THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLVI OCTOBER 1967 NUMBER 8

Convection and Conduction Cooling of Substrates Containing Multiple Heat Sources

By V. L. HEIN

(Manuscript received April 12, 1967)

An analysis is made of the steady-state temperature distribution in a substrate with heat inputs from multiple sources. The problem is of interest in connection with integrated and thin film circuits mounted on ceramic or glass substrates. In these applications, convective heat transfer is present with either conduction along the leads joining the substrate to the heat sink or conduction to one end of the substrate which is heat sinked. A formal three-dimensional solution is obtained which is evaluated for various geometries, thermal conductivities, coefficients of convection and heat-sinking conditions.

I. INTRODUCTION

The remarkable technologies of beam-leaded and thin film integrated circuits have resulted in a new approach to the physical design of electronic circuits.[1] This approach as shown in Fig. 1 consists of bonding beam-leaded integrated circuits to ceramic or glass substrates containing thin film components and conductors. At present, dissipation of the thermal energy generated in these circuits is one of the most severe limitations and affects both device performance and reliability.

The results of this analysis will provide a better understanding of

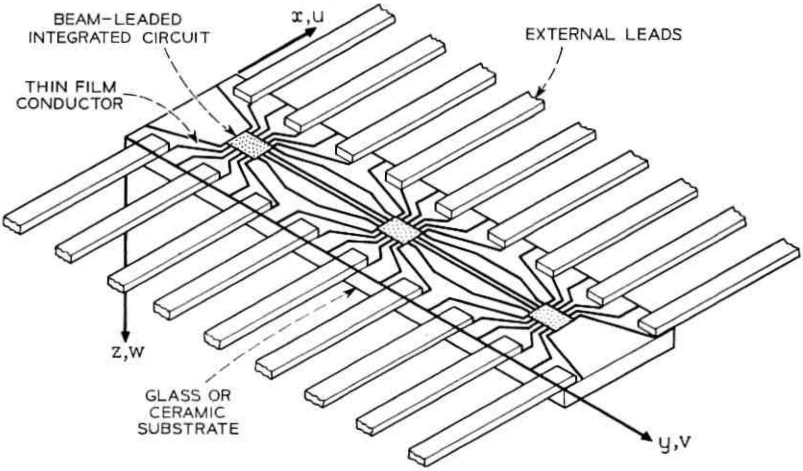


Fig. 1 — Integrated circuit substrate.

the heat transfer phenomena by showing the effects of substrate area, shape, thickness, and thermal conductivity. Included are the effects of the source area, the coefficient of convection and two heat-sinking conditions. The two heat-sinking conditions are: one edge of the substrate is an isothermal boundary, and the leads from the substrate are connected to a heat sink.

Convection is considered on only the two large faces of the substrate since the area of the sides is much smaller. The coefficients of convection are distinguishable for both large faces since some mounting positions will require that these values be different.[2] For instance, the coefficients of convection are quite different for the top and bottom faces of a horizontal plate.

Although the dimensions of most substrates suggest a two-dimensional thin plate model, the very small areas of some heat sources indicate that large thermal gradients in the direction of the normal to the large faces will be present under these sources. This condition is also magnified by the poor thermal conductivity of some substrate materials. Thus, the solution obtained must be three-dimensional to include these effects.

By superposition, multiple sources are considered and the interactions between sources are determined. This provides the necessary information to design the substrate with the desired isolation between temperature sensitive components.

II. MATHEMATICAL MODEL AND BOUNDARY CONDITIONS

The geometry of the problem is shown in Fig. 1 for the case where a substrate is connected to a heat sink through leads. The second heat-sinking condition to be considered is easily visualized if the leads are removed and one edge is considered an isothermal boundary. The temperature in a homogeneous solid of thermal conductivity k satisfies Poisson's equation,

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} = \frac{-Q(x, y, z)}{k}, \tag{1}$$

where Q is the source strength per unit volume. Since there are multiple sources in many of the cases to be considered, it is convenient to use the Green's function approach. It is easily shown[3] that the formal solution can be expressed as

$$\begin{aligned} T(x, y, z) = &\int_0^c \int_0^b \int_0^a G(u, v, w \mid x, y, z) Q(u, v, w)\, du\, dv\, dw \\ &+ k \int_0^b \int_0^a GT_w \Big|_{w=0}^{w=c} - TG_w \Big|_{w=0}^{w=c} du\, dv \\ &+ k \int_0^c \int_0^a GT_v \Big|_{v=0}^{v=b} - TG_v \Big|_{v=0}^{v=b} du\, dw \\ &+ k \int_0^c \int_0^b GT_u \Big|_{u=0}^{u=a} - TG_u \Big|_{u=0}^{u=a} dv\, dw, \end{aligned} \tag{2}$$

where a, b, and c are the substrate dimensions in the x, y, and z directions, respectively, and T_u denotes $\partial T/\partial u$. The Green's function $G(x, y, z \mid u, v, w)$ which is symmetric with $G(u, v, w \mid x, y, z)$, satisfies

$$\frac{\partial^2 G}{\partial x^2} + \frac{\partial^2 G}{\partial y^2} + \frac{\partial^2 G}{\partial z^2} = \frac{-\delta(x-u)\,\delta(y-v)\,\delta(z-w)}{k}. \tag{3}$$

The boundary conditions for $G(x, y, z \mid u, v, w,)$ and $T(x, y, z)$ are

$$\begin{aligned} T_x &= G_x = 0 & & x = a, \\ T_y &= G_y = 0 & & y = 0 \text{ and } b, \\ kT_z &= h_1 T \text{ and } & kG_z = h_1 G \quad & z = 0, \\ -kT_z &= h_2 T \text{ and } & -kG_z = h_2 G \quad & z = c, \end{aligned}$$

and either

$$T = G = 0 \quad \text{or} \quad T_x = G_x = 0, \qquad x = 0. \tag{4}$$

The boundary conditions reference the ambient temperature to zero, and h_1 and h_2 are the coefficients of convection on the two large faces of the substrate. The choice of the last boundary condition depends on whether the substrate under consideration has an isothermal boundary on one end or is supported by leads. If the substrate has leads, the latter boundary condition is used and the heat lost by conduction along the leads is considered by assuming negative sources[3] located at the lead bonding areas. The magnitude of these sources is determined by noting that the temperature difference between the ends of a lead is given by the product of the lead thermal resistance and the amount of heat conducted along that lead.

The case where the leads conduct an appreciable amount of heat is somewhat more difficult and cumbersome than the case of an isothermal boundary at x equals zero or the case where the lead's thermal resistance is of sufficient magnitude to be approximated by an insulated boundary. In what immediately follows, the three cases will be treated separately and generally only one source is considered. The general case of multiple sources follows by superposition and will be treated last.

2.1 *Substrate With High-Resistance Leads*

The solution to this problem is obtained by taking finite cosine transforms[4] of (3) in the x and y directions. The eigenvalues are chosen to satisfy the boundary conditions given by (4). Taking the double cosine transform gives

$$\bar{\bar{G}}_{zz} - (\alpha^2 + \beta^2)\bar{\bar{G}} = \frac{-\delta(z - w) \cos \alpha u \cos \beta v}{k}, \tag{5}$$

where $\alpha = m\pi/a$ $(m = 0, 1, 2, \cdots)$, $\beta = n\pi/b$ $(n = 0, 1, 2, \cdots)$ and $\bar{\bar{G}}$ denotes the transformed dependent variable G. A third transform of (5) is not taken to avoid a triple summation in the final solution. Although the triple summation is no great obstacle, it increases the number of terms needed for numerical evaluation. Thus, (5) is solved for the cases of

$$\alpha^2 + \beta^2 = 0 \quad \text{and} \quad \alpha^2 + \beta^2 \neq 0.$$

If α and β are both zero, by standard methods[5] the solution to (5), satisfying the boundary conditions given by (4) is

$$\begin{aligned} \bar{\bar{G}}(0, 0, z \mid u, v, w) &= \frac{(k/h_2c - w/c + 1)(z + k/h_1)}{k(k/h_1c + k/h_2c + 1)}, \quad z \leqq w, \\ \bar{\bar{G}}(0, 0, z \mid u, v, w) &= \frac{(k/h_2c - z/c + 1)(w + k/h_1)}{k(k/h_1c + k/h_2c + 1)}, \quad z \geqq w. \end{aligned} \tag{6}$$

