THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 49 July–August 1970 Number 6

Copyright © 1970, American Telephone and Telegraph Company

Picturephone® Silicon Target Signal Analysis

By W. E. BEADLE and A. J. SCHORR

(Manuscript received January 5, 1970)

Signal characteristics of the silicon diode array used as the image sensing element of the Picturephone® camera tube have been studied using a numerical computation. Analytical representation of the target is based on the numerical calculation of the depletion region geometry for an array diode unit-cell undergoing discharge. This analysis includes the effects of Si–SiO_2 fixed interface charge density, sea resistance, substrate resistivity, P^+ island geometry, and SiO_2–Si surface inversion phenomenon.

Computed depletion region geometries are used to calculate surface and bulk contributions to array dark current. It is shown that signal current limitations due to surface inversion can be avoided using lower values of sea resistance coupled with higher SiO_2–Si interface fixed charge density. For the resistive sea target structure, inversion effects are less pronounced for targets fabricated with P^+ islands larger than 10 μm diameter and sea sheet resistances less than 10^{14} ohms per square. This signal limiting effect can also be eliminated by using a conductive overlay structure.

Results of analyses of lag characteristics and electron beam limitations are also presented.

I. INTRODUCTION

The silicon target, used as the image sensing element in the *Picturephone*® camera tube,[1-4] is composed of a matrix of diffused diodes.

Signal characteristics of the target are the result of a complex interaction of the material and geometric parameters of the individual array diodes as well as the operating conditions of the target. We have analtyically investigated target signal characteristics by means of numerical computations, solving explicitly for depletion region geometry and capacitance and have determined signal characteristics as a function of Si–SiO_2 interface fixed charge density, sea resistance*, substrate resistivity, P^+ island geometry, SiO_2–Si surface inversion phenomenon, and electron beam acceptance characteristics.

Section II describes the diode array format and the basic relationship of signal current to array capacitance. The model of the reversed biased unit cell and the mathematic analysis are described in Sections II and III. Signal capabilities of a resistive sea diode array structure are discussed in Section IV and extended to include the effects of electron beam acceptance in Section V. An alternate geometric structure using conductive overlays is analyzed in Section VI. The relationship between the array dark current characteristic and target parameters is presented in Section VII. In the concluding section we present experimental verification of the model.

II. GENERAL TARGET OPERATION

The *Picturephone*® camera tube target is an array of 8×10^5 planar P^+N diodes fabricated on a thin N-type silicon membrane, (Fig. 1). The optical sensing mechanism of the target is the discharge

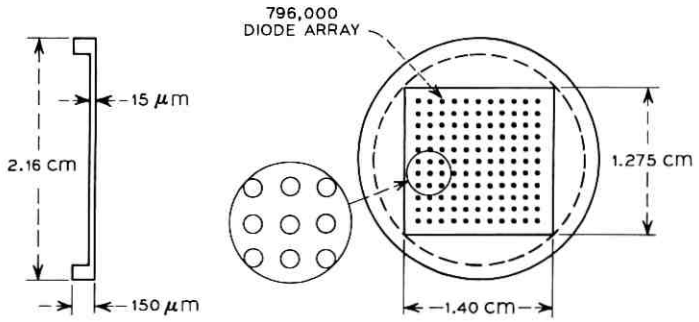


Fig. 1—Silicon camera tube target.

* We use the term "sea" to describe the resistive layer covering the array surface containing the P^+ diffused regions which we will refer to as P^+ islands.

of this array of diodes by light incident on the N^+ surface (side opposite the array). To review the details of the operation it is convenient to consider an individual array element, or unit cell, represented by a 15μm square section of the array containing a centrally located diode.

The back surface of the target is held at a potential V_T. The target face potential is periodically reduced to ground potential by the scanning electron beam (Fig. 2). For ideal beam acceptance, the target unit cell (consisting of the junction plus the oxide capacitance) is reverse biased by an amount V_T immediately after scanning. At this time, the junction depletion region and any existing silicon surface depletion region will be at their geometric maximum. During the time interval (corresponding to a frame time, τ_F), that the electron beam is scanning the remainder of the target, the diode cell is discharged by the photon generated holes which diffuse to the depletion region of the device.

The relationship between light-generated minority carriers available for discharge and the incident photon flux and the subsequent diffusion of these carriers to the diode depletion region has been discussed by others.[4] The individual diode steady state signal current can be related to the diode discharge by

$$I = \frac{1}{\tau_F} \int_{V_F}^{V_T} C(V)\, dV \tag{1}$$

where

$C(V)$ = diode cell capacitance,
V_T = the voltage across the diode cell at the start of a frame,
V_F = the voltage across the diode cell at the completion of a frame.

The upper integration limit will be less than V_T if the tube has capacitive lag. Maximum signal occurs when $V_F = 0$, which corresponds

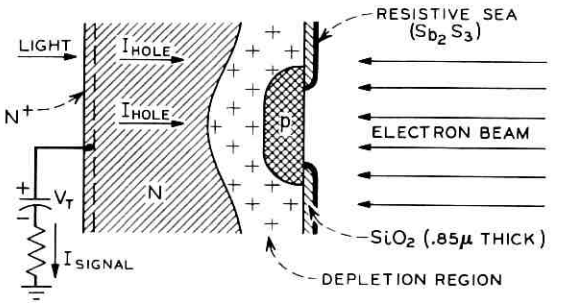


Fig. 2—Target in scanning mode.

to ideal electron beam acceptance. Additional charge storage can be obtained by driving the diode cell into forward bias. Since increased lateral hole spreading causes loss of resolution under this condition, we restrict our analysis to reverse bias conditions.

Figure 3 schematically shows the equivalent components associated with the unit cell structure. These consist of:

(*i*) the diode junction capacitance, C_J ,
(*ii*) the oxide capacitance, C_{ox} , and
(*iii*) the silicon surface capacitance at the silicon-oxide interface, C_{DEP} .

2.1 *Mathematical Model of a Reversed Biased Diode Unit Cell*

For purposes of analysis, an individual diode can be considered as an axisymmetric structure composed of several mathematically distinct regions. Figure 4 is a schematic representation of the analytical model. The regions of interest include:

(*i*) Oxide mask—The oxide region is considered to be a homogeneous, charge free region with dielectric constant ϵ_1 . The potential function in this region is determined by solution of Laplace's equation.

(*ii*) P^+ diffused island—This region is considered to be at a uniform potential.

(*iii*) N-type bulk region—That portion of the N-type bulk region,

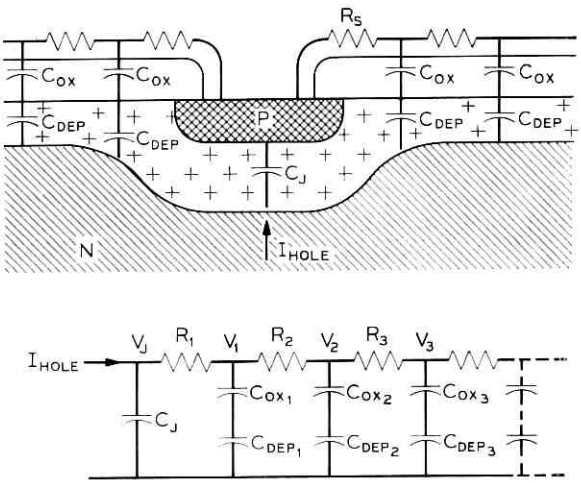


Fig. 3—Target discharge model.

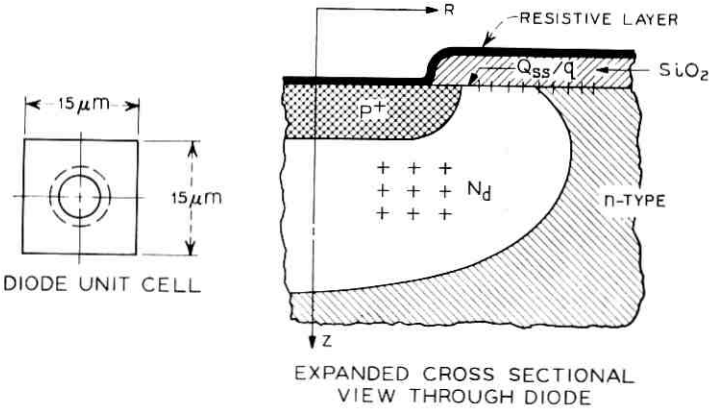


Fig. 4—Diode unit cell geometry.

remote from the depletion region, is considered to be at uniform potential.

(*iv*) Depletion region—The potential distribution in the depletion region is described by Poisson's equation. The charge distribution in the depletion region remote from the oxide-silicon interface is considered constant and set by the doping level of the N-type bulk material. An additional charge density is assumed in the depletion region at the Si–SiO_2 interface to describe the oxide fixed charge. The depletion region has a dielectric constant ϵ_2 .

2.2 *Boundary, Interface Conditions*

The target analysis is complicated because the target consists of axisymmetric diffused regions arranged on a square format and spaced such that the diode depletion regions can connect for some combinations of target parameters. We analyze an axisymmetric structure and correct for the effects of the cell corner regions when necessary.

Except for the boundary separating the depletion and N-type bulk regions (discussed in Section 2.3), the boundary conditions for this problem are straight-forward:

(*i*) There is a radial potential distribution on target surface.

(*ii*) The radial potential gradients along the axis of symmetry are zero.

(*iii*) The condition of conservation of charge flux is applied at the oxide-depletion region interface.

(*iv*) Radial potential gradients at the diode periphery are zero.

2.3 *Method of Solution*

Regions of dissimilar material constants and geometrically complicated boundaries make solution by "closed form" techniques difficult. For this reason and because of its inherent geometric flexibility, a numerical or finite difference solution was adopted.

The entire diode structure is considered to be composed of differential elements spaced on a square mesh (Fig. 5). The potential of each element is described by an appropriate difference equation which takes into account the material constants of the element and its interaction with adjacent elements. For example, the potential of an element in the depletion region at a location (I, J) is given by a difference equation of the type:

$$\epsilon A_z\left[\frac{V(I, J+1) - V(I, J)}{\Delta Z} + \frac{V(I, J-1) - V(I, J)}{\Delta Z}\right] + \epsilon A_-\left[\frac{V(I-1, J) - V(I, J)}{\Delta R}\right] + \epsilon A_+\left[\frac{V(I+1, J) - V(I, J)}{\Delta R}\right] = -q\,\Delta Z A_z N(I, J) \tag{2}$$

where

$V(I, J)$ = potential of point (I, J),
A_z, A_+, A_- = areas of element faces,
ϵ = dielectric constant,
ΔR = radial element spacing,
ΔZ = axial element spacing,
$qN(I, J)$ = charge density.

Similar difference equations can be written to describe the behavior of other regions.

Thus the potential of any element can be described by a linear algebraic equation. If the geometry of interest is described by decomposing the structure into an M by N array of such elements, then the problem is reduced to the solution of an $M \times N$ set of linear equations. For a typical geometry considered in this study, the structure was represented by a 70 by 85 array of elements. This set of equations was solved by accelerated Gauss–Seidel iteration.[5]

An important aspect of the solution of this problem is the specification of the depletion region boundary. The P^+-depletion layer boundary is defined by the doping profile. We assume a P^+ step junction. The depletion region-N-type bulk boundary is a function of impurity distribution, oxide-silicon interface charge density, and resistive sea surface

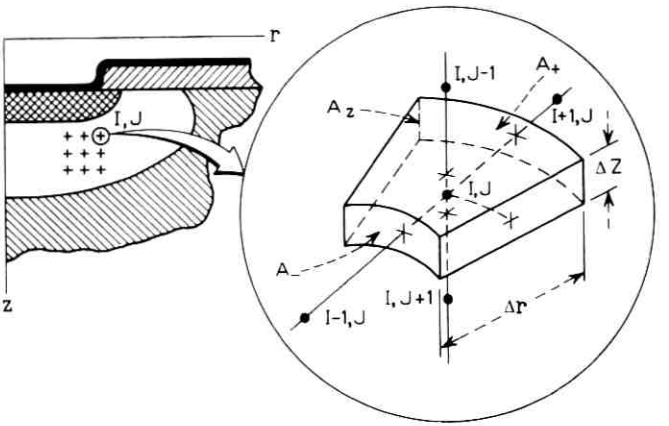


Fig. 5—Difference elements.

potential and must therefore be computed for each new combination of diode parameters. An iterative procedure was used to calculate potential and to define the limits of the depletion region. Beginning at the edge of the P^+ diffused island an iterative sweep through the potential field was made. At those points where the computed potential was less than the applied potential plus the built-in potential of the junction, V_B, Poisson's equation in difference form was applied. When a potential equal to or greater than $V_T + V_B$ was computed, the potential at that point was set equal to $V_T + V_B$. This procedure was repeated until a satisfactory solution was obtained.

2.4 *Capacitance Calculation*

Depletion region capacitance was calculated by treating the depletion boundaries as the surfaces of a two-plate capacitor. The potential field between the plates of such a capacitor was calculated by solving Laplace's equation, taking into account the different dielectric constants for the oxide and silicon regions. Charge density on the surface is given by

$$Q = \epsilon E_n , \tag{3}$$

where E_n is the normal electric field component calculated using difference approximations for the first derivative of the potential, and ϵ is the appropriate dielectric constant.

Total diode capacitance is then equal to the integral of the local charge density divided by the local voltage over the total cell surface area.

III. MATHEMATICAL MODEL OF IDEAL CELL DISCHARGE

The model considered in the analysis of signal capability is shown schematically in Fig. 3. For all subsequent calculations the target is biased at a 12-volt potential relative to the cathode (except as noted). Both cathode potential drop and beam limitations (including the effect of the resistive sea impedance in series with the P^+ island) are neglected in this portion of the analysis. These assumptions imply sufficient beam current to completely restore the target surface to cathode potential after each beam scan, re-establishing the initial 12-volt reverse bias across the junction. (With our convention reverse bias and surface potential sum to 12 volts.) During the time interval during which a diode is disconnected from the electron beam, τ_F, an array diode is discharged by the light generated hole current collected by its depletion region. As the unit cell is discharged, the P^+ island potential increases approaching 12 volts for high illumination levels. The voltage profile on the target surface depends on substrate resistivity, SiO_2–Si interface fixed charge density, target geometry, illumination level, resistive sea sheet resistance and electron beam acceptance characteristic. For the case of a very high resistance sea—that is, no lateral charge flow onto the sea—the potential of the resistive sea surface remote from the diode windows will remain at ground potential and the light generated current will act to discharge the P^+N junction only. Conversely, for the case of a low sheet resistance sea the entire resistive sea surface potential will rise nearly uniformly.

3.1 *Qualitative Description of Model*

The time dependent solution of the diode cell discharge transient is a complicated non-linear problem. However, a reasonable model for this problem should include the voltage dependence of both the junction and SiO_2–Si surface capacitances. The silicon surface capacitance is important since it acts in series with the oxide capacitance to affect the time constant of the resistive sea. For example, in the depleted mode the equivalent oxide surface capacitance can be half the oxide capacitance.

The lateral voltage drop on the diode unit cell surface makes possible the formation of an inversion layer at the SiO_2–Si interface. Since this effect can (for some combination of target parameters) establish the upper limit on sea resistance, the model should include inversion phenomenon. Strong inversion was assumed when the magnitude of the surface potential relative to the bulk semiconductor potential exceeds the junction reverse bias, V_R, by twice the magnitude of the bulk Fermi potential, ϕ_F.[6]

A quasi steady-state solution was used in which the discharge time interval was divided into sub-intervals over which the surface capacitance was assumed time invariant. The dependence of the junction capacitance on voltage was explicitly described.

An outline of the unit diode cell discharge analysis is shown in Fig. 6. For computational purposes, τ_F was divided into several equal segments. The computational sequence used during each time segment was as follows:

(*i*) At the beginning of each segment: First, the unit cell depletion geometry and corresponding surface capacitance were calculated subject to boundary conditions which include the resistive sea surface potential distribution from the previous time step. Second, the surface potential profile was recalculated to adjust for altered surface capacitance by imposing temporal continuity of surface charge.

(*ii*) Having obtained the surface potential and capacitance profiles, a transient solution to the array discharge process for the time segment

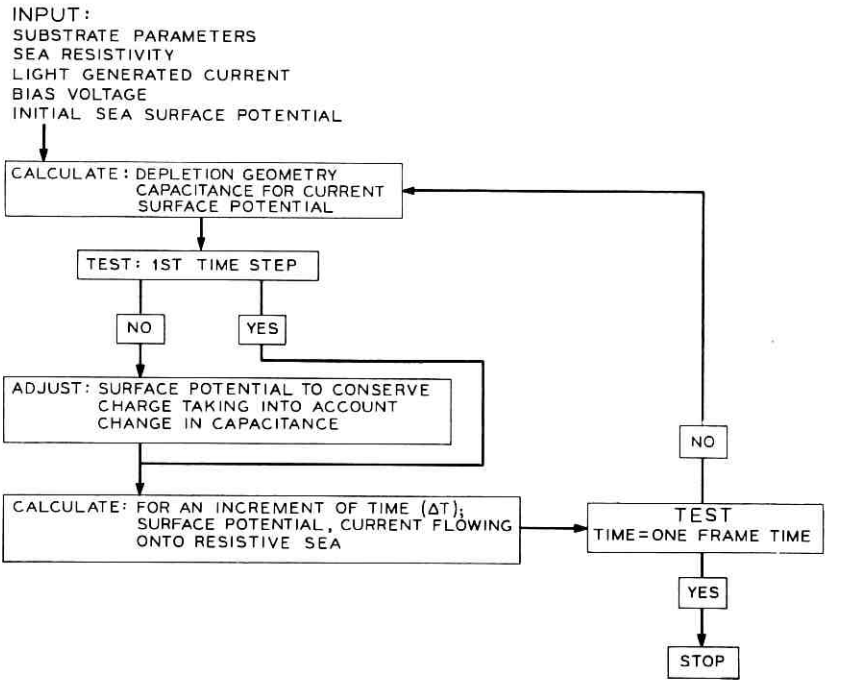


Fig. 6—Unit cell discharge analyses.

was performed with the surface capacitance profile held constant. The voltage dependence of the junction capacitance was accounted for. The time dependent surface potential profile and sea current were then calculated. The final potential profile was used in the first part of step i of the next time interval.

3.2 *Some Particular Details of Target Operation*

Target operation can be most easily characterized by considering operation either below or above total depletion of the silicon surface at the Si–SiO_2 interface. When operating below total surface depletion, variations in sea sheet resistance can alter only the width of the depletion layer at the surface. Above the condition of total surface depletion, the resistive sea sheet resistance can control the formation of a silicon surface inversion layer. It should be noted that because of geometrical effects, total surface depletion may occur in the diode array at voltages significantly below the depletion voltage which would be obtained on a MOS capacitor with the same oxide thickness and interface properties.

3.2.1 *Target Operation Below Total Surface Depletion*

Consider a 10 Ωcm substrate target with an interface fixed charge density (Q_{ss}/q) of $6 \times 10^{11}/cm^2$ and $V_T = 12$ volts. Under these conditions the target will be below total surface depletion. Figures 7 and 8 illustrate the behavior of the depletion region and resistive sea surface potential at various intervals during discharge for two values of sea resistance.

Comparison of the depletion layer geometries at $t = \tau_F$ (corresponding to full discharge) shows that the signal obtained from the junction-only capacitance is not appreciably altered by differences in sea sheet resistance. The surface potential profiles at $t = \tau_F$ clearly indicate the difference in surface charge flow for the two values of sheet resistance. An effective frame time of 1/60 second is assumed.*

Figures 9 and 10 show the corresponding sea current and junction potential variations over one frame time for the same two sea resistance values. In both cases the junction voltage reaches the 12 volt bias potential prior to the completion of the frame time. Therefore, the light discharge current specified in the calculations 6.4 pA was in excess of true saturation, and lateral hole diffusion in the bulk occurs during the latter portion of the discharge cycle. For example, for the 10^{14} Ω/□ resistive sea (Figs. 8 and 10), 3.1 pA of the 6.4 pA light signal corre-

* The system frame time is 1/30 second; however, due to the interlace scan system and large beam diameter, the diodes are charged at a 1/60 second rate.

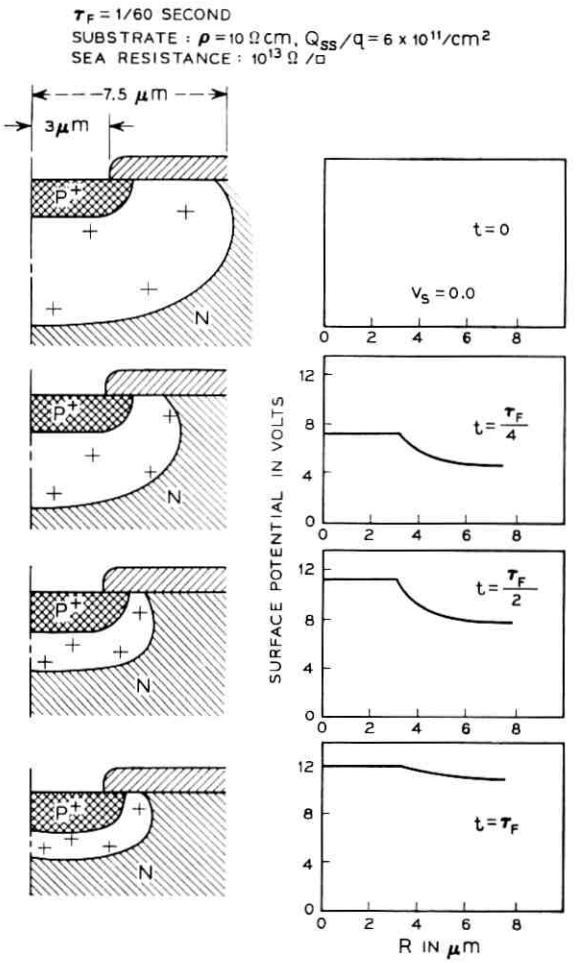


Fig. 7—Calculated discharge conditions for low sea-sheet resistance.

sponds to the discharge of the junction-only capacitance. The time-averaged sea current is 1 pA. The lateral loss due to diffusion is 2.3 pA. By decreasing the sea resistance to 10^{13} Ω/□ (Figs. 7 and 9) the average sea current is increased to about 2 pA-giving a total average cell discharge current of 5.1 pA.

3.2.2 *Target Operation Above Total Surface Depletion*

For target operation at bias voltages where the surface can be totally

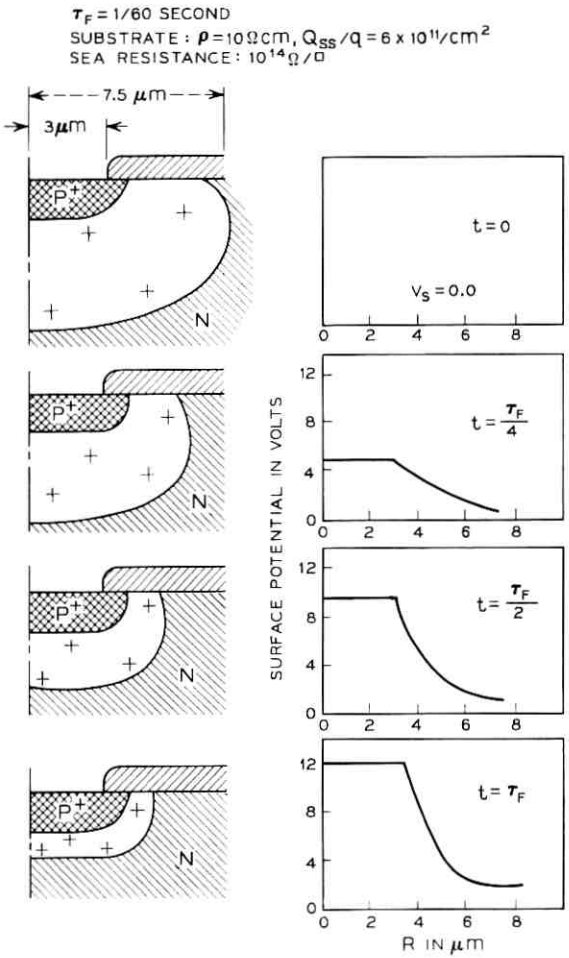


Fig. 8—Calculated discharge conditions for high sea-sheet resistance.

depleted, the sea sheet resistance is more critical. Sea sheet resistance not only controls the coupling of the oxide and junction capacities but can allow the formation of an inversion layer prior to total cell discharge. This effect can result in an additional limitation on maximum useful signal current. For example, consider the final depeltion region configuration (Fig. 11) for a substrate resistivity (ρ) equal to 10 Ωcm, Si–SiO_2 interface fixed charge density (Q_{ss}/q) equal to $3 \times 10^{11}/cm^2$, and a sea sheet resistance (R_s) equal to 10^{14} Ω/□. Because of the lower

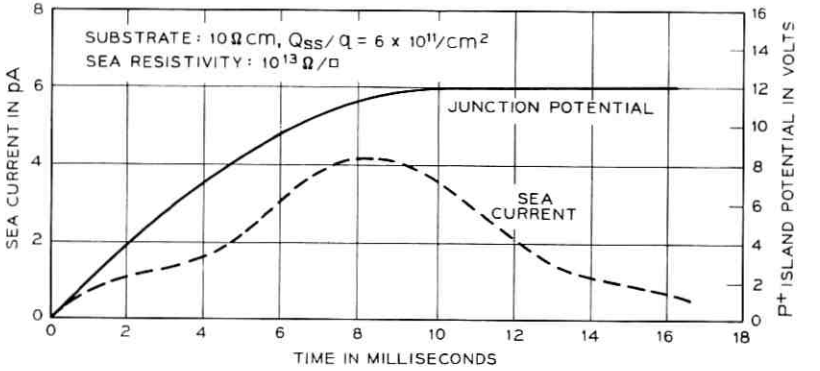


Fig. 9—Calculated discharge characteristics for low sea-sheet resistance.

Q_{ss}/q, the depletion region is markedly different from that of the unit cell previously considered (Figs. 7 and 8). Note that the depletion layer exists even at the end of the frame time. Furthermore, the final discharged condition shows surface inversion under part of the oxide surface. If the inversion layer acts to interconnect adjacent P^+ islands, then loss of resolution will occur.[4] Resolution loss does not occur simply by the existence of an inversion layer but rather when sufficient current flow occurs between the P^+ island and the inversion layer to connect adjacent P^+ islands. To investigate the formation of the inversion layer, it is necessary to examine the Si–SiO_2 interface potential distribution. Figure 12 shows the potential profile (for $Q_{ss}/q = 2 \times 10^{11}/cm^2$ and $R_s = 1.5 \times 10^{14}\ \Omega/\square$) at three time intervals in the discharge cycle. It can be seen that all the light-generated minority carriers which reach

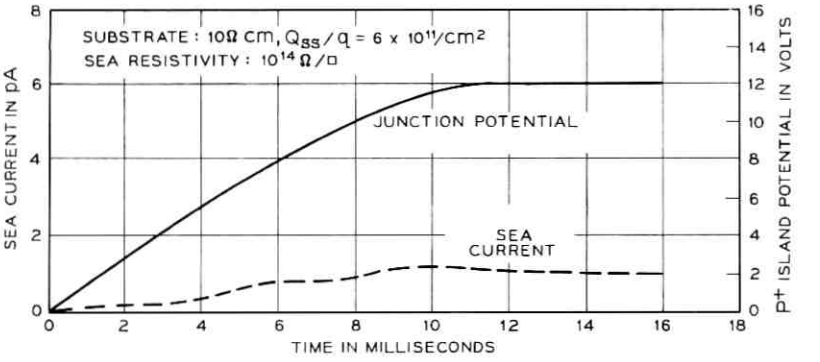


Fig. 10—Calculated discharge characteristics for high sea-sheet resistance.

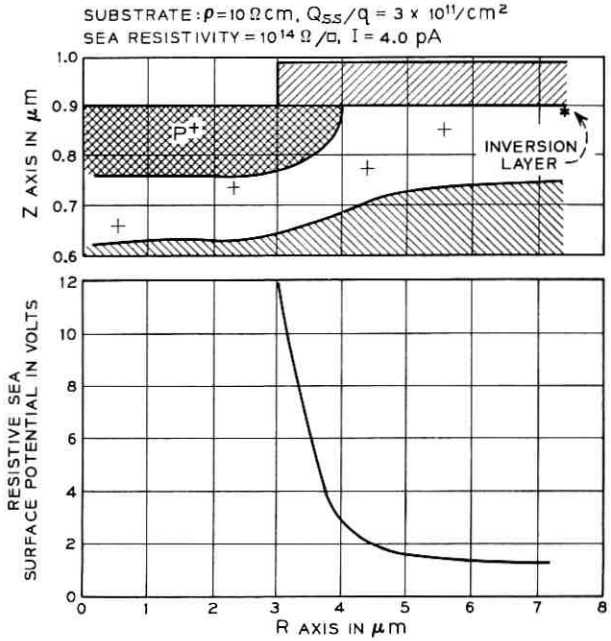


Fig. 11—Calculated discharged conditions for above total surface depletion operation.

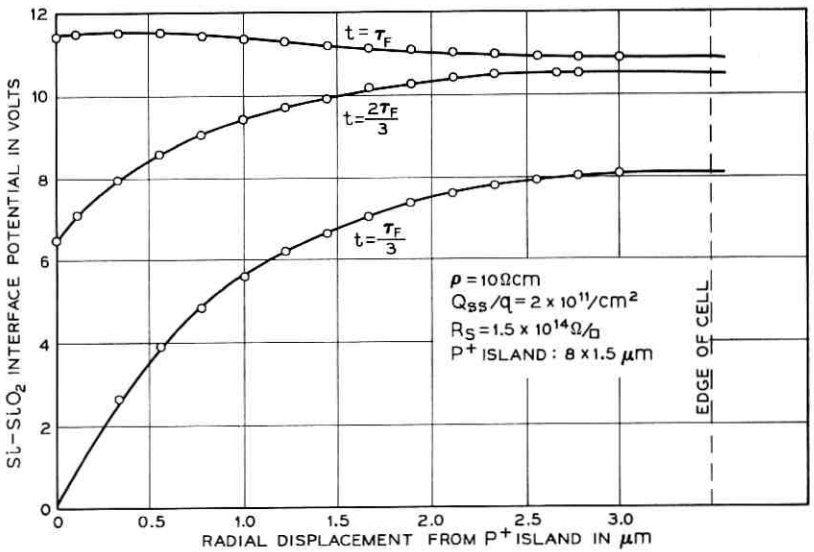


Fig. 12—Si–SiO_2 interface potential at the end of three equal time intervals during discharge.

the depletion region are collected by the P^+ island until that time, late in the discharge cycle, when the silicon surface potential remote from the P^+ island drops to below the P^+ island potential. This condition only occurs for high sea-sheet resistance which causes a large radial voltage gradient on the sea surface.

Figure 13 illustrates the Si–SiO_2 interface potential distribution relative to the P^+ island potential at the end of the discharge cycle for several values of light-generated current. This figure shows that for low current levels a potential barrier for holes exists between the P^+ island and inversion region. The magnitude of this barrier decreases as the junction current is increased. We have assumed that loss of resolution occurs when this potential barrier drops below $2kT$. For this particular example, the diodes effectively interconnect at a light-generated current of 2.1 pA, and the strongly inverted layer need only extend to within 2.6 μm of the P^+ island.

IV. TARGET SIGNAL CAPABILITIES

In this section calculated results are presented which illustrate the general effect of diode parameters and operating bias on target signal capabilities.

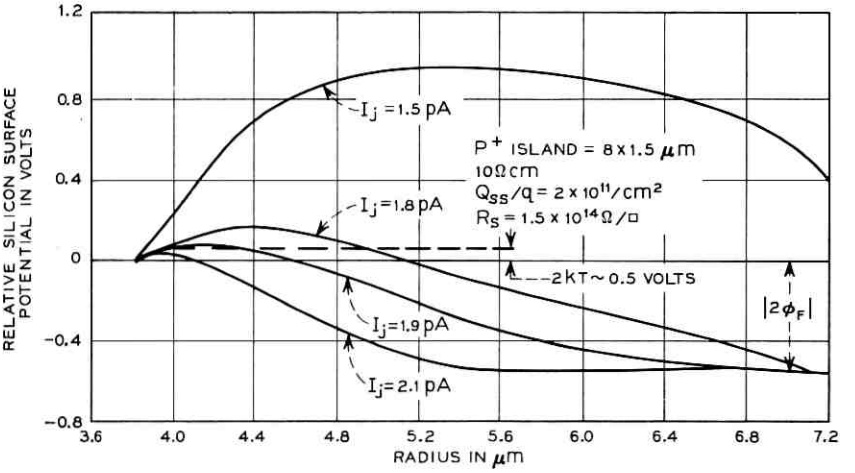


Fig. 13—Si–SiO_2 interface potential relative to P^+ island potential at completion of the discharge cycle for several levels of light-generated discharge current.

4.1 *Effect of Signal-Limiting Mechanisms*

Signal current characteristics of target arrays fabricated on 10 Ωcm substrates with Q_{ss}/q equal to 1, 2, and 3 × 10^{11}/cm^2 are shown in Fig. 14. This figure graphically illustrates the interdependence of the SiO_2–Si

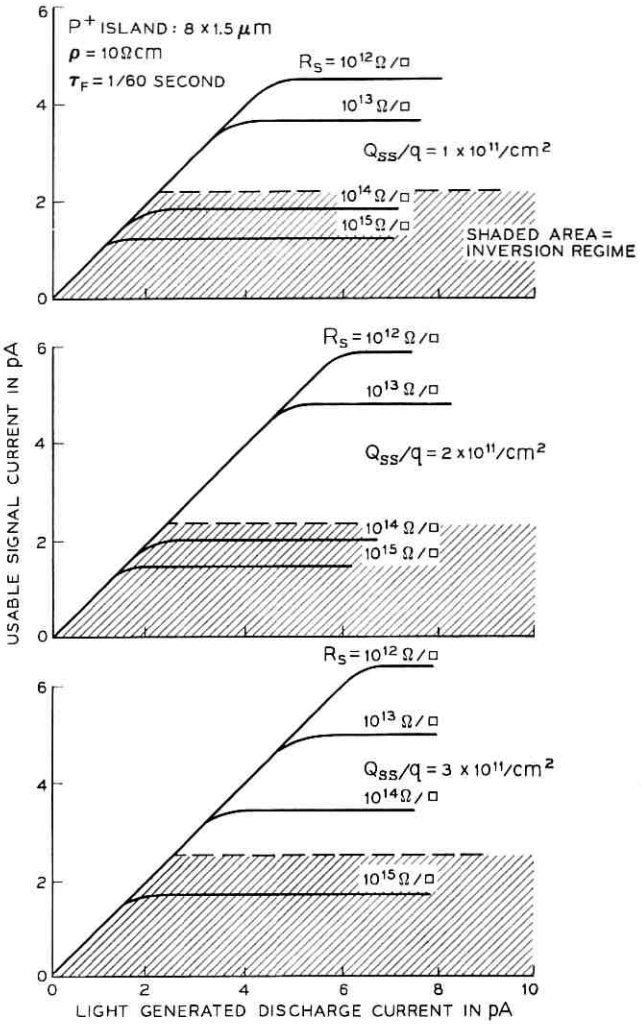


Fig. 14—Effect of Q_{ss}/q on single cell signal current characteristics.

interface charge density and sea sheet resistance on signal current characteristics.

Oxide-silicon interface fixed charge density affects the array signal capacity for all values of sea resistance. For example, consider a target fabricated with a low sea sheet resistance of 10^{12} Ω/□. If this array has an interface fixed charge density of $3 \times 10^{11}/cm^2$, it will provide a 6.4 pA signal current per diode. A similar target having a fixed density of $1 \times 10^{11}/cm^2$ is capable of only 4.5 pA per diode. Using an estimated minimum sea resistance consistent with resolution,[4] of $R_S = 10^{13}$ Ω/□, the maximum usable signal current per diode is 5 pA for $Q_{ss}/q = 3 \times 10^{11}/cm^2$; and 3.6 pA for $Q_{ss}/q = 1 \times 10^{11}/cm^2$.

For high sea-sheet resistances, above 10^{13} Ω/□, the effect of fixed interface charge density on signal is even more pronounced since the inversion mechanism becomes operative. Consider the case of $R_S = 10^{14}$ Ω/□. From Fig. 14 we see that for $Q_{ss}/q = 3 \times 10^{11}/cm^2$, the available signal is 3.4 pA/diode. If, however, the interface fixed charge is reduced to $1 \times 10^{11}/cm^2$, the inversion mechanism becomes operative and the maximum usable signal current is reduced to 1.8 pA.

4.2 *Specific Examples of Geometric Effects*

Certain combinations of P^+ island geometry, interface fixed charge, and sea resistance can result in a limitation in the dynamic signal capability of the target due to diode interconnection caused by silicon

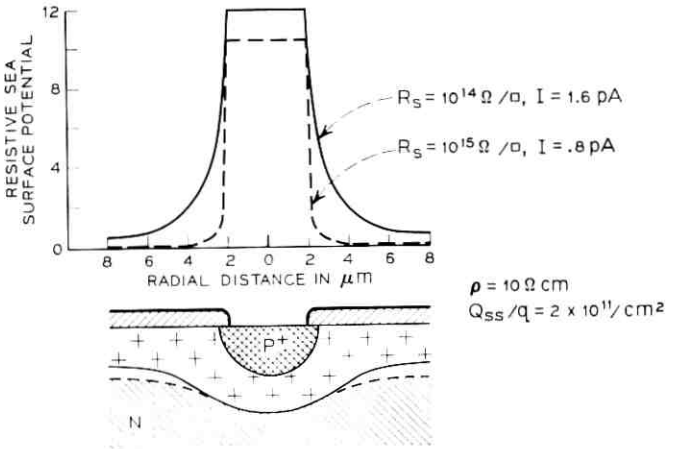


Fig. 15—Discharged conditions for 4 μm window diode.

surface inversion. Consider the three geometries shown in Figs. 15, 16 and 17 which illustrate the effect of two values of sea resistance on maximum signal current. The resistive sea potential immediately after the electron beam scan is assumed to be zero volts. The final surface potential and depletion region geometry are shown for the maximum discharged condition.

In this example, the maximum P^+ island surface potential is limited to 10 volts for the 4 μm window geometry and $R_s = 10^{15}\ \Omega/\square$; this corresponds to a signal current of 0.8 pA. Signal current for this same geometry with $R_s = 10^{14}\ \Omega/\square$ is 1.6 pA. The difference in signal currents results from surface inversion at the higher value of sea resistance. This current-limiting effect is less pronounced for the larger diode diameters as illustrated in Figs. 16 and 17. The effect of P^+ island diameter on signal current characteristics for $Q_{ss}/q = 1$, 2 and 3 $\times$ $10^{11}/cm^2$ is shown in Figs. 18 through 20. These figures graphically illustrate the interdependence of the SiO_2–Si interface charge density and sea sheet resistance on signal current characteristics.

The junction depth for the diode configurations of Figs. 18, 19, and 20 is 2.5 μm. Junction depth has a pronounced effect on signal current capabilities for values of sea sheet resistance greater than $10^{14}\ \Omega/\square$ as shown in Fig. 21. In this example, a 1.5 μm deep junction will experience

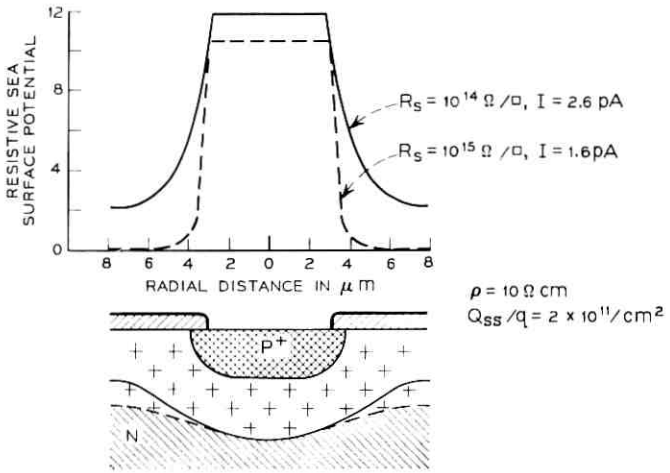


Fig. 16—Discharged conditions for 6 μm window diode.

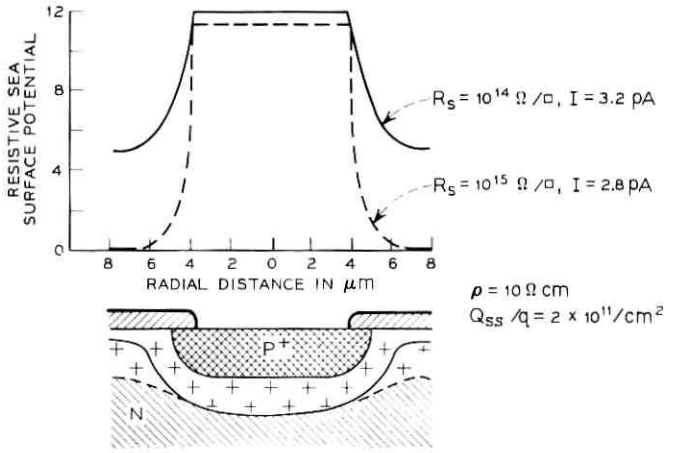


Fig. 17—Discharged conditions for 8 μm window diode.

a 43 percent signal limitation due to surface inversion compared with a 2.5 μm deep junction of the same diameter.

The effect of target bias on usable signal current has also been investigated. Increased target bias reduces usable signal for targets with high sea resistance and low interface fixed-charge density. Signal capabilities of targets operated at 12- and 16-volt bias can be compared

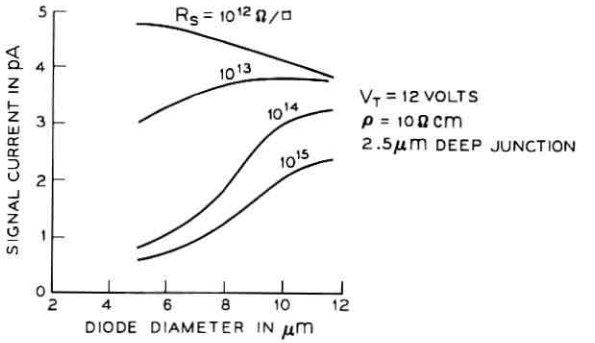


Fig. 18—Single diode signal capability for $Q_{ss}/q = 1 \times 10^{11}$ cm^{-2}.

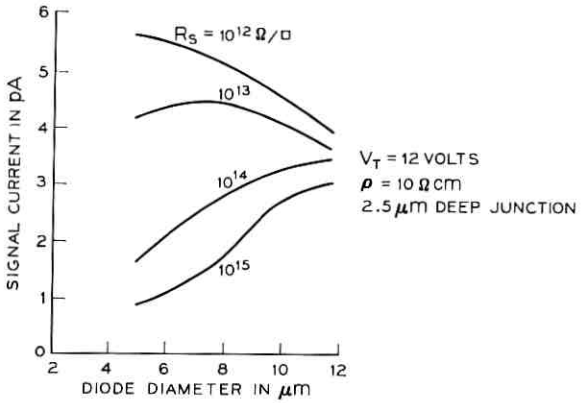


Fig. 19—Single diode signal capability for $Q_{ss}/q = 2 \times 10^{11}$ cm^{-2}.

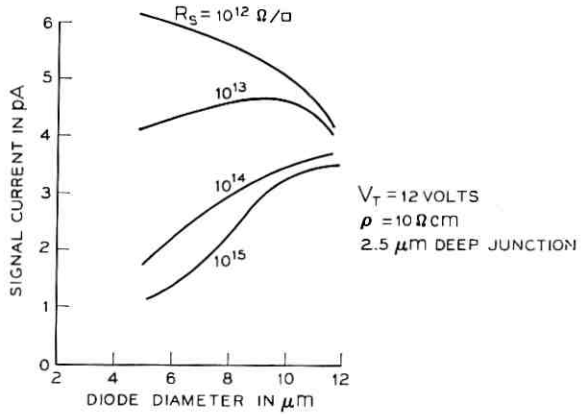


Fig. 20—Single diode signal capability for $Q_{ss}/q = 3 \times 10^{11}$ cm^{-2}.

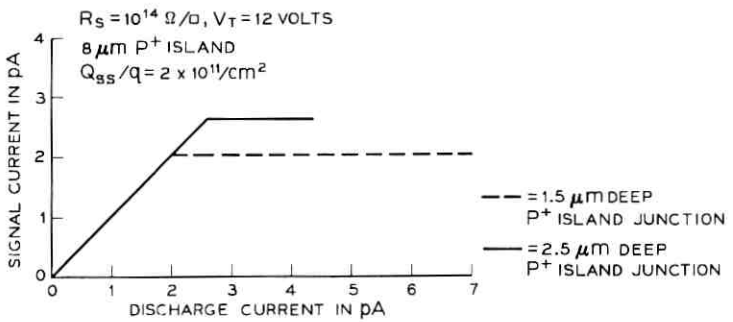


Fig. 21—Effect of P^+ island depth on single diode signal characteristic.

in Fig. 22. For R_s greater than 10^{14} $\Omega/\square$, reduction of usable signal for the higher bias case is due to inversion effects caused by larger sea surface-potential gradients.

V. EFFECT OF ELECTRON BEAM ACCEPTANCE

In order to estimate the effects of electron beam acceptance on signal characteristics, the following experimental-analytic approach was undertaken:

(*i*) A mathematical model of the charge-discharge mechanism of the target in the scanning mode was established.

(*ii*) For several target bias conditions, maximum signal current (maximum light level consistent with resolution) and the residual signal characteristic were measured. The residual signal lag was measured by chopping the target illumination as described in Section 5.2.

(*iii*) Fabrication parameters for this target (Q_{ss}/q, P^+ island geometry, sea resistance and substrate resistivity) were measured.

(*iv*) The junction-only capacitance-voltage characteristic was calculated from measured target parameters. In this instance, sea resistance was sufficiently high so that the junction-only capacitance was the only signal-generation mechanism.

(*v*) Using the capacitance voltage characteristic of the junction, the saturation signal and lag characteristics of the tube, and the mathematical representation of the target charge-discharge mechanism; a beam-landing function was generated by a trial and error procedure which satisfies these measured characteristics. In this calculation the contribution to lag of interface trapping states was neglected.*

(*vi*) Using the beam landing characteristic thus generated and the mathematical model, lag and beam limited signal were calculated for various diode P^+ island geometries.

5.1 *Signal Limitations*

The resulting signal current versus P^+ island diameter for a junction depth of 2.5 μm is given in Fig. 23. Comparison of these results with those for perfect beam landing shows that the beam limitation causes a significant reduction in target signal capability. For example, a diode having 8 μm P^+ island diameter and a 10^{14} $\Omega/\square$ sea sheet resistance has a maximum signal current capability of 2.8 pA. With beam limitation, signal current is reduced to 1.0 pA. It should also be noted that the beam

* G. F. Amelio has shown that under some conditions trapping states at the oxide-silicon interface can contribute to target lag.

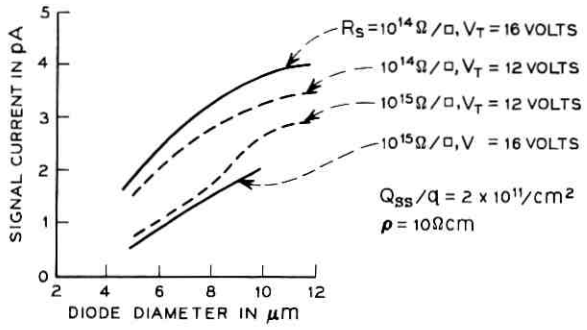


Fig. 22—Effect of bias on single diode signal capability.

current limitation will accentuate surface inversion for the higher values of sea resistance. As one might expect, larger diameter oxide windows improve beam collection efficiency and thus minimize signal limiting.

The effect of beam acceptance on the surface inversion phenomenon is illustrated in Fig. 24. This figure shows the locus of operation voltages as a function of light level for an 8-μm P^+ island diameter and a sea resistance sufficiently high to decouple the oxide capacitance. As light-generated discharge current level increases the maximum target surface potential approaches the target bias. For a sea resistivity of $10^{14}\ \Omega/\square$ the maximum allowable P^+ island potential can rise to the 16 volt bias potential without diode interconnection due to silicon surface inversion. If beam limitation is taken into consideration, the calculated potential swing is 12 to 16 volts resulting in a maximum signal level of 1.2 pA per diode. If the sea sheet resistance is increased to $10^{15}\ \Omega/\square$, surface inver-

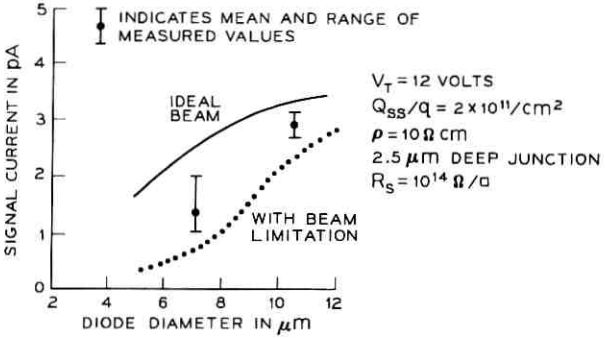


Fig. 23—Effect of electron beam acceptance on single diode signal capability.

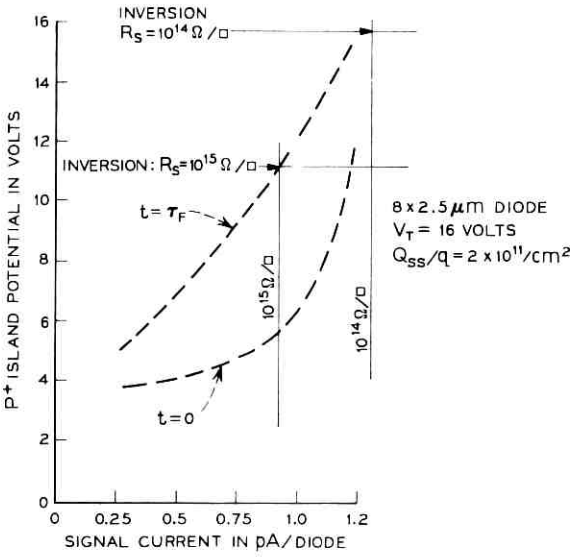


Fig. 24—Calculated P^+ island potential extremes as a function of discharge current.

sion limits the P^+ island potential to 11.0 volts. The maximum signal current for this case is therefore only 0.9 pA.

5.2 *Capacitive Lag—Transient Illumination*

Thus far only time invarient illumination of the target has been considered. Non-ideal beam landing results in capacitive lag for transient illumination conditions. The definition of lag used here is illustrated in Fig. 25. Light is left on for a sufficient time interval for the target to assume steady state operation; that is, $\Delta Q_{\text{charge}} = \Delta Q_{\text{discharge}}$. The light is then turned off and the instantaneous signal current at subsequent scanning intervals determined. Lag is then defined as the ratio of the signal current at 1/15 second after the light is turned off to the steady-state illuminated signal.

Lag is a function of the resistive sea impedance, the diode capacitance, and the electron beam acceptance. The electron beam acceptance is a function of the resistive sea surface potential. Since both the instantaneous diode capacitance and the sea surface potential are determined by the level of light-generated cell discharge current, lag is also a function of signal level as illustrated in Fig. 26. For the case under consideration, two effects are evident:

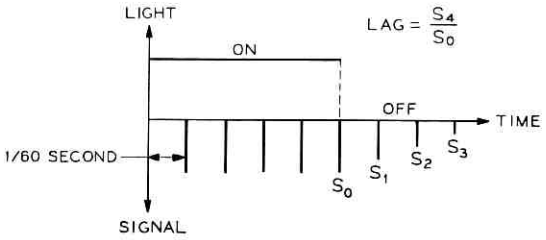


Fig. 25—Definition of lag.

(*i*) At high signal levels lag is enhanced by large target cell capacity.
(*ii*) At low light levels, where cell capacity is reduced, poor beam acceptance increases lag.

The result for this example is a minimum lag condition at less than maximum signal current. Since this effect is a strong function of the electron beam acceptance characteristic, the width of the lag minimum can vary considerably from tube to tube.

VI. ALTERNATE TARGET STRUCTURE

The occurrence of inversion effects can be minimized, but not eliminated, by fabricating targets with greater than 10 μm diameter P^+ islands.

Inversion effects may also be controlled by means of a partial conductive overlay structure.

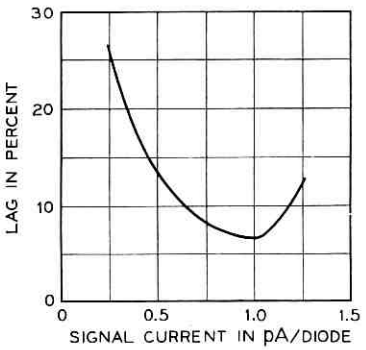


Fig. 26—Calculated lag characteristic.

6.1 *Description of Overlay Structure*

This structure is essentially that of the standard resistive sea silicon target array except conductive buttons have been formed in the oxide windows and extend over the adjacent oxide.

6.2 *Analysis*

The analysis of this structure is performed assuming that the sea region between buttons is at ground potential (a worst case condition) with a uniform potential V_T assigned to the P^+ island and conductive overlay.

Signal capabilities of this structure were also determined by calculation of the capacitance voltage characteristic of the button. Again the sea region was assumed to be at zero potential and thus electrically isolated from the junction button.

6.3 *Conductive Overlay Characteristics*

Depletion region geometries for a range of conductive overlays diameters are shown in Fig. 27. For the assumed 12 volt bias three distinct silicon surface conditions are represented by these examples:

(*i*) Separation of P^+ island from the peripheral surface inversion region by a neutral N region at the interface. This corresponds to a 11.2 μm diameter button.

(*ii*) Continuous depletion layer through the diode unit cell. For a 9.5 μm diameter button the inversion layer formed at the periphery is isolated from the P^+ island by an appreciable potential barrier.

(*iii*) Continuous depletion layer through the diode unit cell and electrical interconnection of the P^+ island and the surface inversion regions. This condition exists for the 8.0 μm diameter button.

For the 9.5 μm button, which represents the minimum diameter necessary to prevent diode interconnection for the example under consideration, the maximum signal current is 1.9 pA/diode. This is slightly less than the current obtainable with a resistive sea structure with 10^{14} $\Omega/\square$ sea sheet resistance.

The major advantage of a properly fabricated conductive overlay structure is the elimination of inversion effects (which include signal limiting and video "blooming" phenomenon) inherent in the resistive sea type targets operated with total surface depletion. While elimination of inversion effects can be accomplished by controlling sea resistance (within the range of 1 to 5 $\times$ 10^{13} $\Omega/\square$ for $Q_{ss}/q \approx 10^{11}/\text{cm}^2$) in con-

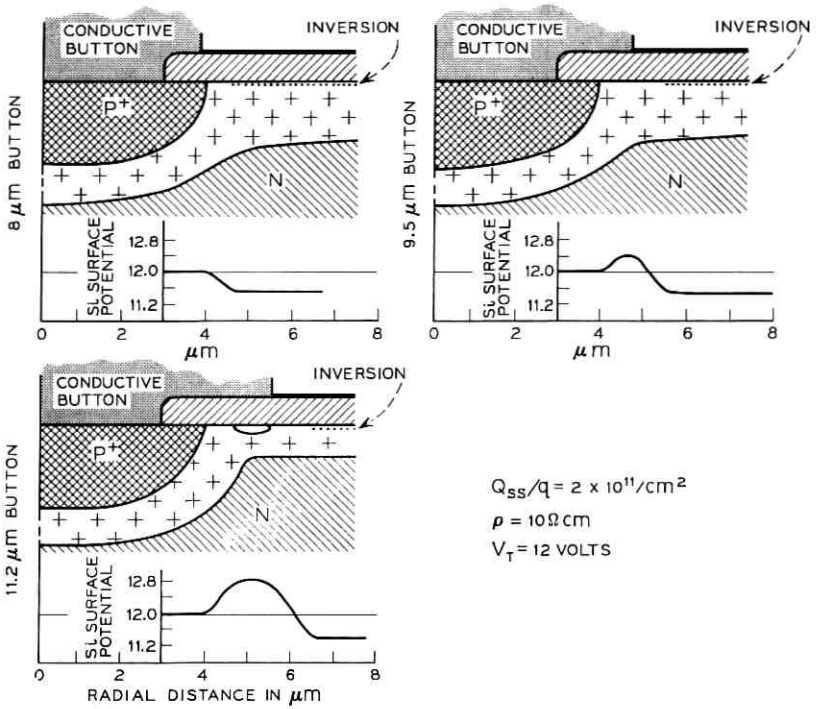


Fig. 27—Effect of conductive overlay diameter on diode discharged condition.

junction with large diameter P^+ islands (>10 μm), the button technique offers several fabrication advantages:

(*i*) Allowable range of sea resistance can be extended to 10^{13} to 10^{15} Ω/□.

(*ii*) For this structure, the critical parameters are easily measured physical dimensions, that is, diffused island and button diameters. Furthermore, the button diameter is not particularly critical provided it is larger than some minimum (approximately 1.5 μm greater than the P^+ island diameter). The maximum diameter is of course restricted to less than the diode center to center spacing, permitting substantial latitude on button size. However, two important fabricational restrictions must be observed. To assure signal uniformity the button diameter must be uniform over the diode array. Non-uniformities will be evident under light saturated signal conditions where the oxide capacitance contributes a significant portion of the signal. The second requirement

is that the P^+ island diameter be controlled to assure that the overlay will extend the required distance past the P^+ island edge to avoid inversion effects.

VII. DARK CURRENT

Surface area and volume components of the unit cell depletion region can be calculated as a function of target bias (see Fig. 28). Assuming that target dark current is the linear summation of the bulk and surface components, this geometric data, with appropriate scaling factors can be used to calculate the dark current characteristic, that is, assuming the surface component of dark current is given[7] by

$$I_s = \frac{N}{2} q n_i S_0 A_s \tag{4}$$

where

S_0 = surface recombination velocity in cm/second,
A_s = depleted silicon surface area of unit cell,
n_i = intrinsic carrier concentration,
N = number of diodes,
q = electronic charge,

and the bulk component is

$$I_{\text{bulk}} = \frac{N}{2} q n_i \frac{V}{\tau_g} \tag{5}$$

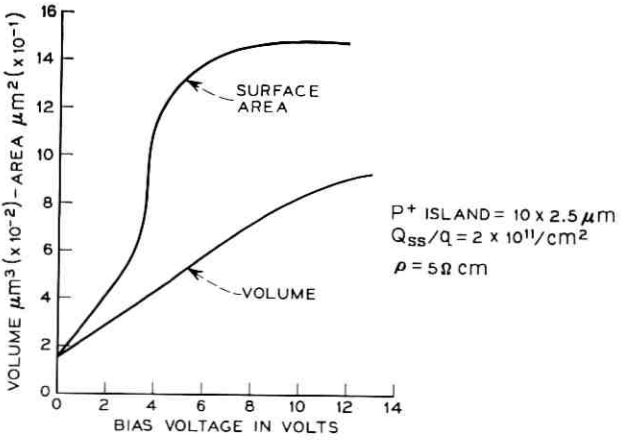


Fig. 28—Calculated depletion region geometric factors.

where

V = depleted unit cell volume,
τ_g = effective bulk generation lifetime.

For example, Fig. 29 illustrates the application of the above using measured values of S_0 and τ_g and compares this result to the measured dark current characteristic.

Figure 30 shows a family of dark current characteristics for a target with 5×10^5 diodes normalized to bulk and surface generation rates for a specific geometry. A 3-volt offset is used to account for beam limitations.

A plot of dark current versus diode diameter for 10 Ωcm material, $Q_{ss}/q = 3 \times 10^{11}/\text{cm}^2$, and $V_T = 12$ V is presented in Figure 31. It is clear, when operating above flatband, that a large P^+ island diameter improves the dark current by reducing the depleted surface area.

VIII. SUMMARY

The material presented thus far represents the results of a mathematical modeling of the silicon diode array. To be of practical use in target design the model must accurately represent target behavior. Since the complicated nature of the problem at hand precludes the normal testing of mathematical models by reduction to simplified cases, an experimental verification of the model's accuracy is the only recourse. In this section we present some comparisons of the calculated and experimentally measured target behavior. We also draw some general conclusions concerning target design.

8.1 *Comparison of Experimental and Calculated Results*

Determination of the target depletion region geometry is fundamental to the analysis of target performance. The most convenient experimental verification of the accuracy of these calculated depletion region geometries is to compare the calculated depletion region capacitance for the case of a uniform surface potential with the measured capacitance of an array covered with a metallic overlay or dot. Figure 32 illustrates such a comparison. The solid curve represents the calculated C–V characteristic for a target with a SiO_2–Si interface fixed charge density of $2 \times 10^{11}/\text{cm}^2$ and a resistivity of 5 Ωcm. The measured values were scaled from a gold dot covering 560 diodes. Notice the excellent agreement between the calculated curve and experimental measurements.

A statistical comparison of theory with measured signal current

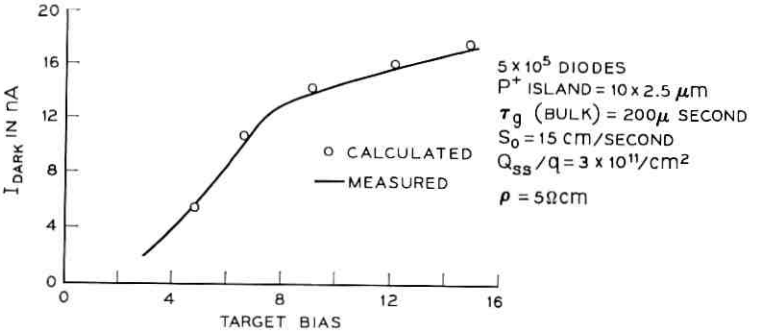


Fig. 29—Comparison of measured and calculated dark current characteristic for 5×10^5 diode array.

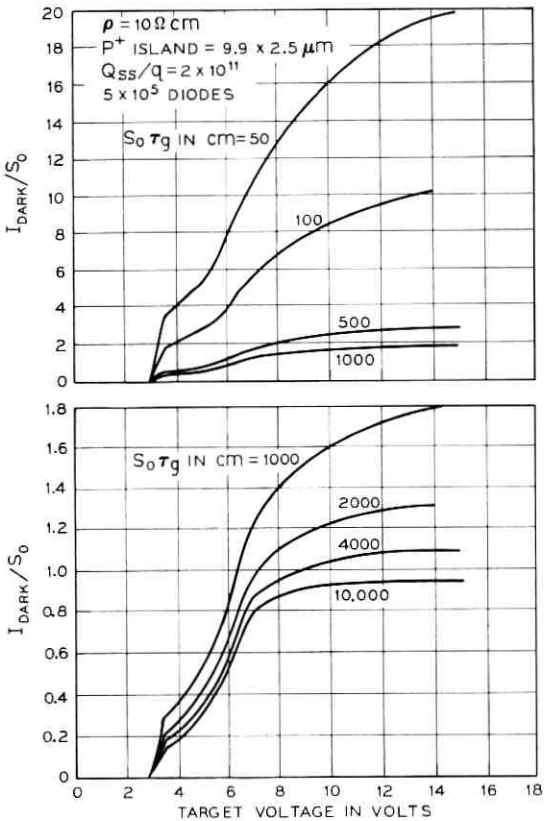


Fig. 30—Calculated normalized dark current characteristics for 5×10^5 diode array.

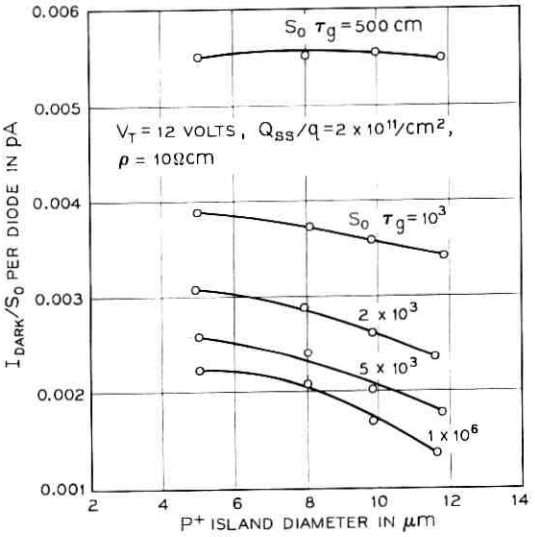


Fig. 31—Calculated relationship between dark current and P^+ island diameter.

versus diode diameter is given in Fig. 23. The range in measured maximum signal current reflects the normal variations in beam acceptance and sea resistance.

A comparison of calculated and measured dark current characteristics is illustrated in Fig. 29 for an array with measured parameters. It can be seen from these results that the bulk and surface components of dark current calculated from the depletion region geometric parameters and measured generation rates provide an adequate model for dark current.

8.2 *Some Generalizations on Target Array Geometry*

In general, several observations of the effect of target diode geometry on signal current can be made:

(*i*) For high sea sheet resistance ($\geqq 10^{14}\ \Omega/\square$) larger diameter diodes have greater signal capabilities.

First, for the case where signal is not limited by inversion this result is due to the simple geometric effect of having a larger capacitance associated with the junction.

Second, for the case where signal is limited due to surface inversion, larger diameter P^+ islands have the effect of delaying the onset of inversion (higher P^+ island potentials are possible).

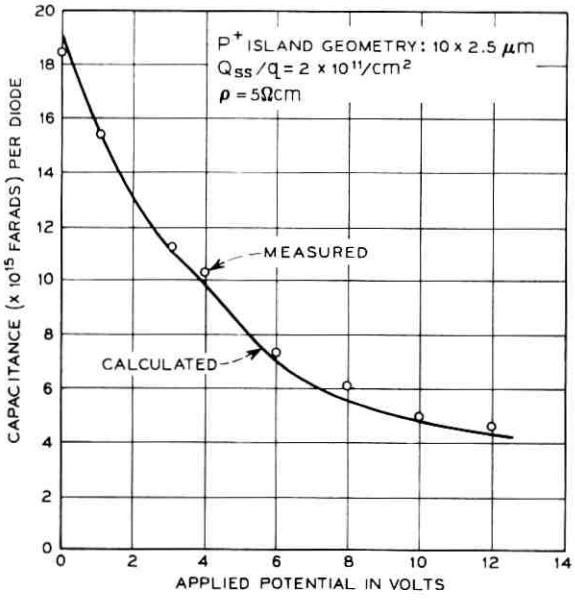


Fig. 32—Comparison of measured and calculated capacitance—voltage characteristic.

Third, increased beam acceptance for large oxide window diameter reduces both lag and signal limitations.

(*ii*) For the case where the major portion of diode leakage current is due to surface generation, increased P^+ island diameter reduces dark current due to the geometric effect of minimizing available current generating surface.

(*iii*) Silicon interface inversion effects can also be controlled by use of conductive overlay whose diameter exceeds the P^+ island diameter by 1.5 μm.

IX. ACKNOWLEDGMENTS

The authors wish to thank Messrs. H. E. Hughes and J. R. Mathews for their support of this work. We would particularly like to express our appreciation for the helpful ideas and suggestions offered by J. R. Mathews, R. D. Plummer, and L. H. Von Ohlsen. The dark current curves were generated by A. A. Yiannoulos.

REFERENCES

1. Reynolds, F. W., "Solid-State Light-Sensitive Storage," U. S. Patent 3-011-089, applied for April 15, 1958, issued November 21, 1961.
2. Crowell, M. H., Buck, T. M., Labuda, E. F., Dalton, J. V., and Walsch, E. J., "A Camera Tube with a Silicon Diode Array Target," B.S.T.J., *46*, No. 2 (February 1967), pp. 491–495.
3. Wendland, P. H., "A Charge-Storage Diode Device Vidicon Camera Tube," IEEE Trans. Elec. Devices, *ED-14*, No. 9 (June 1967), pp. 285–291.
4. Crowell, M. H., and Labuda, E. F., "Silicon Diode Array Camera Tube," B.S.T.J., *48*, No. 5 (May–June 1969), pp. 1481–1528.
5. Issacson, E., and Keller, H. P., *Analysis of Numerical Methods*, New York: John Wiley, 1966, p. 469.
6. Grove, A. S., and Fitzgerald, D. J., "Surface Effects on p-n Junctions: Characteristics of Surface Space Charge Regions Under Non-Equilibrium Conditions," Solid-State Elec., *9*, No. 8 (August 1966), pp. 783–806.
7. Grove, A. S., *Physics and Technology of Semiconductor Devices*, New York: John Wiley, 1967, p.301.

Image Storage and Display Devices Using Fine-Grain, Ferroelectric Ceramics

By A. H. MEITZLER, J. R. MALDONADO, and D. B. FRASER

(Manuscript received January 22, 1970)

Thin plates of ferroelectric ceramic in combination with transparent conductive and photoconductive films have been used to form device structures that can store a real image as a spatial variation in birefringence. This image can be viewed directly by suitably polarized transmitted light or projected onto a viewing screen. The stored image is erasable by the application of appropriate combinations of light and electric fields.

I. INTRODUCTION

Recent publications have called attention to the usefulness of fine-grained, lead zirconate-lead titanate ferroelectric ceramics in several kinds of electro-optic devices.[1-3] This paper reports initial experimental results from ferroelectric picture devices (ferpics) based directly on the electro-optic properties of these new materials. In addition, this paper discusses how the basic device may be used to advantage in several types of display systems.

II. BASIC PRINCIPLES OF OPERATION

A thin, transparent plate of lead zirconate-lead titanate ceramic, as it is initially formed, is optically isotropic. By the process of poling, it can be given a condition of uniaxial birefringence dependent upon the remanent polarization of the material. The poled material has its optic axis parallel to the polarization direction. Figure 1(a) shows how, by means of electrodes applied to its edges, a plate is poled so that the remanent polarization lies unidirectionally in the plane of the plate. In this condition (L-state) the plate exhibits uniform birefringence for polarized light incident normally. An image is stored by switching at least a portion of the domain polarization vectors in the areas where the plate is illuminated to a direction perpendicular to the plane of the plate (T-state), thus reducing the birefringence of these regions. The

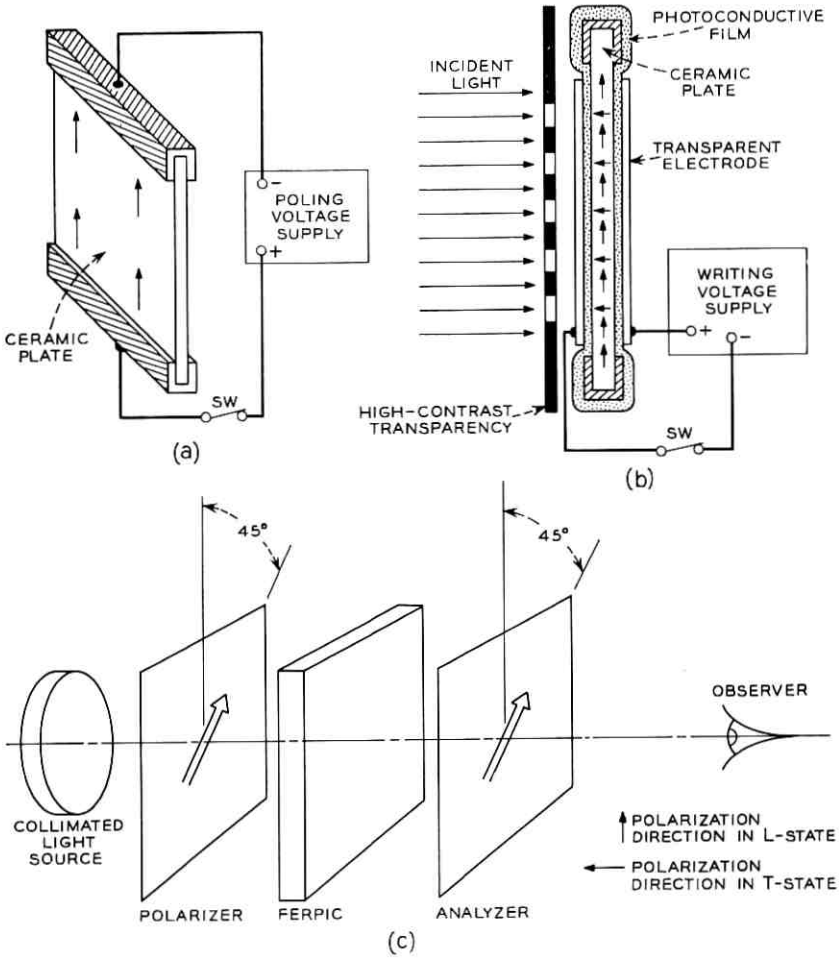


Fig. 1—Construction and operation of an elementary ferpic: (a) poling the ceramic plate, (b) storing an image, (c) observing the recorded image.

required perpendicular switching field can be obtained by means of the arrangement shown in Fig. 1(b). The ceramic plate is coated on both sides with a photoconductive film and sandwiched between transparent electrodes. In front of the ferroelectric picture device (ferpic) is placed a high-contrast transparency illuminated by a beam of collimated incident light. When voltage is applied to the transparent electrodes, the high impedance of the *dark* photoconductive regions prevents the

field inside the ceramic from reaching a value producing a significant amount of switching. In the *illuminated* regions, however, the impedance of the photoconductor is reduced and the field in the ferroelectric increases to the point that appreciable domain switching is produced causing the ceramic underlying the illuminated areas to be switched to the T-state. Thus a stored image is obtained as a spatial modulation of the birefringence of the ceramic. The image can be made visible by inserting the ferpic between a polarizer and analyzer as shown in Fig. 1(c). The image can be erased either by poling the sample in the plane (L-state) or by thermally depoling the material.

III. EXPERIMENTAL RESULTS FROM AN ELEMENTARY FORM OF FERPIC

The details of an experimental device structure that operates in this manner are shown in Fig. 2. The device uses a 50 μm thick plate of ceramic* having 65 percent lead zirconate and 35 percent lead titanate with two atom percent of lanthanum (designated 65/35-2 La). The grain size of this material is approximately one micron. Originally, the plate is poled to have a remanent polarization in the plane of the plate (L-state) by means of an applied field of approximately 20 kV/cm. The sample is poled before the photoconductive film and the transparent electrodes are applied. The specific poling conditions used for a given plate are adjusted to give a state of remanent polarization causing a half-wavelength of phase retardation when polarized light is transmitted at normal incidence through the plate.

After the plate is initially in a condition such that all regions of the plate in the area used for storing the picture are in the L-state, the photoconductive film and transparent conductive films are applied. In the version of a ferpic outlined in Fig. 2, the photoconductive film is PVK† applied simultaneously to both sides of the plate by a dip-coating technique. Transparent conductive electrodes are next applied. In our experimental devices, half-transparent films of Cr-Au are vapor deposited on the two surfaces; in practical devices, more transparent electrodes of tin oxide or indium oxide would be preferred. Fine wire leads are attached to these electrodes and used to connect the device to the voltage source used to supply a switching field in the thickness direction.

Figure 2, in addition to showing the structural features of the

* The ceramic used in our experiments was produced by Clevite Corporation, Cleveland, Ohio 44108.

† The abbreviation PVK stands for polyvinyl carbazole. The material used was obtained from Polyscience, Incorporated, P. O. Box 4, Rydal, Pennsylvania 19046.

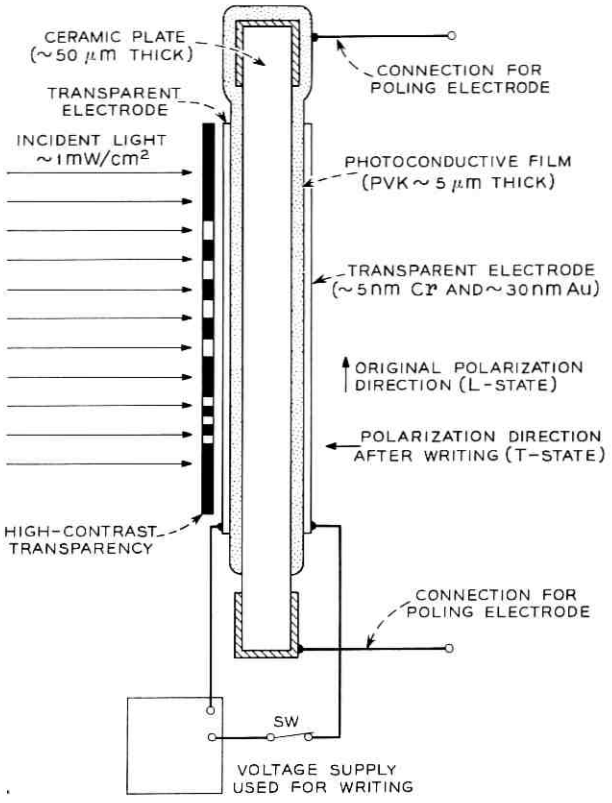


Fig. 2—Construction details of an elementary ferpic and experimental arrangement used to store an image.

device, shows the arrangement used to write the picture information into the ceramic plate. A high contrast transparency is placed immediately in front of the ferpic and illuminated with collimated light. The voltage supply is pulsed on. In the regions where the light passes through the transparency, the photoconductive film becomes conductive and the field in the ceramic becomes large enough to produce switching in the form of 90° polarization rotation. (This mode of operation is described by Land and Thacher in connection with light gate structures using various configurations of metal electrodes.[1] In ferpic devices, the 90° polarization-rotation mode of operation is obtained by the use of photoconductive films and transparent conductive films working in combination with metal electrodes.) Localization of the switching field

requires that the dielectric constant of the ceramic, K_{CER}, be much larger than the dielectric constant of the photoconductor, K_{PVK}. For the materials used in our experimental devices (K_{CER}/K_{PVK}) $\approx$ 400. For a 50 μm thick ceramic plate, the writing conditions were the following: (*i*) a white light flux of 2 mW/cm^2, (*ii*) a 200 V supply, and (*iii*) a voltage pulse duration of approximately 1 minute.

As already described, the stored image can be made visible by inserting the ferpic between a polarizer and analyzer, as indicated in Fig. 1(c), and illuminating it with light from a collimated, monochromatic light source. (The degree of collimation and monochromaticity involved are not critical. Most of our experimental work is done using white light from ordinary incandescent sources and, in fact, the two photographs included with this report were made with this sort of light.) If the phase retardation produced by regions in the L-state is $\lambda/2$, efficient use of light is obtained when polarizer and analyzer are set parallel to each other and at an angle of 45° to the electric polarization vector in the ceramic. The regions in the L-state appear opaque; the regions switched to the T-state present no birefringence to the incident light and appear as bright areas to the viewer.

The ferpic just described can be used as a kind of photographic plate capable of two-level (black and monochrome) image storage with a respectable degree of resolution. Figure 3 shows photographs made

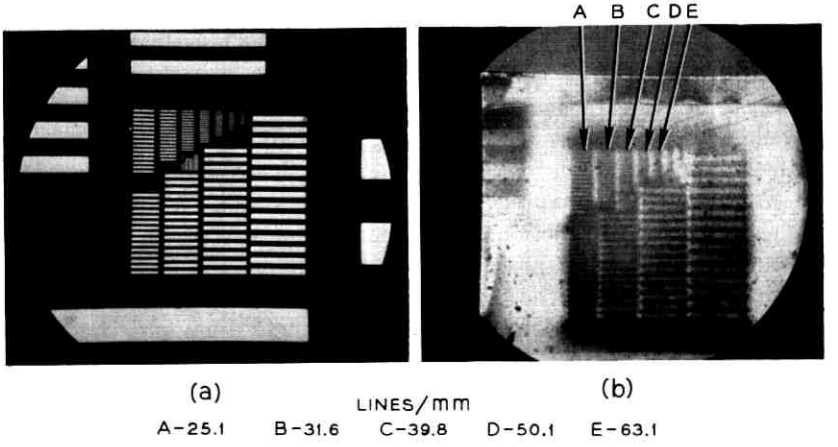


Fig. 3—(a) The original resolution test chart as seen with a microscope using low power magnification.
(b) The resolution test chart image stored in a ferpic and rendered observable by means of a polarizer and analyzer used with the microscope.

using a low power microscope to view a pattern from a resolution test chart stored in a 50 μm thick ceramic plate. The number of lines/mm in the individual columns marked A, B, C, D, and E is indicated in the figure.

This photograph demonstrates that resolutions better than 30 lines/mm are obtainable under the described conditions of operation. A number of experiments have been carried out to establish the principal factor limiting the resolution. Fringing fields within the 50 μm thick plate appear to be the principal factor, since both the optical techniques and the photoconductor used have been demonstrated to have at least an order of magnitude finer resolution capabilities. The present observed resolution is already equivalent to 2 to 3 cycles of variation over a distance equal to the thickness of the plate.

The device described in Fig. 2 has limited usefulness since, like a photographic plate, it can be used only once. The transparent electrodes on the surface prevent the polarization vector under these electrodes from being switched back to the L-state when a voltage is applied to the poling electrodes. We will next consider a form of ferpic that offers the capability of being electrically changeable.

IV. AN ELECTRICALLY CHANGEABLE FERPIC USING AN INTERDIGITAL ELECTRODE ARRAY

An interdigital electrode array deposited on one side of a ceramic plate provides in principle a means of switching the polarization vectors back into the plane of the plate after they have once been switched normal to it. An exploded view of the layer structure proposed for use in a changeable ferpic is shown in Fig. 4. In addition to enabling the plate to be switched into the L-state, use of the interdigital array has the great practical advantage of reducing the voltage required to pole in the longitudinal direction. As already indicated, 20 kV is required to pole over a 1 cm distance. If the line elements have the same spacing as the plate thickness, the same voltage supply ($\sim$200 V for a 50 μm thick plate) can be used to establish both longitudinal and transverse switching fields. The obvious disadvantage to this approach is that the stored image will now be broken into a number of discrete lines. The extent to which the presence of the lines is evident and objectionable will depend on the details of use.

The ferpic structure shown in Fig. 4 constitutes an electrically changeable image storage and display device that functions in three steps: (*i*) RESET, (*ii*) WRITE and (*iii*) VIEW. The first two steps are illustrated

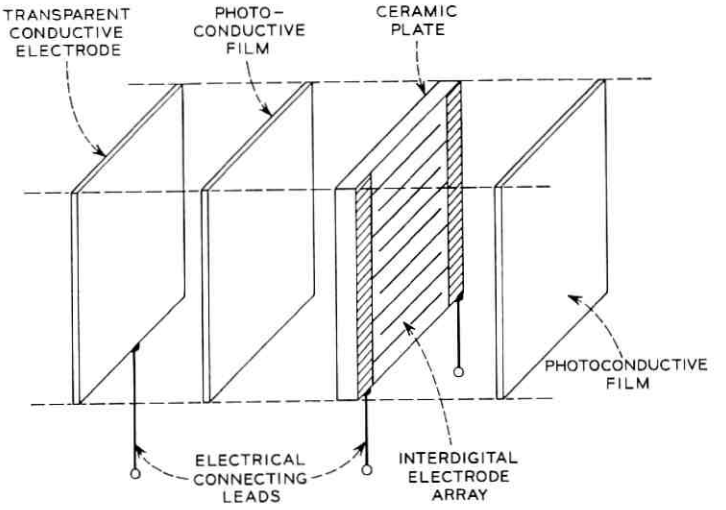


Fig. 4—Exploded view of the layered structure used to form an electrically changeable ferpic.

in Fig. 5(a) and (b); the third step involves essentially the same elements shown earlier in Fig. 1(c). In the RESET step the electrodes of the interdigital array are connected to the voltage supply and the resulting field switches the remanent polarization vectors predominantly into the plane of the plate. In this condition of polarization (L-state), every region of the plate has maximum birefringence for linearly polarized light incident normally. In the WRITE step, the elements of the array are connected in parallel to one terminal of the supply and the other terminal is connected to the transparent conductive electrode. Light is directed at the area to be switched causing the photoconductive layers to conduct and the electric field in the ceramic under the illuminated area to exceed the coercive field.

If the image to be stored in the ferpic is broken into elements, the spacing between lines of the interdigital array should be less than or equal to the size of an element. Writing the image an element at a time, as in a television picture, can be accomplished by modulating either the addressing light beam or the power supply. It is equally conceivable that the picture could be formed all at one time by projecting some desired image on the plate and switching the illuminated elements.

The basic ideas of the interdigital array device were first demonstrated in a device structure using PVK films. The details of this interdigital-

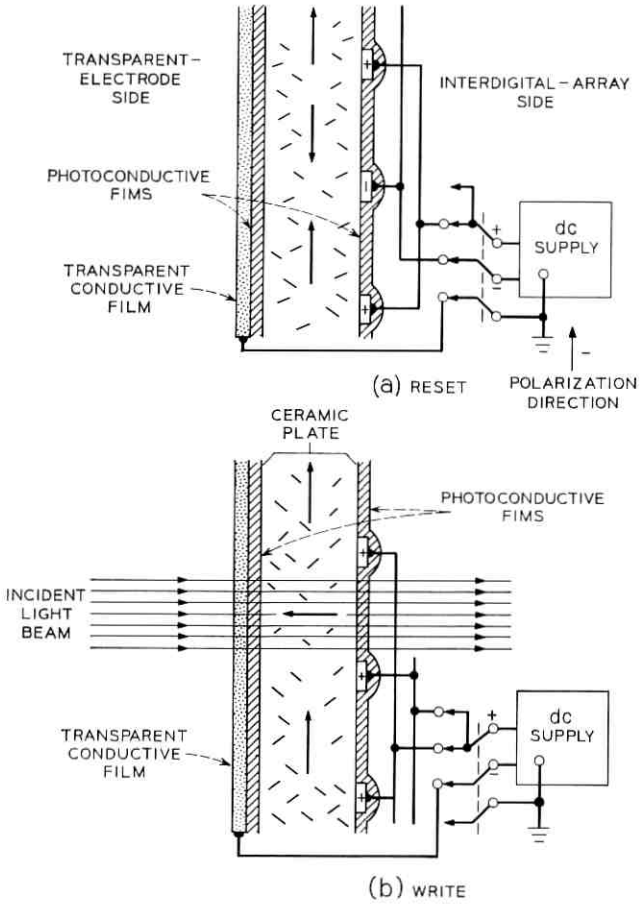


Fig. 5—Details of the RESET and WRITE steps of operation for a ferpic utilizing 90° polarization rotation in the ceramic plate.

array ferpic are shown in Fig. 6, and an example of the image storage obtained is shown in Fig. 7. The upper part of Fig. 7 shows the original, simple, high-contrast image. The lower part shows the image observed through a low-power polarizing microscope. The image was stored in a 50 μm thick ceramic plate with a square working area 0.8 cm on a side.

The experimental structure of Fig. 6 differs from the original structure described in Fig. 4 by the inclusion of an additional transparent conductive electrode on the array side of the ferpic. This additional film is needed because the PVK films do not have a high enough conductivity

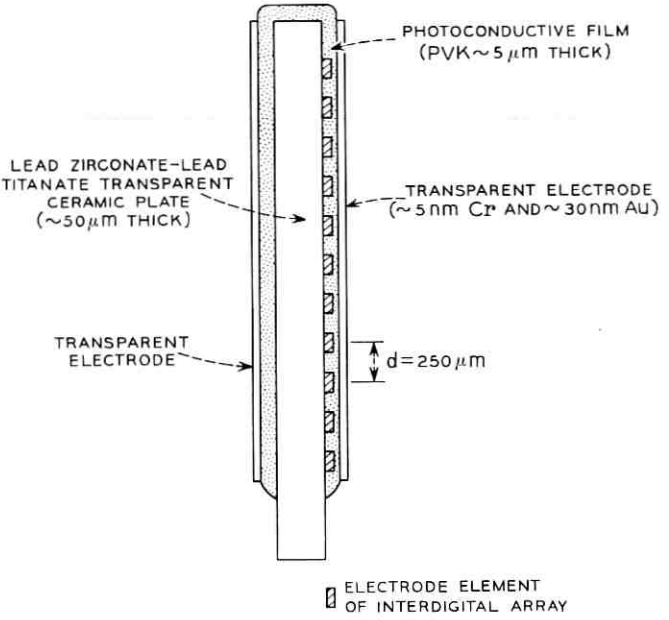


Fig. 6—Construction details of a ferpic using an interdigital array to switch the polarization into the plane of the plate.

when illuminated to establish an equipotential region between the elements of the electrode array. While the addition of the transparent conductive electrode to the array side of the ferpic solves the problem of the low conductivity in the PVK, this modification has the disadvantage that it hinders erasure. The PVK film between the transparent electrode and the array is now subject to breakdown field strengths when the RESET voltage is applied. In addition, experiments with PVK films on devices using the 90° polarization rotation have indicated that the PVK film constrains the motion of domains, probably through a mechanism of trapping polarization charges at the ceramic-PVK interface.

Our most recent experimental efforts have resulted in two significant developments, one related to the performance of photoconductive films used in interdigital-array ferpics and the other related to a new version of a ferpic.

In connection with photoconductive films, we have been successful in realizing CdS films with a high light-to-dark conductivity ratio and a high conductivity when illuminated. These films have been sputter

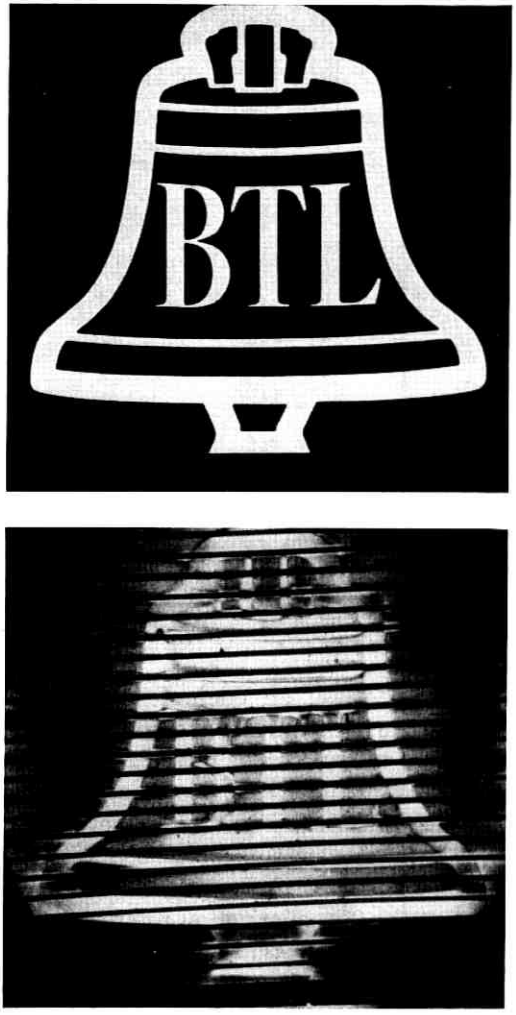


Fig. 7—A simple, high-contrast image stored as a variation in birefringence in a ferpic using an interdigital electrode array: upper, the original image; lower, the image viewed in the ferpic by means of a polarizing microscope. (The horizontal lines apparent in the photograph are formed by the electrodes of the interdigital array and these are spaced 250 μm apart on the actual device.)

deposited on device structures consisting of ceramic plates with interdigital arrays vapor deposited, in registration, on both surfaces. Images with nearly the quality of that shown in Fig. 7 have been successfully stored in the CdS film devices, and, what is more important, these devices have demonstrated a significantly greater degree of changeability than interdigital devices using PVK films. The experiments with these devices demonstrate that the ceramic and the CdS films can work together successfully with interdigital arrays to provide the intended mode of operation.

In connection with the new form of ferpic, a refined structure has been developed which eliminates the necessity of having a separate set of electrodes or interdigital arrays in order to switch the ceramic into the L-state. In this new device structure, a thin ceramic plate and photoconductive film are sandwiched between transparent electrodes, and the stack is bonded to a relatively thick, transparent substrate. The ceramic plate is put in tension along one direction by slightly flexing the substrate. The direction of the tension axis in the ceramic becomes a preferred direction along which the polarization vectors in individual domains tend to align. (Because of the use of a permanent strain to establish this preferred direction, the device is called a "strain-biased ferpic.") In the strained condition, the ceramic can be switched between two states, corresponding to an L- and T-state, by the application of fields in the thickness direction. As in the earlier versions of the device, only the illuminated regions can be switched to the T-state. In order to reset the whole device to the L-state, the whole active area is flooded with illumination while the reset voltage is applied. An important feature of the new device is that its structure permits localized switching to the L-state as well as to the T-state. A paper describing in detail the structural and performance features of the strain-biased ferpic is presently in preparation and will be published at a later time.

V. OPTICAL DISPLAY SYSTEMS WITH LASER BEAM ADDRESSING

Figure 8 shows the essentials of a projection display system using a ferpic in combination with a laser beam addressing module and a source of viewing light. A dichroic mirror that transmits at the projection wavelength and reflects at the write wavelength is used to enable the laser subsystem to be positioned off the main axis. The arrangement shown in the schematic drawing is intended to be an electronically changeable, projection display. In this display system, there are three stages to a complete cycle of operation: (*i*) RESET, (*ii*) WRITE and (*iii*)

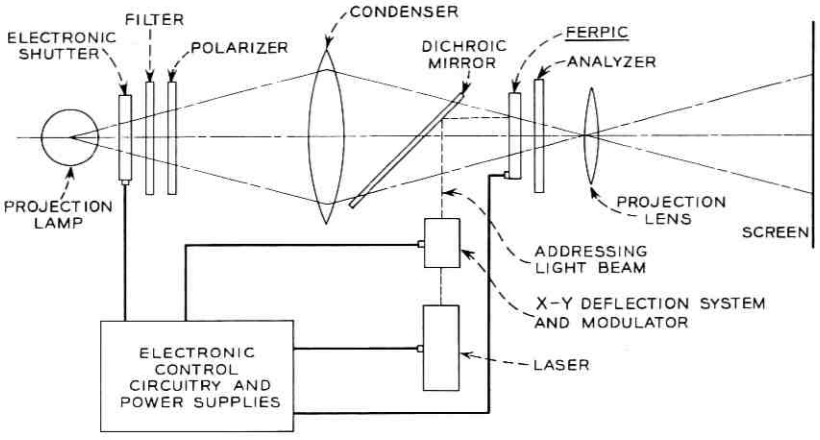


Fig. 8—A projection display system using a ferpic in combination with a laser beam addressing module and a source of viewing light.

