trip delay time is allowable in a packet system, (*v*) what transmit and receive buffer sizes are required, and (*vi*) what packet sizes are attractive for transmission economy.

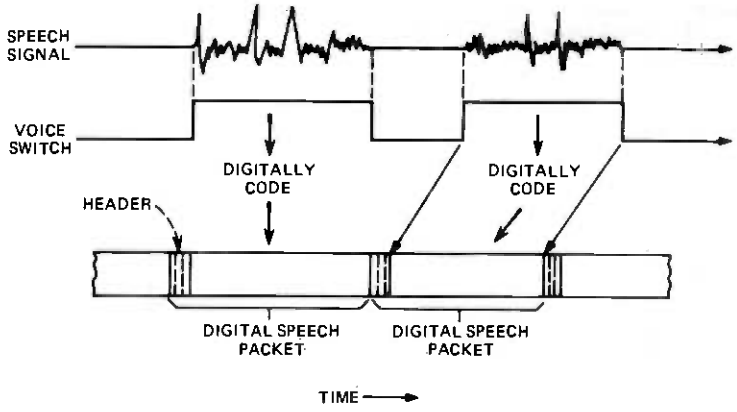


Fig. 1—Speech energy bursts detected by a voice switch, digitally coded, and formed into packets.

Question (*i*) is thoroughly addressed in the extensive literature on Time Assignment Speech Interpolation (TASI) systems. We will add here one more bit of confirmatory data. Questions (*ii*) and (*iii*) dramatically influence the buffer requirements for the system. Our purpose here is to remark about preliminary observations on these issues, and to emphasize these points as candidates for quantitative study.

Extensive data from satellite transmission and echo canceller technology relate to question (*iv*) and suggest that round-trip delays of 0.6 sec, and in some cases up to 1.2 sec, can be used. Questions (*v*) and (*vi*) can properly be addressed only in the context of a complete system design and its optimization. Our remarks, therefore, relate to the points (*i*), (*ii*) and (*iii*). For convenience, all our observations assume that each speech burst is coded as one packet. We implement our experiment by simulation on a laboratory computer, and we process sentence-length signals.

Within-sentence silence time. Our particular voice switch utilizes the Hilbert envelope of the speech signal, and includes a hysteresis logic for positive switch action. The total within-sentence silent time made available by the switching is of course a function of the switch threshold. Too low a threshold provides too little silent time, and too high a threshold eliminates too much signal. Our laboratory observations suggest that within-sentence silent time equal to about 15 percent (of the total sentence duration) can be usefully eliminated. This figure also appears consistent with related studies on voice switching. (Additionally, of course, there are substantial between-sentence silences and natural pauses in conversation flow that can be eliminated.) The sound spectrogram of Fig. 2a shows an input sentence with the significant silent