For α or β not equal to zero, the solution of (5) leads to

$$\bar{\bar{G}}(\alpha, \beta, z \mid u, v, w) = \frac{\cos \alpha u \cos \beta v}{k\gamma} \int_0^z \delta(\zeta - w) \cdot [\phi(\zeta)\psi(z) - \phi(z)\psi(\zeta)]\, d\zeta + A\phi(z) + B\psi(z),$$

where

$$\phi(z) = \sinh \gamma z, \qquad \psi(z) = \cosh \gamma z$$

and

$$\gamma = (\alpha^2 + \beta^2)^{\frac{1}{2}}. \tag{7}$$

Again by standard methods, (7) leads to

$$\begin{aligned} \bar{\bar{G}}(\alpha, \beta, z \mid u, v, w) &= \cos \alpha u \cos \beta v[\phi(z) + (k\gamma/h_1)\psi(z)] \\ &\cdot \frac{\psi(c)[\psi(w) - (h_2/k\gamma)\phi(w)] - \phi(c)[\phi(w) - (h_2/k\gamma)\psi(w)]}{k\gamma[(h_2/k\gamma + k\gamma/h_1)\phi(c) + (1 + h_2/h_1)\psi(c)]}, \quad z \leqq w, \\ \bar{\bar{G}}(\alpha, \beta, z \mid u, v, w) &= \cos \alpha u \cos \beta v[\phi(w) + (k\gamma/h_1)\psi(w)] \\ &\cdot \frac{\psi(c)[\psi(z) - (h_2/k\gamma)\phi(z)] - \phi(c)[\phi(z) - (h_2/k\gamma)\psi(z)]}{k\gamma[(h_2/k\gamma + k\gamma/h_1)\phi(c) + (1 + h_2/h_1)\psi(c)]}, \quad z \geqq w. \end{aligned} \tag{8}$$

The inversion for the double cosine transforms is easily derived[4] and upon substitution into (2) gives

$$\begin{aligned} T(x, y, z) = \int_0^c \int_0^b \int_0^a \Bigg[&\frac{1}{ab} \bar{\bar{G}}(u, v, w \mid 0, 0, z) \\ &+ \frac{2}{ab} \sum_\alpha \bar{\bar{G}}(u, v, w \mid \alpha, 0, z) \cos \alpha x \\ &+ \frac{2}{ab} \sum_\beta \bar{\bar{G}}(u, v, w \mid 0, \beta, z) \cos \beta y \\ &+ \frac{4}{ab} \sum_\alpha \sum_\beta \bar{\bar{G}}(u, v, w \mid \alpha, \beta, z) \cos \alpha x \cos \beta y \Bigg] Q(u, v, w)\, du\, dv\, dw. \end{aligned} \tag{9}$$

It should be noted that the surface integrals of (2) are identically equal to zero because of the boundary conditions and the subsequent choice of eigenvalues. Consequently, these terms do not appear in (9) and the determination of the temperature distribution merely requires the substitution of the appropriate Green's functions from (6) and (8) and the integrations indicated in (9). For the applications cited, the sources are on the z equals zero surface and therefore $Q(x, y, z) =$

$Q(x, y)\ \delta(z)$. Substituting this into (9) gives

$$T(x, y, z) = \int_0^b \int_0^a \left[\frac{1}{ab} \bar{\bar{G}}(u, v, 0 \mid 0, 0, z) + \frac{2}{ab} \sum_\alpha \bar{\bar{G}}(u, v, 0 \mid \alpha, 0, z) \cos \alpha x + \frac{2}{ab} \sum_\beta \bar{\bar{G}}(u, v, 0 \mid 0, \beta, z) \cos \beta y + \frac{4}{ab} \sum_\alpha \sum_\beta \bar{\bar{G}}(u, v, 0 \mid \alpha, \beta, z) \cos \alpha x \cos \beta y \right] Q(u, v)\, du\, dv. \tag{10}$$

Note that in this case $Q(u, v)$ represents real sources since the leads are assumed to be of very high resistance and thereby produce negligible effects.

2.2 *Substrate With Heat Conducting Leads*

The method of solution for this case makes use of the results up to (10). It has been stated that the leads will be treated as negative heat sources and thus by superposition (10) will read

$$T(x, y, z) = \int_0^b \int_0^a \left[\frac{1}{ab} \bar{\bar{G}}(u, v, 0 \mid 0, 0, z) + \frac{2}{ab} \sum_\alpha \bar{\bar{G}}(u, v, 0 \mid \alpha, 0, z) \cos \alpha x + \frac{2}{ab} \sum_\beta \bar{\bar{G}}(u, v, 0 \mid 0, \beta, z) \cos \beta y + \frac{4}{ab} \sum_\alpha \sum_\beta \bar{\bar{G}}(u, v, 0 \mid \alpha, \beta, z) \cos \alpha x \cos \beta y \right] \cdot [Q(u, v) - F(u, v)]\, du\, dv, \tag{11}$$

where $F(u, v)$ represents the heat sources due to leads. It is important to note that both Q and F can represent any arbitrary number of real sources and leads, respectively. This is important since it is quite common for a substrate to have as many as sixteen or eighteen leads. It is obvious at this point that $F(u, v)$ must be specified. The approach used here is to assume a uniform heat source over each lead bond area and determine the magnitudes of these sources by using other available information. The required information is available from (11) since the temperature can be evaluated at each lead location in terms of each source power density. Thus,

$$F(u, v) = f_1g_1(u, v) + f_2g_2(u, v) + \cdots f_ng_n(u, v) \tag{12}$$

represents the sources due to each lead location and $f_1, f_2 \cdots f_n$ are unspecified at this time. The functions $g_i(u, v)$ are a combination of step functions that give unity at the lead bond area and zero elsewhere. To determine $f_1, f_2, \cdots f_n$, the condition that the temperature difference between the two ends of a lead is equal to the product of the thermal power and thermal resistance of that lead is used. Equating these two expressions for the temperature at each lead location gives n equations and n unknowns in f_i. In matrix notation this can be expressed as

$$[A + R][F] = [B], \tag{13}$$

where A is an n by n matrix representing the influence coefficients due to the action of the negative sources. R is also an n by n matrix but the only non-zero elements are those where $i = j$. These elements are the thermal resistances of the leads. The column matrix F consists of $f_1, f_2, \cdots f_n$ and represents the unknowns to be determined. The matrix B is a column matrix which represents the effects of the real heat sources. By substitution of $A + R = C$, then (13) becomes

$$CF = B,$$

and, therefore, (14)

$$C^{-1}CF = C^{-1}B,$$

where C^{-1} is the inverse of C. Since $C^{-1}C = I$, where I is the unit matrix, then

$$IF = C^{-1}B \tag{15}$$

gives the desired results.[6] Using these values of $f_1, f_2, \cdots f_n$ in (11) permits the calculation of the temperature at any point in the substrate. In the results reported here, the temperature was always evaluated at the center of the lead bond area and the flux was assumed constant over that area. This is a reasonable approximation since the lead material generally has a much higher thermal conductivity than the substrate. However, the problem can be solved for other functional representations of the flux if there is evidence that these representations are significantly better approximations of the physical situations. Similarly, the reported results for substrates with leads will be for a single heat source although $Q(u, v)$ and the matrix B are not restricted to such situations and can represent any arbitrary number of sources. Thus, the method of obtaining a formal solution for the temperature distribution in a substrate containing multiple sources and leads has been indicated.