VIEW. During the RESET stage, the viewing light and laser beam are off and the whole ferpic is switched to the L-state. During the WRITE stage, the viewing light remains off while the picture is formed an element at a time by moving the laser beam successively through all the address points. The laser beam is steered by an X-Y deflection system which could be an ultrasonic light deflector system of the sort recently developed for use in an optical memory system[4] or a mechanical scanning system. The choice exists of controlling the switching of a picture element by modulating either the intensity of the laser beam or the intensity of the voltage pulses supplied to the electrodes of the ferpic.

Once the picture has been formed, the viewing source is flashed on. An important advantage of this display system is that, once written, the picture can be held without any further expenditure of electrical power. This feature would make a ferpic display attractive for many applications now met using storage cathode-ray tubes. The fact that the basic light controlling element has intrinsic memory has important implications for practical applications. As a consequence of this intrinsic memory the display is able to hold a given picture for an indefinitely long time. Furthermore, the picture is not volatile in the sense that it is retained in the event of a power failure. Depending upon the details of construction and associated circuitry, it may be possible to break up the stored picture into individual lines or elements and use a "periodic update" mode of operation of interest in connection with bandwidth-reducing techniques of picture transmission.

In the display system proposed here, the image is formed one picture element at a time by a scanned laser beam, but is viewed or projected as a whole using incoherent light in a conventional projection system. This separation of the scanning and viewing functions allows a relatively low-power laser, in any convenient wavelength range, to be used for the scan, while a powerful, efficient incoherent source provides the actual viewed light.

We are not able at the present stage of the device development to state whether or not the device will have the lifetime and switching speed capabilities required to make it useful in real-time display systems. (For this sort of application, ferpics would be in competition with devices like the TITUS tube[5] and the EIDOPHOR tube.[6]) However, there are applications that can make use of a slowly-scanned, high-resolution image storage device that can hold an image for an arbitrarily long time and either continuously display it by an optical projection system or project it onto some form of hard-copy print-out. This sort of application would not require the high cycling rates or extreme lifetimes of a device used in real-time display systems.

With regard to obtaining maximum lifetime in a ferpic, there is an important point that deserves explicit statement: It is not necessary to rotate the polarization vector through a full 90° in order to obtain a useful effect. Operation under conditions producing full 90° polarization-rotation provides the means of obtaining the maximum birefringence change with a minimum material thickness.* On the other hand, such a large change in polarization direction produces strains that could lead to premature failure of the ceramic. In devices using ceramics with sufficiently low optical loss,† there exists the alternative of switching the average polarization through an angle much less than 90°, and then using increased thickness to obtain the desired half-wavelength retardation change.

VI. REFINING THE BASIC DEVICE TO OBTAIN GRAY-SCALE AND COLOR CAPABILITIES

A ferpic, as described in this paper, is basically a two-level (black and monochrome) image storage and display device. However, modification of the device structure and system configuration to include color

* According to Land and Thacher,[1] a Δn of -0.022 is obtained in 65/35 − 2 La ceramic at maximum remanent polarization. This Δn requires only 16 μm of material to produce a half-wavelength of retardation for 6328 Å light.

† A new hot-pressed, lead zirconate-lead titanate ceramic composition has been developed at Sandia Laboratories with lower optical loss and improved electrical switching characteristics.[7]

and a gray-scale appears to be straightforward. For example, the system of Fig. 8 can be modified for color by using a sequential-field, color-scan technique. Instead of one source illuminating the ferpic, there would be three light sources (red, blue, green) with appropriate means to strobe these in sequence. The ferpic would operate as already described except that there would now be three different RESET operations per frame, with a voltage applied to produce a retardation of a half-wavelength for the color flashed during the VIEW stage. In practice the operations RESET-WRITE-VIEW would be repeated in sequence for each color.

Modification of the basic system of Fig. 8 to obtain a gray scale requires that intermediate states of birefringence—and hence polarization—be reproducibly attainable, so that the ceramic plate functions as an intensity modulator and not just as a simple shutter. Experiments[3] with elementary, light-gate devices using the fine-grain ceramic have shown that it is possible to obtain reproducible partial switching with suitable circuitry. The development of a ferpic with gray-scale capability will depend strongly on the switching properties of the ceramic and the photocurrent response characteristics of the photoconductive films with which the ceramic is used. It is still too early to state uniquivocally that a gray-scale device can be obtained with all other desired characteristics, but the possibility is certainly there.

VII. CONCLUSION

The recent development of fine-grain, electro-optic ceramic materials has provided the means of realizing ferroelectric devices capable of storing high-contrast images under the control of electrical voltages. The image is stored as a variation of birefringence in a thin, ceramic plate and can be viewed directly by suitably polarized transmitted light or projected onto a viewing screen. Experimental devices have demonstrated a resolution capability of approximately 50 lines/mm in 50 μm thick ceramic material and have been able to hold the image with no apparent change for times of the order of several months. While our early experimental devices have had only limited electric changeability, the basic material system is certainly capable of providing this essential feature. Images stored in these devices have been projected and viewed by means of simple, desk-top, commercially available, 35 mm projection displays (modified by the inclusion of a polarizer and analyzer). Display systems of the sort considered appear to be well-suited to applications such as high-resolution, slow-scan, and document display.

VIII. ACKNOWLEDGMENTS

The authors take pleasure in acknowledging many helpful discussions of their work with L. K. Anderson and H. Melchior. J. W. Farrell, T. H. Lalonde, and W. J. Nowotarski assisted in the preparation and evaluation of the experimental ferroelectric picture devices.

REFERENCES

1. Land, C. E., and Thacher, P. D., "Ferroelectric Ceramic Electro-optic Materials and Devices," Proc. IEEE, *57*, No. 5 (May 1969), pp. 751–768.
2. Thacher, P. D., and Land, C. E., "Ferroelectric Electro-optic Ceramics with Reduced Scattering," IEEE Trans. Electron Devices, *ED-16*, No. 6 (June 1969), pp. 515–521.
3. Maldonado, J. R., and Meitzler, A. H., "Ferroelectric Ceramic Light Gates Operated in a Voltage-Controlled Mode," IEEE Trans. Electron Devices, *ED-17*, No. 2 (February 1970), pp. 148–157.
4. Pinnow, D. A., Van Uitert, L. G., Warner, A. W., and Bonner, W. A., "Lead Molybdate: A Melt-Grown Crystal with a High Figure of Merit for Acoustooptical Device Applications," Applied Physics Letters, *15*, No. 3 (August 1969), pp. 83–86.
5. Marie, G., "Un Nouveau Dispositif de Restitution d'Images Utilisant un Effet Electro-optique: Le Tube Titus," Philips Res. Repts. *22*, No. 2 (April 1967), pp. 110–132.
6. Labin, E., "The Eidophor Method for Theatre Television," Jour. Soc. Motion Picture and Television Eng., *54*, No. 4 (April 1950), pp. 393–406.
7. Haertling, G. H., and Land, C. E., "Hot Pressed Ferroelectric Ceramics for Electro-optic Applications," Paper Presented at the 1970 Annual Meeting of the American Ceramic Society, May 2–7, 1970, Phila., Pa.

Some Mathematical Properties of a Scheme for Reducing the Bandwidth of Motion Pictures by Hadamard Smearing

By E. R. BERLEKAMP

(Manuscript recieved September 29, 1969)

M. R. Schroeder recently proposed a scheme for compression of motion picture data by taking the difference of two successive frames and then smearing.[1] *The smearing is accomplished by a Hadamard matrix.*

If the Hadamard matrix is of a certain particularly well-understood type, then we show that if the input differential picture consists of a small odd number of large pulses of identical magnitudes (but arbitrary signs), then the output will consist of three components:

(i) Large pulses of equal magnitude and the correct signs, matching each of the input pulses.

(ii) One additional "stray" large pulse, of magnitude equal to the others, but located at a point where the input was zero.

(iii) Scattered pulses of amplitude low relative to the pulses of types i and ii, but so numerous that they consume $(\pi - 2)/\pi$ *of the total energy of the output differential picture.*

We give an explicit formula for the amplitude of each of these pulses.

The problem of determining the distributions of all possible outputs of the proposed system for other classes of inputs is shown to be equivalent to the unsolved problem of finding the weight enumerators for the cosets of the first order Reed-Muller codes.

I. INTRODUCTION

The fact that successive frames of a motion picture are often very nearly alike has led to the consideration of schemes which transmit, for each point of the picture, the difference between the amplitude of the present frame and the amplitude of the previous frame. Since

this differential picture* is frequently zero at many points, there is reason to hope that the bandwidth required for transmission of the differential picture could be greatly reduced by appropriate coding.

One such coding scheme which has been considered by W. K. Pratt, J. Kane, and H. C. Andrews,[2] and refined by M. R. Schroeder[1] is the following: let the differential picture be represented by a real n-dimensional vector, $\mathbf{v}$. (For example, if the picture is represented by a 100×100 grid, then $n = 10000$.) Let $\mathfrak{C}$ be an $n \times n$ *Hadamard* matrix, which is a self-orthogonal real matrix all of whose entries are ± 1, and let the smeared differential picture (or transformed differential picture), $\mathbf{x}$, be defined by $\mathbf{x} = \mathfrak{C}\mathbf{v}/(n)^{\frac{1}{2}}$. Let Q be the power-preserving clipping operator, defined by

$$Q\mathbf{x} = \frac{(||\mathbf{x}||)^{\frac{1}{2}} \operatorname{sgn}(\mathbf{x})}{(n)^{\frac{1}{2}}},$$

where sgn $(\mathbf{x})$ is the n-dimensional vector whose ith component is $+1$ or -1, depending on the sign of the ith component of $\mathbf{x}$. Since we wish the quantizer to have only two outputs, we cannot take sgn $(0) = 0$. Unless stated otherwise, we assume that sgn (0) is undefined. Schroeder has asserted that the vector $\mathbf{y} = Q\mathbf{x} = Q\mathfrak{C}\mathbf{v}/(n)^{\frac{1}{2}}$ provides an appropriate "encoding" of the differential picture $\mathbf{v}$. To "decode" one computes $\mathbf{z} = 1/(n)^{\frac{1}{2}}\mathfrak{C}'\mathbf{y}$. The question to be studied in this paper is the quality of $\mathbf{z}$ as an approximation to $\mathbf{v}$.

Some of the heuristic arguments favoring this proposed scheme are the following: Since successive frames are frequently very similar, the differential picture will have near-zero amplitude at most points. In a typical case when the camera is focused on a moving subject and a fixed background, the differential picture will be identically zero at all background points. If the subject and the background each has uniform color (the simplest plausible case), then the differential picture will be nonzero only at those points on the boundary of the subject. Furthermore, all of the nonzero amplitudes in the differential picture will have equal magnitudes, although their signs will depend on whether they are on the leading or trailing edge of the moving subject. The

* To be precise, the "differential picture" *should* consist of the difference between what the present frame actually is and what the decoder thought the last frame was. Since all of the errors in the system are assumed to arise from quantization, rather than from any sort of unpredictable noise on the communications channel, the encoder may include a replica of the decoder, thereby enabling it to compute what the decoder thought the last frame was. Each transmitted differential picture then includes an attempt to correct the cumulative effects of all previous errors. In this paper, we study only the quantization noise introduced in the encoding and decoding of a single differential picture, ignoring the complicated dynamic questions which arise when one studies the behavior of the system during several successive frames.

conventional manner of encoding the differential picture is to quantize the amplitude at each gridpoint. This scheme will introduce no quantization error at all on the background points, which have zero amplitude, but a relatively high number of quantization levels may be required to keep the quantization errors along the outline of the subject down to a tolerable level. The Hadamard transform of the differential picture, on the other hand, will have its energy spread out relatively uniformly among the grid points. A coarse quantization of the smeared differential picture will introduce quantization errors throughout the differential picture in a relatively uniform manner. When the quantized smeared differential picture is unsmeared, the quantization errors, being somewhat independent, should tend to cancel out. It is thus hoped that a coarser quantization of the smeared differential picture might yield a decoded differential picture of the same fidelity as a substantially finer quantization of the original, unsmeared differential picture.

A somewhat more theoretical discussion of the effects of quantization in the Hadamard transform domain is given in Section VI of Pratt, Kane, and Andrews.[2] The main result is that the Hadamard transformation preserves energy. Hence, if the amplitudes at the various points in the transformed differential picture are independent zero mean gaussian random variables, then the energy of the noise introduced by a two-level quantizer would be $(\pi - 2)/\pi$ of the total energy in the output differential picture, both before and after unsmearing. From this viewpoint, the major attraction of smearing is that it distributes the quantization noise uniformly throughout the picture. If our simple model of a differential picture (which has nonzero amplitude only along the outline of the subject) is coarsely quantized, then all of the quantization noise appears on the outline, where it will tend to blur the subject. However, if this differential picture is smeared, coarsely quantized, and unsmeared, then its quantization noise should be evenly distributed throughout the subject and the background.

We shall now study the relationship between the original differential picture, $\mathbf{v}$, and the decoded differential picture, $\mathbf{z}$. It is clear that the energy in the vector $\mathbf{z}$ is always identical to the energy in the vector $\mathbf{v}$. Hence, for purposes of analysis, it is easiest to compute $\mathbf{z}$ according to the formula

$$\mathbf{z} = A\mathcal{H}' \operatorname{sgn} \mathcal{H}\mathbf{v}$$

where for each frame A is a non-negative scalar chosen to make the energy in $\mathbf{z}$ equal to the energy in $\mathbf{v}$. In this paper, we often omit the actual calculation of A.

In the case where $\mathbf{v}$ has only one nonzero component, then sgn $\mathcal{H}\mathbf{v} = \mathcal{H}\mathbf{v}$ and $\mathbf{z} = A\mathcal{H}'\mathcal{H}\mathbf{v}$. Since $\mathcal{H}'\mathcal{H} = nI$ (where I is the $n \times n$ identity matrix) it follows that $\mathbf{z} = \mathbf{v}$. In other words, the system transmits a single pulse without error.

On the other hand, when $\mathbf{v}$ has only two nonzero components, then the component with the larger amplitude dominates the component with the smaller amplitude. In this case $\mathbf{z}$ again has a single nonzero component, even though $\mathbf{v}$ had two nonzero components. However, the choice $\mathbf{v} = [1, 1-\epsilon, 0, 0, 0, \cdots, 0]$ results in an ambiguity. If we instead write $\mathbf{v} = [1, 1, \epsilon, 0, 0, 0, \cdots, 0]$, then we may actually find that, in the limit as $\epsilon \to 0$, $\mathbf{z} \to [1, 1, 0, 0, \cdots, 0]$. If $\mathbf{v} = [1, 1, 0, 0, \cdots, 0]$, then $\mathbf{z}$ is undefined because it depends on sgn 0, which is either plus or minus. In fact, $\mathbf{z}$ is undefined whenever $\mathbf{v}$ has an even number of nonzero components, all of equal magnitude; but this difficulty might be removable by adding an appropriate background noise function into $\mathbf{v}$, or by choosing sgn(0)'s independently at random. To avoid the necessity of such considerations, we devote our primary attention in this paper to the case in which $\mathbf{v}$ has an *odd* number of nonzero components, all of unit magnitude (but arbitrary sign). In this case, every component of $\mathcal{H}\mathbf{v}$ is an odd integer. Since every component of $\mathcal{H}\mathbf{v}$ must therefore have magnitude at least 1, the sgn function is defined and the analysis remains valid in the presence of a small background noise in any or all components of $\mathbf{v}$.

II. HADAMARD MATRICES

The requirement that any three rows of a Hadamard matrix be pairwise orthogonal leads to the immediate conclusion that if $n > 2$, then an $n \times n$ Hadamard matrix can exist only if n is a multiple of 4. The question of whether or not there actually do exist $n \times n$ Hadamard matrices for all $n \equiv 0 \bmod 4$ is now one of the most intriguing unsolved problems in combinatorial theory. Many ingenious constructions have been proposed, and several of them succeed in obtaining Hadamard matrices for an infinite number of (scattered) values of n. For example, if n is a multiple of 4 and $n - 1$ is a prime-power, then a well-known construction based on quadratic residues in the finite field $GF(n-1)$ yields an $n \times n$ Hadamard matrix. Many other constructions for Hadamard matrices are given in Chapter 14 of Hall,[3] and more recent constructions have been presented by Spence,[4] Goethals and Seidel,[5] and Wallis.[6,7] The smallest value of $n \equiv 0 \bmod 4$ for which no $n \times n$ Hadamard matrix has yet been constructed is $n = 188$.

For many values of n, there exist Hadamard matrices with additional structure. For example, some Hadamard matrices have the property that the first row and first column consist entirely of $+1$'s, and the remaining $(n-1) \times (n-1)$ submatrix has the property that each of its rows is a cyclic shift of the previous row. Such matrices are called *cyclic* Hadamard matrices. They are known to exist whenever $n-1$ is prime, or when $n-1$ is the product of twin primes, or when n is a power of 2. A computer search by Thoene & Golomb[8] and some calculations by Baumert[9] have shown that no cyclic Hadamard matrices of other orders less than 1000 exist, with the possible exceptions of $n = 400, 496, 628, 652, 784, 976$.

From the viewpoint of an algebraic coding theorist, a shortened Hadamard matrix (obtained from a standard Hadamard matrix by multiplying each row by an appropriate sign to make the first column all $+1$'s, and then deleting the first column) is equivalent to an *equidistant binary code*. The n codewords are taken as the rows of the shortened Hadamard matrix, with each $+1$ replaced by 0 and each -1 replaced by 1. Since the dot product of any pair of rows in the shortened Hadamard matrix is -1, the distance between any pair of words in the binary code is $(n+1)/2$. Further discussion of such codes is given in Section 13.5 of Berlekamp.[10]

The best-understood class of equidistant binary codes is the maximal-length shift-register codes, which are also called shortened first-order Reed-Muller codes. In addition to being cyclic and equidistant, they are *linear* over the binary field, which means that the component by component binary sum of any pair of codewords is another codeword. Stated in terms of the original Hadamard matrices, this property means that the componentwise product of any pair of rows of the Hadamard matrix is another row of the same Hadamard matrix. Although Hadamard matrices with this property are relatively rare, they exist for every n which is a power of 2. Because of their correspondence to the Reed-Muller codes, these matrices are comparatively well-understood, and we shall henceforth confine our discussion to Hadamard matrices of this type. Such matrices may be taken as cyclic.

III. THE INDUCED COORDINATIZATION

A $2^k \times 2^k$ Hadamard matrix corresponding to a Reed-Muller code induces a (non-unique) coordinatization on the 2^k components of each row, associating each component with a k-dimensional vector over $GF(2)$. A set of 2^i coordinates is said to form an affine subspace iff the

corresponding 2^j k-dimensional vectors form an affine j-dimensional subspace over $GF(2)$. Similarly, a set of m coordinates are said to be linearly (or affine) independent iff the corresponding k-dimensional vectors are linearly (or affine) independent.

If a set of m binary k-tuples α_1 , α_2 , $\cdots$ α_m are affine independent, then they span an $(m - 1)$-dimensional affine subspace, each element of which has a unique representation of the form

$$\sum_{i=1}^{2j+1} \alpha_{i_1}$$

for some j. A set of binary vectors are affine independent iff no subset containing an even number of vectors sums to zero. An affine basis for the set of all 2^k binary k-dimensional vectors may be selected in various ways. The "standard" basis consists of the all-zero vector and each of the k "unit" vectors. In general, any $k + 1$ affine independent vectors may be chosen as a basis.

If k is very large, then the probability that a randomly chosen set of $k + 1$ k-dimensional binary vectors will be affine independent is $\prod_{i=1}^{\infty} (1 - 2^{-i})$, or about 29 percent. The probability that k randomly chosen vectors will be affine independent is about 58 percent; for $k - 1$, it is 76 percent. If $m \ll k$, then almost every set of m different k-dimensional binary vectors is affine independent.

The first row of the $2^k \times 2^k$ Hadamard matrix may be taken as all $+1$'s. The 2^{k-1} $+1$'s in each of the other rows occur in the components corresponding to some $(k - 1)$-dimensional subspace of the k-dimensional binary vectors, and the -1's occur in the components corresponding to the complementary $(k - 1)$-dimensional affine subspace.

The coordinatization induced by the Hadamard matrix is invariant under all changes of affine basis. When translated into coding terminology, this is equivalent to Theorem 15.35 of Berlekamp.[10]

IV. MAIN RESULT AND DISCUSSION

4.1 *Theorem*

Let $\mathbf{v}$ be a 2^k-dimensional vector whose only nonzero components are $2m + 1$ units occurring at components corresponding to k-dimensional binary vectors α_0 , α_1 , α_2 , $\cdots$ α_{2m} which are affine independent. Let the vector $\mathbf{z}$ be defined by the equation

$$\mathbf{z} = \mathcal{H}^t \operatorname{sgn} \mathcal{H}\mathbf{v}.$$

Then the value of z_β , the component of $\mathbf{z}$ corresponding to the k-

dimensional binary vector β, is given by

$$z_\beta = \begin{cases} 0 & \text{if } \beta \text{ is not in the affine subspace spanned by} \\ & \alpha_0 , \alpha_1 , \alpha_2 , \cdots , \alpha_{2m} \\ \dfrac{S 2^{k-2m}(2m-2j)!\,(2j)!}{j!\,m!\,(m-j)!} & \text{if } \beta = \sum_{l=1}^{2j+1} \alpha_{i_l} \end{cases}$$

and the sign is given by

$$S = (-1)^j \prod_{l=1}^{2j+1} v_{\alpha_{i_l}}.$$

4.2 *Remarks*

We first notice that the answers do not significantly depend on k, but on $2m + 1$, the number of units in the input vector $\mathbf{v}$.

Since $\mathbf{z}$ is defined by an equation of the form $\mathbf{z} = \mathfrak{IC}^t\mathbf{y}$, where the energy of $\mathbf{y}$ is n, the energy of $\mathbf{z}$ is $n^2 = 4^k$. For $i = 0, 1, \cdots 2m$ we have

$$| z_{\alpha_i} |^2 = \left[2^{k-2m} \binom{2m}{m} \right]^2 .$$

For large m, $| z_{\alpha_i} |^2$ is closely approximated by $2^{2k}/\pi m$. Thus only $2/\pi$ of the total energy in $\mathbf{z}$ is located in those components in which $\mathbf{v}$ has units; the remaining $(\pi - 2)/\pi$ of the total energy is distributed throughout the affine subspace. For example, if $m = 5$, $k = 10$, we have the following output:

Input Value	Corresponding Output Value	$\left(\text{Output Value}\right)^2$	Number of such Coordinates
1	252	16×3969	$\binom{11}{1} = 11$
0	28	16×49	$\binom{11}{3} + \binom{11}{9} = 220$
0	12	16×9	$\binom{11}{5} + \binom{11}{7} = 792$
0	252	16×3969	$\binom{11}{11} = 1$
Totals		16×2^{16}	1024

For all m, we notice that if $\beta = \sum_{i=0}^{2m} \alpha_i$, then

$$| z_\beta | = | z_{\alpha_i} |.$$

In other words, if the input to the coding-transmission-decoding system consists of $2m + 1$ pulses of equal magnitudes (and arbitrary signs)

located at positions which are affine independent then the output may be written as the sum of the following three terms:

(*i*) $2m + 1$ pulses of equal magnitude (and correct signs) matching the input.
(*ii*) One stray pulse of the same magnitude as the $2m + 1$ correct ones.
(*iii*) Other errors scattered throughout the affine subspace, having maximum amplitude $[1/(2m - 1)]$ times as large as the correct pulses.

Since each of the errors of type *iii* has a relative amplitude approaching zero for sufficiently large m, one might consider the proposed system "successful" in some sufficiently broad sense of that term even though these errors consume $(\pi - 2)/\pi$ of the energy in the output signal. The error of type *ii* poses a different problem, even though it consumes a negligible fraction of the energy. Further research may be required to decide whether these difficulties might be removed by replacing the operator Q by another quantizer with more levels. Further study will also be required to determine how these quantization errors propagate in successive frames in a dynamic system. (See footnote on page 970.)

V. RELATIONSHIP TO THE WEIGHT ENUMERATION PROBLEM FOR RM COSETS

In the previous sections we calculated the vector $\mathbf{z} = \mathcal{H}'Q\mathcal{H}\mathbf{v}$ for certain specific choices of $\mathbf{v}$. These vectors $\mathbf{v}$ were chosen to be "typical" in some intuitive sense, and yet sufficiently simple in form to enable us to carry through the calculation in closed form, even when the dimensions of the $\mathcal{H}$ matrix $(n \times n)$ were large.

Instead of assuming some ad hoc form for the vector $\mathbf{v}$, we might instead ask, what is the range of the operator $\mathcal{H}'Q\mathcal{H}$? Except for the scalar factor, this is equivalent to determining the range of the operator $\mathcal{H}'$ sgn. For, if there exists a vector $\mathbf{x}$ such that $\mathbf{z} = \mathcal{H}'$ sgn $\mathbf{x}$, then n sgn $\mathbf{x}$ = sgn $\mathcal{H}\mathbf{z}$ and $\mathbf{z} = (\mathcal{H}'$ sgn $\mathcal{H}\mathbf{z})/n$. Hence, every vector in the range of $\mathcal{H}'$ sgn is proportional to a vector in the range of $\mathcal{H}'Q\mathcal{H}$, and every vector in the range of $\mathcal{H}'Q\mathcal{H}$ is a stationary point of this operator. Stated another way, $(\mathcal{H}'Q\mathcal{H})^2 = \mathcal{H}'Q\mathcal{H}$.

In more practical terms, an investigation of the vectors in the range of $\mathcal{H}'$ sgn is actually an investigation of the ensemble of possible differential pictures which the proposed system might produce as output. This set is identical to the set of differential pictures which the system will encode and decode with zero error.

If $\mathcal{H}$ is an $n \times n$ Hadamard matrix, then there are 2^n vectors in the range of $\mathcal{H}^t$ sgn. For reasonable values of n, 2^n is so large that a complete listing of all of these vectors is not feasible. Fortunately, however, these 2^n vectors fall into a relatively small number of classes, each class consisting of those vectors which have the same *distribution* of magnitudes of component amplitudes. The problem of determining the possible distributions of magnitudes of the component amplitudes of a vector in the range of $\mathcal{H}^t$ sgn turns out to be identical to the problem of determining the weight enumerators for the cosets of the Reed-Muller code. This equivalence is seen as follows: If $\mathbf{y}$ is a real vector whose components have unit magnitude, then the number of components of $\mathcal{H}^t\mathbf{y}$ with magnitude $|A|$ is the number of rows of $\mathcal{H}$ whose dot product with $\mathbf{y}$ is $\pm A$. On the other hand, if we convert 1 to 0 and -1 to 1, changing $\mathcal{H}$ to $\mathcal{G}$ and $\mathbf{y}$ to $\mathbf{R}$, then the weight of the binary vector sum of $\mathbf{R}$ and each codeword of the extended max-length feedback shift register code gives the distance between the received word $\mathbf{R}$ and the corresponding codeword, and the enumeration of all of these weights for a particular $\mathbf{R}$ is the weight enumerator for the coset containing $\mathbf{R}$. If $\mathbf{c}$ is a binary codeword for which $w(\mathbf{c} + \mathbf{R}) = d$ and if $\mathbf{u}$ is the real vector of ± 1's corresponding to $\mathbf{c}$, then $\mathbf{u}$ and $\mathbf{y}$ disagree in d components and agree in $n - d$ components. Therefore,

$$\mathbf{u} \cdot \mathbf{y} = n - 2d.$$

Since the first order Reed-Muller code contains both the codewords in the extended maximum length feedback shift register code and their complements, there is a one to one correspondence between RM cosets with weight enumerator $d_0, d_1, \cdots, d_n$ and real vectors in the range of $\mathcal{H}^t$ sgn having magnitudes of component amplitudes distributed as follows: d_i components with amplitudes $\pm(n - 2i)$ for $i = 0, 1, 2, \cdots, n/2 - 1$, and $d_{n/2}/2$ components with amplitude zero.

The coset weight enumerators for first order RM codes of lengths up to 16 have been determined by R. Dick and N. J. A. Sloane.[11] Their results, and the corresponding distributions of magnitudes of amplitude components of the output of the differential picture encoding-decoding system are shown in Table I. Those rows which are predicted by our main theorem have been checked, and the relevant values of m have been listed.

The coset weight enumerator for the first order RM code of length 32 was determined by Berlekamp and Welch,[12] and the results are shown in Table II.

The coset weight enumerators for first order RM codes of length

TABLE I—COSET WEIGHT ENUMERATORS FOR FIRST ORDER RM CODES OF LENGTHS 4, 8 AND 16

Number of vectors in range of $\mathfrak{X}'$ Sgn	Number of cosets	Amplitude / Weight									Value of m if predicted by main theorem
		n = 4									
		Amplitude	±4		±2		0				
		Weight	0 & 4		1 & 3		2 (entries halved)				
8	1		1				3				0
8	1				4						1
		n = 8									
		Amplitude	±8		±6		±4		±2	0	
		Weight	0 & 8		1 & 7		2 & 6		3 & 5	4 (entries halved)	
16.1	1		1							7	0
16.8	8				1				7		—
16.7	7						4			4	1
		n = 16									
		Amplitude	±16	±14	±12	±10	±8	±6	±4	±2	0
		Weight	0 & 16	1 & 15	2 & 14	3 & 13	4 & 12	5 & 11	6 & 10	7 & 9	8 (entries halved)
32.1	1		1								15
32.16	16		1							15	
32.120	120				1				7		8
32.560	560					1		3		12	
32.840	840						2		8		6
32.35	35						4				12
32.448	448							6		10	
32.28	28								16		

Value of m if predicted by main theorem
0
1
0
—
1
0
1
2

$\geqq 64$ have not yet been determined. This problem definitely merits further research.

VI. ACKNOWLEDGMENTS

I am indebted to Mr. J. R. Pierce for suggesting this problem and to Mr. N. J. A. Sloane for suggestions which simplified the proof of the identity in Appendix A.

APPENDIX A

A Sketched Proof of Main Result

One of the implications of our main theorem is that if β lies outside of the affine subspace spanned by the relevant α's, then $z_\beta = 0$. A generalization of this result is the following:

Theorem: If the only nonzero components of v all lie in a $(k-1)$ dimensional affine subspace and β lies outside of this subspace, then $z_\beta = 0$.

Proof: Using the elementary properties of RM codes and affine subspaces, the original $2^k \times 2^k$ Hadamard matrix may be partitioned as

$$\mathcal{H} = \begin{bmatrix} \mathcal{G} & \mathcal{G} \\ \mathcal{G} & -\mathcal{G} \end{bmatrix}$$

where $\mathcal{G}$ is a $2^{k-1} \times 2^{k-1}$ Hadamard matrix. In terms of this partition, the last 2^{k-1} components of $\mathbf{v}$ are zero and $\mathcal{H}\mathbf{v}$ is of the form $[\mathbf{w}, \mathbf{w}]^t$ for some appropriate 2^{k-1}-dimensional vector $\mathbf{w}$. We then obtain

$$\mathcal{H}^t \operatorname{sgn} [\mathbf{w}, \mathbf{w}]^t = 2[\mathbf{z}, 0]^t$$

where

$$\mathbf{z}^t = \mathcal{G}^t \operatorname{sgn} \mathbf{w}^t.$$

By repeated application of this theorem, we deduce that $z_\beta = 0$ unless β is in the affine subspace spanned by the coordinates of the nonzero components of $\mathbf{v}$. If the coordinates of the nonzero components of $\mathbf{v}$ span a d-dimensional affine subspace, then an appropriate change of coordinates allows us to work with a $2^d \times 2^d$ Hadamard submatrix, which is also Hadamard. The original output $\mathbf{z}$ vector merely gains a factor of two for each omitted dimension.

Applying these arguments to the main theorem of the text allows us to confine our attention to the case when $k = 2m$.

Since the RM code is invariant under the full affine group, there is

TABLE II—COSET WEIGHT ENUMERATORS FOR FIRST ORDER RM CODES OF LENGTH 32

Boolean function for coset*	Number of such cosets	Weights									Value of m if predicted
Even Cosets		0 32	2 30	4 28	6 26	8 24	10 22	12 20	14 18	16 (halved)	
2345	496 × 1	0	1	0	0	0	0	0	15	16	
2345&12	496 × 120	0	0	0	0	2	2	0	14	14	
2345&23	496 × 35	0	0	0	1	0	3	0	12	16	
2345&23&45	496 × 28	0	0	0	0	0	6	0	10	16	2
2345&12&34	496 × 840	0	0	0	0	0	2	8	14	8	
2345&123	17360 × 2	0	0	1	0	0	0	3	16	12	
2345&123&12	17360 × 24	0	0	0	1	0	1	4	14	12	
2345&123&24	17360 × 18	0	0	0	0	2	0	4	16	10	
2345&123&14	17360 × 192	0	0	0	0	1	2	4	14	11	
2345&123&45	17360 × 32	0	0	0	0	0	4	4	12	12	
2345&123&12&34	17360 × 72	0	0	0	0	0	4	4	12	12	
2345&123&14&35	17360 × 576	0	0	0	0	0	2	8	14	8	
2345&123&12&45	17360 × 96	0	0	0	0	1	0	8	16	7	
2345&123&24&35	17360 × 12	0	0	0	0	0	0	12	16	4	
2345&123&145	13888 × 320	0	0	0	0	1	1	6	15	9	
2345&123&145&45	13888 × 32	0	0	0	1	0	0	6	15	10	
2345&123&145&24&45	13888 × 480	0	0	0	0	0	3	6	13	10	
2345&123&145&35&24	13888 × 192	0	0	0	0	0	1	10	15	6	
123	155 × 8	0	0	1	0	0	0	7	0	24	
123&45	155 × 512	0	0	0	0	0	4	0	28	0	
123&14	155 × 168	0	0	0	0	2	0	8	0	22	
123&14&25	155 × 336	0	0	0	0	0	0	16	0	16	
123&145	868 × 32	0	0	0	1	0	1	0	30	0	
123&145&23	868 × 320	0	0	0	0	1	0	12	0	19	
123&145&24	868 × 480	0	0	0	0	0	4	0	28	0	
123&145&23&24&35	868 × 192	0	0	0	0	0	0	16	0	16	
12	1 × 155	0	0	0	0	4	0	0	0	28	1
12&34	1 × 868	0	0	0	0	0	0	16	0	16	
—	1 × 1	1	0	0	0	0	0	0	0	31	0

* These functions are written in an abbreviated notation. For example, the second line, 2345&12 indicates that this equivalence class of cosets includes one coset whose members are the 64 Boolean functions of the form $X_2X_3X_4X_5 + X_1X_2 + AX_1 + BX_2 + CX_3 + DX_4 + EX_5 + F$, where A, B, C, D, E, and F are arbitrary binary elements.

no loss of generality in assuming that $\alpha_0 = 0$, that $\alpha_1, \alpha_2, \cdots, \alpha_{2m}$ are unit vectors, and that

$$v_{\alpha_l} = +1 \quad \text{for} \quad l = 0, 1, \cdots, 2m.$$

Any other case can be reduced to this case by an appropriate affine transformation of coordinates.

We now determine the distribution of the vector

$$\mathbf{x} = \mathcal{K}\mathbf{v}.$$

TABLE II—Cont'd

Odd Cosets†		1 3 5 7 9 11 13 15 31 29 27 25 23 21 19 17	
—	32 × 1	1 0 0 0 0 0 0 31	
12	32 × 155	0 0 0 1 3 0 0 28	
12&34	32 × 868	0 0 0 0 0 6 10 16	
123	4960 × 1	0 1 0 0 0 0 7 24	
123&12	4960 × 7	0 0 1 0 0 3 4 24	
123&14	4960 × 84	0 0 0 1 1 2 6 22	
123&45	4960 × 64	0 0 0 0 3 1 7 21	
123&14&25	4960 × 336	0 0 0 0 0 6 10 16	
123&12&45	4960 × 448	0 0 0 0 1 3 13 15	
123&12&34	4960 × 84	0 0 0 0 2 4 4 22	
123&145	27776 × 10	0 0 0 1 1 0 12 18	
123&145&12	27776 × 6	0 0 1 0 0 1 10 20	
123&145&23	27776 × 80	0 0 0 1 0 3 9 19	
123&145&45&23	27776 × 16	0 0 0 1 0 1 15 15	
123&145&24	27776 × 180	0 0 0 0 2 2 10 18	
123&145&24&23	27776 × 240	0 0 0 0 1 5 7 19	
123&145&35&24	27776 × 240	0 0 0 0 1 3 13 15	
123&145&35&24&23	27776 × 192	0 0 0 0 0 6 10 16	
123&145&45&35&24&23	27776 × 60	0 0 0 0 0 4 16 12	

† All functions representing odd cosets also contain the term 12345, which is not shown in this table.

Since $\mathbf{x}$ is the sum of the all-ones vector (corresponding to the column of $\mathfrak{IC}$ associated with $\alpha_0 = 0$) and $2m$ columns of $\mathfrak{IC}$ which correspond to linearly independent codewords in the RM code, it is readily seen that there are $\binom{2m}{t}$ components of $\mathbf{x}$ which have the value $2m + 1 - 2t$. It follows that $\mathbf{y} = \operatorname{sgn} \mathbf{x}$ has $\sum_{t=0}^{m} \binom{2m}{t}$ $+1$'s and $\sum_{t=m+1}^{2m} \binom{2m}{t}$ -1's.

For convenience, we may partition the components of $\mathbf{y}$ into $2m + 1$ subsets, each of which corresponds to the components of $\mathbf{x}$ with the same value. We call the set which consists of $\binom{2m}{m}$ components where $\mathbf{x}$ had value $+1$ the "significant" set of components. The sets of insignificant components may be matched up in pairs; the set where $\mathbf{x}$ had value $2m + 1 - 2t$ being matched with the set where $\mathbf{x}$ had value $2m + 1 + 2t$. Each of these matched sets contains $\binom{2m}{t}$ components.

We now let β be a typical $2m$-dimensional binary vector, which is the sum of $2j+1$ α's. By an appropriate permutation of basis vectors, we may assume that

$$\beta = \alpha_0 + \sum_{i=1}^{2j} \alpha_i = \sum_{i=1}^{2j} \alpha_i .$$

The equation $\mathbf{z} = \mathfrak{IC}'\mathbf{y}$ now allows us to compute z_β as the dot product of a particular row of $\mathfrak{IC}'$ and $\mathbf{y}$. The equation $\beta = \sum_{i=1}^{2j} \alpha_i$ and the correspondence to RM codes allows us to express the particular row

of $\mathcal{H}^t$ under consideration as the componentwise product of the first $2j$ rows of $\mathcal{H}^t$, ignoring the zeroth row, which is all plus one. Symbolically, if we let $\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_{2m}$ denote the first $2m$ rows of $\mathcal{H}^t$, then

$$z_\beta = (\mathbf{r}_1 \otimes \mathbf{r}_2 \otimes \cdots \otimes \mathbf{r}_{2j}) \cdot \operatorname{sgn}\left(\sum_{i=1}^{2m} \mathbf{r}_i\right)$$

where "$\otimes$" denotes the componentwise product. Since $2j$ is even, we may also write

$$z_\beta = [(-\mathbf{r}_1) \otimes (-\mathbf{r}_2) \otimes \cdots \otimes (-\mathbf{r}_{2j})] \cdot \operatorname{sgn}\left(\sum_{i=1}^{2m} \mathbf{r}_i\right).$$

The dot product is the sum over all 2^{2m} components of the componentwise product of $(\mathbf{r}_1 \otimes \mathbf{r}_2 \otimes \cdots \otimes \mathbf{r}_{2j})$ and $\mathbf{y} = \operatorname{sgn}(\sum_{i=1}^{2m} \mathbf{r}_i)$. Since there is cancellation of the summands coming from matched sets of components into which we partitioned $\mathbf{y}$, we need only consider the $\binom{2m}{m}$ "significant" components. On each of these, $\mathbf{y}$ takes value $+1$, and the problem reduces to the following: Given a $2m \times \binom{2m}{m}$ matrix, whose columns represent all ways of distributing m plus ones and m minus ones among $2m$ rows, compute the sum of the entries in the componentwise product of the first $2j$ rows of this matrix. The solution is obtained by noting that if there are i minus ones in the first $2j$ rows, then the componentwise product is $(-1)^i$ and this happens in $\binom{2j}{i}\binom{2m-2j}{m-i}$ columns. Therefore,

$$z_\beta = \sum_i (-1)^i \binom{2j}{i}\binom{2m - 2j}{m - i}.$$

Having already explained the other factors in the more general version of the theorem stated in the text, the theorem is reduced to the identity,

$$\sum_i (-1)^i \binom{2j}{i}\binom{2m - 2j}{m - i} \stackrel{?}{=} \frac{(-1)^j (2m - 2j)!\,(2j)!}{j!\,m!\,(m - j)!}.$$

Multiplying through by $(m!)^2/(2m - 2j)!(2j)!$ reduces this to the equivalent identity,

$$\sum_i (-1)^i \binom{m}{i}\binom{m}{2j - i} \stackrel{?}{=} (-1)^j \binom{m}{j},$$

whose proof is given by Riordan (p. 14, line 7 from bottom).[13] Q.E.D.

APPENDIX B

An Example

Suppose that the differential picture consists of a 4×4 grid, the points of which are lettered as follows:

A	B	C	D
E	F	G	H
I	J	K	L
M	N	O	P

The signs of the units in the 16 × 16 cyclic Hadamard matrix with which the differential picture is smeared may be taken as:

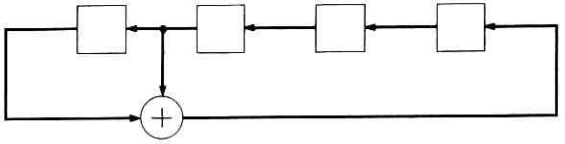
$\mathfrak{H} =$

+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
+	+	+	+	−	+	+	−	−	+	−	+	−	−	−	−
+	+	+	−	+	+	−	−	+	−	+	−	−	−	−	+
+	+	−	+	+	−	−	+	−	+	−	−	−	−	+	+
+	−	+	+	−	−	+	−	+	−	−	−	−	+	+	+
+	+	+	−	−	+	−	+	−	−	−	−	+	+	+	−
+	+	−	−	+	−	+	−	−	−	−	+	+	+	−	+
+	−	−	+	−	+	−	−	−	−	+	+	+	−	+	+
+	−	+	−	+	−	−	−	−	+	+	+	−	+	+	−
+	+	−	+	−	−	−	−	+	+	+	−	+	+	−	−
+	−	+	−	−	−	−	+	+	+	−	+	+	−	−	+
+	+	−	−	−	−	+	+	+	−	+	+	−	−	+	−
+	−	−	−	−	+	+	+	−	+	+	−	−	+	−	+
+	−	−	−	+	+	+	−	+	+	−	−	+	−	+	−
+	−	−	+	+	+	−	+	+	−	−	+	−	+	−	−
+	−	+	+	+	−	+	+	−	−	+	−	+	−	−	−

The induced coordinatization may be read from the 2nd, 3rd, 4th and 5th rows. It is:

Grid Point	Coordinatization	Representation as Sum of Odd Number of Affine Basis Vectors
A	0 0 0 0	H + I + L
B	0 0 0 1	E + H + I + L + P
C	0 0 1 0	H + L + P
D	0 1 0 0	E + I + L
E	1 0 0 1	E
F	0 0 1 1	E + H + L
G	0 1 1 0	E + L + P
H	1 1 0 1	H
I	1 0 1 0	I
J	0 1 0 1	I + L + P
K	1 0 1 1	E + I + P
L	0 1 1 1	L
M	1 1 1 1	H + I + P
N	1 1 1 0	E + H + I
O	1 1 0 0	E + H + P
P	1 0 0 0	P

The coordinates of B through P may also be taken as the successive contents of this shift register:



Now suppose the differential picture is this:

0	0	0	0
+1	0	0	−1
+1	0	0	−1
0	0	0	−1

The coordinates of the nonzero inputs are as follows:

$$\begin{array}{ll} E & 1\ 0\ 0\ 1 \\ H & 1\ 1\ 0\ 1 \\ I & 1\ 0\ 1\ 0 \\ L & 0\ 1\ 1\ 1 \\ \underline{P} & \underline{1\ 0\ 0\ 0} \\ & 0\ 0\ 0\ 1 = \text{sum} \end{array}$$

These are affine independent, so our main theorem applies. The stray output pulse will be located at point B, since this is the point whose coordinates are the vector sum of the coordinates of the other inputs.

We now calculate the output vector step by step, without using the theorem. The input picture is:

$$\mathbf{v} = [0, 0, 0, 0, +1, 0, 0, -1, +1, 0, 0, -1, 0, 0, 0, -1].$$

The "smeared picture" is $\mathbf{x} = \mathfrak{C}\mathbf{v}$, which is given by the sum of the following rows:

$$\begin{array}{cccccccccccccccc} + & - & + & + & - & - & + & - & + & - & - & - & - & + & + & + \\ - & + & + & - & + & - & + & + & + & + & - & - & - & + & - & - \\ + & - & + & - & + & - & - & - & - & + & + & + & - & + & + & - \\ - & - & + & + & + & + & - & - & - & + & - & - & + & + & - & + \\ - & + & - & - & - & + & - & - & + & + & - & + & - & + & + & + \end{array}$$

$$\mathbf{x} = -1,-1,+3,-1,+1,-1,-1,-3,+1,+3,-3,-1,-3,+5,+1,+1$$

The quantized smeared picture is

$$\mathbf{y} = Q\mathbf{x}$$
$$= [-\ -\ +\ -\ +\ -\ -\ -\ +\ +\ -\ -\ -\ +\ +\ +].$$

The decoded differential picture is given by

$$\mathbf{z} = \mathfrak{C}\mathbf{y}$$
$$= [-2, -6, +2, +2, +6, -2, -2, -6, +6, -2, +2, -6, -2, +2, -2, -6].$$

When scaled down by a factor of six and placed on the grid, the output is

$-\frac{1}{3}$	-1	$+\frac{1}{3}$	$+\frac{1}{3}$
$+1$	$-\frac{1}{3}$	$-\frac{1}{3}$	-1
$+1$	$-\frac{1}{3}$	$+\frac{1}{3}$	-1
$-\frac{1}{3}$	$+\frac{1}{3}$	$-\frac{1}{3}$	-1

In order to match the total power of the input, the output must be scaled down slightly more.

REFERENCES

1. Schroeder, M. R., unpublished work.
2. Pratt, W. K., Kane, J., and Andrews, H. C., "Hadamard Transform Image Coding," Proc. IEEE, *57*, No. 1 (January 1969), pp. 58–68.
3. Hall, M., Jr., *Combinatorial Theory*, Waltham, Massachusetts: Blaisdell Publishing Co., 1967.
4. Spence, E., "A New Class of Hadamard Matrices," Glasgow Math. J., *8*, Part I (January 1967), pp. 59–62.
5. Goethals and Seidel, J. J., "Orthogonal Matrices With Zero Diagonal," Canadian J. Math., *19*, No. 5 (1967), pp. 1001–1010.
6. Wallis, J., "A Class of Hadamard Matrices," J. Combinatorial Theory, *6*, No. 1 (January 1969), pp. 40–44.
7. Wallis, J., "Note of a Class of Hadamard Matrices," J. Combinatorial Theory, *6*, No. 2, (March 1969), pp. 222–223.
8. Thoene, R., and Golomb, S. W., "Search for Cyclic Hadamard Matrices," JPL Space Programs Summary, *4*, No. 37–40 (1966), pp. 207–208.
9. Baumert, L. D., "Cyclic Hadamard Matrices," JPL Space Programs Summary, *4*, No. 37-40 (1966), pp. 311–314.
10. Berlekamp, E. R., *Algebraic Coding Theory*, New York: McGraw-Hill, 1968.
11. Dick, R., and Sloane, N. J. A., unpublished work.
12. Berlekamp, E. R., and Welch, L. R., unpublished work.
13. Riordan, J., *Combinatorial Identities*, New York: Wiley, 1968.

Binary Codes Which Are Ideals in the Group Algebra of an Abelian Group

By Mrs. F. J. MacWILLIAMS

(Manuscript received January 13, 1970)

A cyclic code is an ideal in the group algebra of a special kind of Abelian group, namely a cyclic group. Many properties of cyclic codes are special cases of properties of ideals in an Abelian group algebra.

A character of an Abelian group G of order v is, for our purposes, a homomorphism of G into the group of vth roots of unity over $GF(2)$. If G is cyclic with generator x, the character is entirely determined by what it does to x; this effect is kept, and the characters are discarded. If G is not cyclic it is necessary to rehabilitate the characters. Without them the notation is impossible; with them one can prove a number of theorems which reduce in the special case to well-known properties of cyclic codes. Moreover the writer thinks that the general proof is often easier and more suggestive than the proof for the special case. To support this point of view we produce a new theorem, which of course also applies to cyclic codes.

I. INTRODUCTION

A cyclic code is an ideal in the group algebra of a special kind of Abelian group, namely a cyclic group. Many properties of cyclic codes are special cases of properties of ideals in an Abelian group algebra.

A character of an Abelian group G of order v is, for our purposes, a homomorphism of G into the group of vth roots of unity over $GF(2)$. If G is cyclic with generator x, the character is entirely determined by what it does to x; this effect is kept, and the characters are discarded. If G is not cyclic, it is necessary to rehabilitate the characters. Without them the notation is impossible; with them one can prove a number of theorems which reduce in the special case to well-known properties of cyclic codes. Moreover the writer thinks that the general proof is often easier and more suggestive than the proof for the special case. To support this point of view we produce a new theorem, which of course also applies to cyclic codes.

The plan of this paper is as follows: Section II contains a summary of the properties of ideals in an Abelian group algebra. Section III contains a description of the group characters; the reader is assured (and we hope reassured) that an effort has been made to point out the analogies with the cyclic case. In Section IV the characters are extended to the group algebra. This section contains the general cases of several familiar theorems, for example, the dimension of the code, a lower bound on its minimum distance, the Mattson–Solomon mapping, and the identification of the dual code. In Section V the structure of product codes is examined for the general case. Section VI contains the new theorem (which needs too much notation to be explained here) and the special case of this theorem which applies to cyclic codes. The Appendix contains an illustrative example of the smallest possible nontrivial case.

II. GENERAL PROPERTIES OF ABELIAN GROUP ALGEBRAS

Let G be a finite Abelian group of odd order v; the group operation is written as multiplication.

Let $R = FG$ be the group algebra of G over the field $F = GF(2)$. R consists of finite sums

$$A = \sum_{g \in G} a_g g, \qquad a_g \in F.$$

In FG we have two operations, addition and multiplication, defined as follows:

$$A + B = \sum_{g \in G} (a_g + b_g) g,$$

and for $f \in G$,

$$fA = \sum_{g \in G} a_g fg = \sum_{g \in G} a_{f^{-1}g} g.$$

This implies

$$AB = \sum_{h \in G} \sum_{gf=h} a_g b_f h. \tag{1}$$

We use 1 to denote the unit of G, and **1**, **0** to denote the unit and zero of FG.

From the first of these operations we see that FG has the structure of a vector space F^v of dimension v over F. **0** is the zero vector and **1** is the vector $(1\ 0\ 0 \cdots 0)$.

An ideal $\mathfrak{A}$ in FG is defined as follows

$$\mathfrak{A} \text{ is a linear subspace of } F^v,$$

$$A \in \mathfrak{A} \Rightarrow gA \in \mathfrak{A} \quad \text{for all} \quad g \in G.$$

From the general theory of semi-simple group algebras,[1] we know that FG is a principal ideal ring; that is, every ideal is of the form

$$\mathfrak{A} = \{rA, r \in FG\} \quad \text{for some element} \quad A \in FG.$$

We denote the ideal with generator A by $\langle A \rangle$. In fact every ideal has an idempotent generator; $\mathfrak{A} = \langle N \rangle$, where $N = \sum_{g \in G} \eta_g g$ has the properties:

$$N^2 = N,$$

$$r \in \mathfrak{A} \Leftrightarrow rN = r. \tag{2}$$

Since the ground field is $GF(2)$, and G is commutative

$$N^2 = \sum_{g \in G} \eta_g g^2,$$

so that

$$N = \sum_{g \in G} \eta_g g$$

is idempotent if and only if $\eta_g = \eta_{g^2}$ for all $g \in G$.

FG is the direct sum of its minimal ideals,

$$FG = \langle \theta_1 \rangle + \cdots + \langle \theta_t \rangle,$$

and every ideal in FG is the direct sum of a subset of these minimal ideals.[1] The idempotents, θ_i, of the minimal ideals are called primitive idempotents and have the additional properties

$$\sum_{i=1}^{t} \theta_i = \mathbf{1}, \tag{3}$$

$$\theta_i \theta_j = \mathbf{0}, \qquad i \neq j, \tag{4}$$

$$\langle \theta_i \rangle \cap \langle \theta_j \rangle = \mathbf{0}, \qquad i \neq j.$$

Every idempotent in FG is the sum of primitive idempotents. Since we are over $GF(2)$ the sum of idempotents is idempotent, and the set of all idempotents is a vector space I; $\theta_1, \cdots, \theta_t$ are a set of linearly independent basis elements for I, which is thus of dimension t.

We also define a set of "trivial" idempotents as follows:

Let $y_1 = 1 \in G$. Pick $g \in G$, $g \neq 1$, and set

$$y_2 = \{g, g^2, g^4, \cdots, g^{2^s}\}$$

where $g^{2^{s+1}} = g$ (this must happen since G is finite and of odd order). Pick $f \notin y_1 \cup y_2$ and define the set $y_3 = \{f, f^2, f^4, \cdots, f^{2^t}\}$. In this way G

is partitioned into disjoint classes, which we call cycles

$$G = y_1 \cup y_2 \cup y_3 \cdots . \tag{5}$$

Define $Y_i \in FG$ by

$$Y_i = \sum_{\sigma \in y_i} g, \qquad \text{(for example, } Y_2 = g + g^2 + \cdots + g^{2^s}\text{)}.$$

The Y_i are the trivial idempotents. From equation (2) it is clear that every idempotent is the sum of trivial idempotents, and they are obviously linearly independent over F. Hence the trivial idempotents also form a basis for I over F. We have proved the following Lemma:

Lemma 1.1: The number of trivial idempotents is the same as the number of primitive idempotents, and each set is linearly dependent on the other; that is, there exists an invertible $t \times t$ matrix (m_{ij}) over F such that

$$\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_t \end{bmatrix} = (m_{ij}) \begin{bmatrix} Y_1 \\ \vdots \\ Y_t \end{bmatrix}.$$

From a practical point of view it is desirable to find the θ_i. The algorithm for doing this is as described in Ref. 2, except that the group is no longer cyclic. Briefly, we form linear combination of the Y_i in a systematic way until we find t idempotents which satisfy equations (3) and (4). An example is given in Appendix A.

III. GROUP CHARACTERS

Since we shall make extensive use of the characters of the group and the group algebra, we give a brief account of their properties.

For our purposes, a character of G is a homomorphism ψ of G into the vth roots of unity over $GF(2)$. These vth roots of unity lie in an extension field $GF(2^s)$ in which the expression $z^v - 1$ splits into linear factors. They form a cyclic subgroup of the (multiplicative) group of non-zero elements of this field.

Formally

$$\psi(f)\psi(g) = \psi(fg). \tag{6}$$

Hence

$$\psi(1) = 1$$

(the unit of G on the left and of $GF(2^s)$ on the right) and

$$\psi(g^{-1}) = [\psi(g)]^{-1}.$$

If G is a cyclic group of order v, with generator x, a character is a map $x \to \beta$, where β is a vth root of unity. In this case one usually does not distinguish between the character and the value it assigns to x. We define multiplication of characters by

$$(\phi\psi)(g) = \phi(g)\psi(g).$$

Under this operation, the characters form a group, $\mathfrak{X}$. The unit of $\mathfrak{X}$, called the principal character ψ_1, is the map

$$g \to 1 \quad \text{for all} \quad g \in G.$$

The group G and the character group $\mathfrak{X}$ are isomorphic in many ways. We construct a particular isomorphism and use it henceforth.

Theorem 2.1: (Reference 3) The Abelian group G has a unique decomposition as the direct product of cyclic groups of prime power order,

$$G = G_1 \times G_2 \times \cdots \times G_s, \qquad G_i \text{ cyclic of order } p_i^{a_i}.$$

(The primes p_i are not necessarily district.)

Pick a generator x_i for G_i, and a fixed primitive p^{a_i}th root of unity, α_i. Let ψ_{x_i} be the character defined on the generators by

$$\psi_{x_i}(x_i) = \alpha_i, \qquad \psi_{x_i}(x_j) = 1, \qquad i \neq j.$$

By equation (6) this is sufficient to define ψ_{x_i} on any $g \in G$. We may by equation (7) define $\psi_{x_i}^2$

$$\psi_{x_i}^2(g) = [\psi_{x_i}(g)]^2.$$

Lemma 2.2: If φ is any character of G, then ϕ can be represented in the form

$$\varphi = \prod_{i=1}^{s} \psi_{x_i}^{a_i}.$$

Proof: Let $\varphi(x_i) = \beta$. Then

$$\beta^{p_i^{a_i}} = \varphi(x_i)^{p_i^{a_i}} = \varphi(x_i^{p_i^{a_i}}) = \varphi(1) = 1.$$

Thus β is a power of α_i, say $\beta = \alpha_i^{a_i}$. We then see that

$$\varphi(\prod_i x_i^{b_i}) = \prod_i \varphi(x_i^{b_i}) = \prod \alpha_i^{a_i b_i}.$$

Hence

$$\varphi = \prod_i \psi_{x_i}^{a_i}.$$

Set $a = \prod_i x_i^{a_i}$ and denote the character $\varphi = \prod_i \psi_{x_i}^{a_i}$ by φ_a. We then

have

Lemma 2.3: The mapping $a \Leftrightarrow \varphi_a$ *as defined above is an isomorphism between* G *and* $\mathfrak{X}$.

We also use ψ_a to mean the character corresponding to a in this isomorphism.

Lemma 2.4: $\varphi_a(b) = \varphi_b(a)$, *and* $\varphi_{a^{-1}}(b) = \varphi_a(b^{-1})$.

Proof: Let

$$a = x_1^{a_1} \cdots x_s^{a_s}, \qquad b = x_1^{b_1} \cdots x_s^{b_s}.$$

Then

$$\varphi_a(b) = \prod_i [\varphi_{x_i}(b)]^{a_i} = \prod_i \prod_j [\varphi_{x_i}(x_j^{b_j})]^{a_i},$$

$$= \prod \alpha_i^{a_i b_i} = \varphi_b(a).$$

The second statement is proved in a similar way.

We shall need the following theorem which is well known, so the proof is omitted. The skeptical reader may easily construct an elementary proof by using the properties of the roots of unity.

Theorem 2.5:

$$(i) \qquad \sum_{\psi \in \mathfrak{X}} \psi(g) = \begin{cases} v & \text{if} \quad g = 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$(ii) \qquad \sum_{g \in G} \psi(g) = \begin{cases} v & \text{if} \quad \psi = \psi_1, \\ 0 & \text{otherwise.} \end{cases}$$

If G is cyclic, both parts of this theorem reduce to

$$\sum_{i=0}^{v-1} \beta^i = \begin{cases} v & \text{if} \quad \beta = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathfrak{X}(g)$ be a matrix whose columns are labeled by the characters ψ and rows by the group elements g. The entry in row ψ, column g is $\psi(g)$. An example is given in Appendix A.

Lemma 2.6: $\mathfrak{X}(g)\mathfrak{X}^T(g^{-1}) = \text{diagonal}\,[v\,v \cdots v] = vI$. *Hence* $\mathfrak{X}(g)$ *is invertible.*

Proof: A typical entry on the main diagonal is

$$\sum_{g \in G} \psi(g)\psi(g^{-1}) = \sum_{g \in G} \psi(1) = v, \quad \text{by Theorem 2.5 } (ii).$$

A typical off-diagonal entry is

$$\sum_{g \in G} \psi_a(g)\psi_b(g^{-1}) = \sum_{g \in G} \psi_g(a)\psi_g(b^{-1}),$$
$$= \sum_{g \in G} \psi_g(ab^{-1}) = 0, \text{ by Theorem 2.5 } (ii) \text{ since } a \neq b.$$

IV. CHARACTERS OF THE GROUP ALGEBRA

If G is a cyclic group of order v, with generator x, and $A = A(x)$ an element of FG (that is, a polynomial of degree less than v in x) then $A(\beta)$ is the value of the character $x \to \beta$ on A. In the general case, for $A = \sum_{g \in G} a_g g$ in FG, we extend the character to the elements of the group algebra by

$$\psi(A) = \sum_{g \in G} a_g \psi(g).$$

Using the notation of Theorem 2.1 and those that follow, we could write an element of FG as a sum of terms of the form $x_1^{j_1} x_2^{j_2} \cdots x_s^{j_s}$, $0 \leqq j_i < p^{s_i}$. A is a polynomial in the variables $x_1, \cdots, x_s$ with restrictions on the degree of each variable. A character is a mapping $A(x_1 \cdots x_s) \to A(\beta_1, \cdots, \beta_s)$ where β_i is a (p^{s_i})th root of unity. As pointed out in the introduction, there are certain advantages to using this polynomial notation as little as possible.

If G is cyclic, we know that

$$A(x)B(x) \mid_{x=\beta} = AB(x) \mid_{x=\beta}.$$

Analogously for the general case (and with the same proof, using equation (1)),

$$\psi(AB) = \psi(A)\psi(B).$$

If G is cyclic, it is usually the case that $A(\beta_1)A(\beta_2) \neq A(\beta_1\beta_2)$. There is however a vital exception, namely $A(\beta)^2 = A(\beta^2)$. Similarly, in the general case

$$\psi\varphi(A) \neq \psi(A)\varphi(A), \qquad \text{but}$$

Lemma 3.1: $\psi(A)^2 = \psi^2(A) = \psi(A^2)$.

Proof:

$$[\psi(A)]^2 = [\sum_{g \in G} a_g \psi(g)]^2 = \sum_{g \in G} a_g \psi(g)^2,$$
$$= \sum_{g \in G} a_g \psi(g^2).$$

A cyclic code is an ideal in a cyclic group algebra. It is frequently

described as the set of polynomials which vanish on a certain prescribed set S of vth roots of unity:

$$\mathfrak{A} = \{A(x): A(\beta) = 0, \beta \in S\}. \tag{7}$$

Similarly, we can characterize an ideal in the group algebra of an Abelian group as the set of elements of FG which vanish at a prescribed set of characters:

$$\mathfrak{A} = \{A \in FG: \psi(A) = 0, \psi \in S\}. \tag{7'}$$

From Lemma 3.1 we see that in the general case, as in the special case, the maximal set $\hat{S}$ corresponding to a particular ideal must have a special form; in fact it is the union of sets $\{\psi, \psi^2, \psi^4, \cdots\}$.

It is well known that the dimension of the cyclic code associated by equation (7) with the set $\hat{S}$ is the number of vth roots of unity not contained in $\hat{S}$, that is, the number of nonzeros of the code. Similarly in the general case. The following two theorems are proved in Reference 4; we repeat the proofs here for convenience, and also supply an example in Appendix A. Let $g_1, g_2, \cdots, g_r$ be the elements of G. Associate with the element A the $v \times v$ matrix $(a_{g_i^{-1}g_j})$. The entry in row i column j is the coefficient of g_j in g_iA. The ideal $\mathfrak{A} = \langle A \rangle$ is generated as a subspace of $R = F^v$ by the rows of the matrix $(a_{g_i^{-1}g_j})$. The dimension of this ideal is the rank of this matrix.

Theorem 3.2: The dimension of the ideal $\langle A \rangle$ is the number of characters ψ such that $\psi(A) \neq 0$.

Proof: The matrix $\mathfrak{X}^T(g^{-1})(a_{g_i^{-1}g_j})\mathfrak{X}(g)$ has the same rank as $(a_{g_i^{-1}g_j})$, since by Lemma 2.6 $\mathfrak{X}(g)$ is invertible. A typical entry of the product $(a_{g_i^{-1}g_j})\mathfrak{X}(g)$ is of the form

$$n_{ij} = \sum_{g_k \in G} a_{g_i^{-1}g_k}\psi_{g_i}(g_k).$$

Now

$$\sum_{g_k \in G} a_{g_i^{-1}g_k}\psi_{g_i}(g_k) = \sum_{g_k \in G} a_{g_k}\psi_{g_i}(g_ig_k) = \psi_{g_i}(g_i)\cdot\psi_{g_i}(A).$$

Thus

$$n_{ij} = \psi_{g_i}(g_i)\psi_{g_i}(A).$$

In the product $\mathfrak{X}(g^{-1})(n_{ij})$ the diagonal terms are of the form

$$\psi_{g_i}(A)\sum_{\psi \in \mathfrak{X}} \psi(g^{-1})\psi(g) = v\psi_{g_i}(A).$$

The off-diagonal terms are of the form

$$\psi_{g_i}(A) \sum_{\psi \in \mathfrak{X}} \psi(g^{-1})\psi(f) = 0.$$

Thus

$$\mathfrak{X}(g^{-1})^T(a_{g_i^{-1}g_j})\mathfrak{X}(g) = \text{diagonal } [\psi_{g_1}(A), \psi_{g_2}(A), \cdots \psi_{g_v}(A)],$$

and the rank of the matrix $(a_{g_i^{-1}g_j})$ is the number of characters for which $\psi(A) \neq 0$.

We call these characters the non-zeros of the ideal $\langle A \rangle$.

Let D be the $m \times v$ submatrix of $\mathfrak{X}(g)^T$ whose columns are indexed by the group elements and rows by the m characters for which $\psi(A) = 0$. If $\mathbf{a} = (a_1, a_2, \cdots, a_v)$ is a vector of $\langle A \rangle$, then $D\mathbf{a}^T = 0$. If D contains no set of t linearly independent columns, the minimum weight in $\langle A \rangle$ is at least $t + 1$. This is the extension of the BCH bound for cyclic codes. It is generally a very weak lower bound.

Theorem 3.3: (The Mattson-Solomon mapping—see Reference 5.)

(i) *If* $A = \sum a_g g$, then $a_f = \frac{1}{v} \sum_{\psi \in \mathfrak{X}} \psi(A)\psi(f^{-1})$.

(ii) *If* $va_g = \sum_{\psi \in \mathfrak{X}} \beta_g \psi(g^{-1})$, then $\psi_h(A) = \beta_h$.

Proof:

(i)
$$\sum_{\psi \in \mathfrak{X}} \psi(A)\psi(f^{-1}) = \sum_{\psi \in \mathfrak{X}} \sum_{g \in G} a_g \psi(g)\psi(f^{-1}),$$
$$= \sum_{g \in G} a_g \sum_{\psi \in \mathfrak{X}} \psi(gf^{-1}) = va_f .$$

(ii)
$$v\psi_h(A) = \sum_{g \in G} a_g \psi_h(g),$$
$$= \sum_{g \in G} \sum_{\psi_f \in \mathfrak{X}} \beta_g \psi_f(g^{-1})\psi_h(g),$$
$$= \sum_{g \in G} \beta_g \sum_{\psi_f \in \mathfrak{X}} \psi_f(g^{-1})\psi_h(g),$$
$$= \sum_{g \in G} \beta_g \sum_{f \in G} \psi_g(f^{-1})\psi_g(h),$$
$$= \sum_{g \in G} \beta_g \sum_{f \in G} \psi_g(f^{-1}h),$$
$$= \beta_h .$$

Corollary 3.4: A is uniquely determined by the set of values $\psi(A)$.

We divide the group G as in equation (5) into cycles corresponding to the trivial idempotents of FG, and divide the character group $\mathfrak{X}$ into similar classes by the isomorphism of Lemma 2.3.

$$G = y_1 \cup y_2 \cup \cdots \cup y_t \,, \tag{5}$$

$$\mathfrak{X} = \Psi_1 \cup \Psi_2 \cup \cdots \cup \Psi_t \,. \tag{8}$$

$y_1 = 1$ and Ψ_1 contains only the principal character ψ_1. By Lemma 3.1 if $\psi(A) \neq 0$ for some $\psi \in \Psi_i$, then $\psi(A) \neq 0$ for all $\psi \in \Psi_i$. The non-zeros of A are a union of cycles Ψ_i.

The minimal ideals have the smallest possible dimension, so that by Theorem 3.2 the non-zeros of a minimal ideal are, if possible, the characters in a single class Ψ_i. (This is in fact possible; an explicit construction is given in Section V.) If θ_i is the idempotent of this minimal ideal we may define θ_i by the property

$$\psi(\theta_i) = \begin{cases} 1 & \psi \in \Psi_i \,, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Theorem 3.5: The dimension of the ideal $\langle \theta_i \rangle$ is $| \Psi_i |$, the number of elements in Ψ_i.

Since every ideal in FG is the direct sum of minimal ideals, every idempotent is of the form $C = \sum_i \epsilon_i \theta_i$, $i = 0$ or 1. The dimension of C is $\sum \epsilon_i | \Psi_i |$. From equations (2) and (4) we have immediately:

Theorem 3.6: If C_1, C_2 are idempotents with non-zeros Φ_1 and Φ_2, and $\Phi_1 \subset \Phi_2$, then $\langle C_1 \rangle$ is a subideal of $\langle C_2 \rangle$.

The dual code of $\langle N \rangle$ is the set of vectors $b_1, \cdots, b_v$ such that $\sum_{i=1}^{v} a_i b_i = 0$ for all vectors $a_1, \cdots, a_v$ in $\langle N \rangle$. The dimension of the dual code is $v - \dim \langle N \rangle$.

If N is idempotent, the dimension of $\langle (1 + N) \rangle$ is $v - \dim \langle N \rangle$. This follows at once from the fact that $1 = \sum_{i=1}^{t} \theta_i$. For $A = \sum a_g g$, set $A^* = \sum a_g g^{-1}$.

Theorem 3.7: The dual code of $\langle N \rangle$ is $\langle (1 + N)^* \rangle$.

Proof: Let $\sum b_{g^{-1}} g \in \langle (1+N)^* \rangle$; then $\sum b_g g \in \langle (1+N) \rangle$. Since $N(1+N) = 0$, for any $\sum a_g g \in \langle N \rangle$ we have

$$(\sum a_g g)(\sum b_g g) = 0.$$

From the coefficient of 1 in this product

$$\sum a_g b_{g-1} = 0.$$

Therefore, $\langle (1 + N)^* \rangle$ is contained in the dual code of $\langle N \rangle$, and has dimension $v - \dim \langle N \rangle$. Thus it is the dual code.

V. QUASI-CYCLIC AND PRODUCT CODES

Let H be a proper subgroup of order u of the Abelian group G; and let

$$G = k_1H \cup k_2H \cup \cdots \cup k_wH \qquad (k_1 = 1,\ v = uw)$$

be the decomposition of G into cosets of H. In this section we suppose the coordinate places in FG to be arranged in the order

$$k_1h_1\ ,\ k_1h_2\ ,\ \cdots\ ,\ k_1h_u\ ,\ k_1h_1\ ,\ \cdots\ ,\ k_2h_u\ ,\ \cdots\ ,\ k_wh_1\ ,\ \cdots\ ,\ k_wh_u\ .$$

Let $\mathfrak{A}$ be an ideal of FG, and denote by $\mathfrak{A}_i$ the part of $\mathfrak{A}$ which lies in the coordinate places $k_ih_1\ ,\ \cdots\ ,\ k_ih_u$. $\mathfrak{A}_1$ is an ideal of FH (usually several repetitions of an ideal of FH), and since $k_i\mathfrak{A}_1 = \mathfrak{A}_i$ the codes $\mathfrak{A}_i$ are all repetitions of $\mathfrak{A}_1$. Each vector of $\mathfrak{A}$ consists of w vectors of $\mathfrak{A}_1$; these are not in general the same vector, and some of them may be zero. If H is a cyclic group, $\mathfrak{A}$ has the structure of a quasi-cyclic code. Since G contains cyclic subgroups of order p for every prime p which divides v, $\mathfrak{A}$ may have this structure in several different ways.

We make the additional assumption that G is the direct product $G = H \times K$ of subgroups H, K. This means that $H \cap K = 1$, and each element of G can be expressed uniquely as $g = kh$, $k \in K$, $h \in H$. The character group $\mathfrak{X}$ is correspondingly a direct product

$$\mathfrak{X} = \mathfrak{X}_H \times \mathfrak{X}_K\ ,$$

where $\mathfrak{X}_H$, $\mathfrak{X}_K$ are the images of H, K under the isomorphism of Lemma 2.3. Every character can be expressed uniquely as

$$\psi = \varphi_H\varphi_K\ , \qquad \varphi_H \in \mathfrak{X}_H\ , \qquad \varphi_K \in \mathfrak{X}_K\ .$$

We shall need the following result.

Lemma 4.1: $\varphi_H\varphi_K(hk) = \varphi_H(h)\varphi_K(k)$.

Proof: From the isomorphism of Lemma 2.3,

$$\varphi_H(k) = 1, \qquad \varphi_K(h) = 1.$$

Let $A = \sum_{h\in H} a_hh$, $B = \sum_{k\in K} b_kk$ be idempotents in the group algebras FH, FK. Let Φ_H , Φ_K be the non-zeros of A, B respectively. Φ_H , Φ_K correspond to cycles of $\mathfrak{X}_H$, $\mathfrak{X}_K$ which are, of course, also cycles of $\mathfrak{X}$.

The Kronecker product of matrices, M, N, is denoted by $M \times N$ (an example is given in Appendix A).

Theorem 4.2: (*i*) $C = AB$ *is an idempotent of* FG.

(*ii*) *The codes* $\langle C \rangle$ *is the direct product of codes* $\langle A \rangle$, $\langle B \rangle$.
(*iii*) *The non-zeros of* $\langle C \rangle$ *are* $\varphi_H \varphi_K$, $\varphi_H \in \Phi_H$, $\varphi_K \in \Phi_K$.
(*iv*) *The minimum distance of* $\langle C \rangle$ *is the product of those of* $\langle A \rangle$, $\langle B \rangle$.

Proof: (*i*) It is clear that $C = \sum_{k \in K} b_k k \sum_{h \in H} a_h h$ is idempotent.

(*ii*) The first row of the Kronecker product

$$(a) \times (b) = (a_{h_q^{-1} h_r}) \times (b_{k_s^{-1} k_t})$$

consists of the coefficients of C. The second row contains the coefficients of $h_1 C$, and the $(u + 1)$st row the coefficients of $k_2 C$. Without further notation, we see that the rows of this Kronecker product generate the code $\langle C \rangle$ as a subspace of F^v.

(*iii*) $\mathfrak{X}(G) = \mathfrak{X}_H(h) \times \mathfrak{X}_K(k)$. By Theorem 3.2, the non-zeros of C are given by

$$[\mathfrak{X}_H(h^{-1}) \times \mathfrak{X}_K(k^{-1})]^T [(a) \times (b)][\mathfrak{X}_H(H) \times \mathfrak{X}_K(K)]$$
$$= \mathfrak{X}_H^T(h^{-1})(a)\mathfrak{X}_H(H) \times \mathfrak{X}_K^T(k^{-1})(b)\mathfrak{X}_K(K).$$

The triple matrix products are diagonal matrices with ones in the places corresponding to $\varphi \in \Phi_H$ ($\varphi \in \Phi_K$) and zeros elsewhere. Their Kronecker product is a diagonal matrix with ones in the places corresponding to $\varphi_H \varphi_K$, $\varphi_H \in \Phi_H$, $\varphi_K \in \Phi_K$.

(*iv*) This is a well-known property of direct product codes.

Given an idempotent C of FG we would like to know how, if possible, to find subgroups H, K such that $G = H \times K$, and $C = AB$. The following theorem is sometimes helpful.

Theorem 4.3: Let Ψ *be the set of non-zeros of* C; *suppose* Ψ *can be expressed as the product of two sets of cycles* Φ_1 , Φ_2 *where* $\Phi_1 \in \mathfrak{X}_H$, $\Phi_2 \in \mathfrak{X}_K$ *and* $\mathfrak{X} = \mathfrak{X}_H \times \mathfrak{X}_K$. (*Consequently,* $G = H \times K$.)

Then $C = AB$, *where* A, B *are idempotents in* FH *and* FK, *with non-zeros* Φ_1 , Φ_2 ; *consequently the code* $\langle C \rangle$ *is the direct product of codes* $\langle A \rangle$ *and* $\langle B \rangle$.

Proof:

$$C = \sum_{kh \in G} a_{kh} kh = k_1 \sum_{h \in H} a_{k_1 h} h + k_2 \sum_{h \in H} a_{k_2 h} h + \cdots + k_w \sum_{h \in H} a_k f_h h.$$

By Theorem 3.3 (*i*)

$$a_{k_i h} = \sum_{\psi \in \mathfrak{X}} \psi(C) \psi(k_i^{-1} h^{-1}) = \sum_{\varphi_1 \in \Phi_1} \sum_{\varphi_2 \in \Phi_2} \varphi_1 \varphi_2 (k_i h^{-1}),$$

since by hypothesis

$$\psi(C) = \begin{cases} 1 & \text{if } \psi = \varphi_1\varphi_2 \\ 0 & \text{otherwise.} \end{cases}$$

$$= \sum_{\varphi_2 \in \Phi_2} \varphi_2(k_i^{-1}) \sum_{\varphi_1 \in \Phi_1} \varphi_1(h^{-1})$$

by Lemma 4.1. Set $a_h = \sum_{\psi \in \mathfrak{X}H} \alpha_h \psi(h^{-1})$, where

$$\alpha_h = \begin{cases} 1 & \psi \in \Phi_1 \\ 0 & \text{otherwise.} \end{cases}$$

Set $A = \sum_{h \in H} a_h h$; then

$$\psi(A) = \alpha_h = \begin{cases} 1 & \psi \in \Phi_1 \\ 0 & \text{otherwise} \end{cases}$$

by Lemma 3.5 (*ii*). Define B similarly for K. Then A, B are idempotents in FH, FK, and $C = AB$.

If H, K are cyclic groups whose orders are relatively prime, then G is also cyclic. The codes $\langle A \rangle$, $\langle B \rangle$ are cyclic codes in FH, FK respectively, and $\langle C \rangle$ is a cyclic code in FG.

This special case has been thoroughly investigated by Burton and Weldon[6] and Goethals.[7]

The extension to direct products of more than two subgroups is theoretically obvious, but rather hard to visualize. An example for the cyclic case is given in Appendix 2.

VI. A NEW THEOREM

Everything in this paper so far is a natural extension of known results about cyclic codes. This section is not; the special case of Theorem 5 for G cyclic is new and interesting (at least the writer thinks so).

The primitive idempotents θ_i of FG have been defined by the property

$$\psi(\theta_i) = \begin{cases} 1 & \psi \in \Psi_i , \\ 0 & \psi \notin \Psi_i . \end{cases} \tag{10}$$

We recall that the trivial idempotents are defined by the property

$$Y_i = \sum_{g \in G} a_g g, \; a_g = \begin{cases} 1 & g \in y_i , \\ 0 & g \notin y_i . \end{cases} \tag{11}$$

Since these properties look remarkably symmetrical, one expects to

find some symmetry in the matrix (m_{ij}) (Lemma 1.1) which relates θ_i to Y_i. This in fact exists, as follows.

We recall that

$$A^* = \sum_{g \in G} a_g g^{-1}.$$

Theorem 5.1:

$$\theta_i = \sum_{k=1}^{t} r_k Y_k \leftrightarrow Y_i^* = \sum_{k=1}^{t} r_k \theta_k .$$

Proof: Let

$$\theta_i = \sum_{g \in G} b_g g$$

$$= \sum_{g \in G} \sum_{\psi \in \mathfrak{A}} \psi(\theta_i)\psi(g^{-1})g \quad \text{by Lemma 3.3 } i$$

(Note that $1/v = 1$ in characteristic 2.)

$$= \sum_{g \in G} \Big(\sum_{\psi \in \Psi_i} \psi(g^{-1}) \Big) g \quad \text{by (10)}.$$

From definition (8) of Ψ_i, we may suppose that $\Psi_i = \{\psi_f, \psi_{f^q}, \cdots, \psi_{f^{q^{s-1}}}\}$. Then the inner sum is

$$\psi_f(g^{-1}) + \psi_{f^q}(g^{-1}) + \cdots + \psi_{f^{q^{s-1}}}(g^{-1})$$

$$= \psi_g(f^{-1}) + \psi_g(f^{-2}) + \cdots + \psi_g(f^{-2^s}) \quad \text{by Lemma 2.4,}$$

$$= \psi_g(Y_i^*).$$

Thus

$$\theta_i = \sum_{g \in G} \psi_g(Y_i^*) g. \tag{12}$$

(This is the explicit construction for θ_i.) Now suppose

$$Y_i^* = \sum_{k=1}^{t} r_k \theta_k , \qquad r_k \in GF(2).$$

$$\psi_g(Y_i^*) = \sum_{k=1}^{t} r_k \psi_g(\theta_k),$$

$$\psi_g(\theta_k) = \begin{cases} 1, & g \in Y_k , \\ 0, & g \notin Y_k . \end{cases} \quad \text{from (9)}.$$

Hence $\psi_{g_1}(Y_i^*) = r_1$, $\psi_g(Y_i^*) = r_k$ for all $g \in Y_k$. Substituting in equation (11), we obtain

$$\theta_i = r_1 + r_2 \sum_{g \in Y_2} g + \cdots + r_t \sum_{g \in Y_t} g,$$

$$= \sum_{i=1}^{t} r_i Y_i .$$

Let $\psi_k(Y_i)$ be the common value of $\psi(Y_i)$ for $\psi \in \Psi_k$. Equation (12) then becomes

$$\begin{aligned} \theta_i &= \sum_{k=1}^{t} \psi_k(Y_i^*) Y_k , \\ &= \sum_{k=1}^{t} r_k Y_k . \end{aligned} \tag{13}$$

With a slight change of notation, let

$$\theta_i = \sum_{k=1}^{t} m_{ik} Y_k .$$

Let P be a permutation matrix such that P acting on the column vector $(Y_1, Y_2, \cdots, Y_k)^T$ produces $(Y_1^*, Y_2^*, \cdots, Y_t^*)^T$.

Theorem 5.2:

$$(m_{ij})^2 = P.$$

Proof: By Theorem 5.1,

$$\begin{aligned} Y_i^* &= \sum_{k=1}^{t} m_{ik}\theta_k = \sum_{k=1}^{t} m_{ik} \sum_{j=1}^{t} m_{kj} Y_j , \\ &= \sum_{j=1}^{t} \left(\sum_{k=1}^{t} m_{ik} m_{kj} \right) Y_j . \end{aligned}$$

Hence

$$\sum_{k=1}^{t} m_{ik} m_{kj} = \begin{cases} 1 & Y_i^* = Y_j , \\ 0 & \text{otherwise.} \end{cases}$$

We give a brief description of the special case G cyclic of prime order p. FG is now the polynomial ring $R = F[x]/x^p + 1$. Let f be the order of 2 mod p. If $p - 1 = ef$, then

$$2 = g^e$$

for some generator g of the integers mod p. The trivial idempotents, other than 1, are of the form

$$x^i + x^{i \cdot 2} + x^{i \cdot 4} + \cdots + x^{i \cdot 2^{f-1}} .$$

Let σ be the automorphism of R induced by $x \to x^q$; define

$$X_0 = x + x^2 + x^4 + \cdots + x^{2^{f-1}}, \quad X_i = X_{i-1}\sigma, \quad i = 1, \cdots, e - 1.$$

Then

$$X_i = x^a + x^{a\cdot 2} + \cdots + x^{a\cdot 2^{f-1}}, \qquad a = g^i.$$

Since the trivial idempotents were previously called $Y_1, \cdots, Y_t$ we have changed notation; now

$$Y_1 = 1, \qquad Y_2 = X_0, \cdots, Y_t = X_{e-1}.$$

We rename the primitive idempotents correspondingly,

$$\theta_1 = J, \qquad \theta_2 = \eta_0, \cdots, \theta_t = \eta_{e-1}.$$

The characters of G are defined by

$$\psi_{x^k}(x) = \alpha^k,$$

where α is a primitive pth root of unity; thus η_i is defined by

$$\psi_{x^k}(\eta_i) = \begin{cases} 1 & \text{if } k = g^{es+i}, \\ 0 & \text{otherwise.} \end{cases}$$

This may be rewritten as

$$\eta_i(\alpha^{g^{es+k}}) = \begin{cases} 1 & i = k, \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

In particular

$$J(\alpha^i) = \begin{cases} 1 & \alpha^i = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Hence

$$J = \sum_{i=0}^{p-1} x^i.$$

Write

$$\eta_i = m_i + \sum_{k=0}^{e-1} m_{ik} X_k ;$$

$$m_i = X_i^*(J) = f,$$

$$m_{ik} = X_i^*(\alpha^{es+k}).$$

Since $-1 = g^{ef/2}$ we have

$$X_i^* = \begin{cases} X_i , & f \text{ even}, \\ X_{i+e/2} & f \text{ odd}. \end{cases} \tag{15}$$

Now

$$\eta_i\sigma = f + \sum_{k=0}^{e-1} m_{ik}X_{k+1} ,$$

and

$$m_{ik} = X_i^*(\alpha^{g^{es+k}}) = X_{i-1}^*(\alpha^{g^{es+k+1}}),$$

by equation (14) and the definition of X_i. Hence

$$m_{ik} = m_{i-1,k+1} ,$$

and

$$\eta_i\sigma = f + \sum_{j=0}^{e-1} m_{i-1,j}X_j = \eta_{i-1} .$$

Set

$$\theta_0 = f + \sum_{k=0}^{e-1} m_kX_k ;$$

then

$$\theta_1 = \theta_0\sigma^{e-1} = f + \sum_{k=0}^{e-1} m_{k+1}X_k .$$

Clearly

$$J = 1 + \sum_{k=0}^{e-1} X_k .$$

The matrix corresponding to the (m_{ij}) of Theorem 5.2 is of the form

$$\begin{pmatrix} 1 & \mathbf{J} \\ \mathbf{f}^T & M \end{pmatrix},$$

where $\mathbf{J}$, $\mathbf{f}$ are now vectors of length e and

$$M = \begin{pmatrix} m_0 , & m_1 , & \cdots , & m_{e-1} \\ m_1 , & m_2 , & \cdots , & m_0 \\ m_{e-1} , & m_0 , & \cdots , & m_{e-2} \end{pmatrix}.$$

Let P be the permutation matrix which turns the column vector $(1, X_0, \cdots, X_{e-1})^T$ into $(1, X_0^*, \cdots, X_{e-1}^*)^T$. By equation (14)

$$P = I \quad \text{for} \quad f \text{ even},$$

$$P = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0}^T & Q^{e/2} \end{bmatrix} \quad \text{for} \quad f \quad \text{odd}$$

where

$$Q = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & \cdots & \cdots & 0 \end{bmatrix} \quad \text{of size} \quad (e-1) \times (e-1).$$

From Theorem 5.2 we have

$$\begin{bmatrix} 1 + ef, & \mathbf{J} + \mathbf{1}M \\ \mathbf{f}^T + M\mathbf{f}^T & \mathbf{f}^T\mathbf{1} + M^2 \end{bmatrix} = P.$$

For f even, $\mathbf{f}^T\mathbf{1}$ is an $e \times e$ matrix of zeros, hence

$$M^2 = I \quad \text{if } f \text{ is even;}$$

for f odd, $\mathbf{f}^T\mathbf{1}$ is an $e \times e$ matrix of ones, which we denote by K.

$$M^2 = K + Q^{e/2} \quad \text{if } f \text{ is odd.}$$

The matrix M, which is symmetric and circulant in the wrong direction, can be made circulant in the usual way by multiplication by a suitable permutation matrix. Skipping the obvious details we have the following theorem.

Theorem 5.3: *With η_i, X_i defined as above, and*

$$\eta_0 = m_\infty + \sum_{i=0}^{e-1} m_k X_k ,$$

(*i*)
$$X_0^* = m_\infty + \sum_{i=0}^{e-1} m_k \eta_k .$$

(*ii*) *Set*

$$m(y) = m_0 + m_1 y + m_2 y^2 + \cdots + m_{e-1} y^{e-1},$$

$$m(y)^T = m_0 + m_{e-1} y + m_{e-2} y^2 + \cdots + m_1 y^{e-1}.$$

Then

(*i*)
$$m_\infty = \begin{cases} 0 & f \text{ even}, \\ 1 & f \text{ odd}. \end{cases}$$

(*ii*)
$$m(y)m(y)^T = 1 \mod y^e + 1, \quad f \text{ even},$$
$$= \sum_{i=0}^{e-1} y^i + y^{e/2}, \quad f \text{ odd}.$$

Theorem 5.3 has several interesting corollaries of which we mention one.

Let w be the weight (the number of non-zero coordinates) of $m(y)$. The following statements come from Theorem 5.3.

The weight of X_i = The dimension of $\langle \eta_i \rangle = f$.

The weight of η_i = The dimension of $\langle X_i \rangle = wf$, $f = 0(2)$, $wf + 1$, $f = 1(2)$.

Corollary 5.4: *If* $p = 2^k - 1$, $\langle X_i \rangle$ *is a* $(2^k - 1, 2^{k-1})$ code, with minimum weight $\leqq k$.

Proof: For $p = 2^k - 1$, we have $f = k$. Clearly the minimum weight in $\langle X_i \rangle$ is bounded above by that of X_i, which is k.

The minimal ideal $\langle \eta_i \rangle$ is the dual of a Hamming code. Hence η_i (and every other non-zero code word) has weight $(p + 1)/2 = ef/2 + 1$.

Thus $w = e/2$, and the dimension of $\langle X_i \rangle$ is $(p + 1)/2$.

We can use Theorem 5.3 to discover some other remarkably poor cyclic codes; for example

$$p = 251, e = 16, f = 16, w = 9,$$
$$p = 1801, e = 72, f = 25, w = 39.$$

[After the completion of this paper, the writer discovered that Abelian Group Codes have also been investigated by Berman (KIBERNETIKA, vol. 3, no. 3, 1967) and by Paul Camion (to appear).]

VII. CONCLUSION

The writer regretfully admits that she has made no attempt whatsoever to find out whether general Abelian group codes are of any practical value. One obvious thing to do is to make a computer search; the algorithm for finding the primitive idempotents is quite easy to implement. Another direction of research is to look for a class of groups, not cyclic, which produce codes with some desirable practical properties.

VIII. ACKNOWLEDGMENTS

The writer is grateful to her colleagues, especially N. J. A. Sloane, for several excellent suggestions which greatly increased the clarity of this paper.

APPENDIX A

An Example of a Non-Cyclic Abelian Group

Let G be the group of order 9 which is the direct product of two groups of order 3. The elements of G are

$$1, x, x^2, y, xy, x^2y, y^2, xy^2, x^2y^2, \qquad x^3 = y^3 = 1.$$

Let α be a primitive third root of unity over $GF(2)$; then

$$1 + \alpha + \alpha^2 = 0.$$

The matrix $\mathfrak{X}(g)$ is:

	ψ_1	ψ_x	ψ_{x^2}	ψ_y	ψ_{xy}	ψ_{x^2y}	ψ_{y^2}	ψ_{xy^2}	$\psi_{x^2y^2}$
1	1	1	1	1	1	1	1	1	1
x	1	α	α^2	1	α	α^2	1	α	α^2
x^2	1	α^2	α	1	α^2	α	1	α^2	α
y	1	1	1	α	α	α	α^2	α^2	α^2
xy	1	α	α^2	α	α^2	1	α^2	1	α
x^2y	1	α^2	α	α	1	α^2	α^2	α	1
y^2	1	1	1	α^2	α^2	α^2	α	α	α
xy^2	1	α	α^2	α^2	1	α	α	α^2	1
x^2y^2	1	α^2	α	α^2	α	1	α	1	α^2

It is symmetric because the characters are written in the same order as the group elements to which they correspond; the argument does not use the symmetry of $\mathfrak{X}(g)$.

The trivial idempotents are

$$Y_1 = 1;\ Y_2 = x + x^2;\ Y_3 = y + y^2;\ Y_4 = xy + x^2y^2;\ Y_5 = x^2y + xy^2.$$

In order to find the primitive idempotents we need the multiplication table for the Y_i. This also is symmetric and we write only half of it.