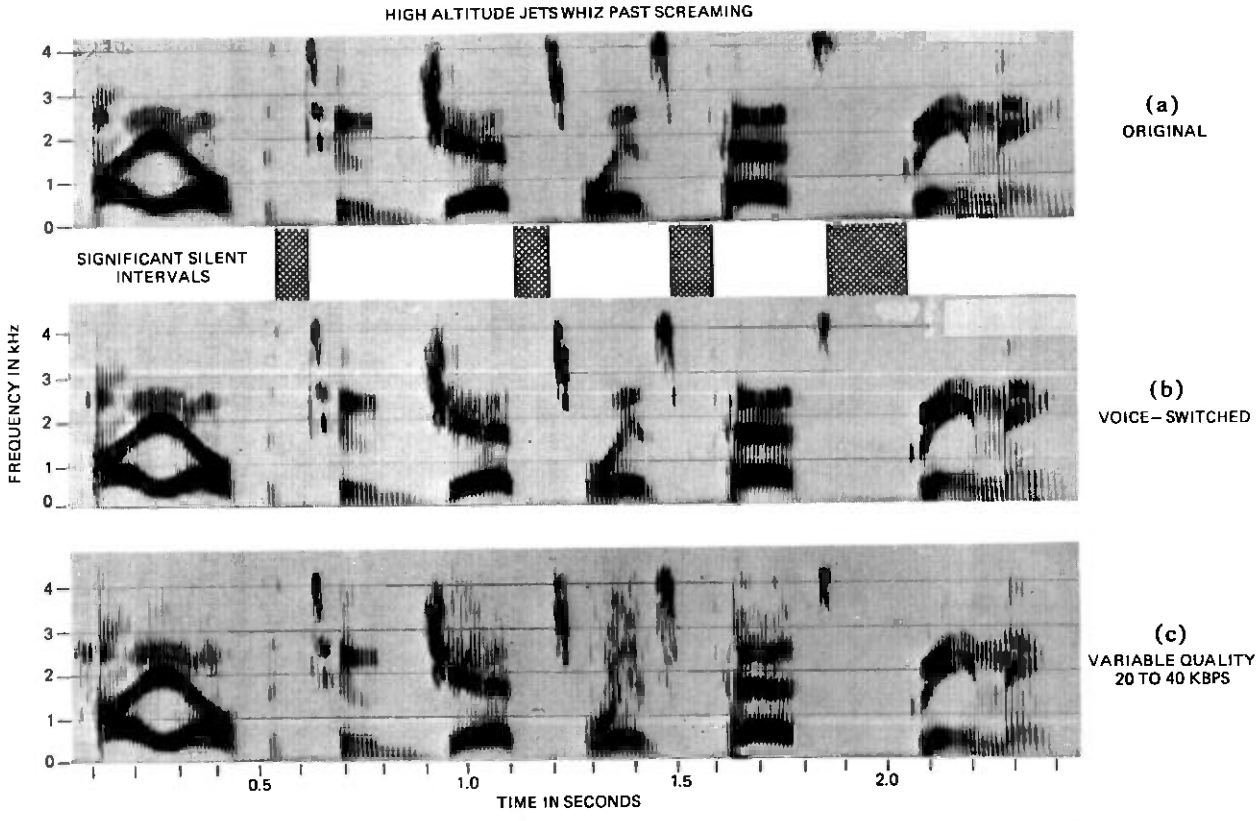


Fig. 2—Experimental transmission of packetized digital speech with variable-quality coding.

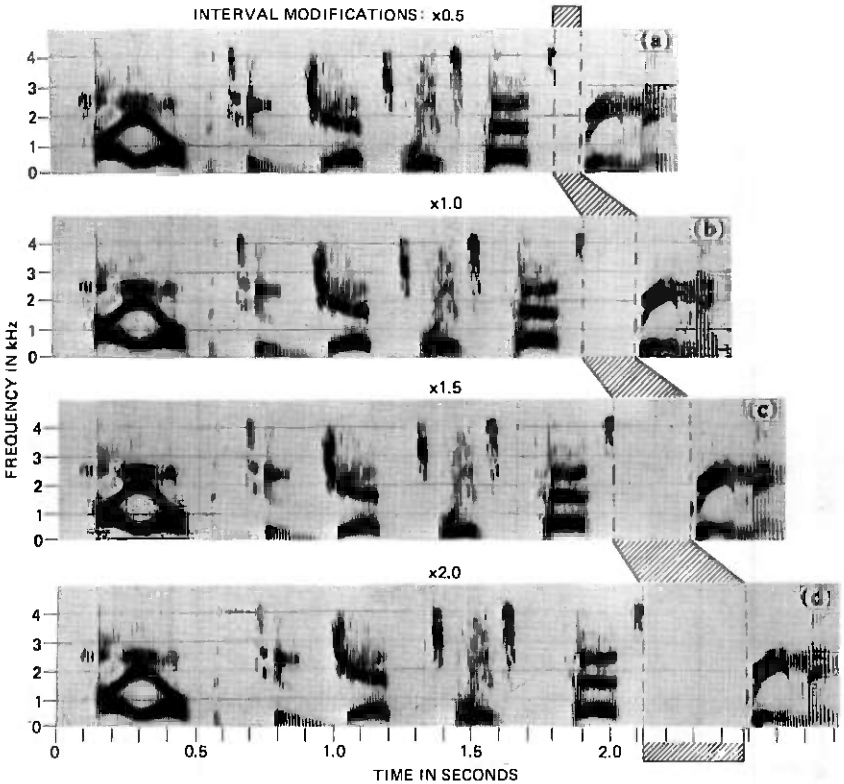


Fig. 3—Modifications of the within-sentence silent intervals for the sentence of Fig. 2. The output is reconstructed from five high-quality packets.

intervals detected by the voice switch. Fig. 2b shows the same signal after passing through the voice switch. In this instance the eliminated silent time is approximately 17 percent of the total duration.

Variable-quality digital coding. We digitally encoded each of the five packets demarked (separated) by the silent intervals in Fig. 2a, using adaptive-differential PCM (ADPCM). The digital coder was also computer simulated. We let the packets be coded successively at bit rates of 40K, 30K, 20K, 30K, and finally back to 40K bits/second, simulating a momentary heavy demand on the transmission system.

The sound spectrogram for this digital coding is shown in Fig. 2c, where the signal packets are reconstructed with silent intervals identical to the original input. One sees that the greatest quantizing noise appears for the momentary quality dip to 20K bits/sec in the third packet. The overall subjective impression of this coding is that the quality is rea-

sonably acceptable.† The perceptual palatability of ADPCM coding also contributes to this result. In this particular instance, the average bit rate for the transmission is 28.6K bits/sec.

Time-interval modification. Latitude in reconstructing the silent intervals in the signal at the receiver can significantly relieve buffer requirements. What modifications in time intervals might be perceptually acceptable? Figure 3 shows receiver reconstruction of high-quality packets with constant, multiplicative modifications of the silent time intervals of 0.5, 1.0, 1.5, and 2.0. (0.0 and 4.0 were also examined, but are not shown here.) One silent interval (the last) is selected and marked for comparison across the signals. Perceptual assessment of these reconstructed packets suggests that interval modifications of the order of ±50 percent are tolerable. This latitude is also large enough to be advantageous in buffer design.* Interval lengthening of more than 200 percent, and shortening down to 0 percent, are clearly not acceptable.

---

† Extensive current work on TASI-D also gives insight about this coding range.

* Additionally, the possibility exists for modifying the durations of the active signal packets (by spectrum-preserving techniques such as the phase vocoder).

Technical material in this note was presented orally to the 93rd meeting of the Acoustical Society of America (J. Acoust. Soc. Am. *61*, S69, June 1977).