2.3 *Substrate With Isotherm on One Boundary*

The solution of the temperature problem for the $x = 0$ boundary being an isotherm is obtained by taking a finite sine transform[4] in the x direction and retaining the cosine transform for the y direction. Applying these transforms to (3) gives

$$\bar{\bar{G}}_{zz} - (\alpha^2 + \beta^2)\bar{\bar{G}} = \frac{-\delta(z - w) \sin \alpha u \cos \beta v}{k}, \tag{16}$$

where $\alpha = (2m - 1)\pi/2a$ $(m = 1, 2, 3, \cdots)$, $\beta = n\pi/b$ $(n = 0, 1, 2, \cdots)$. Comparing (16) and (5) indicates the solution to (5) can be used as the solution to (16) provided sin αu is substituted to replace cos αu and α is now given by (16). The solution to (5) was given by (6) for $\alpha^2 + \beta^2 = 0$ and (8) for $\alpha^2 + \beta^2 \neq 0$. Since $\alpha^2 + \beta^2$ is never equal to zero in (16), (6) is not needed and (8) gives the solution provided the indicated changes are made. The inversion formula for these multiple finite transforms gives the temperature distribution as

$$T(x, y, z) = \int_0^b \int_0^a \left[\frac{2}{ab} \sum_\alpha \bar{\bar{G}}(u, v, 0 \mid \alpha, 0, z) \sin \alpha x \right.$$
$$\left. + \frac{4}{ab} \sum_\alpha \sum_\beta \bar{\bar{G}}(u, v, 0 \mid \alpha, \beta, z) \sin \alpha x \cos \beta y \right] Q(u, v)\, du\, dv, \tag{17}$$

where again $G(u, v, w \mid x, y, z) = G(x, y, z \mid u, v, w)$. It should be noted that in (9), (10), (11), and (16) the summations on α and β are only for the non-zero eigenvalues since the $\alpha = 0$ or $\beta = 0$ terms, if they appear, are already indicated in the inversion formulas.

It should be apparent that similar physical situations such as two opposite boundaries being isotherms poses no new or additional problems. For instance if the boundaries $x = 0$ and $x = a$ are isotherms, (17) is valid provided $\alpha = m\pi/a$ $(m = 1, 2, 3, \cdots)$.

III. NUMERICAL RESULTS

In this section results are given for the three cases discussed in Sections 2.1, 2.2, and 2.3. For the results reported, it is assumed that the heat sources have uniform power density. Thus, they are mathematically represented by a combination of step functions. The results are reported in terms of thermal resistance where thermal resistance

(R_t) is defined as the difference in maximum and minimum temperatures based on one watt of power dissipation. Dimensionless variables are used to present the results in their most general form except for several instances where specific results are desired. The dimensionless variables are $\Delta x/a$, x_0/a, c/a, $\Delta y/b$, y_0/b, c/b, $\Delta z/c$, z_0/c, h_1c/k, and h_0c/k, where Δx, Δy, and Δz are the source dimensions and x_0, y_0, and z_0 the coordinates of the source center. For the applications cited, the sources are always plane sources located on one face of the substrate and thus $\Delta z/c = z_0/c = 0$ in (10), (11), and (17) and consequently, in all the results.

Previously, it was suggested that a three-dimensional solution would be necessary to accurately represent the thermal behavior of a low thermal conductivity substrate containing very small heat sources. This is because the temperature gradient in a direction normal to a plane source must increase at the same rate as the source area decreases if the same amount of heat is dissipated. Thus, small sources require large gradients near the source even though the body temperature may be nearly uniform elsewhere as would be expected if the Biot number is small, i.e., $hc/k \ll 1$. The following physical parameters are chosen as a typical example of a substrate with a small centrally located heat source:

$$\begin{aligned} a/c &= 26.0, & b/c &= 10.0, \\ x_0/a &= 0.50, & y_0/b &= 0.50, \\ \Delta x/a &= 0.0123, & \Delta y/b &= 0.0320 \\ h_1c/k &= 0.93 \times 10^{-3}, & h_2c/k &= 0.93 \times 10^{-3}. \end{aligned}$$

Fig. 2 presents the results where the dependent variable is the dimensionless quantity $ckT(x, y, z)$. These results clearly indicate the problem is three dimensional near the source due to spreading resistance and thus a three-dimension solution is required to accurately describe substrates containing small sources.

The next numerical results are for square and rectangular substrates containing one centrally located heat source and leads that conduct a negligible amount of heat. Thus, the leads can be ignored in the analysis and the equations of Section 2.1 are applicable. Obviously, this model always gives an upper bound for the substrate thermal resistance unless the leads are heat sinked at a higher temperature than the atmosphere surrounding the substrate. Except for this unlikely situation, these results give an easy-to-obtain first approximation of the

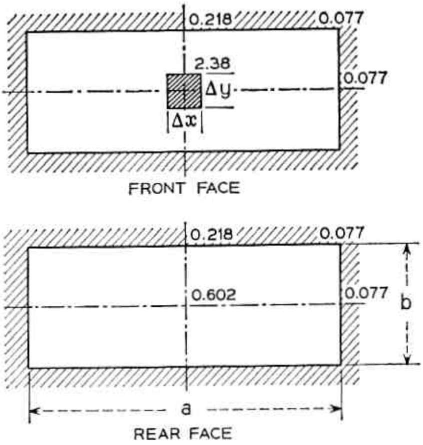


Fig. 2 — Integrated circuit substrate temperature, $ckT(x, y, z)$ as a function of position: $a/c = 26.0$, $b/c = 10.0$, $x_0/a = y_0/b = 0.50$, $\Delta x/a = 0.0123$, $\Delta y/b = 0.0320$, $h_1c/k = h_2c/k = 0.93 \times 10^{-3}$, $P = 1$ watt.

substrate thermal capability and are given in Fig. 3 (a) and (b). Each solid line represents a given substrate whereas each broken line represents a given source. Although the broken lines represent redundant information, they assist in illustrating the effects of various parameters. Thus, Fig. 3(a) and (b) clearly show the effects of heat source and substrate areas.

Results for the case where one end of the substrate is an isothermal boundary are given in Fig. 3(c) and (d). The equation for these results was developed in Section 2.3. By following one of the broken lines, a constant heat source area is being maintained and the effects of changes in the substrate area can be observed. It will be noted that increasing the substrate area provides little benefit since, with a centrally located source, increases in the thermal resistance due to longer conduction paths almost cancel the decreases due to larger convection areas. As with the previous case, it is noted that the square substrate is slightly more efficient than the rectangular one. The Biot number used for the results presented in Fig. 3(a) thru (d) is a typical value for a thin alumina ceramic substrate with free convection from both faces.

Results for a substrate with heat conducting leads are more difficult to generalize since the thermal resistance of the leads, the lead bond areas and locations, and the number of leads are parameters which affect the results. To illustrate the effects of leads, two alumina ceramic substrates containing beam-leaded monolithic integrated circuits are considered. These substrates will be identified as Substrates I and II, and are very similar to the one in Fig. 1 except that both have one heat source. Substrate I is supported by sixteen copper leads, but due to symmetry only one quadrant of the substrate need be considered. Thus, the problem is simplified to a substrate containing one source and four leads. Using the procedure outlined in Section 2.2 gives

$$A = \begin{bmatrix} 608.7 & 500.3 & 476.6 & 466.0 \\ 500.3 & 585.1 & 489.7 & 475.2 \\ 476.6 & 489.7 & 583.7 & 497.1 \\ 466.0 & 475.2 & 497.1 & 602.0 \end{bmatrix}$$