	Y_2	Y_3	Y_4	Y_5
Y_2	Y_2	—	—	—
Y_3	$Y_4 + Y_5$	Y_3	—	—
Y_4	$Y_3 + Y_5$	$Y_2 + Y_5$	Y_4	—
Y_5	$Y_3 + Y_4$	$Y_2 + Y_4$	$Y_2 + Y_3$	Y_5

We have then

$$1 = Y_2 + (1 + Y_2); \qquad Y_3 = Y_3Y_2 + Y_3(1 + Y_2);$$
$$(1 + Y_3) = (1 + Y_3)Y_2 + (1 + Y_3)(1 + Y_2).$$

Thus

$$Y_3 = (Y_4 + Y_5) + (Y_3 + Y_4 + Y_5),$$
$$1 + Y_3 = (Y_2 + Y_4 + Y_5) + (1 + Y_2 + Y_3 + Y_4 + Y_5).$$
$$1 = Y_3 + (1 + Y_3) = (Y_4 + Y_5) + (Y_3 + Y_4 + Y_5)$$
$$+ (Y_2 + Y_4 + Y_5) + (1 + Y_2 + Y_3 + Y_4 + Y_5).$$

We multiply this equation by Y_4 and $(1 + Y_4)$:

$$Y_4 = (Y_2 + Y_3 + Y_4) + (Y_3 + Y_4 + Y_5) + (Y_2 + Y_4 + Y_5) + 0,$$
$$1 + Y_4 = (Y_2 + Y_3 + Y_5) + 0 + 0 + (1 + Y_2 + Y_3 + Y_4 + Y_5).$$

Finally,

$$1 = Y_4 + (1 + Y_4) = (Y_2 + Y_3 + Y_4) + (Y_3 + Y_4 + Y_5)$$
$$+ (Y_2 + Y_4 + Y_5) + (Y_2 + Y_3 + Y_5) + (1 + Y_2 + Y_3 + Y_4 + Y_5).$$

This is a decomposition of 1 into five mutually orthogonal idempotents, which are therefore the primitive idempotents. Set

$$A = Y_2 + Y_3 + Y_4 = x + x^2 + y + y^2 + xy + x^2y^2.$$

We use the table $\mathfrak{X}(g)$ to check that

$$\psi_x(A) = \psi_{x^3}(A) = \psi_y(A) = \psi_{y^3}(A) = \psi_{x^2y}(A) = \psi_{xy^2}(A) = 0$$
$$\psi_{xy}(A) = \psi_{x^2y^2}(A) = 1.$$

Hence

$$Y_2 + Y_3 + Y_4 = \theta_4.$$

Similarly

$$Y_3 + Y_4 + Y_5 = \theta_3 ,$$
$$Y_2 + Y_4 + Y_5 = \theta_2 ,$$
$$Y_2 + Y_3 + Y_5 = \theta_5 ,$$
$$1 + Y_2 + Y_3 + Y_4 + Y_5 = \theta_1 .$$

The matrix $(a_{g_i^{-1}g_j})$ for the trivial idempotent Y_2 is

	1	x	x^2	y	xy	xy^2	y^2	xy^2	x^2y^2
1	0	1	1	0	0	0	0	0	0
x	1	0	1	0	0	0	0	0	0
x^2	1	1	0	0	0	0	0	0	0
y	0	0	0	0	1	1	0	0	0
xy	0	0	0	1	0	1	0	0	0
xy^2	0	0	0	1	1	0	0	0	0
y^2	0	0	0	0	0	0	0	1	1
xy^2	0	0	0	0	0	0	1	0	1
x^2y^2	0	0	0	0	0	0	1	1	0

To save space we write this as the Kronecker product

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} b & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & b \end{bmatrix}, \qquad b = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

and also write $\mathfrak{X}(g)$ as the Kronecker product.

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & \alpha & \alpha^2 \\ 1 & \alpha^2 & \alpha \end{bmatrix} \times \begin{bmatrix} 1 & 1 & 1 \\ 1 & \alpha & \alpha^2 \\ 1 & \alpha^2 & \alpha \end{bmatrix} = \begin{bmatrix} a & a & a \\ a & \alpha a & \alpha^2 a \\ a & \alpha^2 a & \alpha a \end{bmatrix}, \qquad a = \begin{bmatrix} 1 & 1 & 1 \\ 1 & \alpha & \alpha^2 \\ 1 & \alpha^2 & \alpha \end{bmatrix}.$$

$$\mathfrak{X}(g^{-1}) = \begin{bmatrix} a' & a' & a' \\ a' & \alpha^2 a' & \alpha a' \\ a' & \alpha a' & \alpha^2 a' \end{bmatrix}$$

where

$$a' = \begin{pmatrix} 1 & 1 & 1 \\ 1 & \alpha^2 & \alpha \\ 1 & \alpha & \alpha^2 \end{pmatrix}.$$

It is then easy to calculate that

$$\mathfrak{X}(g^{-1})(a_{v_i^{-1}v_i})\mathfrak{X}(g) = \begin{pmatrix} aba' & 0 & 0 \\ 0 & aba' & 0 \\ 0 & 0 & aba' \end{pmatrix}$$

where

$$aba' = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Theorem 3.4 then says that the non-zeros of Y_2 are

$$\psi_x\,,\ (\psi_x)^2,\ \psi_{xv}\,,\ (\psi_{xv})^2,\ \psi_{x^2v}\,,\ (\psi_{x^2v})^2$$

which is obvious from the array $\mathfrak{X}(g)$.

This is also an illustration, though a rather trivial one, of Theorem 4.2. H is the group $(1, x, x^2)$; K is the group $(1, y, y^2)$. A is the ideal $x + x^2$ in FH, and B the ideal 1 in FK. The non-zeros of A are ψ_x , ψ_x^2 and the non-zeros of B are ψ_1 , ψ_v , ψ_v^2 . Clearly $Y_2 = AB$, and the non-zeros of Y_2 are the products $\psi_H\psi_K$, as above.

We can also check Theorems 5.1 and 5.2 from the following table:

$$\begin{aligned}
\theta_1 &= Y_1 + Y_2 + Y_3 + Y_4 + Y_5\,, \\
\theta_2 &= \quad Y_2 \quad\quad + Y_4 + Y_5\,, \\
\theta_3 &= \quad\quad\ Y_3 + Y_4 + Y_5\,, \\
\theta_4 &= \quad Y_2 + Y_3 + Y_4\,, \\
\theta_5 &= \quad Y_2 + Y_3 \quad\quad + Y_5\,.
\end{aligned}$$

It is clear that

$$\begin{aligned}
Y_1 &= \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5\,, \\
Y_2 &= \quad\ \theta_2 \quad\quad + \theta_4 + \theta_5\,, \quad \text{and so on;}
\end{aligned}$$

and

$$(m_{ij})^2 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{bmatrix}^2 = I.$$

APPENDIX B

An Example of the Product of Three Cyclic Codes

Let H, K, L be cyclic groups of orders 3, 5, 7 respectively. Their direct product G is cyclic of order 105. (Unfortunately, this is the smallest possible example.)

Write

$$H = 1, x, x^2; \quad K = 1, y, y^2, y^3, y^4; \quad L = 1, z, z^2, z^3, z^4, z^5, z^6.$$

Let $\langle A_1 \rangle$ (3, 2) and $\langle A_2 \rangle$ (5, 4) be the single parity check codes in FH, FK, with idempotents

$$A_1 = x + x^2; \qquad A_2 = y + y^2 + y^3 + y^4.$$

Let $\langle A_3 \rangle$ (7, 4) be the Hamming code in FL, with idempotent

$$A_3 = 1 + z + z^2 + z^4.$$

The direct product code has idempotent $C = A_1A_2A_3$. $\langle C \rangle$ is a (105, 32) cyclic code, with minimum distance 12. Each vector of C can be represented as a three-dimensional array of ones and zeros, which are situated at the lattice points corresponding to $x^iy^jz^k$ in Fig. 1. (The origin is $x^0y^0z^0$.) The lines of this array which are parallel to the x-axis are vectors of $\langle A_1 \rangle$; those parallel to the y-axis belong to $\langle A_2 \rangle$, and those parallel to the z-axis to $\langle A_3 \rangle$.

It has been suggested (see Ref. 8) that an array like this be used for simultaneous burst and random error correction. It must however be borne in mind that such a code will be highly redundant.

To express C as a cyclic code we write the lattice points in order $1, \mu, \mu^2, \mu^3, \cdots \mu^{104}$, where μ is a generator of the cyclic group G, for example $\mu = xyz$. With this choice $x^iy^jz^k$ becomes μ^n where n is the least integer such that

$$n - i \equiv 0(3); \qquad n - j \equiv 0(5); \qquad n - k \equiv 0(7)$$

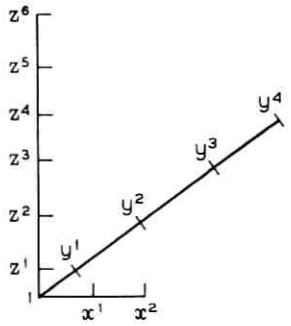


Fig. 1—$x^i y^j z^k$.

for example:

$$x^2y^3z^4 = (xyz)^{53}.$$

REFERENCES

1. Curtis, C. W., and Reiner, I., *Representation Theory of Finite Groups and Associative Algebras*, New York: John Wiley, 1962.
2. MacWilliams, Jessie, "The Structure and Properties of Binary Cyclic Alphabets," B.S.T.J., *44*, No. 2 (February 1965), pp. 303–332.
3. Speiser, Andreas, *Die Theorie der Gruppen von Endlicher Ordnung*, New York: Dover, 1945, Chapter 3.
4. MacWilliams, Jessie, and Mann, H. B., "On the p-Rank of the Design Matrix of a Difference Set," Information and Control, *12*, No. 5–6 (May–June 1968), pp. 474–488.
5. Mattson, M. F., and Solomon, G., "A New Treatment of Bose-Chaudhuri Codes," J. SIAM, *9*, No. 4 (December 1961), pp. 654–669.
6. Burton, H. O., and Weldon, E. J., Jr., "Cylic Product Codes," IEEE Trans. Information Theory, *IT-11*, No. 3 (July 1965), pp. 443–440.
7. Goethals, Jean-Marie, "Factorization of Cylic Codes," IEEE Trans. Information Theory, *IT-13*, No. 2 (April 1967), pp. 242–246.
8. Bridwell, J. D., "Burst Distance and Multiple Burst Correction," B.S.T.J., *49*, No. 5 (May–June 1970), pp. 889–909.

Delta Modulation Codec for Telephone Transmission and Switching Applications

By R. R. LAANE and B. T. MURPHY

(Manuscript received January 12, 1970)

A highly integrable delta modulation codec design, for applications where transmission bandwidth is not at a premium but where an inexpensive and high quality converter is desired, is considered in this paper. An asymmetrical codec integrator is used to improve quantizing noise characteristics. Charge parceling techniques which are used for performing the integrating function, offer advantage over conventional RC integrators.

I. INTRODUCTION

Delta modulation[1] (ΔM) is receiving interest for voiceband analog-to-digital (A/D) conversion applications where coding efficiency is less important than the requirement for economical, but high quality A/D conversion, characteristics. One of the potential applications is in pulse code modulation (PCM) coding systems where the PCM code is formed by first converting the analog signals into a single bit digital code using per terminal ΔM coders. The ΔM bit stream is then converted into a PCM format using digital filtering.[2,3] The technique takes advantage of the simple means of providing A/D conversion with ΔM and utilizes highly integrable digital hardware for providing the ΔM to PCM conversion.

Another promising application of ΔM is in space division switching networks. Analog inputs to the network are converted into a digital code by per terminal ΔM coders. This allows implementation of digital switching networks which are more ideally suited to integrated semiconductor technology than analog networks. Requirements on network loss, signal distortion and crosstalk are significantly relieved.

We describe the design of a single integration ΔM codec (*co*der-*dec*oder) which shows promise of meeting the conversion requirements for both of the above applications. Improved conversion characteristics

are achieved using planned integrator asymmetry. For accurate coding characteristics at high clock (sampling) rates, charge parceling techniques are used in the integrator for reconstructing the anlaog signals. A highly integrable design which offers economical, per line A/D conversion is achieved.

II. DELTA MODULATION CODING REQUIREMENTS

2.1 *Codec Operation*

A block diagram of a delta modulation codec, using single integration, is shown in Fig. 1. To perform the analog-to-digital conversion, an analog input is compared with a reconstructed version of itself from the coder integrator. The relative difference between the signals is translated into a single bit digital code by clocking the output of the comparator stage. The code is then transmitted to both coder and decoder integrators where it forces either a small positive or negative voltage change in the integrator output. Thus, a single bit code is used for controlling the integrator voltage and causes the integrator to produce a close track of the input signal. If a matched integrator is placed at the decoder, a similar track of the analog signal is recovered.

2.2 *Overload Characteristics*

Because the transmitted digital code contains information corresponding to the derivative of the message function, overload characteristics with delta modulation become a function of signal slope instead of amplitude. The overload point occurs when the integrator is forced to produce a similar polarity voltage step during each clock cycle. Thus,

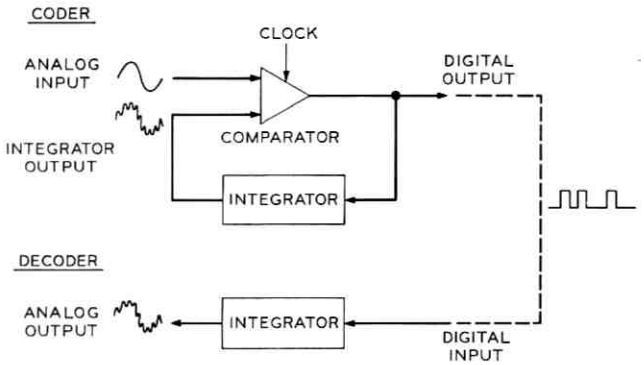


Fig. 1—Delta modulation codec (*co*der-*dec*oder).

for a sampling frequency, f_s, and an integrator step, σ, the maximum integrator voltage slope is given by

$$\sigma f_s .$$

The maximum slope of a sine wave of amplitude, A, and frequency, f, is

$$2\pi f A,$$

and therefore overload occurs when

$$2\pi f A \geqq \sigma f_s . \tag{1}$$

This sets the product of σ and f_s; their relative values are set by quantizing noise requirements for meeting signal to noise objectives.

2.3 *Quantizing Noise*

Granular quantizing noise has been the subject of numerous papers including works by Van de Weg,[4] Wang,[5] and Iwersen.[6] The theory developed by Iwersen predicts a quantizing noise spectrum which is in good agreement with measured noise in delta modulators.[7] To improve noise characteristics of our delta modulator, his theory is used for optimizing coding characteristics. Some of the basic equations and calculations which govern the design of our codec are reviewed in this section. A detailed description is given in Ref. 6.

The effect of an asymmetrical integrator is shown in Fig. 2, where coding of an idle channel input or a dc signal is illustrated. In this example, the positive integrator step, σ_+, is larger than the negative step, σ_-; that is,

$$\sigma_+ = \sigma + \epsilon,$$

$$\sigma_- = -\sigma + \epsilon.$$

As a result, a sawtooth error wave of peak-to-peak amplitude σ is generated. The noise spectrum resulting from coding a steady-state input with unbalanced integrators consists of frequencies given by

$$f_l = | Q[l(1 - \vartheta)/2]f_s | \tag{2}$$

where

$$Q(\alpha) = \alpha - N(\alpha).$$

$N(\alpha)$ is the integer nearest α, and

$$\vartheta = \epsilon/\sigma.$$

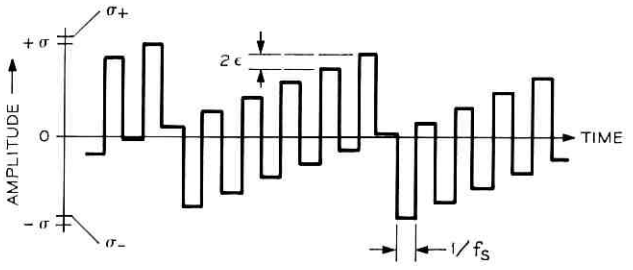


Fig. 2—Integrator output for an asymmetrical coder, shown with $|\sigma_+| > |\sigma_-|$.

The power at the frequency of index l is calculated from

$$P_l = 2\sigma^2/\pi^2 l^2. \tag{3}$$

The frequencies for which l is even are components of the sawtooth wave of peak-to-peak amplitude σ and fundamental frequency ϑf_s. Additional, not so evident, sawteeth are also usually present. For example, components of a second order sawtooth are calculated by choosing values of l equal to aN, where a is a positive integer and N is the odd integer nearest $1/\vartheta$. This sawtooth has a peak-to-peak amplitude of $2\sigma/N \approx 2\epsilon$ and a fundamental frequency of $|\ (1 - N\vartheta)f_s/2\ |$.

It is possible to significantly reduce inband quantizing noise for low level inputs by using a planned integrator imbalance.[8,9] However, to guarantee good noise characteristics, the imbalance must be maintained between a lower and an upper limit. The lower limit must allow an imbalance to force the fundamental component of the fundamental sawtooth wave above voiceband. The change (the spreading) of the noise spectral lines resulting from phase modulation of the idle channel spectrum by the input signal must also be considered.

The upper limit on integrator imbalance is set by quantizing noise objectives. It is difficult to guarantee that the fundamental component of the second order sawtooth wave (of peak-to-peak amplitude $\approx 2\epsilon$) will be kept out of voiceband. Therefore, the magnitude of the integrator step imbalance, ϵ, must be maintained at a level where it will not introduce excessive inband noise problems at low or quiescent input levels.

Iwersen has calculated the quantizing noise as a function of signal level for various step imbalances using a 12-millivolt integrator step size and a 1.544-MHz sampling rate. The calculation is made for a broadband input and uses C-message weighting of the noise in the voiceband. Results are plotted in Fig. 3. The advantage of using an

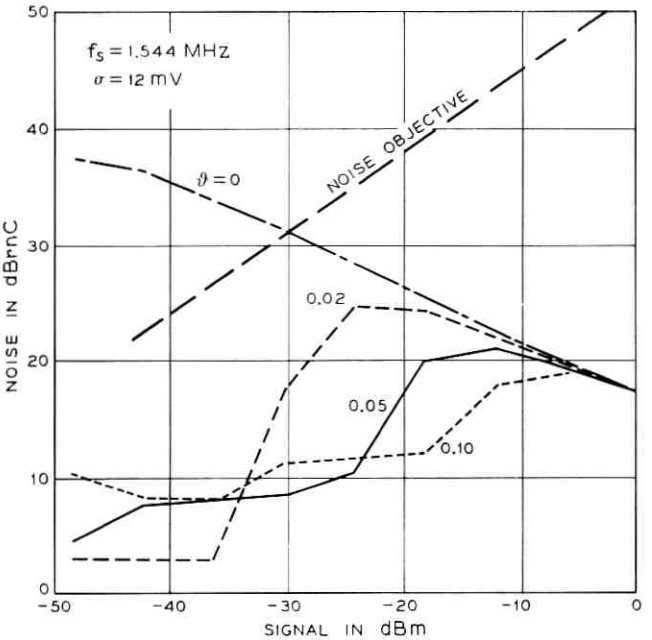


Fig. 3—Calculated quantizing noise versus average speech power for ϑ = 0, 0.02, 0.05, and 0.10.

asymmetrical integrator over a perfectly balanced integrator ($\vartheta = 0$) is clearly evident for low level input signals. An optimal range of integrator imbalance falls into a range from $\vartheta = .02$ to $\vartheta = 0.10$. At higher values of ϑ increased quantizing noise is produced at quiescent levels and at low signal levels. For lower values of ϑ, the fundamental sawtooth frequency is close to voiceband and phase modulation of the spectral lines due to input signals causes excessive quantizing noise energy to fall into voiceband.

2.4 *Transmission Objectives*

Design objectives for this experiment* require that coding a 6-volt peak-to-peak (+7 dBm into 900Ω), 1000-Hz signal will not cause slope overloading. This corresponds to a maximum integrator voltage slope

* The design objectives should not be interpreted as official Bell System transmission objectives. They were established for this experiment as a guide line for providing a reasonable quality of voiceband conversion.

capability of

$$\sigma f_s \approx 19 \times 10^3 \text{ volts/second.}$$

Additional design objectives include a 50-dB dynamic range with a signal-to-noise ratio better than 25 dB at the −43-dBm level, increasing to better than 40 dB at the +7-dBm level. Quantizing noise of idle channels (zero input) should be maintained below 16 dBrn using C-message weighting. Design objectives for gain (loss) variation, over a 200-Hz to 3200-Hz frequency bandwidth, require an average 0.5-dB loss with loss variation maintained within ±0.25 dB.

III. CODEC CIRCUIT DESIGN

The relative simplicity of the delta modulation coding function makes feasible the realization of highly integrable and inexpensive codec configurations. In this design, the codec circuitry utilizes a combination of integrated semiconductor and thin-film capacitor techniques and relies heavily on the characteristics of the two technologies—excellent matching of device characteristics on an integrated circuit (IC) chip and close ratio tolerances between capacitors on thin-film capacitor arrays. These characteristics are of special importance in the key element of the codec, the integrator network, where a high degree of precision is needed for accurately reconstructing an analog signal from a digital input signal.

3.1 *Charge Parceling*

The integrator utilizes a charge parceling circuit (sometimes called the "bucket and dipper" circuit), shown in Fig. 4, to provide the digital-to-analog conversion function. The prime advantage of the charge parceling approach is that it relies only on capacitor ratio tolerances for making the digital-to-analog conversion rather than on absolute resistor and capacitor tolerances required by the more commonly used RC type integrators. Timing problems are also not critical, provided sufficient time is allowed for charging and discharging small charge parceling capacitors in the integrator network.

To add a voltage step to the integrating capacitor, C_I, (Fig. 4) the +1 input (clock input) is applied to produce a positive voltage step of ΔV at the C_{+1} capacitor terminal. This causes an equivalent change in voltage on both of the capacitor terminals until the increase exceeds the threshold of T_2. As T_2 turns on, charge is dumped from C_{+1} to C_I, producing a voltage step on C_I proportional to the two capacitors

$$\Delta V_I = \left[\frac{C_{+1}}{C_I + C_{+1}}\right]\Delta V' \tag{4}$$

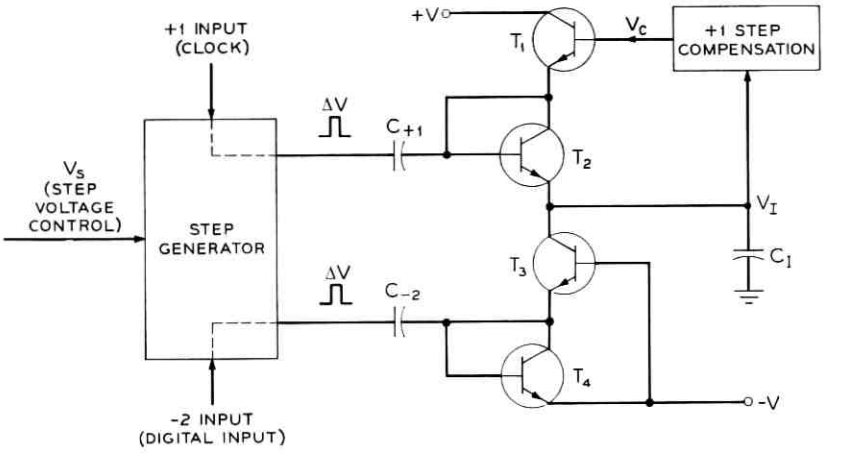


Fig. 4—Charge parceling circuit for ΔM integrator.

where ΔV_I is the voltage change on C_I and $\Delta V'$ is the voltage change at the C_{+1} terminal after T_2 begins to conduct.

As the input returns to its original level, a negative voltage step of ΔV is forced on the C_{+1} terminals. This change causes T_2 to become reverse biased and turns T_1 on to recharge the C_{+1} capacitor. The recharge voltage for C_{+1} is governed by the compensation network which translates a voltage V_c, approximately equal to the integrating capacitor voltage V_I, to the base of T_1. Thus, C_{+1} is recharged to V_c minus the base to emitter drop of T_1, and the net increase in integrator voltage due to a +1 input is

$$\sigma_{+1} = \Delta V_I = [\Delta V - V_{BE_{T_1}} - V_{BE_{T_2}} - (V_I - V_c)] \left(\frac{C_{+1}}{C_I + C_{+1}}\right). \tag{5}$$

Note that the effect of the junction capacitances of T_1 and T_2 also must be considered in the step size calculation, but the effect is secondary and is not described in detail here. Equations including junction capacitances are given in the appendix.

Changes due to temperature in the threshold voltages of $V_{BE_{T_1}}$ and $V_{BE_{T_2}}$ are compensated by controlling the amplitude of the voltage step (ΔV) with a dc voltage V_s and two matching base to emitter (diode) voltages V_D. Therefore, the effective +1 integrator step can be rep-

resented as

$$\sigma_{+1} = [V_s + 2V_D - 2V_{BE} - (V_I - V_c)]\left(\frac{C_{+1}}{C_I + C_{+1}}\right). \quad (6)$$

To produce a negative step to the integrator the -2 input (digital input) is activated. This causes a voltage change on C_{-2} similar to the change produced on C_{+1} by the $+1$ input. However, charge is removed from C_I during the negative slope of ΔV and is dumped from C_{-2} to the $-V$ supply during the positive slope of ΔV. The effective voltage step produced at the integrating capacitor by this input is given by

$$\sigma_{-2} = [V_s + 2V_D - 2V_{BE} - (V_I - V_c)]\left(\frac{\alpha_F C_{-2}}{C_I + \alpha_F C_{-2}}\right) \quad (7)$$

where α_F is the common base current gain of T_3. To minimize the effect of α_F variation, a Darlington transistor pair is planned as a replacement for T_3 in future models.

In the delta modulation codec, a clock controls the $+1$ step input to the integrator, and consequently a σ_{+1} step is added to the integrator during each clock cycle. The digital input to the integrator controls the -2 step input. When present, it decreases the integrator voltage by σ_{-2}. Therefore, whenever a digital input is applied, the net change in the integrator voltage is $\sigma_{+1} - \sigma_{-2}$. When the digital input is not applied, the clock automatically raises the integrator by a σ_{+1} step.

To optimize quantizing noise characteristics, the integrator asymmetry is designed for

$$0.02 < \vartheta < 0.10$$

or, translated to the integrator step requirements (for the case $\sigma_+ > \sigma_-$)

$$\frac{\sigma_{-2}}{\sigma_{+1}} = 1.89 \pm 0.07.$$

An additional feature of the charge parceling integrator is that gain (or loss) between coder and decoder integrators can be easily adjusted. Gain can be adjusted either by changing the step generator voltage, V_s, between integrators or by using a different ratio of integrating to charge parceling capacitors on the coder and decoder integrators. The first technique might be useful as a form of automatic gain control which can be adjusted as a function of voltage. The second technique would be useful when a predetermined amount of gain (or loss) is desired.

3.2 *Compensation Network*

To prevent the decoder output from drifting to either a maximum positive or negative output voltage as a result of differences in the plus

and minus steps between the coder and decoder integrators, the decoder integrator must compensate to automatically adjust its step imbalance to match the coder integrator step imbalance. Compensation is not needed in the coder integrator but is added to help match coding characteristics between coder and decoder.

With the charge parceling circuits described, compensation is achieved by adjusting σ_{+1} of the decoder integrator as a function of the integrator voltage level. The step size is varied by adjusting the recharge voltage for the C_{+1} capacitor.

Figure 5 shows a schematic of the compensation network. A nearly linear compensation as a function of integrator output voltage is achieved by this configuration. Good reproducibility of compensation characteristics between integrators is also possible. Operation is as follows.

With no current through R_2, a voltage equivalent to V_I is translated to the base of T_6 for recharging the +1 step capacitor, C_{+1}. As V_I increases above this level, V_C also increases but begins to lag farther and farther behind V_I, because R_1 can no longer supply all of the current required by the current source I_2. The balance of the current is supplied through R_2, and the voltage drop across R_2 determines the difference between V_C and V_I.

The opposite happens as the integrator voltage drops below the bias point set by R_1 and I_2. Then R_1 will supply more current than accept-

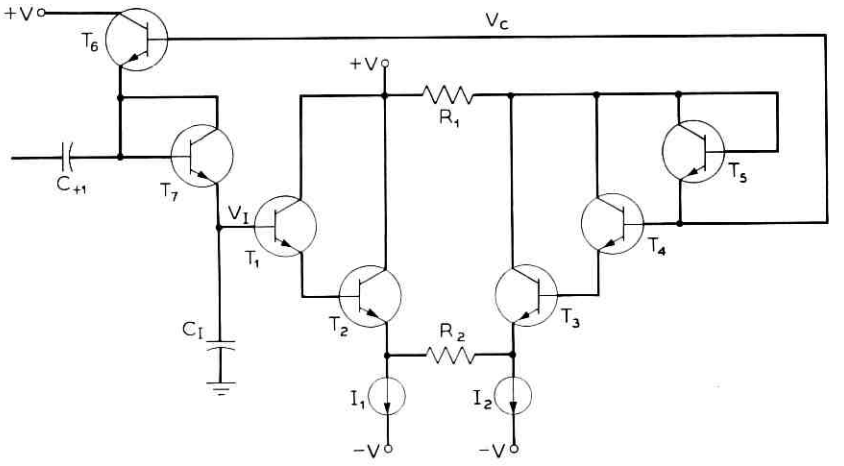


Fig. 5—Integrator step compensation circuit.

able to the current source I_2, and the excess current will be forced to flow in R_2, in this case keeping V_C more positive than V_I. Thus, the amount that C_{+1} is recharged after each voltage step becomes a function of the integrator voltage level, decreasing with increasing integrator voltage and increasing with decreasing voltage. The net effect of compensation on the +1 step size can be expressed as

$$\sigma'_{+1} = \sigma_{+1} - \Delta V_I\left(1 + \frac{R_1}{R_2}\right)^{-1}\left(\frac{C_{+1}}{C_I + C_{+1}}\right) \tag{8}$$

where σ_{+1} is the +1 step at V_I and σ'_{+1} is the +1 step at $V_I + \Delta V_I$.

Gain (loss) variation between coder and decoder is controlled by reproducing the σ_{-2} step size. This step size becomes primarily a function of the dc step control voltage, V_s, the ratio of C_{-2} to C_I, and the common base current gain of the negative step charge parceling transistor. To meet gain (loss) variation requirements, a ± 1 percent tolerance is needed for both the step voltage and the capacitor ratio.

The compensation network automatically adjusts σ_{+1} in the decoder to match the σ_{+1} to σ_{-2} ratio between coder and decoder. Since a small current is derived from C_I as bias for the output buffer stage (T_1 and T_2), a drain is produced on σ_{+1}. Variations in bias current between coder and decoder cause a difference in the effective σ_{+1} step size but are counterbalanced by the compensation network. To minimize drain from C_I, a high impedance connection is made at the analog output terminal from the integrator.

3.3 *Codec Building Blocks*

Figures 6 and 7 are schematics of the comparator and the integrator circuits, respectively. A combination of a comparator and an integrator are required by the coder; only an integrator is required by the decoder. Figure 8 is the block diagram of a complete codec. Thus, a codec is implemented from 61 transistors and diodes, 54 resistors, and 6 capacitors. Transistors, diodes, and resistors are fabricated by IC techniques, the capacitors by thin-film techniques.

The comparator stage (Fig. 6) compares an analog input with an integrator output signal. Their difference is amplified by approximately 200 and is passed to a latching circuit (T_8, T_9, T_{12}). The latch holds the state of the comparator output for the duration of the clock pulse and allows conversion of the output to a single bit digital code by the clocked output gate (T_{10}, T_{11}). To avoid a race condition CLOCK1 input is made to overlap the pulse from CLOCK2 input.

The integrator network (Fig. 7) converts the clock and digital inputs

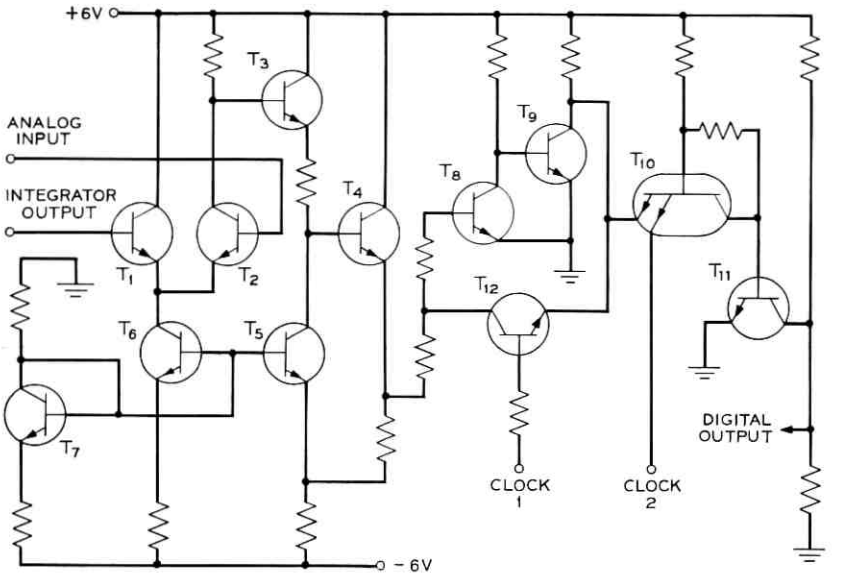


Fig. 6—ΔM comparator.

at the integrator into controlled amplitude voltage steps by first amplifying the input pulses (T_2) and then generating a voltage step (T_3, T_4) to the charge parceling circuit. The step is controlled by a voltage control (V_a = 1 volt) plus two temperature compensating diodes (T_5, T_6). The charge parceling circuit uses 33-pF and 69-pF capacitors for the +1 and −2 steps and a 2690-pF capacitor for the integrator. This allows approximately a 12-millivolt integrator step voltage. Note that the ratio of the charge parceling capacitors represents only a portion of the integrator step imbalance. The effective size of the positive step is decreased by output stage biasing requirements and by the effect of parasitic capacitances in the charge parceling network. The compensation network is designed to provide approximately 2 percent of step size compensation per volt change in the integrator output level.

IV. CODEC OPERATION

Tests were performed on codecs fabricated from discrete beam-lead resistors and transistors on ceramic substrates. Figure 9 shows a photograph of the codec ceramic. The ceramic contains the comparator and the two integrators needed for a complete codec. Discrete capacitors

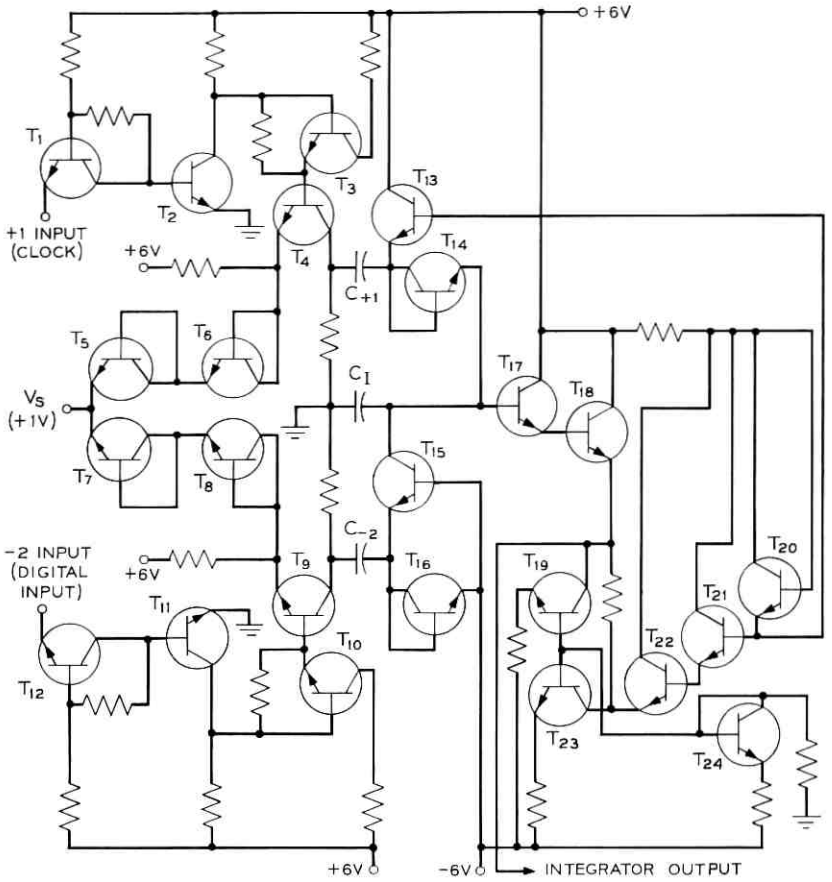


Fig. 7—ΔM integrator.

were used for the charge parceling circuit and were externally mounted to the model.

Measurements were made with a 1.544-MHz sampling frequency and a 12-millivolt integrator step size. Integrator imbalance was set at $\vartheta \approx 0.10$. Figure 10 shows typical signal-to-noise characteristics of the codec for a 3200-Hz input signal. Similar characteristics are observed for other input frequencies. Signal-to-noise ratio is well above transmission objectives. Quiescent quantizing noise is measured at less than 12 dBrnC.

Dips in the signal-to-noise curve are caused by the sawtooth error

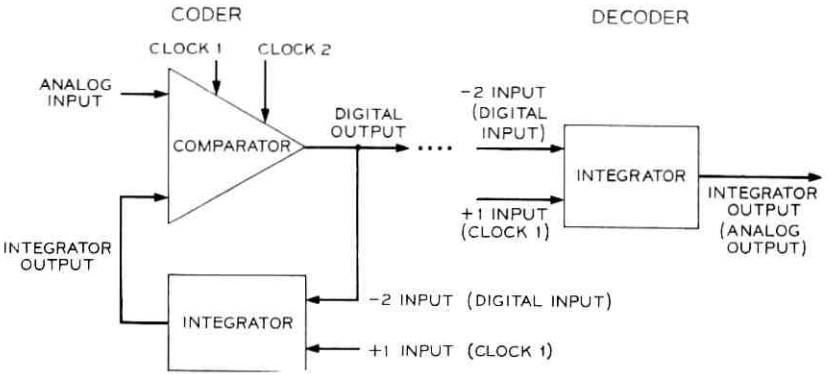


Fig. 8—Interconnecting ΔM comparator and integrator blocks to form a codec.

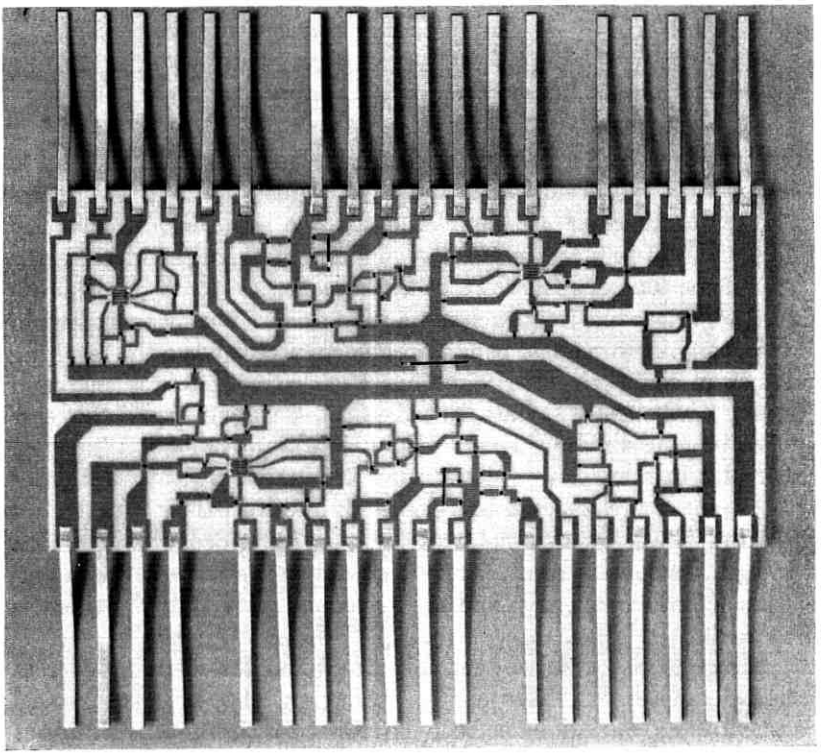


Fig. 9—Discrete beam-lead resistor and transistor model of ΔM codec.

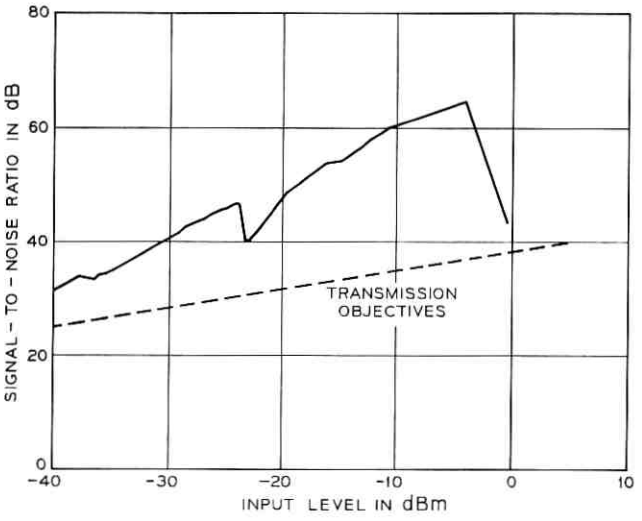


Fig. 10—Measured ΔM signal-to-quantizing noise ratio for a 3200-Hz input signal, f_s = 1.544 MHz, quiescent quantizing noise = 12 dBrnC.

waves generated by the integrator step imbalance.[9] The largest dip in the curve (Fig. 10) occurs when the input signal slope and the fundamental sawtooth slope are approximately the same magnitude. The input signal will reduce the effective slope (and therefore increase the length) of the fundamental sawtooth; and as the sawtooth frequency is reduced to voiceband frequencies, additional noise begins to appear in the voiceband. A worst-case condition is encountered when the resulting fundamental sawtooth error wave has been reduced to the midband frequency of the voice signals. For the 3200-Hz input signal this corresponds to approximately a −23-dBm (900-Ω) input signal. Additional smaller dips in the signal-to-noise curve are caused when higher order sawtooth waves are forced into voiceband; however, their effect is not significant. Thus, signal-to-noise problems are avoided provided that the integrators contain a sufficiently high unbalance to keep the fundamental sawtooth frequency from the voiceband until high signal levels are coded.

A number of codecs have been fabricated and tested using discrete beam-lead devices. All have shown similar signal-to-noise characteristics. Gain (loss) variation has been maintained within ±0.2 dB. Power dissipation is approximately 250 milliwatts per codec.

Design of an integrated version of the codec has also been completed

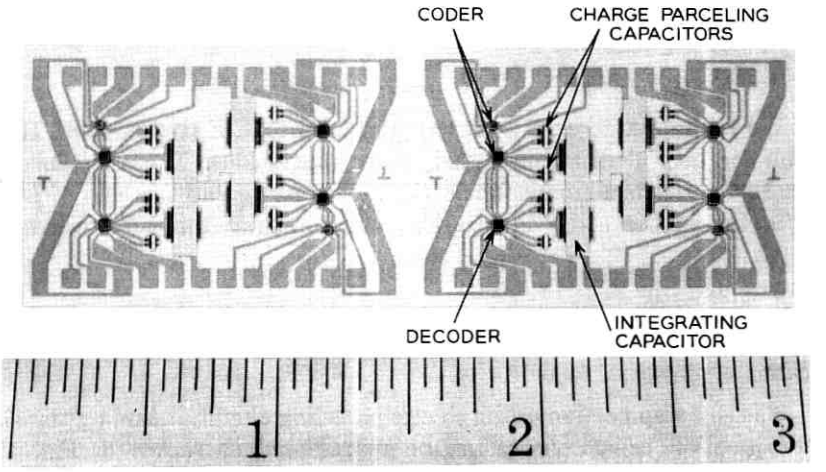


Fig. 11—Four integrated ΔM codecs using thin-film charge parceling capacitors.

and preliminary tests have indicated satisfactory codec operation. The codec is fabricated from three IC chips—one comparator chip and two identical integrator chips. Figure 11 shows a ceramic substrate containing four complete codecs. Thin-film capacitors are used in the charge parceling circuits. To relieve fabrication tolerance requirements, the integrated circuit models use approximately twice the capacitor values of the discrete model.

V. CONCLUSIONS

Delta modulation techniques offer highly integrable and economical analog-to-digital conversion elements. We have described a delta modulation codec design suitable for voiceband applications where conversion economy and quality are more important than the transmission bandwidth requirements.

Techniques are utilized in the design to force a large portion of quantizing noise out of voiceband by using a controlled step imbalance in the integrator. Excellent noise characteristics are achieved. The design attempts to take advantage of integrated circuit techniques and thin-film capacitor techniques by relying on matching of device characteristics and on accurate capacitor ratio tolerances for reproducing coding characteristics. Gain (loss) variations have been maintained within

±0.2 dB using codecs fabricated from discrete beam-lead transistors and resistors.

VI. ACKNOWLEDGMENTS

We would like to express our appreciation to J. E. Iwersen and H. J. Boll for helpful discussions and many suggestions. The assistance of D. E. Gearhart in circuit fabrication, and testing and the help of D. J. D'Stefan in the codec integration, are also gratefully acknowledged.

APPENDIX

Charge Parceling—Effect of Junction Capacitances on Integrator Step Size

The +1 step portion of the charge parceling circuit, including transistor junction capacitances, can be represented as shown in Fig. 12. For $\Delta V' \leqq V_{BE_1} + V_{BE_2}$, the circuit is approximated by Fig. 13. Before T_2 beings to conduct, $\Delta V'$ must increase by $V_{BE_1} + V_{BE_2}$ from its initial value. This requires a voltage swing of $\Delta V''$ at the input, where $\Delta V'' < \Delta V$. The magnitude of $\Delta V''$ is given by

$$(V_{BE_1} + V_{BE_2})(C_{TS_2} + C_{TE_1} + C_{TE_2}) = (\Delta V'' - V_{BE_1} - V_{BE_2})C_{+1}$$

or

$$\Delta V'' = (V_{BE_1} + V_{BE_2})\left(1 + \frac{C_{TS_2} + C_{TE_2} + C_{TE_1}}{C_{+1}}\right).$$

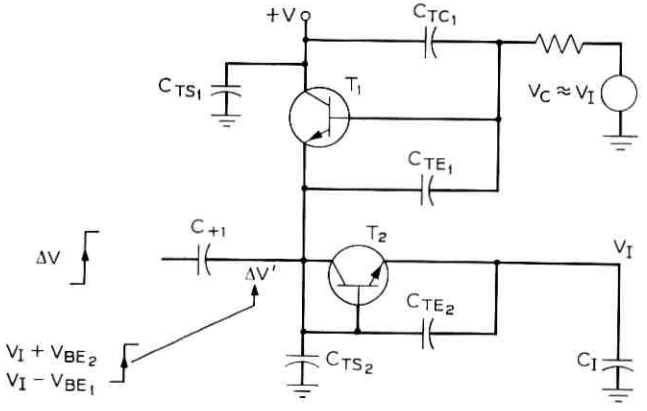


Fig. 12—+1 step portion of the charge parceling circuit.

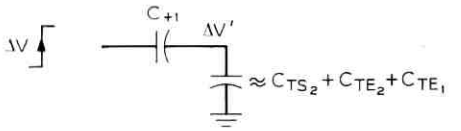


Fig. 13—Approximated circuit (+1 step) for $\Delta V' \leqq V_{BE_1} + V_{BE_2}$.

Thus, charge is dumped on C_I only for a portion of the input swing, $\Delta V - \Delta V''$. After T_2 conducts, the equivalent circuit becomes the one shown in Fig. 14, and

$$(\Delta V - \Delta V'' - \Delta V_I)C_{+1} = \Delta V_I C_I$$

or

$$\Delta V_I = (\Delta V - \Delta V'') \frac{C_{+1}}{C_I + C_{+1}},$$

and

$$\sigma_{+1} = \Delta V_I = \left[\Delta V - (V_{BE_1} + V_{BE2}) \cdot \left(1 + \frac{C_{TS_2} + C_{TE_1} + C_{TE_2}}{C_{+1}}\right)\right] \frac{C_{+1}}{C_I + C_{+1}}.$$

Note that this does not include the change in the effective +1 step size due to biasing current requirements of the output buffer stage.

A voltage step is also produced on C_I due to C_{TE_2}; however, the positive and negative portions are essentially the same and the effect is self-canceling. The −2 step portion of the charge parceling circuit is given by Fig. 15. For $\Delta V' \leqq V_{BE_1} + V_{BE_2}$, the circuit is approximated by Fig. 16.

Before T_1 conducts ΔV must change $\Delta V'$ by $V_{BE_1} + V_{BE_2}$. The

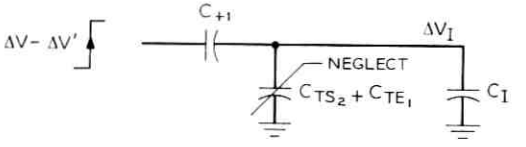


Fig. 14—Equivalent circuit for +1 step charge transfer.

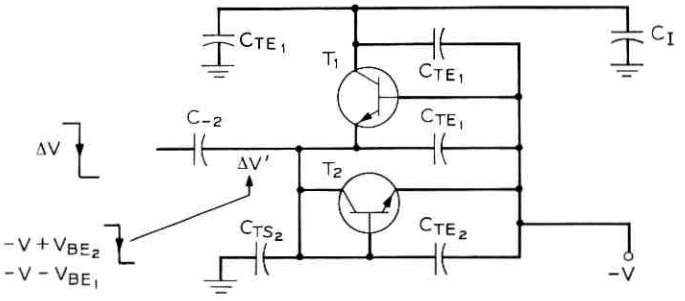


Fig. 15——2 step portion of the charge parceling circuit.

magnitude of this voltage change at input, $\Delta V''$, is given by

$$(V_{BE_1} + V_{BE_2})(C_{TS_2} + C_{TE_1} + C_{TE_2}) = (\Delta V'' - V_{BE_1} - V_{BE_2})C_{-2}$$

or

$$\Delta V'' = (V_{BE_1} + V_{BE_2})\left(1 + \frac{C_{TS_2} + C_{TE_1} + C_{TE_2}}{C_{-2}}\right).$$

Thus, charge is drained from C_I during the voltage swing $\Delta V - \Delta V''$. After T_1 conducts, the equivalent circuit is as shown in Fig. 17, and

$$\Delta V_I = (\Delta V - \Delta V'')\frac{\alpha_2 C_{-2}}{C_I + \alpha_2 C_{-2}}$$

or

$$\sigma_{-2} = \Delta V_I = \left[\Delta V - (V_{BE_1} + V_{BE_2}) \cdot\left(1 + \frac{C_{TS_2} + C_{TE_1} + C_{TE_2}}{C_{-2}}\right)\right]\frac{\alpha_2 C_{-2}}{C_I + \alpha_2 C_{-2}}.$$

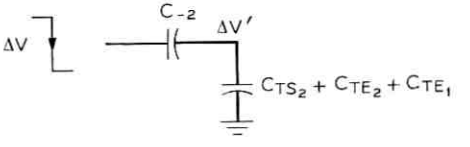


Fig. 16—Approximated circuit (-2 step) for $\Delta V' \leqq V_{BE_1} + V_{BE_2}$.

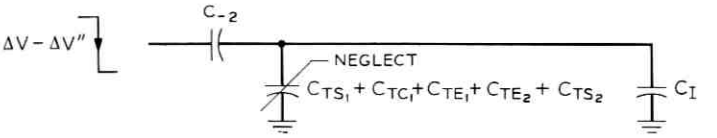


Fig. 17—Equivalent circuit for −2 step charge transfer.

REFERENCES

1. de Jager, F., "Delta Modulation, a Method of PCM Transmission Using the 1-Unit Code," Philips Res. Rep., *7*, (1952), pp. 442–466.
2. Jackson, L. B., Kaiser, J. F., and McDonald, H. S., "An Approach to the Implementation of Digital Filters," IEEE Trans. on Audio and Electroacoustics, *AU-16*, No. 3 (September 1968), pp. 413–421.
3. Goodman, D. J., "The Application of Delta Modulation to Analog-to-PCM Encoding," B.S.T.J., *48*, No. 2 (February 1969), pp. 321–343.
4. Van de Weg, H., "Quantizing Noise of a Single Integration Delta Modulation System with an N-Digit Code," Philips Res. Rep., *8*, (1953), pp. 367–385.
5. Wang, P. P., "Idle Channel Noise of Delta Modulation," IEEE Trans. Commun. Techniques, *16*, No. 10 (October 1968), pp. 737–742.
6. Iwersen, J. E., "Calculated Quantizing Noise of Single Integration Delta Modulation Coders," B.S.T.J., *48*, No. 7 (September 1969), pp. 2359–2389.
7. Laane, R. R., "Measured Quantizing Noise Spectrum for Single Integration Delta Modulation Coders," B.S.T.J., *49*, No. 2 (February 1970), pp. 191–195.
8. Bowers, F. K., "Asymmetric Delta Modulation System," U. S. Patent 2-817-061, applied for June 7, 1955, issued December 17, 1957.
9. Boll, H. J., unpublished work.

On the Performance of Digital Modulation Systems That Expand Bandwidth

By V. K. PRABHU

(Manuscript received December 19, 1969)

It is well known that protection against additive gaussian noise can be obtained in m-ary digital modulation systems by expanding bandwidth or by increasing the channel signal-to-noise ratio. It is also well known that arbitrarily small error probabilities can be attained in digital systems by using long and complex encoding and decoding procedures. Based on the results of Shannon and Slepian, we derive for an optimal system lower bounds to the channel signal-to-noise ratio for various probabilities of error, for various bandwidth expansions, and for a processing interval not greater than the signaling interval of the source. It is assumed that all m characters have equal a priori probabilities and that maximum likelihood detection is used in the receiver. For a bandwidth expansion of two, and for equal energy code words, we also show that the performance of a coherent phase-shift keyed system is as good as that of the optimal system.

I. INTRODUCTION

Various m-ary digital modulation schemes [such as coherent phase-shift keying (CPSK), differentially coherent phase-shift keying (DCPSK), frequency-shift keying (FSK), and others] are currently under investigation for use in satellite, terrestrial, and other radio communication systems.[1,2,3] In such systems, the transmission channel is noisy, and bandlimited, and one is interested in finding an optimum form of modulation for the transmission of information from one point to another. By optimum form of modulation, we mean that we would like to transmit (with a given error rate) as much information as possible in a given band of frequencies and for a given amount of (channel) signal power.

The complexity of the equipment required for particular kinds of modulation, or other considerations in the system, may rule out these optimum transmission schemes in favor of simple and suboptimum

schemes of modulation and demodulation. In order to compare the performance of these simpler practical modulation systems with that of the optimal systems, it is first essential to investigate the performance of these optimal systems.

In this paper we shall assume that the noise in the channel available for communication is gaussian and has a uniform power spectral density over all useful frequency bands. [In terrestrial systems, in the frequency bands above 10 GHz, close spacings of the repeaters are almost always mandatory for reliable communication (during fading conditions produced by rain).[2] If low noise receivers are used in the system, it is possible that the total interference power (due to co-channel and adjacent channel interferers) received by the system may be very much larger than the (thermal) noise power in the system.[3] In this case, note that the total noise corrupting the channel may not be assumed gaussian in all modulation systems (especially if the number of interferers is small).[4,5]]

It is well known that protection against (additive) white gaussian noise can be obtained in m-ary digital modulation systems by expanding bandwidth and/or by increasing the channel signal-to-noise (power) ratio. In fact, Shannon has shown that it is possible to transmit with arbitrarily small error probability the output of a discrete source of entropy R over a channel of bandwidth W perturbed by additive white gaussian noise of average power N by using signals of average power S provided R is less than $W \log_2 (1 + S/N)$ b/s.[6,7] However, such a transmission scheme may involve long and complex encoding and decoding procedures, and to attain these low error rates it may be necessary to provide large storage (or long delay) in the transmitting and receiving equipment. Practical modulation systems presently used for large scale communication do not in general have such large storage capabilities or unlimited bandwidths. Also, we must note that there are practical limitations on the average power of a transmitter and the power that can be received by a receiver.

Since most of the practical modulation systems have a certain bandwidth expansion n and since bandwidth expansion usually improves the (noise) performance of the system, we shall now investigate the optimum performance of digital modulation systems that have a (channel) bandwidth expansion n, a finite channel signal-to-noise power ratio $S/N = \rho^2$, and a processing interval which cannot exceed one signaling interval T of the source.* The message source is assumed to

* It is assumed that the time interval over which the channel is used in decoding one message symbol cannot exceed T.

be of bandwidth W_0, and the channel bandwidth is assumed to be W. Further, it is assumed that all m characters (the output of the discrete message source consists of m different characters or symbols) have equal *a priori* probabilities and that the characters generated by the message source are statistically independent of each other. We also assume that a maximum likelihood detection scheme is used in the receiver.

For such a system, we shall evaluate the lower bounds to the character error probability $P(m, n, \rho^2)$ of the optimal system so that we can compare its performance with that of any practical modulation system. For a given error rate, the difference between the signal-to-noise ratio required by the optimal system and that required by the practical system will then be a measure of the quality of performance of the practical modulation system.

Here, we would like to note that our approach is identical to that of Slepian in Ref. 8 in which he evaluated upper and lower bounds to $P(m, n, \rho^2)$ for n odd, $n \geqq 5$, and $m = 128$ (numerical results for $m = 32$, 64, and 256 can also be found in an unpublished memorandum by Slepian[9]) when $1 \geqq P(m, n, \rho^2) \geqq 10^{-5}$.* In addition to giving numerical results when $P(m, n, \rho^2) < 10^{-5}$ (error rates as low as 10^{-9} are desired in some digital modulation systems[2]) for $m = 2$, 4, 8, 16 and 32, we give a method of evaluating upper and lower bounds to $P(m, n, \rho^2)$ for all values of $n \geqq 2$. We also point out the special significance of $n = 2$, and give closed form solutions for the lower bound when $n = 2$, 3 and 5.

II. COMMUNICATION SYSTEM MODEL†

The m-ary digital modulation system that we shall consider in this paper is assumed to have a signaling interval T. Since we assume Nyquist rate signaling, we shall assume that

$$T = \frac{1}{2W_0}, \tag{1}$$

where W_0 is the bandwidth of the message source. Every T seconds, the message source generates one of m characters or symbols. Since the characters generated by the message source are assumed to be statistically independent, the entropy R of the message source is given by

$$R = 2W_0 \log_2 m \text{ b/s}. \tag{2}$$

* Note that Slepian uses $P(m, n, \rho^2)$ in determining the threshold in *analog* modulation systems that expand bandwidth.

† Compare the communication system model that we discuss in this paper to that given in Ref. 8.

If S is the average power in the channel of bandwidth W, it follows from Ref. 6 that it is possible to transmit with arbitrarily small error probability if and only if

$$2W_0 \log_2 m \leqq W \log_2 (1 + S/N) \tag{3}$$

when there are no restrictions on the way in which the transmitter and receiver operate.

Equation (3) can be shown to yield

$$S/N \geqq m^{2/n} - 1 \tag{4}$$

where

$$n = \frac{W}{W_0} = 2TW = \text{the bandwidth expansion factor.} \tag{5}$$

The lower bound to S/N given by equation (4) is then the smallest signal-to-noise ratio required to transmit (with arbitrarily small error probability) an m-ary digital signal through a channel of bandwidth expansion n when there are no restrictions on the transmitting and receiving equipments. This lower bound to the signal-to-noise ratio S/N is shown in Fig. 1.

In the communication system model we are considering in this paper, there is no provision for the storage of a large number of characters, and hence it is to be expected that we will need signal-to-noise ratios much larger than those given in Fig. 1 (when arbitrarily low error rates are desired). Since the processing interval of our communication system is assumed not to exceed one Nyquist interval (corresponding to the message source), the channel signal corresponding to time T can be used to decode one and only one message symbol.* If the bandwidth of the channel is W, the channel signal can be completely specified by samples taken every $1/2W$ seconds. In the Nyquist interval T, there are then $n = 2WT$ channel samples.† We are then assuming that n channel samples are used to decode one message character, or that each of the m message symbols are mapped into a channel vector having n components.‡ (In the error-free transmission scheme of Shannon, $\ell, \ell \geqq 1$, successive message symbols are mapped into one channel vector. By making ℓ sufficiently large, by choosing the m^l channel vectors appropriately and by decoding appropriately at the receiver, the results given

* This is equivalent to saying loosely that the communication system does not have storage capability for more than one Nyquist interval.[8]

† There are certain subtle points involved in this assumption. Some of these points and their implications are discussed in Refs. 10 and 11.

‡ That is, we construct a dictionary that associates with each of the m message symbols a particular n-dimensional channel vector.

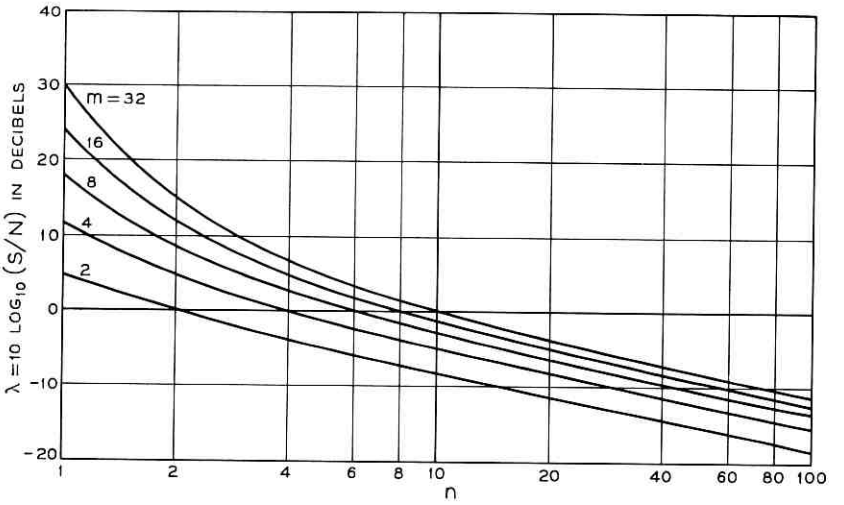


Fig. 1—Lower bound to the signal-to-noise ratio for different bandwidth expansions and ideal signaling.

by Shannon may be obtained.[8] It is to be noted that it is essential to store ℓ message characters before we can generate the appropriate channel vector. In our scheme of transmission, we put $\ell = 1$, and investigate the optimum performance of the system.)

As far as the average power in the channel is concerned, the channel vectors can be chosen in a variety of ways.[12] Since all the characters are assumed to have equal *a priori* probabilities, we will make the assumption that all of them have the same average power S (or the same energy ST).*

Since the noise corrupting the channel is assumed to be white, each component of each channel vector is perturbed independently by the addition of a gaussian variate of mean zero and variance N.

III. EVALUATION OF PROBABILITY OF ERROR $P(m, n, \rho^2)$

Since the channel vectors corresponding to different message symbols have the same average power S, all these n-dimensional vectors ter-

* Other types of restrictions (such as maximum power, maximum average power, and so on) can also be put on the signal vectors to analyze the communication system given in our paper. Since all symbols are assumed to be equally likely, we do not consider a system in which there can be unequal distribution of power among different channel vectors. In particular, some amplitude modulation systems (such as single-sideband AM) do not satisfy the requirement that channel vectors corresponding to different message characters have the same average power S, and hence such systems are not covered in this paper.

minate on the surface of a sphere of radius $(nS)^{\frac{1}{2}}$. By choosing the channel vectors appropriately, and by using maximum likelihood detection receiver, it can be shown[12] that the minimum probability of error $P(m, n, \rho^2)$, averaged over all symbols, satisfies the inequalities

$$Q(m, n, \rho^2) \leqq P(m, n, \rho^2) < \bar{Q}(m, n, \rho^2), \quad \rho^2 = \frac{S}{N}, \tag{6}$$

where[8,12]

$$Q(m, n, \rho^2) = L\left(\theta_{m,n}, n, \rho^2 \frac{n}{2}\right), \tag{7}$$

$$\bar{Q}(m, n, \rho^2) = U\left(\theta_{m,n}, n, \rho^2 \frac{n}{2}\right), \tag{8}$$

$$\frac{2}{m} = \frac{\int_0^{\theta_{m,n}} \sin^{n-2} \mu \, d\mu}{\int_0^{\pi/2} \sin^{n-2} \mu \, d\mu} = I_{\sin^2 \theta_{m,n}}\left(\frac{n-1}{2}, \frac{1}{2}\right), \tag{9}$$

$$L\left(\theta, n, \rho^2 \frac{n}{2}\right) = \int_\theta^\pi p_n(\lambda) \, d\lambda, \tag{10}$$

$$U\left(\theta, n, \rho^2 \frac{n}{2}\right) = L\left(\theta, n, \rho^2 \frac{n}{2}\right) + \Omega(\theta) \int_0^\theta p_n(\lambda) \Omega(\lambda) \, d\lambda, \tag{11}$$

$$\Omega(\lambda) = \frac{(n-1)\pi^{(n-1)/2}}{\Gamma\left(\frac{n+1}{2}\right)} \int_0^\lambda \sin^{n-2} \mu \, d\mu, \tag{12}$$

$$\Gamma(k) = \int_0^\infty e^{-x} x^{k-1} \, dx, \tag{13}$$

$$p_n(\lambda) = \frac{(n-1) \exp\left(-\rho^2 \frac{n}{2} \sin^2 \lambda\right) \sin^{n-2} \lambda}{2^{n/2} (\pi)^{\frac{1}{2}} \Gamma\left(\frac{n+1}{2}\right)} \cdot \int_0^\infty r^{n-1} \exp\left[-\frac{[r - \rho(n)^{\frac{1}{2}} \cos \lambda)^2}{2}\right] dr, \tag{14}$$

and $I_x(\alpha, \beta)$ is the incomplete beta function given by

$$I_x(\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \int_0^x t^{\alpha-1}(1-t)^{\beta-1} \, dt, \qquad 0 \leqq x \leqq 1; \tag{15}$$

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} \, dt. \tag{16}$$

The significance of the inequalities in equation (6) may be explained as follows. No m-ary digital modulation system with a given bandwidth expansion n and a given signal-to-noise ratio ρ^2 can achieve a lower probability of error than that given by $Q(m, n, \rho^2)$.* Also, we observe from equation (6) that m-ary digital modulation systems can be built to have an error probability given by $\bar{Q}(m, n, \rho^2)$.†

The bounds given by equation (6) can, therefore, be used in comparing the performance of practical modulation systems with that of the optimal systems, and in estimating the quality of performance of the practical modulation systems. For a given probability of error, and m and n, we can compare the signal-to-noise ratio required for a practical modulation system with the minimum signal-to-noise ratio predicted by equation (6) for the optimal system.‡

Since we are interested in this kind of comparison and since there seem to be theoretical considerations which show[10] that $\bar{Q}(m, n, \rho^2)$ is a very weak bound [it is shown in Ref. 10 that some explicit codes can be constructed to make $P(m, n, \rho^2)$ very close to $Q(m, n, \rho^2)$], we shall not discuss the upper bound $\bar{Q}(m, n, \rho^2)$ any more in this paper.

IV. EVALUATION OF LOWER BOUND $Q(m, n, \rho^2)$

For the sake of comparing the performance of proposed modulation systems with that of the optimal systems, it is essential to evaluate the lower bound $Q(m, n, \rho^2)$ for different values of m and n. The evaluation of $Q(m, n, \rho^2)$ is rather difficult and is usually done by using a digital computer. Slepian[8] has given methods of evaluating this bound when n is *odd*, and numerical values are given for $Q(m, n, \rho^2)$ when $1 \geqq Q(m, n, \rho^2) \geqq 10^{-5}$, $m = 32, 64, 128$ and 256, and $n \geqq 5$.

Since error rates of less than 10^{-5} are desired in digital systems and since no numerical results are available when $m < 32$ (in general, it is easier to build digital modulation systems with low values of m), we shall give some further numerical values for the cases considered by Slepian. In addition, we shall give a method to evaluate $Q(m, n, \rho^2)$ when n is even and point out its special significance when $n = 2$.

First we review briefly Slepian's method of evaluating $Q(m, n, \rho^2)$.[10] For any given values of m and n, we evaluate by interpolation the value of $\theta_{m,n} (0 < \theta_{m,n} \leqq \pi/2)$ from the set of tables for $I_x(\alpha, \beta)$ given in Ref.

* Of course, we assume that the digital modulation system satisfies other requirements given in this paper.

† For $n = 1$ or 2, it can be shown that we can make $Q(m, n, \rho^2) = P(m, n, \rho^2)$. Also, for all (integral) n, and $m = 2$, we can make $Q(2, n, \rho^2) = P(2, n, \rho^2)$.

‡ In making this comparison, we have to assume that we can estimate the bandwidth expansion factor n for the practical modulation system.

13. Since it has been shown[10] that

$$L(\theta, n, \sigma^2) = L(\theta, n-2, \sigma^2) + \cos\theta G(\theta, n-2, \sigma^2), \qquad n > 3, \quad (17)$$

$$\sigma^2 = \rho^2 \frac{n}{2}, \quad (18)$$

$$G(\theta, n, \sigma^2) = \sigma \cos\theta \sin\theta b_n G(\theta, n-1, \sigma^2) + \frac{n-2}{n-1} \sin^2\theta G(\theta, n-2, \sigma^2), \qquad n > 2, \quad (19)$$

$$b_n = \frac{n-2}{n-1} b_{n-2}, \qquad n > 2, \quad (20)$$

and

$$b_1 = \pi^{\frac{1}{2}}, \quad (21)$$

$$b_2 = \frac{2}{\pi^{\frac{1}{2}}}, \quad (22)$$

$$G(\theta, 1, \sigma^2) = \tfrac{1}{2} \exp(-\sigma^2 \sin^2\theta)[2 - \operatorname{erfc}(\sigma \cos\theta)], \quad (23)$$

$$G(\theta, 2, \sigma^2) = \frac{1}{\pi} \sin\theta \exp(-\sigma^2) + \frac{2\sigma}{\pi^{\frac{1}{2}}} \sin\theta \cos\theta G(\theta, 1, \sigma^2), \quad (24)$$

$$L(\theta, 3, \sigma^2) = \tfrac{1}{2} \operatorname{erfc}(\sigma) + \cos\theta G(\theta, 1, \sigma^2), \quad (25)$$

where

$$\operatorname{erfc}(x) = \frac{2}{\pi^{\frac{1}{2}}} \int_x^\infty \exp(-t^2)\, dt = 1 - \operatorname{erf}(x), \quad (26)$$

$Q(m, n, \rho^2)$ can be evaluated for *odd* n from equations (7), (9) and (17) through (25). However, for even n, we cannot use Slepian's method of evaluating $Q(m, n, \rho^2)$ unless we can find an explicit expression for $L(\theta, 2, \sigma^2)$.

Now it has been proved[14,15,16] that moderately high values of $n (2 < n < 5)$ are required for some digital modulation systems in order to optimize transmission rates per unit bandwidth. Since we would like to compare these systems (and other systems with similar bandwidth expansions) with the optimal systems, we shall first express $Q(m, n, \rho^2)$ explicitly for odd n, and $n \leqq 5$ before we discuss the evaluation of $Q(m, n, \rho^2)$ for even n.

4.1 *Lower Bound* $Q(m, n, \rho^2)$ *for* $n = 3, 5$

For $n = 3$, equations (7), (9), (10), (13) and (14) can be shown to yield

$$Q(m, 3, \rho^2) = \frac{1}{2}\left(\operatorname{erfc}\left[\rho\left(\tfrac{3}{2}\right)^{\frac{1}{2}}\right] + (1 - 2/m)\right.$$
$$\left.\cdot \exp\left[-\frac{6\rho^2}{m}\left(1 - \frac{1}{m}\right)\right]\left\{2 - \operatorname{ercf}\left[\rho\left(\tfrac{3}{2}\right)^{\frac{1}{2}}\left(1 - \frac{2}{m}\right)\right]\right\}\right). \quad (27)$$

For $m = 2^\ell$, $1 \leqq \ell \leqq 5$, we have evaluated $Q(m, 3, \rho^2)$ and have shown the results in Fig. 2.

Let us now compare the performance of an FSK system (using square-wave modulation, ideal discrimination detection with an integrate-and-dump circuit as the post-detection filter) with a bandwidth expansion of 3 with the performance of the optimal system. For the FSK system the symbol error probability P_{FSK} can be shown[16,17] to be given by

$$P_{\text{FSK}} \sim \frac{1}{(2\pi\rho^2)^{\frac{1}{2}}} \frac{\cot\left\{\frac{\pi}{4}\frac{n-2}{m-1}\right\}}{\left\{\cos\left[\frac{\pi}{2}\frac{n-2}{m-1}\right]\right\}^{\frac{1}{2}}} \exp\left[-2\rho^2 \sin^2\left\{\frac{\pi}{4}\frac{n-2}{m-1}\right\}\right],$$
$$\rho^2 \gg 1, \quad n < m + 1. \quad (28)$$

For $n = 3$ and $m = 4$ and 8 we have plotted P_{FSK} and $Q(m, 3, \rho^2)$ in Fig. 3. Noting that the error rate of the optimal system can be made close to $Q(m, n, \rho^2)$ for small n, it can be observed from Fig. 3 that the error performance of the FSK system is inferior to the optimal system by several decibels. However, note that the formula in equation (28) is an asymptotic formula and that we have calculated the bandwidth expansion factor for the FSK system by Carson's rule. Also, note that we have taken the bandwidth of the message source to be $1/2T$, where T is the signaling interval of the source. If all these assumptions are reasonable, we must conclude from Fig. 3 that the performance of the FSK system is far from being optimum.

Let us now consider $n = 5$. For $n = 5$, we have

$$Q(m, 5, \rho^2)$$
$$= \frac{1}{2}\left[\operatorname{erfc}\left[\rho\left(\tfrac{5}{2}\right)^{\frac{1}{2}}\right] + \frac{5\rho \sin^2\theta \cos^2\theta}{(10\pi)^{\frac{1}{2}}} \exp\left\{-\rho^2\tfrac{5}{2}\right\}\right.$$
$$+ \tfrac{1}{2}\cos\theta \exp\left\{-\rho^2\tfrac{5}{2}\sin^2\theta\right\}(\sin^2\theta + 5\rho^2\cos^2\theta\sin^2\theta + 2)$$
$$\left.\cdot\left\{2 - \operatorname{erfc}\left[\rho\left(\tfrac{5}{2}\right)^{\frac{1}{2}}\cos\theta\right]\right\}\right], \quad (29)$$

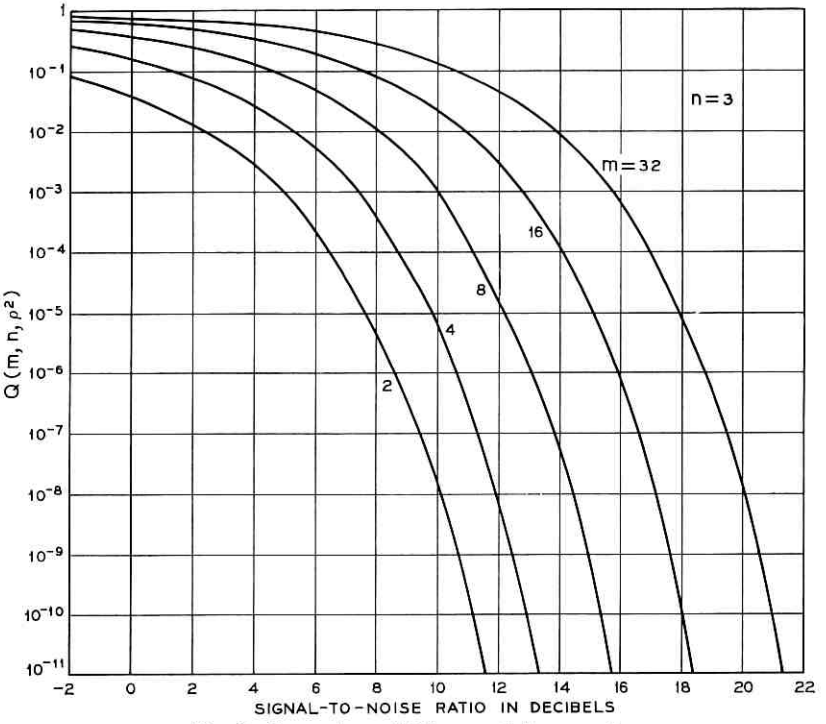


Fig. 2—Lower bound $Q(m, n, \rho^2)$ for $n = 3$.

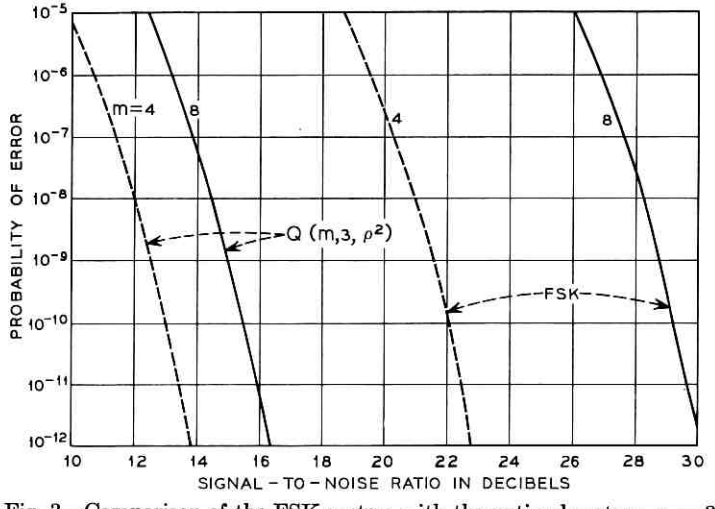


Fig. 3—Comparison of the FSK system with the optimal system, $n = 3$.

where

$$\cos \theta = 2 \cos \left[\frac{4\pi}{3} + \frac{1}{3} \cos^{-1} \left(\frac{2}{m} - 1 \right) \right], \tag{30}$$

and

$$\sin \theta = +(1 - \cos^2 \theta)^{\frac{1}{2}}. \tag{31}$$

For $m = 2^\ell$, $1 \leqq \ell \leqq 5$, we have plotted $Q(m, 5, \rho^2)$ in Fig. 4.

For n odd, and $n > 5$, derivation of an expression for $Q(m, n, \rho^2)$ becomes rather tedious and long, and we shall not give these expressions. However, for $n = 7$, 9, 11, 13 and 17, we have calculated $Q(m, n, \rho^2)$ for $m = 2^\ell$, $1 \leqq \ell \leqq 5$, and the results are given in Figs. 5, 6, 7, 8 and 9. These numerical results which add to the results given by Slepian were obtained by using his method (see Appendix A for an alternative method).

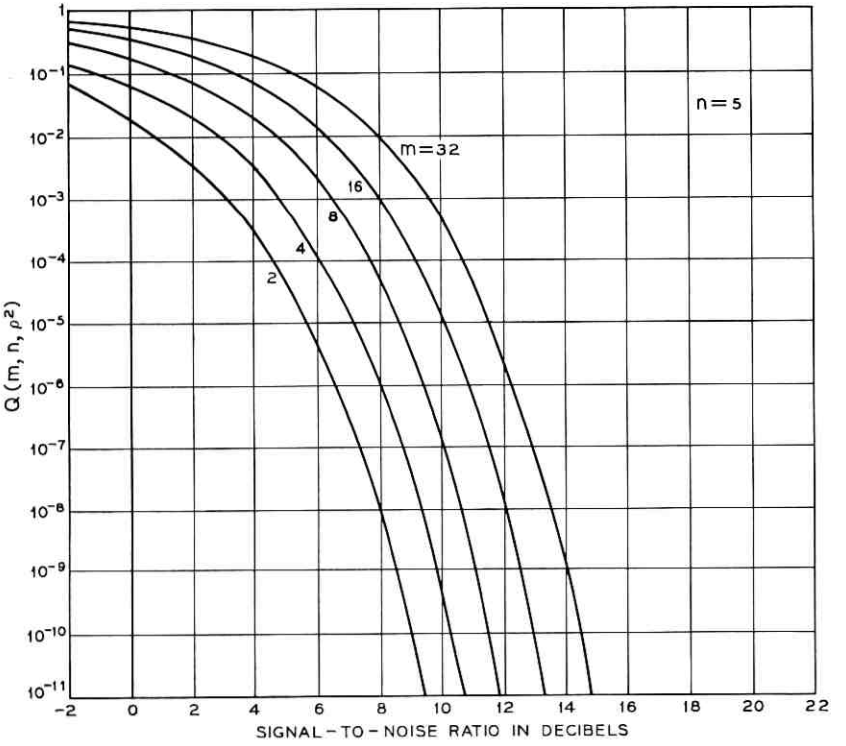


Fig. 4—Lower bound $Q(m, n, \rho^2)$ for $n = 5$.

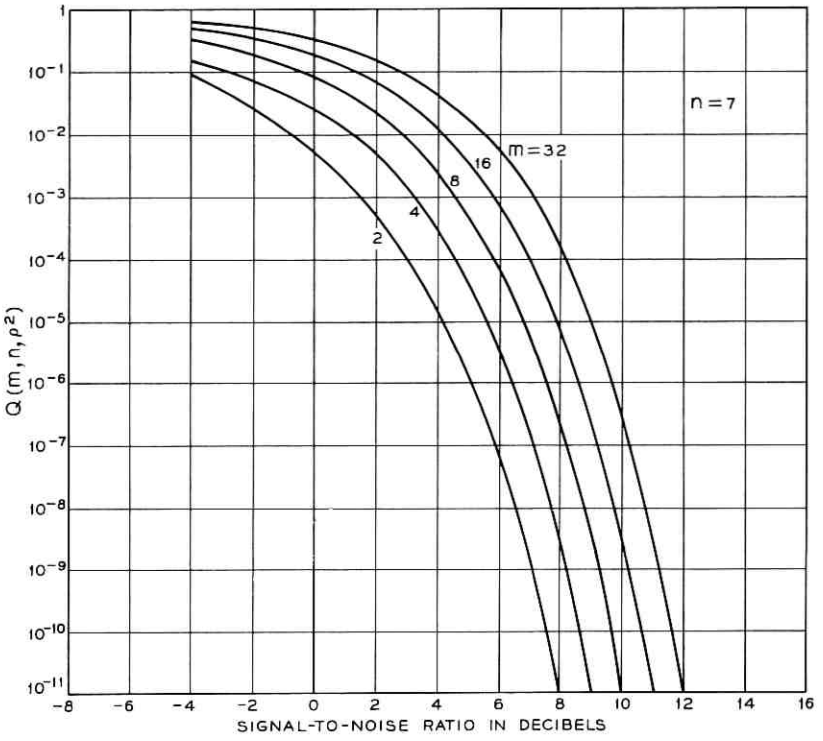


Fig. 5—Lower bound $Q(m, n, \rho^2)$ for $n = 7$.

4.2 *Lower Bound $Q(m, n, \rho^2)$ for n even*

Since we can calculate $Q(m, n, \rho^2)$ from $L(\theta, n, \sigma^2)$ and since the recurrence equation relates $L(\theta, n, \sigma^2)$ to $L(\theta, n - 2, \sigma^2)$ [see equation (17)], we can calculate $Q(m, n, \rho^2)$ for even n if we can calculate $L(\theta, 2, \sigma^2)$.

It can be shown that

$$L(\theta, 2, \sigma^2) = \int_{\theta}^{\pi} p_2(\lambda, \sigma^2)\, d\lambda \tag{32}$$

where

$$p_2(\lambda, \sigma^2) = \frac{1}{\pi}\left[e^{-\sigma^2} + \sigma(\pi)^{\frac{1}{2}} \cos\lambda e^{-\sigma^2 \sin^2\lambda}\{1 + \operatorname{erf}(\sigma \cos\lambda)\}\right]. \tag{33}$$

Noting that $p_2(\lambda, \sigma^2)/2$ is the probability density of the phase angle of a sinusoidal carrier of amplitude $(2A)^{\frac{1}{2}}$ corrupted by random gaussian noise of average power N (signal-to-noise ratio $\sigma^2 = A/N$), we have

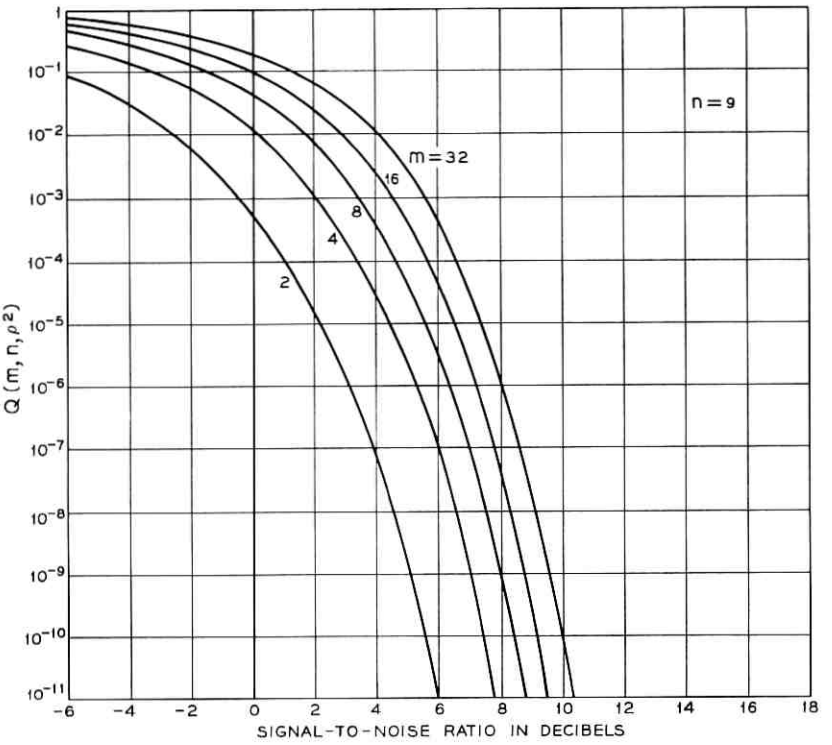


Fig. 6—Lower bound $Q(m, n, \rho^2)$ for $n = 9$.

shown in Appendix B that $L(\theta, 2, \sigma^2)$ can be calculated for any θ. Hence we can calculate $Q(m, n, \rho^2)$ for any even n. Now it can be shown[4] (see Appendix C) that*

$$L(\theta, 2, \sigma^2) = \tfrac{1}{2} \operatorname{erfc} (\sigma), \qquad \theta = \pi/2; \tag{34}$$
$$L(\theta, 2, \sigma^2) = \operatorname{erfc} [\sigma/(2)^{\frac{1}{2}}] - \tfrac{1}{4} \operatorname{erfc}^2 [\sigma/(2)^{\frac{1}{2}}], \qquad \theta = \pi/4$$

and

$$\tfrac{1}{2} \operatorname{erfc} (\sigma \sin \theta) + \max \left\{ 0, \tfrac{1}{2} \operatorname{erfc} (\sigma \sin \theta) - \frac{\tan \theta}{\pi} \exp (-\sigma^2)[1 - \pi^{\frac{1}{2}} \sigma \exp (\sigma^2) \operatorname{erfc} (\sigma)] \right\}$$
$$\leqq L(\theta, 2, \sigma^2) < \operatorname{erfc} (\sigma \sin \theta), \qquad 0 < \theta \leqq \pi/2; \tag{35}$$

* Some of these results can be obtained from Ref. 4 by putting $\Omega = 0$.

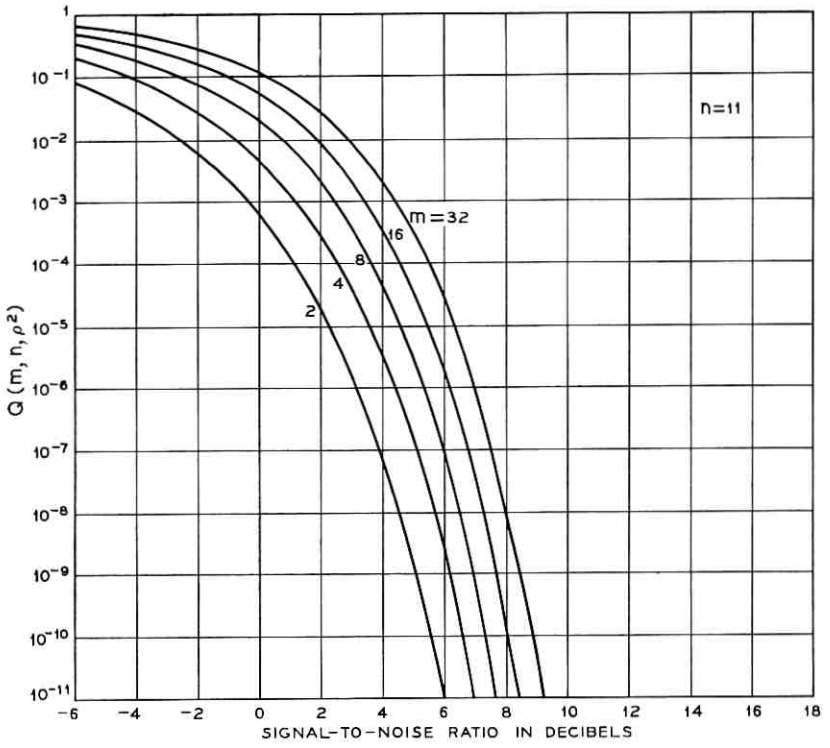


Fig. 7—Lower bound $Q(m, n, \rho^2)$ for $n = 11$.

where

$$\max \{a, b\} = \begin{cases} a, & a \geqq b; \\ b, & a < b. \end{cases} \tag{36}$$

Since the upper and lower bounds to $L(\theta, 2, \sigma^2)$ cannot differ by more than a factor of two and since all quantities involved in equations (17) through (25) are positive, we shall now write a modified bound

$$Q'(m, n, \rho^2) = L'(\theta_{m,n}, n, \sigma^2) \tag{37}$$

where

$$L'(\theta, n, \sigma^2) = L'(\theta, n-2, \sigma^2) + \cos\theta G(\theta, n-2, \sigma^2), \qquad n > 3, \tag{38}$$

and

$$L'(\theta, 2, \sigma^2) = \tfrac{1}{2} \operatorname{erfc}(\sigma \sin\theta). \tag{39}$$

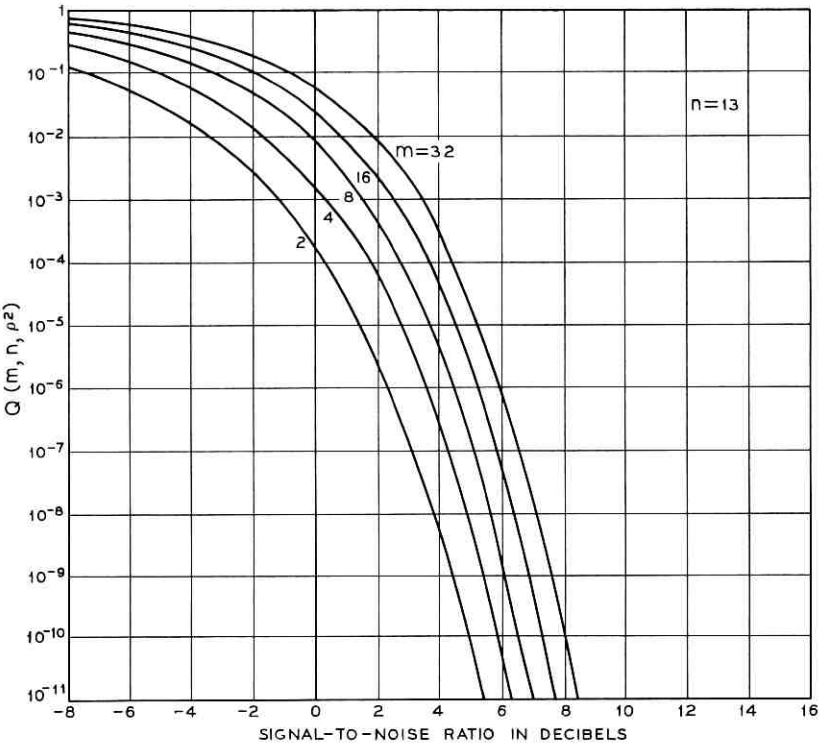


Fig. 8—Lower bound $Q(m, n, \rho^2)$ for $n = 13$.

Since

$$L(\theta, 2, \sigma^2) \geqq L'(\theta, 2, \sigma^2), \qquad 0 < \theta \leqq \pi/2 \tag{40}$$

note that

$$P(m, n, \rho^2) \geqq Q'(m, n, \rho^2). \tag{41}$$

Let us now consider the particular case $n = 2$. For $n = 2$,

$$\frac{2}{m} = \frac{\int_0^{\theta_{m,2}} d\mu}{\int_0^{\pi/2} d\mu} = \frac{2\theta_{m,2}}{\pi} \tag{42}$$

or

$$\theta_{m,2} = \frac{\pi}{m}. \tag{43}$$

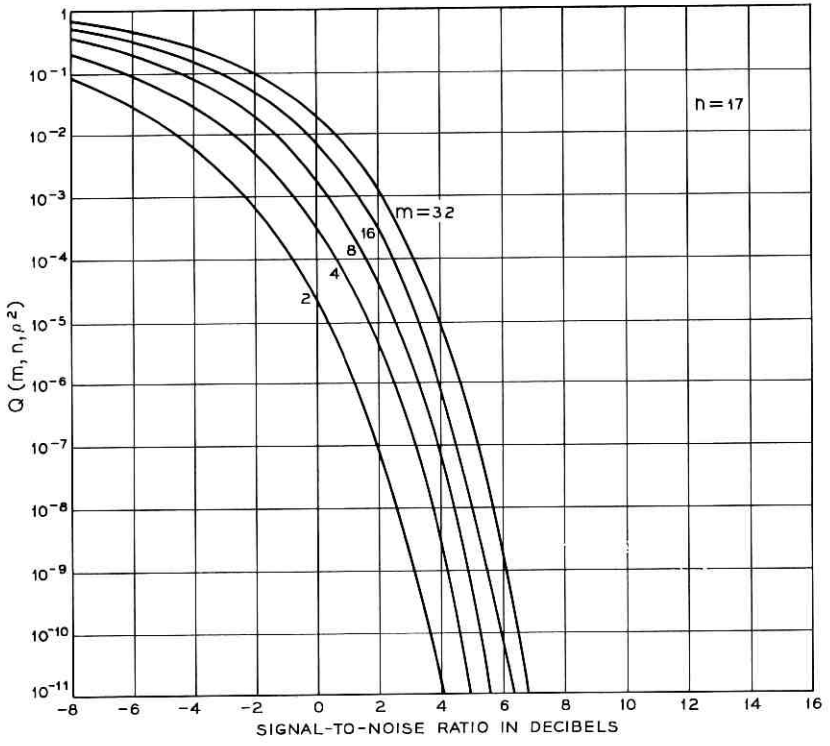


Fig. 9—Lower bound $Q(m, n, \rho^2)$ for $n = 17$.

Since $\theta_{m,2} = \pi/m$, and $p_2(\lambda, \sigma^2)$ is given by equation (33), it can be shown[18,19,20] from equations (7) and (10) that $Q(m, 2, \rho^2)$ is equal to the error probability obtained in m-ary coherent phase-shift keyed (CPSK) systems. Also, for $n = 2$, it can be shown[12] that

$$P(m, 2, \rho^2) = Q(m, 2, \rho^2). \tag{44}$$

It, therefore, follows that the error rates obtained in m-ary coherent phase-shift keyed systems are identical to those obtained in any m-ary digital modulation system that has a bandwidth expansion of two.* Hence we conclude that the error rates of any digital modulation system with a bandwidth expansion of two cannot be lower than the error rates of CPSK systems for all m.

Since the error rates of CPSK systems have been investigated in

* CPSK systems can be shown [20,21] to have approximately a bandwidth expansion of two.

detail,[18,19,20] we shall not give numerical values of $Q(m, 2, \rho^2)$ in this paper. However, we would like to note that[18,19,20]

$$Q(2, 2, \rho^2) = \tfrac{1}{2} \operatorname{erfc} (\rho), \tag{45}$$

$$Q(4, 2, \rho^2) = \operatorname{erfc} [\rho/(2)^{\frac{1}{2}}] - \tfrac{1}{4} \operatorname{erfc}^2 [\rho/(2)^{\frac{1}{2}}], \tag{46}$$

and

$$\tfrac{1}{2} \operatorname{erfc} (\rho \sin \pi/m) + \max \left\{0, \tfrac{1}{2} \operatorname{erfc} (\rho \sin \pi/m) - \frac{\tan \pi/m}{\pi} \exp (-\rho^2)[1 - \pi^{\frac{1}{2}} \rho \exp (\rho^2) \operatorname{erfc} (\rho)]\right\}$$

$$\leqq Q(m, 2, \rho^2) < \operatorname{erfc} (\rho \sin \pi/m), \qquad m > 2. \tag{47}$$

For signal-to-noise ratios greater than 5 dB, it can be shown[4] that

$$Q(m, 2, \rho^2) \approx \operatorname{erfc} (\rho \sin \pi/m), \qquad m \geqq 4, \tag{48}$$

and that the error in this approximation is less than 5 percent.

For $n > 2$, $Q(m, n, \rho^2)$ can be evaluated by using methods presented in Appendix B and using equations (17) through (25). However, this is usually difficult and tedious, and since $Q(m, n, \rho^2)$ and $Q'(m, n, \rho^2)$ can at most differ by a factor of two, we shall use the modified bound $Q'(m, n, \rho^2)$. Observe that $Q'(m, n, \rho^2)$ can easily be evaluated from equations (18) through (24), and (37) through (39).*

For $n = 4, 8, 12$ and 16, and $m = 2^\ell$, $1 \leqq \ell \leqq 5$, we have evaluated $Q'(m, n, \rho^2)$ and the results are shown in Figs. 10, 11, 12 and 13.

V. DISCUSSION AND CONCLUSIONS

Based on the results of Shannon and Slepian, we have derived, for different probabilities of error, lower bounds to the channel signal-to-noise ratio required by optimal systems to transmit the output of an m-ary message source through a channel of bandwidth expansion n. We assume that the channel is perturbed by additive white gaussian noise, all channel signals have the same average power S, and that the processing interval for decoding one message symbol is not greater than one signaling interval. When this interval can be arbitrary and when the transmission rate is not greater than the channel capacity, it is well known that the probability of error can be made arbitrarily

* For any n, note that $\theta_{2,n} = \pi/2$, and that $Q(2, n, \rho^2) = Q'(2, n, \rho^2) = \frac{1}{2} \operatorname{erfc} [\rho(n/2)^{1/2}]$.

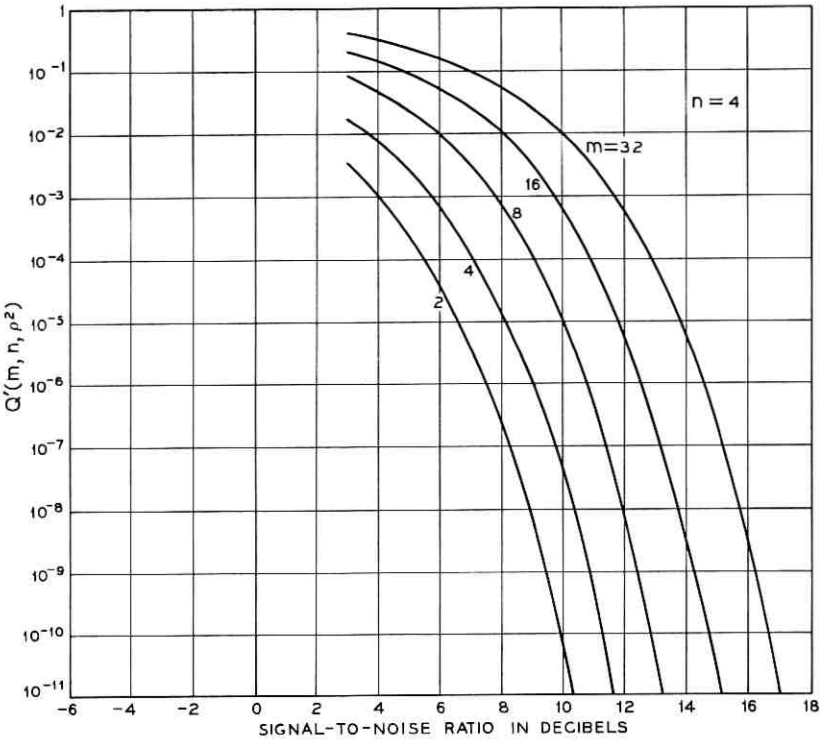


Fig. 10—Lower bound $Q'(m, n, \rho^2)$ for $n = 4$.

close to zero by using long and complex encoding and decoding procedures.

For different practical modulation systems, we can then compare the signal-to-noise ratio required for different probabilities of error with the lower bound given in this paper for optimal systems. This will aid us in deciding about the optimality or nonoptimality of different systems, and in evaluating the quality of performance of different modulation systems.

By using Slepian's method, we evaluate this lower bound for odd n, and $m = 2^\ell$, $1 \leqq \ell \leqq 5$. We also give a method of evaluating the lower bound for even n, and derive a simpler modified lower bound for n even, and $n > 2$. This modified lower bound has been evaluated for n even, and $m = 2^\ell$, $1 \leqq \ell \leqq 5$.

For a bandwidth expansion of two, the performance of a coherent

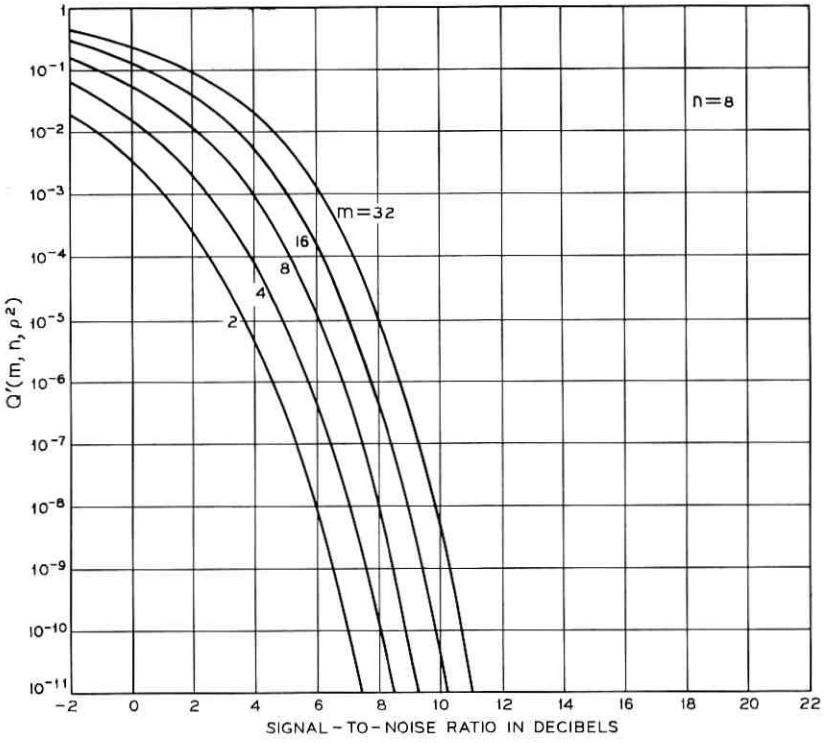


Fig. 11—Lower bound $Q'(m, n, \rho^2)$ for $n = 8$.

phase-shift keyed system has been shown to be as good as that of the optimal system.

A particular FSK system with a bandwidth expansion of three has been compared to the optimal system, and it appears that its performance is substantially inferior to that of the optimal system.

APPENDIX A

Evaluation of Lower Bound $Q(m, n, \rho^2)$

In this appendix, we shall give a second method to evaluate $Q(m, n, \rho^2)$. It can easily be shown from equation (9) that

$$\theta_{2,n} = \frac{\pi}{2} \quad \text{for all} \quad n, \tag{49}$$

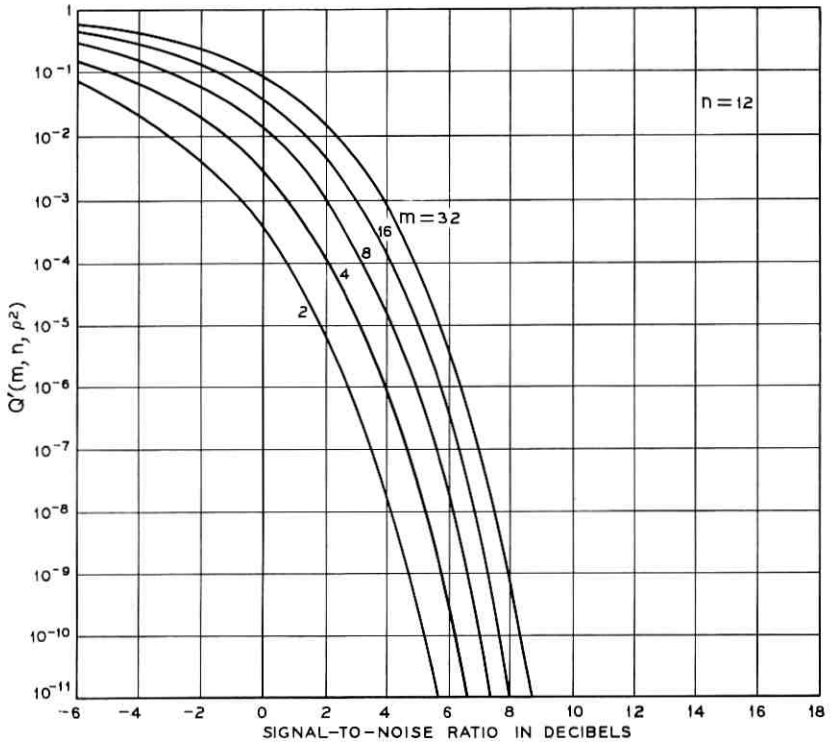


Fig. 12—Lower bound $Q'(m, n, \rho^2)$ for $n = 12$.

and that

$$P(2, n, \rho^2) = Q(2, n, \rho^2) = \tfrac{1}{2} \operatorname{erfc} [\rho(n/2)^{\frac{1}{2}}]. \tag{50}$$

Hence, we can write

$$L\left(\frac{\pi}{2}, n, \sigma^2\right) = \tfrac{1}{2} \operatorname{erfc} (\sigma). \tag{51}$$

Since

$$L(\theta, n, \sigma^2) = \int_\theta^\pi p_n(\lambda)\, d\lambda, \tag{52}$$

or

$$L\left(\theta, n, \rho^2 \frac{n}{2}\right) = \int_\theta^{\pi/2} p_n(\lambda)\, d\lambda + \int_{\pi/2}^\pi p_n(\lambda)\, d\lambda, \tag{53}$$

$$L\left(\theta, n, \rho^2 \frac{n}{2}\right) = T\left(\theta, n, \rho^2 \frac{n}{2}\right) + \tfrac{1}{2} \operatorname{erfc} [\rho(n/2)^{\frac{1}{2}}], \tag{54}$$

$$T(\theta, n, \sigma^2) = \frac{(n-1) \exp(-\sigma^2)}{2^{n/2}(\pi)^{\frac{1}{2}}\Gamma\left(\frac{n+1}{2}\right)} \int_0^\infty dr \int_\theta^{\pi/2} r^{n-1} \exp(-r^2/2)$$

$$\cdot \exp(r\sigma(2)^{\frac{1}{2}} \cos\lambda) \sin^{n-2}\lambda \, d\lambda. \tag{55}$$

Expanding $\exp(r\sigma(2)^{\frac{1}{2}} \cos\lambda)$ into a Taylor series and integrating term by term, we have

$$T(\theta, n, \sigma^2) = \frac{(n-1) \exp(-\sigma^2)}{2^{n/2}(\pi)^{\frac{1}{2}}\Gamma\left(\frac{n+1}{2}\right)} \cdot \sum_{\ell=1}^{\infty} \frac{[\sigma(2)^{\frac{1}{2}}]^\ell}{\ell!} \int_0^\infty r^{\ell+n-1} \exp - r^2/2_{dr}$$

$$\cdot \int_\theta^{\pi/2} \cos^\ell\lambda \sin^{n-2}\lambda \, d\lambda,$$

$$= \frac{(n-1) \exp(-\sigma^2)}{2(\pi)^{\frac{1}{2}}\Gamma\left(\frac{n+1}{2}\right)} \sum_{\ell=0}^{\infty} \frac{(2\sigma)^\ell}{\ell!} \Gamma\left(\frac{\ell+n}{2}\right)$$

$$\cdot \frac{1}{2} B\left(\frac{\ell+1}{2}, \frac{n-1}{2}\right) I_{\cos^2\theta}\left(\frac{\ell+1}{2}, \frac{n-1}{2}\right). \tag{56}$$

Equation (56) can be simplified to

$$T(\theta, n, \sigma^2) = \frac{\exp(-\sigma^2)}{2(\pi)^{\frac{1}{2}}} \sum_{\ell=0}^{\infty} \frac{(2\sigma)^\ell}{\ell!} \Gamma\left(\frac{\ell+1}{2}\right) I_{\cos^2\theta}\left(\frac{\ell+1}{2}, \frac{n-1}{2}\right). \tag{57}$$

Equations (54) and (57) yield

$$L(\theta, n, \sigma^2) = \tfrac{1}{2} \operatorname{erfc}(\sigma) + \frac{1}{2(\pi)^{\frac{1}{2}}} \exp(-\sigma^2)$$

$$\cdot \sum_{\ell=0}^{\infty} \frac{(2\sigma)^\ell}{\ell!} \Gamma\left(\frac{\ell+1}{2}\right) I_{\cos^2\theta}\left(\frac{\ell+1}{2}, \frac{n-1}{2}\right). \tag{58}$$

For m not too large and for large n, it can be shown that

$$\delta = \frac{\frac{\pi}{2} - \theta_{m,n}}{\pi/2} \tag{59}$$

is small. If δ is small, we can prove that the series given in equation

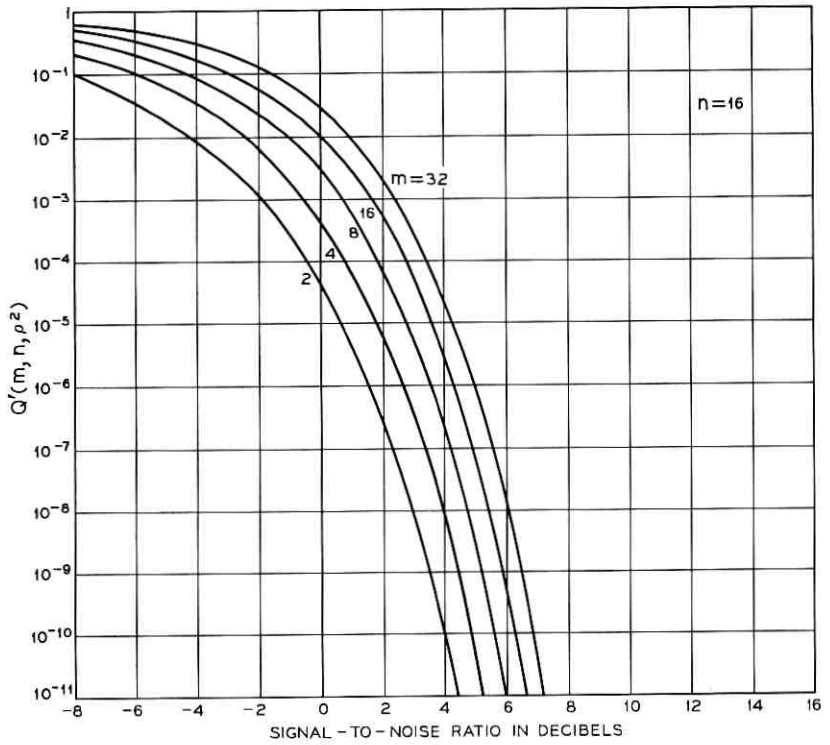


Fig. 13—Lower bound $Q'(m, n, \rho^2)$ for $n = 16$.

(58) converges rapidly and that we may alternatively calculate $Q(m, n, \rho^2)$ from equations (7), (9) and (58).

APPENDIX B

Evaluation of Distribution Function $L(\theta, 2, \sigma^2)$

Equations (10) and (14) can be shown to yield

$$L(\theta, 2, \sigma^2) = 1 - \int_{-\theta}^{\theta} p(\mu)\, d\mu, \tag{60}$$

where

$$p(\lambda) = \frac{1}{2\pi}\left[\exp(-\sigma^2) + \sigma(\pi)^{\frac{1}{2}} \cos\lambda \exp(-\sigma^2 \sin^2\lambda) \cdot\{1 + \operatorname{erf}(\sigma \cos\lambda)\}\right]. \tag{61}$$

Note[18-20] that $p(\lambda)$ is the probability density function of the phase angle λ, $0 \leqq \lambda < 2\pi$, of the sum of a sinusoidal carrier (of unit amplitude) and gaussian noise [of average power $1/(2\sigma^2)$].

We can also show[19] that

$$L(\theta, 2, \sigma^2) = 1 - \frac{\theta}{\pi} - \sum_{k=1}^{\infty} \frac{2h_k}{k} \sin k\theta, \tag{62}$$

where

$$h_{2\ell+1} = \frac{(\pi\sigma^2)^{\frac{1}{2}}}{2\pi} \exp(-\sigma^2/2)[I_\ell(\sigma^2/2) + I_{\ell+1}(\sigma^2/2)],$$

$$\ell = 0, 1, 2, \cdots \tag{63}$$

$$h_{2\ell} = \frac{1}{\pi} \sum_{n=-\infty}^{\infty} A_{2n+1} B_{2\ell-(2n+1)}, \qquad \ell = 1, 2, 3, \cdots; \tag{64}$$

$$A_{-2s-1} = A_{2s+1} = \frac{(\pi\sigma^2)^{\frac{1}{2}}}{2} \exp(-\sigma^2/2)[I_s(\sigma^2/2) + I_{s+1}(\sigma^2/2)],$$

$$s = 0, 1, 2, \cdots; \tag{65}$$

and

$$B_{-2p-1} = B_{2p+1} = (-1)^p \frac{1}{\pi} \frac{(\pi\sigma^2)^{\frac{1}{2}}}{2p+1} \exp(-\sigma^2/2)$$

$$\cdot [I_p(\sigma^2/2) + I_{p+1}(\sigma^2/2)], \qquad p = 0, 1, 2, \cdots. \tag{66}$$

$I_n(x)$ is the modified Bessel function of the first kind and of order n.

Since all h_k's can be calculated using either a set of tables or a digital computer and since the series given in equation (62) converges, we can calculate $L(\theta, 2, \sigma^2)$ for all σ and θ.

APPENDIX C

Evaluation of Upper and Lower Bounds

From equations (10) and (14), and Appendix B, we observe that $L(\theta, 2, \sigma^2)$ is the probability that the phase angle λ, $0 \leqq \lambda < 2\pi$, of a sinusoidal carrier of zero initial phase and unit amplitude lies outside the range $-\theta \leqq \lambda \leqq \theta$ when it is corrupted by random white gaussian noise of average power $1/2\sigma^2$.

When $\theta = \pi/2$, we can show[19,20] that

$$L(\theta, 2, \sigma^2) = \tfrac{1}{2} \operatorname{erfc}(\sigma), \qquad \theta = \pi/2. \tag{67}$$

When $\theta = \pi/4$, we can also show[18,20] that

$$L(\theta, 2, \sigma^2) = \text{erfc}\,[\sigma/(2)^{\frac{1}{2}}] - \tfrac{1}{4}\,\text{erfc}^2\,[\sigma/(2)^{\frac{1}{2}}]. \tag{68}$$

When $0 \leqq \theta < \pi/2$, let the sinusoidal carrier be represented by phasor OS in Fig. 14. Let x_u and x_v represent the in-phase and quadrature components of white gaussian noise corrupting the sinusoidal carrier. The quantity $L(\theta, 2, \sigma^2)$ is, therefore, given by the probability that the terminus of the vector **OT** lies in areas marked 1, 2 and 3.

We, therefore, have[4]

$$\text{erfc}\,(\sigma \sin\theta) - \frac{\tan\theta}{\pi}\exp(-\sigma^2)[1 - \pi^{\frac{1}{2}}\sigma \exp(\sigma^2)\,\text{erfc}\,(\sigma)]$$

$$\leqq L(\theta, 2, \sigma^2) < \text{erfc}\,(\sigma \sin\theta). \tag{69}$$

Also, since $L(\theta, 2, \sigma^2)$ is greater than the probability that the terminus of the vector **OT** lies in areas marked 1 and 2 (or 2 and 3), we can write

$$L(\theta, 2, \sigma^2) > \tfrac{1}{2}\,\text{erfc}\,(\sigma \sin\theta). \tag{70}$$

Combining equations (69) and (70), we get equation (35).

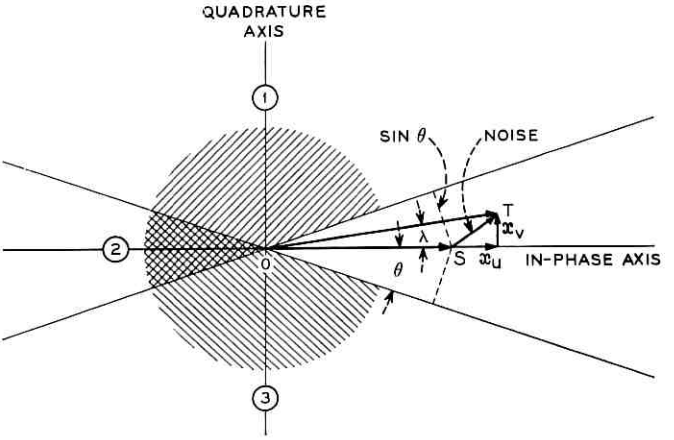


Fig. 14—Derivation of bounds to $L(\theta, 2, \sigma^2)$.

REFERENCES

1. Tillotson, L. C., "A Model of a Domestic Communication Satellite System," B.S.T.J., *47*, No. 10 (December 1968), pp. 2111–2136.
2. Tillotson, L. C., "Use of Frequencies Above 10 GHz for Common Carrier Applications," B.S.T.J., *48*, No. 6 (July–August 1969), pp. 1563–1576.
3. Ruthroff, C. L., and Tillotson, L. C., "Interference in a Dense Radio Network," B.S.T.J., *48*, No. 6 (July–August 1969), pp. 1727–1743.

4. Prabhu, V. K., "Error Rate Considerations for Coherent Phase-Shift Keyed Systems with Co-Channel Interference," B.S.T.J., *48*, No. 3 (March 1969), pp. 743–767.
5. Prabhu, V. K., and Enloe, L. H., "Interchannel Interference Considerations in Angle-Modulated Systems," B.S.T.J., *48*, No. 7 (September 1969), pp. 2333–2358.
6. Shannon, C. E., "Communication in the Presence of Noise," Proc. IRE, *37*, No. 1 (January 1949), pp. 1–12.
7. Shannon, C. E., "A Mathematical Theory of Communication," B.S.T.J., *27*, No. 3 (July 1948), pp. 379–423; *27*, No. 4 (October 1948), pp. 623–656.
8. Slepian, D., "The Threshold Effect in Modulation Systems that Expand Bandwidth," IRE Trans. on Information Theory, *IT-8*, No. 5 (September 1962), pp. 122–127.
9. Slepian, D., unpublished work.
10. Slepian, D., "Bounds on Communication," B.S.T.J., *42*, No. 3 (May 1963), pp. 681–707.
11. Landau, H. J., and Pollak, H. O., "Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty—III: The Dimension of the Space of Essentially Time- and Band-Limited Signals," B.S.T.J., *41*, No. 4 (July 1962), pp. 1295–1336.
12. Shannon, C. E., "Probability of Error for Optimal Codes in a Gaussian Channel," B.S.T.J., *38*, No. 3 (May 1959), pp. 611–656.
13. Pearson, K., *Tables of the Incomplete Beta-Function*, Cambridge University Press, London, England, 1934, pp. 1–15.
14. Pierce, J. R., unpublished work.
15. Rowe, H. E., unpublished work.
16. Mazo, J. E., Rowe, H. E., and Salz, J., "Rate Optimization for Digital Frequency Modulation," B.S.T.J., *48*, No. 9 (November 1969), pp. 3021–3030.
17. Mazo, J. E., and Salz, J., "Theory of Error Rates for Digital FM," B.S.T.J., *45*, No. 9 (November 1966), pp. 1511–1535.
18. Cahn, C. R., "Performance of Digital Phase-Modulation Communication Systems," IRE Trans. on Communication Systems, *CS-7*, No. 1 (May 1959), pp. 3–6.
19. Prabhu, V. K., "Error-Rate Considerations for Digital Phase-Modulation Systems," IEEE Trans. on Communication Technology, *COM-17*, No. 1 (February 1969), pp. 33–42.
20. Lucky, R. W., Salz, J., and Weldon, E. J., Jr., *Principles of Data Communication*, New York: McGraw-Hill, 1968, pp. 248–251.
21. Prabhu, V. K., unpublished work.

Transit-Time Variations in Line-of-Sight Tropospheric Propagation Paths

By D. A. GRAY

(Manuscript received December 12, 1969)

We present in this paper transit-time variations in line-of-sight propagation paths and systems operating at frequencies up to 30 GHz. We discuss variations due to both atmospheric changes (no precipitation) and rain and point out some relationships to PCM systems.

I. INTRODUCTION

In a recent paper[1], J. R. Pierce considered stable synchronization of large digital transmission networks, and pointed out that the realization of such a synchronized network calls for, among other things, more information concerning network transit-time variations. In this paper, we seek extreme values for these variations in line-of-sight propagation paths in order to provide some of this information. Estimates are given for (*i*) the maximum variation in transit-time, $\Delta\tau_{max}$, which one might expect over the period of a year, and (*ii*) the maximum time derivative, $\dot{\tau}_{max}$, which one might encounter. The estimated values are related to digital systems, with most specific examples given for a 500 megabit transmission rate. Variations due to changes in the atmosphere (no precipitation), and those due to rain are discussed separately. In the longer line-of-sight paths achieved in tandem systems, repeaters are assumed to be stable, that is, the concern herein is with atmospheric variations only. Delays associated with selective fading are not discussed, but they are believed not to exceed the given estimates of the maximum variations.

The transit-time τ is given by the familiar relationship

$$\tau = \frac{1}{c}\int_{P_1}^{P_2} n\, ds \tag{1}$$

where c is the velocity of light, n the medium refractive index, ds the differential path length, and the limits P_1 and P_2 represent the end points

of the path. If one assumes Δn_{max}, the maximum index of refraction change which is expected over a given period of time, and if one assumes that the changes in n over the entire path S are perfectly correlated and equal to Δn_{max}, then the corresponding maximum transit-time variation, $\Delta \tau_{max}$, is given by

$$\Delta \tau_{max} = \frac{1}{c} (\Delta n_{max}) S. \tag{2}$$

In a subsequent section, $\Delta \tau_{max}$ will be computed after estimating Δn_{max}. A similar set of assumptions concerning the time derivative $\dot{n}$ leads to an estimate of $\dot{\tau}_{max}$, namely,

$$\dot{\tau}_{max} = \frac{1}{c} \dot{n}_{max} S. \tag{3}$$

II. MAGNITUDE OF TRANSIT-TIME VARIATIONS IN THE ATMOSPHERE

Letting the superscript a denote the atmosphere, the estimate of Δn^a_{max} to be used in this section is based on data found in Bean and Dutton.[2] The data comprise eight years of point observations at six locations in the United States. The locations are listed in Table I where the maximum, minimum, and range Δn of the index of refraction are given in N units, with $N = (n - 1) \times 10^6$. A value which approximately strikes an average for the six locations listed will be used, namely, $\Delta n^a_{max} = 1.0 \times 10^{-4}$. Substituting in equation (2),

$$\Delta \tau^a_{max} = 0.333 \times 10^{-9} S \tag{4}$$

where S is in kilometers. Equation (4) is plotted in Fig. 1. Since the atmospheric index of refraction, n^a, may be regarded as independent of frequency up to 30 GHz[2], the curve in Fig. 1 may be considered applicable up to this frequency.

TABLE I—EXTREMES IN INDEX OF REFRACTION FOR SIX LOCATIONS IN THE UNITED STATES

Location	max N	min N	ΔN
Washington, D. C.	393	277	116
San Antonio, Texas	387	258	129
Bismarck, N. D.	368	260	108
Colorado Springs, Col.	307	214	93
Salt Lake City, Utah	323	230	93
Tatoosh Island, Wash.	354	292	62

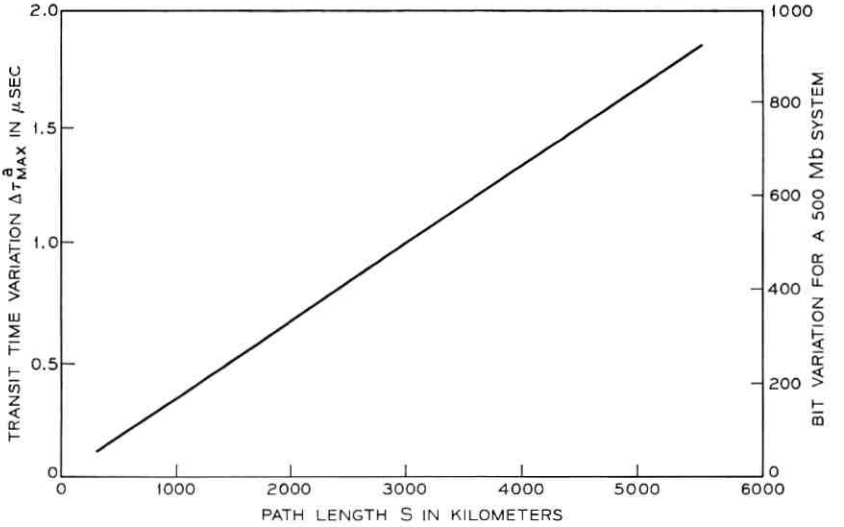


Fig. 1—Estimated maximum atmospheric (no precipitation) transit-time variations for long path length line-of-sight tropospheric communications systems.

For long paths, $\Delta\tau^a_{max}$ is overestimated because the assumption that Δn is perfectly correlated and equal over the entire path becomes unrealistic as S increases. However, the overestimation is not excessive because those components in Δn which primarily are due to seasonal and diurnal variations are highly correlated and combine to be of the order of 50 N units (approximately 30 N units for seasonal and 20 N units for diurnal). Thus, $\Delta\tau^a_{max}$ is conservative to within a factor of two for the continental United States. As an example, for a 3000-mile path in Fig. 1, $\Delta\tau^a_{max} = 1.6$ μs; certainly, it would be no less than 0.8 μs. In terms of a 500 Mb system, $\Delta\tau^a_{max} = 1.6$ μs is equivalent to an 800 bit variation.

For short path lengths, more detailed data are needed and these are in Fig. 2. One notes that, for short paths, a Δn^a which typifies a given region should be used instead of the Δn^a_{max} of 100 N units. Therefore, Fig. 2 shows $\Delta\tau_{max}$ versus path length for each of the six locations listed in Table I.

Now consider the problem of synchronizing two clocks separated by a distance S. For a digital system of pulse spacing T, two clocks may be considered synchronized if they are in phase to within a factor f of a pulse spacing. For the sake of argument in this discussion, we choose $f = 0.1$, and we draw on Fig. 2 the lines $f \cdot T$ for pulse rates of 50, 100,

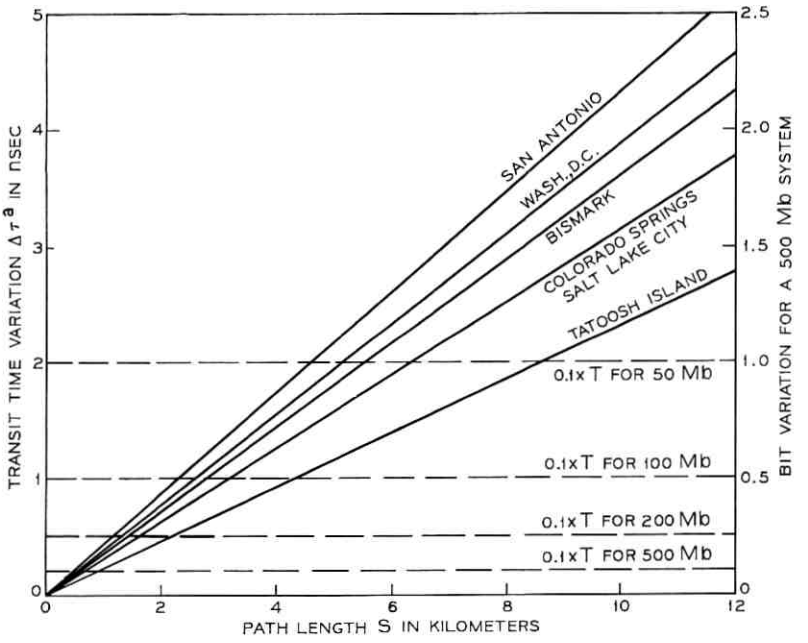


Fig. 2—Maximum atmospheric (no precipitation) transit-time variations for short paths. Variations are shown for six locations in the United States.

200, and 500 Mb. The intersections of the lines $f \cdot T$ with the curves for $\Delta\tau_{max}$ determine the maximum path lengths for which atmospheric transit-time variations may be neglected. For instance, these lengths are respectively 5, 2.5, 1.25 and 0.5 km for the 50, 100, 200, and 500 Mb rates when the region of concern is Washington, D. C. For some other criterion of synchronization, that is for some other f, these path lengths, of course, will be different.

From Fig. 2, it is evident that for any but the shortest path lengths, some compensation for refractive index changes must be incorporated in a digital system running on a universal clock. Figures 1 and 2, however, do not provide information on how rapidly one must compensate. This question is answered by first considering the time derivative of transit-time variations which are due to turbulence, and then by considering those due to the motion of synoptic scale air masses.

III. TIME DERIVATIVE OF TRANSIT-TIME VARIATIONS

Table II shows the rms values of the hourly transit-time variations[3,4] which are due to atmospheric turbulence. Because of the randomness

TABLE II—RMS HOURLY TRANSIT-TIME VARIATIONS IN THE TURBULENT ATMOSPHERE

RMS Delay (seconds)	Path Length (miles)	Frequency (GHz)
1.5×10^{-12}	2.25	10
4×10^{-12}	3.5	1
15×10^{-12}	10	1
22×10^{-12}	60	1

of turbulent motion, one would expect the rms values to be proportional to the square root of the path length S. Thus, on an hourly basis, the variations for a transcontinental path work out to be 0.2 ns, which is one tenth of the pulse spacing in a 500 Mb system. On the basis of a five minute interval, measurements show that rms variations are two orders of magnitude smaller than the hourly changes.[4] Thus it appears that compensation for turbulence-induced atmospheric transit-time variations can be made on the order of tens of minutes for transcontinental links, and even more slowly for shorter paths.

In contrast with the turbulence-related phenomena discussed above, the motion of synoptic-scale air masses brings about changes in index-of-refraction which are correlated over large regions, that is, for regions covering hundreds of kilometers, changes in n will be proportional to path length, S, not to $(S)^{\frac{1}{2}}$. For instance, an advancing cold front brings with it a decrease in temperature θ, a decrease in water vapor pressure e, an increase in total pressure P, and an accompanying change ΔN_F which is well correlated over the entire frontal advance.

For the purpose of discussion, a model front is shown in Fig. 3. The model front passes transversely across a microwave transmission path. The index of refraction change ΔN_F occurs over a distance S_Δ, and the

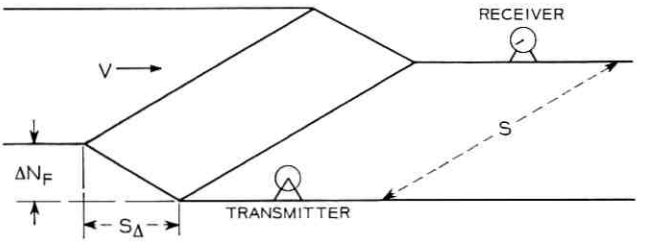


Fig. 3—Model of a frontal system moving across a line-of-sight transmission path.

front moves at velocity v. This model permits one to estimate the maximum rate of change of transit-time $\dot{\tau}_{\max}$.

Proceeding with the estimate of $\dot{\tau}_{\max}$, from Fig. 3,

$$\dot{n}_{\max} = 10^{-6} \frac{\Delta N_F}{S_\Delta} \cdot v \Big|_{\max} = 10^{-6} \frac{\Delta N_F}{t_\Delta} \Big|_{\max}$$

where $t_\Delta = S_\Delta / v$. Substituting in equation (3),

$$\dot{\tau}_{\max} = 0.33 \times 10^{-11} \frac{\Delta N_F}{t_\Delta} \Big|_{\max} S. \tag{5}$$

Estimates of the quantity $\Delta N_F / t_\Delta \mid_{\max}$ have been obtained from temperature, pressure, and relative humidity data continuously recorded at Crawford Hill since January 1, 1967. The records were examined for events of rapidly changing atmospheric conditions. N was computed for conditions just prior to and just after these events using the formula[2]

$$N = 77.6 \frac{P}{\theta} + 3.73 \times 10^5 \frac{e}{\theta^2} \tag{6}$$

where θ is expressed in kelvins, and P and e in millibars. ΔN_F was taken as the difference between the two computed values for each event. Table III gives the date, time, ΔN_F, t_Δ, and $\Delta N_F / t_\Delta$ for the largest events recorded. It is noted from the Table that $\Delta N_F / t_\Delta \mid_{\max} = .0788$. Rounding off to .08, and substituting in equation (5),

TABLE III—INDEX OF REFRACTION CHANGES ACCOMPANYING RAPIDLY VARYING ATMOSPHERIC CONDITIONS

Date	Time	ΔN_F	t_Δ (sec)	$\Delta N_F / t_\Delta$
2/28/67	10:30 PM	7.23	900	.00804
7/14/67	3:00 PM	3.94	600	.00657
10/3/67	9:00 PM	−13.67	300	.0456
2/17/68	2:00 PM	13.63	300	.0455
3/29/68	3:30 PM	15.93	900	.0177
4/30/68	6:00 PM	7.55	300	.0252
6/3/68	4:00 PM	13.75	600	.0229
7/2/68	8:00 PM	22.18	900	.0246
7/24/68	2:30 PM	19.84	300	.0662
8/7/68	2:30 AM	18.67	600	.0311
8/15/68	4:00 AM	−31.75	1200	.0264
8/17/68	3:30 AM	6.18	300	.0206
8/22/68	9:40 PM	−2.42	300	.00807
11/29/68	2:00 AM	3.67	180	.0204
12/5/68	2:00 PM	14.18	180	.0788
6/13/69	2:30 PM	10.12	600	.0169
6/24/69	5:30 PM	7.49	300	.0250

$$\dot{\tau}_{max} = 2.67 \times 10^{-13} S. \quad (7)$$

The maximum value S over which good correlation of N can occur will be taken as 500 km (this value will be discussed subsequently). From equation (7), $\dot{\tau}_{max} \cong 1.36 \times 10^{-10}$ sec/sec. In terms of a 500 Mb system, $\dot{\tau}_{max} \cong .0667$ b/s, or 1/16th of a bit per second.

The derived $\dot{\tau}_{max}$ is thought to be larger than would be encountered in practice. Complete parallelism (or coincidence), as shown in Fig. 3, becomes highly improbable when the path length S approaches 500 km. Primary reasons are: (*i*) a front and a transmission path can be oriented at angles ranging over a large fraction of π radians; (*ii*) nature will not produce fronts composed of straight line segments, but rather of curves. Since a line-of-sight propagation route is made up of straight line segments, it will be improbable that a front would coincide with it. Consequently, it appears that the choice of $S = 500$ km is extreme, and that $\dot{\tau}_{max} = 1/16$ b/s at 500 Mb is a good upper bound.

IV. TRANSIT-TIME VARIATIONS DUE TO RAIN

Letting the superscript r denote rain, then Δn^r is specified using computations of the medium refractive index for given rainfall rates under the assumption of a Laws & Parsons drop size distribution.[5] These computations include the index of refraction of the medium for 6, 16 and 30 GHz for the rain fall rates 0.25, 1.25. 2.5, 5.0, 12.5, 25.0, 50.0, 100.0, and 150.0 mm/hr. For rainfall rates exceeding 150 mm/hr, proportional scaling of the 150 mm/hr medium index will be used.

The nature of rainfall is such that higher fall rates are associated with smaller areas of coverage. For this reason, long and short path length transit-time variations will be treated differently. For path lengths of the order of hundreds of kilometers, we will assume that average-path rainrates of about 10 mm/hr can occur. Using a 10 mm/hr rainfall rate, Δn^r is equal to 0.83 N units at 6 and 16 GHz, and 0.67 N units at 30 GHz. The corresponding values of $\Delta\tau^r$ are plotted versus path length in Fig. 4. Comparison of Fig. 4 with Fig. 1 shows that the transit-time variations one expects from rain are more than an order of magnitude smaller than those expected from the atmosphere itself. This implies that as far as the dynamic range of compensation equipment in synchronized digital systems is concerned, atmospheric variations (other than rain) are the determining factor for long paths.

For short paths, rainfall rates much greater than 10 mm/hr often occur. Using results for New Jersey,[6] transit-time variations for short paths are computed using the refractive index values found in Table

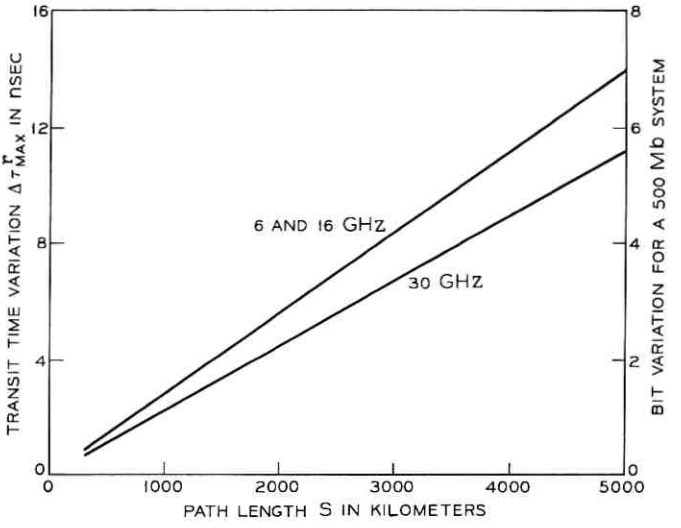


Fig. 4—Estimated maximum rain-caused transit-time variations for long path length line-of-sight communication systems.

IV; they are plotted in Fig. 5 with P as a parameter, where P equals the number of minutes per year the given variation will be exceeded. Variations to be exceeded 0.5 min/year and 5 min/year are shown, and as before, the rainfall variations are an order of magnitude smaller than the atmospheric variations (Fig. 2) for the corresponding path lengths.

When viewed in terms of the nature of convective showers (showers exhibiting high rainfall rates), Fig. 5 aids in estimating the upper bounds for the time derivative of rain-caused transit-time fluctuations. Convective showers can introduce large attenuation in a path rapidly with onsets of the order of tens of seconds. In Fig. 5 the intersection of the criterion line for a 500 Mb system with the 16 GHz curve for 0.5 min/yr occurs at a path length of 10 km. If the onset of the rain over this path were to occur in 10 seconds, then $\dot{\tau}_{max} = .01$ b/s, which is a factor of 6 smaller than the time derivative for the fronts considered previously.

IV. CONCLUSIONS

The maximum transit-time variations encountered in the troposphere are due primarily to changes in the gaseous atmosphere rather than rain, and may amount to 1.6 μs for a transcontinental path; this converts to 800 bits in a 500 Mb system. The maximum time derivative of the

TABLE IV—INDEX OF REFRACTION VARIATIONS EXCEEDED P MINUTES PER YEAR FOR VARIOUS PATH LENGTHS

(a) P = 5 min/yr.

Path Length (km)	Path Average Rainfall Rate (mm/hr)	Δn^r at 6 GHz (N units)	Δn^r at 16 GHz (N units)	Δn^r at 30 GHz (N units)
1.3	140	9.44	7.42	4.64
2.6	135	9.11	7.18	4.51
5.2	110	7.46	5.98	3.86
7.8	90	6.14	5.02	3.3
10.4	80	5.48	4.54	3.0

(b) P = 0.5 min/yr.

Path Length (km)	Path Average Rainfall Rate (mm/hr)	Δn^r at 6 GHz (N units)	Δn^r at 16 GHz (N units)	Δn^r at 30 GHz (N units)
1.3	190	12.8	10.1	6.21
2.6	175	11.78	9.3	5.72
5.2	150	10.1	7.9	4.9
7.8	130	8.78	6.94	4.38
10.4	110	7.46	5.98	3.86

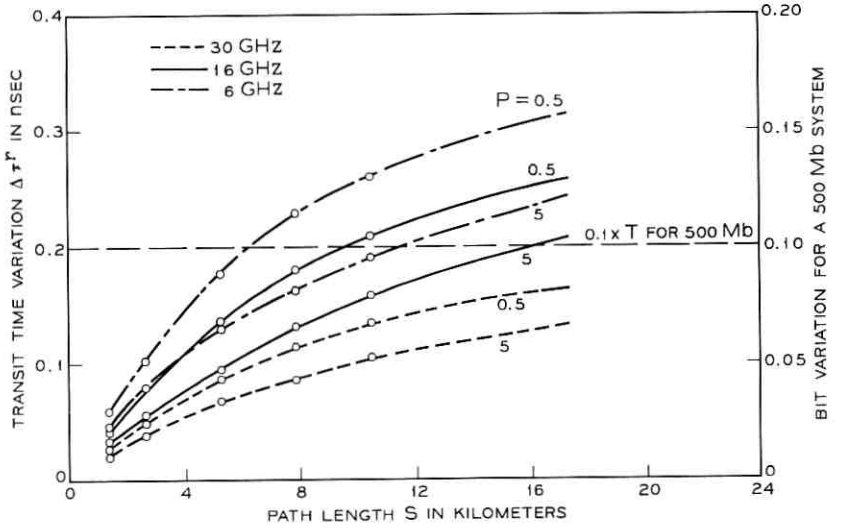


Fig. 5—Rain-caused transit-time variations for short path lengths. The parameter P equals the number of minutes/year the given variation will be exceeded.

transit-time variation is due to the motion of fronts, and has been derived herein: $\dot{\tau}_{max} = 1.36 \times 10^{-10}$ sec/sec which is equivalent to 1/16th of a bit at 500 Mb. Neither $\Delta\tau^a_{max}$ nor $\dot{\tau}_{max}$ appears so large as to prohibit the synchronization of digital systems transmitted over line-of-sight propagation paths.

V. ACKNOWLEDGMENTS

The author expresses his appreciation to J. R. Pierce for stimulating this study, to D. C. Hogg for discussions and to D. Setzer for computations.

REFERENCES

1. Pierce, J. R., "Synchronizing Digital Networks," B.S.T.J., *48*, No. 3 (March 1969), p. 615–636.
2. Bean, B. R., and Dutton, E. J., *Radio Meteorology*, Monograph 92, Washington, D. C.: National Bureau of Statistics, 1966, pp. 78, 409–414.
3. Beard, C. I., "Phase Quadrature Components of 10.4 GHz Scattered Field on a Short Tropospheric Path," Proc. IEEE, *56*, No. 8 (August 1968), pp. 1398–1399.
4. Herbstreit, J. W. and Thompson, M. C., "Measurement of the Phase of Radio Waves Received over Transmission Paths with Electrical Lengths Varying as a Result of Atmospheric Turbulence," Proc. IRE, *43*, No. 10 (October 1955), pp. 1391–1401.
5. Setzer, D., "Computed Transmission Characteristics of Rain at Microwave and Visible Frequencies," to be published in October 1970 B.S.T.J.
6. Hogg, D. C., "Statistics on Attenuation of Microwave by Intense Rain," B.S.T.J., *48*, No. 9 (November 1969), pp. 2949–2962.

Joint Optimization of Automatic Equalization and Carrier Acquisition for Digital Communication

By ROBERT W. CHANG

(Manuscript received January 22, 1970)

In this paper, we analyze single-sideband amplitude-modulation digital communication systems to develop a method for jointly and optimally setting the carrier phase and the automatic transversal equalizer of such systems. Mean-square equalization error is used as the performance criterion. We develop a simple receiver structure and study the convergence of the method. Exact locations of the stationary points in the parameter space are determined and the classifications of the stationary points are obtained. We show that the mean-square equalization error has only global minima and saddlepoints, but not local minima and maxima. Thus, the mean-square equalization error will converge to the absolute minimum by the proposed method, regardless of the initial settings of the parameters. A simple condition on the step sizes of the adjustments is also obtained which ensures the convergence of the process. Explicit formulas of the joint optimum parameter settings and of the corresponding minimum mean-square error are obtained. For illustration purposes, a single-sideband digital communication system using a five- or nine-tap transversal equalizer is simulated on a computer. Both theory and simulation show that the equalization error depends critically on the carrier phase when the number of equalizer taps is not large, and that the minimum equalization error can be obtained by using the proposed method.

I. INTRODUCTION

In single-sideband amplitude-modulation digital communication systems with transversal filter equalization,[1,2] the adjustment of the carrier phase is critical to the system's performance when the number of equalizer taps is not large. In this paper, a method is proposed for setting the carrier phase jointly with the automatic equalizer to minimize the mean-square equalization error.

We formulate a mathematical model of this study in Section II; in Sections III and IV, we analyze the system and develop a receiver structure. The problem of convergence is studied in Sections V and VI to determine if the equalization error will converge to the absolute minimum by the proposed method and whether such convergence depends on the initial settings of the parameters. We also consider step sizes of the adjustments. In Section VII, we derive explicit formulas for evaluating the system's performance. A voiceband data communication system is simulated on a computer to test the proposed method and the results are described in Section VIII.

II. MATHEMATICAL MODEL

As shown in Fig. 1, a single-sideband amplitude-modulation system is considered. When an impulse $\delta(t)$ is applied to the transmitter input, a signal $a(t)$ is received at the receiving filter output. The Fourier transform of $a(t)$ is denoted by $A(f)$. (The Fourier transform of a function will be consistently denoted by the appropriate capital letter.) It is assumed that $A(f)$ is band-limited between f_1 and f_2, that is,

$$A(f) \not\equiv 0, \quad \text{only for} \quad f_1 < |f| < f_2 . \tag{1}$$

The signal $a(t)$ is demodulated as shown in Fig. 1, where the demodulating carrier frequency is f_c. In single-sideband systems

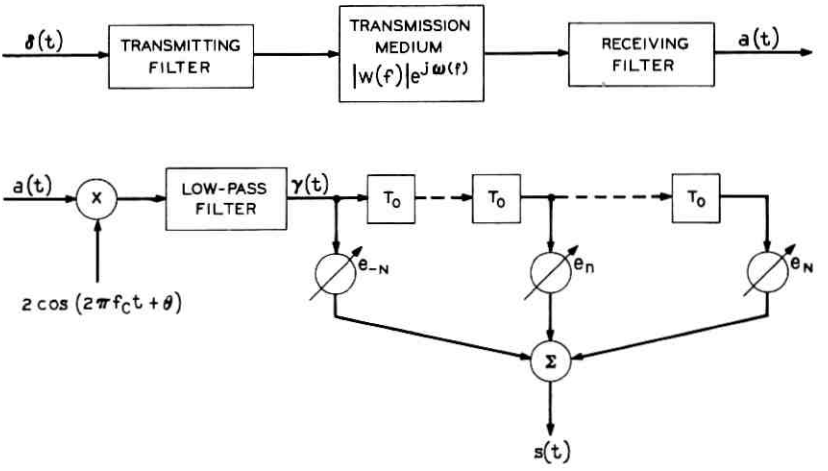


Fig. 1—An amplitude modulation system with coherent detection and transversal filter equalization.

$$f_c \leqq f_1 \tag{2}$$

or

$$f_c \geqq f_2 . \tag{3}$$

The demodulating carrier phase is denoted by θ. The demodulator output is a signal $\gamma(t)$ with Fourier transform $\Gamma(f)$. Since $A(f)$ is band-limited, $\Gamma(f)$ is also band-limited, that is,

$$\Gamma(f) = 0, \qquad |f| > f_0 \tag{4}$$

where f_0 is $f_2 - f_1$ when f_c is f_1 or f_2 (f_0 is larger than $f_2 - f_1$ if f_c is less than f_1 or larger than f_2).

As shown in Fig. 1, the channel is equalized by a conventional transversal equalizer consisting of $2N + 1$ taps with gains e_n, $n = -N$ to N, spaced at T_0-second intervals, where

$$T_0 = \frac{1}{2f_0} \quad \text{seconds.} \tag{5}$$

When an impulse $\delta(t)$ is applied at the transmitter input, the transversal equalizer output is $s(t)$. Clearly $s(t)$ is the overall impulse response of the system. The receiver parameters (that is, the demodulating carrier phase θ and the equalizer tap gains e_n, $n = -N$ to N) will be set to minimize the difference between the overall impulse response $s(t)$ and a desired impulse response $q(t)$. The familiar mean-square error criterion is used. That is, θ and e_n, $n = -N$ to N, will be jointly set to minimize the mean-square error

$$\mathcal{E}_0 = \int_{-\infty}^{\infty} [s(t) - q(t)]^2 \, dt. \tag{6}$$

For brevity, the tap gains e_n, $n = -N$ to N, will be abbreviated $\{e_n\}$ in the sequel.

III. ANALYSIS

In this section, we analyze the system to develop a receiver structure for jointly setting θ and $\{e_n\}$ by the method of steepest descent.

In analyzing carrier signals, it is most convenient to use Hilbert transform techniques. As is well known,[1] the demodulator output $\gamma(t)$ is related to the input $a(t)$ by

$$\gamma(t) = \cos(2\pi f_c t + \theta)a(t) + \sin(2\pi f_c t + \theta)\hat{a}(t) \tag{7}$$

where $\hat{a}(t)$ is the Hilbert transform of $a(t)$. {When dealing with lengthy

time functions, we shall sometimes use the sign $\mathcal{H}$, that is, $\mathcal{H}[a(t)] = \hat{a}(t)$.]

It is seen from Fig. 1 that

$$s(t) = \sum_{n=-N}^{N} e_n \gamma(t - nT_0 - NT_0). \tag{8}$$

We shall use the partial derivatives $\partial \mathcal{E}_0/\partial\theta$ and $\partial \mathcal{E}_0/\partial e_n$ in the method of steepest descent. Since $q(t)$ is independent of θ, we obtain from equation (6)

$$\frac{\partial \mathcal{E}_0}{\partial \theta} = \int_{-\infty}^{\infty} 2[s(t) - q(t)] \frac{\partial s(t)}{\partial \theta}\, dt. \tag{9}$$

In writing equation (9), the order of differentiation and integration has been exchanged [the underlying conditions for making such an exchange are easily satisfied by $s(t)$ and $q(t)$ encountered in communication systems]. Substituting equation (7) into equation (8) and taking the partial derivative give

$$\frac{\partial s(t)}{\partial \theta} = \sum_{n=-N}^{N} e_n \{-\sin [2\pi f_c(t - nT_0 - NT_0) + \theta] a(t - nT_0 - NT_0) + \cos [2\pi f_c(t - nT_0 - NT_0) + \theta] \hat{a}(t - nT_0 - NT_0)\}. \tag{10}$$

From equations (8) and (7),

$$\hat{s}(t) = \sum_{n=-N}^{N} e_n [\mathcal{H}\{\cos [2\pi f_c(t - nT_0 - NT_0) + \theta] a(t - nT_0 - NT_0)\} + \mathcal{H}\{\sin [2\pi f_c(t - nT_0 - NT_0) + \theta] \hat{a}(t - nT_0 - NT_0)\}]. \tag{11}$$

In single-sideband modulation, we have either inequality (2) or inequality (3). Let us consider inequality (2) first. When inequality (2) holds, the frequency spectrum of $a(t)$, $A(f)$, does not overlap the spectra of $\cos 2\pi f_c t$ and $\sin 2\pi f_c t$. Furthermore, $A(f)$ occupies a higher frequency band; therefore, equation (11) becomes

$$\mathcal{H}[\cos (2\pi f_c t + \theta) a(t)] = \cos (2\pi f_c t + \theta) \hat{a}(t),$$

and

$$\mathcal{H}[\sin (2\pi f_c t + \theta) \hat{a}(t)] = -\sin (2\pi f_c t + \theta) a(t), \qquad f_c \leqq f_1 .$$

Substituting the above into equation (11) gives

$$\hat{s}(t) = \sum_{n=-N}^{N} e_n\{\cos\,[2\pi f_c(t - nT_0 - NT_0) + \theta]\hat{a}(t - nT_0 - NT_0) - \sin\,[2\pi f_c(t - nT_0 - NT_0) + \theta]a(t - nT_0 - NT_0)\}, \qquad f_c \leqq f_1 . \tag{12}$$

Comparing equation (12) with equation (10) shows that

$$\hat{s}(t) = \frac{\partial s(t)}{\partial \theta}, \qquad f_c \leqq f_1 . \tag{13}$$

Substituting equation (13) into equation (9) gives

$$\frac{\partial \mathcal{E}_0}{\partial \theta} = \int_{-\infty}^{\infty} 2[s(t) - q(t)]\hat{s}(t)\,dt, \qquad f_c \leqq f_1 . \tag{14}$$

Since a function and its Hilbert transform are orthogonal, equation (14) reduces to

$$\frac{\partial \mathcal{E}_0}{\partial \theta} = -2\int_{-\infty}^{\infty} q(t)\hat{s}(t)\,dt, \qquad f_c \leqq f_1 . \tag{15}$$

Thus, $\partial\mathcal{E}_0/\partial\theta$ can be generated by correlating $q(t)$ with $\hat{s}(t)$. Note that the transversal equalizer output will be $\hat{s}(t)$ instead of $s(t)$ if the demodulating carrier $\cos\,(2\pi f_c t + \theta)$ is replaced by $\cos\,[2\pi f_c t + \theta + \pi/2]$. However, even though $\hat{s}(t)$ can be generated in this simple fashion, we prefer not to generate $\hat{s}(t)$ because the system must be used instead to generate $s(t)$ to compute the other partial derivatives $\partial\mathcal{E}_0/\partial e_n$. Therefore, we convert equation (15) into the form

$$\frac{\partial \mathcal{E}_0}{\partial \theta} = 2\int_{-\infty}^{\infty} \hat{q}(t)s(t)\,dt, \qquad f_c \leqq f_1 . \tag{16}$$

This step can be verified by Parseval's theorem. Now we need only to correlate $s(t)$ with an easily generated $\hat{q}(t)$ to obtain $\partial\mathcal{E}_0/\partial\theta$.

The above is for the case $f_c \leqq f_1$. In the other case, $f_c \geqq f_2$, the frequency spectrum of $a(t)$ occupies a frequency band lower than f_c; therefore, the two equations above equation (12) should be rewritten as

$$\mathcal{H}[\cos\,(2\pi f_c t + \theta)a(t)] = \sin\,(2\pi f_c t + \theta)a(t),$$

$$\mathcal{H}[\sin\,(2\pi f_c t + \theta)\hat{a}(t)] = -\cos\,(2\pi f_c t + \theta)\hat{a}(t), \qquad f_c \geqq f_2 .$$

Repeating the steps from equation (12) to equation (16), we get

$$\frac{\partial \mathcal{E}_0}{\partial \theta} = -2\int_{-\infty}^{\infty} s(t)\hat{q}(t)\,dt, \qquad f_c \geqq f_2 . \tag{17}$$

Note from equations (17) and (16) that the sign of the correlator output must be reversed when one shifts the carrier frequency from one side of $A(f)$ to the other side.

About the equalizer tap gains, it is seen from equations (6) and (8) that

$$\frac{\partial \mathcal{E}_0}{\partial e_n} = \int_{-\infty}^{\infty} 2[s(t) - q(t)] \frac{\partial s(t)}{\partial e_n} dt,$$

$$= \int_{-\infty}^{\infty} 2[s(t) - q(t)]\gamma(t - nT_0 - NT_0)\, dt, \qquad n = -N \text{ to } N. \tag{18}$$

Thus, $\partial \mathcal{E}_0/\partial e_n$ can be generated by correlating the error signal $[s(t) - q(t)]$ with the output $\gamma(t - nT_0 - NT_0)$ of the nth tap. This is the concept introduced in Ref. 3 where the problem of setting the tap gains $\{e_n\}$ was considered.

IV. RECEIVER STRUCTURE

It has been shown in the previous section that $\partial \mathcal{E}_0/\partial \theta$ can be easily generated simultaneously with $\partial \mathcal{E}_0/\partial e_n$, $n = -N$ to N. Therefore, the method of steepest descent can be used to adjust simultaneously θ and $\{e_n\}$. In the training period prior to data transmission, isolated test pulses are transmitted. For instance, $\delta(t)$ in Fig. 1 may be one of the test pulses. The transmission of $\delta(t)$ generates a signal $s(t)$ at the equalizer output. From $s(t)$ the partial derivatives $\partial \mathcal{E}_0/\partial \theta$ and $\partial \mathcal{E}_0/\partial e_n$, $n = -N$ to N, are computed. The parameters θ and $\{e_n\}$ are then changed by amounts proportional to the partial derivatives, that is,

$$\theta_{\text{new}} = \theta_{\text{old}} - \alpha \frac{\partial \mathcal{E}_0}{\partial \theta},$$

$$(e_n)_{\text{new}} = (e_n)_{\text{old}} - \beta \frac{\partial \mathcal{E}_0}{\partial e_n}, \qquad n = -N \text{ to } N,$$

where α and β are positive proportional constants which may vary from one adjustment to another (the choice of their values will be considered in Section VI). After the changes are all made, another test pulse is transmitted and the process is repeated. The process is terminated after a prefixed number of test pulses.

The receiver structure is shown in Fig. 2. The partial derivatives $\partial \mathcal{E}_0/\partial \theta$ and $\partial \mathcal{E}_0/\partial e_n$ are computed according to equations (17) and (18), respectively ($f_c \geqq f_2$ is assumed). If $f_c \leqq f_1$, the correlator output in Fig. 2 will be $-\partial \mathcal{E}_0/\partial \theta$.

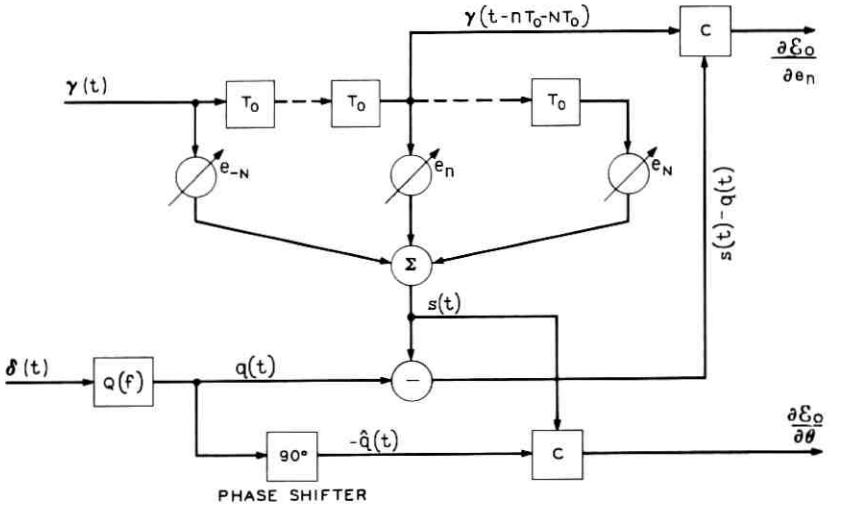


Fig. 2—Block diagram of the joint method. (C in the block denotes correlator.)

V. FURTHER ANALYSIS

We now analyze the system to answer the questions: (*i*) Will the mean-square error converge to the absolute minimum by the proposed adjustments? (*ii*) Is the convergence ensured regardless of how the parameters θ and $\{e_n\}$ are set prior to the starting of the process? (*iii*) What is the minimum mean-square error that can be obtained by the proposed method? To answer these questions, it is necessary to locate the stationary points of $\mathcal{E}_0$, to distinguish between various types of stationary points, to determine conditions under which $\mathcal{E}_0$ will converge to a global minimum, to derive explicit solutions of $\mathcal{E}_0$ when the parameters are set jointly or independently, and to obtain numerical data by simulating a real typical channel. In this section, we determine the location of the stationary points and distinguish between various types of stationary points.

It is noted from equations (16) and (17) that $\partial\mathcal{E}_0/\partial\theta$ changes sign when the carrier frequency f_c is shifted from one side of $A(f)$ (that is, $f_c \leqq f_1$) to the other side ($f_c \geqq f_2$). To avoid this complication, a quantity ρ defined by

$$\begin{aligned} \rho &= \theta, \quad \text{when} \quad f_c \leqq f_1 , \\ &= -\theta, \quad \text{when} \quad f_c \geqq f_2 , \end{aligned} \tag{19}$$

will be used instead of θ in the following. By this change, the results obtained in the sequel will hold for both $f_c \leqq f_1$ and $f_c \geqq f_2$. Hence, the position of f_c will not be identified further.

Since it is more convenient to use matrix notations in the following, time samples will be used. Let

$$\begin{aligned} s_k &= s(kT_0 + NT_0), \\ q_k &= q(kT_0 + NT_0), \\ \gamma_k &= \gamma(-kT_0). \end{aligned} \tag{20}$$

Then equation (6) becomes

$$\mathcal{E}_0 = \frac{1}{2f_0}\,\mathcal{E}$$

where

$$\mathcal{E} = \sum_{k=-\infty}^{\infty} (s_k - q_k)^2. \tag{21}$$

Since f_0 is a constant, minimizing $\mathcal{E}$ minimizes $\mathcal{E}_0$.

It can be easily shown that equation (21) can be written in the following matrix form

$$\mathcal{E} = \mathbf{E}'\mathbf{y}\mathbf{E} - 2\mathbf{E}'\mathbf{v} + \sum_{k=-\infty}^{\infty} q_k^2 \tag{22}$$

where the prime denotes transpose, the vector $\mathbf{E}$ and $\mathbf{v}$ and the matrix $\mathbf{y}$ are defined by

$$\mathbf{E} = \begin{bmatrix} e_{-N} \\ e_{-N+1} \\ \vdots \\ e_N \end{bmatrix}, \qquad \mathbf{v} = \begin{bmatrix} v_{-N} \\ v_{-N+1} \\ \vdots \\ v_N \end{bmatrix}, \tag{23}$$

$$\mathbf{y} = \begin{bmatrix} y_{-N,-N} & \cdots & y_{-N,N} \\ \vdots & & \vdots \\ y_{N,-N} & \cdots & y_{N,N} \end{bmatrix}, \tag{24}$$

$$y_{h,i} = \sum_{k=-\infty}^{\infty} \gamma_{h-k}\gamma_{i-k}\,, \tag{25}$$

$$v_n = \sum_{k=-\infty}^{\infty} q_k\gamma_{n-k}\,. \tag{26}$$

Clearly, $\mathbf{E}$ and $\mathbf{v}$ are $(2N + 1) \times 1$ vectors, and $\mathbf{y}$ is a $(2N + 1) \times (2N + 1)$ matrix. One can easily show that $\mathbf{y}$ is nonsingular.

We now evaluate the partial derivatives $\partial \mathcal{E}/\partial \rho$ and $\partial \mathcal{E}/\partial e_n$, $n = -N$ to N, to determine the optimum parameter settings. From equation (22),

$$\frac{\partial \mathcal{E}}{\partial \rho} = \frac{\partial}{\partial \rho} \mathbf{E}'\mathbf{y}\mathbf{E} - 2 \frac{\partial}{\partial \rho} \mathbf{E}'\mathbf{v}. \tag{27}$$

To evaluate the terms in equation (27), we note from sampling theorem that equation (25) can be written as

$$y_{h,i} = 2f_0 \int_{-\infty}^{\infty} \gamma(t)\gamma[t - (h - i)T_0]\, dt. \tag{28}$$

By Parseval's theorem, equation (28) becomes

$$y_{h,i} = 2f_0 \int_{-\infty}^{\infty} |\Gamma(f)|^2 \exp[-j2\pi f(h - i)T_0]\, df. \tag{29}$$

For single-sideband systems, the demodulating carrier phase ρ appears only in the phase characteristic of $\Gamma(f)$, while the amplitude characteristic $|\Gamma(f)|$ of $\Gamma(f)$ is independent of ρ. Therefore, from equation (29), $y_{h,i}$ is independent of ρ and

$$\frac{\partial}{\partial \rho} \mathbf{E}'\mathbf{y}\mathbf{E} = 0. \tag{30}$$

This greatly simplifies the results. (It is important to note that this simplification is not possible for double-sideband and vestigial-sideband systems because ρ appears in $|\Gamma(f)|$ in such systems.) Substituting equation (30) into equation (27) gives

$$\frac{\partial \mathcal{E}}{\partial \rho} = -2 \frac{\partial}{\partial \rho} \mathbf{E}'\mathbf{v}. \tag{31}$$

To evaluate equation (31), we convert the elements v_n of $\mathbf{v}$ into explicit functions of ρ. For single-sideband systems, $\Gamma(f)$ can be decomposed into the form

$$\begin{aligned} \Gamma(f) &= H(f) \exp(-j\rho), \qquad f \geqq 0; \\ &= H(f) \exp(j\rho), \qquad f \leqq 0; \end{aligned} \tag{32}$$

where $H(f)$ is independent of ρ.* From equations (26), (20), (4), and

* $H(f)$ is the Fourier transform of $\gamma(t)$ when $\rho = 0$. The amplitude and phase of $H(f)$ will be denoted by $|H(f)|$ and $\eta(f)$, respectively, that is, $H(f) = |H(f)| \exp[j\eta(f)]$.

(32), it can be shown that

$$v_n = \mu_n \exp(-j\rho) + \nu_n \exp(j\rho), \qquad n = -N \text{ to } N \tag{33}$$

where

$$\mu_n = \sum_{k=-\infty}^{\infty} q_k \int_0^{f_0} H(f) \exp[j2\pi f(k-n)T_0]\, df \tag{34}$$

and

$$\nu_n = \sum_{k=-\infty}^{\infty} q_k \int_{-f_0}^{0} H(f) \exp[j2\pi f(k-n)T_0]\, df. \tag{35}$$

Note that μ_n and ν_n are independent of ρ. From equation (33), we can evaluate the term $\partial/\partial\rho \mathbf{E}'\mathbf{v}$ in equation (31) to obtain

$$\frac{\partial \mathcal{E}}{\partial \rho} = 2j[\exp(-j\rho)\mathbf{E}'\boldsymbol{\mu} - \exp(j\rho)\mathbf{E}'\boldsymbol{\nu}] \tag{36}$$

where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{-N} \\ \mu_{-N+1} \\ \vdots \\ \mu_N \end{bmatrix}, \qquad \boldsymbol{\nu} = \begin{bmatrix} \nu_{-N} \\ \nu_{-N+1} \\ \vdots \\ \nu_N \end{bmatrix}. \tag{37}$$

The above is for $\partial\mathcal{E}/\partial\rho$. The other partial derivatives can be obtained from equation (22) as

$$\frac{\partial \mathcal{E}}{\partial \mathbf{E}} = 2\mathbf{y}\mathbf{E} - 2\mathbf{v}. \tag{38}$$

A necessary condition for a specific ρ and $\mathbf{E}$ to be jointly optimum (that is, to jointly minimize $\mathcal{E}$) is that they satisfy

$$\frac{\partial \mathcal{E}}{\partial \rho} = 0 \tag{39}$$

and

$$\frac{\partial \mathcal{E}}{\partial \mathbf{E}} = \mathbf{0}. \tag{40}$$

There are special cases where the optimum setting of ρ is arbitrary (for instance, if an infinite length tapped delay line is used, the taps can always be adjusted to reduce the mean-square error $\mathcal{E}$ to zero

regardless of how ρ is set). We shall not consider such special cases here. This implies, as can be shown from equations (36), (38), (39), and (40), that the special cases where $\mathbf{u}'\mathbf{y}^{-1}\mathbf{u}$ is zero are disregarded.

From equation (38), equation (40) can be written as

$$\mathbf{E} = \mathbf{y}^{-1}\mathbf{v}. \tag{41}$$

From equations (36) and (41), equation (39) can be written as

$$\exp(-j\rho)\mathbf{u}'\mathbf{y}^{-1}\mathbf{v} - \exp(j\rho)\mathbf{v}'\mathbf{y}^{-1}\mathbf{v} = 0. \tag{42}$$

One can show from equation (33) that equation (42) is equivalent to

$$\frac{\mathbf{u}'\mathbf{y}^{-1}\mathbf{u}}{\mathbf{v}'\mathbf{y}^{-1}\mathbf{v}} = \exp(j4\rho). \tag{43}$$

It can be shown from equations (34) and (35) that

$$\operatorname{Re}[\mathbf{u}'\mathbf{y}^{-1}\mathbf{u}] = \operatorname{Re}[\mathbf{v}'y^{-1}\mathbf{v}],$$

and

$$\operatorname{Im}[\mathbf{u}'\mathbf{y}^{-1}\mathbf{u}] = -\operatorname{Im}[\mathbf{v}'\mathbf{y}^{-1}\mathbf{v}], \tag{44}$$

where Re and Im denote real and imaginary parts of the complex number, respectively. From equation (44), one can show that equation (43) is satisfied if and only if

$$\rho = m_0\frac{\pi}{2} + \tfrac{1}{2}\tan^{-1}\frac{\operatorname{Im}[\mathbf{u}'\mathbf{y}^{-1}\mathbf{u}]}{\operatorname{Re}[\mathbf{u}'\mathbf{y}^{-1}\mathbf{u}]} \tag{45}$$

where m_0 can be any integer including zero. Substituting equation (45) into equation (33) and substituting the resulting equation into equation (41) give

$$\mathbf{E} = \exp\left[-j\left(m_0\frac{\pi}{2} + \tfrac{1}{2}\tan^{-1}\frac{\operatorname{Im}[\mathbf{u}'\mathbf{y}^{-1}\mathbf{u}]}{\operatorname{Re}[\mathbf{u}'\mathbf{y}^{-1}\mathbf{u}]}\right)\right]\mathbf{y}^{-1}\mathbf{u}$$
$$+ \exp\left[j\left(m_0\frac{\pi}{2} + \tfrac{1}{2}\tan^{-1}\frac{\operatorname{Im}[\mathbf{u}'\mathbf{y}^{-1}\mathbf{u}]}{\operatorname{Re}[\mathbf{u}'\mathbf{y}^{-1}\mathbf{u}]}\right)\right]\mathbf{y}^{-1}\mathbf{v}. \tag{46}$$

The value of ρ and $\mathbf{E}$ which satisfies the necessary conditions (39) and (40) is given by equations (45) and (46), respectively. It is clearly seen from equations (45) and (46) that, since m_0 can be any integer, there is more than one set of solutions of ρ and $\mathbf{E}$ from conditions (39) and (40). In the following, we determine which of these solutions actually minimizes $\mathcal{E}_0$.

In conventional terminology,[4] each solution of conditions (39) and

(40) is a stationary point of $\mathcal{E}_0$. We shall determine which of the stationary points is a global minimum.

We first identify the minima, maxima, and saddlepoints. As is well known,[4] a sufficient condition for a stationary point to be a local minimum is that the quadratic form

$$F = \mathbf{h}'\mathbf{Q}\mathbf{h} \tag{47}$$

be positive-definite at that stationary point, where $\mathbf{Q}$, known as the Hessian matrix, is

$$\mathbf{Q} = \begin{bmatrix} \frac{\partial^2 \mathcal{E}_0}{\partial \rho^2} & \frac{\partial^2 \mathcal{E}_0}{\partial \rho\, \partial e_{-N}} & \cdots & \frac{\partial^2 \mathcal{E}_0}{\partial \rho\, \partial e_N} \\ \frac{\partial^2 \mathcal{E}_0}{\partial e_{-N}\, \partial \rho} & \frac{\partial^2 \mathcal{E}_0}{\partial e_{-N}^2} & \cdots & \frac{\partial^2 \mathcal{E}_0}{\partial e_{-N}\, \partial e_N} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 \mathcal{E}_0}{\partial e_N\, \partial \rho} & \frac{\partial^2 \mathcal{E}_0}{\partial e_N\, \partial e_{-N}} & \cdots & \frac{\partial^2 \mathcal{E}_0}{\partial e_N^2} \end{bmatrix}$$

and

$$\mathbf{h}' = [h_1 \quad h_2 \cdots h_{(2N+2)}].$$

A stationary point is a local maximum if F is negative-definite at that point. At a saddlepoint, F is indefinite. To evaluate F, note from equations (16), (17), and (19) that

$$\frac{\partial \mathcal{E}_0}{\partial \rho} = 2 \int_{-\infty}^{\infty} \hat{q}(t) s(t)\, dt \tag{48}$$

and

$$\frac{\partial s(t)}{\partial \rho} = \hat{s}(t). \tag{49}$$

From equations (48), (49), (8), and (18), we get

$$\frac{\partial^2 \mathcal{E}_0}{\partial \rho^2} = 2 \int_{-\infty}^{\infty} q(t) s(t)\, dt; \tag{50}$$

$$\frac{\partial^2 \mathcal{E}_0}{\partial \rho\, \partial e_n} = 2 \int_{-\infty}^{\infty} \gamma(t - nT_0 - NT_0) \hat{q}(t)\, dt, \qquad n = -N \text{ to } N; \tag{51}$$

and

$$\frac{\partial^2 \mathcal{E}_0}{\partial e_i\, \partial e_j} = 2 \int_{-\infty}^{\infty} \gamma(t - iT_0 - NT_0) \gamma(t - jT_0 - NT_0)\, dt, \qquad i, j = -N \text{ to } N. \tag{52}$$

Substituting equations (50), (51), and (52) into equation (47) and rearranging, we obtain after some steps

$$F = 2\int_{-\infty}^{\infty}\left[h_1\hat{q}(t) + \sum_{n=-N}^{N} h_{2+n+N}\gamma(t - nT_0 - NT_0)\right]^2 dt + 2h_1^2\int_{-\infty}^{\infty} q(t)s(t)\,dt - 2h_1^2\int_{-\infty}^{\infty} q^2(t)\,dt. \tag{53}$$

Now we may substitute equations (45) and (46) into equation (53) and evaluate the resulting expression over all possible h_1 to $h_{(2N+2)}$ to determine whether F is positive-definite, negative-definite, or indefinite at a given stationary point. This determines if that point is a local minimum, a local maximum, or a saddlepoint. While these steps are important in the analysis, they are rather complex and are therefore given in Appendix A. It is also shown in that appendix that all the local minima are equal and hence all are global minima. The results in Appendix A are summarized in the following proposition.

Proposition 1. $\mathcal{E}_0$ has a global minimum when ρ and $\mathbf{E}$ are given, respectively, by equations (45) and (46), with m_0 in these equation being an even integer. $\mathcal{E}_0$ has a saddlepoint when ρ and $\mathbf{E}$ are given, respectively, by equations (45) and (46), with m_0 in these equations being an odd integer.

It is seen from this proposition that there is an infinite number of global minima, each one corresponding to an even integer m_0. The distance in ρ between two adjacent global minima is therefore π. There is also an infinite number of saddlepoints, each one corresponding to an odd integer m_0. The distance in ρ between two adjacent saddlepoints is also π. The distance in ρ between a saddlepoint and its adjacent global minimum is $\pi/2$. It is instructive to illustrate these with an example and a figure. Consider a transversal equalizer with only one tap e_0 (such a single tap serves as an automatic gain control and the problem is to jointly set the automatic gain control and carrier phase to minimize the mean-square error). For simplicity, suppose that the term $\tan^{-1}$ [Im $(\underline{\mu}'\mathbf{y}^{-1}\underline{\mu})$]/[Re $(\underline{\mu}'\mathbf{y}^{-1}\underline{\mu})$] in equation (45) turned out to be zero. Then from the proposition $\mathcal{E}_0$ has global minima at $\rho = 0, \pm\pi, \pm 2\pi, \cdots$, and $\mathcal{E}_0$ has saddlepoints at $\rho = \pm\pi/2, \pm 3\pi/2, \cdots$. The global minima at $\rho = 0$ and π are illustrated by points 1 and 3 in Fig. 3 and the saddlepoint at $\rho = \pi/2$ is illustrated by point 2.

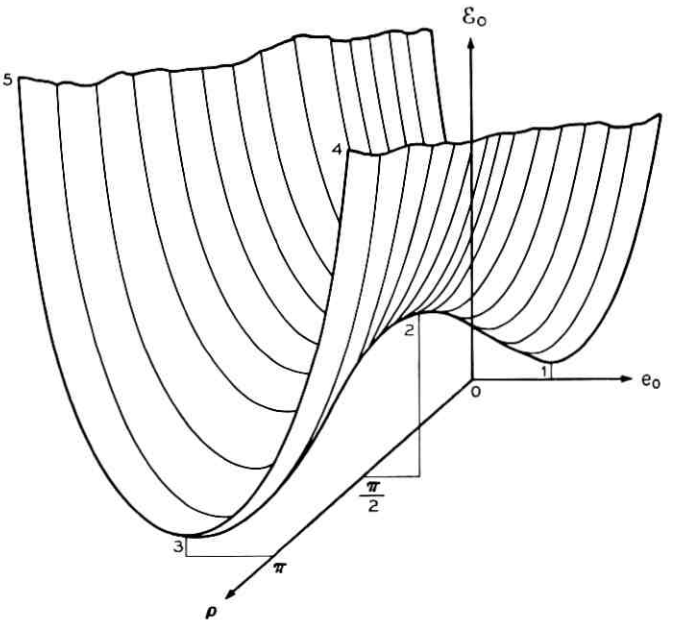


Fig. 3—An example illustrating the global minima and saddlepoints of $\mathcal{E}_0$.

The curved surface in the figure illustrates the variation of $\mathcal{E}_0$ with ρ and e_0. For instance, when ρ is fixed at π, $\mathcal{E}_0$ varies with e_0 as shown by the convex curve passing through points 4, 3, and 5. (For fixed ρ, $\mathcal{E}_0$ is a convex function of the tap gains e_{-N} to e_N.) As can be seen, point 2 is a saddlepoint because varying e_0 away from point 2 with ρ constant increases $\mathcal{E}_0$, while varying ρ away from point 2 with e_0 constant decreases $\mathcal{E}_0$.

To summarize, in this section we have located the global minima and saddlepoints and proved that there is no local minimum or maximum. There is also no valley[5] since the global minima are all distinct. These results will be used in the next section for the study of convergence and in Sections VII and VIII for the computation and comparison of performances.

VI. A CONDITION OF CONVERGENCE

As described in Section IV, after a test pulse is transmitted and the partial derivatives are computed, the carrier phase and the tap gains are adjusted according to the equation

$$\theta_{\text{new}} = \theta_{\text{old}} - \alpha \frac{\partial \mathcal{E}_0}{\partial \theta}, \tag{54}$$

$$(e_n)_{\text{new}} = (e_n)_{\text{old}} - \beta \frac{\partial \mathcal{E}_0}{\partial e_n}, \tag{55}$$

where α and β are positive proportional constants which may vary from one adjustment to another. As is well known,[5] the process always converges when α and β are sufficiently small, but may not converge if α and β are too large. In this section, we derive a condition on α and β which ensures the convergence of the process.

To facilitate analysis, we shall, as in Section V, replace θ with ρ, $\mathcal{E}_0$ with $\mathcal{E}$, the tap gains e_{-N} to e_N with the vector $\mathbf{E}$, and the partial derivatives $\partial\mathcal{E}/\partial e_{-N}$ to $\partial\mathcal{E}/\partial e_N$ with the vector $\partial\mathcal{E}/\partial\mathbf{E}$. The adjustment made after the kth test pulse will be called the kth adjustment ($k = 1, 2, \cdots$). The values of ρ, $\mathbf{E}$, $\mathcal{E}$, $\partial\mathcal{E}/\partial\rho$, and $\partial\mathcal{E}/\partial\mathbf{E}$ prior to the kth adjustment will be denoted by ρ_k, $\mathbf{E}_k$, $\mathcal{E}(\rho_k, \mathbf{E}_k)$, $(\partial/\partial\rho)\mathcal{E}(\rho_k, \mathbf{E}_k)$ and $(\partial/\partial\mathbf{E})\mathcal{E}(\rho_k, \mathbf{E}_k)$, respectively. The α and β used in the kth adjustment will be called α_k and β_k, respectively (note that $\alpha_k > 0$ and $\beta_k > 0$ for all k so that the adjustments will be made in the negative gradient direction). The values of ρ, $\mathbf{E}$, $\mathcal{E}$, $\partial\mathcal{E}/\partial\rho$, and $\partial\mathcal{E}/\partial\mathbf{E}$ after the kth adjustment will be denoted by ρ_{k+1}, $\mathbf{E}_{k+1}$, $\mathcal{E}(\rho_{k+1}, \mathbf{E}_{k+1})$, $(\partial/\partial\rho)\,\mathcal{E}(\rho_{k+1}, \mathbf{E}_{k+1})$, and $(\partial/\partial\mathbf{E})\,\mathcal{E}(\rho_{k+1}, \mathbf{E}_{k+1})$, respectively.

With the above notations, equations (53) and (54) can be written as

$$\rho_{k+1} = \rho_k - \alpha_k \frac{\partial}{\partial\rho}\,\mathcal{E}(\rho_k, \mathbf{E}_k), \tag{56}$$

$$\mathbf{E}_{k+1} = \mathbf{E}_k - \beta_k \frac{\partial}{\partial\mathbf{E}}\,\mathcal{E}(\rho_k, \mathbf{E}_k). \tag{57}$$

The decrease in mean-square error due to the kth adjustment is denoted by $\Delta\mathcal{E}_k$, that is,

$$\Delta\mathcal{E}_k = \mathcal{E}(\rho_k, \mathbf{E}_k) - \mathcal{E}(\rho_{k+1}, \mathbf{E}_{k+1}). \tag{58}$$

Clearly $\Delta\mathcal{E}_k$ approaches zero when the partial derivatives approach zero. A stronger statement that may sometimes hold is that "$\Delta\mathcal{E}_k$ approaches zero *only* when the partial derivatives approach zero." By this statement is meant that for every $\epsilon > 0$, we can find a $\delta > 0$ such that

$$|\,\Delta\mathcal{E}_k\,| \geqq \delta \tag{59}$$

if

$$\left[\frac{\partial}{\partial \rho}\,\mathcal{E}(\rho_k\,,\,\mathbf{E}_k)\right]^2 + \left[\frac{\partial}{\partial \mathbf{E}}\,\mathcal{E}(\rho_k\,,\,\mathbf{E}_k)\right]'\left[\frac{\partial}{\partial \mathbf{E}}\,\mathcal{E}(\rho_k\,,\,\mathbf{E}_k)\right] \geqq \epsilon. \tag{60}$$

We shall say that $\mathcal{E}$ converges to a stationary point if for every $\epsilon > 0$ there is an N such that

$$\left[\frac{\partial}{\partial \rho}\,\mathcal{E}(\rho_k\,,\,\mathbf{E}_k)\right]^2 + \left[\frac{\partial}{\partial \mathbf{E}}\,\mathcal{E}(\rho_k\,,\,\mathbf{E}_k)\right]'\left[\frac{\partial}{\partial \mathbf{E}}\,\mathcal{E}(\rho_k\,,\,\mathbf{E}_k)\right] < \epsilon \tag{61}$$

for all $k > N$.

The following lemma is needed in our discussion.

Lemma. If $\Delta\mathcal{E}_k > 0$ for all k and $\Delta\mathcal{E}_k$ approaches zero only when the partial derivatives approach zero, $\mathcal{E}$ must converge to a stationary point.

Proof: The proof of this lemma is simple. Assume that $\mathcal{E}$ does not converge to a stationary point, that is, for an $\epsilon > 0$ it is not possible to find an N such that equation (61) holds for all $k > N$. Then equation (59) must hold for an infinite number of k's. From this and the assumptions in the lemma, we see that $\Delta\mathcal{E}_k \geqq \delta > 0$ for infinite number of k's. This implies that $\mathcal{E}$ reduces without bound as k increases, contradicting the fact that the mean-square error cannot be less than zero. Hence the lemma holds.

The lemma provides a means of determining α_k and β_k. According to the lemma, $\mathcal{E}$ will converge to a stationary point if we can determine α_k and β_k such that $\Delta\mathcal{E}_k > 0$ for all k and $\Delta\mathcal{E}_k$ approaches zero only when the partial derivatives approach zero. Theoretically, such α_k and β_k can be determined using known mathematical techniques.[5] But unfortunately these techniques were developed for use with computers and are too complicated to be used in a receiver (for a discussion, see Appendix B). In the following proposition, it is shown that α_k and β_k can be determined rather easily if $\mathbf{E}$ and ρ are set in some alternate fashion.

Proposition 2. Let the set of positive integers be divided arbitrarily into two disjoint subsets $\mathcal{K}_1$ and $\mathcal{K}_2$, each containing an infinite number of positive integers. Let $\alpha_k = 0$ when $k \in \mathcal{K}_1$, and $\beta_k = 0$ when $k \in \mathcal{K}_2$. Let λ_1 denote the maximum eigenvalue of $\mathbf{y}$ ($\lambda_1 > 0$ since $\mathbf{y}$ is positive definite), and let $[\int_{-\infty}^{\infty} s^2(t)\,dt]_{\mathbf{E}_k}$ denote the value of $\int_{-\infty}^{\infty} s^2(t)\,dt$ when $\mathbf{E} = \mathbf{E}_k$. The mean-square error $\mathcal{E}$ will converge to a stationary point if

$$0 < \beta_k < \frac{1}{\lambda_1} \tag{62}$$

for $k \in \mathcal{K}_1$, and

$$0 < \alpha_k < \frac{2^{\frac{1}{2}}}{2f_0\left[\int_{-\infty}^{\infty} s^2(t)\,dt\right]_{\mathbf{E}_k} + 2f_0 \int_{-\infty}^{\infty} q^2(t)\,dt} \tag{63}$$

for $k \in \mathcal{K}_2$.

The proof of this proposition is complex and is given in Appendix C. The proposition states that ρ and $\mathbf{E}$ may be adjusted in any alternating fashion† and the mean-square error $\mathcal{E}$ will converge to a stationary point if β_k satisfies equation (62) during the adjustment of $\mathbf{E}$, and α_k satisfies equation (63) during the adjustment of ρ. The term λ_1 in equation (62) and the terms $[\int_{-\infty}^{\infty} s^2(t)\,dt]_{\mathbf{E}_k}$ and $\int_{-\infty}^{\infty} q^2(t)\,dt$ in equation (63) can be estimated or measured. Consider first λ_1, the maximum eigenvalue of $\mathbf{y}$. There are various methods of estimating the maximum eigenvalue of a matrix.[6,7] For example, it is possible to estimate λ_1 from amplitude characteristics of the transmission medium. [Note from Fig. 1 that the amplitude and phase characteristics of the transmission medium are denoted by $|W(f)|$ and $\omega(f)$, respectively.] It is seen from equation (29) that the elements $y_{h,i}$ of $\mathbf{y}$ can be computed from $|\Gamma(f)|$. For single-sideband systems, $|\Gamma(f)|$ depends on $|W(f)|$, but not on $\omega(f)$ and the demodulating carrier phase ρ. Thus, if statistics of $|W(f)|$ are available (for example, in a voiceband system), $y_{h,i}$ can be computed from equation (29) and λ_1 can be estimated. The maximum possible value of λ_1 then can be used instead of λ_1 in equation (62).

The factor $\int_{-\infty}^{\infty} q^2(t)dt$ in equation (63) is known because the desired signal $q(t)$ is given. The other factor $[\int_{-\infty}^{\infty} s^2(t)dt]_{\mathbf{E}_k}$ is simply the energy of the signal $s(t)$ prior to the kth adjustment, and can be easily measured at the equalizer output.

Summarizing the above, a condition of convergence has been described in the lemma. Based on the lemma, a specific condition of convergence has been obtained in Proposition 2. The upper bound in equation (62) can be estimated from *a priori* channel statistics and the upper bound in equation (63) can be easily determined prior to each adjustment. Thus, α_k and β_k can be set accordingly prior to the adjustments to ensure that $\mathcal{E}$ will converge to a stationary point.

It has been shown in the previous section that a stationary point must be a global minimum or a saddlepoint. Thus, $\mathcal{E}$ may converge

† For example, one may fix ρ and adjust $\mathbf{E}$ until $\partial\mathcal{E}/\partial\mathbf{E}$ approaches zero, then fix $\mathbf{E}$ and adjust ρ until $\partial\mathcal{E}/\partial\rho$ approaches zero, and repeat the cycle until both $\partial\mathcal{E}/\partial\mathbf{E}$ and $\partial\mathcal{E}/\partial\rho$ approach zero.

to a saddlepoint instead of a global minimum. Fortunately, such a possibility is remote. A great advantage of gradient methods is that they will inherently stay away from saddlepoints.[5] It has been found by researchers that gradient search computer program avoids saddlepoints so dependably that the only way they could test their program for exploring the neighborhood of a pass was to start the search there. Wilde and Beightler suggested the reason by sketching a bimodal surface to show that only one gradient line out of the infinite number possible actually passes through the saddlepoint.[5] The other gradient lines all lead directly to a minimum or a maximum. Hence the possibility of converging to a saddlepoint is remote.