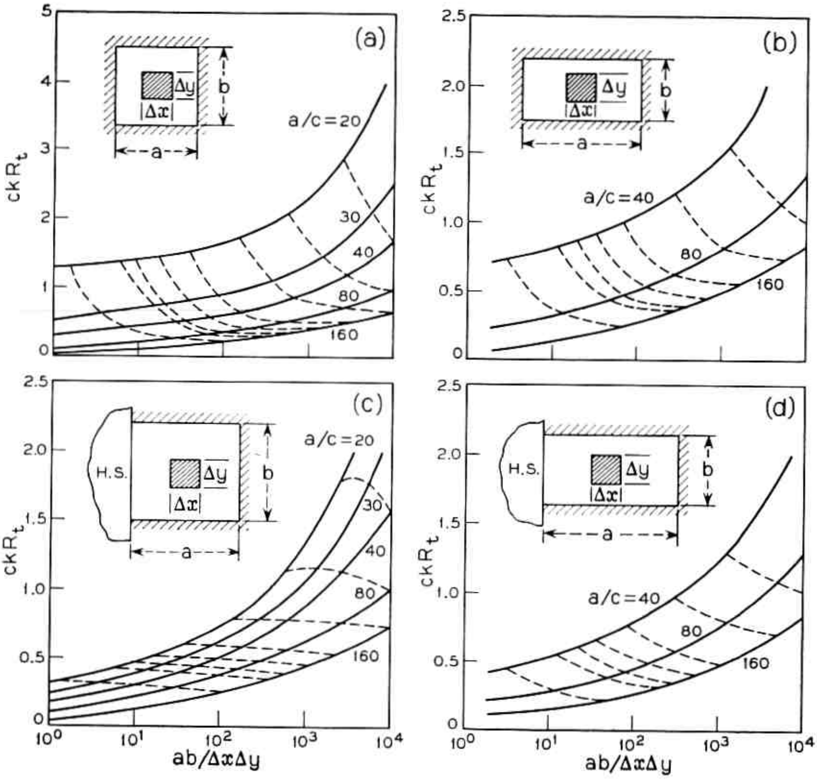


Fig. 3—Substrate thermal resistance: (a) and (c) $a/c = b/c$, $\Delta x/a = \Delta y/b$, $x_0/a = y_0/b = 0.50$, $h_1c/k = h_2c/k = 0.93 \times 10^{-3}$; (b) and (d) $a/c = 2b/c$, $\Delta x/a = \Delta y/2b$, $x_0/a = y_0/2b = 0.50$, $h_1c/k = h_2c/k = 0.93 \times 10^{-3}$.

and

$$B = \begin{bmatrix} 124.2 \\ 121.3 \\ 118.1 \\ 116.1 \end{bmatrix}, \tag{20}$$

where the total output from the heat source has been taken as one-fourth watt since only one-fourth of the real heat source lies in the quadrant being considered. Lead spacing was 0.190 cm (0.075 inch), and each lead bond area was 0.038 cm (0.015 inch) by 0.038 cm. To continue the analysis and determine the f_i's, it is necessary to specify the thermal resistances of the leads as these values are the diagonal elements of the matrix R. To obtain a better understanding of the effects of the leads, it is useful to present the substrate thermal resistance as a function of the lead thermal resistance. These results are given in Fig. 4 and clearly indicate that the substrate thermal resistance can be substantially reduced by heat sinking the leads. Fig. 4 also gives results for 0.107 by 0.107, 0.157 by 0.157, and 0.208 by 0.208 cm square (0.042 by 0.042, 0.062 by 0.062, and 0.082 by 0.082 inches square, respectively) sources to illustrate the effects of source size for substrates with leads. For these physical situations, the change in thermal resistance due to changes in source size is within 1°C/watt of

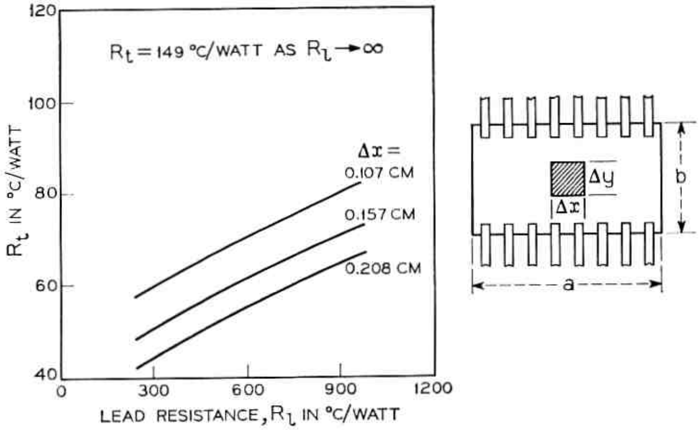


Fig. 4 — Thermal resistance of Substrate I: a = 1.61 cm, b = 0.89 cm, c = 0.0635 cm, $\Delta x = \Delta y$, k = 0.202 watt/cm-°C, $h_1 = h_2$ = 0.003 watt/cm²-°C.

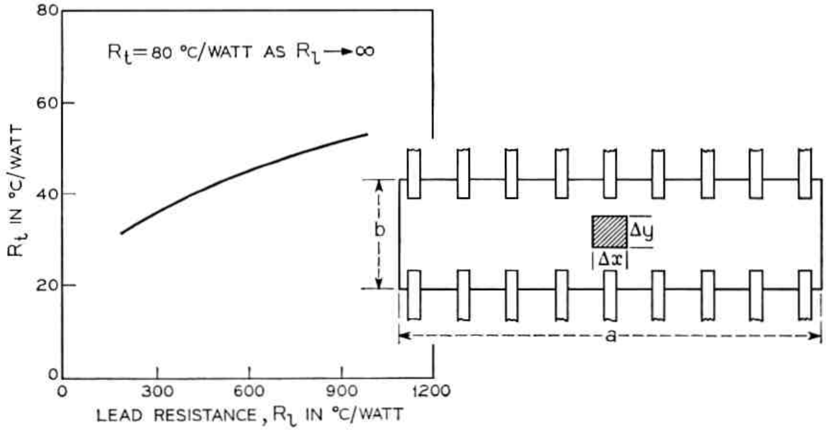


Fig. 5 — Thermal resistance of Substrate II: a = 3.30 cm, b = 0.89 cm, c = 0.0635 cm, $\Delta x = \Delta y$, k = 0.202 watt/cm-°C, $h_1 = h_2$ = 0.003 watt/cm^2-°C.

the results obtained for substrates without leads. This indicates that for small sources the source area effects are local and the effects of moderate changes in source area can be approximated by the results for substrates without heat conducting leads.

The results for Substrate II are given in Fig. 5. This design is very similar to the previous one except the substrate is approximately twice as long, contains eighteen leads on 0.381 cm (0.150 inch) spacing and each lead bond area is 0.038 cm (0.015 inch) by 0.076 cm (0.030 inch). These results indicate a thermal resistance of 80°C/watt for this substrate as compared to 149°C/watt for the previous one if the effects of the leads are not included. These substrate resistances are reduced to 45 and 62°C/watt, respectively, if the thermal resistance of each lead is 600°C/watt. To illustrate the effects of the leads, Fig. 6 shows the percent reduction in the substrate thermal resistance due to each lead. In this figure the leads are numbered beginning with the lead closest to the heat source. Although results for these two designs do not permit a generalization of lead effects, they do illustrate what effects may be expected for geometries that are reasonably similar.

This significant effect of the leads illustrates one of the disadvantages of glass and glazed ceramic substrates. At present most leads bonded to these substrates have had much higher thermal resistances than those bonded to unglazed ceramics. This is because of the inability to repeatedly bond thick copper leads to glass or glazed substrates with-

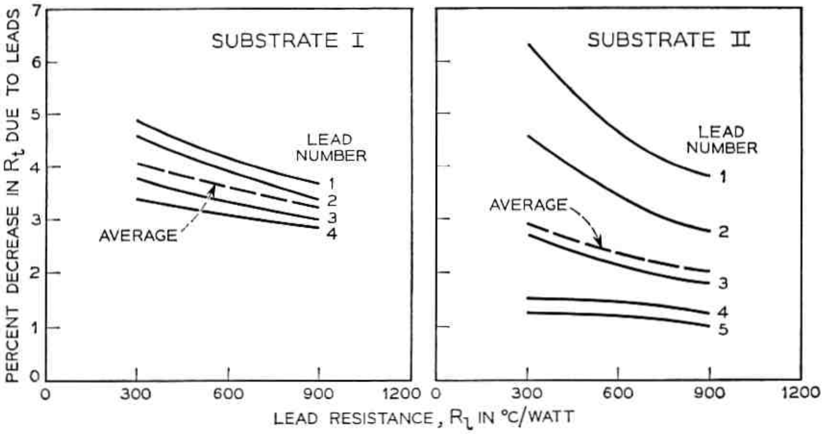


Fig. 6 — Percent decrease in substrate thermal resistance due to leads.

out fractures at the time of bonding or upon subsequent thermal cycling. Thus, the effect of replacing thick copper leads with thin gold ribbon, as is commonly used, can easily result in an additional thermal resistance of 20 to 40°C/watt. These results also suggest that when necessary, ceramic substrates could be glazed in only those areas as required for component performance and thus permit the use of low thermal resistance leads.

A substrate dimension of obvious interest is the thickness. Results considering this parameter are given in Fig. 7(a) and (b) for the indicated rectangular substrates. For the substrate without leads, the change in thermal resistance due to a 40 percent increase in substrate thickness is insignificant whereas for the substrate with an isothermal boundary on one edge a 20 percent improvement in thermal resistance is possible for the small substrate areas. These conclusions are made by comparing the results of Fig. 3(b) and (d) with Fig. 7(a) and (b), where again the Biot number has been chosen to represent a thin alumina ceramic substrate with free convection from both faces.

The coefficient of convection and the substrate thermal conductivity are also parameters that have significant effects on substrate heat transfer characteristics. If the dependent variable is chosen as ckR_t, it is possible to consider the effects of changes in the coefficient of convection and the substrate thermal conductivity by considering various values of the Biot number since these two parameters always appear in this non-dimensional form. However, for the applications cited the thermal conductivities are approximately 0.202 watt/cm-°C for alumina ceramic, 1.35 watt/cm-°C for beryllia ceramic and 0.020 watt/cm-°C for glass while the coefficients of convection range from 0.003 to 0.010 watt/cm²-°C. These coefficients of convection represent the range from free to very moderate forced convection. Since this range is much smaller than the range of thermal conductivities, the results from the evaluation of these two parameters are presented separately. To illustrate the effects of the coefficient of convection, a rectangular substrate without leads is chosen and the results are presented in Fig. 8. As specific examples, the Biot numbers chosen can represent a thin alumina ceramic with free convection in equipment, free convection under laboratory conditions with small substrates and moderate forced convection. These results illustrate the need for adequate air volume and velocity around the substrate since the coefficient of convection[2] is dependent upon both.

All previous results were for Biot numbers in a range applicable to thin alumina ceramic substrates subjected to free or moderate forced convection. The following results will be applicable for beryllia ceramic and glass substrates under the same conditions. Since typical glasses used for thin film substrates have thermal conductivities ten to twenty times smaller than those of alumina ceramics, the results are obviously less favorable. However, if the average temperature term of (10) is considered, it will be noted that this term is most strongly influenced by the reciprocal of the product of the coefficient of con-

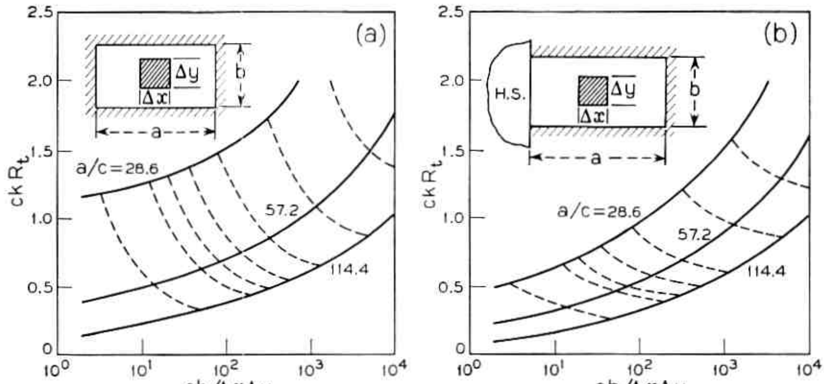


Fig. 7 — Substrate thermal resistance: (a) and (b) $a/c = 2b/c$, $\Delta x/a = \Delta y/2b$, $x_0/a = y_0/b = 0.50$, $h_1c/k = h_2c/k = 1.3 \times 10^{-3}$.

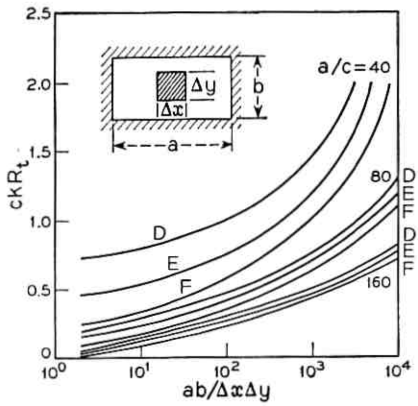


Fig. 8—Substrate thermal resistance for several Biot numbers: $a/c = 2b/c$, $\Delta x/a = \Delta y/2b$, $x_0/a = y_0/b = 0.50$, $h_1c/k = h_2c/k$; (D) $h_1c/k = 0.93 \times 10^{-3}$; (E) $h_1c/k = 1.6 \times 10^{-3}$; (F) $h_1c/k = 3.2 \times 10^{-3}$.

vection and the substrate area. Thus, the substrate thermal conductivity has a minor effect on this term. As the source area-to-substrate area ratio approaches unity, this term is the significant term in the answer since the substrate begins to approach a uniform temperature which is also the average temperature. Thus, the conclusion can be made that if the source area is approximately equal to the substrate

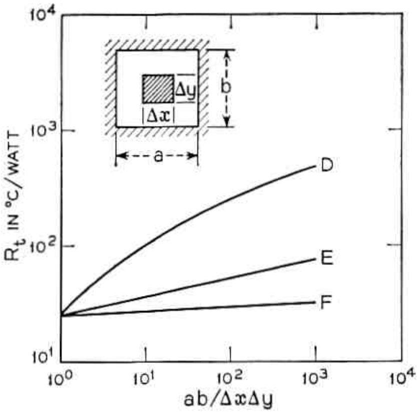


Fig. 9—Thermal resistance of glass, alumina and beryllia ceramic substrates: $a = b = 2.54$ cm, $c = 0.0635$ cm, $\Delta x/a = \Delta y/b$, $x_0/a = y_0/b = 0.50$, $h_1 = h_2 = 0.003$ watt/cm²-°C; (D) $k = 0.020$ watt/cm-°C; (E) $k = 0.202$ watt/cm-°C; (F) $k = 1.35$ watt/cm-°C.

area, the effects of low thermal conductivity substrates will be minimal but if the source area is very small, as is typical with semiconductor sources, the increase in substrate thermal resistance will be most significant. Fig. 9 illustrates these effects for a 2.54 by 2.54 cm (1.0 by 1.0 inch) square substrate.

A direct comparison can be made between alumina and beryllia substrates by comparing the results given in Fig. 10(a) thru (d) with those previously given in Fig. 3(a) thru (d). By choosing several points from the corresponding figures, one can quickly demonstrate the advantage of the high thermal conductivity substrate since the thermal conductivities differ by a factor of 6.7 if the same thickness and coefficient of convection values are used. Specific results giving