It has been shown in the previous section that the distance in ρ between a saddlepoint and its adjacent global minima is $\pi/2$. From this, several tests can be devised for distinguishing between global minima and saddlepoints. The following one is particularly simple. At the design stage of the system, one may compute the value of $\mathcal{E}$ at global minima and saddlepoints (the detailed steps in Section VII may be used. The value of $\mathcal{E}$ at global minima is equal to $\mathcal{E}_{min}$ in equation (64), while the value of $\mathcal{E}$ at saddlepoints is equal to $\mathcal{E}_{ind}$ evaluated at $\Delta\rho = \pi/2$.) As illustrated in Section 8.1 and Fig. 4, the value of $\mathcal{E}$ at saddlepoints can be many times larger than that at global minima. Consequently, a threshold can be set up such that when $\mathcal{E}$ converges to a value above the threshold it may be concluded that a saddlepoint

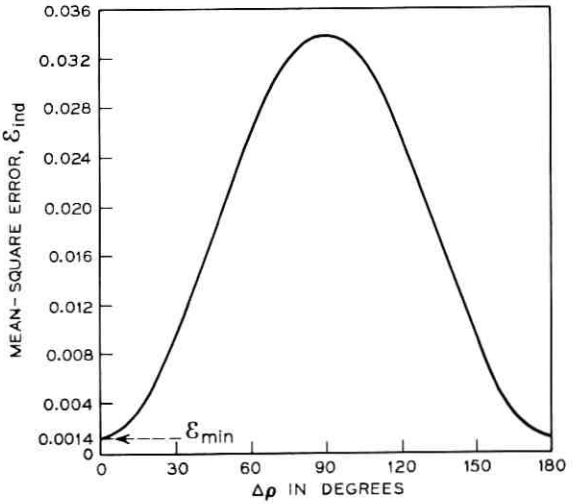


Fig. 4—Variation of $\mathcal{E}_{ind}$ with Δp. (See example in Section VIII.)

is reached. Then ρ will be shifted by $\pi/2$ and $\mathbf{E}$ readjusted. If $\mathcal{E}$ converges to a value below the threshold, the process is terminated. This and other possible tests all require additional circuitry. Since the possibility of converging to a saddlepoint is remote, one should decide from actual trials whether such a test should be employed.

VII. PERFORMANCE COMPARISON

In the previous sections, we have considered jointly setting ρ and $\mathbf{E}$ for minimizing the mean-square error $\mathcal{E}$. It has been shown in Proposition 1 that the joint optimum settings of ρ and $\mathbf{E}$ are given, respectively, by equation (45) and (46) (with m_0 in these equations being an even integer). Substituting these optimum settings into equation (22), we obtain the minimum mean-square error

$$\mathcal{E}_{\min} = \sum_{k=-\infty}^{\infty} q_k^2 - 2 \mid \underline{\mathbf{u}}'\mathbf{y}^{-1}\underline{\mathbf{u}} \mid - 2\underline{\mathbf{u}}'\mathbf{y}^{-1}\mathbf{v}. \tag{64}$$

It has been shown that this $\mathcal{E}_{\min}$ can be obtained by setting ρ and $\mathbf{E}$ jointly. Now we wish to compare $\mathcal{E}_{\min}$ with what can be obtained by another scheme. In single-sideband systems, it is possible to transmit a carrier pilot to the receiver to demodulate the received signal. The demodulating carrier phase therefore is

$$\rho = \rho_c + \rho_f,$$

where ρ_c is the phase of the received carrier pilot and ρ_f is a fixed phase shift that is sometimes introduced for signal shaping. With ρ fixed at $\rho_c + \rho_f$, $\mathbf{E}$ can be adjusted to minimize $\mathcal{E}$. The value of $\mathcal{E}$ thus obtained will be denoted by $\mathcal{E}_{\text{ind}}$, where the subscript "ind" indicates that ρ and $\mathbf{E}$ are set independently. We now compare $\mathbf{E}_{\text{ind}}$ with $\mathbf{E}_{\min}$.

To determine $\mathbf{E}_{\text{ind}}$, let the difference between $\rho_c + \rho_f$ and the optimum setting of ρ be denoted by $\Delta\rho$. From equation (45),

$$\Delta\rho = \rho_c + \rho_f - \left[m_0 \frac{\pi}{2} + \tfrac{1}{2} \tan^{-1} \frac{\text{Im } [\underline{\mathbf{u}}'\mathbf{y}^{-1}\underline{\mathbf{u}}]}{\text{Re } [\underline{\mathbf{u}}'\mathbf{y}^{-1}\underline{\mathbf{u}}]} \right] \tag{65}$$

where m_0 can be any even integer (including zero). Since $\mathbf{E}$ is set to minimize $\mathcal{E}$ after ρ is set to $\rho_c + \rho_f$, $\mathbf{E}$ is equal to $\mathbf{y}^{-1}\mathbf{v}$ evaluated at $\rho = \rho_c + \rho_f$, that is,

$$\mathbf{E} = \mathbf{y}^{-1}\mathbf{v}_c \tag{66}$$

where $\mathbf{v}_c$ is $\mathbf{v}$ evaluated at $\rho = \rho_c + \rho_f$. Substituting equation (66) into equation (22) gives

$$\mathcal{E}_{\text{ind}} = \sum_{k=-\infty}^{\infty} q_k^2 - \mathbf{v}_c' \mathbf{y}^{-1} \mathbf{v}_c \,. \tag{67}$$

It can easily be shown from equation (33) that

$$\mathbf{v}_c' \mathbf{y}^{-1} \mathbf{v}_c = \exp\left[-j2(\rho_c + \rho_f)\right] \underline{\mu}' \mathbf{y}^{-1} \underline{\mu} + 2\underline{\mu}' \mathbf{y}^{-1} \mathbf{v} + \exp\left[j2(\rho_c + \rho_f)\right] \mathbf{v}' \mathbf{y}^{-1} \mathbf{v}. \tag{68}$$

Substituting equation (65) into equation (68), and substituting the resulting equation into equation (67), one can obtain after some manipulations that

$$\mathcal{E}_{\text{ind}} = \sum_{k=-\infty}^{\infty} q_k^2 - 2(\cos 2\,\Delta\rho)\, |\, \underline{\mu}' \mathbf{y}^{-1} \underline{\mu}\, | - 2\underline{\mu}' \mathbf{y}^{-1} \mathbf{v}. \tag{69}$$

From equations (64) and (69), the difference between $\mathcal{E}_{\text{ind}}$ and $\mathcal{E}_{\text{min}}$ is

$$\mathcal{E}_{\text{ind}} - \mathcal{E}_{\text{min}} = 2[1 - \cos 2\Delta\rho]\, |\, \underline{\mu}' \mathbf{y}^{-1} \underline{\mu}\, |. \tag{70}$$

Note that $\underline{\mu}'$ and $\mathbf{y}^{-1}$ are independent of $\Delta\rho$. Thus, the only term in equation (70) that depends on $\Delta\rho$ is $\cos 2\Delta\rho$. Now we can make the following observations.

First, $\mathcal{E}_{\text{ind}} - \mathcal{E}_{\text{min}}$ is nonnegative, meaning that independently setting ρ and $\mathbf{E}$ increases the mean-square error. Second, $\mathcal{E}_{\text{ind}} - \mathcal{E}_{\text{min}}$ varies periodically with $\Delta\rho$ with a period of π. Third, because of the nature of cosine, $\mathcal{E}_{\text{ind}} - \mathcal{E}_{\text{min}}$ is small when $\Delta\rho$ is small, but increases rapidly when $\Delta\rho$ increases (note the factor two in $\cos 2\Delta\rho$).

For a given system, one may compute $\mathcal{E}_{\text{ind}}$, $\mathcal{E}_{\text{min}}$, and their difference from equations (64) and (69). The computation can best be carried out by a computer program in the following steps (see also the example in the next section). First, specify f_0, N, and the desired signal $q(t)$. Determine the time samples $\{q_k\}$ of $q(t)$. Second, determine $H(f)$ from transfer functions of the transmitting filter, the transmission medium, and the receiving filter. Third, compute the elements $y_{h,i}$ of $\mathbf{y}$ from the equation

$$y_{h,i} = 4f_0 \int_0^{f_0} |\, H(f)\, |^2 \cos\left[2\pi f(h - i)T_0\right] df. \tag{71}$$

Compute the elements $a_{h,i}$ of $\mathbf{y}^{-1}$ from $y_{h,i}$. Fourth, compute the terms $|\, \underline{\mu}' \mathbf{y}^{-1} \underline{\mu}\, |$ and $\underline{\mu}' \mathbf{y}^{-1} \mathbf{v}$ in equations (64) and (69) using the explicit equations

$$|\, \underline{\mu}' \mathbf{y}^{-1} \underline{\mu}\, | = [(B_1 - B_2)^2 + 4B_3^2]^{\frac{1}{2}}, \tag{72}$$

$$\underline{\mu}' \mathbf{y}^{-1} \mathbf{v} = B_1 + B_2 \,, \tag{73}$$

where

$$B_1 = \sum_{k=-N}^{N} \sum_{i=-N}^{N} a_{k,i}\xi_k\xi_i , \tag{74}$$

$$B_2 = \sum_{k=-N}^{N} \sum_{i=-N}^{N} a_{k,i}\zeta_k\zeta_i , \tag{75}$$

$$B_3 = \sum_{k=-N}^{N} \sum_{i=-N}^{N} a_{k,i}\xi_k\zeta_i , \tag{76}$$

$$\xi_i = \sum_{k=-\infty}^{\infty} q_k \int_0^{f_0} | H(f) | \cos [\eta(f) + 2\pi f(kT_0 - iT_0)]\, df, \tag{77}$$

$$\zeta_i = \sum_{k=-\infty}^{\infty} q_k \int_0^{f_0} | H(f) | \sin [\eta(f) + 2\pi f(kT_0 - iT_0)]\, df. \tag{78}$$

Finally, compute $\mathcal{E}_{\min}$ from equation (64) and $\mathcal{E}_{\text{ind}}$ from equation (69).

VIII. COMPUTER SIMULATIONS

We now apply the previous results to a single-sideband partial-response voiceband data communication system using a five- or nine-tap transversal equalizer. Since the results are largely similar, we shall describe only the five-tap case.

Two computer programs have been written for this system. In the first program, $\mathcal{E}_{\min}$ and $\mathcal{E}_{\text{ind}}$ are computed using the formulas in the previous section. In the second one, $\mathcal{E}_{\min}$ and $\mathcal{E}_{\text{ind}}$ are obtained using the method of steepest descent. The results of these two programs are described separately in the following subsections.

8.1 *Comparison of* $\mathcal{E}_{\min}$ *and* $\mathcal{E}_{\text{ind}}$ *Using the Explicit Formulas in Section VII*

Consider the voiceband data communication system described in Ref. 8. The desired signal $q(t)$ of such a system is the class IV partial-response signal,[9] that is,

$$q(t) = \frac{2^{\frac{1}{2}}\pi \sin 2\pi f_0(t - t_0)}{[2\pi f_0(t - t_0)]^2 - \pi^2} , \tag{79}$$

where a time delay t_0 is included to take into account the time delay in the channel. From Ref. 8,

$$f_0 = 1200 \text{ Hz}. \tag{80}$$

Hence, $T_0 = 1/2f_0 = 1/2400$ seconds. We shall consider a transversal

equalizer with five taps, that is, $N = 2$. It has been defined that $H(f)$ is the Fourier transform of $\gamma(t)$ when $\rho = 0$. The amplitude and phase of $H(f)$ have been denoted by $|H(f)|$ and $\eta(f)$, respectively, that is,

$$H(f) = |H(f)| \exp [j\eta(f)]. \tag{81}$$

It is assumed that the amplitude and phase distortions in the filters are negligible compared with those in the transmission medium. Consequently, $|H(f)|$ and $\eta(f)$ can be determined from amplitude and phase characteristics of the transmission medium (which is a voice channel in this case). From Ref. 10 and Fig. 4 of Ref. 8, a typical $|H(f)|$ is found to be

$$|H(f)| = [\sin 2\pi T_0 f] 10^{0.3T_0(f_0-2f)}, \qquad 0 \leqq f \leqq f_0. \tag{82}$$

The component $\sin 2\pi T_0 f$ in equation (82) represents the desired amplitude characteristic of the class IV partial-response signal, while the term $10^{0.3T_0(f_0-2f)}$ represents typical amplitude distortion in a voice channel. The delay characteristic of five links of K carrier shown in Fig. 24 of Ref. 11 is representative of the delay characteristic of a voice channel. Therefore, it will be used here and

$$\eta(f) = 9.89 \sin [2\pi(f + f_0) \cdot 0.00019 - 2.203], \qquad 0 \leqq f \leqq f_0. \tag{83}$$

Constant phases and time delays in the transmission medium and the filters are omitted in writing $\eta(f)$ because their values are not available and because they only change the phase and time origins in the computations. However, such an omission makes it impossible to determine the phase ρ_c of the received carrier pilot because the carrier pilot does not travel the same path as the signal. The signal is transmitted through the transmitting and receiving filters, while the carrier pilot is transmitted outside of these filters (these filters theoretically should have infinite attenuations at the carrier frequency). Furthermore, the carrier pilot is recovered at the receiver through a separate narrowband filter or a phase-lock loop, while the signal is demodulated and passes through a low-pass filter. With these differences and without detailed phase characteristics of the filters, it is not possible to determine ρ_c here. Therefore, we shall simply leave ρ_c and $\Delta\rho$ as variables and compute the variation of $\mathcal{E}_{\text{ind}}$ with $\Delta\rho$. (It should be noted that $\Delta\rho$ may assume small values in some real systems.)

The variation of $\mathcal{E}_{\text{ind}}$ with $\Delta\rho$ has been determined for various values of t_0. The curve obtained at $t_0 = -2T_0$ is typical and is presented in Fig. 4. The value of $\mathcal{E}_{\min}$ is also indicated in the figure. It can be seen that $\mathcal{E}_{\text{ind}}$ can be as large as 0.034, while $\mathcal{E}_{\min}$ is only 0.0014. Thus,

the mean-square error can increase $0.034/0.0014 = 24.3$ times if the demodulating carrier phase ρ is not set properly. Note that this large increase is obtained for a five-tap transversal equalizer. A similar result has been obtained for a nine-tap transversal equalizer. These results demonstrate that the mean-square error depends critically on the carrier phase setting when the number of equalizer taps is not large. (It should be noted, however, that when a large number of taps are used, the mean-square error will not be sensitive to the carrier phase setting.)

8.2 *Mean-Square Errors Obtained by Method of Steepest Descent*

A computer program has been written to simulate the system described in Section 8.1. The receiver parameters are adjusted by the method of steepest descent described in Section IV. Five hundred test pulses are used in each training period. The initial setting of the center tap e_0 is unity, while the initial settings of the other taps are zero. The parameter t_0 is fixed at $-2T_0$ as mentioned in Section 8.1. The mean-square errors obtained with various initial settings of ρ are shown in Fig. 5.

The points marked by "×" in Fig. 5 are obtained by adjusting ρ and $\mathbf{E}$ jointly using $\alpha_k = 0.523$ and $\beta_k = 0.1$ for all k. For instance, point A is obtained by initially setting ρ to 40 degrees above the optimum

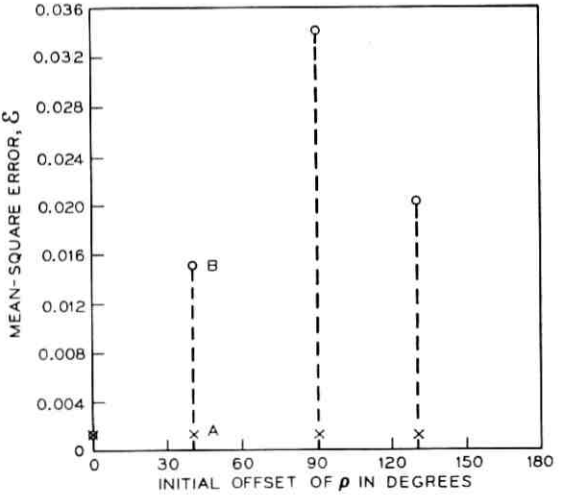


Fig. 5—Mean-square errors obtained by method of steepest descent.

ρ and then adjusting ρ and **E** jointly by the method of steepest descent. It can be seen from these points that the mean-square error $\mathcal{E}$ always converges to the value of $\mathcal{E}_{min}$ determined in Section 8.1. This demonstrates the fact that $\mathcal{E}_{min}$ can be obtained by jointly setting ρ and **E**.

The other points marked by circles are obtained by fixing ρ and adjusting **E** only ($\alpha_k = 0$ and $\beta_k = 0.1$ for all k). For instance, point B is obtained by initially setting ρ to 40 degrees above the optimum ρ and then adjusting **E** only by the method of steepest descent. It can be seen, by comparing these circled points with the value of $\mathcal{E}_{ind}$ in Fig. 4, that the mean-square error $\mathcal{E}$ converges, as expected, to $\mathcal{E}_{ind}$.

Only 500 test pulses are used in each training period because $\mathcal{E}$ converges within this period in all cases. For instance, the convergence of $\mathcal{E}$ to the values shown by points A and B in Fig. 5 are illustrated, respectively, by curves A and B in Fig. 6. It can be seen that $\mathcal{E}$ converges rapidly to the final values.

IX. SUMMARY AND CONCLUSIONS

It is shown in this paper that in single-sideband systems the transversal equalizer and the carrier phase can be set jointly by the method of steepest descent to minimize the mean-square equalization error.

The system is analyzed and a receiver structure is developed. The receiver structure is theoretically as simple as a conventional one.

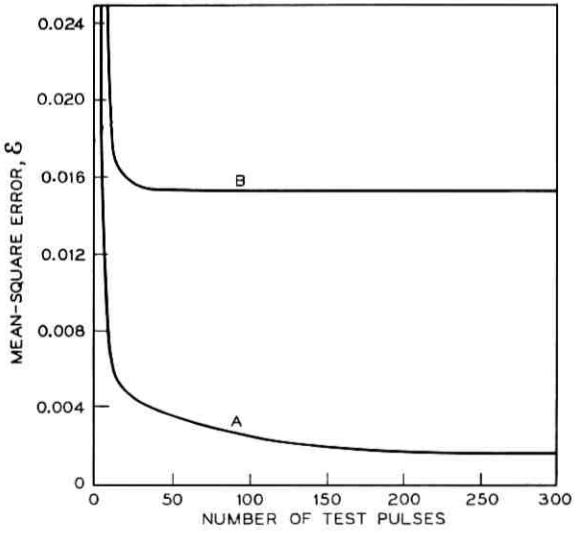


Fig. 6—An illustration of convergence.

A well-known problem of the method of steepest descent is that the function to be minimized may converge to a local minimum instead of to a global minimum. To prove that this troublesome problem does not arise here, the variation of the mean-square error in the parameter space is analyzed. Exact locations of the stationary points in the parameter space are determined and the classifications of the stationary points are obtained. It is shown that the mean-square error has only global minima and saddlepoints, but not local minimum or maximum. Thus, it is not possible for the adjustments to converge to a local minimum, regardless of the initial parameter settings. This completely eliminates the problem.

It is also shown that two adjacent global minima (or two adjacent saddlepoints) are separated by 180 degrees in demodulating carrier phase, while a global minimum is separated from its adjacent saddlepoints by 90 degrees in demodulating carrier phase. From this, a test is described to determine whether a global minimum or a saddlepoint is reached by the adjustments and to correct the settings if a saddlepoint is reached. This test may not be necessary because both previous experience and theoretical considerations have shown that the method of steepest descent inherently stays away from saddlepoints.

The choice of the step sizes of the adjustments is also considered. There are methods for determining the step sizes; however, they require complicated computations. As an alternative, it is shown in this paper that the step sizes can be easily determined if the equalizer and the demodulating carrier phase are adjusted in different steps (the steps can be alternated in any manner).

Closed-form expressions of the joint optimum parameter settings and of the corresponding minimum mean-square error are obtained for computation of system performance. For illustration purposes, a single-sideband data communication system using a five- or nine-tap transversal equalizer is simulated on a computer. Both theoretical and simulation results show that the equalization error can increase by ten times or more when the carrier phase is not properly set. These demonstrate that when the number of equalizer taps is not large, the equalization error depends critically on the carrier phase setting. The computer simulation also verifies the theory that the equalization error can be minimized by using the joint method described in this paper.

X. ACKNOWLEDGMENTS

I wish to thank Miss A. C. Weingartner for writing a computer program to simulate the data communication system described in Section

VIII, and Miss C. A. Reichenstein for the computer program used in Section 8.1.

APPENDIX A

Classification of Stationary Points

In this appendix, we prove the statements that:

(*i*) F is positive definite when ρ and $\mathbf{E}$ are given, respectively, by equations (45) and (46), and m_0 in these equations is an even integer;
(*ii*) F is indefinite when ρ and $\mathbf{E}$ are given, respectively, by equations (45) and (46), and m_0 in these equations is an odd integer; and
(*iii*) All the local minima of $\mathcal{E}_0$ are equal.

Consider the first statement. Let Ω be the set of all $\mathbf{h}$ except $\mathbf{h} = \mathbf{0}$. Let Ω be decomposed into two disjoint sets Ω_1 and Ω_2, where Ω_1 contains the h's with $h_1 \neq 0$, and Ω_2 contains the $\mathbf{h}$'s with $h_1 = 0$. Consider first Ω_1. Since $h_1 \neq 0$ in Ω_1, equation (53) can be written as

$$F = 2h_1^2 \int_{-\infty}^{\infty} [\sigma(t) - q(t)]^2 \, dt + 2h_1^2 \int_{-\infty}^{\infty} q(t)s(t) \, dt - 2h_1^2 \int_{-\infty}^{\infty} q^2(t) \, dt \tag{84}$$

where

$$\sigma(t) = \sum_{n=-N}^{N} e_n' \hat{\gamma}(t - nT_0 - NT_0) \tag{85}$$

and

$$e_n' = \frac{h_{2+n+N}}{h_1}, \qquad n = -N \text{ to } N. \tag{86}$$

For the sake of brevity, we shall denote the entire right side of equation (45) by ρ_0 and the entire right side of equation (46) by $\mathbf{E}_0$. For any function G, the symbol $[G]_{\rho_0}$ denotes the value of G evaluated at $\rho = \rho_0$, the symbol $[G]_{\rho_0, \mathbf{E}_0}$ denotes the value of G evaluated at $\rho = \rho_0$ and $\mathbf{E} = \mathbf{E}_0$, and the symbol Min G denotes the minimum value of G in Ω_1. These symbols and notations may be used jointly. For instance, Min $[G]_{\rho_0, \mathbf{E}_0}$ denotes the minimum value of $[G]_{\rho_0, \mathbf{E}_0}$ in Ω_1.

Since $h_1 \neq 0$ in Ω_1, $[F]_{\rho_0, \mathbf{E}_0} > 0$ in Ω_1 if and only if $[F/2h_1^2]_{\rho_0, \mathbf{E}_0} > 0$ in Ω_1, or if and only if Min $[F/2h_1^2]_{\rho_0, \mathbf{E}_0} > 0$. From equation (84),

$$\text{Min}\left[\frac{F}{2h_1^2}\right]_{\rho_0,\mathbf{E}_0} = \text{Min}\left[\int_{-\infty}^{\infty}[\sigma(t) - q(t)]^2\,dt\right]_{\rho_0} + \left[\int_{-\infty}^{\infty} q(t)s(t)\,dt\right]_{\rho_0,\mathbf{E}_0} - \left[\int_{-\infty}^{\infty} q^2(t)\,dt\right]. \tag{87}$$

In writing the above equation, we have used the fact that $\sigma(t)$ is independent of $\mathbf{E}$, $s(t)$ is independent of $\mathbf{h}$, and $q(t)$ is independent of ρ, $\mathbf{E}$, and $\mathbf{h}$.

Consider the first term on the right side of equation (87). From equation (85), $\sigma(t)$ is a function of $\hat{\gamma}(t)$, where $\hat{}$ indicates Hilbert transform. It can be shown from equations (7) and (19) that

$$[\hat{\gamma}(t)]_{\rho_0} = [\gamma(t)]_{\rho_0+\pi/2}\,. \tag{88}$$

Now we have

$$\begin{aligned}
&\text{Min}\left[\int_{-\infty}^{\infty}[\sigma(t) - q(t)]^2\,dt\right]_{\rho_0} \\
&= \text{Min}\int_{-\infty}^{\infty}\left[\sum_{n=-N}^{N} e_n'[\hat{\gamma}(t - nT_0 - NT_0)]_{\rho_0} - q(t)\right]^2 dt, \\
&= \text{Min}\int_{-\infty}^{\infty}\left[\sum_{n=-N}^{N} e_n'[\gamma(t - nT_0 - NT_0)]_{\rho_0+\pi/2} - q(t)\right]^2 dt, \\
&= \left\{\text{Min}\int_{-\infty}^{\infty}\left[\sum_{n=-N}^{N} e_n'\gamma(t - nT_0 - NT_0) - q(t)\right]^2 dt\right\}_{\rho_0+\pi/2}.
\end{aligned} \tag{89}$$

Note that equation (88) is used in the second step above. Since the term $\sum_{n=-N}^{N} e_n'\gamma(t - nT_0 - NT_0)$ in the last expression above is similar to $s(t)$ in equation (8), the whole term in the $\{\ \}$ in the last expression above is equal to $\mathcal{E}_0$ minimized with respect to $\mathbf{E}$. This proves that Min $[\int_{-\infty}^{\infty}[\sigma(t) - q(t)]^2\,dt]_{\rho_0}$ is equal to $\mathcal{E}_0$ minimized with respect to $\mathbf{E}$ and evaluated at $\rho = \rho_0 + \pi/2$. It can be easily shown from equations (21) and (22) that $\mathcal{E}_0$ minimized with respect to $\mathbf{E}$ is equal to $1/2f_0[\sum_{k=-\infty}^{\infty} q_k' - \mathbf{v}'\mathbf{y}^{-1}\mathbf{v}]$; hence,

$$\begin{aligned}
&\text{Min}\left[\int_{-\infty}^{\infty}[\sigma(t) - q(t)]^2\,dt\right]_{\rho_0} \\
&= \frac{1}{2f_0}\sum_{k=-\infty}^{\infty} q_k^2 - \frac{1}{2f_0}[\mathbf{v}'\mathbf{y}^{-1}\mathbf{v}]_{\rho_0+\pi/2}\,, \\
&= \frac{1}{2f_0}\sum_{k=-\infty}^{\infty} q_k^2 - \frac{1}{2f_0}[2\underline{\mathbf{u}}'\mathbf{y}^{-1}\mathbf{v} - \exp(-j2\rho_0)\underline{\mathbf{u}}'\mathbf{y}^{-1}\underline{\mathbf{u}} - \exp(j2\rho_0)\mathbf{v}'\mathbf{y}^{-1}\mathbf{v}],
\end{aligned} \tag{90}$$

where equation (33) has been used in the last step.

We have evaluated the first term in equation (87). Consider now the second term in equation (87). From equations (20), (23), and (26),

$$\int_{-\infty}^{\infty} q(t)s(t)\,dt = \frac{1}{2f_0}\,\mathbf{v}'\mathbf{E}. \tag{91}$$

Substituting equations (33), (45), and (46) into equation (91) and simplifying, we obtain, after some steps,

$$\left[\int_{-\infty}^{\infty} q(t)s(t)\,dt\right]_{\rho_0,\mathbf{E}_0} = \frac{1}{f_0}\,\mathfrak{u}'\mathbf{y}^{-1}\mathbf{v} + \frac{1}{2f_0}\exp(-j2\rho_0)\mathfrak{u}'\mathbf{y}^{-1}\mathfrak{u} + \frac{1}{2f_0}\exp(j2\rho_0)\mathfrak{v}'\mathbf{y}^{-1}\mathbf{v}. \tag{92}$$

The last term in equation (87) is

$$\int_{-\infty}^{\infty} q^2(t)\,dt = \frac{1}{2f_0}\sum_{k=-\infty}^{\infty} q_k^2\,. \tag{93}$$

Substituting equations (90), (92), and (93) into (87) and canceling out the terms having opposite signs, we obtain

$$\text{Min}\left[\frac{F}{2h_1^2}\right]_{\rho_0,\mathbf{E}_0} = \frac{1}{f_0}\exp(-j2\rho_0)\mathfrak{u}'\mathbf{y}^{-1}\mathfrak{u} + \frac{1}{f_0}\exp(j2\rho_0)\mathfrak{v}'\mathbf{y}^{-1}\mathbf{v}. \tag{94}$$

Note that so far m_0 can be any integer. Let $|\,\mathfrak{u}'\mathbf{y}^{-1}\mathfrak{u}\,|$ denote the magnitude of $\mathfrak{u}'\mathbf{y}^{-1}\mathfrak{u}$. It can be shown from equation (44) that when m_0 is an even integer, the right side of equation (94) is equal to $2/f_0|\,\mathfrak{u}'\mathbf{y}^{-1}\mathfrak{u}\,|$. Hence, when m_0 is an even integer (including zero),

$$\text{Min}\left[\frac{F}{2h_1^2}\right]_{\rho_0,\mathbf{E}_0} = \frac{2}{f_0}\,|\,\mathfrak{u}'\mathbf{y}^{-1}\mathfrak{u}\,|\,. \tag{95}$$

Since $|\,\mathfrak{u}'\mathbf{y}^{-1}\mathfrak{u}\,| \neq 0$, equation (95) shows that $\text{Min}\,[F/2h_1^2]_{\rho_0,\mathbf{E}_0} > 0$. Thus, $[F]_{\rho_0,\mathbf{E}_0} > 0$ in Ω_1 when m_0 is an even integer.

Now consider Ω_2. Since $h_1 = 0$ in Ω_2, equation (53) is reduced to

$$F = 2\int_{-\infty}^{\infty}\left[\sum_{n=-N}^{N} h_{2+n+N}\gamma(t - nT_0 - NT_0)\right]^2 dt. \tag{96}$$

Since $\gamma(t - nT_0 - NT_0)$, $n = -N$ to N, are linearly independent and h_2 to h_{2N+2} cannot be all zero in Ω_2, the integrand in equation (96) cannot be zero for all t. Hence, $F > 0$ in Ω_2. This implies, of course, $[F]_{\rho_0,\mathbf{E}_0} > 0$ in Ω_2.

We have shown that when m_0 is an even integer, $[F]_{\rho_0,\mathbf{E}_0} > 0$ in Ω_1 and Ω_2. Hence, when m_0 is an even integer, $[F]_{\rho_0,\mathbf{E}_0} > 0$ for all $\mathbf{h}$ except $\mathbf{h} = 0$. This proves the first statement at the beginning of this Appendix.

Now we prove the second statement at the beginning of this Appendix. Note that the derivations from equations (84) through (94) hold for all integer m_0. It can be shown from equation (44) that when m_0 is an odd integer, the right side of equation (94) is equal to $-2/f_0|\ \mathbf{\mu}'\mathbf{y}^{-1}\mathbf{\mu}\ |$. Hence, when m_0 is an odd integer, equation (94) becomes

$$\text{Min}\left[\frac{F}{2h_1^2}\right]_{\rho_0,\mathbf{E}_0} = -\frac{2}{f_0}\ |\ \mathbf{\mu}'\mathbf{y}^{-1}\mathbf{\mu}\ |\ . \tag{97}$$

Since $|\ \mathbf{\mu}'\mathbf{y}^{-1}\mathbf{\mu}\ | \neq 0$, Min $[F/2h_1^2]_{\rho_0,\mathbf{E}_0} < 0$. Thus, $[F]_{\rho_0,\mathbf{E}_0}$ can be negative in Ω_1 (and hence in Ω). It can also be positive in Ω (for instance, $[F]_{\rho_0,\mathbf{E}_0} > 0$ in Ω_2). Therefore, F is indefinite when ρ and $\mathbf{E}$ are given, respectively, by equations (45) and (46), and m_0 in these equations is an odd integer. This proves the second statement.

Finally, consider the third statement. We have shown that F is positive definite and $\mathcal{E}_0$ has a local minimum when ρ and $\mathbf{E}$ are given, respectively, by equations (45) and (46), and m_0 in these equations is an even integer. Substituting equations (45) and (46) into equation (22) and letting m_0 be an even integer, we see that the local minimum of $\mathcal{E}_0$ is

$$\mathcal{E}_0 = \frac{1}{2f_0}\sum_{k=-\infty}^{\infty} q_k^2 - \frac{1}{f_0}\ |\ \mathbf{\mu}'\mathbf{y}^{-1}\mathbf{\mu}\ | - \frac{1}{f_0}\,\mathbf{\mu}'\mathbf{y}^{-1}\mathbf{\nu}. \tag{98}$$

Since the right side of equation (98) is independent of m_0, all the local minima of $\mathcal{E}_0$ are equal and hence are all global minima. This proves the third statement at the beginning of this Appendix.

APPENDIX B

Discussion of Existing Method

A number of methods are described in Ref. 5 for determining the proportional constant in steepest descent adjustments. It has been pointed out that these methods can be used to solve certain nonlinear equations on computer.[12]

These methods require elaborate computations. For example, consider the possibility of determining α_k and β_k using the theorem on page 31 of Ref. 5. Since it is assumed in that theorem that a single proportional constant is used for all parameters, we change the scale of ρ or $\mathbf{E}$ such that it is also appropriate to use a single proportional constant in our case. After this is accomplished, we can use a single constant α_k in equations (56) and (57). To determine α_k, define

$$g(\rho_k , \mathbf{E}_k , \alpha_k) = \frac{\Delta \mathcal{E}_k}{\alpha_k \left\{ \left[\frac{\partial}{\partial \rho} \mathcal{E}(\rho_k , \mathbf{E}_k) \right]^2 + \sum_{n=-N}^{N} \left[\frac{\partial}{\partial e_n} \mathcal{E}(\rho_k , \mathbf{E}_k) \right]^2 \right\}} . \quad (99)$$

Choose a constant δ in the range $0 < \delta \leqq \frac{1}{2}$. Compute $g(\rho_k , \mathbf{E}_k , 1)$. If $g(\rho_k , \mathbf{E}_k , 1) < \delta$, choose $\alpha_k < 1$ so that $\delta \leqq g(\rho_k , \mathbf{E}_k , \alpha_k) \leqq 1 - \delta$ (it has been shown that this choice is always possible). If $g(\rho_k , \mathbf{E}_k , 1) \geqq \delta$, choose $\alpha_k = 1$.

It can be seen from this description that the method requires elaborate computations and α_k must be determined on a trial and error basis when $g(\rho_k , \mathbf{E}_k , 1) < \delta$. It is therefore difficult to use this method in a receiver.

APPENDIX C

Proof of Proposition 2

In this Appendix, we prove Proposition 2. Consider first $k \in \mathcal{K}_1$. Since $\alpha_k = 0$ when $k \in \mathcal{K}_1$, we have $\rho_{k+1} = \rho_k$ and

$$\begin{aligned} \Delta \mathcal{E}_k &= \mathcal{E}(\rho_k , \mathbf{E}_k) - \mathcal{E}(\rho_k , \mathbf{E}_{k+1}), \\ &= \mathbf{E}_k' \mathbf{y} \mathbf{E}_k - 2\mathbf{E}_k' \mathbf{v} + \sum_{h=-\infty}^{\infty} q_h^2 \\ &\quad - \left[\mathbf{E}_{k+1}' \mathbf{y} \mathbf{E}_{k+1} - 2\mathbf{E}_{k+1}' \mathbf{v} + \sum_{h=-\infty}^{\infty} q_h^2 \right]. \end{aligned} \quad (100)$$

From equations (57) and (38)

$$\mathbf{E}_{k+1} = \mathbf{E}_k - \beta_k (2\mathbf{y}\mathbf{E}_k - 2\mathbf{v}). \quad (101)$$

Substituting equation (101) into equation (100), we rearrange the resulting equation into the form

$$\Delta \mathcal{E}_k = 4\beta_k (\mathbf{y}\mathbf{E}_k - \mathbf{v})' [\mathbf{I} - \beta_k \mathbf{y}] (\mathbf{y}\mathbf{E}_k - \mathbf{v}) \quad (102)$$

where $\mathbf{I}$ is the identity matrix. Let the eigenvalues of $\mathbf{y}$ be denoted, in the order of decreasing magnitude, by λ_i, $i = 1$ to $2N + 1$, so that

$$\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_{2N+1} . \quad (103)$$

It can be shown that $\mathbf{y}$ is positive definite; hence,

$$\lambda_i > 0, \qquad i = 1 \text{ to } 2N + 1. \quad (104)$$

Let $\mathbf{u}_i$, $i = 1$ to $2N + 1$, be a set of orthonormal eigenvectors of $\mathbf{y}$. Let

$$\mathbf{Q} = [\mathbf{u}_1 \mathbf{u}_2 \cdots \mathbf{u}_{2N+1}]. \quad (105)$$

That is, the ith column of $\mathbf{Q}$ is $\mathbf{u}_i$. From well-known matrix properties,

$$\mathbf{y} = \mathbf{QDQ}' \tag{106}$$

where $\mathbf{D}$ is a diagonal matrix whose ith diagonal element is λ_i. Furthermore,

$$\mathbf{QQ}' = \mathbf{I}. \tag{107}$$

Substituting equations (106) and (107) into equation (102), one can write

$$\Delta \mathcal{E}_k = 4\beta_k \sum_{i=1}^{2N+1} (1 - \beta_k\lambda_i)[(\mathbf{yE}_k - \mathbf{v})'\mathbf{u}_i]^2. \tag{108}$$

If β_k satisfies equation (62), we can write

$$1 - \beta_k\lambda_1 \geqq P_1 > 0. \tag{109}$$

Then, from equations (103), (108), and (109), we have

$$\Delta \mathcal{E}_k \geqq 4\beta_k P_1 \sum_{i=1}^{2N+1} [(\mathbf{yE}_k - \mathbf{v})'\mathbf{u}_i]^2. \tag{110}$$

It can be shown from equations (38), (107), and (105) that

$$\left[\frac{\partial}{\partial \mathbf{E}} \mathcal{E}(\rho_k , \mathbf{E}_k)\right]'\left[\frac{\partial}{\partial \mathbf{E}} \mathcal{E}(\rho_k , \mathbf{E}_k)\right] = 4 \sum_{i=1}^{2N+1} [(\mathbf{yE}_k - \mathbf{v})'\mathbf{u}_i]^2. \tag{111}$$

Comparing equations (110) and (111) gives

$$\Delta \mathcal{E}_k \geqq \beta_k P_1 \left[\frac{\partial}{\partial \mathbf{E}} \mathcal{E}(\rho_k , \mathbf{E}_k)\right]'\left[\frac{\partial}{\partial \mathbf{E}} \mathcal{E}(\rho_k , \mathbf{E}_k)\right]. \tag{112}$$

Now we have the conclusion:

Conclusion 1. If equation (62) holds, $\Delta\mathcal{E}_k$ is bounded below by equation (112) which shows that $\Delta\mathcal{E}_k$ is positive and approaches zero only when the partial derivatives $(\partial/\partial\mathbf{E})\mathcal{E}(\rho_k , \mathbf{E}_k)$ approach zero.

The above is for $k \in \mathcal{K}_1$. Next we consider $k \in \mathcal{K}_2$. Since $\beta_k = 0$ when $k \in \mathcal{K}_2$, we have $\mathbf{E}_{k+1} = \mathbf{E}_k$. From equations (58) and (22)

$$\begin{aligned} \Delta \mathcal{E}_k &= \mathcal{E}(\rho_k , \mathbf{E}_k) - \mathcal{E}(\rho_{k+1} , \mathbf{E}_k), \\ &= -2\mathbf{E}_k'(\mathbf{v}_k - \mathbf{v}_{k+1}), \end{aligned} \tag{113}$$

where

$$\mathbf{v}_k = \exp(-j\rho_k)\mathbf{u} + \exp(j\rho_k)\mathbf{v}, \tag{114}$$

$$\mathbf{v}_{k+1} = \exp(-j\rho_{k+1})\mathbf{u} + \exp(j\rho_{k+1})\mathbf{v}. \tag{115}$$

From equation (56) and the two equations above,

$$\mathbf{v}_k = \mathbf{v}_{k+1} = \left\{1 - \cos\left[\alpha_k \frac{\partial}{\partial \rho} \mathcal{E}(\rho_k , \mathbf{E}_k)\right]\right\}[\exp(-j\rho_k)\mathbf{u}_u + \exp(j\rho_k)\mathbf{v}] - j \sin\left[\alpha_k \frac{\partial}{\partial \rho} \mathcal{E}(\rho_k , \mathbf{E}_k)\right][\exp(-j\rho_k)\mathbf{u} - \exp(j\rho_k)\mathbf{v}]. \tag{116}$$

The vectors $\mathbf{u}$ and $\mathbf{v}$ are defined in equation (37) and their elements are defined in equations (34) and (35). It can be seen from equations (34) and (35) that the elements of $\mathbf{u}$ and $\mathbf{v}$ are complex numbers. Let $\mathbf{u}$ be written as

$$\mathbf{u} = \xi + j\zeta \tag{117}$$

where the elements of ξ and ζ are real numbers that can be determined from equation (34). Comparing equation (35) with equation (34), we see that the elements of $\mathbf{v}$ are complex conjugates of those of $\mathbf{u}$. Hence $\mathbf{v}$ can be written as

$$\mathbf{v} = \xi - j\zeta. \tag{118}$$

Substituting equations (117) and (118) into equation (116) and then substituting the resulting equation into equation (113), we obtain after some steps,

$$\Delta \mathcal{E}_k = 4\left\{\cos\left[\alpha_k \frac{\partial}{\partial \rho} \mathcal{E}(\rho_k , \mathbf{E}_k)\right] - 1\right\}[\mathbf{E}'_k\xi \cos \rho_k + \mathbf{E}'_k\zeta \sin \rho_k] - 4 \sin\left[\alpha_k \frac{\partial}{\partial \rho} \mathcal{E}(\rho_k , \mathbf{E}_k)\right][\mathbf{E}'_k\zeta \cos \rho_k - \mathbf{E}'_k\xi \sin \rho_k]. \tag{119}$$

From equations (36), (117), and (118),

$$\begin{aligned} \frac{\partial}{\partial \rho} \mathcal{E}(\rho_k , \mathbf{E}_k) &= 2j\mathbf{E}'_k(\exp(-j\rho_k)\mathbf{u} - \exp(j\rho_k)\mathbf{v}), \\ &= 2j\mathbf{E}'_k(-2j \sin \rho_k\xi + 2j \cos \rho_k\zeta), \\ &= 4[\mathbf{E}'_k\xi \sin \rho_k - \mathbf{E}'_k\zeta \cos \rho_k]. \end{aligned} \tag{120}$$

It is clear from equation (56) that, if $(\partial/\partial\rho)\mathcal{E}(\rho_k , \mathbf{E}_k) = 0$, then $\rho_{k+1} = \rho_k$ and from equation (115), $\Delta\mathcal{E}_k = 0$. So we need to evaluate $\Delta\mathcal{E}_k$ only for $(\partial/\partial\rho)\mathcal{E}(\rho_k , \mathbf{E}_k) \neq 0$. It can be seen from equation (120) that, when $(\partial/\partial\rho)\mathcal{E}(\rho_k , \mathbf{E}_k) \neq 0$, $\mathbf{E}'_k\xi$ and $\mathbf{E}'_k\zeta$ cannot be simultaneously zero. Hence, we can define a quantity

$$\phi = \tan^{-1} \frac{\mathbf{E}_k'\zeta}{\mathbf{E}_k'\xi} \tag{121}$$

and a quantity

$$C = [(\mathbf{E}_k'\xi)^2 + (\mathbf{E}_k'\zeta)^2]^{\frac{1}{2}} > 0. \tag{122}$$

From equations (121) and (122), equation (120) can be rewritten as

$$\frac{\partial}{\partial \rho} \mathcal{E}(\rho_k , \mathbf{E}_k) = 4C \sin (\rho_k - \phi). \tag{123}$$

Substituting equation (123) into equation (119), using equations (121) and (122), and rearranging, we obtain, after a number of steps,

$$\Delta\mathcal{E}_k = 4C\{\cos [\rho_k - \phi - 4C\alpha_k \sin (\rho_k - \phi)] - \cos (\rho_k - \phi)\} \tag{124}$$

or

$$\Delta\mathcal{E}_k = 8C \sin [\rho_k - \phi - 2C\alpha_k \sin (\rho_k - \phi)] \sin [2C\alpha_k \sin (\rho_k - \phi)]. \tag{125}$$

It may be assumed that $0 \leqq \rho_k - \phi \leqq 2\pi$ (a factor of 2π or its multiple can be dropped). Since $(\partial/\partial\rho)\mathcal{E}(\rho_k , \mathbf{E}_k) \neq 0$, $\rho_k - \phi$ cannot be 0, π, or 2π [see equation (123)]. Hence, we have either

$$0 < \rho_k - \phi < \pi \tag{126}$$

or

$$\pi < \rho_k - \phi < 2\pi. \tag{127}$$

We shall consider equation (126) first.

As can be seen from equation (56), the constant α_k determines the size of each adjustment (note that $\alpha_k > 0$ is required so that the adjustments will be made in the negative gradient direction). We wish to determine the permissible range of α_k, that is, to determine a number δ such that $\Delta\mathcal{E}_k > 0$ for every α_k in the range $0 < \alpha_k < \delta$. It can be shown from equations (124) and (126) that δ can only be as large as $1/2C$; hence, the permissible range of α_k is

$$0 < \alpha_k < \frac{1}{2C}. \tag{128}$$

From equations (128) and (126), we obtain

$$0 < 2C\alpha_k \sin (\rho_k - \phi) < 1. \tag{129}$$

It can be easily shown from equation (129) that

$$\sin [2C\alpha_k \sin (\rho_k - \phi)] > \frac{1}{\pi} 4C\alpha_k \sin (\rho_k - \phi) > 0. \tag{130}$$

It can be shown from equations (126) and (128) that $[\rho_k - \phi - 2C\alpha_k \sin(\rho_k - \phi)]$ must be in the range

$$0 < [\rho_k - \phi - 2C\alpha_k \sin(\rho_k - \phi)] \leq \frac{\pi}{2} \tag{131}$$

or in the range

$$\frac{\pi}{2} < [\rho_k - \phi - 2C\alpha_k \sin(\rho_k - \phi)] < \pi. \tag{132}$$

When equation (131) holds,

$$\sin [\rho_k - \phi - 2C\alpha_k \sin(\rho_k - \phi)] \geqq \frac{2}{\pi} [\rho_k - \phi - 2C\alpha_k \sin(\rho_k - \phi)]. \tag{133}$$

Combining (133) and (130) and using equations (125) and (123), we obtain, after some simplification,

$$\Delta \mathcal{E}_k > \frac{2\alpha_k}{\pi^2} [2 - 4C\alpha_k] \left[\frac{\partial}{\partial \rho} \mathcal{E}(\rho_k , \mathbf{E}_k)\right]^2. \tag{134}$$

It can be shown in a similar manner that when equation (132) holds,

$$\Delta \mathcal{E}_k > \frac{8C\alpha_k^2}{\pi^2} \left[\frac{\partial}{\partial \rho} \mathcal{E}(\rho_k , \mathbf{E}_k)\right]^2. \tag{135}$$

It has been shown from equations (126) and (128) that either equation (134) or equation (135) holds. In a similar manner, it can be shown from equations (127) and (128) that either equation (134) or equation (135) holds. Thus, we have the conclusion:

Conclusion 2. When equation (128) holds, either equation (134) or equation (135) holds. Physically this means that, when equation (128) holds, $\Delta\mathcal{E}_k$ is positive and approaches zero only when $(\partial/\partial\rho)\mathcal{E}(\rho_k , \mathbf{E}_k)$ approaches zero.

It is not easy to determine whether equation (128) is satisfied in practice because the constant C depends on $\mathbf{E}_k$, ξ, and ζ and is somewhat difficult to compute. Hence, we wish to replace equation (128) with inequality (63) in Proposition 2.

From the definition of $\mathbf{v}_k$ in equation (114) and from equations (117) and (118), we can write

$$2\mathbf{E}_k'\mathbf{v}_k = 4[\cos \rho_k]\mathbf{E}_k'\xi + 4[\sin \rho_k]\mathbf{E}_k'\zeta. \tag{136}$$

From equations (136) and (22), one can show that

$$L_k - \mathcal{E}(\rho_k, \mathbf{E}_k) = 4\mathbf{E}_k'\xi \cos \rho_k + 4\mathbf{E}_k'\zeta \sin \rho_k \tag{137}$$

where

$$L_k = \mathbf{E}_k'\mathbf{y}\mathbf{E}_k + \sum_{h=-\infty}^{\infty} q_h^2 .$$

Note that the terms $\mathbf{E}_k'\mathbf{y}\mathbf{E}_k$, $\sum_{h=-\infty}^{\infty} q_h^2$, $\mathbf{E}_k'\xi$, and $\mathbf{E}_k'\zeta$, in equation (137) are independent of ρ. Furthermore, equation (137) holds for all ρ_k. Letting $\rho_k = 0$ in equation (137), we obtain

$$L_k - [\mathcal{E}(\rho_k, \mathbf{E}_k)]_{\rho_k=0} = 4\mathbf{E}_k'\xi \tag{138}$$

where $[\mathcal{E}(\rho_k, \mathbf{E}_k)]_{\rho_k=0}$ is $\mathcal{E}(\rho_k, \mathbf{E}_k)$ evaluated at $\rho_k = 0$. Since $\mathcal{E}(\rho_k, \mathbf{E}_k) \geqq 0$ for all ρ_k, we have from equation (138)

$$L_k \geqq 4\mathbf{E}_k'\xi. \tag{139}$$

It can be shown from equation (22) that

$$[\mathcal{E}(\rho_k, \mathbf{E}_k)]_{\rho_k=0} < 2L_k . \tag{140}$$

From equations (140) and (138),

$$-L_k < 4\mathbf{E}_k'\xi. \tag{141}$$

From equations (139) and (141),

$$(\mathbf{E}_k'\xi)^2 \leqq \tfrac{1}{16}L_k^2 . \tag{142}$$

We have obtained equation (142) by letting $\rho_k = 0$ in equation (137). If we let $\rho_k = \pi/2$ in equation (137), we obtain, in a similar manner,

$$(\mathbf{E}_k'\zeta)^2 \leqq \tfrac{1}{16}L_k^2 . \tag{143}$$

From equations (142), (143), and (122),

$$C \leqq \frac{1}{2(2)^{\frac{1}{2}}} L_k . \tag{144}$$

Using sampling theory, we can verify that

$$L_k = 2f_0\left[\int_{-\infty}^{\infty} s^2(t)\, dt\right]_{\mathbf{E}_k} + 2f_0 \int_{-\infty}^{\infty} q^2(t)\, dt \tag{145}$$

where $[\int_{-\infty}^{\infty} s^2(t)\, dt]_{\mathbf{E}_k}$ is $\int_{-\infty}^{\infty} s^2(t)\, dt$ evaluated at $\mathbf{E} = \mathbf{E}_k$.

It can be seen from equations (144) and (145) that, if equation (63) holds, equation (128) is satisfied. Thus, we have the conclusion:

Conclusion 3. When equation (63) holds, equation (128) holds and, from Conclusion 2, either equation (134) or equation (135) holds. Hence,

when equation (63) holds, $\Delta\mathcal{E}_k$ is positive and approaches zero only when $(\partial/\partial\rho)\mathcal{E}(\rho_k, \mathbf{E}_k)$ approaches zero.

Now, let us summarize the results presented in this appendix. It has been shown that when equation (62) holds in $\mathcal{K}_1$, $\Delta\mathcal{E}_k$ is bounded below by equation (112) (see Conclusion 1). It has also been shown that when equation (63) holds in $\mathcal{K}_2$, $\Delta\mathcal{E}_k$ is bounded below by either equation (134) or equation (135) (see Conclusion 3). It is seen from the lower bounds in equations (112), (134), and (135) that $\Delta\mathcal{E}_k$ is positive and approaches zero only when the partial derivatives approach zero. Since $\mathcal{K}_1$ and $\mathcal{K}_2$ each contain an infinite number of k's and since the mean-square error $\mathcal{E}$ cannot reduce without bound, $\mathcal{E}$ must converge to a stationary point. This proves Proposition 2 in Section VI.

REFERENCES

1. Lucky, R. W., Salz, J., and Weldon, E. J., Jr., *Principles of Data Communication*, New York: McGraw-Hill, 1968, pp. 128–165, 170.
2. Bennett, W. R., and Davey, J. R., *Data Transmission*, New York: McGraw-Hill, 1965, pp. 269–273.
3. Lucky, R. W., and Rudin, H. R., "An Automatic Equalizer for General-Purpose Communication Channels," B.S.T.J., *46*, No. 9 (November 1967), pp. 2179–2208.
4. Wilde, D. J., and Beightler, C. S., *Foundations of Optimization*, Englewood Cliffs, N. J.: Prentice-Hall, 1967, pp. 24–27, 296.
5. Goldstein, A. A., *Constructive Real Analysis*, New York: Harper and Row, 1967, pp. 26–39.
6. Marcus, M., and Minc, H., *A Survey of Matrix Theory and Matrix Inequalities*, Boston: Allyn and Bacon, 1964, pp. 139–148.
7. Gersho, A., "Adaptive Equalization of Highly Dispersive Channels for Data Transmission," B.S.T.J., *48*, No. 1 (January 1969), pp. 55–70.
8. Becker, F. K., Kretzmer, E. R., and Sheehan, J. R., "A New Signal Format for Efficient Data Transmission," B.S.T.J., *45*, No. 5 (May–June 1966), pp. 755–758.
9. Kretzmer, E. R., "Generalization of a Technique for Binary Data Communication," IEEE Trans. on Comm. Tech., *COM-14*, No. 1 (February 1966), pp. 67–68.
10. Alexander, A. A., Gryb, R. M., and Nast, D. W., "Capabilities of the Telephone Network for Data Transmission," B.S.T.J., *39*, No. 3 (May 1960), pp. 431–476.
11. Gibby, R. A., "An Evaluation of AM Data System Performance by Computer Simulation," B.S.T.J., *39*, No. 3 (May 1960), pp. 675–704.
12. Gersho, A., "Solving Nonlinear Network Equations Using Optimization Techniques," B.S.T.J., *48*, No. 9 (November 1969), pp. 3135–3138.

A General Approach to Twin-T Design and Its Application to Hybrid Integrated Linear Active Networks

By G. S. MOSCHYTZ

(Manuscript received January 2, 1970)

In this paper we approach twin-T design with a view to controlling the sensitivity of the transmission zero with respect to component variations, according to criteria that are of particular interest in the design of hybrid integrated linear active networks. We give design examples and derive conditions that relate null depth and component characteristics with expected zero displacement in the s-plane.

I. INTRODUCTION

In 1934, H. W. Augustadt invented the twin-T network while carrying out investigations for an economical rectifier filter for phonograph amplifiers.[1] The two main fields of application of the twin-T network were introduced in 1938 by H. H. Scott who discussed its uses as a feedback network to obtain highly selective amplifiers and stable oscillators.[2] In the following years the circuit was thoroughly analyzed in the unloaded state[3-6] and, later, in the loaded state when driven from a nonideal voltage source.[7-9] Consideration was also given to the network's selectivity properties and to the effects of loading and network asymmetry.[10-12] In the early 1960s, a somewhat new application was introduced for the twin-T when synthesis methods based on root locus techniques were developed to employ the twin-T as a compensation network in dc servo systems.[13-15] These investigations were limited to the symmetrical twin-T with fixed source and load resistances. They were later expanded to include wide ranges of source and load impedances[16] and to provide prescribed pole-zero locations[17] using parameter plane techniques.

Recently, with the advent of linear integrated circuits, interest in the twin-T network has been revived yet again, this time by network theoreticians attempting to generate, by RC network synthesis tech-

niques, a still wider range of pole-zero configurations than had hitherto been possible.[18–21] At the same time, numerous methods of active RC filter synthesis were developed that rely on the basic frequency characteristics of a twin-T network or modifications thereof, to provide the required filtering properties.[22–30] These methods depend, for their frequency stability on the stability of the twin-T network. To ensure a very high degree of stability the twin-T has been realized by tantalum thin film components and then combined with silicon integrated active circuits to produce hybrid integrated filter networks.[31] In applications of this kind null frequency and null depth tuning procedures become very critical, particularly because thin film resistors can only be adjusted in the increasing direction; furthermore the null characteristics (gain and phase) become more important than in more conventional applications, and adjustments of these characteristics should not only be possible but also simple. It is with respect to these problems that the twin-T network is reexamined once again here.

The requirement that the six components of a twin-T network provide a perfect null, that is, a pair of imaginary zeros, at a particular frequency imposes only two design constraints on the network. A third results from the impedance scaling factor chosen for the network. Thus, three parameters remain to be chosen by whatever criteria seem most important for a given application. Most often circuit simplicity dominates this choice, resulting in the symmetrical twin-T. In other instances, practical considerations requiring that either all the resistors or all the capacitors be equal will determine the choice. In those cases where the network is synthesized to provide other than standard pole-zero locations, no choice exists at all, since all the network parameters are generally accounted for.

In this paper, we select the three unconstrained design parameters in such a way as to control the null characteristics of the twin-T according to criteria of particular importance in the design of linear active networks. In such networks the twin-T is generally part of a positive or negative feedback configuration whose closed loop poles are closely tied to the open loop zeros on the $j\omega$-axis. The latter are generated by the transmission null of the twin-T network. The higher the Q of the network, the closer the tie between the closed loop poles and the open loop zeros and, consequently, the more critical the sensitivity and stability of the twin-T transmission null. To obtain a measure for both, the zero sensitivity functions for the commonly used and for the general twin-T configurations are derived first. By selecting the three design parameters remaining in the sensitivity functions of the general twin-T

appropriately, it is found that a relatively wide range of sensitivity criteria can be met. Some of these are useful in contributing to the overall stability of an active feedback network incorporating a twin-T. Others are of interest in considerations pertaining to useful frequency and null depth tuning strategies in the vicinity of a perfect twin-T null.

To guarantee stability, conditions are also derived here that prevent the twin-T transmission zeros from drifting to the right half s-plane. This implies a maximum permissible null depth of the twin-T that can be expressed in terms of the twin-T design parameters and the temperature and aging characteristics of its components.

II. CIRCUIT ANALYSIS OF THE GENERAL TWIN-T

The voltage transfer function of the general twin-T shown in Fig. 1 is given by the ratio of two third-order polynomials, namely

$$\frac{E_2}{E_1} = T(s) = \frac{N(s)}{D(s)} = \frac{1 + a_1 s + a_2 s^2 + a_3 s^3}{1 + b_1 s + b_2 s^2 + b_3 s^3} \tag{1}$$

where

$$a_1 = R_3(C_1 + C_2), \tag{2a}$$

$$a_2 = R_3(R_1 + R_2)C_1C_2 , \tag{2b}$$

$$a_3 = R_1R_2R_3C_1C_2C_3 , \tag{2c}$$

$$b_1 = R_3(C_1 + C_2) + R_2C_2 + R_1(C_2 + C_3), \tag{2d}$$

$$b_2 = R_3[R_1C_3(C_1 + C_2) + (R_1 + R_2)C_1C_2] + R_1R_2C_2C_3 , \tag{2e}$$

$$b_3 = R_1R_2R_3C_1C_2C_3 = a_3 , \tag{2f}$$

and

$$s = \sigma + j\omega.$$

The null, or transmission zero, of the twin-T is defined by the roots

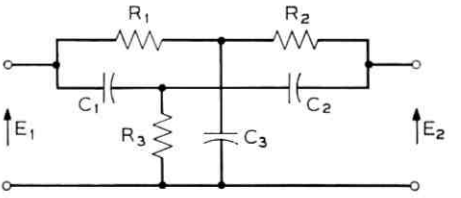


Fig. 1—General twin-T network.

of $N(s)$ on the imaginary axis. Thus for

$$N(s)\,|_{s=j\omega} = 0 \tag{3}$$

the null frequency or conjugate complex transmission zeros $z_{1,2} = \pm j\omega_N$ are obtained.

Using the substitutions

$$C_s = \frac{C_1 C_2}{C_1 + C_2}, \tag{4a}$$

$$C_p = C_1 + C_2, \tag{4b}$$

$$R_s = R_1 + R_2, \tag{4c}$$

$$R_p = \frac{R_1 R_2}{R_1 + R_2}, \tag{4d}$$

the following two conditions for the twin-T null frequency result from equation (3)

$$\omega_N^2 = \frac{1}{R_1 R_2 C_s C_3} \tag{5}$$

and

$$\frac{C_3}{R_3} = \frac{C_p}{R_p}. \tag{6}$$

These can be combined as follows

$$\omega_N^2 = \frac{1}{R_1 R_2 C_s C_3} = \frac{1}{R_s R_3 C_1 C_2}. \tag{7}$$

Thus, for a perfect null the transmission zero of each of the two bridged-Ts obtained by disconnecting R_3 and C_3, respectively, of the twin-T (see Fig. 2) must be equal. This fact has been used to develop a 2-step tuning method for the twin-T.[32] Substituting equations (5) and (6) into equation (1) gives the transfer polynomials of the nulled twin-T

$$N(s) = \frac{R_3 C_p}{\omega_N^2}\left(s + \frac{1}{R_3 C_p}\right)(s^2 + \omega_N^2) \tag{8}$$

and

$$D(s) = \frac{R_3 C_p}{\omega_N^2}\left(s + \frac{1}{R_s C_p}\right)\left[s^2 + \left(\omega_N^2 R_1 C_3 + \frac{1}{C_1 R_3}\right)s + \omega_N^2\right]. \tag{9}$$

The open-circuit impedance matrix of the perfectly nulled twin-T

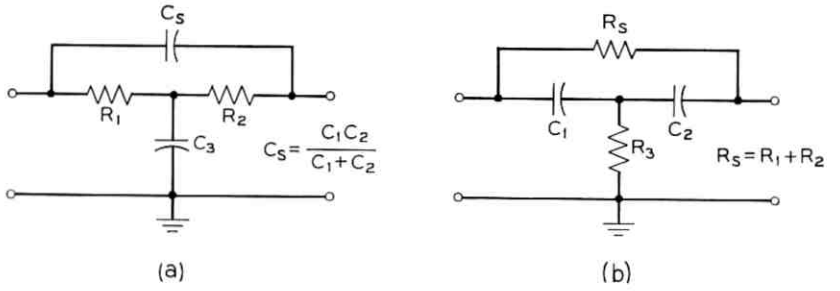


Fig. 2—Bridged-T networks derived from general twin-T shown in Fig. 1, whose natural frequencies coincide with twin-T null frequency. (a) R_3 disconnected; (b) C_3 disconnected.

is given in Appendix A. The voltage transfer function follows from equations (8) and (9), namely

$$T_N(s) = \frac{s^2 + \omega_N^2}{s^2 + \frac{\omega_N}{q_N} s + \omega_N^2} \tag{10}$$

where ω_N is given by equation (7),

$$q_N = \frac{\omega_N}{2\sigma_N} = \frac{1}{\omega_N(R_1C_3 + R_2C_2)} \tag{11}$$

and

$$2\sigma_N = \omega_N^2 R_1 C_3 + \frac{1}{R_3C_1} = \frac{1}{R_2C_2} + \frac{1}{R_3C_1}. \tag{12}$$

Inspection of equations (8) and (9) shows that the two third-order polynomials of the transfer function of a general twin-T are simplified by one degree due to pole-zero cancellation when the conditions for a perfect null are satisfied. It is shown in Appendix B that even when the twin-T null is *not* perfect (that is, the transmission zeros are not on but only close to the $j\omega$-axis), this pole-zero cancellation still takes place, provided that $R_1C_1 = R_2C_2$.

The most frequently used twin-T is structurally (and electrically) symmetrical (see Appendix A). For this case (see Fig. 3a) $R_1 = R_2 = R$, $R_3 = R/2$, $C_1 = C_2 = C$, $C_3 = 2C$, and the coefficients of equation (10) are $\omega_N = 1/RC$, $2\sigma_N = 4/RC$ and $q_N = \frac{1}{4}$. Another commonly used version of the twin-T is the potentially symmetrical configuration (see Appendix A). This is obtained by impedance scaling one-half of the symmetrical twin-T by some factor ρ. The resulting twin-T elements are

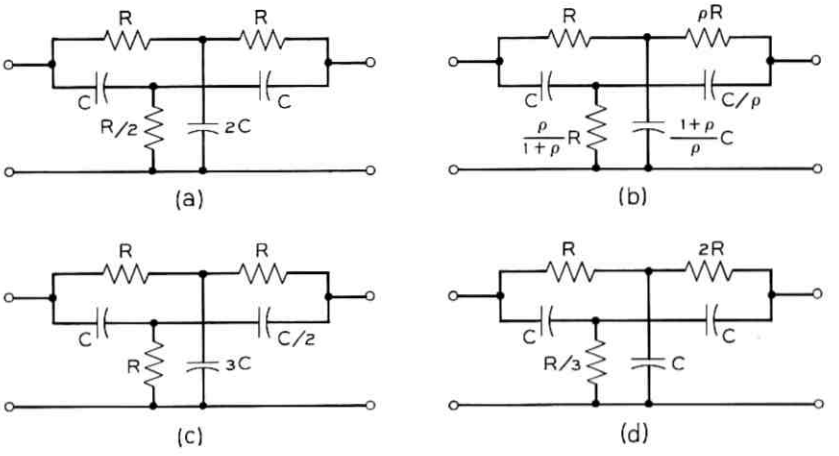


Fig. 3—Frequently used twin-T configurations. (a) Symmetrical; (b) Potentially symmetrical; (c) Equal resistors; (d) Equal capacitors.

(see Fig. 3b) $R_1 = R$, $R_2 = \rho R$, $R_3 = \rho R/(1 + \rho)$, $C_1 = C$, $C_2 = C/\rho$, and $C_3 = (1 + \rho)C/\rho$. The coefficients of the transfer function [equation (10)] for this case are $\omega_N = 1/RC$, $2\sigma_N = (2/RC)(1 + \rho)/\rho$ and $q_N = \frac{1}{2}\rho/(1 + \rho)$. Notice that for the extreme asymmetrical case for which $\rho \gg 1$, q_N takes on its maximum value of 0.5.

Sometimes it is useful to make all the resistors of the twin-T equal. This enables one to gang three variable resistors in order to vary the null frequency. The twin-T elements are then (see Fig. 3c) $R_1 = R_2 = R_3 = R$, $C_1 = C$, $C_2 = C/2$, $C_3 = 3C$, and the coefficients of the transfer function are the same as those of the symmetrical twin-T. Similarly, if the three capacitors are to be made equal for ganging or other purposes, the twin-T elements are (see Fig. 3d) $R_1 = R$, $R_2 = 2R$, $R_3 = R/3$. Here again the coefficients of the transfer function are the same as those of the symmetrical twin-T.

III. SENSITIVITY OF TWIN-T NULL CHARACTERISTICS TO COMPONENT VARIATIONS

The null or zero sensitivity of the twin-T to variations of any component x gives a measure for the degree of change of the transmission characteristics in the vicinity of the twin-T null frequency as a result of variations of a component x. Referring to the complex frequency plane, the zero (or pole) sensitivity gives a measure for the zero (or pole) displacement due to an incremental change in the value of the component

x. It is defined by[33]

$$S_x^z = \frac{dz}{dx/x} \tag{13}$$

where z is the complex null frequency of the network.

The zero displacement dz in the s-plane has a real and an imaginary component. It defines a vector or the direction in which a zero (or pole) travels from its initial location with incremental changes of a component x. Since x and therefore dx/x must be real, the zero (pole) sensitivity defines a vector in the same direction as the zero (pole) displacement dz. Herein lies the importance of knowing the root (that is, pole or zero) sensitivity of a network since it provides insight into the stability of a system with respect to the component x that is expected to vary the most. It also provides information relevant to network tuning since it relates adjustments of any component x to its effect on the roots of the transfer function. Conversely, as we shall see later, a network can be designed to provide a prescribed sensitivity between some parameter of the transfer function such as the displacement of a specific transmission zero and the variation of some preselected component x. The choice of sensitivity may be such as to result in a network that responds to a simplified tuning strategy or whose characteristics may be adjusted in a desirable way by the component x.

The most important characteristic of the twin-T is its behavior in the region of the frequency null. In the s-plane this behavior is characterized by the sensitivity of the transmission zero, which is initially located on the imaginary axis for a perfect null. The sensitivities of this zero, that is, $z = j\omega_N$, to each of the six components of the general twin-T have therefore been derived in Appendix C and listed in Table I. In doing so it has been found useful to characterize the general twin-T by the following parameters

$$\lambda = \frac{R_1}{R_1 + R_2}, \tag{14}$$

$$\nu = \frac{C_2}{C_1 + C_2}, \tag{15}$$

and

$$\alpha = \frac{\omega_1}{\omega_N} = \left(\frac{R_a C_a}{R_b C_b}\right)^{\frac{1}{2}}. \tag{16}$$

λ and ν give a measure for the degree of symmetry of the series elements of the twin-T; α relates the series elements to the shunt elements.

TABLE I—ZERO-SENSITIVITY FUNCTIONS FOR A GENERAL NULLED TWIN-T

$S_{R_1}^{j\omega_N} = \frac{\alpha\omega_N}{2(1+\alpha^2)}\left[1-\lambda-j\left(\frac{1}{\alpha}+\alpha\lambda\right)\right]$	$S_{C_1}^{j\omega_N} = -\frac{\alpha\omega_N}{2(1+\alpha^2)}\left[1-\nu+j\left(\alpha+\frac{\nu}{\alpha}\right)\right]$
$S_{R_2}^{j\omega_N} = \frac{\alpha\omega_N}{2(1+\alpha^2)}\left[\lambda-j\left\{\frac{1}{\alpha}+\alpha(1-\lambda)\right\}\right]$	$S_{C_2}^{j\omega_N} = -\frac{\alpha\omega_N}{2(1+\alpha^2)}\left[\lambda+j\left\{\alpha+\frac{(1-\nu)}{\alpha}\right\}\right]$
$S_{R_3}^{j\omega_N} = -\frac{\alpha\omega_N}{2(1+\alpha^2)}(1+j\alpha)$	$S_{C_3}^{j\omega_N} = \frac{\alpha\omega_N}{2(1+\alpha^2)}\left(1-\frac{j}{\alpha}\right)$

where:

$$\omega_N^2 = \frac{1}{R_1R_2C_sC_3} = \frac{1}{R_sR_3C_1C_2} \qquad R_s = R_1 + R_2\ ; \qquad R_p = \frac{R_1R_2}{R_1+R_2}$$

$$\omega_1 = \frac{1}{R_3C_p} = \frac{1}{R_pC_3} \qquad C_s = \frac{C_1C_2}{C_1+C_2}\ ; \qquad C_p = C_1 + C_2\ .$$

$$\alpha = \frac{\omega_1}{\omega_N}\ ; \qquad \lambda = \frac{R_1}{R_1+R_2}\ ; \qquad \nu = \frac{C_2}{C_1+C_2}$$

A useful check for the validity of the expressions in Table I is that they must satisfy the following condition for the root sensitivity of a passive RC network if the relative resistor and capacitor changes are assumed to track[33]

$$\sum \mathbb{S}^{z}_{R_i} = \sum \mathbb{S}^{z}_{C_i} = -z. \tag{17}$$

Let us now consider the zero sensitivity of the commonly used twin-T configurations shown in Fig. 3.

3.1 *Symmetrical Twin-T*

This case is characterized by the parameters $\omega_N = 1/RC$, $\alpha = 1$, and $\lambda = \nu = 0.5$. The resulting zero sensitivity functions are listed in Table II, Part 1 and the corresponding zero displacements in the complex frequency plane are shown graphically in Fig. 4a*. This diagram also demonstrates the realization of the condition for passive RC networks given by equation (17). Therefore, by assuming tracking and equal but opposite temperature coefficients of the resistors and capacitors, temperature drift of the null frequency can theoretically be eliminated completely. If tracking of like component variations does not occur, the drift displacement due to the symmetrical elements R_1 and R_2 and of C_1 and C_2 are identical. The displacements due to the shunt elements R_3 and C_3 have approximately the same value but follow a somewhat less steep slope. Thus if the twin-T is being used in a feedback network to generate conjugate complex poles in the left half plane close to the $j\omega$-axis, the stability of the network will be more sensitive to drift in the shunt elements than to drift of those in series.

3.2 *Potentially Symmetrical Twin-T*

This case is characterized by the network parameters $\omega_N = 1/RC$, $\alpha = 1$, and $\lambda = \nu = 1/(1 + \rho)$. The resulting zero-sensitivity functions are listed in Table II, Part 2. Since they depend on the symmetry coefficient ρ, some control on the sensitivity can be expected by this coefficient. Table II, Parts 2a and 2b list the sensitivity functions for the two extremes, that is, when ρ is much larger and much smaller than unity, respectively. The corresponding zero displacements are shown in Fig. 4b. These two complementary cases can be thought of as having evolved from the symmetrical case (Fig. 4a) by spreading out the displacement

* As pointed out earlier, the root sensitivity function given by equation (13) defines a differential vector in the complex s-plane. It can be shown that this vector lies on the branch of the root locus with respect to a component x that starts at z or, in other words, that the root displacement Δz and the root sensitivity have the same argument.

TABLE II—ZERO-SENSITIVITY FUNCTIONS FOR COMMONLY USED TWIN-T CONFIGURATIONS

1. Symmetrical Twin-T

$(R_1 = R_2 = 2R_3 = R;\ C_1 = C_2 = C_3/2 = C)$: $\omega_N = \dfrac{1}{RC}$; $q_N = \frac{1}{4}$; $\alpha = 1$; $\lambda = \nu = \frac{1}{2}$; $\gamma = \frac{1}{4}$

$\mathbb{S}_{R_1}^{j\omega_N} = \mathbb{S}_{R_2}^{j\omega_N} = \dfrac{\omega_N}{4}\left(\frac{1}{2} - \frac{3}{2}j\right) = 0.395\,\omega_N/\underline{-71°30'}$	$\mathbb{S}_{C_1}^{j\omega_N} = \mathbb{S}_{C_2}^{j\omega_N} = -\dfrac{\omega_N}{4}\left(\frac{1}{2} + \frac{3}{2}j\right) = -0.395/\underline{71°30'}$
$\mathbb{S}_{R_3}^{j\omega_N} = -\dfrac{\omega_N}{4}(1 + j) = -0.354/\underline{45°}$	$\mathbb{S}_{C_3}^{j\omega_N} = \dfrac{\omega_N}{4}(1 - j) = 0.354/\underline{-45°}$

2. Potentially Symmetrical Twin-T

$\begin{bmatrix} R_1 = R_2/\rho = R_3(1+\rho)/\rho = R \\ C_1 = \rho C_2 = C_3\rho/(1+\rho) = C \end{bmatrix}$: $\omega_N = \dfrac{1}{RC}$; $q_N = \dfrac{1}{2}\left(\dfrac{\rho}{1+\rho}\right)$; $\alpha = 1$; $\lambda = \nu = \dfrac{1}{1+\rho}$; $\gamma = \dfrac{\rho}{(1+\rho)^2}$

$\mathbb{S}_{R_1}^{j\omega_N} = \dfrac{\omega_N}{4}\left(\dfrac{\rho}{1+\rho} - j\dfrac{2+\rho}{1+\rho}\right)$	$\mathbb{S}_{C_1}^{j\omega_N} = -\dfrac{\omega_N}{4}\left(\dfrac{\rho}{1+\rho} + j\dfrac{2+\rho}{1+\rho}\right)$
$\mathbb{S}_{R_2}^{j\omega_N} = \dfrac{\omega_N}{4}\left(\dfrac{1}{1+\rho} - j\dfrac{1+2\rho}{1+\rho}\right)$	$\mathbb{S}_{C_2}^{j\omega_N} = -\dfrac{\omega_N}{4}\left(\dfrac{1}{1+\rho} + j\dfrac{1+2\rho}{1+\rho}\right)$
$\mathbb{S}_{R_3}^{j\omega_N} = -\dfrac{\omega_N}{4}(1 + j)$	$\mathbb{S}_{C_3}^{j\omega_N} = \dfrac{\omega_N}{4}(1 - j)$

TABLE II—Cont'd

2A. Potentially Symmetrical Twin-T $(\mathbb{S} \gg 1)$: $(\rho \gg 1)$: $q_N \to \frac{1}{2}$: $\alpha = 1$; $\lambda = \nu \to 0$; $\gamma \to 0$

$\mathbb{S}_{R_1}^{j\omega_N} \approx \frac{\omega_N}{4}(1 - j) = 0.354\,\omega_N/\underline{-45°}$	$\mathbb{S}_{C_1}^{j\omega_N} \approx -\frac{\omega_N}{4}(1 + j) = -0.354/\underline{45°}$
$\mathbb{S}_{R_2}^{j\omega_N} \approx -j\frac{\omega_N}{2} = 0.5\,\omega_N/\underline{-90°}$	$\mathbb{S}_{C_2}^{j\omega_N} \approx -j\frac{\omega_N}{2} = 0.5\,\omega_N/\underline{-90°}$
$\mathbb{S}_{R_3}^{j\omega_N} = -\frac{\omega_N}{4}(1 + j) = -0.354/\underline{45°}$	$\mathbb{S}_{C_3}^{j\omega_N} = \frac{\omega_N}{4}(1 - j) = 0.354/\underline{-45°}$

2B. Potentially Symmetrical Twin-T $(\mathbb{S} \ll 1)$: $(\rho \ll 1)$: $q_n \to 0$; $\alpha = 1$; $\lambda = \nu \to 1$; $\gamma \to 0$

$\mathbb{S}_{R_1}^{j\omega_N} \approx -j\frac{\omega_N}{2} = 0.5\,\omega_N/\underline{-90°}$	$\mathbb{S}_{C_1}^{j\omega_N} \approx -j\frac{\omega_N}{2} = 0.5\,\omega_N/\underline{-90°}$
$\mathbb{S}_{R_2}^{j\omega_N} \approx \frac{\omega_N}{4}(1 - j) = 0.354/\underline{-45°}$	$\mathbb{S}_{C_2}^{j\omega_N} \approx -\frac{\omega_N}{4}(1 + j) = -0.354/\underline{45°}$
$\mathbb{S}_{R_3}^{j\omega_N} = -\frac{\omega_N}{4}(1 + j) = -0.354/\underline{45°}$	$\mathbb{S}_{C_3}^{j\omega_N} = \frac{\omega_N}{4}(1 - j) = 0.354/\underline{-45°}$

TABLE II—Cont'd

3. Twin-T with Equal Resistors

$(R_1 = R_2 = R_3 = R;\ C_1 = 2C_2 = C_3/3 = C)$: $\omega_N = \dfrac{1}{RC}$; $q_N = \frac{1}{4}$; $\alpha = \frac{2}{3}$; $\lambda = \frac{1}{2}$; $\nu = \frac{1}{3}$; $\gamma = \frac{1}{9}$

$\mathbb{S}_{R_1}^{j\omega_N} = \mathbb{S}_{R_2}^{j\omega_N} = \dfrac{\omega_N}{13}(1.5 - 5.5j) = 0.44\,\omega_N/\underline{-75°}$	$\mathbb{S}_{C_1}^{j\omega_N} = -\dfrac{\omega_N}{13}(2 + 3.5j) = -0.276\,\omega_N/\underline{60°}$
$\mathbb{S}_{R_3}^{j\omega_N} = -\dfrac{\omega_N}{13}(3 + 2j) = -0.278\,\omega_N/\underline{34°}$	$\mathbb{S}_{C_2}^{j\omega_N} = -\dfrac{\omega_N}{13}(1 + 5j) = -0.392\,\omega_N/\underline{79°}$
	$\mathbb{S}_{C_3}^{j\omega_N} = \dfrac{\omega_N}{13}(3 - 4.5j) = 0.416\,\omega_N/\underline{-56°}$

4. Twin-T with Equal Capacitors

$(R_1 = R_2/2 = 3R_3 = R;\ C_1 = C_2 = C_3 = C)$: $\omega_N = \dfrac{1}{RC}$; $q_N = \frac{1}{4}$; $\alpha = \frac{3}{2}$; $\lambda = \frac{1}{3}$; $\nu = \frac{1}{2}$; $\gamma = \frac{1}{2}$

$\mathbb{S}_{R_1}^{j\omega_N} = \dfrac{\omega_N}{13}(2 - 3.5j) = 0.276\,\omega_N/\underline{-60°}$	$\mathbb{S}_{C_1}^{j\omega_N} = \mathbb{S}_{C_2}^{j\omega_N} = -\dfrac{\omega_N}{13}(1.5 + 5.5j) = -0.44\,\omega_N/\underline{75°}$
$\mathbb{S}_{R_2}^{j\omega_N} = \dfrac{\omega_N}{13}(1 - 5j) = 0.392\,\omega_N/\underline{-79°}$	$\mathbb{S}_{C_3}^{j\omega_N} = \dfrac{\omega_N}{13}(3 - 2j) = 0.278\,\omega_N/\underline{-34°}$
$\mathbb{S}_{R_3}^{j\omega_N} = -\dfrac{\omega_N}{13}(3 + 4.5j) = -0.416\,\omega_N/\underline{56°}$	

vectors corresponding to the series components R_1, R_2 and C_1, C_2, respectively, and leaving the vectors corresponding to the shunt components R_3 and C_3 unchanged. The noteworthy feature of these two extreme cases is that they both provide for null frequency adjustment (if only over differentially small frequency ranges) by a single component, namely by R_2 or C_2 when $\rho \gg 1$ and by R_1 or C_1 when $\rho \ll 1$. Of the two cases, the former is preferable since it simultaneously provides higher selectivity than a symmetrical twin-T (that is, q_N approaches its maximum value). More will be said about zero-sensitivity and its effect on network adjustability in Section IV.

3.3 *Twin-T With Equal Resistors*

For this case $\omega_N = 1/RC$, $\alpha = \frac{2}{3}$, $\lambda = \frac{1}{2}$, and $\nu = \frac{1}{3}$. The corresponding zero-sensitivity functions are given in Table II, Part 3, and the zero displacements are shown graphically in Fig. 4c. Since the series resistors are equal in this case, the corresponding sensitivity functions are also equal. However, the sensitivity function of the shunt resistor is smaller in value and flatter in slope. Therefore, in order to shift the null frequency accurately, a high precision ganged 3-resistor potentiometer must be used whose resistor values track very closely.

3.4 *Twin-T With Equal Capacitors*

Here we have the design parameters $\omega_N = 1/RC$, $\alpha = \frac{3}{2}$, $\lambda = \frac{1}{3}$, and $\nu = \frac{1}{2}$. The corresponding sensitivity functions are listed in Table II, Part 4. Since this case is the dual of the equal-resistor case discussed above, the displacement vectors are negative and conjugate to those of Fig. 4c. Generally high precision ganged resistors are more readily available than capacitors, so that for variable null-frequency tuning purposes this case is less practical than the preceding one.

IV. NOVEL TWIN-T NETWORKS WITH PRESCRIBED TUNING CHARACTERISTICS

In the preceding section, the null sensitivity of component variations was discussed with respect to the most commonly used twin-T configurations. In this section the expressions for the general nulled twin-T, that is, those satisfying only the null conditions given by equations (5) and (6), are reexamined in relation to the corresponding zero sensitivity functions listed in Table I. In particular we investigate how the remaining twin-T parameters that are not constrained by the two null conditions can be utilized to modify the dependence of the null characteristics to adjustments of certain twin-T components in such a way as to

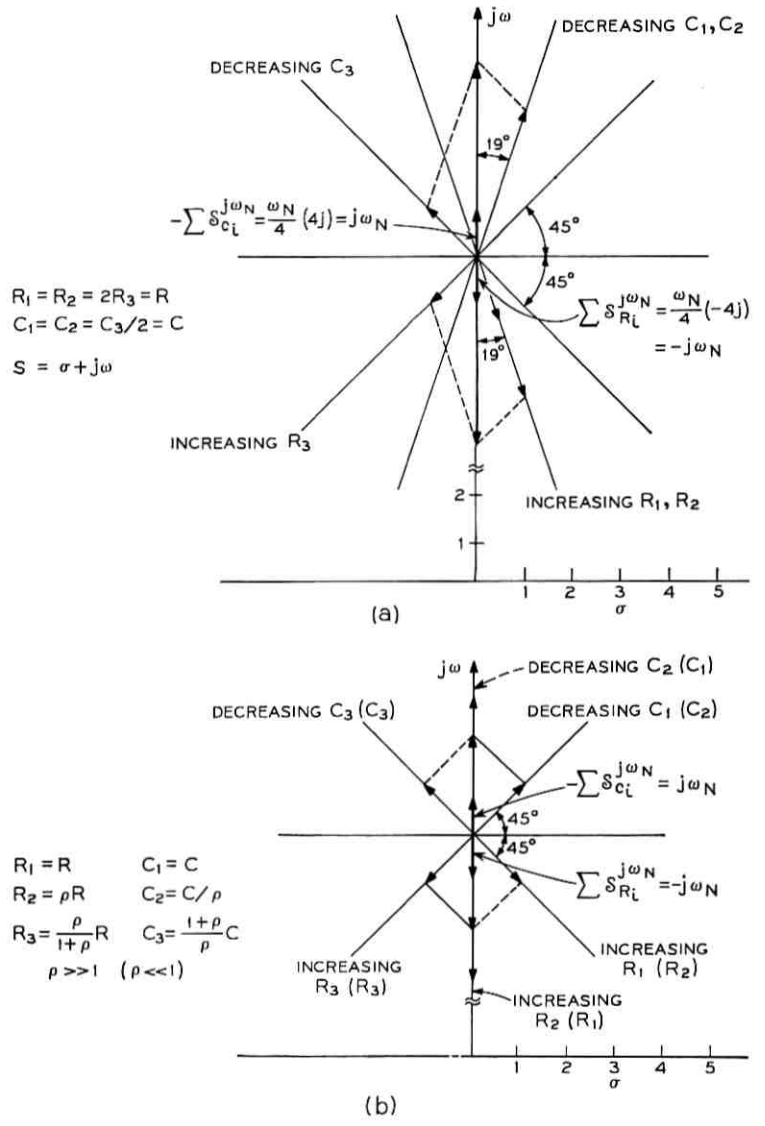
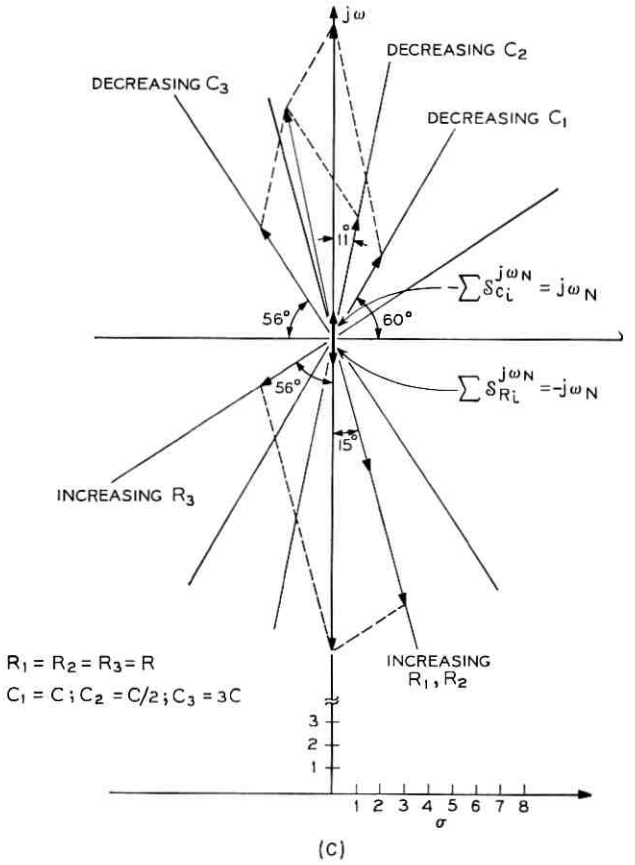


Fig. 4—Zero displacements for frequently used twin-T configurations. (a) Symmetrical; (b) Potentially symmetrical; (c) Equal resistors.



(C)

satisfy various tuning strategies that are particularly useful in practical applications.

The dependence of twin-T transmission characteristics in the vicinity of the null frequency on variations of any component x are essentially determined by the zero sensitivity functions listed in Table I. Design equations for twin-T networks providing a specified dependence of null characteristics on the adjustment of a particular component x may, therefore, be found by setting corresponding constraints on the zero sensitivity functions and solving the resulting equations for the twin-T components. Instead of designing a twin-T to a specified *dependence of transmission characteristics in the vicinity of the null frequency* to variations of a given component x, it therefore suffices for us to design a twin-T to a specified *zero sensitivity* with respect to x.

The expressions listed in Table I show that, after the null frequency ω_N has been specified, there are basically three parameters left to determine in order to design a twin-T to a prescribed zero sensitivity. These are the frequency ratio α, the resistor ratio λ, and the capacitor ratio ν. A parameter that sometimes provides clearer physical insight into the design of a twin-T than the frequency ratio α is the ratio of series to shunt capacitors γ, namely

$$\gamma = \frac{C_s}{C_3} = \nu \frac{C_1}{C_3}. \tag{18}$$

The two parameters are related by the expression

$$\alpha = \left[\frac{\gamma}{\lambda(1 - \lambda)}\right]^{\frac{1}{2}} \tag{19}$$

α has been plotted in Fig. 5 as a function of λ with the parameter γ. The values of γ for the common twin-T configurations are included in Table II. From the defining equations, the limits on the four twin-T parameters are

$$0 < \alpha < \infty, \tag{20a}$$

$$0 < \gamma < \infty, \tag{20b}$$

$$0 < \lambda < 1, \tag{20c}$$

$$0 < \nu < 1. \tag{20d}$$

A fundamental characteristic of the twin-T is its ability to reject a narrow frequency band centered at the null frequency f_N and to pass, substantially unattenuated, the frequencies on either side of this band. A useful parameter characterizing the selectivity of frequency rejection is the inverse damping factor q_N [see equation (11)] also known as the pole Q. Physically q_N is the ratio of the center frequency f_N divided by the bandwidth at which 3 dB attenuation occurs* (see Fig. 6a).

It is important, while examining the effects of the parameters listed in equation (20) on the zero sensitivity of the twin-T to keep an eye on their effect on the twin-T selectivity as expressed by q_N. Obviously, poor selectivity might be too high a price to pay for any set of controlled sensitivity functions. On the other hand, because the twin-T is a passive RC network, the selectivity factor q_N only has a narrow range of realiza-

* This definition is only accurate for unloaded twin-T networks such as those being considered here. For the case of a loaded twin-T with an unsymmetrical frequency response, it has been found more useful to define selectivity as the slope of the phase ϕ at the null frequency, that is, by $\tau(\omega_N) = d\phi/d\omega|_{\omega=\omega_N}$.

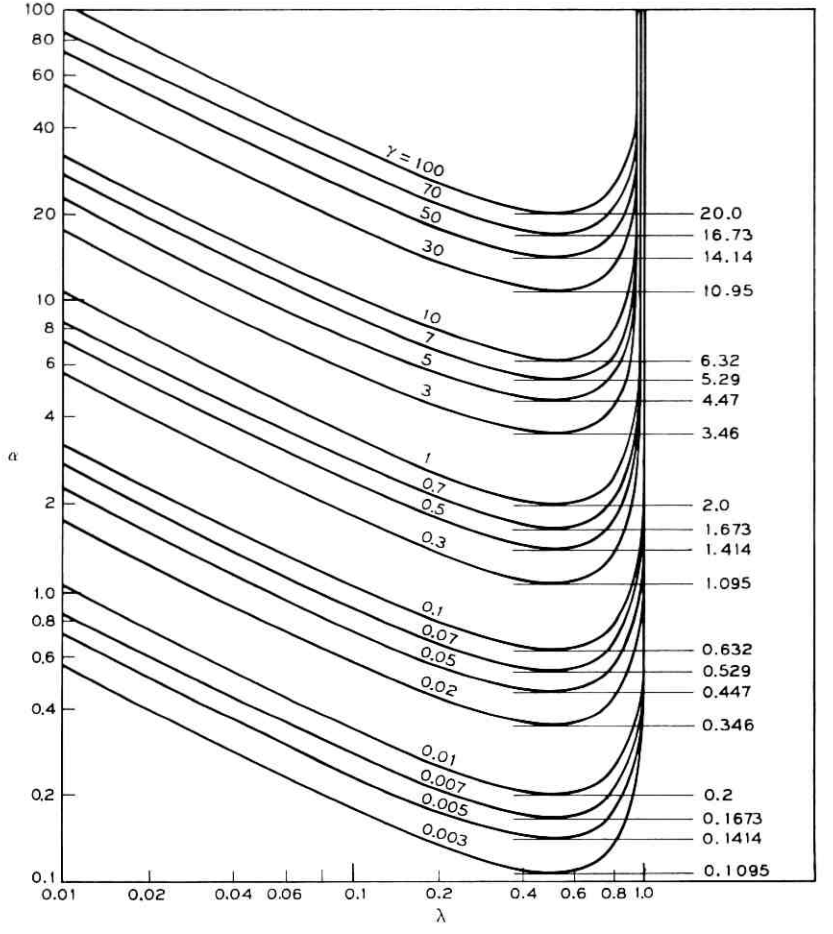


Fig. 5—Frequency ratio α as a function of λ and parameter γ.

bility as it is, that is,

$$0 < q_N < 0.5 \tag{21}$$

whereby the value 0.25 is realized the most frequently, namely with the symmetrical as well as with the equal resistor or equal capacitor twin-T configurations. However even within the limited range given by equation (21), the difference in actual frequency selectivity can be quite significant. This is illustrated in Fig. 6b where twin-T frequency

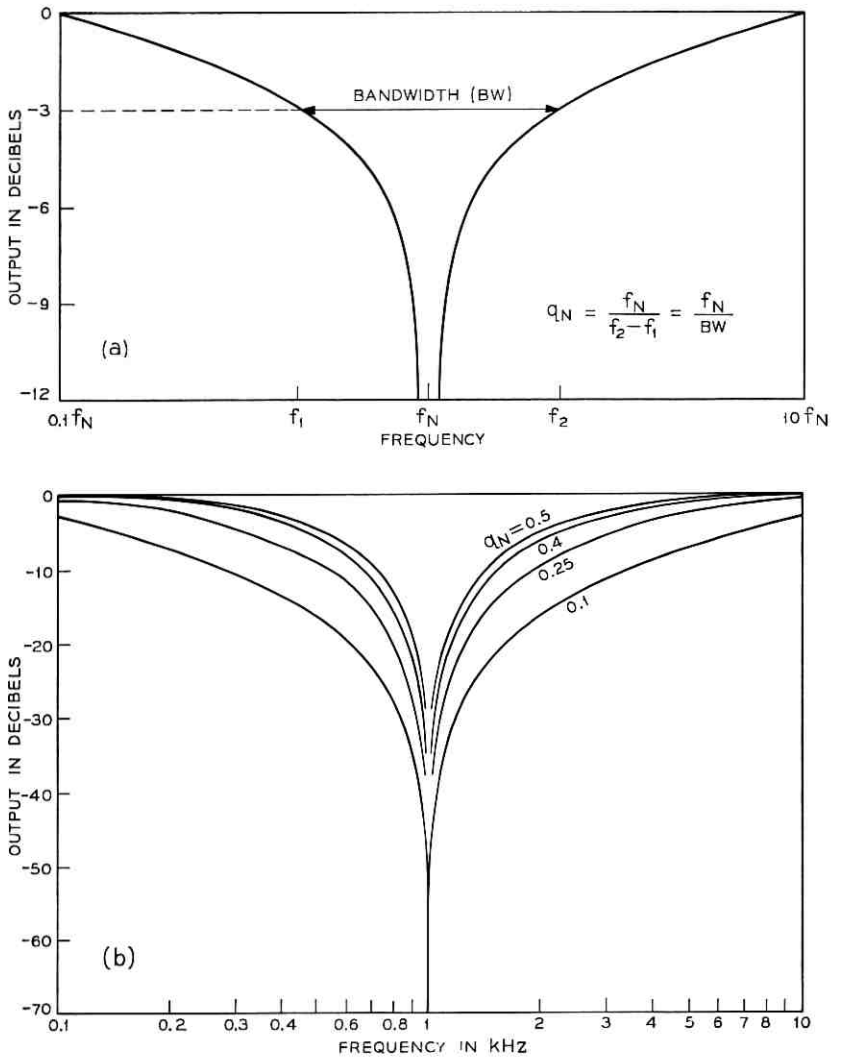


Fig. 6—Twin-T selectivity. (a) Definition of q_N; (b) Frequency response for various q_N-values.

response curves have been plotted as a function of the parameter q_N.