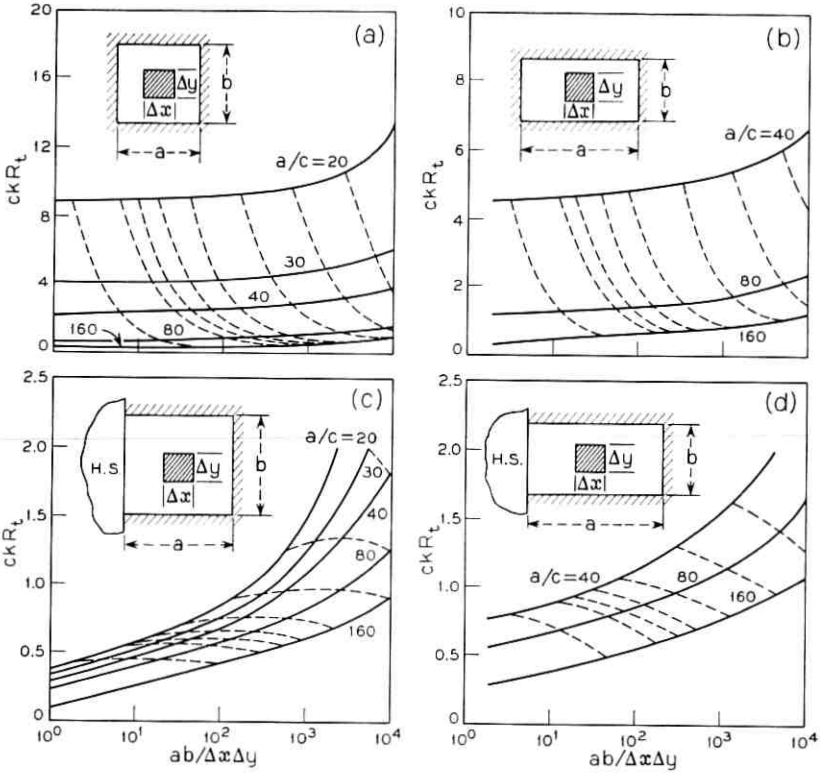


Fig. 10—Substrate thermal resistance: (a) and (c) $a/c = b/c$, $\Delta x/a = \Delta y/b$, $x_0/a = y_0/b = 0.50$, $h_1c/k = h_2c/k = 0.14 \times 10^{-3}$; (b) and (d) $a/c = 2b/c$, $\Delta x/a = \Delta y/2b$, $x_0/a = y_0/b = 0.50$, $h_1c/k = h_2c/k = 0.14 \times 10^{-3}$.

such a direct comparison are given in Fig. 11(a) and (b). These figures demonstrate that for very small sources or for substrates with an isothermal boundary on one edge, the beryllia substrate can make possible approximately a 60°C/watt reduction in the substrate thermal resistance. Since adequate power dissipation is one of the present problems of thin film and integrated circuits, the use of high conductivity substrates should be considered.

The final numerical results illustrate the thermal interaction between sources which must be considered if some components have temperature sensitive parameters. A synchronous clock logic circuit is chosen for this example. One version of this circuit consists of four flat packages appliqued to an alumina substrate and inserted into a socket. The socket design is such that the edge of the substrate which makes contact closely approximates an isothermal boundary. Each flat package contains two or three integrated circuit chips and dissipates approximately one-fourth watt. To complete this analysis the only required additional computation is to calculate the temperatures at other source locations and superimpose these effects. Fig. 12 gives the temperature profile of the substrate based on a one-fourth watt power dissipation from each flat package. It should be emphasized that the temperatures given are referenced to a heat sink and ambient of zero degrees centigrade and that the flat package and integrated circuit chip thermal resistances must be added to the values obtained for

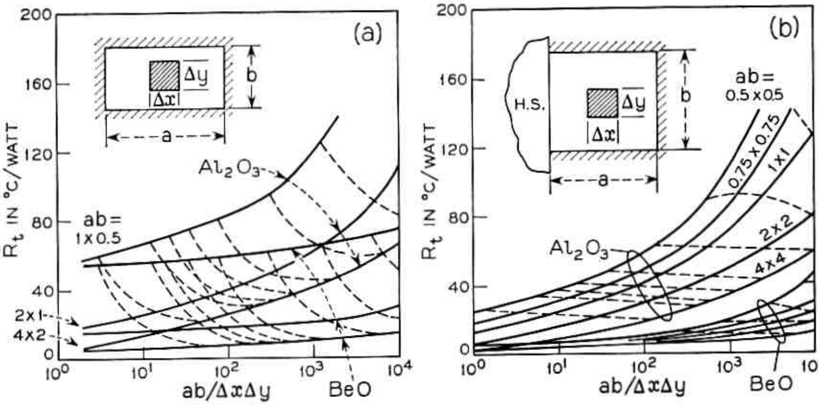


Fig. 11—Substrate thermal resistance: (a) $a/c = 2b/c$, $\Delta x/a = \Delta y/2b$, $x_0/a = y_0/b = 0.50$, $h_1c/k = h_2c/k = 0.14 \times 10^{-3}$, $c = 0.0635$ cm, $k = 0.202$ watt/cm-°C; (b) $a/c = b/c$, $\Delta x/a = \Delta y/b$, $x_0/a = y_0/b = 0.50$, $h_1c/k = h_2c/k = 0.14 \times 10^{-3}$, $c = 0.0635$ cm, $k = 0.202$ watt/cm-°C.

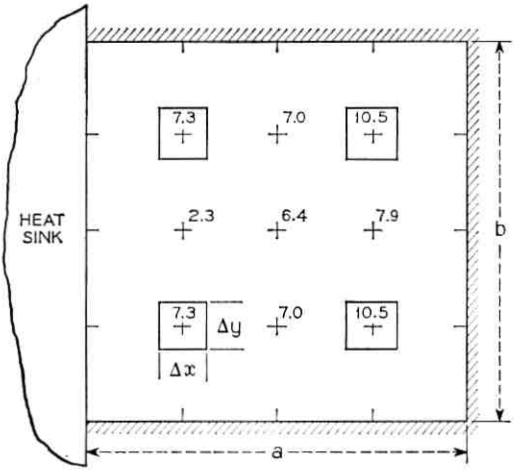


Fig. 12—Temperature profile (°C) per watt for synchronous clock logic circuit: $a/c = b/c = 80$, $\Delta x/a = \Delta y/b = 0.20$, $x_0/a = 0.25$ and 0.75, $y_0/b = 0.25$ and 0.75, $h_1c/k = h_2c/k = 0.63 \times 10^{-3}$, $c = 0.0635$ cm, $k = 0.202$ watt/cm-°C.

the substrate to obtain the overall thermal resistance. A reasonable temperature rise for the flat package and integrated circuit chip is 9.5°C so that the hottest temperature in the circuit should be no greater than 85°C if it is assumed that the maximum ambient temperature is 65°C. Less than a 5°C difference in temperatures due to interaction between the four flat packages is also expected. Thus, the temperature sensitive components of this circuit should track reasonably well. This ability to predict interaction effects between sources is obviously important for circuits with temperature sensitive components.

IV. CONCLUSIONS

The three-dimensional thermal problem of convection from the two large faces of a substrate with heat conducting leads has been solved for a substrate with isothermal edges, insulated edges or combinations of these boundary conditions. Results from this solution show the effects of substrate area, shape, thickness and thermal conductivity. Included are the effects of source area and the coefficient of convection. Two of the more important conclusions concerning heat transfer characteristics of glass and ceramic substrates follow:

(*i*) The coefficient of convection has a significant effect on the thermal resistance of substrates with insulated edges or high resistance leads and becomes the dominant parameter as the heat source area approaches the substrate area.

(*ii*) The thermal conductivity of the substrate is the dominant parameter affecting the thermal resistance of a substrate containing small area heat sources which are typical of beam-leaded integrated circuits.

V. ACKNOWLEDGMENTS

The author is grateful to H. E. Hughes for his helpful discussions and interest in this work.

REFERENCES

1. Morton, J. A., Bell Laboratories Record, Oct./Nov., 1966, pp. 290–291.
2. McAdams, W. H., *Heat Transmission,* McGraw-Hill Book Company, Inc., New York, 1952, pp. 165–183.
3. Morse, P. M. and Feshbach, H., *Methods of Theoretical Physics,* McGraw-Hill Book Company, Inc., New York, 1953, Part I, pp. 791–886.
4. Sneddon, I. N., *Fourier Transforms,* McGraw-Hill Book Company, Inc., New York, 1951, pp. 71–82.
5. Hildebrand, F. B., *Advanced Calculus for Applications,* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1962, pp. 25–29.
6. Hildebrand, F. B., *Methods of Applied Mathematics,* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1963, p. 17.

Characteristics of Superconductor Strip Transmission Lines with Periodic Load

By J. K. HSIAO

(Manuscript received June 9, 1967)

The characteristic impedance and propagation constant of a thin film superconducting strip transmission line has been derived by use of London's two fluid model. It is shown that this line at moderate frequency has negligible attenuation and dispersion. A periodically loaded cross film cryotron circuit is also analyzed. The attenuation, phase constant, and characteristic impedance of this loaded line is given and related to the parameters of the unloaded line by the factor K which is the ratio of the gate separation to the gate width.