Expressing q_N by the same parameters as are used for the zero sensitivity functions in Table I we have

$$q_N = \frac{\alpha(1-\nu)(1-\lambda)}{\alpha^2(1-\lambda)+(1-\nu)} = \frac{(1-\nu)[\gamma\lambda(1-\lambda)]^{\frac{1}{2}}}{\gamma+\lambda(1-\nu)}. \tag{22}$$

With equation (22) we can easily observe the effects on selectivity that any sensitivity constraints on the parameters λ, ν, α or γ may have.

After these preliminary remarks, let us now consider in what ways and by which criteria it would be useful in practice to control the zero sensitivity functions given in Table I. Due to the RC self-duality of the twin-T,[34] we need only consider either the resistor or the capacitor functions. Since both discrete and (thin film) integrated RC networks are generally tuned by variable or anodizable *resistors*, the zero sensitivity functions with respect to these will be examined.

4.1 *Frequency Tuning by One Component*

By making the real part of any one of the three sensitivity functions go to zero, it is possible to shift the null frequency accurately over a limited frequency range by varying only the one corresponding resistor rather than two as would be necessary in general.

4.1.1 *Frequency Tuning with R_1*

Here we require that

$$\operatorname{Re} \mathcal{S}_{R_1}^{j\omega_N} \to 0. \tag{23}$$

By inspection of Table I this condition is fulfilled if $\lambda \to 1$, or $R_1 \gg R_2$. Then

$$\mathcal{S}_{R_1}^{j\omega_N} \approx -j\frac{\omega_N}{2}, \tag{24a}$$

$$\mathcal{S}_{R_2}^{j\omega_N} \approx \frac{\alpha\omega_N}{2(1+\alpha^2)}\left(1-\frac{j}{\alpha}\right), \tag{24b}$$

$$\mathcal{S}_{R_3}^{j\omega_N} = \frac{\alpha\omega_N}{2(1+\alpha^2)}(1+j\alpha). \tag{24c}$$

Equation (24c) remains the same as for the general case, which is independent of λ. However, from equation (22) we find that

$$q_N|_{\lambda\to 1} \to 0. \tag{25}$$

Therefore, condition (23) can only be realized at the expense of selectivity. Incidentally, the potentially symmetrical twin-T for which $\rho \ll 1$ (see Table II, Part 2b) is a special case of the one treated here, namely, that for which $\alpha = 1$.

4.1.2 *Frequency Tuning with R_2*

We require that

$$\operatorname{Re} \mathcal{S}_{R_2}^{j\omega_N} \to 0. \tag{26}$$

This condition can be approached if $\lambda \to 0$, or $R_2 \gg R_1$. We then obtain:

$$S_{R_1}^{j\omega_N} \approx \frac{\alpha\omega_N}{2(1+\alpha^2)}\left(1 - \frac{j}{\alpha}\right), \tag{27a}$$

$$S_{R_2}^{j\omega_N} \approx -\frac{j\omega_N}{2}, \tag{27b}$$

$$S_{R_3}^{j\omega_N} = -\frac{\alpha\omega_N}{2(1+\alpha^2)}(1 + j\alpha). \tag{27c}$$

Aside from interchanging the sensitivities with respect to R_1 and R_2, these expressions are the same as the preceding ones (equations 24a, b, and c). However, there is one important difference, namely in the selectivity which may now actually be larger than the "symmetrical" value of 0.25. From equation (22) we have

$$q_N\,|_{\lambda\to 0} = \frac{\alpha(1-\nu)}{\alpha^2 + 1 - \nu}. \tag{28}$$

Depending on the choice of α and ν, q_N can be made to approach its maximum value of 0.5. Here again the potentially symmetrical configuration for which $\rho \gg 1$ is a special case, namely, that for which $\alpha = 1$ and ν approaches zero in the same manner as λ does. This is one of a variety of possible cases for which equation (28) approaches 0.5. Other combinations of α and ν are best obtained by plotting equation (28) on semilog paper as shown in Fig. 7. By setting the derivative of equation (28) with respect to α equal to zero one obtains

$$\alpha_{\max} = (1-\nu)^{\frac{1}{2}} \tag{29}$$

and

$$q_{\max} = \frac{(1-\nu)^{\frac{1}{2}}}{2}. \tag{30}$$

Expression (30) is also shown in Fig. 7 by the dashed curve. Clearly there is a wide practical range of twin-T networks, with good-to-excellent selectivity, that will satisfy condition (26) and thus provide simple frequency tuning over a restricted frequency range.

One of the disadvantages of the twin-T configurations described here is that R_2, the frequency tuning resistor, is "floating," that is, it does not have one of its terminals connected to ground. Thus if it should be desired to switch various values of R_2 in order to filter or to generate different discrete frequencies, hard-contact switches would generally be

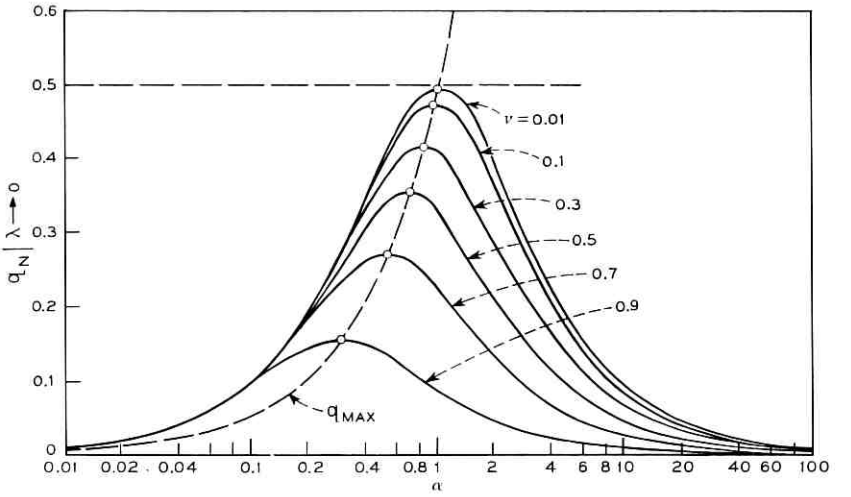


Fig. 7—Selectivity factor q_N as function of α and parameter γ, for special case that $\lambda \to 0$.

necessary since rapid semiconductor switching circuits are difficult to design in the floating mode, especially if transformers are to be avoided. For this reason the next case is of particular interest.

4.1.3 *Frequency Tuning with R_3*

The required condition here is that

$$\text{Re } s_{R_3}^{j\omega_N} \to 0. \tag{31}$$

Inspection of Table I shows that this condition cannot be realized under any circumstances. However, the following equivalent condition can be realized

$$\text{Re } s_{R_3}^{j\omega_N} \ll I_m s_{R_3}^{j\omega_N} \tag{32}$$

if $\alpha \gg 1$. Referring to Fig. 5, it is clear that this condition can be obtained in two ways, namely either by letting λ approach zero or unity with a medium value for γ or simply by letting γ become very large. However, by inspection of equation (22), both methods result in low q_N values. Thus, although the tuning resistor has one terminal grounded which does have certain advantages in terms of circuit implementation, these may be offset by the low selectivity obtainable.

4.2 *Null-Depth Tuning by One Component*

In some applications it may be desirable to make adjustments in the null depth of a twin-T after it has been initially tuned for a perfect null. This can be achieved with a single component (for example, resistor) if the imaginary part of the sensitivity function with respect to that component can be made equal to zero. By considering the general sensitivity functions given in Table I and again restricting ourselves to variable resistors for practical reasons (and because of RC-duality of the twin-T), we obtain the following three cases:

4.2.1 *Null-Depth Tuning with R_1*

Here we require that

$$I_m S_{R_1}^{j\omega N} \to 0. \tag{33}$$

The minimum of the imaginary part of $S_{R_1}^{j\omega N}$ occurs when $\lambda = 1/\alpha^2$ in which case the sensitivity functions become

$$S_{R_1}^{j\omega N} = \frac{\alpha\omega_N}{2(1+\alpha^2)}\left(1 - \frac{1}{\alpha^2} - j\frac{2}{\alpha}\right)\Bigg|_{\alpha\gg 1} \approx \frac{\omega_N}{2\alpha}, \tag{34a}$$

$$S_{R_2}^{j\omega N} = \frac{\alpha\omega_N}{2(1+\alpha^2)}\left(\frac{1}{\alpha^2} - j\alpha\right)\Bigg|_{\alpha\gg 1} \approx -j\frac{\omega_N}{2}, \tag{34b}$$

$$S_{R_3}^{j\omega N} = -\frac{\alpha\omega_N}{2(1+\alpha^2)}(1 + j\alpha)\Bigg|_{\alpha\gg 1} \approx -j\frac{\omega_N}{2}. \tag{34c}$$

As shown in the above expressions, condition (33) is realized by (34a) if $\alpha \gg 1$. Furthermore, the other two sensitivity functions turn out to be orthogonal to equation (34a) enabling independent null-frequency and null-depth control by two individual resistors (for example, R_1 and R_2 or R_1 and R_3).

It will be remembered that any sensitivity functions requiring large values of α were dismissed as impractical in the cases presented in Sections 4.1.1 to 4.1.3 due to the resulting decrease in selectivity. This was quite realistic since at least the case in Section 4.1.2 could be realized accurately while maintaining a free choice in the selectivity constant q_N. We will see in this section that no such freedom exists in any of the cases discussed and that any configuration allowing null-depth tuning by one component invariably results in selectivity deterioration. Practical implementation will therefore require a compromise between the realization of any one of the sensitivity functions and selectivity. However, as will be seen, not all the cases discussed here are equally disadvantageous with respect to this compromise.

Substituting $\lambda = 1/\alpha^2$ into equation (22), we obtain

$$q_N = \frac{(\alpha^2 - 1)(1 - \nu)}{\alpha(\alpha^2 - \nu)}. \tag{35}$$

To obtain as large a value as possible for q_N, we let $\nu \to 0$ in which case $q_N \approx 1/\alpha$. Thus the more accurately we wish to realize expression (34a), the smaller the selectivity will be.

4.2.2 *Null-Depth Tuning with R_2*

The requirement here means that

$$I_m \mathcal{S}_{R_2}^{j\omega_N} \to 0. \tag{36}$$

This would be accurately realizable if we could let

$$\lambda = 1 + \frac{1}{\alpha^2}. \tag{37}$$

Due to the restriction given by inequality (20c), this is not possible. Instead, equation (37) can be approached approximately by letting:

$$\lambda \to 1 \tag{38a}$$

and

$$\alpha \gg 1. \tag{38b}$$

By inspection of equation (19) and Fig. 5, inequality (38b) follows directly from condition (38a). More specifically, the imaginary part of $\mathcal{S}_{R_2}^{j\omega_N}$ has a minimum when

$$\lambda = 1 - \frac{1}{\alpha^2} \tag{39}$$

which is of course realizable. With equation (39), the sensitivity functions are then

$$\mathcal{S}_{R_1}^{j\omega_N} = \frac{\alpha\omega_N}{2(1 + \alpha^2)} \left(\frac{1}{\alpha^2} - j\alpha\right)\Bigg|_{\alpha \gg 1} \approx -j\frac{\omega_N}{2}, \tag{40a}$$

$$\mathcal{S}_{R_2}^{j\omega_N} = \frac{\alpha\omega_N}{2(1 + \alpha^2)} \left(1 - \frac{1}{\alpha^2} + j\frac{2}{\alpha}\right)\Bigg|_{\alpha \gg 1} \approx \frac{\omega_N}{2\alpha}, \tag{40b}$$

$$\mathcal{S}_{R_3}^{j\omega_N} = -\frac{\alpha\omega_N}{2(1 + \alpha^2)} (1 + j\alpha)\Bigg|_{\alpha \gg 1} \approx -j\frac{\omega_N}{2}. \tag{40c}$$

Thus equation (40b) provides the desired sensitivity function and also, as in the preceding case, the other two functions are orthogonal to it

allowing for independent frequency and null depth tuning with two individual resistors.

Substituting equation (39) into equation (22), the selectivity function becomes

$$q_N = \frac{1 - \nu}{\alpha(2 - \nu)} \tag{41}$$

which is largest with respect to ν when $\nu \to 0$. Then $q_N \approx 1/2\alpha$ which, as in the previous case, is all the smaller the more accurately the desired sensitivity function (40b) is to be realized. However, the preceding case (namely, that in Section 4.2.1) is preferable if it is simply a question of finding a component with which to adjust null depth, since for a given value of α the selectivity coefficient is twice as large as here.

4.2.3 *Null-Depth Tuning with R_3*

Here we require that

$$I_m S_{R_3}^{i_\omega N} \to 0. \tag{42}$$

From Table I it is evident that to satisfy this condition $\alpha \to 0$. From Fig. 5 we see that α has a minimum when $\lambda = 0.5$. Furthermore, since α is proportional to γ, we can select $\gamma \ll 1$ in order for α to approach zero. From equation (22) we obtain the selectivity coefficient

$$q_N = \frac{\alpha(1 - \nu)}{\alpha^2 + 2(1 - \nu)} \tag{43}$$

which is maximum with respect to ν when $\nu \to 0$. Therefore from equation (43) we obtain $q_N \approx \alpha/2$. The condition for α, in this case, is the inverse of that for the two preceding cases. Taking this into consideration while comparing the corresponding selectivity coefficients shows the case in Section 4.2.1 to have the highest selectivity. It, like that in 4.2.2, has the added advantage of orthogonal sensitivity functions providing for both the desired null-depth tuning as well as frequency tuning by single components in the vicinity of the null frequency. On the other hand, if a variable component with one terminal grounded is preferred for the reasons given in Section 4.1.2, then the case in Section 4.2.3 may be used, provided the selectivity, which is smaller by a factor of two, is still acceptable.

4.3 *Orthogonal Tuning With Two Components*

Orthogonality between two zero sensitivity functions simplifies null adjustments in the vicinity of a perfect null, particularly if the two

functions are parallel with the real and imaginary axes. Some of the configurations described in the previous sections provided the latter type of orthogonality, but only at the cost of selectivity. General orthogonality, which is discussed here, may be of interest for a variety of reasons, for example, the 90 degree phase reference required for tuning purposes may be easier to generate than any other arbitrary phase reference.

Two vectors $\mathbf{q} = u + jv$ and $\mathbf{p} = w + jz$ are orthogonal if

$$uw + vz = 0. \tag{44}$$

Thus, to obtain orthogonality between pairs of the functions listed in Table I, we must investigate if they can be made to satisfy this condition.

4.3.1 *Orthogonal Tuning Between R_1 and R_2*

This requires that

$$\lambda(1 - \lambda) + \left(\frac{1}{\alpha} + \alpha\lambda\right)\left[\frac{1}{\alpha} + \alpha(1 - \lambda)\right] = 0. \tag{45}$$

Solving for the roots of this equation one obtains $\alpha_{1,2} = \pm j$ which is not physically realizable.

4.3.2 *Orthogonal Tuning Between R_1 and R_3*

To satisfy the condition for orthogonality here, we require that

$$-(1 - \lambda) + \alpha\left(\frac{1}{\alpha} + \alpha\lambda\right) = 0. \tag{46}$$

Solving equation (46) for α results in the same nonrealizable roots as were obtained in Section 4.3.1. However, one additional solution exists here, namely $\lambda = 0$. This condition can only be approximated [see inequality (20c)] and has been dealt with in Sections 4.1.2 and 4.2.1 where, as expected, $S_{R_1}^{j\omega N}$ is orthogonal to $S_{R_3}^{j\omega N}$.

4.3.3 *Orthogonal Tuning Between R_2 and R_3*

It is required that

$$-\alpha + \alpha\left[\frac{1}{\alpha} + \alpha(1 - \lambda)\right] = 0 \tag{47}$$

which is satisfied when $\lambda = 1$. As in the preceding case, this condition can only be approximated; it has been dealt with in Sections 4.1.1 and 4.2.2.

It is evident from the above that, apart from the cases of orthogonality already discussed in earlier sections, the condition for general orthog-

onality given by equation (44) does not produce any new realizable twin-T configurations.

4.4 *Design Examples of Twin-T Networks With Controlled Tuning Characteristics*

The design equations for twin-T networks with the tuning characteristics described in Sections 4.1 and 4.2 have been compiled in Table III. Results from Section 4.3 have not been included since they did not

TABLE III—TWIN-T DESIGN EQUATIONS FOR CONTROLLED ZERO SENSITIVITY

Design Equations for Controlled Sensitivity and Sensitivity Functions	Design Equations for Maximum Selectivity	Remarks
1A)		
Re $S_{R_1}^{j\omega_N} \to 0$: $\lambda \to 1$	$q_N \mid_{\nu\to 0} \approx \frac{\alpha(1-\lambda)}{\alpha^2(1-\lambda)+1}$	$q_N \ll 0.5$
$S_{R_1}^{j\omega_N} \approx -j\frac{\omega_N}{2}$		Orthogonality between
$S_{R_2}^{j\omega_N} \approx \frac{\alpha\omega_N}{2(1+\alpha^2)}\left(1-\frac{j}{\alpha}\right)$	$q_{N\max} \mid_{\nu\to 0} \approx \frac{\sqrt{1-\lambda}}{2}$	$S_{R_2}^{j\omega_N}$ and $S_{R_3}^{j\omega_N}$
	where:	
$S_{R_3}^{j\omega_N} = -\frac{\alpha\omega_N}{2(1+\alpha^2)}\left(1-\frac{j}{\alpha}\right)$	$\alpha = \frac{1}{\sqrt{1-\lambda}}$	
1B)		
Re $S_{R_2}^{j\omega_N} \to 0$: $\lambda \to 0$	$q_N \approx \frac{\alpha(1-\nu)}{\alpha^2+1-\nu}$	$0 < q_N < 0.5$
$S_{R_1}^{j\omega_N} \approx \frac{\alpha\omega_N}{2(1+\alpha^2)}\left(1-\frac{j}{\alpha}\right)$	$q_{N\max} \approx \frac{(1-\nu)^{\frac{1}{2}}}{2}$	Orthogonality between
$S_{R_2}^{j\omega_N} \approx -\frac{j\omega_N}{2}$	where:	$S_{R_1}^{j\omega_N}$ and $S_{R_3}^{j\omega_N}$
$S_{R_3}^{j\omega_N} = -\frac{\alpha\omega_N}{2(1+\alpha^2)}(1+j\alpha)$	$\alpha_{\max} = (1-\nu)^{\frac{1}{2}}$	
1C)		
Re $S_{R_3}^{j\omega_N} \to 0$: $\alpha \gg 1$	$q_N \approx \frac{1-\nu}{\alpha}$	$q_N \ll 0.5$
$S_{R_1}^{j\omega_N} = \frac{\alpha\omega_N}{2(1+\alpha^2)}\left[1-\lambda-j\left(\frac{1}{\alpha}+\alpha\lambda\right)\right]$	$q_{N\max} \approx \frac{1}{\alpha}$	Variable resistor (R_3) has one common (that is, grounded) terminal
$S_{R_2}^{j\omega_N} = \frac{\alpha\omega_N}{2(1+\alpha^2)}\left[\lambda-j\left\{\frac{1}{\alpha}+\alpha(1-\lambda)\right\}\right]$	for $\nu \to 0$	
$S_{R_3}^{j\omega_N} \approx -j\frac{\omega_N}{2}$		

TABLE III—Cont'd

Design Equations for Controlled Sensitivity and Sensitivity Functions	Design Equations for Maximum Selectivity	Remarks
2A)		
Im $S_{R_1}^{j\omega N} \to 0$: $\alpha \gg 1$, $\lambda = \frac{1}{\alpha^2} \ll 1$	$q_N \approx \frac{(\alpha^2 - 1)(1 - \nu)}{\alpha(\alpha^2 - \nu)}$	$q_N \ll 0.5$
$S_{R_1}^{j\omega N} \approx \frac{\omega_N}{2\alpha}$	$q_{N\max} \approx \frac{1}{\alpha}$	Orthogonality between
$S_{R_2}^{j\omega N} \approx -j\frac{\omega_N}{2}$	for $\nu \to 0$	$S_{R_1}^{j\omega N}$ and $S_{R_2}^{j\omega N} = S_{R_3}^{j\omega N}$
$S_{R_3}^{j\omega N} \approx -j\frac{\omega_N}{2}$		
2B)		
Im $S_{R_3}^{j\omega N} \to 0$: $\alpha \gg 1$, $\lambda = \left(1 - \frac{1}{\alpha^2}\right) \to 1$	$q_N \approx \frac{1 - \nu}{\alpha(2 - \nu)}$	$q_N \ll 0.5$
$S_{R_1}^{j\omega N} \approx -j\frac{\omega_N}{2}$	$q_{N\max} \approx \frac{1}{2\alpha}$	Orthogonality between
$S_{R_2}^{j\omega N} \approx \frac{\omega_N}{2\alpha}$	for $\nu \to 0$	$S_{R_3}^{j\omega N}$ and $S_{R_1}^{j\omega N} = S_{R_2}^{j\omega N}$
$S_{R_3}^{j\omega N} \approx -j\frac{\omega_N}{2}$		
2C)		
Im $S_{R_3}^{j\omega N} \to 0$: $\alpha \to 0$, $\lambda = 0.5$	$q_N \approx \frac{\alpha(1 - \nu)}{\alpha^2 + 2(1 - \nu)}$	$q_N \ll 0.5$
$S_{R_1}^{j\omega N} \approx \frac{\omega_N}{2}\left(\frac{\alpha}{2} - j\right)$	$q_{N\max} \approx \frac{\alpha}{2}$	Variable resistor (R_3) has one common (that is, grounded) terminal
$S_{R_2}^{j\omega N} \approx \frac{\omega_N}{2}\left(\frac{\alpha}{2} - j\right)$	for $\nu \to 0$	
$S_{R_3}^{j\omega N} \approx -\frac{\alpha\omega_N}{2}$		

produce anything not already obtained in the two previous sections.

Using the design equations listed in Table III, the detailed procedure for the design of two twin-T networks with prescribed tuning characteristics follows.

4.4.1 *Twin-T With Null Frequency Tunable by R_2*

To satisfy condition (26), we find from Table III, Part 1B, that $\lambda \ll 1$

and therefore select $\lambda = 0.01$. Furthermore, assuming that $q_N = 0.25$, $\omega_N = 2\pi 1$ kHz, and $R_1 = 1$ KΩ, we find $R_2 = 99$ KΩ. From Fig. 7 we find that the q_{max}-curve passes through the value 0.25 when $\alpha = 0.5$. However, with $\lambda = 0.01$, γ takes on a simpler value for $\alpha = 0.55$ (see Fig. 5), namely

$$\gamma = \alpha^2\lambda(1 - \lambda) = 0.003.$$

Solving equation (28) for ν we obtain

$$\nu = 1 - \frac{\alpha^2}{4\alpha - 1} = 0.75$$

and from equation (18)

$$C_1 = \frac{\gamma}{\nu} \cdot C_3 = 0.004C_3 .$$

From equations (5) and (18)

$$C_3 = \frac{1}{\omega_N(\gamma R_1 R_2)^{\frac{1}{2}}} = 0.292\ \mu\text{F}$$

and

$$C_1 = \frac{\gamma}{\nu} C_3 = 1.168\ \text{nF}.$$

Finally, with equation (15)

$$C_2 = \frac{\nu}{1 - \nu} C_1 = 3.504\text{nF}$$

and, from equation (6)

$$R_3 = \frac{C_3}{C_p} \cdot R_p = 62\text{K}.$$

The corresponding sensitivity functions can be calculated directly by substituting the values obtained above into the expressions listed in Table I. Considering only the relative values of the resistor sensitivity functions, one obtains

$$S_{R_1}^{j\omega_N} = \frac{\alpha\omega_N}{2(1 + \alpha^2)} (0.99 - 1.83j),$$

$$S_{R_2}^{j\omega_N} = \frac{\alpha\omega_N}{2(1 + \alpha^2)} (0.01 - 2.365j),$$

$$S_{R_3}^{j\omega_N} = \frac{\alpha\omega_N}{2(1 + \alpha^2)} (1 + 0.55j).$$

The twin-T network resulting from the above calculations, as well as the zero displacement vectors given above, are shown in Fig. 8. In Fig. 9a, measurements of the frequency response of this twin-T are compared with those of a symmetrical twin-T nulled at the same frequency. The initial frequency response of the two ideally nulled networks are identical since both have a selectivity constant q_N equal to 0.25. On varying R_2, however, the null depth of the symmetrical network decreases considerably with varying null frequency compared to that of the non-symmetrical configuration. This is also apparent from Fig. 9b, where the percentage frequency shift of the null is shown as a function of the relative resistor change R_2. Whereas this curve is not appreciably different for the two configurations, the simultaneous variation in null depth is.

4.4.2 *Twin-T with Null Depth Tunable by R_3*

To satisfy condition (42), we find from Table III, Part 2C, that $\alpha \ll 1$ and select $\alpha = 0.1$. Furthermore, with $\lambda = 0.5$ and letting $\nu = 0.01$, for maximum selectivity, we find from equation (22) that $q_N = 0.048$.

As in the previous example, we assume that $\omega_N = 2\pi 1$ kHz and, because $\lambda = 0.5$, we select $R_1 = R_2 = 10$ KΩ. From equation (19) we find $\gamma = 0.0025$ and, in precisely the same way as in the preceding example, $C_3 = 0.318$ μF, $C_1 = 79.5$ nF, $C_2 = 0.804$ nF, and $R_3 = 19.8$ K. The corresponding zero sensitivity functions follow directly from Table I. The relative values of the resistor sensitivity functions are

$$S_{R_1}^{j\omega_N} = \frac{\alpha\omega_N}{2(1+\alpha^2)}(0.5 - 10.05j),$$

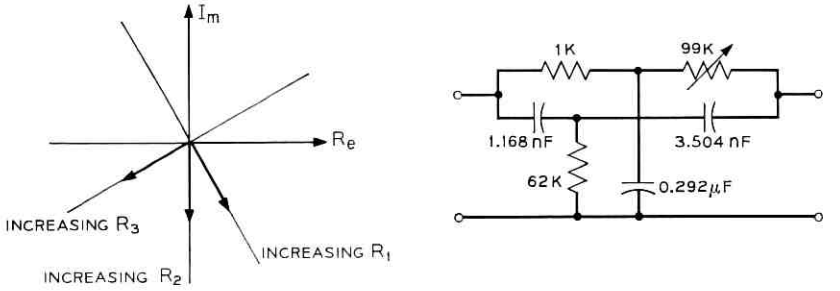


Fig. 8—Zero displacement and twin-T configuration for $\omega_N = 2\pi \cdot 1$ kHz, $q_N = 0.25$, $\lambda = 0.01$, $\nu = 0.75$, $\alpha = 0.55$.

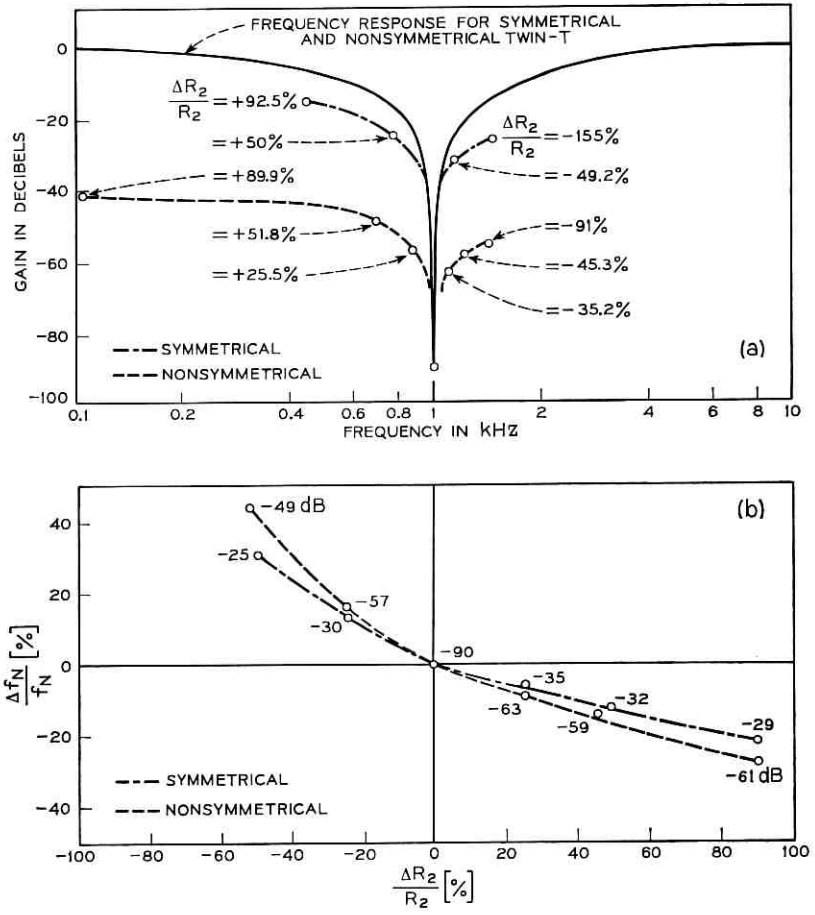


Fig. 9—Comparison of measurements conducted on twin-T shown in Fig. 8 and symmetrical twin-T. (a) Null-frequency shift and null-depth variation with variation of R_2; (b) Twin-T null frequency as function of percentage change in R_2.

$$S_{R_2}^{j\omega_N} = \frac{\alpha\omega_N}{2(1+\alpha^2)}(0.5 - 10.05j),$$

$$S_{R_3}^{j\omega_N} = -\frac{\alpha\omega_N}{2(1+\alpha^2)}(1 + 0.1j),$$

The resulting twin-T network and the zero displacement vectors are shown in Fig. 10. Measurements made on the twin-T are shown in Fig. 11a where they are compared with those of a symmetrical network

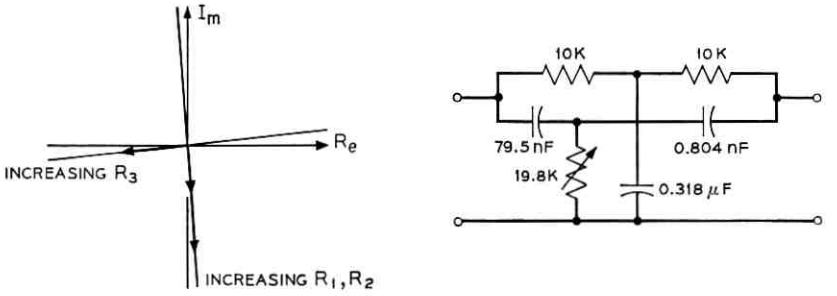


Fig. 10—Zero displacement and twin-T configuration for $\omega_N = 2\pi \cdot 1$ kHz, $q_N = 0.048$, $\lambda = 0.5$, $\nu = 0.01$, $\alpha = 0.1$.

with the same null frequency. The initial frequency response of the two configurations differs here, since the selectivity coefficient of the symmetrical twin-T is 0.25, and that of the other is 0.048. The null depth of the symmetrical twin-T can be decreased by more than 50 dB from an initial 90-dB null with no measurable change in the null frequency. This compares with over 1 percent variation of null frequency for the symmetrical configuration. This is shown again in Fig. 11b where the null depth variation is plotted versus relative change in the resistor R_3.

V. TWIN-T NULL STABILITY USING THIN FILM COMPONENTS

The twin-T is frequently used to provide stable zeros in the design of hybrid integrated linear active networks. If a high degree of stability is required, thin film components must be used for the twin-T network. Just what degree of null stability can be expected with thin film components whose temperature coefficients and aging characteristics are known, follows directly from the sensitivity functions discussed in the previous sections. This is shown in the following.

A displacement dz in the transmission zero z of a twin-T network can be expressed in terms of the zero sensitivity defined by equation (13) as follows

$$dz = \sum_{i=1}^{3} S_{R_i}^{z} \frac{dR_i}{R_i} + \sum_{i=1}^{3} S_{C_i}^{z} \frac{dC_i}{C_i}. \tag{48}$$

As shown in Fig. 12, if the twin-T transmission zero is close to the $j\omega$-axis it can be considered purely imaginary for purposes of computing sensitivities, thus

$$S_x^z \approx S_x^{j\omega_N}. \tag{49}$$

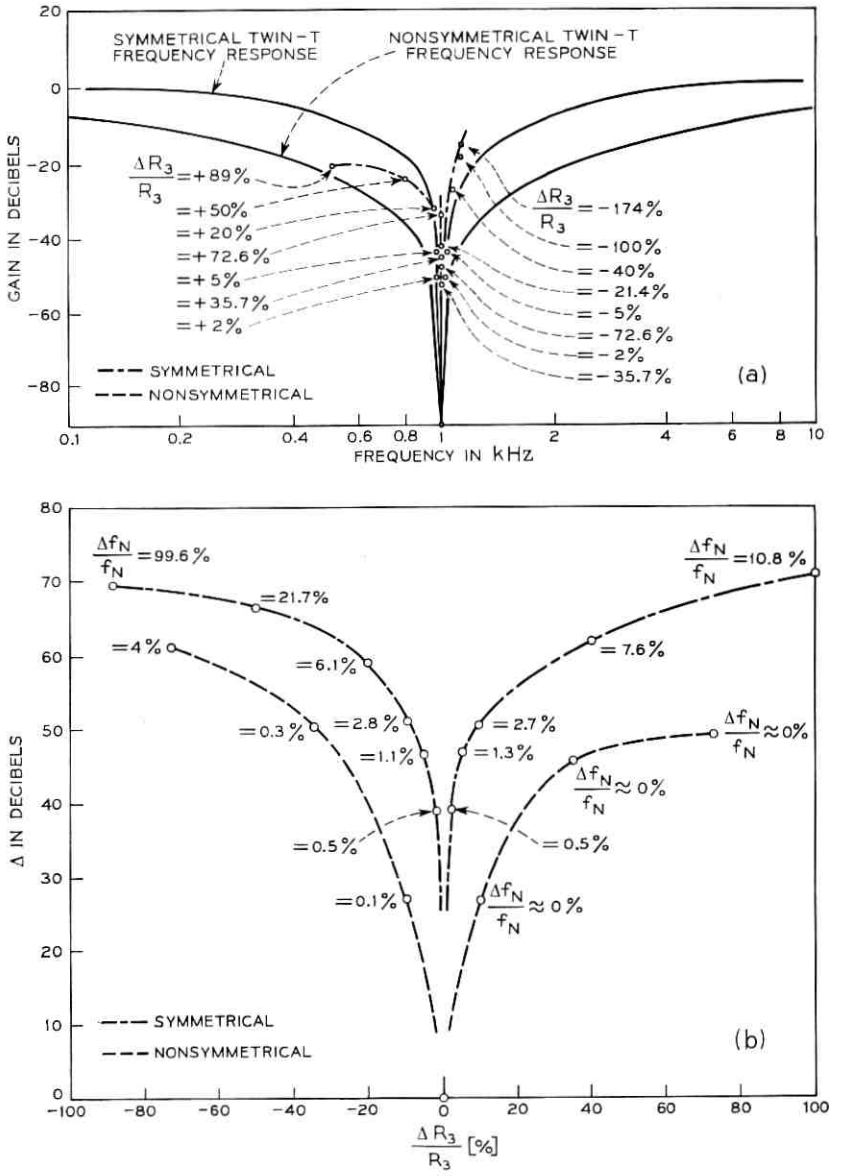


Fig. 11—Comparison of measurements conducted on twin-T shown in Fig. 10 and symmetrical twin-T. (a) Null-frequency shift and null-depth variation with variation of R_3; (b) Twin-T null depth as function of percentage change in R_3.

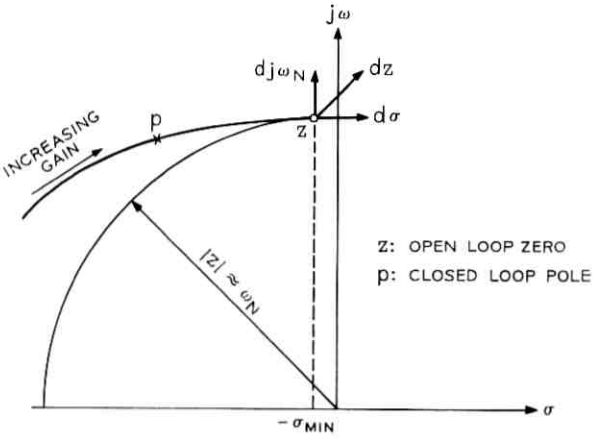


Fig. 12—Typical root locus of feedback network using a twin-T to provide the open loop zeros.

Furthermore, referring again to Fig. 12, the zero displacement is given as

$$dz = d\sigma + dj\omega_N . \tag{50}$$

Equating (48) and (50) and substituting the expressions of Table I in equation (48), it is possible to solve for $d\sigma$ and $dj\omega_N$.

First, however, some characteristics peculiar to thin film integrated circuitry must be considered. Due to the batch processing techniques used, component uniformity can be guaranteed much more accurately than with discrete components. Above all, component variations tend to track very closely on a given glass or ceramic substrate and these variations can be precisely predicted and controlled. These features permit a considerable simplification in the following calculations without any loss in accuracy. Thus, we can write

$$\frac{\Delta R_i}{R_i} = [\delta_r \pm \epsilon_r]\,\Delta T + \kappa_r = \frac{\Delta R}{R} \tag{51}$$

and

$$\frac{\Delta C_i}{C_i} = [\delta_c \pm \epsilon_c]\,\Delta T + \kappa_c = \frac{\Delta C}{C}. \tag{52}$$

The temperature coefficients of the resistors and capacitors are δ_r and δ_c, respectively; ϵ_r and ϵ_c are the tracking ratios between the three resistors and the three capacitors, respectively; ΔT is the temperature

range under consideration; and κ_r and κ_c are the percentage resistor and capacitor aging, respectively. Substituting equations (51) and (52) as well as the expressions in Table I into equation (48), we obtain for the real part of the finite zero displacement $\Delta z = \Delta\sigma + j\Delta\,\omega_N$

$$\frac{\Delta\sigma}{\omega_N} = \frac{\alpha}{2(1+\alpha^2)}\left[\epsilon_c(1-\nu) - \epsilon_r(1-\lambda)\right]\Delta T. \tag{53}$$

Notice that the zero displacement parallel to the real axis depends only on the *amount of mistracking between components* and not on the absolute drift of the individual components. In other words, if all components of a kind drift by the same percentage, the null depth of a tuned twin-T will not change. This is to be expected since *equal component drift corresponds to a frequency scaling process.*

The imaginary part of the finite zero displacement Δz is obtained in the same way as the real displacement above. Thus

$$-\frac{\Delta\omega_N}{\omega_N} = (\delta_r + \delta_c)\,\Delta T + \kappa_r + \kappa_c + \frac{\epsilon_r\,\Delta T}{2}\left[\frac{1+\alpha^2(2-\lambda)}{1+\alpha^2}\right] + \frac{\epsilon_c\,\Delta T}{2}\left[\frac{\alpha^2+2-\nu}{1+\alpha^2}\right]. \tag{54}$$

Thus, the zero displacement along the $j\omega$-axis depends on the actual drift of the individual components. Clearly, if the drift coefficients of the resistors can be made equal but opposite to those of the capacitors, drift along the $j\omega$-axis can be practically eliminated.

In various active filter schemes the network poles are tied closely to the transmission zeros generated by a twin-T. Thus, in high Q networks, uncontrolled drift of the twin-T zero into the right-half s-plane could pull the poles over with it, causing oscillation. Similarly, drift of the twin-T zero along the $j\omega$-axis would cause frequency drift in the active filter.

To prevent oscillation due to drift into the right half plane, the transmission zero of the twin-T must be located left of the $j\omega$-axis by some distance $\sigma_{\min}$ such that, under worst case component drift, it will not travel across the $j\omega$-axis. Referring to Fig. 12, this implies that

$$\sigma_{\min} \geqq \mathrm{Re}\,(\Delta z_{\max}) = \Delta\sigma_{\max}\,. \tag{55}$$

This condition, in turn, implies that the twin-T null depth may not exceed a certain maximum attenuation $T_{N\,\max}$ which can now be calculated directly.

It follows from Appendix B that the transfer function of a twin-T

with a nonperfect null can be approximated as follows

$$T_N(s) \approx \frac{s^2 + 2\sigma s + \omega_N^2}{s^2 + \frac{\omega_N}{q_N} s + \omega_N^2}. \tag{56}$$

With equation (55), the maximum null attenuation for left half plane transmission zeros is then

$$T_{N\max}\,|_{s=j\omega_N} = 2\frac{q_N}{\omega_N}\sigma_{\min} = 2\frac{q_N}{\omega_N}\Delta\sigma_{\max}\,. \tag{57}$$

With equations (22) and (53) this becomes

$$T_{N\max} = \frac{\alpha^2}{1+\alpha^2}\left(\frac{(1-\nu)(1-\lambda)}{\alpha^2(1-\lambda)+(1-\nu)}\right)[\epsilon_r(1-\lambda)+\epsilon_c(1-\nu)]\,\Delta T. \tag{58}$$

In active filter applications where the twin-T transmission zero z represents the open loop zero of the root locus of a pole p with respect to gain (see Fig. 12), the highest attainable Q of the network is all the more limited the larger $\sigma_{\min}$ has to be chosen for stability. In the limit, as the loop gain approaches infinity, the closed loop pole p coincides with z. The upper limit on Q is therefore given by

$$Q_{\max} < \frac{\omega_N}{2\sigma_{\min}} \tag{59}$$

or, with equation (53)

$$Q_{\max} < \frac{(1+\alpha^2)}{\alpha\Delta T}\left(\frac{1}{\epsilon_r(1-\lambda)+\epsilon_c(1-\nu)}\right). \tag{60}$$

Thus, with the type of active network design represented by the root locus in Fig. 12, both network stability and maximum Q ultimately depend on the stability of the twin-T network.

As an example of the above, we shall consider the stability of a symmetrical twin-T network fabricated with tantalum thin film resistors and capacitors. The required ambient temperature range is assumed to be from 0°C to 60°C. From equation (58) we obtain

$$T_{N\max}\Big|_{\substack{\lambda=\nu=0.5\\ \alpha=1}} = \tfrac{1}{16}[\epsilon_r + \epsilon_c]\,\Delta T. \tag{61}$$

Typically, for tantalum thin film resistors and capacitors $\epsilon_r = \pm 5$ ppm/°C and $\epsilon_c = \pm 15$ ppm/°C. Therefore $T_{N\max} = 7.5 10^{-5} = -83$ dB.

The frequency drift for a symmetrical twin-T results from equation (54) as

$$-\frac{\Delta\omega_N}{\omega_N} = [(\delta_r + \delta_c) + \tfrac{5}{8}(\epsilon_r + \epsilon_c)]\,\Delta T + \kappa_r + \kappa_c\,. \tag{62}$$

Typically, for tantalum thin film components $\kappa_r = \kappa_c = 0.1\%$ and for tantalum thin film capacitors $\delta_c = 200$ ppm/°C. The TC of tantalum thin film resistors can be controlled by oxygen doping during the sputtering process.[35] It may therefore be of interest to solve equation (62) for the required TCR, that is, δ_r, when a maximum acceptable frequency drift is specified. Assuming that $(\Delta\omega_N/\omega_N)_{max} \leqq 0.5\%$ we obtain $\delta_r = (-215 \pm 50)$ ppm/°C.

VI. CONCLUSIONS

Design equations have been presented that provide twin-T configurations with null characteristics that depend on individual component variations in a predictable and controlled manner. This is achieved by deriving the zero sensitivity functions with respect to each component of the general nulled twin-T. The network parameters required to obtain a twin-T that can be tuned by a desired procedure and the extent to which such a procedure can at all be realized results directly from inspection of the general sensitivity functions.

Special null tuning procedures are considered that are useful in linear active networks incorporating a twin-T in the feedback path. One twin-T configuration permits the null frequency to be shifted over a limited range by the variation of one component only while the null depth remains constant. Another enables the null depth to be varied by one component with negligible variation in the null frequency. The possibility of independent null frequency and null depth tuning by two individual components is also investigated. Orthogonal sensitivity functions that are parallel to the real and imaginary axis are required to do this. It is shown that, apart from the orthogonality that is obtained as a by-product in the first two cases, any other or more general kind of orthogonality in the sensitivity functions cannot be realized. Design examples for the first two cases are given. Measurements conducted on the resulting twin-T configurations are presented and compared with similar measurements made on a conventional, that is, symmetrical, twin-T. This comparison demonstrates the effectiveness of the given design equations.

The stability of the null characteristics of a twin-T with given

component characteristics results directly from the sensitivity functions. Limits on the permissible null depth of a twin-T are derived for the case that left-half plane zeros must be guaranteed under worst case component drift conditions. Similarly, expressions are derived that permit the limits on resistor and capacitor temperature coefficients and aging characteristics to be established in order not to exceed a given maximum frequency drift of the transmission null. A numerical example is given using the characteristics of tantalum thin film resistors and capacitors.

VII. ACKNOWLEDGMENT

Thanks are due to G. Malek who carried out the twin-T measurements shown in Figs. 9 and 11.

APPENDIX A

Twin-T Impedance Matrix

In order to calculate the open circuit impedance matrix of the twin-T, it is useful to first obtain its general equivalent π-network. This can be simply obtained by converting each of the two T-networks of the twin-T into its equivalent π-network. This is shown in Figs. 13a and 13b, assuming sinusoidal input signals. The two resulting π-networks can then be connected in parallel (as shown in Fig. 13c), and the resulting impedance directly calculated. With the two conditions for a perfect null given by equations (5) and (6), we get a simple π-network as shown in Fig. 14. The corresponding impedances are given by

$$Z_a = \frac{R_1}{1 + \frac{\tau_1}{\tau}} \left(1 + \frac{1}{s\tau}\right), \tag{63}$$

$$Z_b = \frac{R_2}{1 + \frac{\tau_2}{\tau}} \left(1 + \frac{1}{s\tau}\right), \tag{64}$$

and

$$Z_c = R_s \frac{\left(1 + \frac{1}{s\tau}\right)}{s\tau_s + \frac{1}{s\tau}}, \tag{65}$$

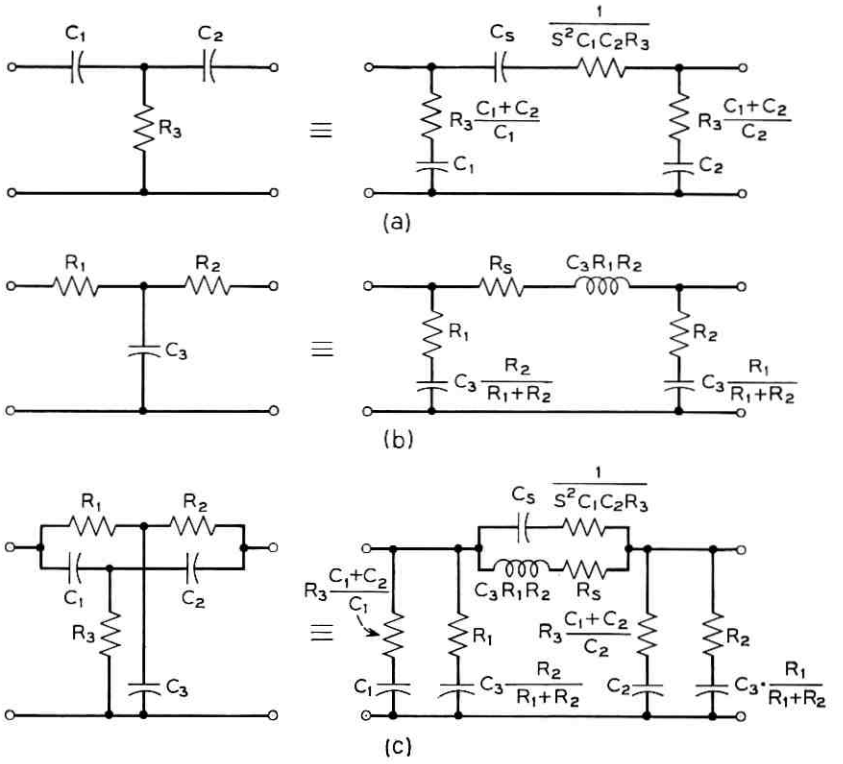


Fig. 13—(a) and (b), Conversion of the two T-networks of the twin-T into equivalent π-networks; (c), Equivalent π-network of the twin-T.

where

$$\tau_1 = R_1C_1 ,$$

$$\tau_2 = R_2C_2 ,$$

$$\tau_a = R_aC_a ,$$

$$\tau = R_pC_3 = R_3C_p .$$

In terms of these impedances, the open-circuit impedance matrix for the twin-T simply follows as

$$[z] = \frac{1}{Z_a + Z_b + Z_c}\begin{bmatrix} Z_a(Z_b + Z_c) & Z_aZ_b \\ Z_aZ_b & Z_b(Z_a + Z_c) \end{bmatrix}. \tag{66}$$

So far we have considered a general twin-T that has an infinite null

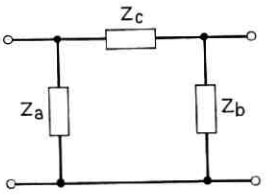


Fig. 14—General π-network.

at a specified frequency ω_N. Because the corresponding general transfer function [see equation (10)] has the form of a quadratic fraction, it has geometric symmetry around ω_N.

The most frequently used twin-T is structurally (and electrically) symmetrical. For this case.

$$\tau_1 = \tau_2 = \tau_s = \tau = RC, \tag{67}$$

$$R_3 = R/2, \tag{68}$$

and

$$C_3 = 2C. \tag{69}$$

The impedances of Fig. 14 then become

$$Z_a = Z_b = \frac{R}{2}\left(1 + \frac{1}{s\tau}\right) \tag{70}$$

and

$$Z_c = 2R\,\frac{(1 + s\tau)}{1 + s^2\tau^2}. \tag{71}$$

The open-circuit impedance matrix consequently becomes

$$(z)_s = \frac{R}{4}\begin{bmatrix} \dfrac{1 + 4s\tau + s^2\tau^2}{s\tau(1 + s\tau)} & \dfrac{1 + s^2\tau^2}{s\tau(1 + s\tau)} \\ \dfrac{1 + s^2\tau^2}{s\tau(1 + s\tau)} & \dfrac{1 + 4s\tau + s^2\tau^2}{s\tau(1 + s\tau)} \end{bmatrix}. \tag{72}$$

The voltage transfer function for the symmetrical twin-T then follows as

$$T_{N_s}(s) = \frac{z_{21_s}}{z_{11_s}} = \frac{s^2\tau^2 + 1}{s^2\tau^2 + 4\tau s + 1}. \tag{73}$$

Comparing with the transfer function given by equation (10), we have

$$\omega_{N_0} = \frac{1}{RC}, \tag{74}$$

$$2\sigma_{N_0} = \frac{4}{RC} \tag{75}$$

and the inverse damping factor

$$q_{N_0} = 0.25. \tag{76}$$

It can be shown that the selectivity, that is, the inverse damping factor, can be increased by modifying the symmetrical twin-T into a potentially symmetrical network. This is possible with any structurally symmetrical network for which Bartlett's bisection theorem holds. A symmetrical network can be converted into a potentially symmetrical network by impedance scaling one half of the network by some factor ρ. This is shown for the twin-T in Fig. 15. The corresponding z-matrix then becomes:

$$(z)_{ps} = \frac{\rho}{1+\rho}\cdot\frac{R}{2}\begin{bmatrix} \dfrac{s^2\tau^2 + 2\left(1+\dfrac{1}{\rho}\right)s\tau + 1}{s\tau(1+s\tau)} & \dfrac{1+s^2\tau^2}{s\tau(1+s\tau)} \\ \dfrac{1+s^2\tau^2}{s\tau(1+s\tau)} & \dfrac{s^2\tau^2 + 2(1+\rho)s\tau + 1}{s\tau(1+s\tau)} \end{bmatrix} \tag{77}$$

and the voltage transfer function results as

$$T_{N_{ps}} = \frac{z_{21_{ps}}}{z_{11_{ps}}} = \frac{s^2\tau^2 + 1}{s^2\tau^2 + 2\left(1+\dfrac{1}{\rho}\right)s\tau + 1}. \tag{78}$$

In terms of the transfer function (10) we find:

$$\omega_{N_{ps}} = \frac{1}{RC}, \tag{79}$$

$$2\sigma_{N_{ps}} = \frac{2}{RC}\left(\frac{\rho+1}{\rho}\right) \tag{80}$$

and

$$q_{N_{ps}} = \frac{1}{2}\frac{\rho}{1+\rho}. \tag{81}$$

ρ gives a measure of the twin-T symmetry. For the extreme asymmetrical

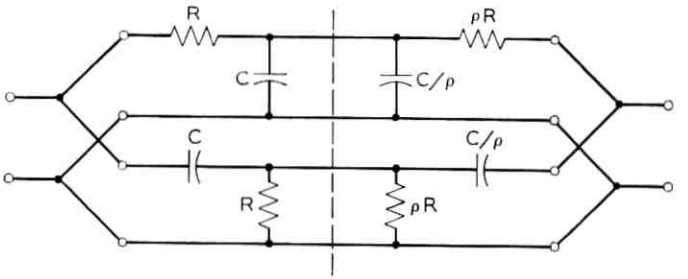


Fig. 15—Potentially symmetrical twin-T resulting from symmetrical twin-T after one-half of the twin-T has been impedance-scaled by a factor ρ.

case for which $\rho \gg 1$, q_{N_p} takes on its maximum value, namely

$$q_{N_p}\,|_{\rho\to\infty} \to \tfrac{1}{2}. \tag{82}$$

APPENDIX B

Twin-T with a Finite Null

Inspection of equations (8) and (9) shows that the transfer function of a general twin-T is simplified by one degree due to the pole-zero cancellation on the negative real axis at $\omega_1 = 1/R_3C_p$ when the conditions for a perfect null given by equations (5) and (6) are satisfied. We investigate here the conditions necessary to ensure this pole-zero cancellation when the null-conditions are only approximately satisfied, that is, when the twin-T has a finite null. To do so we derive the sensitivity of the pole and zero at ω_1 with respect to the six parameters of the twin-T and investigate under which conditions the respective pole and zero sensitivities are the same.

Writing the twin-T transfer function in the bilinear form with respect to a parameter x we obtain:

$$T_N(s) = \frac{N(s)}{D(s)} = \frac{A_x(s) + xB_x(s)}{U_x(s) + xV_x(s)}. \tag{83}$$

The sensitivity of a zero z with respect to x is then given by:

$$S_x^z = \frac{(s-z)A_x(s)}{N(s)}\bigg|_{s=z} = -x\,\frac{(s-z)B_x(s)}{N(s)}\bigg|_{s=z} \tag{84}$$

that of a pole p with respect to x by

$$S_x^p = \frac{(s-p)U_x(s)}{D(s)}\bigg|_{s=p} = -x\,\frac{(s-p)V_x(s)}{D(s)}\bigg|_{s=p}. \tag{85}$$

From equations (84) and (85) we obtain

$$S_x^z \Big|_{z=-\omega_1} = \frac{\omega_1}{1+\alpha^2} \cdot A_x(-\omega_1) \tag{86}$$

and

$$S_x^p \Big|_{p=-\omega_1} = \frac{\omega_1}{1+\alpha^2-\dfrac{\alpha}{q_N}} \cdot U_x(-\omega_1) + \frac{\omega_1}{\left(1-\dfrac{1}{1-\lambda}\right)+\alpha^2\left(1-\dfrac{1}{1-\nu}\right)} U_x(-\omega_1) \tag{87}$$

where

$$\alpha = \frac{\omega_1}{\omega_N}, \tag{88}$$

$$\lambda = \frac{R_1}{R_1+R_2}, \tag{89}$$

$$\nu = \frac{C_1}{C_1+C_2}, \tag{90}$$

and

$$q_N = \frac{\omega_N}{2\sigma_N} = \frac{\alpha(1-\nu)(1-\lambda)}{\alpha^2(1-\lambda)+(1-\nu)}. \tag{91}$$

Calculating the respective $A_x(-\omega_1)$ and $U_x(-\omega_1)$ functions we obtain the zero and pole sensitivities listed in Table IV. Comparing the functions listed in the two columns of the table it is clear from inspection that they will be equal when

$$\frac{\lambda(1-\nu)}{\nu(1-\lambda)} = 1. \tag{92}$$

With equations (88) through (90), this condition becomes

$$R_1C_1 = R_2C_2. \tag{93}$$

Thus for all twin-T configurations in which the time constants of the series elements are the same, pole-zero cancellation on the negative real axis is maintained for differentially small perturbations of any element of the twin-T. For positive element changes (that is, increasing values) the dipole frequency will decrease, that is, move in the direction of the origin of the s-plane. Twin-T networks that satisfy equation (93)

TABLE IV—SENSITIVITY FUNCTIONS FOR THE NEGATIVE REAL POLE AND ZERO OF THE TWIN-T LOCATED AT $-\omega_1$

Zero Sensitivity	Pole Sensitivity
$S_{R_1}^{-\omega_1} = \omega_1 \frac{\alpha^2}{1+\alpha^2}(1-\lambda)$	$S_{R_1}^{-\omega_1} = \omega_1 \frac{\alpha^2}{\frac{\lambda(1-\nu)}{\nu(1-\lambda)}+\alpha^2}(1-\lambda)$
$S_{R_2}^{-\omega_1} = \omega_1 \frac{\alpha^2}{1+\alpha^2}\lambda$	$S_{R_2}^{-\omega_1} = \omega_1 \frac{\alpha^2}{\frac{\lambda(1-\nu)}{\nu(1-\lambda)}+\alpha^2}\lambda$
$S_{R_3}^{-\omega_1} = \omega_1 \frac{1}{1+\alpha^2}$	$S_{R_3}^{-\omega_1} = \omega_1 \frac{1}{1+\alpha^2\frac{\nu(1-\lambda)}{\lambda(1-\nu)}}$
$S_{C_1}^{-\omega_1} = \omega_1 \frac{1-\nu}{1+\alpha^2}$	$S_{C_1}^{-\omega_1} = \omega_1 \frac{(1-\nu)}{1+\alpha^2\frac{\nu(1-\lambda)}{\lambda(1-\nu)}}$
$S_{C_2}^{-\omega_1} = \omega_1 \frac{\nu}{1+\alpha^2}$	$S_{C_2}^{-\omega_1} = \omega_1 \frac{\nu}{1+\alpha^2\frac{\nu(1-\lambda)}{\lambda(1-\nu)}}$
$S_{C_3}^{-\omega_1} = \omega_1 \frac{\alpha^2}{1+\alpha^2}$	$S_{C_3}^{-\omega_1} = \omega_1 \frac{\alpha^2}{\frac{\lambda(1-\nu)}{\nu(1-\lambda)}+\alpha^2}$

include all symmetrical configurations in which the series elements are identical as well as potentially symmetrical configurations in which the series elements are characterized by relations of the type

$$\begin{aligned} R_2 &= aR_1, \\ C_2 &= C_1/a. \end{aligned} \tag{94}$$

APPENDIX C

Twin-T Zero Sensitivity

Expressing the numerator $N(s)$ of the twin-T transfer function

$$N(s) = A_x(s) + xB_x(s) \tag{95}$$

the null return difference $F_x^0(s)$ with respect to x is given by

$$F_x^0(s) = \frac{N(s)}{A_x(s)} = 1 + x\,\frac{B_x(s)}{A_x(s)}. \tag{96}$$

With equations (1) and (2), the null return difference of the twin-T with respect to its six components can be calculated directly.

To obtain the null return difference of the nulled twin-T, equation (8) can be substituted into equation (96), namely

$$F_x^0(s) = \frac{(s + \omega_1)(s^2 + \omega_N^2)}{\omega_1\omega_N^2 A_x(s)} \tag{97}$$

where, $\omega_1 = 1/R_3C_p = 1/R_pC_3$. The corresponding zero sensitivity then results as

$$S_x^{j\omega_N} = \frac{s - j\omega_N}{F_x^0(s)}\bigg|_{s=j\omega_N} = \frac{(s - j\omega_N)\omega_1\omega_N^2 A_x(s)}{(s + \omega_1)(s^2 + \omega_N^2)}\bigg|_{s=j\omega_N} \tag{98}$$

which simplifies to

$$S_x^{j\omega_N} = -\frac{\alpha(1 + j\alpha)}{2(1 + \alpha^2)}\,\omega_N \cdot A_x(j\omega_N) \tag{99}$$

where $\alpha = \omega_1/\omega_N$. The individual $A_x(j\omega_N)$ functions follow directly from equations (1) and (2). Substituting these into equation (99), the zero sensitivity of the nulled twin-T with respect to its six components is obtained. These are listed in Table I.

REFERENCES

1. Augustadt, H. W., Electric Filter, U. S. Patent No. 2, 106, 785, February 1, 1938; see also: "Circuit Classics Electric Filter," Elec. Equipment Eng., No. 33 (February 1965), p. 35.
2. Scott, H. H., "A New Type of Selective Circuit and Some Applications," Proc. IRE, *26*, No. 2 (February 1938), pp. 226–235.
3. Tuttle, W. N., "Bridged-T and Parallel-T Null Circuits for Measurements at Radio Frequencies," Proc. IRE, *28*, No. 1 (January 1940), pp. 23–30.
4. Hastings, A. E., "Analysis of a Resistance-Capacitance Parallel-T Network and Applications," Proc. IRE, *34*, No. 3 (March 1946), pp. 126–129.
5. Stanton, L., "Theory and Application of Parallel-T Resistance Capacitance Frequency-Selective Networks," Proc. IRE, *34*, No. 7 (July 1946), pp. 447–456.
6. Wolf, A., "Note on a Parallel-T Resistance-Capacitance Network," Proc. IRE, *34*, No. 9 (September 1946), p. 659.
7. Cowles, L. G., "The Parallel-T Resistance-Capacitance Network," Proc. IRE, *40*, No. 12 (December 1952), pp. 1712–1717.
8. Oono, Y., "Design of Parallel-T Resistance-Capacitance Networks," Proc. IRE, *43*, No. 9 (May 1955), pp. 617–619.
9. Smith, D. H., "The Characteristics of Parallel-T RC Networks," Electronic Engineering, *29*, No. 348 (February 1957), pp. 71–77.
10. Bolle, A. P., "Theory of Twin-T RC Networks and their Application to Oscillators," J. British IRE, *13*, No. 12 (December 1953), pp. 571–587.
11. Dutta Roy, S. C., "On the Design of Parallel-T Resistance-Capacitance Networks for Maximum Selectivity," J. Inst. Telecommunication Engineers (India), *8*, No. 5 (September 1962), pp. 218–223.

12. Mehta, V. B., "Comparison of RC Networks for Frequency Stability in Oscillators," Proc. IEE, *112*, No. 2 (February, 1965), pp. 296–300.
13. Slaughter, J. B., and Rosenstein, A. B., "Twin-T Compensation Using Root Locus Methods," AIEE Trans., *81*, Part II, No. 64 (January 1963), pp. 339–350.
14. Lazear, T. J. and Rosenstein, A. B., "Pole-Zero Synthesis and the General Twin-T" AIEE Trans., *83*, Part II, No. 11 (November 1964), pp. 389–393.
15. Barker, A. C. and Rosenstein, A. B., "S-Plane Synthesis of the Symmetrical Twin-T Network," AIEE Trans., Part II, *83*, No. 11 (November 1964), pp. 382–388.
16. Hollister, F. H., and Thaler, G. J., "Symmetrical Parallel-Tee Network—Parameter Plane Analysis and Synthesis," Proc. Nat. Elec. Conf., *21* (October 1965), pp. 430–438.
17. Hollister, F. H., and Thaler, G. S., "Loaded and Null Adjusted Symmetrical Parallel-Tee Network," Proc. Nat. Elec. Conf., *21* (October 1965), pp. 753–758.
18. Mitra, S. K., "A Note on the Design of RC Notch Networks with Maximum Gain," Proc. IEEE, *54*, No. 10 (October 1966), p. 1487.
19. Shenoi, B. A., "A New Technique for Twin-T RC Network Synthesis," IEEE Trans. Circuit Theory, *CT-11*, No. 3 (September 1964), pp. 435–436.
20. Hakimi, S. L., and Seshu, S., "Realization of Complex Zeros of Transmission by Means of RC Networks," Proc. Nat. Elec. Conf., *13* (October 1957), pp. 1013–1025.
21. Holt, A. G. J., and Reineck, K. M., "Synthesis of RC Zero Sections," Radio and Electronic Engineer, *33*, No. 1 (January 1967), pp. 9–15.
22. Hillan, A. B., "The Parallel-T Bridge Amplifier," J. Inst. Elec. Eng., *94*, Part III, No. 27 (January 1947), pp. 42–51.
23. Sallen, R. P., and Key, E. L., "A Practical Method of Designing RC Active Filters," IRE Trans. Circuit Theory, *CT-2*, No. 1 (March 1955), pp. 74–85.
24. Bachmann, A. E., "Transistor Active Filters Using Twin-T Rejection Networks," Proc. IEEE *106*, Part B, No. 26 (March 1959), pp. 170–174.
25. Balabanian, N., and Cinkilie, I., "Expansion of an Active Synthesis Technique," IEEE Trans. Circuit Theory, *CT-10*, No. 2 (June 1963), pp. 290–298.
26. Balabanian, N., and Patel, B., "Active Realization of Complex Zeros," IEEE Trans. Circuit Theory, *CT-10*, No. 2 (June 1963), pp. 299–300.
27. Piercey, R. N. G., "Synthesis of Active RC Filter Networks," ATE J., *21*, No. 2 (April 1965), pp. 61–75.
28. Kerwin, W. J., and Huelsman, L. P., "The Design of High Performance Active RC Bandpass Filters," IEEE Int. Conv. Record, 14, Part 10 (March 1966), pp. 74–80.
29. Moschytz, G. S., "Active RC Filter Building Blocks Using Frequency emphasizing Networks," IEEE J. Solid-State Circuits, *SC-2*, No. 2 (June 1967), pp. 59–62.
30. Moschytz, G. S., "Sallen and Key Filter Networks with Amplifier Gain Larger than or Equal to Unity," IEEE J. Solid-State Circuits, *SC-2*, No. 3 (September 1967), pp. 114–116.
31. Moschytz, G. S., "Miniaturized Filter Building Blocks Using Frequency Emphasizing Networks," Proc. Nat. Elec. Conf., *23*, (October 1967), pp. 364–369, 1967; also "FEN Filter Design Using Hybrid Integrated Building Blocks," Proc. IEEE, *58*, No. 4 (April 1970), pp. 550–566.
32. Moschytz, G. S., "Two-Step Precision Tuning of a Twin-T Notch Filter," Proc. IEEE, *54*, No. 5 (May 1966), pp. 811–812.
33. Mitra, S. K., *Analysis and Synthesis of Linear Active Networks*, New York: John Wiley and Sons, Inc., 1969, pp. 172, 202.
34. Hakimi, S. L., and Cruz, J. B., "On Minimal Realization of RC Two-Ports," Proc. Nat. Elec. Conf., *16* (October 1960), pp. 258–267.
35. Parisi, G., "Control of Temperature Coefficient of Resistance by Reactive Sputtering of Tantalum with Nitrogen and Oxygen Simultaneously," Proc. Elec. Components Conf., 1969. Washington, D. C., April 1969, pp. 366–371.

Thermal and Electrical Properties of Coated Conductive Substrates for Integrated Circuit Chip Mounting

By H. G. MATTES

(Manuscript received February 10, 1970)

Presently used substrates for integrated circuit chip mounting and interconnection provide limited heat sink capability and preclude the effective use of matched impedance transmission line interconnection. These characteristics can be improved by the use of very narrow microstrip lines on thin dielectric layers backed by thermally and electrically conductive material. We propose a model for calculating the effective thermal resistance of layered substrates and make calculations for a variety of materials. The thermal problem is considered for heat extraction either through the edge or through the face of the substrate. Improvements in substrate thermal conductance of an order of magnitude appear realizable compared to presently used substrates (for example, 0.625 mm alumina). We analyze the electrical parameters of narrow (0.02–0.1 mm) microstrip lines on coated substrates in the frequency range 10 MHz to 10 GHz. The characteristic impedance, propagation delay time, and attenuation are found to be frequency dependent, and efforts to minimize this frequency dependence by magnetic loading result in greater, though more constant, delay times. Losses may be significant (~2-5 dB/cm in many cases), but short line lengths due to dense circuit packing allowed by improved heat dissipating capability will minimize this disadvantage.

I. INTRODUCTION

A necessary component in realizing the potential of integrated circuit technology is the substrate upon which the integrated circuit chip is mounted. The substrate must meet several requirements to avoid deterioration of the system's performance:

(*i*) The substrate must act as a rigid, reliable mounting platform for the chip.

(*ii*) The substrate must allow the dissipation of heat produced on the chip without causing an increase in temperature which would detrimentally effect circuit operation.

(*iii*) The substrate must permit electrical connections to and interconnections between chips in a manner compatible with desired circuit performance.

(*iv*) The substrate fabrication and chip mounting must be realizable at a cost consistent with the cost requirements of the system.

Presently used substrates (for example, 0.625 mm thick Al_2O_3 ceramic wafers) are limited in their attainment of objectives *ii* and *iii* outlined above. Heat sink capability is limited by the low thermal conductivity and relatively large thickness of material necessary for mechanical rigidity and by the difficulty in extracting heat from the substrate itself. High frequency (>500 MHz) performance of beam-leaded integrated circuits suffers from the transmission line mismatch necessary with conventional substrates as is discussed below.

A new materials configuration, as shown in Fig. 1, is envisioned which would simultaneously satisfy the four basic substrate requirements listed above. A thin dielectric layer (0.001 mm to 0.05 mm thick) deposited or grown on a massive electrically and thermally conductive ground plane could allow efficient heat dissipation, effective electrical interconnection and provide a rigid, inexpensive bonding plane. It is the purpose of this paper to calculate the thermal and electrical properties of the coated substrate configuration suggested here.

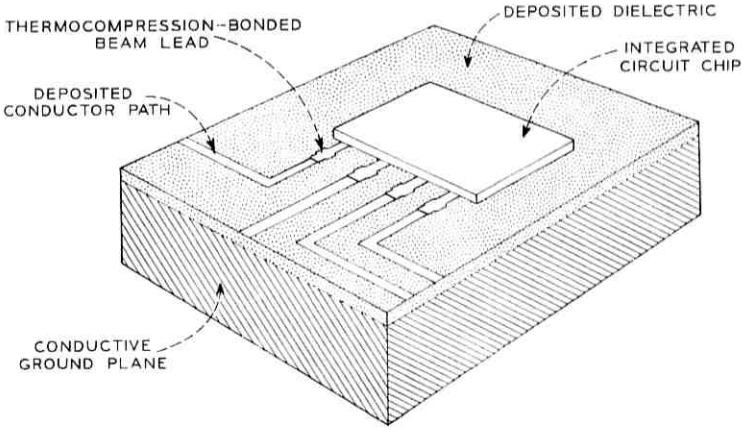


Fig. 1—Schematic drawing of layered substrate.

II. THERMAL PROPERTIES

The model used to represent the proposed substrate for calculation of thermal properties is shown in Fig. 2. All constants may be arbitrarily set, and the heat flow may be subjected to either of two distinct boundary conditions: the sides of the disk may be held fixed at a constant temperature and the top and bottom faces insulated, or the bottom face may be held at a constant temperature and the sides and top insulated. These cases closely approximate two heat sinking configurations frequently found in practice. The cylindrical geometry was chosen for ease of calculation (the square or rectangular problem, although algebraically simpler, results in a double series which presents formidable convergence problems when one is interested in a numerical result). The specific geometry should have little effect on the calculated values, and any reasonable geometry (held constant throughout the calculations)

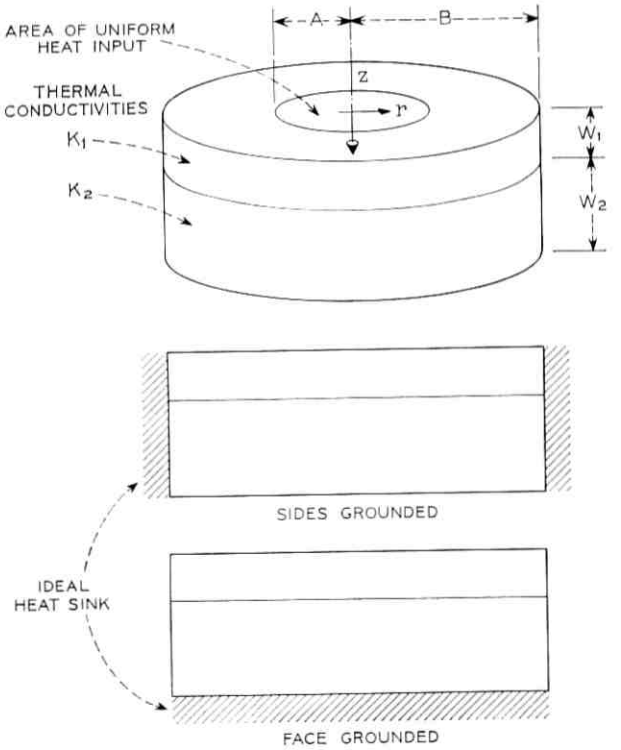


Fig. 2—Model used in calculations of thermal properties.

should provide a reliable comparison between alternate substrates. The heat input was assumed uniform over a circle of radius A, although other heat input distributions might have been chosen. As in the choice of geometry, this should have a minimal effect on the results. (Currently used beam leaded chips distribute heat over an area larger than the silicon dimensions. It can be assumed that the entire chip-beam lead structure thus forms a relatively uniform heat source on a substrate.[1])

2.1 *Algebraic Solution*

The thermal problem has already been solved for a homogeneous cylinder.[2] In the case of a layered cylinder, it is necessary to solve Laplace's equation simultaneously in both the upper and lower regions and match boundary conditions at the interface. In the calculations described in this paper, no discrete thermal resistance was inserted at the interface so the matching of boundary conditions reduced to

$$K_1 \frac{\partial T_1}{\partial z} = K_2 \frac{\partial T_2}{\partial z} \quad \text{at} \quad z = W_1 \tag{1}$$

where K_1, T_1, K_2, T_2 are the thermal conductivities and temperatures in regions 1 (upper) and 2 (lower). The insertion of a thermal resistance at the interface is trivial but was approximated as zero due to the intimate interface contact obtained by the method of materials preparation envisioned for these substrates.

Laplace's equation was solved for the two sets of boundary conditions (sides grounded and face grounded) by the method of separation of variables. In cylindrical coordinates, $\nabla^2 T = 0$ reduces to

$$T = \Psi_0 + \Psi_1 z + \sum_m \sum_n \Psi_{m,n}[A_m J_m(\delta_n r) + B_m N_m(\delta_n r)]\begin{Bmatrix}\sin \\ \cos\end{Bmatrix} m\theta \, [\exp(\pm \delta_n z)] \tag{2}$$

where boundary conditions exist to impose eigenfunction solutions in the r direction, and J_m and N_m are Bessel functions of the first and second kind respectively. The boundary condition on the grounded surfaces was given as $T = 0$ while $\partial T/\partial$ (normal) was set equal to zero on the insulated surfaces. On the top surface ($z = 0$) the boundary condition was written:

$$\frac{\partial T_1}{\partial z} = \begin{cases} \dfrac{-f}{k_1} & 0 < r < A, \quad z = 0; \\ 0 & A < r < B, \quad z = 0. \end{cases} \tag{3}$$

f is the value of heat input in (watts/unit area). This boundary condition was met by using the orthogonality properties of the Bessel functions in a Fourier-Bessel series as outlined in Morse and Feshbach and other texts.[3]

For the case in which the sides are grounded at $T = 0$, the temperature distribution is found to be

$$T = \sum_n \frac{2fA}{\beta_n^2 K_1} \frac{J_1\left(\beta_n \frac{A}{B}\right) J_0\left(\beta_n \frac{r}{B}\right)}{J_1^2(\beta_n)} \left[c_1 \cosh\left(\frac{\beta_n z}{B}\right) + c_2 \sinh\left(\frac{\beta_n z}{B}\right)\right] \quad (4)$$

where β_n is the nth zero of $J_0(r)$ and

$$c_1 = \begin{cases} \dfrac{\frac{K_2}{K_1}\tanh^2\left(\frac{\beta_n W_1}{B}\right) - \left(\frac{K_2}{K_1} - 1\right)\tanh\left(\frac{\beta_n W_1}{B}\right)\tanh\left(\frac{\beta_n (W_1 + W_2)}{B}\right) - 1}{\tanh\left(\frac{\beta_n W_1}{B}\right)\left[1 - \frac{K_2}{K_1} - \tanh\left(\frac{\beta_n W_1}{B}\right)\tanh\left(\frac{\beta_n (W_1 + W_2)}{B}\right)\right] + \frac{K_2}{K_1}\tanh\left(\frac{\beta_n (W_1 + W_2)}{B}\right)}, \\ \qquad \text{for } 0 < z < W_1; \\ \dfrac{\tanh^2\left(\frac{\beta_n W_1}{B}\right) - 1}{\tanh\left(\frac{\beta_n W_1}{B}\right)\left[1 - \frac{K_2}{K_1} - \tanh\left(\frac{\beta_n W_1}{B}\right)\tanh\left(\frac{\beta_n (W_1 + W_2)}{B}\right)\right] + \frac{K_2}{K_1}\tanh\left(\frac{\beta_n (W_1 + W_2)}{B}\right)}, \\ \qquad \text{for } W_1 < z < W_1 + W_2; \end{cases} \quad (5)$$

$$c_2 = \begin{cases} 1, \quad \text{for } 0 < z < W_1; \\ \dfrac{\tanh\left(\frac{\beta_n (W_1 + W_2)}{B}\right)\left[1 - \tanh^2\left(\frac{\beta_n W_1}{B}\right)\right]}{\tanh\left(\frac{\beta_n W_1}{B}\right)\left[1 - \frac{K_2}{K_1} - \tanh\left(\frac{\beta_n W_1}{B}\right)\tanh\left(\frac{\beta_n (W_1 + W_2)}{B}\right)\right] + \frac{K_2}{K_1}\tanh\left(\frac{\beta_n (W_1 + W_2)}{B}\right)}, \\ \qquad \text{for } W_1 < z < W_1 + W_2. \end{cases} \quad (6)$$

For the case in which the bottom face is grounded at $T = 0$, the temperature distribution is found to be

$$T = \frac{A^2}{B^2} f(d_1 + d_2 Z) + \sum_n \frac{2fA}{\alpha_n^2 K_1} \frac{J_1\left(\alpha_n \frac{A}{B}\right) J_0\left(\alpha_n \frac{r}{B}\right)}{J_0^2(\alpha_n)} \left[d_3 \cosh\left(\frac{\alpha_n z}{B}\right) + d_4 \sinh\left(\frac{\alpha_n z}{B}\right)\right] \quad (7)$$

where α_n is the nth zero of $J_1(r)$ and

$$d_1 = \begin{cases} -\dfrac{W_1}{K_1} - \dfrac{W_2}{K_2} & 0 < z < W_1 \\ -\dfrac{W_1 + W_2}{K_2} & W_1 < z < W_1 + W_2 \end{cases}; \quad (8)$$

$$d_2 = \begin{cases} \dfrac{1}{K_1} & 0 < z < W_1 \\ \dfrac{1}{K_2} & W_1 < z < W_1 + W_2 \end{cases}; \tag{9}$$

$$d_3 = \begin{cases} \dfrac{\tanh\left(\frac{\alpha_n(W_1+W_2)}{B}\right) - \tanh\left(\frac{\alpha_n W_1}{B}\right)\left\{\frac{K_2}{K_1}\left[\tanh\left(\frac{\alpha_n W_1}{B}\right)\tanh\left(\frac{\alpha_n(W_1+W_2)}{B}\right) - 1\right] + 1\right\}}{\tanh^2\left(\frac{\alpha_n W_1}{B}\right) + \left(\frac{K_2}{K_1} - 1\right)\tanh\left(\frac{\alpha_n W_1}{B}\right)\tanh\left(\frac{\alpha_n(W_1+W_2)}{B}\right) - \frac{K_2}{K_1}}, \\ \qquad\qquad \text{for } 0 < z < W_1; \\ \dfrac{\tanh\left(\frac{\alpha_n(W_1+W_2)}{B}\right)\left[1 - \tanh^2\left(\frac{\alpha_n W_1}{B}\right)\right]}{\tanh^2\left(\frac{\alpha_n W_1}{B}\right) + \left(\frac{K_2}{K_1} - 1\right)\tanh\left(\frac{\alpha_n W_1}{B}\right)\tanh\left(\frac{\alpha_n(W_1+W_2)}{B}\right) - \frac{K_2}{K_1}}, \\ \qquad\qquad \text{for } W_1 < z < W_1 + W_2; \end{cases} \tag{10}$$

$$d_4 = \begin{cases} 1 \quad \text{for } 0 < z < W_1 \\ \dfrac{\tanh^2\left(\frac{\alpha_n W_1}{B}\right) - 1}{\tanh^2\left(\frac{\alpha_n W_1}{B}\right) + \left(\frac{K_2}{K_1} - 1\right)\tanh\left(\frac{\alpha_n W_1}{B}\right)\tanh\left(\frac{\alpha_n(W_1+W_2)}{B}\right) - \frac{K_2}{K_1}}, \\ \qquad\qquad \text{for } W_1 < z < W_1 + W_2. \end{cases} \tag{11}$$

A quantitative measure of the power dissipating capabilities of a substrate is given by the thermal resistance

$$R_{\text{therm}}\left(\frac{\text{degrees C}}{\text{watt}}\right) = \frac{\Delta T(^\circ\text{C})}{q\ (\text{watts input})} = \frac{T_{\max}}{\pi A^2 f} \tag{12}$$

$T_{\max}$ clearly occurs at $z = 0$, $r = 0$ so

$$R_{\text{therm}}\ (\text{side grounded}) = \frac{2}{\pi K_1 A}\sum_n \frac{1}{\beta_n^2}\frac{J_1\left(\beta_n \frac{A}{B}\right)}{J_1^2(\beta_n)} \cdot \frac{\frac{K_2}{K_1}\tanh^2\left(\frac{\beta_n W_1}{B}\right) - \left(\frac{K_2}{K_1} - 1\right)\tanh\left(\frac{\beta_n W_1}{B}\right)\tanh\left(\frac{\beta_n(W_1+W_2)}{B}\right) - 1}{\tanh\left(\frac{\beta_n W_1}{B}\right)\left[1 - \frac{K_2}{K_1} - \tanh\left(\frac{\beta_n W_1}{B}\right)\tanh\left(\frac{\beta_n(W_1+W_2)}{B}\right)\right] + \frac{K_2}{K_1}\tanh\left(\frac{\beta_n(W_1+W_2)}{B}\right)} \tag{13}$$

and

$$R_{\text{therm}}\ (\text{face grounded}) = \frac{W_1 K_2 + W_2 K_1}{\pi B^2 K_1 K_2} + \frac{2}{\pi K_1 A}\sum_n \frac{1}{\alpha_n^2}\frac{J_1\left(\alpha_n \frac{A}{B}\right)}{J_0^2(\alpha_n)} \cdot \frac{\tanh\left(\frac{\alpha_n(W_1+W_2)}{B}\right) - \tanh\left(\frac{\alpha_n W_1}{B}\right)\left\{\frac{K_2}{K_1}\left[\tanh\left(\frac{\alpha_n W_1}{B}\right)\tanh\left(\frac{\alpha_n(W_1+W_2)}{B}\right) - 1\right] + 1\right\}}{\tanh^2\left(\frac{\alpha_n W_1}{B}\right) + \left(\frac{K_2}{K_1} - 1\right)\tanh\left(\frac{\alpha_n W_1}{B}\right)\tanh\left(\frac{\alpha_n(W_1+W_2)}{B}\right) - \frac{K_2}{K_1}}. \tag{14}$$

2.2 *Computer Program*

A FORTRAN IV computer program was written for an IBM 360-50 to evaluate $R_{\text{therm-side}}$ and $R_{\text{therm-face}}$. The first forty values of alpha and beta (the zeros of J_1 and J_0) are inserted in the program from the tabulated values of Watson.[4] The values of the higher zeros are obtained in the program by analysis of the asymptotic approximations for J_1 and J_0.

Since the expressions for R_{therm} take the form of infinite series, no completely precise evaluation can be made. Figure 3 shows the approximation to $R_{\text{therm-face}}$ given by truncating the summation after varying numbers of terms. Plotting data similar to Figure 3 for a range of all input parameters (A, B, K_1, K_2, W_1, W_2) disclosed several trends in the convergence of the series. The "period" of the oscillation was inversely correlated to the ratio A/B. This is mathematically consistent with the use of a Fourier-Bessel series to describe the top face of the disk. For ratios of A/B less than 0.02, the first maxima in the expression for $R_{\text{therm-side}}$ is not reached until partial sums with greater than 300 terms are evaluated. The amplitude of the oscillations in the function R_{therm} (number of terms) was positively correlated to K_2/K_1 and W_1/W_2. The damping of the oscillation (per cycle) appeared to be relatively

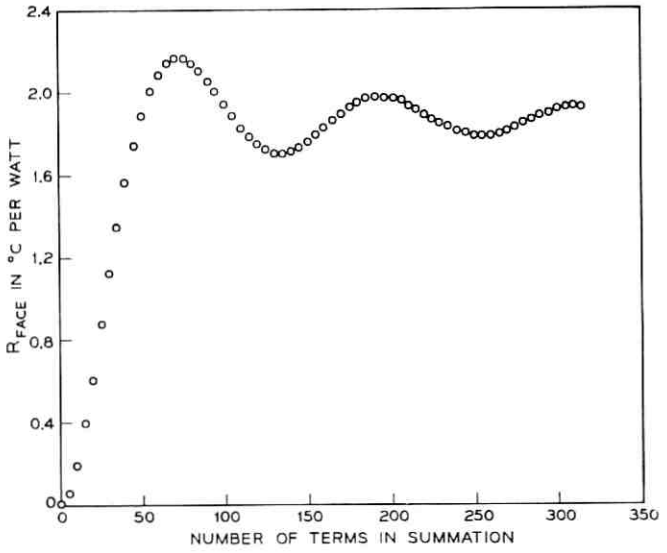


Fig. 3—Partial sums approximating R_{face} for $A = 0.0625$, $B = 3.75$, $K_1 = 0.20$, $K_2 = 2.05$, $W_1 = 0.0005$, $W_2 = 0.0625$.

independent of input data. These variabilities in the behavior of R_{therm} made determination of a precise value difficult. Additionally, only the first 314 terms of the series could be evaluated. The operation of the program was terminated at this point by a floating point overflow which occurred in the IBM-SSP subroutine used to generate the Bessel functions. Since the largest arguments of the Bessel functions are intrinsic to the problem, and are not functions of the input parameters, nothing could be done to circumvent this difficulty. It thus became necessary to evaluate R_{therm} on the basis of a truncated series of 314 terms of less.

A graphical outline of the method used to extract an approximation to the asymptotic value of R_{therm} from a finite number of terms in shown in Fig. 4. The first two local maxima, M_1 and M_3, and the first local minimum, M_2, were recorded; and the averages, A_1 and A_2, were calculated where

$$A_1 = \frac{M_1 + M_2}{2}; \tag{15}$$

$$A_2 = \frac{M_2 + M_3}{2}. \tag{16}$$

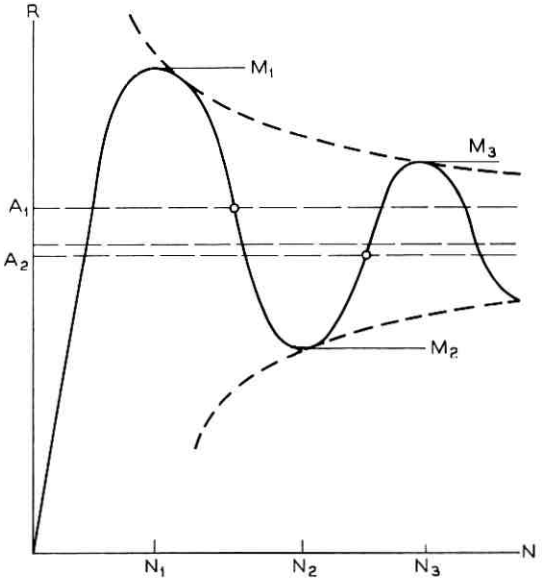


Fig. 4—Pictorial of extremum averaging technique used for improving series convergence.

In Appendix A, a theorem is stated and proved to show that the asymptotic value of R_{therm} is less than A_1 and greater than A_2. An average of the two, $(A_1 + A_2)/2$, was used as a most probable value of R_{therm}.

As a test of the accuracy of the derivation of $R_{therms\text{-}side}$ and $R_{therm\text{-}face}$ and the computer programs for generating these values, K_2 was set equal to K_1 (equivalent to a uniform, single layer cylinder) and values of R_{therm} computed for $W_1 = 0.625$ mm, $W_2 = 0$; $W_1 = 0.375$ mm, $W_2 = 0.25$ mm; and $W_1 = 0$, $W_2 = 0.625$ mm. The three cases were in complete mutual agreement, and agreed to the accuracy given with the published results of D. P. Kennedy for heat conduction in a homogeneous, isotropic cylinder.[2]

2.3 *Numerical Evaluation*

Several materials combinations were chosen for evaluation as possible substrates. Some of the considerations for inclusion were compatibility with integrated circuit techniques (for example, silicon), smooth surfaces for microdeposition of conductor paths (for example, polished silicon, cold-rolled aluminum), low dielectric constant (for example, SiON), and thermal expansion coefficient. The materials considered and the associated values of K used with them are shown in Table I. $A = 0.625$ mm and $B = 3.75$ cm were chosen as representative of typical integrated circuit chips and substrate sizes and were used in most of the calculations.

The results of these calculations for a variety of layered substrates are shown in Figs. 5 through 10. With high conductivity dielectrics (for example, Al_2O_3) the thermal resistivity of the ground plane contributes significantly to the total thermal resistance, but resistances as low as that of beryllia are attainable for a considerable range of dielectric thicknesses. Low conductivity, glassy (for example, SiON, SiO_2) dielectrics, can result in thermal resistances equal to or greater than that of alumina if glassy layers of greater than 0.03 mm thick are used.

TABLE I—THERMAL CONDUCTIVITY OF SUBSTRATE MATERIALS

Material	K (watts/cm °C)
Al	2.05
Al_2O_3 (deposited)	0.20
Al_2O_3 (sintered)	0.29
BeO	1.95
Si	0.88
Si_3N_4	0.012
SiO_2	0.012
SiON	0.012

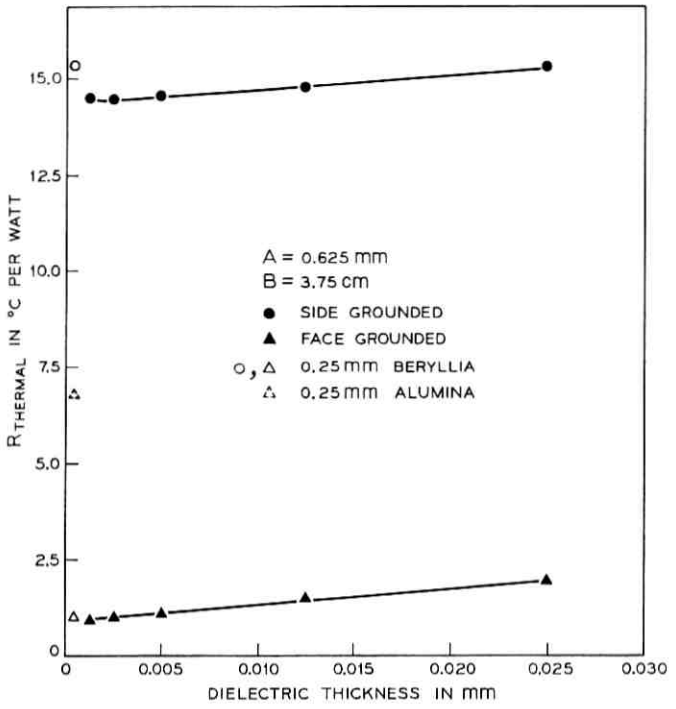


Fig. 5—Aluminum oxide on 0.25 mm aluminum.

For comparison, thermal resistances of presently available homogeneous substrates were calculated and are given in Table II.

Figure 11 shows the effect of substrate radius on $R_{\text{therm-side}}$ (the effect on $R_{\text{therm-face}}$ was $<2\%$ for $B \geqq 0.5$ cm). In the range shown, heat is presumably flowing radially at $r = B$ so increases in B add to R_{therm} at a rate proportional to $1/B$. Consequently, a "critical radius" exists above which $R_{\text{therm-side}}$ is relatively constant.

Radiative and convective heat losses will be negligible for the high conductivity layered substrates discussed in this paper. A lumped value of convective and radiation coefficient of 0.003 watts/cm.^{2} – °C results in an equivalent parallel thermal resistance of greater than 50°C/watt for all cases considered here.

2.4 *Thermal Results*

Layered substrates consisting of dielectric layers backed by thermally conducting ground planes can provide substantial improvements in heat

sink capability over alumina ceramic substrates. Ceramic substrates are commonly used in 0.625 mm thickness to provide needed strengths. Metal ground planes could be reduced to 0.25 mm thickness while maintaining physical strength greater than 0.625 mm ceramic. For situations where heat sinking the back of the substrate is possible, Al_2O_3 on 0.25 mm aluminum can provide up to twice the heat sink capability of 0.625 mm beryllia. Metal substrates have the additional advantage of convenient, low thermal barrier heat sink mounting or even direct incorporation into the heat sink structure (for example, heat pipes, finned substrates).

Although the above values of R_{therm} are significant, as given, for relative comparison of substrate materials and geometries, their absolute significance can only be assessed in relation to other parameters intrinsic to an integrated circuit heat dissipating system. V. E. Holt

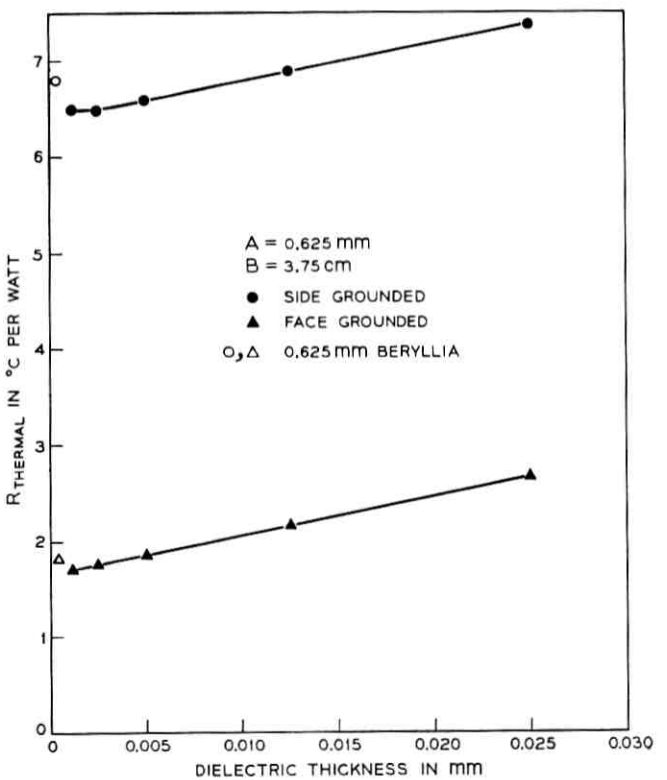


Fig. 6—Aluminum oxide on 0.625 mm aluminum.

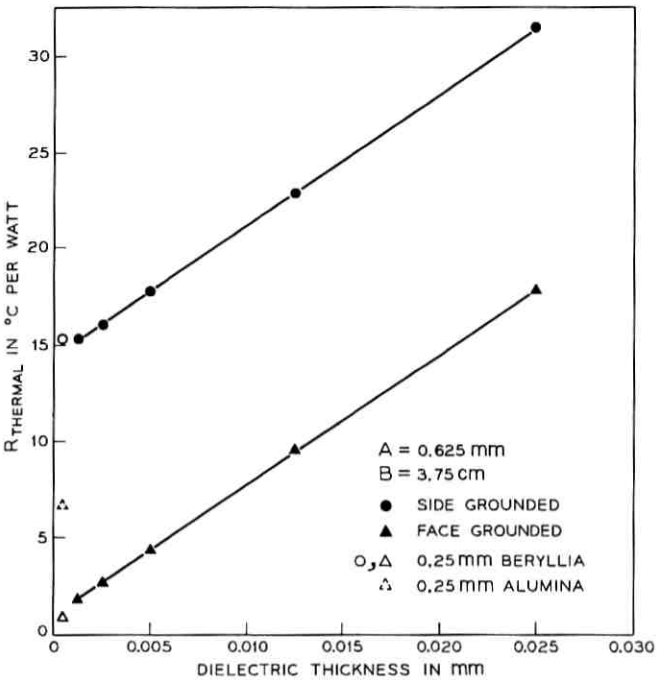


Fig. 7—Silicon oxynitride on 0.25 mm aluminum.

has found thermal barriers of 32°C/watt for 1.25 mm chips conducting heat to a substrate by means of beam leads and a resin bonding layer.[1] This value is reduced to 3 to 5°C/watt by AuSi eutectic bonding. Imperfections in the heat sink will have to be assessed in each situation encountered.

III. ELECTRICAL PROPERTIES

3.1 *Microstrip Transmission Lines*

Electrical interconnections on the proposed substrates would best be made in the form of standard microstrip transmission lines.[5] They can provide transmission line interconnects necessary for high speed pulse and microwave electronics while maintaining low values of crosstalk between adjacent lines and still provide dc and low frequency paths for other circuit requirements. The primary incompatibility of presently used substrates and beam-leaded integrated circuit chips is due to the impedance mismatch between the narrow (0.02 to 0.05 mm) beam leads

and the physically large transmission line (typically 50Ω implies 0.625 mm line width on 0.625 mm alumina).[5] The removal of this discontinuity would require narrow conductors deposited on a dielectric layer of appropriate thickness to maintain the desired transmission line impedance. The physical configuration contemplated for small microstrip lines is outlined in Fig. 12. To minimize the impedance discontinuity at the interconnection with beam leaded chips, W is restricted to 0.1 mm maximum. To minimize crosstalk, this would suggest $t_d \leqq 0.05$ mm[6]. These figures, coupled with the electrical parameters of dielectric materials sufficiently smooth to allow the reliable deposition of 0.02 mm conducting paths, are consistent with the desire to develop lines with a characteristic impedance of 50Ω, the most commonly used interface impedance for circuitry operating above 500 MHz. Under certain conditions, as seen below, it is desirable to raise the series inductance of the transmission line. For this reason, provision has been made for placing a thin layer of high permeability material between the conducting strip and the dielectric. In practice the adhesion metal of a sandwiched interconnect (dielectric—adhesion metal—gold) might serve this purpose.

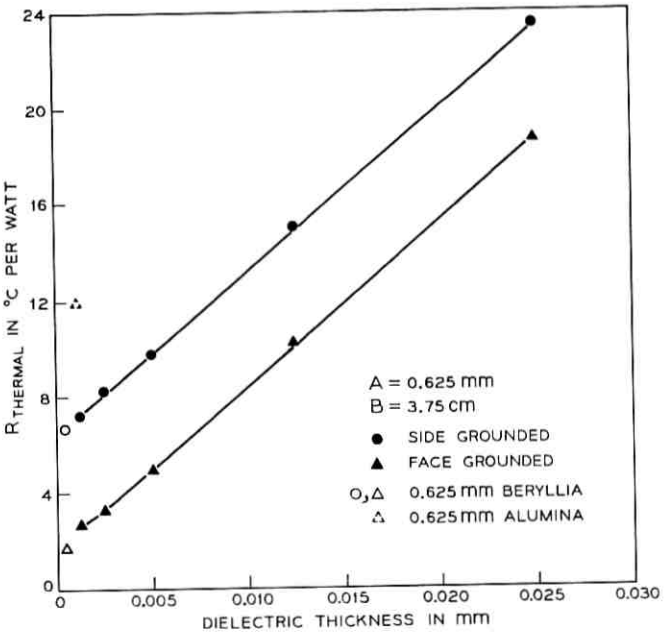


Fig. 8—Silicon oxynitride on 0.625 mm aluminum.

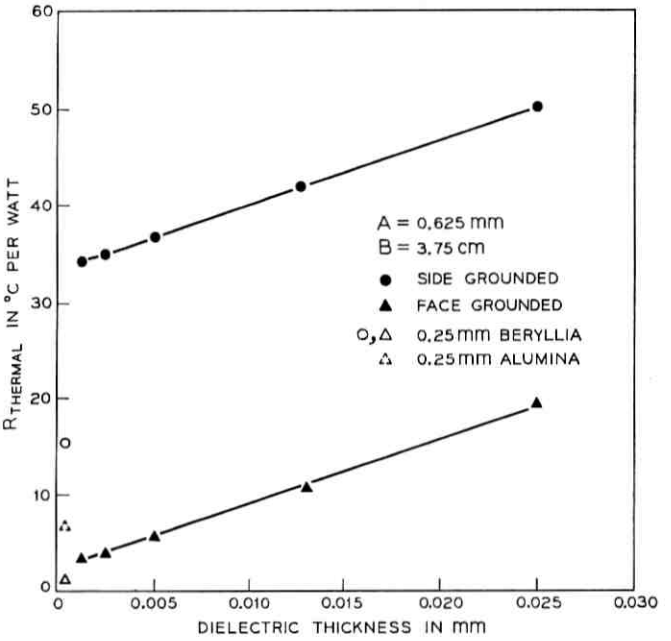


Fig. 9—Silicon dioxide on 0.25 mm silicon.

3.2 *Transmission Line Parameters*

Although many transmission modes are necessarily present in an unsymmetrical transmission line such as an unshielded microstrip, the TEM mode will be considered of dominant interest in the frequency range considered here and the line parameters will be calculated for this mode. Several authors (for example, Assadourian and Rimai,[7] and H. A. Wheeler[8]) have derived expressions for the characteristic impedance and other pertinent parameters for microstrip transmission lines. Although these formulae have been generally verified in their range of approximation[5,9,10] the physical scale of lines contemplated in the last section requires the inclusion of several terms which had been ignored in the previous derivations.

The characteristic impedance, Z_0, and propagation constant, γ, of any power transmission system are given by[11]

$$Z_0 = \left(\frac{R + j\omega L}{G + j\omega C}\right)^{\frac{1}{2}}, \tag{17}$$

$$\gamma = [(R + j\omega L)(G + j\omega C)]^{\frac{1}{2}}. \tag{18}$$

Ignoring losses (R, $G \rightarrow 0$), the propagation delay time can then be written

$$\tau = \frac{-j\gamma}{\omega} = (LC)^{\frac{1}{2}}. \tag{19}$$

The inductance of the line is given by two components: the inductance due to magnetic field energy storage in the strip conductor and the ground plane (the internal inductance); and the inductance due to magnetic field energy storage between the strip conductor and the ground plane (the external inductance). The internal inductance of the strip conductor can be derived from Ramo and Whinnery[12] and is given by

$$L_s = \frac{1}{W\omega}\left[\frac{\rho_s}{\delta_s}\frac{\sinh\left(\frac{2t_s}{\delta_s}\right) - \sin\left(\frac{2t_s}{\delta_s}\right)}{\cosh\left(\frac{2t_s}{\delta_s}\right) - \cos\left(\frac{2t_s}{\delta_s}\right)}\right] \tag{20}$$

where ρ_s is the resistivity of the strip conductor material and δ_s is the

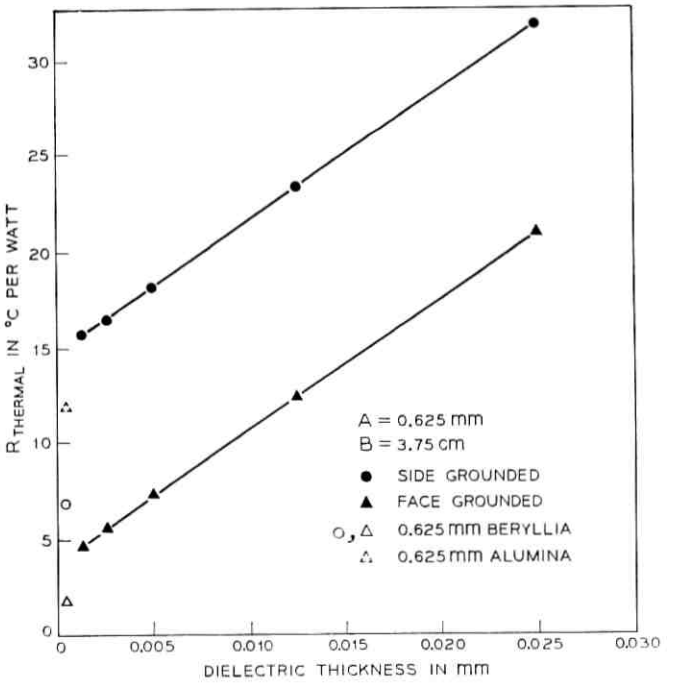


Fig. 10—Silicon dioxide on 0.625 mm silicon.

TABLE II—THERMAL RESISTANCE OF CONVENTIONAL SUBSTRATES

	$R_{\text{therm-side}}$ (degrees C/watt)	$R_{\text{therm-face}}$ (degrees C/watt)
Alumina		
0.25 mm	103	6.8
0.625 mm	46	12
Beryllia		
0.25 mm	15.3	1.02
0.625 mm	6.8	1.8

skin depth of the strip conductor at the frequency, ω, of interest. [$\delta = (2\rho/\omega\mu)^{\frac{1}{2}}$]. For the frequency range considered in this paper ($f > 10$ MHz) current spreading in the ground plane can be ignored for lines as narrow as 0.02 mm if the ground material has a resistivity $\leqq 10^{-5}$ ohm-cm. For higher resistivity materials this approximation is less valid, especially at the lower frequencies, but the approximation will be made for convenience. A more rigorous analysis in the higher resistivity case would give increased internal inductance and increased

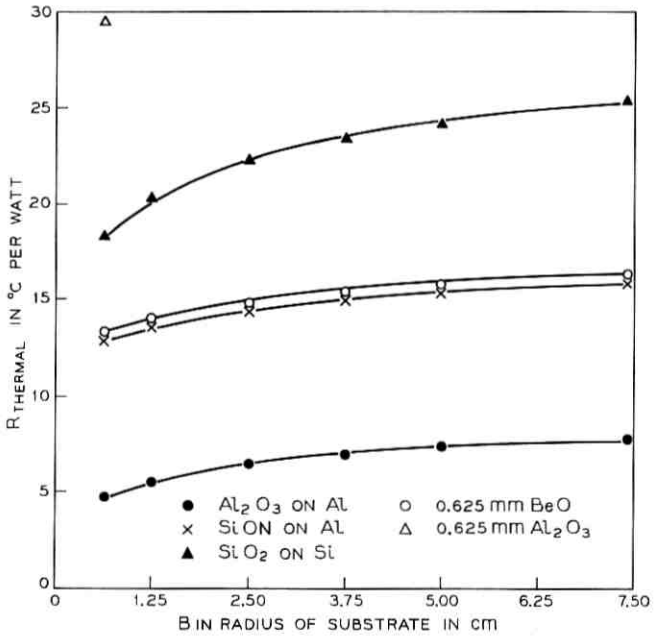


Fig. 11—Thermal resistance with side grounded $A = 0.625$ mm $W_1 = 0.0125$ mm, $W_2 = 0.625$ mm.

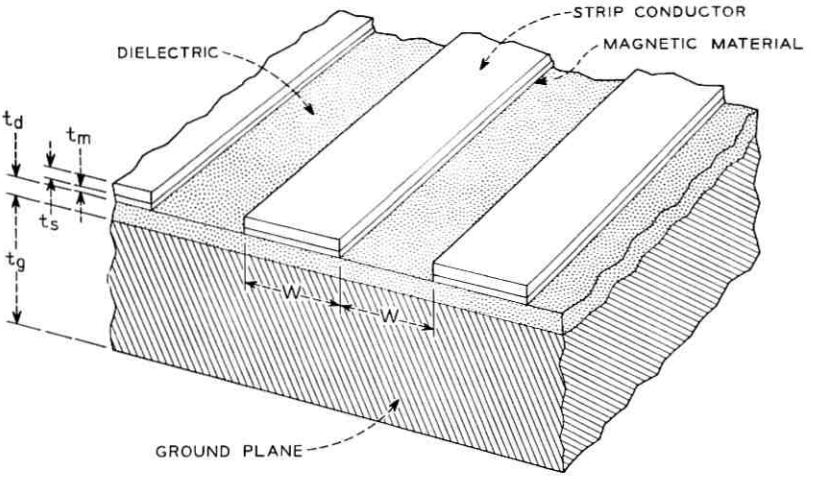


Fig. 12—Schematic drawing of microstrip transmission lines on layered substrate.

crosstalk through magnetic fields generated by spreading currents in the ground plane. The internal inductance per unit length of line due to fields in the ground plane can then be given by a formula similar to (20)

$$L_g = \frac{1}{W\omega}\left[\frac{\rho_g}{\delta_g}\frac{\sinh\left(\frac{2t_g}{\delta_g}\right) - \sin\left(\frac{2t_g}{\delta_g}\right)}{\cosh\left(\frac{2t_g}{\delta_g}\right) - \cos\left(\frac{2t_g}{\delta_g}\right)}\right] \tag{21}$$

where ρ_g and δ_g are the resistivity and skin depth of the ground plane. For the dimensional range of interest ($W > 2t_d$) the external inductance per unit length is approximately given by English and McNichol[13] as

$$L_{ext} = \frac{\mu_0}{W}(t_d + \mu_R t_m) \tag{22}$$

where μ_R is the relative permeability of the magnetic material. Total L is given by the sum

$$L = L_s + L_g + L_{ext}\,. \tag{23}$$

The total capacitance per unit length between the strip conductor and the ground plane is the sum of four components

$$C = C_T + C_g + C_f + C_e\,. \tag{24}$$

C_T is the capacitance per unit length between the top of the strip and the ground at infinity[14].

$$C_T = \frac{4\epsilon_0}{\pi} \ln 2. \tag{25}$$

C_g is the capacitance per unit length, neglecting fringing, between the face of the strip and the ground plane.

$$C_g = \epsilon_0 \epsilon_R \frac{W}{t_d} \tag{26}$$

where ϵ_R is the relative dielectric constant of the insulating layer. The additional capacitance per unit length due to the fringing fields at the edges of the strip is given by Joines[15]

$$C_f = \frac{4\epsilon_0 \epsilon_R}{\pi} \ln \left[1 + \left(1 - \exp \frac{-\pi W}{2t_d} \right)^{\frac{1}{2}} \right]. \tag{27}$$

For the case where the strip conductor is not of negligible thickness, an additional contribution to the capacitance is due the capacitance between the edges of the strip and the ground plane. The expression for this term, derived in Appendix B, is

$$C_s = \frac{2\epsilon_0 \epsilon_R}{\pi} \ln \left\{ \frac{2t_d + 2t_s + [2(t_d + t_s)(t_d + 2t_s)]^{\frac{1}{2}}}{t_d} \right\}. \tag{28}$$

Losses on these microstrip lines will be considered as a perturbation in the calculation of the other parameters although this approximation is quite poor in several of the cases considered below. Losses in the dielectric will be ignored, being much less than the conductor losses for the geometries discussed in this paper. Under considerations similar to those for deriving the internal inductance, the equivalent series resistance is given by[11,12]

$$R = \frac{1}{W} \left[\frac{\rho_s}{\delta_s} \frac{\sinh \left(\frac{2t_s}{\delta_s} \right) + \sin \left(\frac{2t_s}{\delta_s} \right)}{\cosh \left(\frac{2t_s}{\delta_s} \right) - \cos \left(\frac{2t_s}{\delta_s} \right)} + \frac{\rho_g}{\delta_g} \frac{\sinh \left(\frac{2t_g}{\delta_g} \right) + \sin \left(\frac{2t_g}{\delta_g} \right)}{\cosh \left(\frac{2t_g}{\delta_g} \right) - \cos \left(\frac{2t_g}{\delta_g} \right)} \right]. \tag{29}$$

The attenuation per unit length of line is given by[11]

$$\alpha = \frac{R}{2Z_0} \text{ (nepers)}. \tag{30}$$

Although this expression is rigorously valid only for $| R/\omega | \ll | L |$, it will be used in all computations made in this paper. In the case where

$| R/\omega | \gtrsim | L |$, the line will have a significant reactive component in its characteristic impedance. Equation (30) for α will then contain a contribution for phase shift as well as attenuation. In this paper $| Z_0 |$, Re (Z_0), and Re (α) will be calculated.

The equations for crosstalk between adjacent, parallel strip lines derived by Kordos[6] apply directly to microstrip lines of the dimensions considered here if the characteristic impedance is calculated by the means outlined above. In these cases, Kordos derives an expression for the near end crosstalk of properly terminated parallel lines of length, l.

$$\frac{V_{ne}(s)}{V_g(s)} \simeq \frac{30 \ln [1 + (t_d/W)^2]^{\frac{1}{2}}}{\text{Re } (Z_0)(\epsilon_R)^{\frac{1}{2}}} [1 - \exp (-2s\tau l)]. \tag{31}$$

$V_{ne}(s)$ and $V_g(s)$ are the Laplace transforms of the near-end crosstalk voltage and the input signal respectively. The far-end crosstalk is approximately zero. This formula is only valid for $t_m = 0$. For $t_m > 0$, V_{ne} would be further reduced by the "keeper" effect of the magnetic material on the magnetic field of the driven line.

3.3 *Numerical Evaluation*

The formulas given above have been applied to layered substrates of the type discussed earlier. The calculations assumed that the strip conductors (Au $- 2.5 \times 10^{-6}\Omega-$cm.) should exhibit a resistance of $\lesssim 0.004\Omega$/sq. to give a dc resistance $< 1.5\Omega$/cm for the narrowest line considered, 0.02 mm. This results in $t_s = 0.00625$ mm. Silicon and aluminum were considered as possible ground plane materials. The maximum doping level in silicon which allows the preservation of a smooth surface after processing is approximately 2×10^{19} atoms/cm^3 which results in a bulk resistivity of approximately $2.5 \times 10^{-3}\Omega-$cm. The constant parameters used in the calculations are shown in Table III.

The expressions given in Section 3.2 were evaluated for a spectrum of frequencies between 10 MHz and 10 GHz for several physically realizable dielectric-ground plane materials combinations with varying

TABLE III—ELECTRICAL PARAMETERS FOR STRIPLINE CALCULATIONS

(All dimensions in cm)

ρ_s (Au) $= 2.5 \times 10^{-6}$	$t_g = 6.25 \times 10^{-2}$
ρ_g (Al) $= 2.5 \times 10^{-6}$	ϵ_R (sintered Al_2O_3) $= 9.6$
ρ_g (Si) $= 2.5 \times 10^{-3}$	ϵ_R (deposited Al_2O_3) $= 9.0$
$\mu_R = 100$	ϵ_R (SiON) $= 3.8$
$t_s = 6.25 \times 10^{-4}$	ϵ_R (SiO_2) $= 3.6$

thicknesses of magnetic material. A representative sampling of the results is shown in Figs. 13 through 20. These data are not meant to be comprehensive or complete but merely to represent the effects and trends observed in the calculations.

Figures 13 through 16 give Re (Z_0) and τ as a function of dielectric thickness and/or line width. The effect of the large internal inductance of Si on Z_0 and τ is especially apparent for thin dielectrics (where the external inductance is small). It is noted in Figs. 13 and 14 that τ passes through a local minimum at $t_d \sim 0.01$ mm mils. Below this value (for aluminum ground planes) the constant internal inductance is greater than the external inductance; and the capacitance, which decreases with t_d, causes τ to decrease with t_d. Above the minimum, as the increasing external inductance predominates, the fringing capacitance does not decrease linearly with t_d so τ rises. As t_d increases further, the equations used in this paper cease to be valid and the inductance will not increase linearly with t_d.

In Figs. 17 and 18, the solid lines show the frequency dependence of the characteristic impedance. The variation of Re (Z_0) with ω is particu-

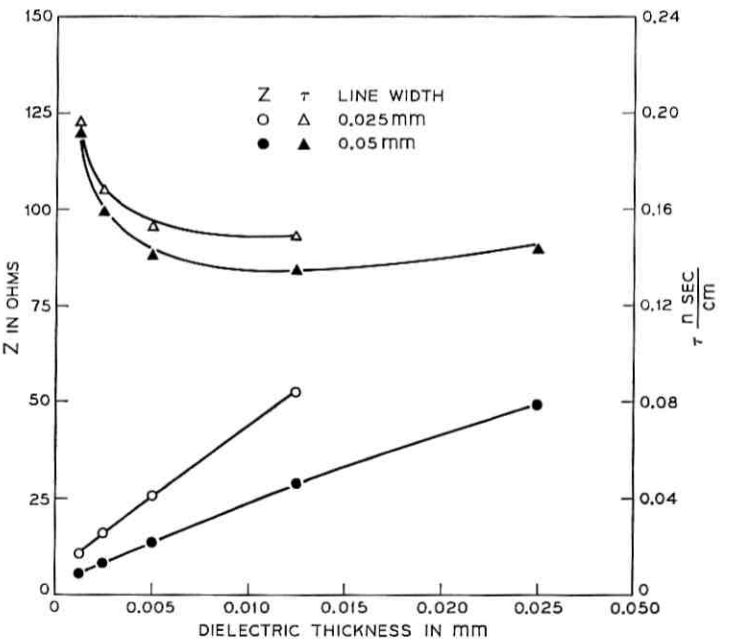


Fig. 13—Impedance and delay time at 1 GHz for aluminum oxide dielectric on aluminum ground plane.

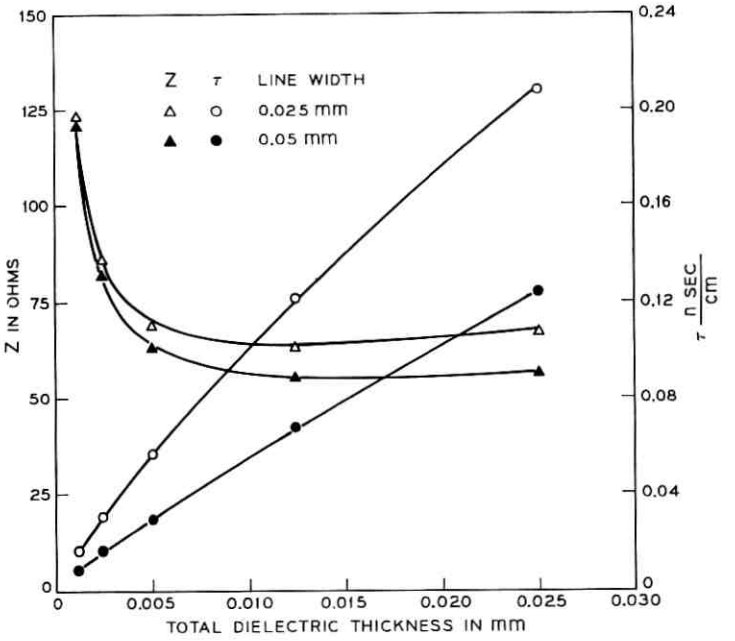


Fig. 14—Impedance and delay time at 1 GHz for silicon oxynitride dielectric on aluminum ground plane.

larly large in the case of Si since the frequency independent external inductance is small compared to the internal inductance. For pulse applications in which an impedance match is desirable over a broad frequency range, the situation may be improved somewhat by increasing the frequency independent component of the inductance. This may be done by loading the line with magnetic material, the results of which are shown by the dashed lines in Figs. 17 and 18. The open triangles show the unavoidably greater propagation delay times associated with the loaded lines. It can be seen from the deriving formulas that only the product $\mu_R t_m$ was considered significant, so materials of other permeabilities and appropriate thicknesses would be equally effective.

The losses associated with these lines have been calculated, and a sample of the results is shown in Fig. 19. Attenuation (dB/cm) is plotted for two materials configurations at each of two frequencies for a range of substrate thicknesses. The losses are particularly great for silicon ground planes (0.0025Ω-cm) and for small dielectric thicknesses (which result in low values of Z_0). The losses decrease rapidly as line width increases.

The near-end crosstalk for semi-infinite lines is shown in Fig. 20 as calculated without the influence of the magnetic material. For all geometries considered in this paper the near-end crosstalk is below 4 percent and is typically 1 percent or less. The presence of magnetic material would reduce this further.

For comparison, the transmission parameters of a 50Ω microstrip transmission line on a conventional 0.625 mm thick Alumina substrate are[5,15] given in Table IV.

3.4 *Electrical Results*

Effective transmission line interconnections for chip to chip on a layered integrated circuit substrate appear realizable. The major disadvantages are variation of impedance with frequency, appreciable delay times (and variation of delay time with frequency causing pulse distortion) and significant losses. These problems can be minimized by

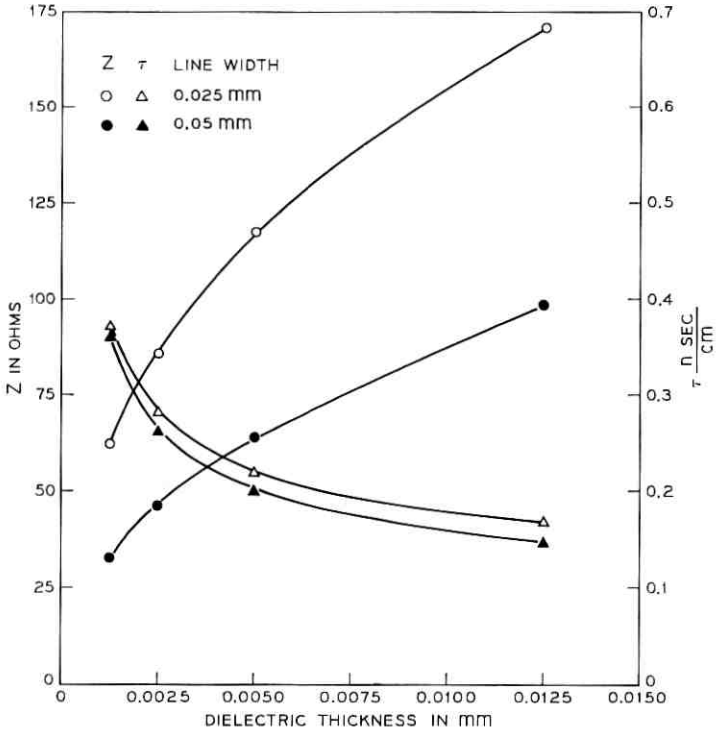


Fig. 15—Impedance and delay time at 1 GHz for silicon dioxide dielectric on silicon ground plane.

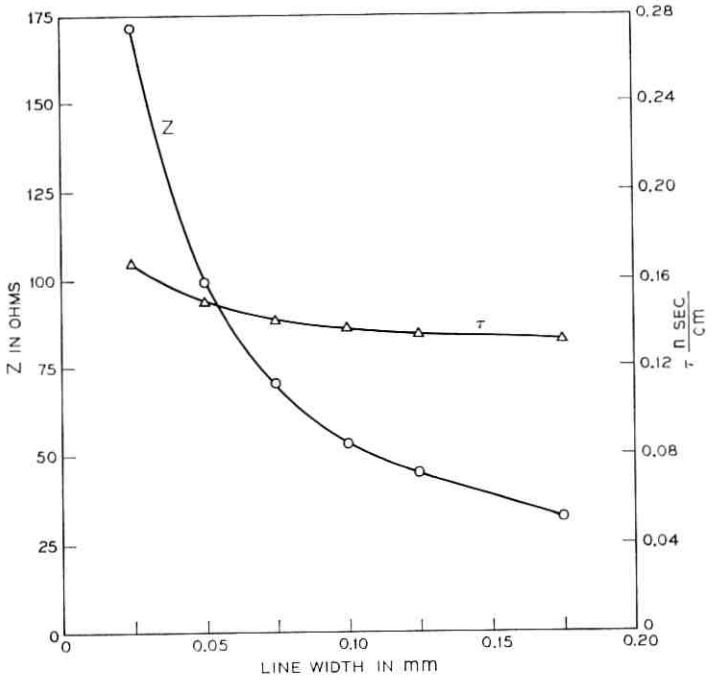


Fig. 16—Impedance and delay time at 1 GHz for 0.0125 mm thick silicon dioxide dielectric on silicon ground plane.

careful design (for example, judicious choices to t_m and μ_R). The problems of delay times and attenuation (characteristics directly proportional to the length of the line) will also be reduced by the much greater chip packing densities allowable with substrates of high thermal conductivity. Nonetheless, the high losses encountered with Si ground planes (along with its lower thermal conductivity) make this material a dubious substrate choice* except for possible special cases in which wide lines could be used for all portions of the circuitry where attenuation need be considered and where the ground returns for high current lines (for example, power supply lines) could be deposited as surface metallization.

* This configuration (0.0025 Ω-cm Si as a ground plane) should not be confused with the possible use of high resistivity Si (1500 Ω-cm) as a dielectric with gold surface metallization and a metal ground plane. T. M. Hyltin has measured the dielectric loss in microstrips with high resistivity Si dielectrics.[16] Such a dielectric, when used in the configuration discussed in this paper, would contribute an additional 0.5 dB/cm attenuation to the attenuation calculated above for aluminum ground planes.

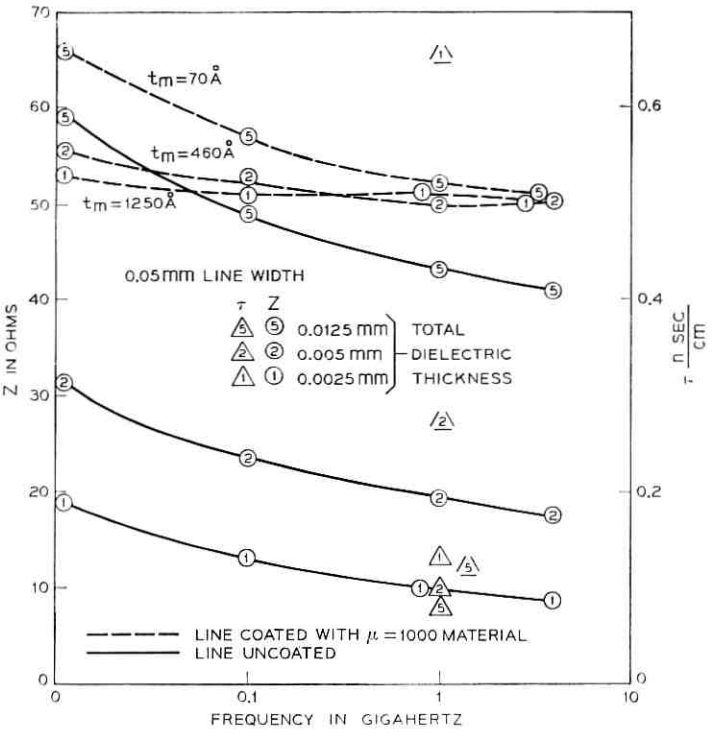


Fig. 17—Impedance and delay time for silicon oxynitride dielectric on aluminum ground plane.

In the case of aluminum ground planes, the losses on narrow (2 mil) lines are probably tolerable for present circuitry with the shortened lines due to higher packing. For longer lines, power leads, and so on, it may be necessary to use wider lines and connect to the beam leads of the integrated circuit chips through tapered section "transformers". The impedance transition need not be as great as the desired width change would indicate if t_m could be increased under the wide lines to compensate for the decreased L and increased C.

Many present applications for lower speed integrated circuit logic do not require matched transmission lines for all or any of the interconnections. The primary electrical requirement is then the completion of interconnects between integrated circuit chips with a minimum capacitance to ground. For such applications the narrow conductors envisioned in this paper would nearly compensate for the increased capacitance per unit length due to reduced t_d compared with conven-

tional 0.625 mm alumina substrates. Use of SiON dielectric ($\epsilon_R = 3.4$) would additionally improve this situation. Resistive signal loss in this application should be negligible since typical input impedances are 1–2K ohms. The possibility of shortened lead length due to higher packing densities on high thermal conductivity substrates then suggests the possibility of improved electrical performance.

IV. CONCLUSIONS

The advantages to be gained from a layered substrate are substantial from thermal considerations. The modifications of the electrical properties resulting from the layered geometry do not appear to present a significant obstacle to the development of such substrates. Indeed many improvements in electrical performance appear realizable with thin line metallization and transmission line interconnections. Layered

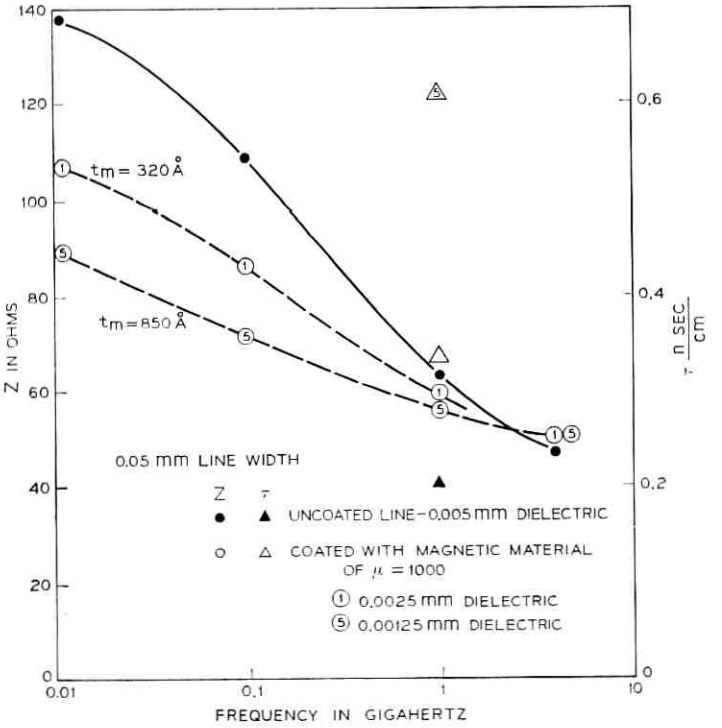


Fig. 18—Impedance and delay time for silicon dioxide dielectric on silicon ground plane.

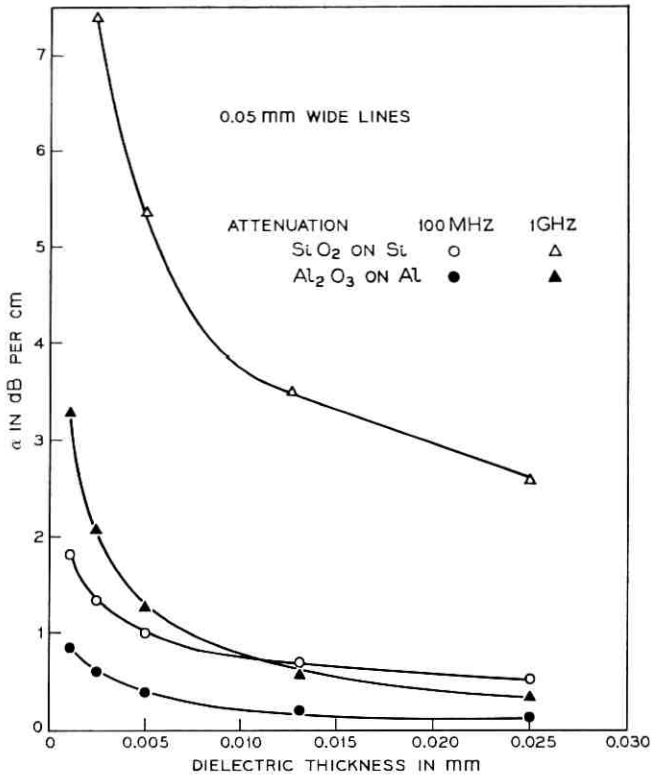


Fig. 19—Attenuation at 100 MHz and 1 GHz.

substrates with massive metal ground planes will adequately satisfy rigid mounting plane requirements, and the physical configuration of unbalanced microstrip lines allows the use of reliable, inexpensive thermocompression bonding of a beam-leaded integrated circuit chip to the substrate.

V. ACKNOWLEDGMENTS

The author would like to thank W. B. Grupen and T. P. Tignor for their comments on preliminary drafts of this paper, and A. W. Rose for his assistance in preparing the computer programs.

APPENDIX A

Averaging Technique for Improved Series Conversion

Theorem: Let $T(N)$ be a function of varying sign with zero crossings enumerated $z_0, z_1, \cdots, z_i, z_{i+1}, \cdots$ where i is chosen such that

$$\left.\frac{dT}{dN}\right|_{N=Z_i} > 0. \tag{32}$$

If

$$\left|\sum_{N=Z_i}^{Z_{i+1}} T(N)\right| > \left|\sum_{N=Z_{i+1}}^{Z_{i+2}} T(N)\right| > \left|\sum_{N=Z_{i+2}}^{Z_{i+3}} T(N)\right|. \tag{33}$$

Then

$$\frac{\sum_{N=0}^{Z_i} T(N) + \sum_{N=0}^{Z_{i+1}} T(N)}{2} < \sum_{N=0}^{\infty} T(N) < \frac{\sum_{N=0}^{Z_{i+1}} T(N) + \sum_{N=0}^{Z_{i+2}} T(N)}{2}. \tag{34}$$

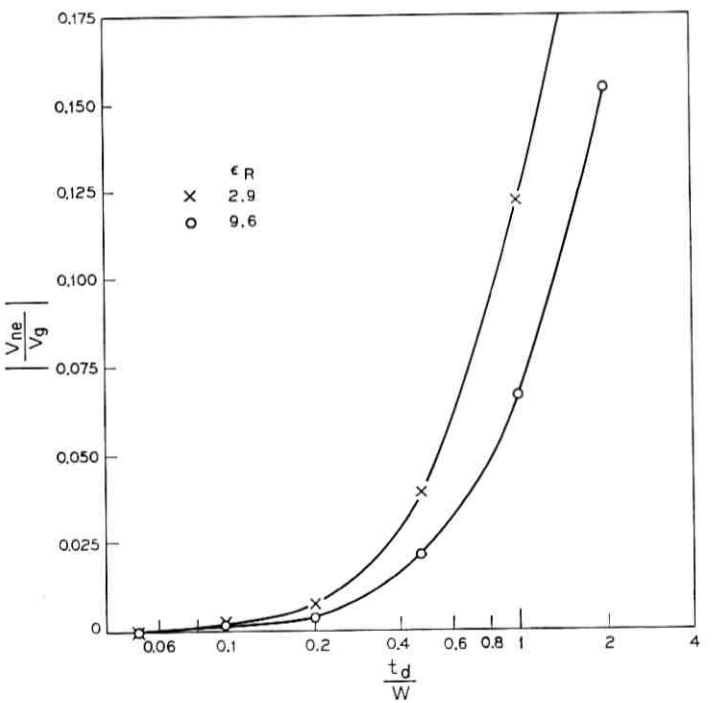


Fig. 20—Near end crosstalk between adjacent, parallel 50Ω microstrip lines ($t_m = 0$.)

TABLE IV—50Ω STRIPLINES ON 0.625 MM ALUMINA SUBSTRATES

Line width	0.625 mm
Delay time	0.05 nsec/cm
Attenuation	0.048 db/cm
Crosstalk	6.75%

Proof: (See Fig. 21)

Write

$$\sum_{N=0}^{Z_i} T(N) = S_i \tag{35}$$

from inequality (33),

$$\sum_{N=i}^{i+1} T(N) > \sum_{N=i+2}^{i+3} T(N) \tag{36}$$

so

$$S_{i+1} - S_i > S_{i+3} - S_{i+2}\,; \tag{37}$$

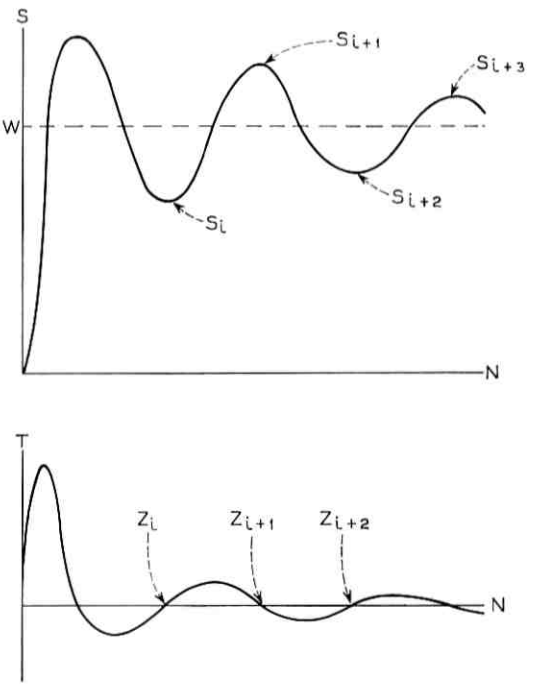


Fig. 21—Enumeration of extrema and zero crossings.

from inequality (32)

$$| S_i - S_{i+1} | > | S_{i+2} - S_{i+3} | \tag{38}$$

so

$$| S_i - S_{i+2} | > | S_{i+1} - S_{i+3} |; \tag{39}$$

similarly

$$| S_{i+1} - S_{i+3} | > | S_{i+2} - S_{i+4} | \tag{40}$$

so

$$| S_i - S_{i+2} | > | S_{i+1} - S_{i+3} | > | S_{i+2} - S_{i+4} |. \tag{41}$$

Write

$$S_\infty = W. \tag{42}$$

Inequality (41) holds for any i chosen in agreement with inequality (32) as $i \rightarrow \infty$. Then

$$| S_i - W | > | S_{i+1} - W | > | S_{i+2} - W | \tag{43}$$

define

$$A_i = \frac{S_i + S_{i+1}}{2} \tag{44}$$

but from inequality (43)

$$| S_i - W | > | S_{i+1} - W |$$

and from inequality (32)

$$S_i < S_{i+1}$$

so

$$W - S_i > S_{i+1} - W, \tag{45}$$

giving

$$S_i + S_{i+1} < 2W \tag{46}$$

so

$$A_i < \frac{2W}{2} = W. \tag{47}$$

From inequality (44)

$$A_{i+1} = \frac{S_{i+1} + S_{i+2}}{2} \tag{48}$$

but from inequality (43)

$$| S_{i+1} - W | > | S_{i+2} - W |$$

and from inequality (32)

$$S_{i+2} < S_{i+1}$$

so

$$S_{i+1} - W > W - S_{i+2}\,, \tag{49}$$

giving

$$S_{i+1} + S_{i+2} > 2W \tag{50}$$

so

$$A_{i+1} > \frac{2W}{2} = W \tag{51}$$

from inequality (47) and (51)

$$A_i < W < A_{i+1} \tag{52}$$

and the theorem is proved.

APPENDIX B

Edge-ground Capacitance for Microstrip

The capacitance from the edges of the finite thickness strip above the infinite ground plane is equivalent to the capacitance problem represented in Fig. (22a). This is equivalent to twice the capacitance represented in Fig. (22b). The conformal transformation $W = \sin^{-1}(z/t_d)$ results in the configuration of Fig. (22c). The capacitance per unit length of this configuration can be calculated by equations (26) and (27) of the text. This results in

$$C \sim \frac{\epsilon_0 \epsilon_R}{\pi} \left[\cosh^{-1} \left(1 + \frac{t_s}{t_d}\right) + \ln \left(1 + \{1 - \exp\left[-\cosh^{-1}\left(1 + t_s/t_d\right)\right]\}^{\frac{1}{2}}\right)\right].$$

For $t_s \gtrsim t_d$,

$$\cosh^{-1}\left(1 + \frac{t_s}{t_d}\right) \simeq \ln\left(2 + \frac{2t_s}{t_d}\right).$$

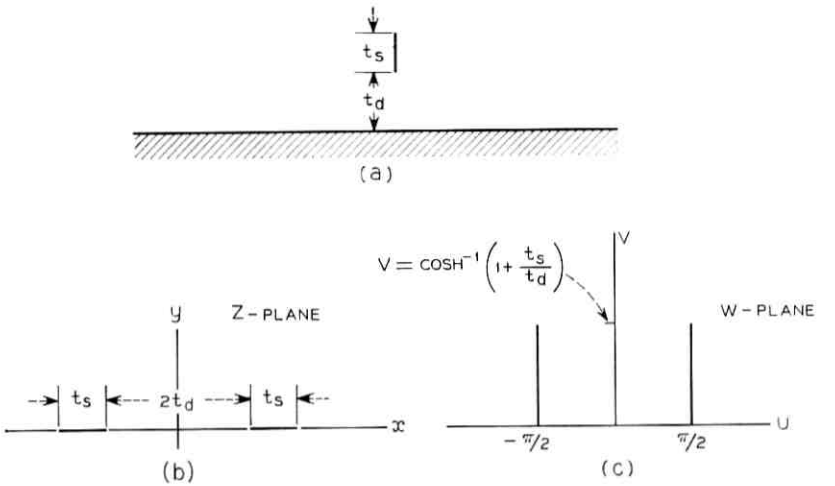


Fig. 22—Conformal mappings.

Then

$$C \simeq \frac{\epsilon_0\epsilon_R}{\pi}\left\{\ln\left(\frac{2t_d + 2t_s}{t_d}\right) + \ln\left[1 + \left(\frac{t_d + 2t_s}{2t_d + 2t_s}\right)^{\frac{1}{2}}\right]\right\},$$

$$= \frac{\epsilon_0\epsilon_R}{\pi}\ln\left\{\frac{2t_d + 2t_s + [2(t_d + t_s)(t_d + 2t_s)]^{\frac{1}{2}}}{t_d}\right\}.$$

The edge-ground capacitance is then

$$C_e = \frac{2\epsilon_0\epsilon_R}{\pi}\ln\left\{\frac{2t_d + 2t_s + [2(t_d + t_s)(t_d + 2t_s)]^{\frac{1}{2}}}{t_d}\right\}.$$

REFERENCES

1. Holt, V. E., unpublished work.
2. Kennedy, D. P., "Heat Conduction in a Homogeneous Solid Circular Cylinder of Isotropic Media," IBM Report TR 00.15072.699, December 4, 1959.
3. Morse, P. M., and Feshbach, H., *Methods of Theoretical Physics*, New York: McGraw-Hill and Company, 1953.
4. Watson, G. N., *A Treatise on the Theory of Bessel Functions*, Cambridge: Cambridge University Press, 1952.
5. Schneider, M. V., "Microstrip Lines for Microwave Integrated Circuits," B.S.T.J., *48*, No. 5 (May–June 1969), pp. 1421–1444.
6. Kordos, R. W., "Analysis of Crosstalk between Microstrip Lines," Record of IEEE Region 3 Convention (April 22–24, 1968), pp. 14.3.1–14.3.3.
7. Assadourian, F., and Rimai, E., "Simplified Theory of Microstrip Transmission Systems," Proc. IRE, *40*, No. 12 (December 1952), p. 1651.
8. Wheeler, H. A., "Transmission-Line Properties of Parallel Wide Strips by a Conformal-Mapping Approximation," IEEE Trans. of Microwave Theory and Techniques, *MTT-12*, No. 3 (May 1964), p. 280.

9. Seckelmann, R., "On Measurements of Microstrip Properties," Microwave Journal, *61*, No. 1 (January 1968), pp. 61–64.
10. Caulton, M., Hughes, J. J., and Sobol, H., "Measurements on the Properties of Microstrip Transmission Lines for Microwave Integrated Circuits," RCA Review, *27*, No. 3 (September 1966), p. 377–391.
11. Moore, R. K., *Traveling-Wave Engineering*, New York: McGraw-Hill, 1960.
12. Ramo, S., and Whinnery, J. R., *Fields and Waves in Modern Radio*, New York: John Wiley, 1956.
13. English, T. D., and McNichol, J. J., "Integrated Transmission Lines for Magnetic Thin Film Memories," IEEE Trans. on Magnetics, *MAG-1*, No. 4 (December 1965), p. 272.
14. Thomson, J. J., *Recent Researches in Electricity and Magnetism*, London: Clarendon Press, 1893.
15. Joines, W. T., unpublished work.
16. Hyltin, T. M., "Microstrip Transmission on Semiconductor Dielectrics," IEEE Trans. on Microwave Theory and Techniques, *MTT-13*, No. 6 (November 1965), p. 777.

The Field Singularity at the Edge of an Electrode on a Semiconductor Surface

By J. A. LEWIS and E. WASSERSTROM*

(Manuscript received February 18, 1970)

Near the edge of a charged electrode on the surface of a semiconductor, the field in the semiconductor may become very large because of the accumulation of charge at the electrode edge. Such large local fields are undesirable, not only because they may cause local breakdown, but also because they make the behavior of a semiconductor device difficult to predict.

In the present paper we consider a simple mathematical model of an electrode edge-semiconductor-insulator configuration and derive conditions under which large local fields may be avoided. More accurately, since the electrode edge and semiconductor corner angles are assumed to be perfectly sharp, we derive conditions under which the local field in the semiconductor is nonsingular. It is necessary to include the effect of the surrounding insulator, because even for small insulator-semiconductor dielectric constant ratios, a field singularity in the insulator will be coupled back into the semiconductor.

I. INTRODUCTION

Beneath a charged electrode situated on the surface of a semiconductor, at points far from the electrode edge, the electrostatic field is regular and quasi-one-dimensional, with its maximum value at the electrode. Near the electrode edge, however, the field may become very large because of the accumulation of surface charge at the sharply curved electrode edge.[1] Also, the jump in dielectric constant between the semiconductor and the surrounding insulating material may produce a large local field intensity. Such a field may be so large as to cause avalanche breakdown near the edge, but, in any case, the presence of such an edge effect makes the behavior of a semiconductor device difficult to predict.

* On leave from the Technion-Israel Institute of Technology, Haifa, Israel, when this work was performed.

In this paper we consider a simple mathematical model of an electrode edge-semiconductor-insulator configuration, namely a sharp-edged electrode on top of a semiconductor mesa, as in Fig. 1. We study the behavior of the potential, or rather its singular part, in the two wedge-shaped semiconductor and insulator regions shown in the inset circle of Fig. 1, assuming that the potential is locally planar and that its singular part satisfies Laplace's equation, both in the insulator and in the semiconductor. Since the treatment is local and the electrode edge and mesa corner are replaced by mathematically sharp wedges, our analysis can only predict the existence or nonexistence of a singular field at the edge and cannot produce an estimate of local field strength, which depends on conditions far from the edge.

We derive an estimate of the order of the singularity in the potential of the form $\varphi = 0(r^p)$, where r is the distance from the corner and $p > 0$. The local field is thus of order r^{p-1}, singular for $p < 1$. We consider p as a function of the semiconductor wedge angle α, the electrode wedge angle β, and the insulator-semiconductor permittivity ratio η, for α and β between zero and 180° and η between zero and one.

We find that, to avoid a field singularity, we must make β greater than 90° and α less than 90°. In particular, if we take $\beta = 180°$, any α less than 90° yields a nonsingular field. Such a configuration might be realized, for example, by using an overhanging electrode on an undercut semiconductor mesa, as in Fig. 2. The length of the overhang must be several Debye lengths for small electrode potential and several depletion layer thicknesses for large reverse bias, in order that the present theory be applicable.

Figure 3 summarizes our principal results. It gives the range of α, for $90° \leqq \beta \leqq 180°$, $0 \leqq \eta \leqq 1$, within which the field is nonsingular.

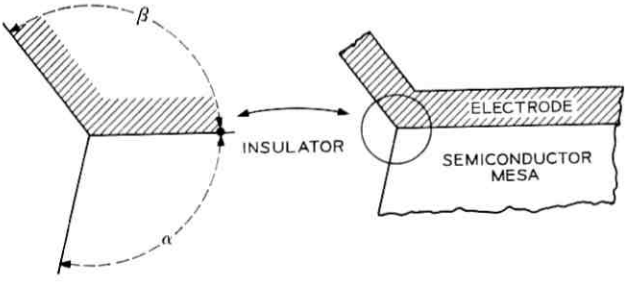


Fig. 1—Mesa with sharp-edged electrode.

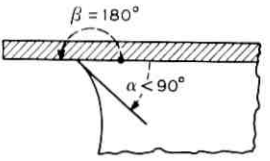


Fig. 2—Overhanging electrode on undercut mesa.

It should be noted that our results apply without modification to edge fields in capacitors and, with the appropriate interpretation, to steady temperature fields in conductors. In the latter case, it would be interesting to study the corresponding thermal stresses, as, for example, in a glass-to-metal seal.

II. A MODEL OF THE ELECTRODE EDGE

We consider the electric field in a current-free semiconductor, near an electrode edge whose cross-section is sketched in Fig. 4. The sketch shows a typical semiconductor mesa with corner angle α, surrounded by insulating material, and supporting an electrode with corner angle β. For given insulator-semiconductor permittivity ratio $\eta = \epsilon_0/\epsilon_1$ (≈ 0.1 for air-silicon, 0.3 for silica-silicon), we wish to choose α and β to avoid local field singularities.

In the semiconductor the dimensionless potential φ satisfies an equation of the general form

$$\nabla^2\varphi = f(\varphi), \tag{1}$$

where the Laplacian operator is made dimensionless with the Debye length for small potential and by the depletion layer thickness for large reverse bias (see, for example, Ref. 2). "Large" or "small" distance then means large or small with respect to one of these typical lengths. In particular, we shall assume that the electrode is so large in a direction perpendicular to the cross-section that the local field may be treated as planar.

We seek solutions of equation (1) which have field singularities at the vertex $r = 0$, that is, a bounded potential φ such that $|\nabla\varphi|$ is unbounded at $r = 0$. In the neighborhood of such a singularity the individual second derivatives which make up the Laplacian will be very large, although they must combine to make the Laplacian equal to the bounded function $f(\varphi)$. As far as the singular part of the solution is concerned then, the specific form of $f(\varphi)$ is unimportant and we can in fact set $f(\varphi)$ equal to zero. The singular solution then satisfies Laplace's

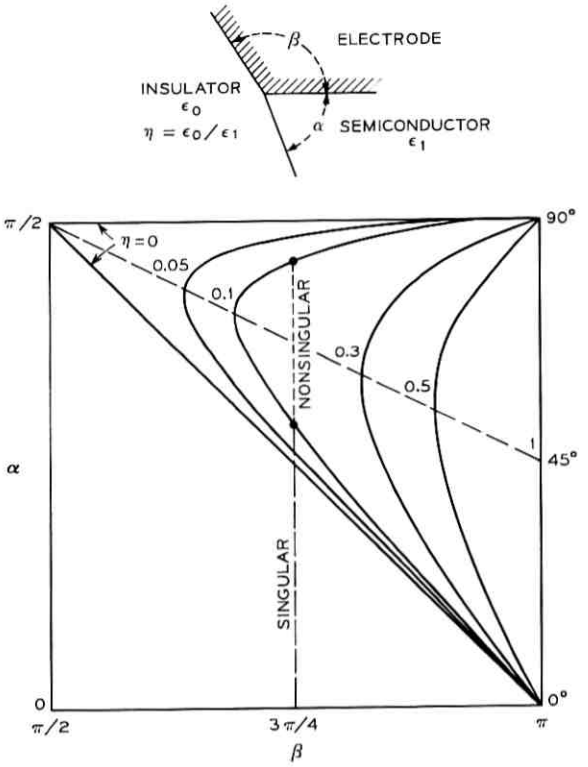


Fig. 3—Bounds on semiconductor angle α as a function of electrode angle β for a nonsingular field.

equation

$$\nabla^2\varphi = \partial^2\varphi/\partial r^2 + \partial\varphi/r\,\partial r + \partial^2\varphi/r^2\,\partial\theta^2 = 0, \tag{2}$$

in the neighborhood of $r = 0$, both in the semiconductor ($0 < \theta < \alpha$) and in the insulator ($\alpha < \theta < 2\pi - \beta$). At the electrode faces we have the boundary conditions

$$\varphi(r, 0) = \varphi(r, 2\pi - \beta) = 1, \tag{3}$$

while at the semiconductor-insulator interface, in the absence of surface charge, we have the continuity conditions*

* As we shall see, no matter how small η is, equation (5) couples any singularity in the insulator back into the semiconductor. Satisfaction of this condition is an essential feature of the problem.

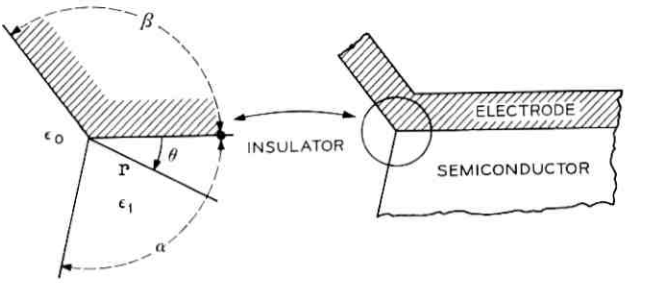


Fig. 4—Semiconductor-electrode edge configuration.

$$\varphi(r, \alpha^-) = \varphi(r, \alpha^+), \tag{4}$$

$$\partial\varphi(r, \alpha^-)/\partial\theta = \eta\partial\varphi(r, \alpha^+)/\partial\theta. \tag{5}$$

We now expand φ in positive powers of r, setting

$$\varphi(r, \theta) = \begin{cases} 1 + \sum_{k=1}^{\infty} A_k r^{p_k} \sin p_k\theta, & \text{for } 0 \leqq \theta \leqq \alpha, \\ 1 + \sum_{k=1}^{\infty} B_k r^{p_k} \sin p_k(2\pi - \beta - \theta), & \text{for } \alpha \leqq \theta \leqq 2\pi - \beta, \end{cases}$$

satisfying equation (2) and boundary condition (3). The A's, B's, and p's are chosen to satisfy the continuity conditions (4) and (5), which take the form

$$A \sin p\alpha - B \sin p(2\pi - \beta - \alpha) = 0,$$

$$A \cos p\alpha + \eta B \cos p(2\pi - \beta - \alpha) = 0.$$

These equations have a nontrivial solution only if the coefficient determinant vanishes, giving the characteristic equation for p

$$\eta \sin p\alpha \cos p(2\pi - \beta - \alpha) + \cos p\alpha \sin p(2\pi - \beta - \alpha) = 0. \tag{6}$$

We wish to find the smallest positive value of p which will satisfy this equation, as a function of α, β, and η. For this value of p we have

$$\varphi - 1 = 0(r^p),$$

$$| \nabla\varphi | = 0(r^{p-1}),$$

singular for $p < 1$, in the neighborhood of $r = 0$.

III. THE FLAT SURFACE AND THE MESA

Before treating the general problem, let us consider two special cases, in order to gain insight into the behavior of p as a function of α, β, and η. In the very common case of a thin electrode ($\beta = 0$) on a flat semiconductor surface ($\alpha = \pi$), as in Fig. 5, the characteristic equation (6) reduces to

$$\sin p\pi \cos p\pi = 0, \tag{7}$$

for all η. It is satisfied by $p = 1$, $p = \frac{1}{2}$, the latter being the smallest value of p. In this case, near the electrode edge the field is singular, like $r^{-\frac{1}{2}}$, no matter what insulating material is used. This is the same singularity as that obtained in the classical Weber problem of the disk electrode. The same local behavior was also found in Ref. 3, where a closed-form solution of the above problem was derived for the linearized semiconductor equation and $\eta = 0$, $\eta = 1$.

Now let us attempt to reduce the singularity by cutting away the semiconductor, placing the electrode on top of a mesa, as shown in Fig. 6 for $\alpha = \pi/2$. For $\eta = 0$ this configuration gives a one-dimensional, regular field, normal to the electrode, so that we might expect that it would be advantageous for small η. With $\beta = 0$, $\alpha = \pi/2$, equation (6) becomes

$$\eta \sin \frac{p\pi}{2} \cos \frac{3p\pi}{2} + \cos \frac{p\pi}{2} \sin \frac{3p\pi}{2} = 0, \tag{8}$$

satisfied by $p = 1$, for all η. However, note that, for $\eta = 0$, it also has the smaller root $p = \frac{2}{3}$. If we examine the equations for A and B, we find that, in this case, $A = 0$, so that the root $p = \frac{2}{3}$ gives a singularity only in the insulator for $\eta = 0$. However, a perturbation for small positive η shows that A is of order η, so that the singularity is coupled back into the semiconductor for any positive η, no matter how small.

On the other hand, for $\eta = 1$, the smallest root of equation (8) is $p = \frac{1}{2}$. In this case p is independent of α, for $\eta = 1$ corresponds to a single dielectric filling the whole space around the electrode. Now, if

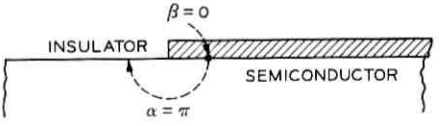


Fig. 5—Flat semiconductor surface.

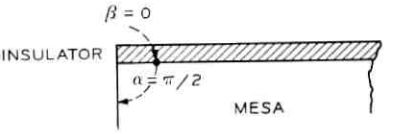


Fig. 6—Right-angled mesa corner.

p is a continuous function of η, its value for $0 < \eta < 1$ must lie between the value for $\eta = 0$ and the value for $\eta = 1$; that is, we must have

$$\tfrac{1}{2} < p < \tfrac{2}{3},$$

for $\beta = 0$, $\alpha = \pi/2$, and $0 < \eta < 1$. For small η, the singularity is weakened in some sense, but not removed, by the formation of a mesa.

These two special cases indicate that the calculation of $p = p(\alpha, \beta, \eta)$ is not completely straightforward. They also suggest that the simple cases $\eta = 0$, $\eta = 1$ can be used as a framework for the general calculation.

IV. LIMITING VALUES OF PERMITTIVITY RATIO

Let us first consider the case $\eta = 1$. In this case equation (6) becomes

$$\sin p(2\pi - \beta) = 0,$$

independent of α, as one would expect. Its smallest positive root is

$$p = p(\alpha, \beta, 1) = p_1(\beta) = \pi/(2\pi - \beta), \tag{9}$$

singular for $\beta < \pi$. Those familiar with potential theory will recognize this as the singularity at the tip of a wedge-shaped electrode, protruding into a uniform dielectric. It has been studied in detail by Wasow, Lehman, and Joyce.[4-6]

The other limiting case $\eta = 0$ gives two roots. Equation (6) becomes

$$\cos p\alpha \sin p(2\pi - \beta - \alpha) = 0,$$

with the roots

$$p = p(\alpha, \beta, 0) = p_0^-(\alpha) = \pi/2\alpha, \tag{10}$$

$$p = p(\alpha, \beta, 0) = p_0^+(\alpha, \beta) = \pi/(2\pi - \beta - \alpha). \tag{11}$$

The first of these roots gives a singular field in the semiconductor for $\alpha > \pi/2$; the second gives a singular field for $\alpha + \beta < \pi$ (in the insulator only, for $\eta = 0$) but weakly coupled back into the semiconductor for small positive η.

For $0 < \eta < 1$, p must lie between p_1 and the smaller of the two values p_0^-, p_0^+. Now $p_0^- > p_0^+$ for $\alpha < (2\pi - \beta)/3$, so that p lies between p_1 and p_0^+, for $0 \leqq \alpha \leqq (2\pi - \beta)/3$, and between p_1 and p_0^-, for $(2\pi - \beta)/3 \leqq \alpha \leqq \pi$. Also $p_0^- < p_1$, for $\alpha > (2\pi - \beta)/2$, so that we finally obtain the series of bounds listed below:

$$\frac{\pi}{2\pi - \beta} < p < \frac{\pi}{2\pi - \beta - \alpha}, \quad \text{for} \quad 0 < \alpha < \frac{2\pi - \beta}{3}, \tag{12}$$

$$\frac{\pi}{2\pi - \beta} < p < \frac{\pi}{2\alpha}, \quad \text{for} \quad \frac{2\pi - \beta}{3} < \alpha < \frac{2\pi - \beta}{2}, \tag{13}$$

$$\frac{\pi}{2\alpha} < p < \frac{\pi}{2\pi - \beta}, \quad \text{for} \quad \frac{2\pi - \beta}{2} < \alpha < \pi. \tag{14}$$

Now the largest value of the upper bound is attained at the point where the upper bound of equation (12), an ascending hyperbola as a function of α, meets the upper bound of equation (13), a descending hyperbola, that is at the point $\alpha = (2\pi - \beta)/3$, where $p = 3\pi/2(2\pi - \beta) = p_{max}$. For $\beta < \pi/2$ this maximum is less than unity, so that a field singularity can be avoided only by choosing the electrode angle β greater than 90°. Similarly, the upper bound of equation (13) implies that the semiconductor angle α must be chosen less than 90° to avoid a field singularity.

A particularly simple way of satisfying these requirements is the combination of overhanging electrode ($\beta = \pi$) and slightly undercut semiconductor mesa ($\alpha \leqq \pi/2$) shown in Fig. 7. In order that the theory be applicable, the length of the overhang must be several characteristic lengths, i.e. several Debye lengths for small electrode potential and several depletion layer thicknesses for large electrode potential. The mesa corner needs to be undercut only enough to be certain that α is

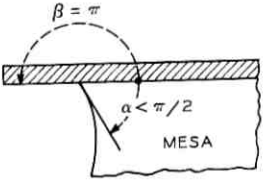


Fig. 7—Overhanging electrode on undercut mesa.

never greater than 90°, within fabrication tolerances. In Section V we document these preliminaries.

V. ARBITRARY PERMITTIVITY RATIO

For detailed calculations it is convenient to rewrite the characteristic equation (6) in the form

$$(1 + \eta) \sin p(2\pi - \beta) + (1 - \eta) \sin p(2\pi - \beta - 2\alpha) = 0 \qquad (15)$$

from which it is a simple matter to calculate the derivative

$$\frac{\partial p}{\partial \alpha} = \frac{2(1-\eta)p \cos p(2\pi-\beta-2\alpha)}{(1+\eta)(2\pi-\beta) \cos p(2\pi-\beta)+(1-\eta)(2\pi-\beta-2\alpha) \cos p(2\pi-\beta-2\alpha)}. \qquad (16)$$

Now, whereas it is difficult to solve equation (6) directly for p as a function of α, β, η, because of the existence of neighboring higher roots, it is a simple matter to integrate equation (16), starting from $\alpha = 0$, where $p = \pi/(2\pi - \beta)$ for all η. This calculation was carried out for $\eta = 0, 0.1, 0.3, 0.5, 0.7, 1.0$ and for $\beta = 0, \pi/2, 3\pi/4, \pi$. The results are shown in Figs. 8 through 11, for the two lowest branches of p.* As Section IV predicts, a nonsingular field ($p \geqq 1$) becomes possible only for $\beta \geqq \pi/2$, with the range of permissible values of α and η, increasing from $\alpha = \pi/2$, $\eta = 0$, at $\beta = \pi/2$, to $0 < \alpha < \pi/2$, $0 < \eta < 1$, at $\beta = \pi$.

This range of values of α for given β and η is easily found. If we set $p = 1$ in equation (15), we find that it is satisfied by two values of α, which bound the permissible range for a nonsingular field. Figure 3 shows the result of this elementary calculation. For example, for $\beta = 3\pi/4$, $\eta = 0.1$, the field is singular for $\alpha < 0.291\pi = 52.5°$, regular for $0.291\pi < \alpha < 0.459\pi = 82.5°$, and singular again for larger α. The slanting dashed line gives the minimum value of β, and the corresponding value of α, for which the field is nonsingular for given η.

VI. ACKNOWLEDGMENT

This problem was suggested originally by S. Sze. The authors profited from several illuminating discussions with J. McKenna. The calculations were carried out by Miss Judith B. Seery.

* Two branches are shown to indicate the topology, although of course only the lower branch is of interest.

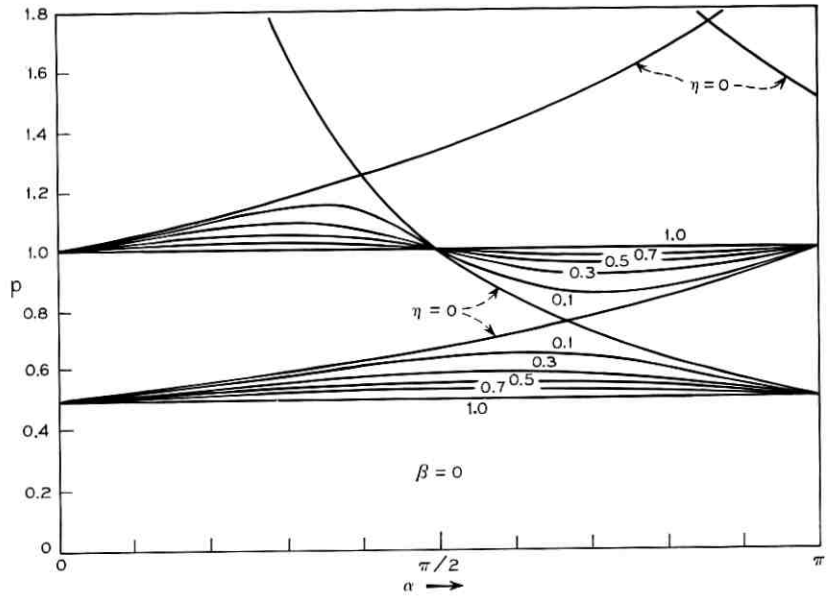


Fig. 8—$p(\alpha)$ for $\beta = 0$ and various η.

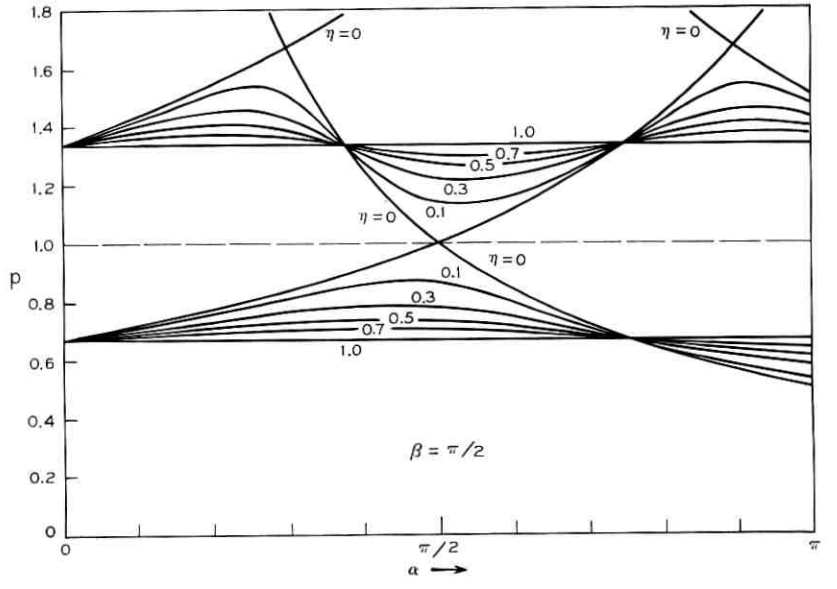


Fig. 9—$p(\alpha)$ for $\beta = \pi/2$ and various η.

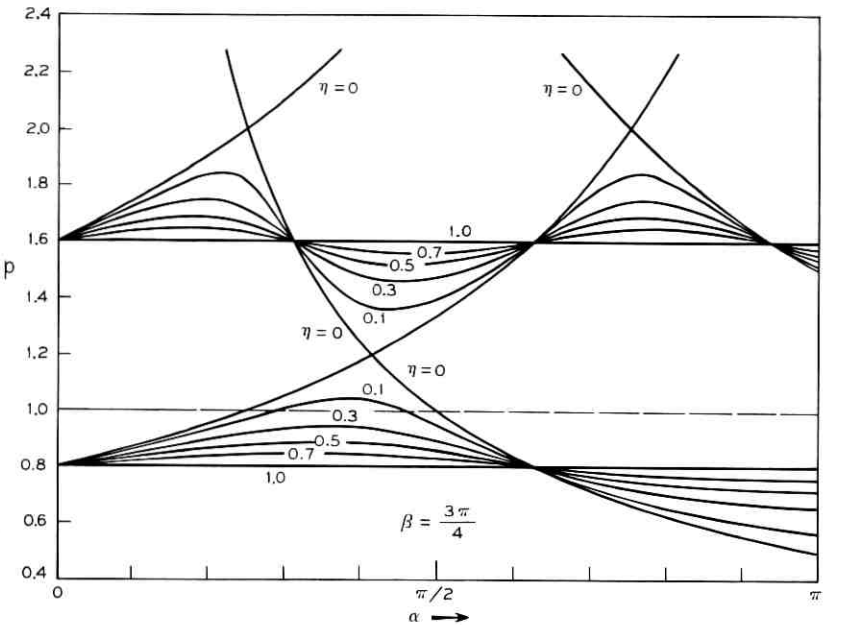


Fig. 10—$p(\alpha)$ for $\beta = 3\pi/4$ and various η.

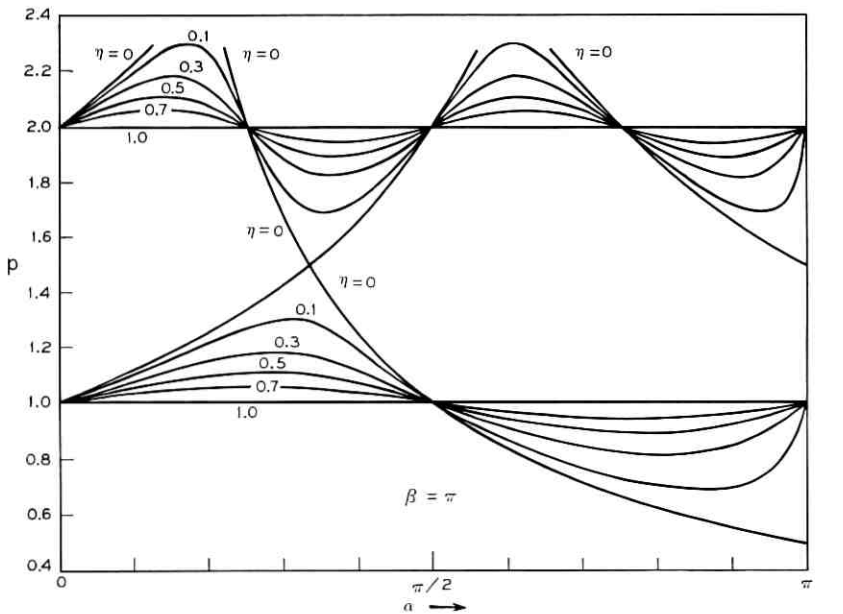


Fig. 11—$p(\alpha)$ for $\beta = \pi$ and various η.

REFERENCES

1. McKenna, J., and Wasserstrom, E., "The Potential Due to a Charged Metallic Strip on a Semiconductor Surface," B.S.T.J., *49*, No. 5 (May-June 1970), pp. 853–877.
2. Lewis, J. A., McKenna, J., and Wasserstrom, E., unpublished work.
3. Lewis, J. A., unpublished work.
4. Wasow, W. R., "Asymptotic Development of the Solution of Dirichlet's Problem at Analytic Corners," Duke Math. J., *24*, No. 1 (March 1957), pp. 47–56.
5. Lehman, S. R., "Developments at an Analytic Corner of Solutions of Elliptic Partial Differential Equations," J. Math. Mech., *8*, No. 5 (September 1959), pp. 729–760.
6. Joyce, W. B., unpublished work.

Determination of the Shape of the Human Vocal Tract from Acoustical Measurements

By B. GOPINATH and M. M. SONDHI

(Manuscript received January 14, 1970)

In this paper we describe methods for determining the cross-sectional area function of the human vocal tract from acoustical measurements made at one end. The pressure and volume velocity are assumed to obey Webster's horn equation, which is valid for frequencies below 3.5 kHz. Acoustical properties below 3.5 kHz do not uniquely specify the area function. This paper shows how high frequency information may be incorporated into the mathematical model in a manner consistent with a priori information about the vocal tract. Some results of application of the methods by computer simulation are presented. It is interesting to see from the figures that nine numbers (namely, length, four formants, and four residues) specify the area function quite well for practical purposes.

I. INTRODUCTION

In recent years there has been considerable interest in the modelling of speech production in terms of the motion of the articulators. This interest has stimulated work on the determination of the shape of the human vocal tract as a function of the utterance. For frequencies less than 3500 Hz, wave motion in the vocal tract is essentially planar, so that the shape is effectively specified by the cross-sectional area as a function of distance from one end of the tract (say from the glottis).

During the past two decades X-ray techniques have been used to determine these area functions. These techniques suffer from two major drawbacks: (*i*) In order to keep the exposure to X-rays within safe dosage limits, only a small number of measurements can be made on any one subject; (*ii*) The interpretation of X-ray data is a complex and difficult art, and a number of assumptions must be made in order to convert this data to area functions. The accuracy with which area functions are reconstructed is rather limited.

In 1965, Mermelstein and Schroeder suggested the new approach of inferring the area functions from acoustic information.[1] Under the usual assumptions of lossless plane wave propagation, they showed that if the area function $A(x)$ of a vocal tract of length l is of the form

$$\log A(x) = \log A_0 + \sum_{n=1}^{\infty} a_n \cos \frac{n\pi x}{l} \tag{1}$$

then in the limit as $a_n \to 0$ for all n, the nth eigenfrequency (with the tract closed at $x = 0$ and open at $x = 1$), is given by

$$\lambda_n = \lambda_{0n}(1 - \tfrac{1}{2}a_{2n-1}) \tag{2}$$

where λ_{0n} is the nth eigenfrequency of the uniform tract ($a_n \equiv 0$, $n = 1, \cdots$). Likewise, to the same approximation the nth eigenfrequency with the tract closed at both ends, is given by

$$\mu_n = \mu_{0n}(1 - \tfrac{1}{2}a_{2n}). \tag{3}$$

Using equation (2) to obtain a_{2n+1} from ω_n, Mermelstein and Schroeder obtained antisymmetric approximations to area functions from a knowledge of formant frequencies alone. This work was extended by Schroeder[2] and Mermelstein[3] to include the even-order coefficients [using measured values of the poles and zeroes of the input admittance at the lips, which correspond respectively to the λ's and μ's in equations (2) and (3)]. An extension was also made by Mermelstein[3] who devised an iterative algorithm to compute the first $2m$ coefficients in equation (1) from a knowledge of $(\lambda_1 \cdots \lambda_m, \mu_1 \cdots \mu_m)$, when the perturbations of the eigenvalues are too large for first order perturbation theory to be accurate. Another iterative scheme has been obtained by J. Heinz, by applying perturbation theory to tracts of arbitrary shape.[4]

These methods are applications of very general techniques (namely, perturbation theory and steepest descents) which do not make use of the special characteristics of the problem at hand. They also leave unanswered certain mathematical questions such as the convergence of the iterative procedures and uniqueness of the solution.

In this paper, we describe two (noniterative) methods for computing the area function from acoustical data. Apart from clarifying the physical and mathematical aspects of the problem, these methods provide solutions in a form suitable for analyzing the sensitivity of the reconstructed area functions to inaccuracies of the data. They also enable us to answer such basic questions as: "What tube has all but a finite number of eigenvalues identical to those of a uniform tube?"

In Section II we introduce the wave equation and Webster's horn equation and list the basic properties of the solutions and eigenfunctions of the horn equation under homogeneous boundary conditions.

In Section III we present a method for computing the area function, based upon the factorization of the kernel of an integral operator which transforms solutions of the horn equation to the solutions of the equation for a uniform tract. The existence of such a transform was proved by Marchenko,[5] and the transform has been used in the solution of the inverse-Sturm Liouville problem by Gelfand and Levitan.[6]

In Section IV we present an alternative method for computing the area function based upon the solution of an integral equation whose kernel is the driving point response to an impulse at one end of the tract. This integral equation was introduced without derivation by Krein in a paper on an application of his theory of extensions of positive definite kernels.[7,8] Our derivation is physically motivated and uses only elementary theory of forced motion of a second order system.

In Section V we present preliminary results of an application of our methods to the determination of vocal tract shapes and a comparison with X-ray derived data. Figures 3 through 9 show the results of these computations.

II. MATHEMATICAL PRELIMINARIES

For a tube of variable cross-sectional area $A(x)$ the equations relating acoustical pressure p and volume velocity V are

$$\frac{\partial p}{\partial x} = -\frac{\rho}{A(x)} \frac{\partial V}{\partial t}, \tag{4a}$$

$$\frac{\partial V}{\partial x} = -\frac{A(x)}{\rho c^2} \frac{\partial p}{\partial t}, \tag{4b}$$

under the assumption of lossless plane wave propagation in the tube. These assumptions are accurate for the vocal tract for frequencies up to about 4 kHz. For convenience we will choose units such that the velocity of sound $c = 1$, the density of air $\rho = 1$ and the length of the tube is π. Then elimination of V in equations (4a) and (4b) gives

$$\frac{\partial}{\partial x} A(x) \frac{\partial p}{\partial x} = A(x) \frac{\partial^2 p}{\partial t^2} \qquad 0 \leqq x \leqq \pi; \tag{5}$$

and for sinusoidal time dependence, such that $p = \phi(x, \lambda)e^{i\lambda t}$, the function $\phi(x, \lambda)$ satisfies

$$\frac{\partial}{\partial x} A(x) \frac{\partial \phi(x, \lambda)}{\partial x} + \lambda^2 A(x)\phi(x, \lambda) = 0 \qquad 0 \leqq x \leqq \pi \tag{6}$$

which is Webster's horn equation. Throughout this paper it will be assumed that $A(x) > 0$ $(0 \leqq x \leqq \pi)$, $A(0) = 1$, and that $A(x)$ has continuous first and second derivatives except at a finite number of points in $[0, \pi]$. At the points of discontinuity $A(x)$ and its first two derivatives are assumed to have finite right and left limits. Under these conditions on $A(x)$ the following lemma holds.

Lemma 1: The solution of equation (6) *satisfying the initial conditions*

$$\phi(0, \lambda) = 1, \qquad \phi'(0, \lambda) = 0 \tag{7}$$

exists, and

$$| [A(x)]^{\frac{1}{4}}\phi(x, \lambda) - [A_0(x)]^{\frac{1}{4}}\psi(x, \lambda) | \leqq \frac{K}{\lambda} \tag{8}$$

where

$$0 \leqq K < \infty, \qquad 0 \leqq x \leqq \pi,$$

and $\psi(x, \lambda)$ is the solution of equation (6) *with the same initial conditions and $A(x)$ replaced by a canonical shape $A_0(x)$. The function $A_0(x)$ is such that $A_0(0) = A(0) = 1$, $A_0(x)$ is constant everywhere in $[0, \pi]$ except at points of discontinuity of $A(x)$ where it jumps by the same factor as does $A(x)$.*

The proof of this lemma is given in Appendix A.

The solutions of equation (6) satisfying the initial conditions (7) become eigenfunctions if they satisfy some homogeneous boundary condition at $x = \pi$. These eigenfunctions and eigenvalues have well known properties which for the specific case $\phi(\pi, \lambda) = 0$ we summarize in the following lemma.

Lemma 2: If $A(x)$ satisfies the conditions described above, then there exists a sequence λ_i (the eigenvalues) satisfying

(i) $\lambda_1 > 0$, $\lambda_i \to \infty$ as $i \to \infty$;

(ii) $\phi(x, \lambda_i)$ are solutions of equation (6) *satisfying the initial conditions* (7) *and the condition* $\phi(\pi, \lambda_i) = 0$; (8a)

(iii)
$$\int_0^\pi A(x)\phi(x, \lambda_i)\phi(x, \lambda_j)\, dx = 0, \qquad i \neq j,$$
$$= \alpha_i^2, \qquad i = j, \tag{9}$$

with

$$0 < \alpha_i^2 < \infty;$$

(iv) $\phi(x, \lambda_i)$ *are complete in the space* $L_2(0, \pi)$ *of square integrable functions.*

An immediate consequence of Lemmas 1 and 2 is the following corollary.

Corollary 2.1: If μ_i *is the sequence of eigenvalues for the canonical tube* $A_0(x)$, *with the conditions at* $x = 0$ *and* $x = \pi$ *as in Lemmas* 1 *and* 2, *then*

$$\lambda_i = \mu_i + 0\left(\frac{1}{i}\right) \tag{10}$$

and the α_i^2 *of Lemma* 2 *satisfy*

$$\begin{aligned} \alpha_i^2 &= \int_0^\pi A_0(x)\psi^2(x, \mu_i)\, dx + 0\left(\frac{1}{i^2}\right) \\ &= \gamma_i^2 + 0\left(\frac{1}{i^2}\right) \end{aligned} \tag{11}$$

where ψ *is as defined in Lemma* 1.

Finally we will require the following lemma.

Lemma 3: There exists a function $H(x, t)$ *such that*

$$[A_0(x)]^{\frac{1}{4}}\psi(x, \lambda) = [A(x)]^{\frac{1}{4}}\phi(x, \lambda) + \int_0^x H(x, t)[A(t)]^{\frac{1}{4}}\phi(t, \lambda)\, dt. \tag{12}$$

This can be proved by substituting equation (12) into equation (6). After trivial, but involved, algebra (see Appendix B) it turns out that for (12) to be true, $H(x, t)$ must satisfy the following:

$$\frac{\partial^2 H(x, t)}{\partial x^2} - \frac{\partial^2 H(x, t)}{\partial t^2} + \frac{\{[A(t)]^{\frac{1}{4}}\}''}{[A(t)]^{\frac{1}{4}}} H(x, t) = 0 \tag{13}$$

$$H(x, x) = -\frac{1}{2}\int_0^x \frac{\{[A(t)]^{\frac{1}{4}}\}''}{[A(t)]^{\frac{1}{4}}}\, dt - \{[A(x)]^{\frac{1}{4}}\}' \,|_{x=0} \tag{14}$$

$$\{[A(t)]^{\frac{1}{4}}\}' H(x, t) \,|_{t=0} - \left.\frac{\partial H(x, t)}{\partial t}\right|_{t=0} = 0. \tag{15}$$

The theory of partial differential equations guarantees the existence of a solution to equation (13) under the boundary conditions (14) and (15).

III. DERIVATION OF $A(x)$ FROM THE SPECTRAL FUNCTION OR TWO SETS OF EIGENVALUES

The spectral function is defined as a staircase function of λ with jumps of α_i^2 at λ_i $(i = 1, \cdots)$. Thus to say that the spectral function is known is equivalent to saying that the pairs (λ_i, α_i^2) $i = 1, \cdots$ are known. Appendix C shows that if λ_i $(i = 1, \cdots)$ are the eigenvalues of the same tube for the conditions $\phi'(0, \lambda) = 0, \phi(0, \lambda) = 1, a\phi(\pi, \lambda) + b\phi'(\pi, \lambda) = 0$, $(b \neq 0)$, then a knowledge of the pairs (λ_i, λ_i'), $i = 1, \cdots$ specifies the spectral function. Also, in Section IV it will turn out [see equation (27), with $x = 0$] that λ_i is the ith pole of the driving point impedance, and $1/2\alpha_i^2$ the corresponding residue.

We now derive $A(x)$, given the spectral function. In cases where $A(x)$ has continuous first and second derivatives the spectral function suffices to uniquely determine $A(x)$. In cases where $A(x)$ has a finite number of discontinuities, the locations and magnitudes of the jumps are also assumed to be known. Note that equation (12) may be written in symbolic form as

$$[A_0(x)]^{\frac{1}{2}}\psi(x, \lambda) = (I + H)\phi(x, \lambda)[A(x)]^{\frac{1}{2}} \tag{16}$$

where I is the identity on $(L^2[0, \pi])$ and H the integral operator such that

$$g = Hf \Leftrightarrow g(x) = \int_0^x H(x, t)f(t)\, dt. \tag{17}$$

Define the operator U which takes a square summable sequence of real numbers f_i, $i = 1, 2, \cdots$ to a square integrable function $f(x)$ [on $(0, \pi)$] defined as

$$f(x) = \sum_{i=1}^{\infty} f_i[A(x)]^{\frac{1}{2}}\phi(x, \lambda_i)/\alpha_i^2 . \tag{18}$$

Define the adjoint operator U^* which takes a square integrable function $f(x)$ to a square summable sequence f_i given by

$$f_i = \frac{1}{\alpha_i^2}\int_0^\pi f(x)\phi(x, \lambda_i)[A(x)]^{\frac{1}{2}}\, dx \qquad i = 1, \cdots . \tag{19}$$

Let R and R^* be defined analogously to U and U^*, with $[A(x)]^{\frac{1}{2}}\phi(x, \lambda_i)$ replaced by $\psi(x, \lambda_i)(A_0)^{\frac{1}{2}}$ in equations (18) and (19).

Then

$$R = (I + H)U$$

and

$$RR^* = (I + H)UU^*(I + H)^*. \tag{20}$$

However, from the completeness and orthogonality of the $[A(x)]^{\frac{1}{4}}\phi(x, \lambda_i)$, it follows that $UU^* = I$. Thus

$$RR^* = (I + H)(I + H)^*. \tag{21}$$

Note that

$$[(RR^* - I)f](x) = \int_0^\pi f(t) \sum_{i=1}^{\infty} \left(\frac{\psi(x, \lambda_i)\psi(t, \lambda_i)}{\alpha_i^2} - \frac{\psi(x, \mu_i)\psi(t, \mu_i)}{\gamma_i^2} \right) \cdot [A_0(t)A_0(x)]^{\frac{1}{4}}\, dt \tag{22}$$

(because $[A_0(x)]^{\frac{1}{4}}\psi(x, \mu_i)/\gamma_i$ is an orthonormal and complete sequence).

From the asymptotic formulae of Lemma 1, it is seen that the kernel of the operator $RR^* - I$ is of the Hilbert-Schmidt type [that is, square integrable on the square $(0 \leqq x, t \leqq \pi)$.] Therefore, if the λ_i, α_i correspond to those of a tube with appropriate boundary conditions then the factorization of equation (21) is always possible.

This essentially completes our derivation. For the kernel of RR^* can be constructed if the λ_i's and α_i's are known [and in the case of discontinuous $A(x)$, the positions and magnitudes of the jumps are also known]. The factorization (21) then gives $H(x, t)$. Finally, since $\phi(x, 0) = 1$ and $\psi(x, 0) = 1$ $(0 \leqq x \leqq \pi)$, equation (12) gives

$$[A(x)]^{\frac{1}{4}} + \int_0^x H(x, t)[A(t)]^{\frac{1}{4}}\, dt = [A_0(x)]^{\frac{1}{4}} \tag{23}$$

which can be solved for $[A(x)]^{\frac{1}{4}}$.