I. INTRODUCTION

This paper presents a study of the high-frequency performance of thin film superconducting transmission circuits. Particular attention is given to transmission lines between cryotron elements and between substrates each carrying many cryotrons.

These interconnections are microstrip lines with very low characteristic impedances. Since the separation between the transmitting strip and the ground plane is very small in comparison to the width of the strip, edge effects can be neglected. This simplifies the analysis and gives an easy understanding of the propagation phenomena.

II. MICROSTRIP TRANSMISSION LINE WITH SUPERCONDUCTING STRIP

Due to the finite conductivity of the strip and ground plane in a non-superconductive microstrip transmission line, the phase characteristic has some dispersion and the attenuation is frequency dependent.[1] If the strip and ground plane are superconducting, the phase dispersion and the attenuation will disappear at frequencies below 1 GHz. This was shown by Swihart[2] using Maxwell's and London's equation. Using a very simple but not rigorous approach, the characteristic impedance and propagation constant are presented to give an understanding of this type of transmission line. The inductance, capacitance, and conductance in the dielectric region are the same as for non-superconducting strip line and have the following values:[1]

$$l^e = \frac{\mu_e h}{b} \tag{1}$$

$$c = \frac{\epsilon_e b}{h} \tag{2}$$

$$g = \frac{\sigma_e b}{h}\,, \tag{3}$$

provided $b/h \ll 1$,
where

l^e is the inductance per unit length,
c is the capacitance per unit length,
g is the conductance per unit length,
b is the width of the strip,
h is the distance between the strip and the ground plane, and
μ_e, ϵ_e, and σ_e are, respectively, the permeability, permitivity, and conductance of the dielectric material between the strip and ground plane.

The internal impedance of the conducting strip and ground plane of the superconducting line are different. These are found by the following manipulation. Assume that a superconducting strip of thickness $x = d$ and infinite width y, has a current flowing in the z-direction as shown in Fig. 1. The superconducting strip is immersed in a uniform dielectric material. The current is uniformly distributed along the y direction. London's equation that is based on a two fluid model[3] includes the following relations:

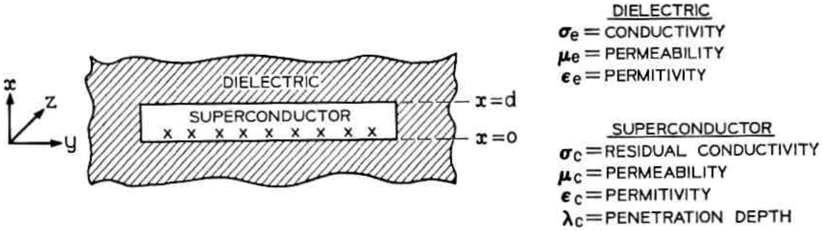


Fig. 1 — Current flowing in a superconductor strip.

$$\mathbf{J} = \mathbf{J}_n + \mathbf{J}_s \tag{4}$$

$$\nabla \times \lambda_c^2 \mathbf{J}_s = \mathbf{H} \tag{5}$$

$$\mu_c \frac{d}{dt} (\lambda_c^2 \mathbf{J}_s) = \mathbf{E}, \tag{6}$$

where

$\mathbf{J}_s$ is the superconducting current,
$\mathbf{J}_n$ is the normal conducting current,
$\mathbf{J}$ is the total current,
H and **E** are, respectively, magnetic and electric fields,
λ_c is the penetration depth of the superconductor, and
μ_c is the permeability of the superconductor.

Maxwell's equations applicable to both the superconducting and dielectric regions are as follows:

$$\nabla \cdot \mathbf{D} = \rho \qquad \mathbf{D} = \epsilon \mathbf{E} \tag{7}$$

$$\nabla \cdot \mathbf{B} = 0 \qquad \mathbf{B} = \mu \mathbf{H} \tag{8}$$

$$\nabla \times \mathbf{E} = -\mu \frac{d\mathbf{H}}{dt} \tag{9}$$

$$\nabla \times \mathbf{H} = \epsilon \left(\frac{d\mathbf{E}}{dt} \right) + \mathbf{J}. \tag{10}$$

These equations are based on MKS units.

From these equations, in the superconductor region, replacing J by $J_n + J_s$ and then using (6) and (7) through (10), we have a general expression of London's equation.

$$\nabla^2 \mathbf{E}_c = \frac{1}{\lambda_c^2} \mathbf{E}_c + \sigma_c \mu_c \frac{d\mathbf{E}_c}{dt} + \epsilon_c \mu_c \frac{d^2 E_c}{dt^2}, \tag{11}$$

where σ_c is the residual normal conductivity. This is the conductivity measured just above critical temperature. E_c, μ_c, and ϵ_c are, respectively, the E-field, permeability, and permittivity in the superconductor region.

The three terms on the right side of this equation are contributions of superconducting current, normal current, and the displacement current, respectively. In the superconductor region the displacement current term can be neglected. Hence, (11) becomes

$$\nabla^2 \mathbf{E}_c = \frac{1}{\lambda_c^2} \mathbf{E}_c + \sigma_c \mu_c \frac{d\mathbf{E}_c}{dt}. \tag{12}$$

In the dielectric region, we have

$$\nabla^2 \mathbf{E}_e = \sigma_e \mu_e \frac{d\mathbf{E}_e}{dt} + \epsilon_e \mu_e \frac{d^2\mathbf{E}_e}{dt^2}. \tag{13}$$

Use the coordinates defined by Fig. 1 and assuming the fields to be sinusoidal with respect to time,

$$\frac{d^2 E_{zc}}{dx^2} = \left(\frac{1}{\lambda_c^2} + j\omega\sigma_c\mu_c\right) E_{zc} \tag{14}$$

$$\frac{d^2 E_{ze}}{dx^2} = (-\omega^2 \epsilon_e \mu_e + j\omega\sigma_e\mu_e) E_{ze} \ . \tag{15}$$

The solutions of the above equations are, respectively,

$$E_{zc} = A_1 \cosh k_1 x + {}_{z}A_2 \sinh k_1 x \qquad 0 \leqq x \leqq d$$

$$E_{ze} = B \exp(-k_2 x) \qquad d \leqq x,$$

where k, and k_2 are defined as

$$k_1 = \frac{1}{\lambda_c} (1 + j\omega\sigma_c\mu_c\lambda_c^2)^{\frac{1}{2}} \tag{16}$$

$$k_2 = j\omega\sqrt{\mu_e\epsilon_e}\left(1 - j\frac{\sigma_e}{\omega\epsilon_e}\right)^{\frac{1}{2}}. \tag{17}$$

In the superconducting region, we retain two solutions in order to match the boundary condition at $x = d$, while in dielectric region we retain only one solution because we assume that the dielectric material is uniform and extends to infinity, and there is no reflection wave.

The H-fields in both regions are, respectively,

$$j\omega\mu_c H_{yc} = -A_1 k_1 \sinh k_1 x - A_2 k_1 \cosh k_1 x$$

$$j\omega\mu_e H_{ye} = B k_2 \exp(-k_2 x).$$

At the boundary $x = d$

$$E_{zc} = E_{ze} \ , \qquad H_{ye} = H_{yc} \ .$$

Hence,

$$A_2/A_1 = -\frac{k_1/k_2 \sinh k_1 d + \cosh k_1 d}{\sinh k_1 d + k_1/k_2 \cosh k_1 d} \quad \text{for} \quad \mu_c = \mu_e \ . \tag{18}$$

At

$$x = 0$$

$$E_{z_c} = A_1$$

$$H_{y_c} = -\frac{1}{j\omega\mu_c} A_2 k_1 .$$

By use of Ampere's Law the current inside the superconductor can be found. For unit width, it is

$$H_{y_c}\big|_{x=0} - H_{y_c}\big|_{x=d}$$

$$= \frac{1}{j\omega\mu_c}(-A_2k_1 + A_1k_1 \sinh k_1d + A_2k_1 \cosh k_1d). \qquad (19)$$