Although, in general, the factorization of equation (18) is difficult, we will show in Section V an effective method of computation when all but a finite number of λ_i's and α_i's are identical to the corresponding μ_i's and γ_i's.

IV. DERIVATION OF THE AREA FUNCTION FROM THE INPUT IMPEDANCE

In this section, for simplicity, the area function and its first two derivatives will be assumed continuous. Consider the forced pressure response $y(x, t)$ in the tube, due to a unit ramp $r(t)$ of volume velocity at $x = 0$. This may be obtained by including a term $\delta(x)r(t)$ on the right hand side of equation (4b). The resulting equation for $y(x, t)$ is

$$\frac{\partial}{\partial x} A(x) \frac{\partial y(x, t)}{\partial x} - A(x) \frac{\partial^2 y(x, t)}{\partial t^2} = -\delta(x)u(t) \tag{24}$$

where $u(t)$ is the unit step. Integrating equation (24) over x from 0 to a gives

$$A(a)\left.\frac{\partial y}{\partial x}\right|_{x=a} - A(0)\left.\frac{\partial y}{\partial x}\right|_{x=0} - \int_0^a A(x)Z(x, t)\,dx = -1 \quad t \geqq 0. \tag{25}$$

In this equation we have put $u(t)\ \partial^2 y/\partial t^2 = Z(x, t)$; it is the transfer impedance (in the time domain) from the input end to the point x. However $(\partial y/\partial x) = 0$ at $x = 0$ because of the boundary condition. Also, since the velocity of sound has been normalized to unity, for $t \leqq a$ the region of the tube beyond $x = a$ is undisturbed. Thus for $t \leqq a$, $A(x)\ \partial y/\partial x = 0$ for $x \geqq a$. Thus equation (25) becomes

$$\int_0^a A(x)Z(x, t)\,dx = 1 \qquad 0 \leqq t \leqq a. \tag{26}$$

By expansion in terms of $\phi(x, \lambda_i)$ it can be verified that

$$Z(x, t) = \frac{\partial^2 y(x, t)}{\partial t^2} = \sum_{i=1}^{\infty} \frac{\phi(x, \lambda_i)\cos \lambda_i t}{\alpha_i^2} \qquad t \geqq 0 \tag{27}$$

where the convergence is assumed to be in the sense of distributions, and ϕ_i , α_i , λ_i are as defined in Section II.

Let $f(t)$ be a function such that

$$\int_0^a f(t)Z(x, t)\,dt = 1 \qquad x \leqq a. \tag{28}$$

Then by substitution into equation (26) it follows that

$$\int_0^a f(t)\,dt = \int_0^a A(x)\,dx. \tag{29}$$

The interesting duality in equations (26), (28), and (29) enables determination of $A(x)$ in terms of $Z(0, t)$ rather than $Z(x, t)$. Multiplying equation (28) by $A(x)Z(x, s)$, integrating over x and changing order of integration on the left-hand side we get

$$\int_0^a f(t)\,dt \int_0^a A(x)Z(x, t)Z(x, s)\,dx = \int_0^a A(x)Z(x, s)\,dx \quad t < a. \tag{30}$$

For $s < a$, the right-hand side equals unity by virtue of equation (27). On the left-hand side the integration limits on x can be changed to $(0, \pi)$ since $Z(x, t) = 0$ for $x > t$. Then substituting for $Z(x, t)$, $Z(x, s)$ from equation (27) and using the orthogonality equation (9) we get

$$\int_0^a f(t) \sum \frac{\cos \lambda_i s \cos \lambda_i t}{\alpha_i^2}\,dt = 1, \qquad s \leqq a. \tag{31}$$

[Equation (31) may also be obtained by multiplying equation (28) by $[A(x)]^{\frac{1}{2}}$ and using the linear transformation of Section III.] Defining $\hat{f}(t) = f(|t|)$, $|t| \leqq a$ we note that $\hat{f}(t)$ satisfies

$$\int_{-a}^{a} \hat{f}(t) \sum \frac{\cos \lambda_i s \cos \lambda_i t + \sin \lambda_i s \sin \lambda_i t}{\alpha_i^2} dt = 1 \qquad |s| \leqq a \quad (32)$$

since $\sin \lambda_i t$ is odd. From an elementary trigonometric identity it then follows that

$$\int_{-a}^{a} \hat{f}(t) Z(0, |t - s|) \, dt = 1 \qquad s \leqq a \quad (33)$$

and, from equation (29)

$$\int_{0}^{a} \hat{f}(t) \, dt = \int_{0}^{a} A(x) \, dx. \quad (34)$$

Thus, if $Z(0, t)$ (which is the driving point impedance function at $x = 0$) is known, or measured, then solution of equation (33) for each a gives the area function. [Note that to get $A(x)$ for $x \leqq a$, $Z(0, t)$ is required for $t \leqq 2a$, as expected from physical considerations.]*

We close this section by noting that although we have discussed the method in terms of measurements made at $x = 0$, where the boundary condition corresponds to a closed end, trivial modifications are needed if measurements are to be made at an open end. In the latter case since the pressure vanishes, $\partial v/\partial x = 0$. Therefore, the same method is applicable to the horn equation for volume velocity, with a measurement of driving point admittance (instead of impedance). The driving point impedance (admittance) may, of course, be evaluated from measurements at any end with an arbitrary, known, termination.

V. APPLICATION OF THE METHODS TO DETERMINING VOCAL TRACT AREA FUNCTIONS

As noted in the introduction, the one dimensional Webster's horn equation is an accurate description of wave propagation in the vocal tract, only for frequencies less than about 3.5 kHz. Hence the λ_i of Section II have no physical counterpart whenever they exceed 3.5. We therefore start with the λ_i and α_i $(i = 1, \cdots n)$ as measured data, and assume that for $i > n$, the λ_i and α_i are identical with those of some canonical tube. In view of the asymptotic formulae of Section II, this assumption is reasonable.

* For another derivation of equation (33) see Appendix D where it is further shown that $f(a) = [A(a)]^{1/2}$.

We know of no *a priori* method for choosing the canonical shape so as to give the best match between the computed area functions and those of the actual vocal tract. For simplicity one might assume the canonical tube to be uniform. However the experimental area functions published by Fant[9] all show a sharp discontinuity at the epiglottis, which suggests choosing a canonical tube with such a discontinuity. We have tried both these canonical shapes.

Once a canonical shape has been chosen, α_i , λ_i , $\psi(x, \lambda_i)$ and $\psi(x, \mu_i)$, $i = 1, \cdots n$ may be computed. Under the assumptions of this section, the factorization of equation (21) can then be carried out in the following manner.

We use the vector notation

$$\mathbf{k}_1^T(x) = [\psi(x, \lambda_i)/\alpha_1 , \cdots , \psi(x, \lambda_n)/\alpha_n , \psi(x, \mu_i)/\gamma_1 , \cdots , \psi(x, \mu_n)/\gamma_n]$$

$$\mathbf{k}_2^T(x) = [\psi(x, \lambda_i)/\alpha_1 , \cdots , \psi(x, \lambda_n)/\alpha_n , -\psi(x, \mu_1)/\gamma_1 , \cdots , -\psi(x, \mu_n)/\gamma_n]$$

where $\mathbf{k}_1(x)$ and $\mathbf{k}_2(x)$ are n-dimensional column vectors and the superscript T denotes transposition. Then the kernel of equation (22) becomes $k_1^T(x)k_2(t)$, and it is easily seen that $H(x, t)$ has the form

$$H(x, t) = \mathbf{k}_1^T(x)\mathbf{h}(t) \qquad x > t \tag{35}$$

where $h(t)$ is some vector function of t. Then equation (18) becomes

$$\mathbf{k}_1^T(x)\mathbf{k}_2(t) = \mathbf{k}_1^T(x)\mathbf{h}(t) + \mathbf{k}_1^T(x)\left[\int_0^t \mathbf{h}(\sigma)\mathbf{h}^T(\sigma)\, d\sigma\right]\mathbf{k}_1(t) \tag{36}$$

as long as $\lambda_i \neq \mu_i$, $i = 1, \cdots n$. (If $\lambda_i = \mu_i$ for some i, a slight modification is necessary.) However, from the linear independence of the components of $\mathbf{k}_1(x)$ it follows that

$$\mathbf{k}_2(t) = \mathbf{h}(t) + \left[\int_0^t \mathbf{h}(\sigma)\mathbf{h}^T(\sigma)\, d\sigma\right]\mathbf{k}_1(t). \tag{37}$$

Equation (37) can be solved for $\mathbf{h}(t)$ by the analog circuit shown in Fig. 1, or by an equivalent computer simulation. Also, since equation (23) now becomes

$$[A(x)]^{\frac{1}{2}} + \mathbf{k}_1^T(x) \int_0^x \mathbf{h}(t)[A(t)]^{\frac{1}{2}}\, dt = [A_0(x)]^{\frac{1}{2}} \tag{38}$$

the analog circuit of Fig. 2 yields $[A(x)]^{\frac{1}{2}}$.

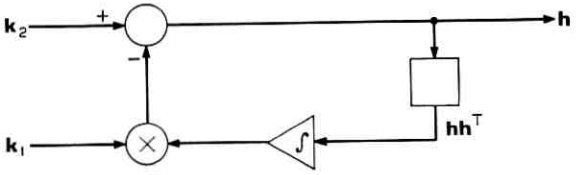


Fig. 1—Analog computer circuit for computing h.

The results of such computations for some area functions published by Fant[9] are shown in Figs. 3 through 9.

We close this section by noting that if the experimental data is in the form of a driving point impulse response, then the simplest procedure is to use the method of Section IV [that is, to solve equation (33) for various values of a]. We have not computed area functions by this method so far, but propose to do so, using impedance tube or other experimental data. The limitations due to the inapplicability of the horn equation at high frequencies apply to this method as well. The effect of low-pass filtering the driving point response is being investigated.

VI. CONCLUSIONS AND DISCUSSION

The comparison between measured and computed area functions of Figs. 3 through 9, indicates that knowledge of the first few (λ, α) pairs is sufficient to get reasonable estimates of the area function. The λ's may be obtained directly from the speech output, since they can be computed with reasonable accuracy from the formant frequencies. The α's on the other hand cannot be computed directly from the speech waveform, and impedance tube or other equivalent measurement would appear to be necessary. However, the vocal tract has physical constraints which might be reflected in a functional dependence of the α's on the λ's. The possibility of such functional dependence is being

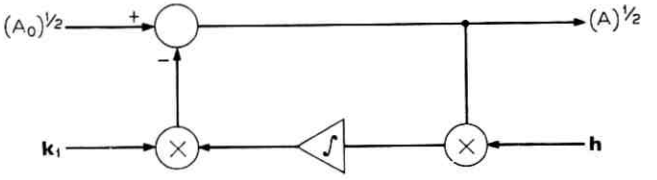


Fig. 2—Analog computer circuit for computing A.

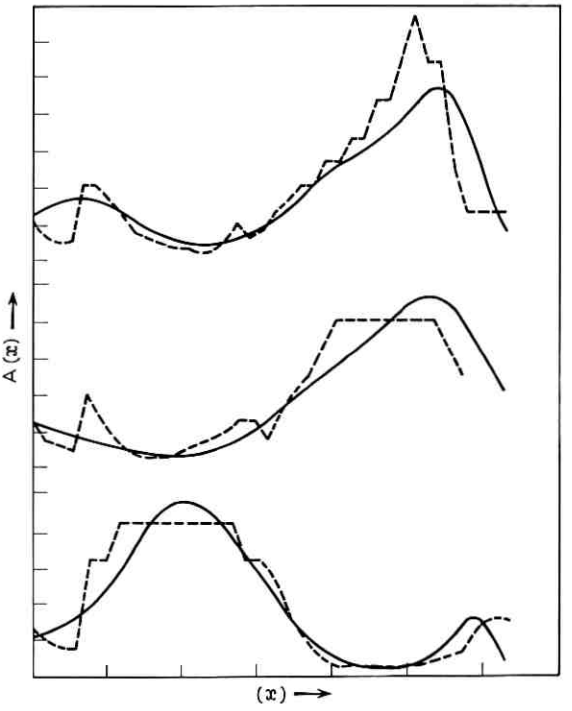


Fig. 3—Area functions reconstructed from the first *three* poles and residues of input impedance by the method of Section V using a uniform canonical tube. Dashed curves are the X-ray measurements.

investigated. The sensitivity of the computed area functions to changes in the α's is also being investigated.

If one is willing to make acoustical measurements at the lips, then the method of Section IV is the most direct way of computing the area function. It has the added advantage that the length of the vocal tract need not be known. Some preliminary results on the effect of band-limiting the impulse response have been obtained and will be reported in a later paper.

APPENDIX A

Proof of Lemma 1

Under the assumptions of this lemma, equation (6) may be transformed to:

$$\{[A(x)]^{\frac{1}{4}}\varphi(x, y)\}'' + \lambda^2[A(x)]^{\frac{1}{4}}\varphi(x, \lambda) = \{[A(x)]^{\frac{1}{4}}\}''\varphi(x, \lambda) \qquad (39)$$

except at the points $x_1, \cdots x_k$, where $A(x)$ is discontinuous. With $x_0 = 0$ and $x_{k+1} = \pi$, equation (39) gives

$$[A(x)]^{\frac{1}{2}}\varphi(x, \lambda) = f_i(x, \lambda) + \int_{x_i}^{x} \frac{\sin \lambda(x - t)}{\lambda} \frac{\{[A(t)]^{\frac{1}{2}}\}''}{[A(t)]^{\frac{1}{2}}} [A(t)]^{\frac{1}{2}}\varphi(t, \lambda)\, dt,$$
$$x_i < x \leqq x_{i+1}, \quad i = 0, 1, \cdots, k \tag{40}$$

where

$$f_i(x, \lambda) = a_i(\lambda) \cos \lambda x + b_i(\lambda) \sin \lambda x.$$

The coefficients $a_i(\lambda)$, $b_i(\lambda)$ are to be determined so as to make $\varphi(x, \lambda)$ and $A(x)\varphi'(x, \lambda)$ everywhere continuous. (The conditions at $x = 0$ give $a_0(\lambda) = 1$, $b_0(\lambda) = 0$.) Clearly for every λ there exists a bound $m_i(\lambda) = \sup |[A(x)]^{\frac{1}{2}}\varphi(x, \lambda)|$, $x_i \leqq x \leqq x_{i+1}$. Then from equation (40),

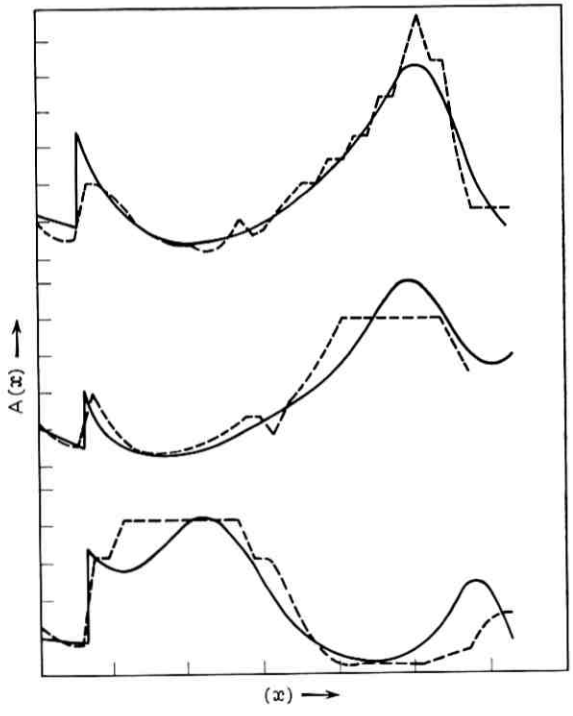


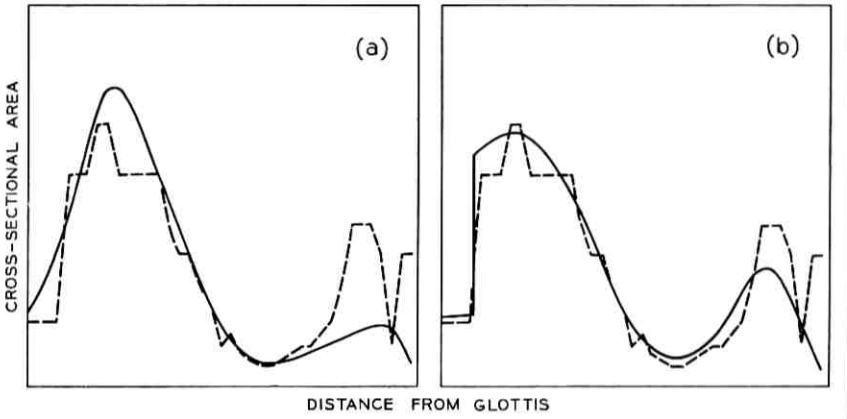
Fig. 4—Same as in Fig. 3, except the canonical tube was chosen with a discontinuity at the epiglottis.

$$m_i(\lambda) \leqq |\, f_i(x, \lambda)\,| + \frac{1}{\lambda}\, m_i(\lambda) \int_{x_i}^{x} |\sin \lambda(x - t)| \left| \frac{\{[A(t)]^{\frac{1}{2}}\}''}{[A(t)]^{\frac{1}{2}}} \right| dt$$

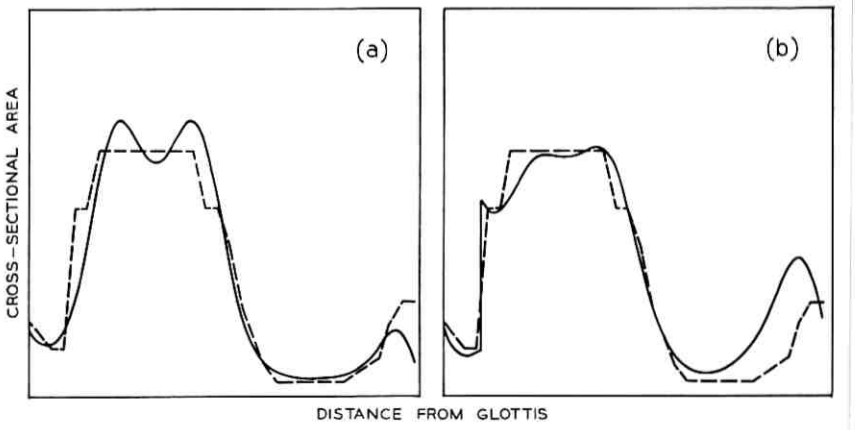
$$\leqq [a_i^2(\lambda) + b_i^2(\lambda)]^{\frac{1}{2}} + M m_i(\lambda)/\lambda \tag{41}$$

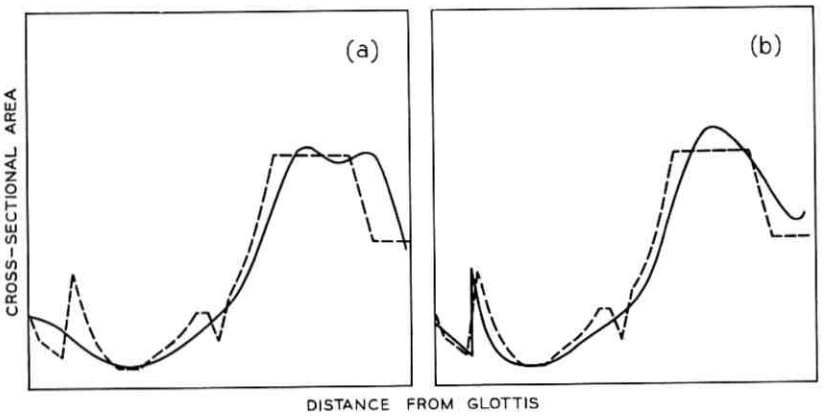
where M is a bound on the integral for all i. Thus for $\lambda > 2M$,

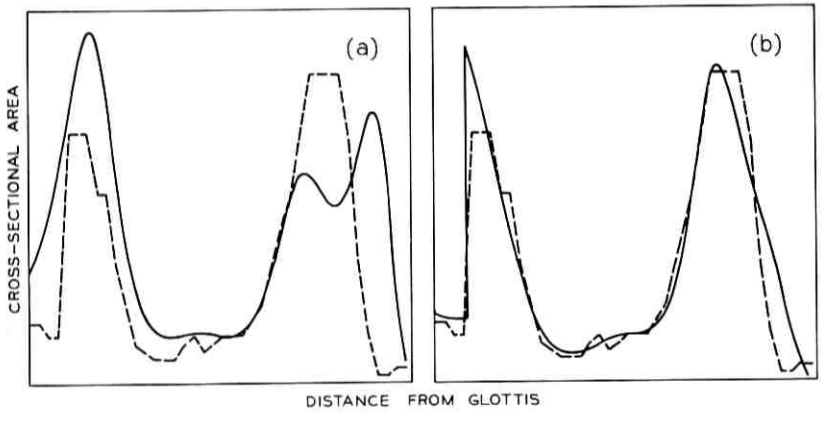
$$|\, [A(x)]^{\frac{1}{4}} \varphi(x, \lambda)\,| \leqq 2[a_i^2(\lambda) + b_i^2(\lambda)]^{\frac{1}{2}} \triangleq 2\gamma_i(\lambda). \tag{42}$$

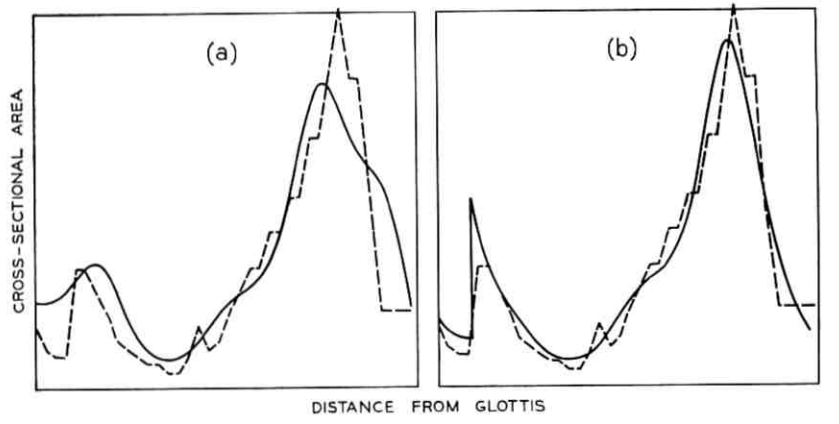


Figs. 5–9—Area functions reconstructed from the first *four* poles and residues: (a) the reconstruction using a uniform canonical tube, and (b) the reconstruction with a discontinuous canonical tube as in Section V. Dashed curves are X-ray measurements.




(a)
(b)
CROSS-SECTIONAL AREA
DISTANCE FROM GLOTTIS


(a)
(b)
CROSS-SECTIONAL AREA
DISTANCE FROM GLOTTIS


(a)
(b)
CROSS-SECTIONAL AREA
DISTANCE FROM GLOTTIS

Then from equation (40),

$$[A(x)]^{\frac{1}{2}}\varphi(x, \lambda) = f_i(x, \lambda) + \frac{\gamma_i(\lambda)c_i(x)}{\lambda}, \qquad x_i < x < x_{i+1} \tag{43}$$

where $c_i(x)$ is bounded. Differentiating equation (40) and using a similar argument gives,

$$\frac{[A(x)]^{\frac{1}{2}}}{\lambda}\frac{\partial\varphi(x, \lambda)}{\partial x} = \frac{1}{\lambda}\frac{\partial f_i(x, \lambda)}{\partial x} - \gamma_i(\lambda)\, d_i(x)/\lambda, \qquad x_i < x < x_{i+1} \tag{44}$$

with $d_i(x)$ bounded. Defining

$$\mathbf{k}_i \equiv \begin{bmatrix} [A(x_i+)/A(x_i-)]^{\frac{1}{2}} & 0 \\ 0 & [A(x_i-)/A(x_i+)]^{\frac{1}{2}} \end{bmatrix}$$

and

$$\mathbf{R}_i \equiv \begin{bmatrix} \cos\lambda x_i & \sin\lambda x_i \\ \sin\lambda x_i & -\cos\lambda x_i \end{bmatrix}$$

the continuity conditions at x_i give:

$$\begin{bmatrix} a_i(\lambda) \\ b_i(\lambda) \end{bmatrix} = \mathbf{R}_i\mathbf{k}_i\mathbf{R}_i\begin{bmatrix} a_{i-1}(\lambda) \\ b_{i-1}(\lambda) \end{bmatrix} + \frac{\gamma_{i-1}(\lambda)}{\lambda}\cdot\mathbf{R}_i\mathbf{k}_i\begin{bmatrix} c_{i-1}(x_i) \\ d_{i-1}(x_i) \end{bmatrix} - \frac{\gamma_i(\lambda)}{\lambda}\mathbf{R}_i\begin{bmatrix} c_i(x_i) \\ d_i(x_i) \end{bmatrix}. \tag{45}$$

Since the norm of $\mathbf{R}_i$ is unity and of $\mathbf{k}_i$ finite it follows upon taking the lengths of the vectors on either side of equation (45) that for large enough λ, $\gamma_i(\lambda) \leqq K'\gamma_{i-1}(\lambda)$, for some finite constant K'. Since $\gamma_0(\lambda) = 1$, it follows that $\gamma_i(\lambda)$ is bounded for all i, as $\lambda \to \infty$. Then from equation (45)

$$\begin{bmatrix} a_i \\ b_i \end{bmatrix} = \mathbf{R}_i\mathbf{k}_i\mathbf{R}_i\begin{bmatrix} a_{i-1} \\ b_{i-1} \end{bmatrix} + 0\left(\frac{1}{\lambda}\right). \tag{46}$$

However, if a_i, b_i satisfy equation (46), then $f_i(x, \lambda) = [A_i(x)]^{\frac{1}{2}}\psi(x, \lambda) + 0(1/\lambda)$ for $x_i \leqq x \leqq x_{i+1}$, therefore from equation (43)

$$[A(x)]^{\frac{1}{2}}\varphi(x) = [A_0(x)]^{\frac{1}{2}}\psi(x) + \frac{c(x)}{\lambda} \tag{47}$$

with $c(x)$ bounded.

APPENDIX B

Existence of the Linear Transformation of Equation (12)

In this appendix we prove the existence of the linear transformation of equation (12). Consider first the region $0 \leqq x \leqq x_1$, where x_1 is the first point of discontinuity of $A(x)$. Then with $y(x) = [A(x)]^{\frac{1}{4}}\phi(x, \lambda)$ and $q(x) = \{[A(x)]^{\frac{1}{4}}\}''/[A(x)]^{\frac{1}{4}}$, equation (6) becomes

$$y''(x) = -\lambda^2 y(x) + q(x)y(x). \tag{48}$$

Consider the function

$$\mathfrak{N}(x) = \int_0^x H(x, t)y(t)\, dt. \tag{49}$$

Then

$$\mathfrak{N}''(x) = \frac{d}{dx}[H(x, x)y(x)] + \frac{\partial H(x, t)}{\partial x} y(t)\Big|_{t=x} + \int_0^x \frac{\partial^2 H(x, t)}{\partial x^2} y(t)\, dt. \tag{50}$$

After integrating twice by parts we arrive at the identity

$$\int_0^x \frac{\partial^2 H(x, t)}{\partial t^2} y(t)\, dt = \frac{\partial H(x, t)}{\partial t} y(t)\Big|_{t=0}^{t=x} - H(x, t)\frac{dy}{dt}\Big|_{t=0}^{t=x} + \int_0^x H(x, t)\frac{d^2y}{dt^2}\, dt. \tag{51}$$

Substituting for d^2y/dt^2 from equation (48) into equation (51) and adding equation (50) we get

$$\mathfrak{N}''(x) + \lambda^2\mathfrak{N} = \int_0^x \left[\frac{\partial^2 H(x, t)}{\partial x^2} - \frac{\partial^2 H(x, t)}{\partial t^2} + q(t)H(x, t)\right] \cdot y(t)\, dt + 2y(x)\frac{d}{dx}H(x, x) - y(t)\frac{\partial H(x, t)}{\partial t}\Big|_{t=0} + H(x, t)\frac{dy(t)}{dt}\Big|_{t=0}. \tag{52}$$

If now $H(x, t)$ satisfies the differential equation (13) with boundary conditions (14) and (15), then equation (52) shows that $\mathfrak{N}(x) + y(x)$ is some linear combination of $\cos \lambda x$ and $\sin \lambda x$. Matching of boundary conditions at $x = 0$ then shows that

$$\mathfrak{N}(x) + y(x) = \cos \lambda x = [A_0(x)]^{\frac{1}{4}}\psi(x, \lambda), \qquad 0 \leqq x \leqq x_1. \tag{53}$$

The proof may be extended to $x > x_1$ by a similar procedure. Thus, for example, in the range $x_1 \leqq x \leqq x_2$, $H(x, t)$ must satisfy the differential equation (13), the boundary conditions (14) and (15) for $x > x_1$, and boundary conditions at $x = x_1$ imposed by the continuity requirements on $\phi(x, \lambda)$, $\psi(x, \lambda)$, $A(x)\phi'(x, \lambda)$ and $A_0(x)\psi'(x, \lambda)$.

APPENDIX C

Spectral Function from Two Sets of Eigenvalues

We outline a method of getting a spectral function from two sets of eigenvalues. Let $\varphi(x, \lambda)$ be the solution of equation (6) such that

$$\varphi(\pi, \lambda) = \alpha, \qquad A(\pi)\varphi'(\pi, \lambda) = \beta \tag{54}$$

for every λ. Let $\psi(x, \lambda)$ be the solution such that

$$\psi(\pi, \lambda) = \gamma, \qquad A(\pi)\psi'(\pi, \lambda) = \delta. \tag{55}$$

Let $\lambda_1^2, \lambda_2^2, \cdots$ be the values of λ^2 for which $a\varphi(0, \lambda) + bA(0)\varphi'(0, \lambda) = 0$ and let $\mu_1^2, \mu_2^2, \cdots$ be the values of μ^2 for which $a\psi(0, \mu) + bA(0)\psi'(0, \mu) = 0$. Let

$$m(\lambda) = \frac{a\psi(0, \lambda) + bA(0)\psi'(0, \lambda)}{a\varphi(0, \lambda) + bA(0)\varphi'(0, \lambda)}. \tag{56}$$

Then the zeroes of $m(\lambda)$ are $\mu_1, \mu_2, \cdots$ and the poles are $\lambda_1, \lambda_2, \cdots$. If $X(x, \lambda)$ is any solution of equation (6), then it is easily shown that

$$(\lambda^2 - \lambda_n^2) \int_0^\pi A(x)X(x, \lambda)\varphi(x, \lambda_n)\, dx$$

$$= A(x)[X(x, \lambda)\varphi'(x, \lambda_n) - \varphi(x, \lambda_n)X'(x, \lambda)]_0^\pi. \tag{57}$$

Choosing $X(x, \lambda) = \psi(x, \lambda) - m(\lambda)\varphi(x, \lambda)$ in equation (57) and using the boundary conditions on $\psi(x, \lambda)$ and $\varphi(x, \lambda)$, we get

$$(\lambda^2 - \lambda_n^2) \int_0^\pi A(x)[\psi(x, \lambda) - m(\lambda)\varphi(x, \lambda)]\, dx = \beta\gamma - \alpha\delta \tag{58}$$

for all λ. As $\lambda \to \lambda_n$

$$\alpha_n^2 = \int_0^\pi A(x)\varphi^2(x, \lambda_n)\, dx = \lim_{\lambda \to \lambda_n} (\alpha\delta - \beta\gamma)/[(\lambda^2 - \lambda_n^2)m(\lambda)]. \tag{59}$$

Thus, given λ_1, λ_2, and μ_1, μ_2, one obtains $m(\lambda)$ and hence $\alpha_1, \alpha_2, \cdots$.

APPENDIX D

Derivation of Integral Equation (33)

We give here a derivation of the integral equation (33) based upon the results of Section III. For simplicity we will assume that $A(x)$ [and hence $A_0(x)$] has no discontinuities. Then from equation (23)

$$(I + H)A^{\frac{1}{2}}(x) = u(x) \tag{60}$$

where $u(x)$ is equal to 1 for all $x > 0$. Thus if $f(x)$ is a function such that

$$[(I + H)^{-1}u](x) = g(x) \tag{61}$$

then

$$\int_0^\pi g^2(x)\,dx = \int_0^\pi A(x)\,dx. \tag{62}$$

Notice that if

$$(I + H)(I + H)^* = I + K \tag{63}$$

then

$$\begin{aligned}\int_0^\pi A(x)\,dx &= \int_0^\pi [(I + K)^{-1}u](x)\,dx \\ &= \int_0^\pi f(x)\,dx\end{aligned} \tag{64}$$

where $f(x)$ satisfies the equation

$$[(I + K)f](x) = u(x). \tag{65}$$

The kernel of $I + K$ is recognized as that of equation (31) with $a = \pi$. Equation (33) therefore follows (for $a = \pi$) from the symmetrization of $f(x)$, exactly as in Section IV. However, the argument given here is independent of the length π, which may be replaced by a.

Using equations (60), (63) and (65), we have

$$(I + H)^* f(\cdot) = [A(\cdot)]^{\frac{1}{2}}. \tag{66}$$

Therefore from equation (17)

$$f(a) = [A(a)]^{\frac{1}{2}}. \tag{67}$$

REFERENCES

1. Mermelstein, P., and Schroeder, M. R., "Determination of Smoothed Cross-Sectional Area Functions of the Vocal Tract from Formant Frequencies," paper A-24, Proceedings of the Fifth International Congress on Acoustics,

1965, Liege, Belgium, D. E. Commins, editor (Imprimerie Georges Thone, Liege, 1965), Vol. 1a.
2. Schroeder, M. R., "Determination of the Geometry of the Human Vocal Tract," J. Acoust. Soc. of Amer., *41*, No. 4 (April 1967), pp. 1002–1010.
3. Mermelstein, P., "Determination of the Vocal-Tract Shape from Measured Formant Frequencies," J. Acoust. Soc. of Amer., *41*, No. 5 (May 1967), pp. 1283–1294.
4. Heinz, J., "Perturbation Functions for the Determination of Vocal Tract Area Functions from Vocal Tract Eigenvalues," Quarterly Progress and Status Report, April 15, 1967, Speech Transmission Laboratory, Royal Inst. of Tech., Stockholm, Sweden, pp. 1–14.
5. Marchenko, V. A., "Some Questions in the Theory of Second Order Differential Operators," Doklady Akademii Nauk, SSSR (N.S.), *72*, No. 3 (May 21, 1950), pp. 457–460.
6. Gelfand, I. M., and Levitan, B. M., "On the Determination of a Differential Equation from its Spectral Function," Izvestia Akademii Nauk, SSSR (Seria Matematicheskaya), *15*, No. 4 (July–August 1951), pp. 309–360. English Translation: Amer. Math. Soc. Translations, Ser. 2, *1* (1955), pp. 253–304.
7. Krein, M. G., "Solution of the Inverse Sturm-Liouville Problem," Doklady Akademii Nauk SSSR (N.S.), *76*, No. 1 (January 1, 1951), pp. 21–24.
8. Krein, M. G., "Determination of the Density of a Nonuniform Symmetric String from its Frequency Spectrum," Doklady Akademii Nauk SSSR (N.S.), *76*, No. 3 (January 21, 1951), pp. 345–348.
9. Fant, G. M., "Acoustic Theory of Speech Production," The Hague, Netherlands: Mouton and Co., p. 115.

Analysis of a Thin Circular Loop Antenna Over a Homogeneous Earth

By S. C. MOORTHY

(Manuscript received March 31, 1969)

In this paper, the current distribution on a bare conducting loop, situated in free space over a semi-infinite medium, is obtained for arbitrary time harmonic excitations. The loop is assumed to be thin, perfectly conducting and the standard one-dimensional integral equation and its Fourier series solution are used as the starting points. The field due to the current in the loop, where the semi-infinite medium is absent, is expressed as a superposition of plane waves. The tangential component of the field reflected by the interface, of the semi-infinite medium, is evaluated using appropriate Fresnel reflection coefficients. This reflected field serves as a new source for the loop and induces a current on the loop. The field due to the induced current is treated in the same manner, and this process is repeated indefinitely. The summation of the original current and all the induced currents gives the steady-state current on the loop.

I. INTRODUCTION

It is well known that a high altitude nuclear burst generates an intense electromagnetic transient which covers a large geographical area.[1] This transient field induces currents in communication circuits and, if these are large enough, adversely affects communication channels. One problem of particular interest in land-line communication is the coupling to large loops formed by cables.

The loops formed by cables deployed in practical communication systems are very complex and cannot be analyzed easily. Typically, they run for many miles over inhomogeneous terrain and contain many junction points; nevertheless a great deal of insight into the behavior of these irregular loops can be obtained by studying the behavior of a large regular loop over a homogeneous ground. In this paper, the theoretical foundations for an analysis of a circular loop over a homogeneous ground are developed, for time harmonic excitations. The

response of the loop for transient fields may be obtained by standard Fourier transform techniques.

Various problems related to the thin circular loop have been considered by numerous authors. These may be broadly classified under two categories: loop in an infinite homogeneous medium and loop in a stratified medium.

In the first category, Pocklington, Oseen, Hallen, Storer, and Wu have analyzed the problem of a bare thin perfectly conducting circular loop in free space.[2–6] All these authors use the Fourier series expansion to solve the integral equation for the current in the loop. A good analysis of the problem is given in Wu's paper. Adachi and Mushiake analyze the same problem by solving the integrodifferential equation for the current using an iterative method.[7,8] Mei, Baghdasarian and Angelakos, and Tang have discussed the direct numerical solution of the integral equation.[9–11] A variational approach for determining the scattering cross section of a loop is given by Kouyoumjian.[12] Problems concerning loaded loops are considered by Iizuka, Harrington and Ryerson, and Harrington and Mautz.[13–15] The analysis of a loop in a conducting medium is an extension of the analysis of a loop in free space and lends itself to certain approximations. Kraichman, Chen and King and King, and Harrison and Tingley have discussed the bare loop in a dissipative medium;[16–18] Galejs has discussed an insulated loop in a dissipative medium.[19] Finally, the solution to the problem of two identical coaxial coupled loops in a homogeneous medium has been solved by Iizuka, King and Harrison.[20]

In the second category the literature is mostly on small loops or magnetic dipoles over different types of media and is very extensive. (See Ref. 21 for an extensive bibliography.) Wait has considered the problem of loops over a homogeneous earth;[22,23] recently, Sinha and Bhattacharya have analyzed the problem of a vertical magnetic dipole buried inside the earth.[24]

The treatment of a small current carrying loop as a magnetic dipole, while satisfactory for many purposes, is nevertheless inexact. Moreover, we do encounter situations where the loop diameter is comparable to the wavelength of the excitation frequency and here we cannot assume the current to be uniformly distributed on the loop. A typical example would be the excitation of a loop by a narrow electromagnetic pulse which contains a broad spectrum of frequencies. The purpose of this paper is to solve the problem of a bare loop over a homogeneous earth taking into account the current distribution on the loop.

Specifically, the system under consideration is a bare, thin, perfectly

conducting circular loop of mean radius b, formed by bending a cylindrical wire of radius a, situated in free space with its plane parallel to and at a distance d from the interface of a semi-infinite, linear homogeneous, isotropic medium (Fig. 1). The loop is excited by an electromagnetic wave of harmonic (exp $j\omega t$) time variation (the slice generator used to compute the admittance being a limiting case). It is assumed that $k_o a \ll 1$ and $a \ll b$, so that, if the loop were situated in a homogeneous medium the current distribution induced by a specified time harmonic excitation is given by the so-called one-dimensional integral equation.[6] In addition it is assumed that $d \gg a$.

It is desired to determine the current distribution $I(\phi)$ on the loop in the aforementioned system. This is accomplished in the following three, more or less self-contained, sections. In Section II the field of a circular filamentary current in free space is expressed as a superposition of plane waves. Section III evaluates the reflected field when an arbitrary field, of the general form obtained in Section II, is incident on the interface between free space and the semi-infinite medium. In Section IV the results of Sections II and III are combined with the integral equation for the current on a loop in free space to determine the steady state current $I(\phi)$ on the loop using a recurrent "reflection-induction" scheme.

II. THE ELECTROMAGNETIC FIELD OF A CIRCULAR FILAMENTARY CURRENT

Consider a circular filament of current $I(\phi)$ (Fig. 2a) of radius b situated in free space. The coordinate system is so chosen that the loop is parallel to the xy plane at a distance d from it. The loop current may be expressed as a surface current density $\mathbf{K}$, in the $z = d$ plane, in the following manner.

$$\mathbf{K} = \mathbf{a}_\phi I(\phi)\delta(\rho - b)\delta(z - d) \tag{1}$$

where ρ, ϕ and z are the cylindrical coordinates and $\mathbf{a}_\phi$ the unit vector in the ϕ direction. The electromagnetic field due to $\mathbf{K}$ may be expressed as a superposition of plane waves as follows.[25]

$$\mathbf{H}^{(i)} = \iint_{-\infty}^{+\infty} [\pm P, \pm Q, (lP + mQ)(1 - l^2 - m^2)^{-\frac{1}{2}}] \cdot \exp\{jk_o[lx + my \mp (1 - l^2 - m^2)^{\frac{1}{2}}(z - d)]\}\, dl\, dm, \qquad z \gtrless d. \tag{2}$$

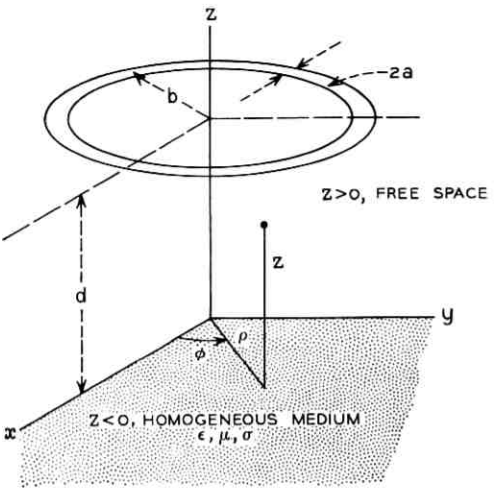


Fig. 1—A thin circular loop over a homogeneous semi-infinite medium.

$$\mathbf{E}^{(i)} = \eta_o \iint_{-\infty}^{\infty} \{[lmP + (1 - l^2)Q](1 - l^2 - m^2)^{-\frac{1}{2}} - [(1 - m^2)P + lmQ](1 - l^2 - m^2)^{-\frac{1}{2}}, \pm(lQ - mP)\} \cdot \exp\{jk_o[lx + my \mp (1 - l^2 - m^2)^{\frac{1}{2}}(z - d)]\}\, dl\, dm, \quad z \gtrless d, \tag{3}$$

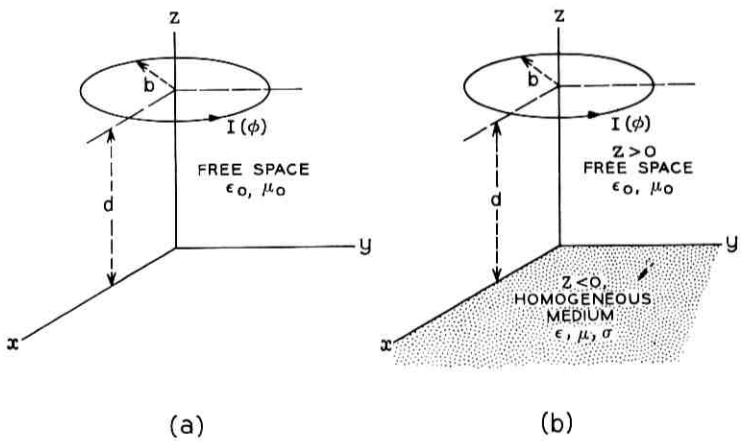


Fig. 2—(a) A filamentary loop current in free space, (b) A filamentary loop current over a homogeneous semi-infinite medium.

where

$$P(l, m) = (k_o^2/8\pi^2) \iint_{-\infty}^{+\infty} K_y(x, y) \exp [-jk_o(lx + my)]\, dx\, dy, \quad (4)$$

$$Q(l, m) = -(k_o^2/8\pi^2) \iint_{-\infty}^{+\infty} K_x(x, y) \exp [-jk_o(lx + my)]\, dx\, dy, \quad (5)$$

$$k_o^2 = \omega^2 \mu_o \epsilon_o \quad \text{and} \quad \eta_o = (\mu_o/\epsilon_o)^{\frac{1}{2}}. \quad (6)$$

The quantities l and m are in general complex and the integrals in equations (2) and (3) are contour integrals, the choice of contours being dictated by physical considerations. A possible choice of l and m is defined by the following transformation

$$\left.\begin{aligned} l &= \tau \cos \psi, \\ m &= \tau \sin \psi, \end{aligned}\right\} \quad 0 \leqq \tau, \quad \psi \text{ being complex.} \quad (7)$$

Let

$$I(\phi) = \sum_{n=-\infty}^{+\infty} I_n \exp (jn\phi). \quad (8)$$

Substitution of equations (1), (7) and (8) into equations (4) and (5), and changing from rectangular to cylindrical coordinates yields the following equations

$$P(\tau, \psi) = (k_o^2 b/8\pi) \sum_{n=-\infty}^{+\infty} (-j)^n \exp (jn\psi) J_n(k_o b\tau)(I_{n-1} + I_{n+1}), \quad (9)$$

$$Q(\tau, \psi) = -j(k_o^2 b/8\pi) \sum_{n=-\infty}^{+\infty} (-j)^n \exp (jn\psi) J_n(k_o b\tau)(I_{n-1} - I_{n+1}), \quad (10)$$

where J_n denotes the Bessel function of order n.

III. CIRCULAR FILAMENTARY CURRENT OVER A SEMI-INFINITE MEDIUM

Here again we consider the filamentary loop of Section II, but instead of being situated in free space it is situated over the homogeneous semi-infinite medium $z \leq 0$ (Fig. 2b). The total field in the region $z \geqq 0$ consists of the primary field $\mathbf{E}^{(i)}$, $\mathbf{H}^{(i)}$ of the filamentary current [equations (2) and (3)] and the field $\mathbf{E}^{(r)}$, $\mathbf{H}^{(r)}$ reflected by the interface $z = 0$. We proceed as follows to evaluate the latter. Let

$$\mathbf{H}^{(i)}(x, y, z) = \iint_{-\infty}^{+\infty} \mathbf{H}_o(l, m) \exp [-jk_o \mathbf{n}_o(l, m) \cdot \mathbf{r}]\, dl\, dm, \qquad z \leqq d, \quad (11)$$

$$\mathbf{E}^{(i)}(x, y, z) = \iint_{-\infty}^{+\infty} \mathbf{E}_o(l, m) \exp [-jk_o \mathbf{n}_o(l, m) \cdot \mathbf{r}] \, dl \, dm, \qquad z \leqq d, \tag{12}$$

where

$$\mathbf{H}_o(l, m) = [-\mathbf{a}_x P - \mathbf{a}_y Q + \mathbf{a}_z(lP + mQ)(1 - l^2 - m^2)^{-\frac{1}{2}}] \cdot \exp [-jk_o d(1 - l^2 - m^2)^{\frac{1}{2}}], \tag{13}$$

$$\mathbf{E}_o(l, m) = \eta_o\{\mathbf{a}_x[lmP + (1 - l^2)Q](1 - l^2 - m^2)^{-\frac{1}{2}} + \mathbf{a}_y[(m^2 - 1)P - lmQ](1 - l^2 - m^2)^{-\frac{1}{2}} + \mathbf{a}_z(mP - lQ)\} \cdot \exp [-jk_o d(1 - l^2 - m^2)^{\frac{1}{2}}], \tag{14}$$

$$\mathbf{n}_o(l, m) = -\mathbf{a}_x l - \mathbf{a}_y m - \mathbf{a}_z(1 - l^2 - m^2)^{\frac{1}{2}}, \tag{15}$$

and

$$\mathbf{r} = \mathbf{a}_x x + \mathbf{a}_y y + \mathbf{a}_z z. \tag{16}$$

Each one of the constituent plane waves propagates in the "direction" $\mathbf{n}_o(l, m)$. Let $R_\perp(l, m)$ and $R_\parallel(l, m)$ represent the Fresnel reflection coefficients for the cases where the incident electric field is perpendicular and parallel respectively to the plane of incidence. Evaluation of the reflected field is achieved by resolving each plane wave into components with the electric field perpendicular and parallel to the plane of incidence. To this end we define a local coordinate system† as shown in Fig. 3.

The plane of incidence is defined by the unit vectors $\mathbf{a}_z$ and $\mathbf{n}_o(l, m)$. Let

$$\begin{aligned} \mathbf{a}_1 &= (\mathbf{a}_z \times \mathbf{n}_o)/[1 - (\mathbf{a}_z \cdot \mathbf{n}_o)^2]^{\frac{1}{2}}, \\ \mathbf{a}_2 &= \mathbf{a}_z \times \mathbf{a}_1 . \end{aligned} \tag{17}$$

Then $\mathbf{a}_z$, $\mathbf{a}_1$ and $\mathbf{a}_2$ form a right-handed coordinate system, $\mathbf{a}_1$ is normal to the plane of incidence and $\mathbf{a}_z$ and $\mathbf{a}_2$ lie in the plane of incidence. In terms of $\mathbf{a}_x$ and $\mathbf{a}_y$ we have

$$\mathbf{a}_1 = (l^2 + m^2)^{-\frac{1}{2}}(m\mathbf{a}_x - l\mathbf{a}_y), \tag{18}$$

$$\mathbf{a}_2 = (l^2 + m^2)^{-\frac{1}{2}}(l\mathbf{a}_x + m\mathbf{a}_y). \tag{19}$$

† The propagation vector $\mathbf{n}_0$ becomes complex for certain values of l and m, the associated plane wave being inhomogenous. When this happens some of the terms used in the analysis (for example, plane of incidence, coordinate system, normal, and so on) become inaccurate and should be interpreted in a generalized sense. The results obtained are quite general and valid.

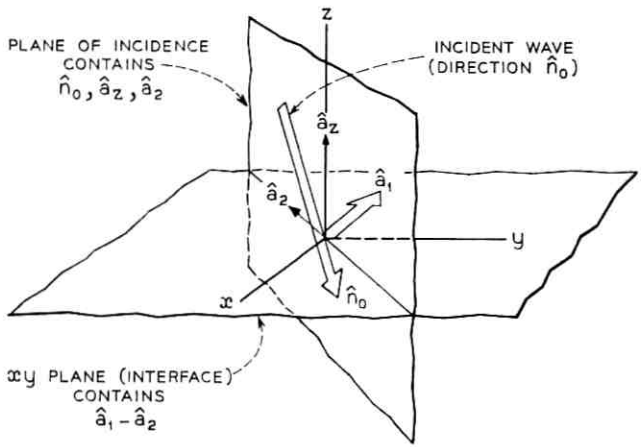


Fig. 3—Local coordinate system.

The field vectors $\mathbf{H}_o$ and $\mathbf{E}_o$ are now resolved into components perpendicular ($\mathbf{H}_{o\perp}$, $\mathbf{E}_{o\perp}$) and parallel ($\mathbf{H}_{o\parallel}$, $\mathbf{E}_{o\parallel}$) to the plane of incidence as follows

$$\mathbf{H}_o = \mathbf{H}_{o\perp} + \mathbf{H}_{o\parallel}, \tag{20}$$

where

$$\mathbf{H}_{o\perp} = \mathbf{a}_1(l^2 + m^2)^{-\frac{1}{2}}(lQ - mP) \exp [-jk_o d(1 - l^2 - m^2)^{\frac{1}{2}}], \tag{21}$$

$$\mathbf{H}_{o\parallel} = [\mathbf{a}_z(1 - l^2 - m^2)^{-\frac{1}{2}} - \mathbf{a}_2(l^2 + m^2)^{-\frac{1}{2}}](lP + mQ) \cdot \exp [-jk_o d(1 - l^2 - m^2)^{\frac{1}{2}}], \tag{22}$$

$$\mathbf{E}_o = \mathbf{E}_{o\perp} + \mathbf{E}_{o\parallel}, \tag{23}$$

where

$$\mathbf{E}_{o\perp} = \mathbf{a}_1\eta_o(l^2 + m^2)^{-\frac{1}{2}}(1 - l^2 - m^2)^{-\frac{1}{2}}(lP + mQ) \exp [-jk_o d(1 - l^2 - m^2)^{\frac{1}{2}}], \tag{24}$$

$$\mathbf{E}_{o\parallel} = \eta_o[\mathbf{a}_z - \mathbf{a}_2(l^2 + m^2)^{-\frac{1}{2}}(1 - l^2 - m^2)^{\frac{1}{2}}](mP - lQ) \cdot \exp [-jk_o d(1 - l^2 - m^2)^{\frac{1}{2}}]. \tag{25}$$

It may be easily verified that

$$\mathbf{H}_{o\parallel} = \eta_o^{-1}(\mathbf{n}_o \times \mathbf{E}_{o\perp}), \tag{26}$$

$$\mathbf{E}_{o\parallel} = -\eta_o(\mathbf{n}_o \times \mathbf{H}_{o\perp}). \tag{27}$$

The reflected field corresponding to the incident field defined in equations (11) and (12) may be represented as

$$\mathbf{H}^{(r)}(x, y, z) = \iint_{-\infty}^{+\infty} \mathbf{H}_2(l, m) \exp [-jk_o\mathbf{n}_2(l, m)\cdot\mathbf{r}]\, dl\, dm, \qquad z \geqq 0, \tag{28}$$

$$\mathbf{E}^{(r)}(x, y, z) = \iint_{-\infty}^{+\infty} \mathbf{E}_2(l, m) \exp [-jk_o\mathbf{n}_2(l, m)\cdot\mathbf{r}]\, dl\, dm, \qquad z \geqq 0, \tag{29}$$

where

$$\mathbf{n}_2(l, m) = -\mathbf{a}_x l - \mathbf{a}_y m + (1 - l^2 - m^2)^{\frac{1}{2}}\mathbf{a}_z . \tag{30}$$

Let

$$\mathbf{H}_2 = \mathbf{H}_{2\perp} + \mathbf{H}_{2\parallel} , \tag{31}$$

$$\mathbf{E}_2 = \mathbf{E}_{2\perp} + \mathbf{E}_{2\parallel} . \tag{32}$$

According to Fresnel's laws

$$\mathbf{E}_{2\perp} = R_\perp(l, m)\mathbf{E}_{o\perp} , \tag{33}$$

$$\mathbf{H}_{2\perp} = R_\parallel(l, m)\mathbf{H}_{o\perp} , \tag{34}$$

where

$$R_\perp(l, m) = \{\mu k_o(1 - l^2 - m^2)^{\frac{1}{2}} - \mu_o[k^2 - k_o^2(l^2 + m^2)]^{\frac{1}{2}}\}/ \{\mu k_o(1 - l^2 - m^2)^{\frac{1}{2}} + \mu_o[k^2 - k_o^2(l^2 + m^2)]^{\frac{1}{2}}\}, \tag{35}$$

$$R_\parallel(l, m) = \{\mu_o k^2(1 - l^2 - m^2)^{\frac{1}{2}} - \mu k_o[k^2 - k_o^2(l^2 + m^2)]^{\frac{1}{2}}\}/ \{\mu_o k^2(1 - l^2 - m^2)^{\frac{1}{2}} + \mu k_o[k^2 - k_o^2(l^2 + m^2)]^{\frac{1}{2}}\}, \tag{36}$$

$$k^2 = -j\omega\mu(\sigma + j\omega\epsilon). \tag{37}$$

We also have the relations

$$\mathbf{H}_{2\parallel} = \eta_o^{-1}(\mathbf{n}_2 \times \mathbf{E}_{2\perp}), \tag{38}$$

$$\mathbf{E}_{2\parallel} = -\eta_o(\mathbf{n}_2 \times \mathbf{H}_{2\perp}). \tag{39}$$

Substitution of equations (33), (34), (38) and (39) into equations (31) and (32) yields

$$\mathbf{E}_2 = R_\perp\mathbf{E}_{o\perp} - \eta_o R_\parallel(\mathbf{n}_2 \times \mathbf{H}_{o\perp}), \tag{40}$$

$$\mathbf{H}_2 = R_\parallel\mathbf{H}_{o\perp} + \eta_o^{-1}R_\perp(\mathbf{n}_2 \times \mathbf{E}_{o\perp}). \tag{41}$$

Equations (40), (41), (28) and (29) specify the reflected field completely. The ϕ component of the electric field is of special interest since it induces a current in an actual circular loop. Taking the ϕ component of equation (40), using polar coordinates for x, y, and equation (7), we obtain

$$E_{2\phi} = \eta_o[R_\parallel(\tau)(1 - \tau^2)^{\frac{1}{2}}(P \sin \psi - Q \cos \psi) \sin (\psi - \phi) - R_\perp(\tau)(1 - \tau^2)^{-\frac{1}{2}}(P \cos \psi + Q \sin \psi) \cos (\psi - \phi)] \cdot \exp [-jk_o d(1 - \tau^2)^{\frac{1}{2}}], \tag{42}$$

where

$$R_\parallel(\tau) = [\mu_o k^2(1 - \tau^2)^{\frac{1}{2}} - \mu k_o(k^2 - k_o^2\tau^2)^{\frac{1}{2}}]/[\mu_o k^2(1 - \tau^2)^{\frac{1}{2}} + \mu k_o(k^2 - k_o^2\tau^2)^{\frac{1}{2}}], \tag{43}$$

$$R_\perp(\tau) = [\mu k_o(1 - \tau^2)^{\frac{1}{2}} - \mu_o(k^2 - k_o^2\tau^2)^{\frac{1}{2}}]/[\mu k_o(1 - \tau^2)^{\frac{1}{2}} + \mu_0(k^2 - k_o^2\tau^2)^{\frac{1}{2}}]. \tag{44}$$

$P(\tau, \psi)$, $Q(\tau, \psi)$ are defined in equations (9) and (10), and after rearrangement yield the following equations

$$P \cos \psi + Q \sin \psi = j(4\pi)^{-1}k_o^2 b \sum_{n=-\infty}^{+\infty} I_n J_n'(k_o b\tau) \exp \left[jn\left(\psi - \frac{\pi}{2}\right)\right], \tag{45}$$

$$P \sin \psi - Q \cos \psi = (4\pi\tau)^{-1} n k_o \sum_{n=-\infty}^{+\infty} I_n J_n(k_o b\tau) \exp \left[jn\left(\psi - \frac{\pi}{2}\right)\right], \tag{46}$$

where the prime denotes differentiation with respect to the argument. The ϕ component of the reflected field is contained in equation (29) and is explicitly given by

$$E_\phi^{(r)}(\rho, \phi, z) = \int_0^\infty \int_C E_{2\phi}(\tau, \psi) \cdot \exp [jk_o\rho\tau \cos (\psi - \phi) - jk_o z(1 - \tau^2)^{\frac{1}{2}}]\, d\psi\tau\, d\tau. \tag{47}$$

The contour of integration C in the complex ψ plane is to be chosen from physical considerations. An examination of equations (42), (45) and (46) shows that the ψ integrals $\mathcal{J}_1$, $\mathcal{J}_2$ are of the form

$$\mathcal{J}_{1,2} = \int_C \begin{matrix}\sin \\ \cos\end{matrix} (\psi - \phi) \exp \left[jn\left(\psi - \frac{\pi}{2}\right) + jk_o\rho\tau \cos (\psi - \phi)\right] d\psi. \tag{48}$$

Since $\mathcal{J}_{1,2}$ are solutions of Helmholtz equation in cylindrical coordinate system, we expect them to be cylindrical functions (compare with Ref. 26, pp. 367–368). The requirement that they be bounded when the argument of the Bessel functions approach zero determines that they are the ordinary J functions. Thus we obtain,

$$\mathcal{J}_1 = 2\pi n(k_o\rho\tau)^{-1}J_n(k_o\rho\tau)\exp(jn\phi), \tag{49}$$

$$\mathcal{J}_2 = -j2\pi J_n'(k_o\rho\tau)\exp(jn\phi). \tag{50}$$

Substitution of equations (42), (45), (46), (48), (49) and (50) into equation (47) yields the following expression for the ϕ component of the reflected field

$$E_\phi^{(r)}(\rho, \phi, z) = 2\pi \sum_{n=-\infty}^{+\infty} I_n z_n(\rho, z)\exp(jn\phi), \tag{51}$$

where

$$\begin{aligned} z_n(\rho, z) = (4\pi)^{-1}k_o^2 b\eta_o \int_0^\infty & [n^2(k_o^2\rho b)^{-1}J_n(k_o\rho\tau)J_n(k_o b\tau)\tau^{-1}(1-\tau^2)^{\frac{1}{2}}R_\parallel(\tau) \\ & - J_n'(k_o b\tau)J_n'(k_o\rho\tau)\tau(1-\tau^2)^{-\frac{1}{2}}R_\perp(\tau)] \\ & \cdot\exp[-jk_o(z+d)(1-\tau^2)^{\frac{1}{2}}]\,d\tau. \end{aligned} \tag{52}$$

In particular,

$$\begin{aligned} z_n(b, d) = (4\pi)^{-1}k_o^2 b\eta_o \int_0^\infty & [n^2(k_o b)^{-2}J_n^2(k_o b\tau)\tau^{-1}(1-\tau^2)^{\frac{1}{2}}R_\parallel(\tau) \\ & - J_n'^2(k_o b\tau)\tau(1-\tau^2)^{-\frac{1}{2}}R_\perp(\tau)]\exp[-j2k_o d(1-\tau^2)^{\frac{1}{2}}]\,d\tau. \end{aligned} \tag{53}$$

IV. CONDUCTING CIRCULAR LOOP OVER A SEMI-INFINITE MEDIUM

In this section we consider a perfectly conducting thin circular loop of mean radius b situated over the semi-infinite medium (Fig. 1) with its plane parallel to the interface. The radius of the wire forming the loop is a and the height of the loop above the interface is d.

4.1 *Current Distribution*

Let $E_\phi^{(o)}(\phi)$ be the ϕ component of the applied electric field, $I^{(o)}(\phi)$ the current distribution that would be created by $E_\phi^{(o)}(\phi)$ on the loop if it were situated in free space and $I(\phi)$ the current distribution caused by $E_\phi^{(o)}(\phi)$ when the loop is situated as shown in Fig. 1. Typical examples

of $E_\phi^{(o)}(\phi)$ are the slice generator (used in admittance computations) and the ϕ component of the electric field that would exist at the loop location in the absence of the loop (as in scattering problems).

The relation between the applied tangential electric field and the currents induced on a loop takes the form of coupled integral equations which are extremely difficult to solve.[6] However if the loop is "thin" ($a \ll b$, $a \ll \lambda$) the current distributions on the loop is given accurately by the so-called "one-dimensional integral equation." Thus we have

$$E_\phi^{(o)}(\phi) = \int_0^{2\pi} G(\phi - \phi') I^{(o)}(\phi')\, d\phi', \tag{54}$$

where

$$G(x) = j(4\pi)^{-1}\eta_o \left[k_o b \cos x + (k_o b)^{-1} \frac{\partial^2}{\partial x^2}\right](2\pi)^{-1} \cdot \int_{-\pi}^{+\pi} \{R(x) \exp[-jk_o R(x)]\}_{a\to 2a\sin\theta}\, d\theta, \tag{55}$$

$$R(x) = b[4\sin^2(x/2) + (a^2/b^2)]^{\frac{1}{2}}. \tag{56}$$

A formal solution to equation (54) is obtained by using Fourier series representations as follows. Let

$$I^{(o)}(\phi) = \sum_{n=-\infty}^{+\infty} I_n^{(o)} \exp(jn\phi), \tag{57}$$

$$E_\phi^{(o)}(\phi) = \sum_{n=-\infty}^{+\infty} \alpha_n^{(o)} \exp(jn\phi), \tag{58}$$

$$G(\phi - \phi') = \sum_{n=-\infty}^{+\infty} \beta_n \exp[jn(\phi - \phi')], \tag{59}$$

where

$$I_n^{(o)} = (2\pi)^{-1} \int_0^{2\pi} I(\phi) \exp(-jn\phi)\, d\phi, \tag{60}$$

$$\alpha_n^{(o)} = (2\pi)^{-1} \int_0^{2\pi} E_\phi^{(o)}(\phi) \exp(-jn\phi)\, d\phi, \tag{61}$$

$$\beta_n = (2\pi)^{-1} \int_0^{2\pi} G(x) \exp(-jnx)\, dx. \tag{62}$$

Substitution of equation (57), (58) and (59) into equation (54) yields the following relation between the Fourier coefficients

$$I_n^{(o)} = (2\pi)^{-1}(\alpha_n^{(o)}/\beta_n). \tag{63}$$

The current distribution is given by

$$I^{(o)}(\phi) = (2\pi)^{-1} \sum_{n=-\infty}^{+\infty} (\alpha_n^{(o)}/\beta_n) \exp(jn\phi). \tag{64}$$

Let $E_\phi^{(1)}(\phi)$ be the tangential electric field, at the loop, of the reflected field whose incident field is caused by $I^{(o)}(\phi)$ and let $I^{(1)}(\phi)$ be the current distribution on the loop caused by $E_\phi^{(1)}(\phi)$. Let

$$E_\phi^{(1)}(\phi) = \sum_{n=-\infty}^{+\infty} \alpha_n^{(1)} \exp(jn\phi), \tag{65}$$

$$I^{(1)}(\phi) = \sum_{n=-\infty}^{+\infty} I_n^{(1)} \exp(jn\phi). \tag{66}$$

Then

$$I_n^{(1)} = (2\pi)^{-1}[\alpha_n^{(1)}/\beta_n] \tag{67}$$

[compare with equation (63)]. Also from equations (51) and (65) we obtain

$$\alpha_n^{(1)} = 2\pi I_n^{(o)} z_n(b, d). \tag{68}$$

Hence

$$I_n^{(1)} = I_n^{(o)}[z_n(b, d)/\beta_n]. \tag{69}$$

In general, if $I_n^{(k)}$ denotes the nth Fourier coefficient of current distribution $I^{(k)}(\phi)$, induced by the kth reflected field we have

$$I_n^{(k)} = I_n^{(k-1)}[z_n(b, d)/\beta_n], \qquad k = 1, 2, \cdots. \tag{70}$$

Let

$$I(\phi) = \sum_{n=-\infty}^{+\infty} I_n \exp(jn\phi). \tag{71}$$

Then

$$I_n = \sum_{k=0}^{\infty} I_n^{(k)}, \tag{72}$$

that is,

$$\begin{aligned} I_n &= I_n^{(o)} + I_n^{(1)} + I_n^{(2)} + \cdots, \\ &= I_n^{(o)}\{1 + [z_n(b, d)/\beta_n] + [z_n(b, d)/\beta_n]^2 + \cdots\}, \\ &= I_n^{(o)}\{1 - [z_n(b, d)/\beta_n]\}^{-1}, \quad \text{provided} \quad |[z_n(b, d)/\beta_n]| < 1. \end{aligned} \tag{73}$$

Henceforth, we assume that $|z_n(b, d)/\beta_n| < 1$. Substituting for $I_n^{(o)}$

in equation (73) from equation (63) we obtain

$$I_n = (2\pi)^{-1}\alpha_n^{(o)}[\beta_n - z_n(b, d)]^{-1}, \tag{74}$$

$$I(\phi) = (2\pi)^{-1} \sum_{n=-\infty}^{+\infty} \alpha_n^{(o)}[\beta_n - z_n(b, d)]^{-1} \exp jn\phi. \tag{75}$$

The following expression for β_n is obtained by simplifying equation (62):

$$\beta_n = j(4\pi k_o b^2)^{-1}\eta_o(2/\pi) \cdot \int_0^{\pi/2} \{\tfrac{1}{2}(k_o b)^2[M_{n+1}(a) + M_{n-1}(a)] - n^2 M_n(a)\}_{a\to 2a\sin\theta}\, d\theta, \tag{76}$$

where

$$M_n(x) = \frac{1}{\pi}\int_0^{\pi/2} (\sin^2\theta + x^2/4b^2)^{-\frac{1}{2}} \cdot \exp[-j2k_o b(\sin^2\theta + x^2/4b^2)^{\frac{1}{2}}] \cos(2n\theta)\, d\theta. \tag{77}$$

Let

$$K_n = (2/\pi)\int_0^{\pi/2} M_n(a)\Big|_{a\to 2a\sin\theta} d\theta. \tag{78}$$

Then

$$\beta_n = j(4\pi k_o b^2)^{-1}\eta_o[\tfrac{1}{2}(k_o b)^2(K_{n+1} + K_{n-1}) - n^2 K_n]. \tag{79}$$

4.2 *Input Admittance*

Let the primary source be a slice generator (delta function source) of voltage V located at $\phi = 0$. That is

$$E_\phi^{(o)}(\phi) = [V\delta(\phi)/b]. \tag{80}$$

Substituting equation (80) into equation (61), we obtain

$$\alpha_n^{(o)} = (V/2\pi b). \tag{81}$$

Substitution of equation (81) into equation (75) yields

$$I(\phi) = V(4\pi^2 b)^{-1} \sum_{n=-\infty}^{+\infty} [\beta_n - z_n(b, d)]^{-1} \exp(jn\phi). \tag{82}$$

The admittance Y at the input terminals at $\phi = 0$ is given by

$$Y = I(0)/V = (4\pi^2 b)^{-1} \sum_{n=-\infty}^{+\infty} [\beta_n - z_n(b, d)]^{-1}. \tag{83}$$

The use of the delta function generator will give rise to an infinite input admittance[6] so that the series in equation (83) is divergent. However this difficulty is overcome by computing the difference between the admittances of two loops of different radii.[27]

V. SPECIAL CASES

5.1 *The Magnetic Dipole Over a Semi-Infinite Medium*

When the radius, b, of the loop becomes very small compared to the wavelength (that is, $k_o b \ll 1$) the current distribution on the loop becomes uniform. This enables us to retain only the zeroth terms in the infinite series representing the different quantities of interest. The field of a dipole over ground is well discussed in the literature and will not be considered here. The input impedance of a dipole over a semi-infinite medium is of considerable interest and may be obtained from equation (83). Thus we obtain

$$Z_{in} = 4\pi^2 b[\beta_o - z_o(b, d)]. \tag{84}$$

The term $4\pi^2 b\beta_o$ represents the input impedance of the loop in the absence of the semi-infinite medium and the term $-4\pi^2 b z_o$ represents the contribution of the semi-infinite medium. That is

$$Z_{in} = Z_{pri} + Z_{sec}, \tag{85}$$

where

$$Z_{pri} = 4\pi^2 b\beta_o, \tag{86}$$

$$\begin{aligned} Z_{sec} &= -4\pi^2 b z_o(b, d) \\ &= \pi(k_o b)^2 \eta_o \int_0^\infty J_1^2(k_o b\tau)\tau(1 - \tau^2)^{-\frac{1}{2}} R_\perp(\tau) \\ &\qquad \cdot \exp\left[-j2k_o d(1 - \tau)^{\frac{1}{2}}\right] d\tau. \end{aligned} \tag{87}$$

5.2 *Thin Circular Loop Over a Perfectly Conducting Plane*

Let

$$z_n^*(b, d) = \lim_{\sigma \to \infty} z_n(b, d). \tag{88}$$

When $\sigma \to \infty$ the reflection coefficients simplify to

$$R_\parallel(\tau) = +1, \qquad R_\perp(\tau) = -1. \tag{89}$$

Substituting equation (89) into equation (53), we obtain

$$z_n^*(b, d) = (4\pi)^{-1}k_o^2 b\eta_o \cdot \int_0^\infty [n^2(k_o b)^{-2}J_n^2(k_o b\tau)\tau^{-1}(1 - \tau^2)^{\frac{1}{2}} + J_n'^2(k_o b\tau)\tau(1 - \tau^2)^{-\frac{1}{2}}] \cdot \exp[-j2k_o d(1 - \tau^2)^{\frac{1}{2}}]\, d\tau. \quad (90)$$

The integral in equation (90) may be simplified, by making use of the properties of Bessel functions, and yields

$$z_n^*(b, d) = (4\pi b)^{-1}\eta_o[\tfrac{1}{2}(k_o b)^2(\zeta_{n-1} + \zeta_{n+1}) - n^2\zeta_n], \quad (91)$$

where

$$\zeta_n = \int_0^\infty \tau(1 - \tau^2)^{-\frac{1}{2}}J_n^2(k_o b\tau) \exp[-j2k_o d(1 - \tau^2)^{\frac{1}{2}}]\, d\tau. \quad (92)$$

An alternate expression for ζ_n is obtained as follows:

$$J_n^2(k_o b\tau) = \frac{1}{\pi}\int_0^\pi J_o(2k_o b\tau \sin\theta) \cos(2n\theta)\, d\theta; \quad (93)$$

substituting equation (93) into equation (92) and changing the order of integration we obtain

$$\zeta_n = j(k_o b)^{-1}M_n(2d), \quad (94)$$

where $M_n(x)$ is defined by equation (77). Therefore,

$$z_n^*(b, d) = j(4\pi k_o b^2)^{-1} \cdot \eta_o\{(\tfrac{1}{2})(k_o b)^2[M_{n+1}(2d) + M_{n-1}(2d)] - n^2M_n(2d)\}. \quad (95)$$

A comparison of equations (95) and (76) reveals a strong similarity between the expressions for β_n and $z_n^*(b, d)$. The input admittance of a thin circular loop over a perfect ground plane is given by

$$Y = (4\pi^2 b)^{-1}\sum_{n=-\infty}^{+\infty}[\beta_n - z_n^*(b, d)]^{-1}. \quad (96)$$

The above formula agrees, with that derived by Iizuka and others,[20] for the input admittance $Y^{(a)}$ of a loop in the presence of an identical coaxial loop carrying a current distribution which has an opposite phase. They use the simpler Kernel given by Storer[5] to computer β_n, a procedure satisfactory for small loops. However they make use of the similarity between β_n and $z_n^*(b, d)$ to compute the latter using the approximate expressions given by Storer. This procedure will yield erroneous results for large separation, d, for the following reason.

The approximate expressions for β_n given by Storer, or for that matter Wu, are valid for $k_o a \ll 1$. The corresponding condition to be

imposed in the evaluation of $z_n^*(b, d)$ is, $(2k_o d) \ll 1$. Thus it is seen that the approximate expressions of Storer give accurate values of $z_n^*(b, d)$ only for very small separations.

VI. NUMERICAL COMPUTATIONS

Equation (76) was further analyzed and approximate expressions for the β_n coefficients were derived in terms of Bessel and Legendre functions. The integral defining the z_n coefficients could not be expressed in terms of known functions and so was evaluated by numerical integration. However, in the evaluation of z_n^* it was possible to use some of the formulae developed for β_n for small values of d [compare with equation (95)]. The numerical integration was carried out by using the Romberg integration scheme. All the computations were done by FORTRAN programs on a GE-635 computer.

Figure 4 shows the variation of the input admittance of a loop, when it is in free space, over moist earth and over an infinitely con-

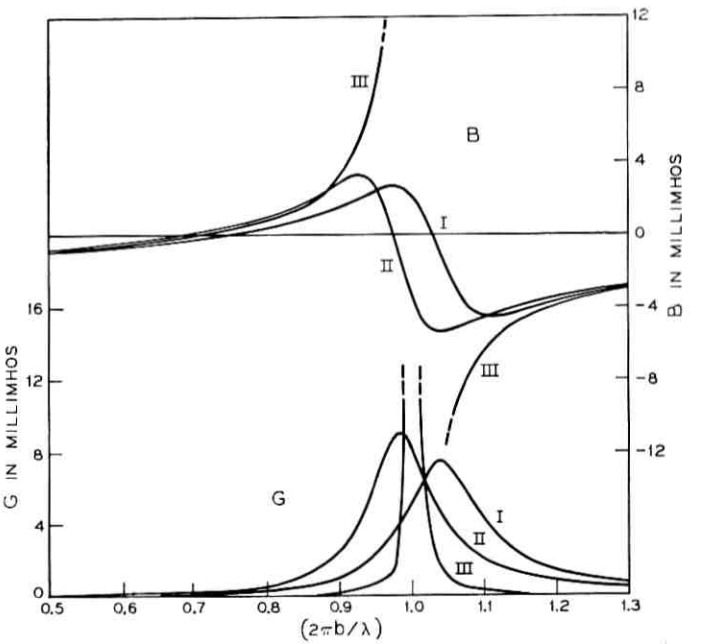


Fig. 4—Input admittance of a circular loop, over different media, as a function of frequency ($d/b = 0.25$; $a/b = 0.002$). G—conductance; B—susceptance. I—loop in free space; II—loop over moist earth; III—loop over perfect ground ($\sigma \to \infty$).

ductive ground plane, as a function of frequency. The values of the various parameters used in these calculations are $2\pi b = 30$ meters, $f = 5$ MHz to 13 MHz, $d/b = 0.25$, $a/2b = 0.001$, $\sigma = 5$ millimhos/meter, and $\epsilon/\epsilon_0 = 15$. The last two parameters characterize the moist earth. The frequency range was deliberately chosen so that the moist earth cannot be approximated either as a highly conductive medium (low frequency approximation) or as a lossless dielectric (high frequency approximation).

The real part G of the input admittance shows its characteristic peak near $k_0 b = 1$ (these occur for values of $k_0 b$ near 1, 2, 3, $\cdots$) but the exact location of the peak as well as its magnitude depends on the medium below the loop. When the loop is located above a highly conducting ground plane the resonance is particularly sharp at $k_0 b = 1$ since at this frequency the loop and its image are exactly half wavelength apart. The imaginary part B of the input admittance changes from inductive to capacitive near $k_0 b = 0.7$ and back to inductive near $k_0 b = 1$. Here again the transition at $k_0 b = 1$ for the loop over a highly conducting ground plane is almost discontinuous.

Figure 5 shows the variation of the input admittance of a loop over a highly conductive ground plane as a function of the distance between the loop and the ground plane. The curves are plotted for $k_0 b = 1$, $a/2b = 0.001$ and d/b ranging from 0.25 to 5.0. It is observed that as d/b increases, the input admittance approaches the free space value in an oscillatory manner.

The aforementioned calculations are presented only as examples of the different types of investigations that may be carried out based on the theory developed. The computer programs developed in this connection are very general and may be used for computations of loops as large as $k_0 b = 10$.

VII. SUMMARY

The problem of a thin, perfectly conducting, circular loop situated in free space over a semi-infinite homogeneous isotropic medium was solved. Expressions for the current distribution on the loop caused by an arbitrary time harmonic source [equation (75)] and the input admittance [equation (83)] were derived. The results are applied to special cases to evaluate the input impedance of a vertical magnetic dipole over a semi-infinite medium [equation (84)] and the input admittance of a circular loop over a perfectly conducting ground plane [equation (96)]. Some numerical results are also given. The analysis for the general

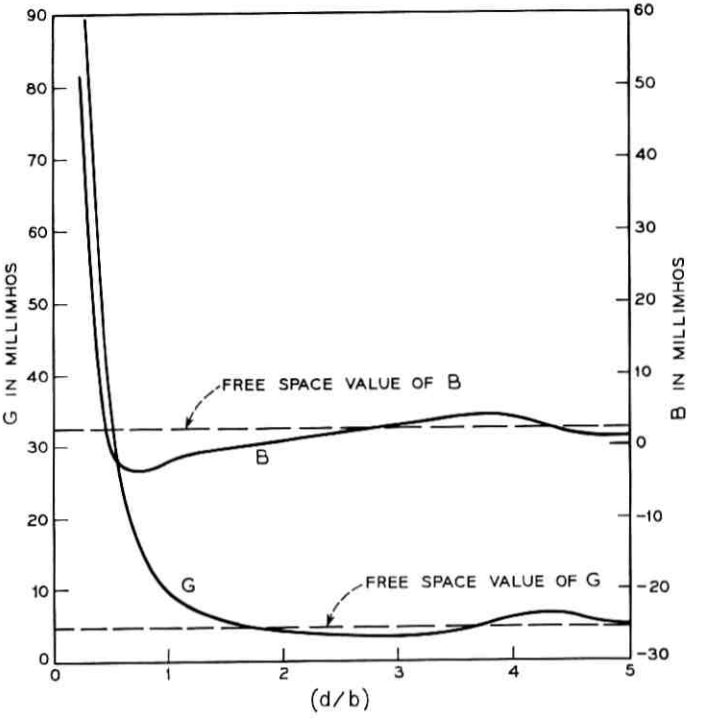


Fig. 5—Input admittance of a circular loop over an infinitely conducting ground plane as a function of height ($k_0 b = 1$, $a/b = 0.002$). G—conductance; B—susceptance.

case of a circular loop in an arbitrary homogeneous medium (as opposed to free space assumed here) over another homogeneous medium can be easily done by modifying the parameters.

VIII. ACKNOWLEDGMENTS

The author wishes to thank Messrs. D. A. Alsberg, J. W. Carlin, G. T. Hawley, and R. D. Tuminaro for encouragement and numerous helpful comments. Computational assistance provided by S. Renna is gratefully acknowledged.

REFERENCES

1. Karzas, W. J., and Latter, R., "Detection of the Electromagnetic Radiation from Nuclear Explosions in Space," Phys. Rev., *137*, No. 5B (March 1965), pp. 1369–1378.
2. Pocklington, H. C., "Electrical Oscillations in Wires," Proc. Camb. Phil. Soc., *9*, (1897), pp. 324–332.

3. Oseen, C. W., "Uber die Electromagnetishe Spektrum einen Dunnen Ringes," Ark. Mat. Astr. Fys., *9*, (1913), pp. 1–34.
4. Hallèn, E., "Theoretical Investigation into Transmitting and Receiving Qualities of Antennae," Nova. Acta. Uppsala., *2*, No. 4 (November 1938), pp. 1–44.
5. Storer, J. E., "Impedance of Thin Wire Loop Antennas," Trans. AIEE, *75*, No. 11, part 1 (November 1956), pp. 606–619.
6. Wu, T. T., "Theory of the Thin Circular Loop Antenna," J. Math. Phys., *3*, No. 6 (December 1962), pp. 1301–1304.
7. Adachi, S., and Mushiake, Y., "Theoretical Formulation for Circular Loop Antennas by Integral Equation Method," Sci. Rep. Res. Inst. Tohoku University, *9*, No. 1, Series B (Elec. Comm.) (June 1957), pp. 9–18.
8. Adachi, S., and Mushiake Y., "Studies of Large Circular Loop Antennas," Sci. Rep. Res. Inst. Tohoku University, *9*, No. 2, series B (Elec. Comm.) (September 1957), pp. 79–103.
9. Mei, K. K., "On the Integral Equation of Thin Wire Antennas," IEEE Trans. on Antennas and Propagation, *AP-13*, No. 3 (May 1965), pp. 374–378.
10. Baghdasarian, A., and Angelakos, D. J., "Scattering from Conducting Loops and Solution of Circular Loop Antennas by Numerical Methods," Proc. IEEE, *53*, No. 8 (August 1965), pp. 818–822.
11. Tang, C. H., "Input Impedance of Arc Antennas and Helical Radiators," IEEE Trans. on Antennas and Propagation, *AP-12*, No. 1 (January 1964), pp. 2–9.
12. Kouyoumjian, R. G., "Backscattering from a Circular Loop," Appl. Sci. Res., *6*, section B (1956), pp. 165–179.
13. Iizuka, K., "The Circular Loop Antenna Multiloaded with Positive and Negative Resistors," IEEE Trans. on Antennas and Propagation, *AP-13*, No. 1 (January 1965), pp. 7–20.
14. Harrington, R. F., and Ryerson, J. L., "Electromagnetic Scattering by Loaded Wire Loops," Radio Science, *1* (New Series), No. 3 (March 1966), pp. 347–352.
15. Harrington, R. F., and Mautz, J., "Electromagnetic Behavior of Circular Wire Loops with Arbitrary Excitation and Loading," Proc. IEE, *115*, No. 1 (January 1968), pp. 68–77.
16. Kraichman, M. B., "Impedance of a Circular Loop in an Infinite Conducting Medium," J. Res. NBS, *66D*, No. 4 (August 1962), pp. 499–503.
17. Chen, C. L., and King, R. W. P., "The Small Bare Loop Antenna Immersed in a Dissipative Medium," IEEE Trans. on Antennas and Propagation, *AP-11*, No. 3 (May 1963), pp. 266–269.
18. King, R. W. P., Harrison, Jr., C. W., and Tingley, D. G., "The Admittance of Bare Circular Loop Antennas in a Dissipative Medium," IEEE Trans. on Antennas and Propagation, *AP*-12, No. 4 (July 1964), pp. 434–438.
19. Galejs, J., "Admittance of Insulated Loop Antennas in a Dissipative Medium," IEEE Trans. on Antennas and Propagation, *AP-13*, No. 2 (March 1965), pp. 229–235.
20. Iizuka, K., King, R. W. P., and Harrison, Jr., C. W., "Self and Mutual Admittances of Two Identical Circular Loop Antennas in a Conducting Medium and in Air," IEEE Trans. on Antennas and Propagation, *AP-14*, No. 4 (July 1966), pp. 440–450.
21. Banōs, Jr., A., *Dipole Radiation in the Presence of a Conducting Half-Space*, New York: Pergamon Press, 1966.
22. Wait, J., "Mutual Electromagnetic Coupling of Loops Over a Homogeneous Earth," Geophysics, *20*, No. 3 (July 1955), pp. 630–637.
23. Wait, J., "The Magnetic Dipole Over a Horizontally Stratified Earth," Canadian J. Phys., *29*, No. 6 (November 1951), pp. 577–592.
24. Sinha, A. K., and Bhattacharya, P. K., "Vertical Magnetic Dipole Buried Inside a Homogeneous Earth," Radio Science, *1* (New Series), No. 3 (March 1966), pp. 379–395.
25. Clemmow, P. C., *The Plane Wave Spectrum Representation of Electromagnetic Fields*, New York: Pergamon Press, 1966.
26. Stratton, J. A., *Electromagnetic Theory*, New York: McGraw-Hill, 1941.
27. Moorthy, S. C., unpublished work.

Contributors to This Issue

W. E. Beadle, B.S.(M.E.), 1954, M.S.(E.E.), 1958, Montana State University; Electrical Engineer, 1963, Stanford University; Research Associate, Montana State University, 1958–1961; Bell Telephone Laboratories, 1963—. At Bell Laboratories, Mr. Beadle has been working primarily in the area of semiconductor device technology. He initially worked on germanium microwave transistor development. More recently his work has been on the silicon diode arrays for optical image sensing. He is Supervisor, Integrated Circuits Department. Member, IEEE, Sigma Xi, Tau Beta Pi.

E. R. Berlekamp, B.S. and M.S. (electrical engineering), 1962, from Massachusetts Institute of Technology on the cooperative program with Bell Telephone Laboratories; Ph. D., 1964, Massachusetts Institute of Technology; taught at the University of California, Berkeley, 1964–1967; Bell Telephone Laboratories, 1967—. Mr. Berlekamp has been engaged in research in algebraic coding theory and related combinatorial mathematics. Member, IEEE, American Mathematical Association, editorial boards of *Information and Control* and the *American Mathematical Monthly*.

Robert W. Chang, B.S.E.E., 1955, National Taiwan University; M.S.E.E., 1960, North Carolina State University; Ph.D., 1965, Purdue University; Bendix Corporation, 1960–1963; Bell Telephone Laboratories, 1965—. Mr. Chang has worked on a variety of problems in data transmission and communication system theory. Member, Phi Kappa Phi, Eta Kappa Nu, Sigma Xi, Association for Computing Machinery, IEEE.

D. B. Fraser, B.Sc. (Hons.), 1954, University of Manitoba; M.A., 1955, and Ph.D., 1958, University of Toronto; Bell Telephone Laboratories, 1959—. At Bell Laboratories, Mr. Fraser has studied materials for ultrasonic devices and holographic storage. He is now engaged in studies of sputtered thin films and ferroelectric ceramic studies related to optical display devices. Member, American Physical Society, A.A.A.S.

B. GOPINATH, M.S. (mathematical physics), 1964, University of Bombay, India; M.S.E.E. and Ph.D.(E.E.), 1968, Stanford University; postdoctoral research associate, Stanford, 1967–1968; Bell Telephone Laboratories, 1968—. Mr. Gopinath's primary interest, as a member of the Systems Theory Research Group, is in the applications of mathematical methods to physical problems.

D. A. GRAY, B.S.E.E., 1963, Tufts University; M.S.E.E., 1965, and Ph.D.(E.E.), 1969, Stanford University; Bell Telephone Laboratories, 1969—. Mr. Gray is working on studies of the propagation of millimeter waves through rainfall. Member, Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

R. R. LAANE, B.S.E.E., 1962, University of Illinois; M.S.E.E., 1964, New York University; Bell Telephone Laboratories, 1962—. Mr. Laane has worked on the application of superconductive switches to data processing systems and has investigated the application of optical processing techniques to data processing systems. Since 1967, he has been engaged in exploratory work on the application of semiconductor devices to telephone switching networks and on analog-to-digital conversion techniques. He is presently also working on solid-state *Picturephone*® switching networks. Member, IEEE.

J. A. LEWIS, B.S., 1944, Worcester Polytechnic Institute; M.S., 1948, and Ph.D., 1950, Brown University; Bell Telephone Laboratories, 1951—. Mr. Lewis has worked on problems in piezoelectricity, elasticity, and heat conduction. He is currently concerned with semiconductor problems. Member, American Mathematical Society, Society for Industrial and Applied Mathematics, American Institute of Aeronautics and Astronautics.

MRS. F. J. MACWILLIAMS, B.A., 1939, and M.A., 1941, Cambridge University, England; Ph.D., 1962, Harvard University; Bell Telephone Laboratories, 1956—. Mrs. MacWilliams has worked in transmission networks development and data communications engineering, and is now in mathematics research. Member, Mathematical Association of America, American Mathematical Society.

J. R. MALDONADO, D.C.F.M., 1961, University of Havana, Cuba; Ph.D., 1968, University of Maryland; Bell Telephone Laboratories, 1968—. Since joining Bell Laboratories, Mr. Maldonado has worked with ferroelectric ceramic materials for optical device applications. Member, American Physical Society, Sigma Pi Sigma, Sigma Xi.

HANS G. MATTES, B.S.E.E., 1964, California Institute of Technology; M.S.E.E., 1966, and Ph.D., 1968, University of Southern California; Bell Telephone Laboratories, 1968—. Mr. Mattes is engaged in the development of substrates for integrated electronics. Member, Sigma Xi.

A. H. Meitzler, B.S., 1951, Muhlenberg College; M.S., 1953, and Ph.D., 1955, Lehigh University; Bell Telephone Laboratories, 1955—. Mr. Meitzler has worked in the areas of ultrasonic devices, acoustic losses in solids, piezoelectric and ferroelectric transducer materials, and most recently, in the area of display devices using ferroelectric materials. Member, American Physical Society, IEEE, Acoustical Society of America, Sigma Xi.

S. C. MOORTHY, B. S., 1961, Kerala University, India; M. S., 1963, and Ph. D., 1966, University of Pennsylvania; Bell Telephone Laboratories, 1967—. He has worked in phased-array antennas and electromagnetic pulse effects and is currently doing research in the area of millimeter waveguides. Member, I.E.E.E, American Physical Society, Sigma Xi.

G. S. MOSCHYTZ, M.S.E.E., 1958, and Ph.D., Electrical Engineering, 1960, Federal Institute of Technology, Zurich, Switzerland; Bell Telephone Laboratories, 1963—. Since joining Bell Laboratories, Mr. Moschytz has investigated methods of synthesizing linear and digital circuits to be used in data transmission equipment that can be microminiaturized by combining thin film and silicon integrated circuit elements. This work has included the design of active RC data modems and filter schemes suitable for hybrid integrated circuit implementation as well as the design of silicon integrated linear and digital devices. He is Supervisor of the Active Filter Group in the Data Communications Laboratory. Member, IEEE.

B. T. MURPHY, B.S., 1953, Ph.D., 1959, University of Leeds, England; Bell Telephone Laboratories, 1963—. Mr. Murphy worked in the field of medical physics at the University of Leeds, on electron beam studies at Mullard Research Laboratories, and since 1959 has been engaged in work on semiconductor devices. At Bell Laboratories, he has worked on semiconductor device modeling, the circuit and the structural aspects of semiconductor integrated circuits, and millimeter wave IMPATT diodes. He is Head of the Exploratory Device Department. Member, IEEE, American Physics Society.

V. K. PRABHU, B.E. (Dist.), 1962, Indian Institute of Science, Bangalore, India; S.M., 1963, and Sc.D., 1966, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1966—. Mr. Prabhu has been concerned with various theoretical problems in solid-state microwave devices, noise, and optical communication systems. Member, IEEE, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, AAAS.

A. J. SCHORR, B.S.(M.E.), 1960, Carnegie Mellon; M.S.(M.E.), 1964, University of Pittsburgh; Bell Telephone Laboratories, 1964—. Mr. Schorr has studied glass-metal systems for miniature diodes, analyzed thermal conductance of multi-material structures, performed thermal conductivity studies, and has been involved in the numerical solution of problems relating to the characterization of semiconductor devices. Member, Pi Tau Sigma.

M. M. SONDHI, B.S. (Honours), 1950, Delhi University (Delhi, India); D.I.I.Sc., 1953, Indian Institute of Science (Bangalore, India); M.S., 1955, and Ph.D., 1957, University of Wisconsin; Bell Telephone Laboratories, 1962—. Mr. Sondhi is working on problems concerning the processing and transmission of speech signals and modeling the detection of auditory and visual signals by human beings.

E. WASSERSTROM, B.S., 1956, M.S., 1960, Technion-Israel Institute of Technology; Ph.D., 1964, Brown University; Division of Sponsored Research at M.I.T., 1962–1964; Department of Aeronautical Engineering at the Technion, 1964–1968, 1969—; Bell Telephone Laboratories (on leave of absence from the Technion), 1968–1969. He is currently engaged in numerical analysis.