The internal impedance of a conductor is defined as[4]

$$Z^i = \frac{E_0}{I} \text{ ohm/m}^2, \qquad (20)$$

where E_0 is the surface E-field and I is the total current in the conductor for unit width. Hence,

$$Z^i = \frac{j\omega\mu_c}{k_1} \frac{k_2/k_1 + \coth k_1d}{1 + k_2/k_1 \tanh k_1d/2}. \qquad (21)$$

The classic skin-effect depth of a normal conductor is defined as[4]

$$\delta = \frac{1}{\sqrt{\pi f \mu_c \sigma_c}}.$$

From this and (16), for the superconducting region

$$k_1 = \frac{1}{\lambda_c}\left(1 + 2j\frac{\lambda_c^2}{\delta^2}\right)^{\frac{1}{2}}.$$

For several common superconducting materials, the classic skin-effect penetration depth δ are listed as follows:

	Transition[5] temperature	Conductivity[5] at transition temperature	δ at $f = 10^8$ Hz	δ at $f = 10^9$ Hz
Lead	7.22°K	0.52×10^{10} mho/m	0.7×10^4Å	0.216×10^4Å
Tin	3.74	0.896×10^{10}	0.53×10^4	0.167×10^4
Tantalum	4.38	0.806×10^9	1.77×10^4	0.56×10^4
Indium	3.374–3.432	0.36×10^9	2.65×10^4	0.84×10^4

The pentration depth λ_c is in the vicinity of thousand Å (for lead, it is 500Å, while for tin it is 1500 Å). Hence, in general, at a frequency

10^8 Hz $\lambda_c/\delta \approx 1/10$. This is assumed at a temperature T which is at least 0.1°K below transition temperature. Under this condition, for a good approximation

$$k_1 \approx \frac{1}{\lambda_c}.$$

The $\sigma_e/\omega\epsilon_e$ term in (17) is defined as the loss tangent of dielectric material. For most dielectric material at room temperature, it is approximately 10^{-3} to 10^{-4}.[6] For SiO, it is 10^{-2} at 1500 Hz. There is no available data at helium temperature. However, for reasonable approximation we can say that

$$\frac{\sigma_e}{\omega\epsilon_e} \ll 1.$$

Hence,

$$\begin{aligned} k_2 &\approx j\omega\sqrt{\mu_e\epsilon_e} \\ &\approx j\frac{2\pi}{\Lambda_e}, \end{aligned}$$

where Λ_e is the wave length in the dielectric. For SiO its relative permitivity is approximately 5.[7] Hence, $\Lambda_e \approx 74.5$ cm at 10^8 Hz.

The $|k_2|/|k_1|$ ratio is then

$$|k_2|/|k_1| \approx \frac{\lambda_c}{\Lambda_e} \approx 10^{-7}.$$

Under this assumption, (21) can be approximated as

$$Z^i = j\omega\mu_c\lambda_c \coth d/\lambda_c \text{ ohm/m}^2. \tag{22}$$

Next let us assume that a superconducting strip transmission line is formed by two strips immersed in a dielectric material as shown in Fig. 2. For this line, its series impedance is the sum of the Z^i and the inductance in the dielectric; hence, its series impedance and parallel admittance are, respectively

$$Z = j\omega\mu_c\frac{h}{b}\left(1 + \frac{\lambda_{c_1}}{h}\coth d_1/\lambda_{c_1} + \frac{\lambda_{c_2}}{h}\coth d_2/\lambda_{c_2}\right) \tag{23}$$

$$y = j\omega\epsilon_e\frac{b}{h}. \tag{24}$$

If the dielectric loss and classic skin-effect loss is not negligible, then

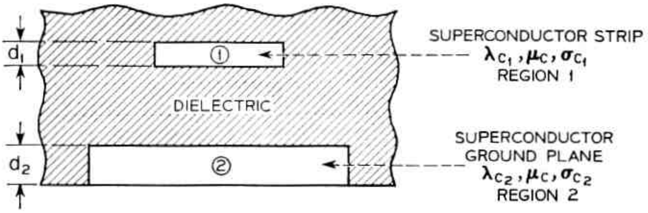


Fig. 2 — Superconductor Strip Transmission Line.

$$Z = j\omega\mu_c \frac{h}{b}\left[1 + \frac{\lambda_{c_1}}{h}\left(1 + 2j\,\frac{\lambda_{c_1}^2}{\delta_1^2}\right)^{-\frac{1}{2}} \coth k_1 d + \frac{\lambda_{c_2}}{h}\left(1 + 2j\,\frac{\lambda_{c_2}^2}{\delta_2^2}\right)^{-\frac{1}{2}} \coth k_2 d\right] \tag{23a}$$

$$y = j\omega\epsilon_e \frac{b}{h}\left(1 - j\,\frac{\sigma_e}{\omega\epsilon_e}\right). \tag{24a}$$

The characteristic impedance and propagation constant of this line can be found by use of the following relations:

$$Z_c = \sqrt{Z/y}$$

$$\gamma_c = \sqrt{Zy}.$$

Hence, by use of (23) and (24), we get

$$Z_c = \sqrt{\frac{\mu_c}{\epsilon_e}}\,\frac{h}{b}\left(1 + \frac{\lambda_{c_1}}{h}\coth d_1/\lambda_{c_1} + \frac{\lambda_{c_2}}{h}\coth d_2/\lambda_{c_2}\right)^{\frac{1}{2}} \tag{25}$$

$$\gamma_c = j\omega\sqrt{\mu_c\epsilon_e}\left(1 + \frac{\lambda_{c_1}}{h}\coth d_1/\lambda_{c_1} + \frac{\lambda_{c_2}}{h}\coth d_2/\lambda_{c_2}\right)^{\frac{1}{2}}. \tag{26}$$

Set $\gamma_c = j\beta_c$.

Then the phase velocity of this strip line is

$$V_p = \frac{\omega}{\beta_c}$$

$$= \frac{1}{\sqrt{\mu_c\epsilon_e}}\left(1 + \frac{\lambda_{c_1}}{h}\coth d_1/\lambda_{c_1} + \frac{\lambda_{c_2}}{h}\coth d_2/\lambda_{c_2}\right)^{-\frac{1}{2}} \text{ m/sec.}$$

And its delay time is

$$\tau_c = \sqrt{\mu_c\epsilon_e}\left(1 + \frac{\lambda_{c_1}}{h}\coth d_1/\lambda_{c_1} + \frac{\lambda_{c_2}}{h}\coth d_2/\lambda_{c_2}\right)^{\frac{1}{2}} \text{ sec/m.}$$

While by use of (23a) and (24a) we get

$$Z_c = \sqrt{\frac{\mu_c}{\epsilon_e}}\frac{h}{b} \cdot \frac{\left[1+\frac{\lambda_{c_1}}{h}(1+2j\lambda_{c_1}^2/\delta_1^2)^{-\frac{1}{2}}\coth k_1d+\frac{\lambda_{c_2}}{h}(1+2j\lambda_{c_2}^2/\delta_2^2)^{-\frac{1}{2}}\coth k_2d\right]^{\frac{1}{2}}}{(1-j\sigma_e/\omega\epsilon_e)^{\frac{1}{2}}} \tag{25a}$$

$$\gamma_c = j\omega\sqrt{\mu_c\epsilon_e}\,(1-j\sigma_e/\omega\epsilon_e)^{\frac{1}{2}}\left[1+\frac{\lambda_{c_1}}{h}(1+2j\lambda_{c_1}^2/\delta_1^2)^{-\frac{1}{2}}\coth k_1d + \frac{\lambda_{c_1}}{h}(1+2j\lambda_{c_2}^2/\delta_2)^{-\frac{1}{2}}\coth k_2d\right]^{\frac{1}{2}}. \tag{26a}$$

At frequencies below 1 GHz with the temperature at least 0.1°K below transition temperature, (25) and (26) give fairly accurate results, providing the dielectric loss is negligible. Then the characteristic impedance is a real number with negligible frequency dependence. The propagation constant is directly proportional to frequency; hence, its group velocity and phase velocity are the same and there is no attenuation.

Fig. 3 shows the characteristic impedance and delay time of some superconducting strip lines with various dielectric thickness (h).

It is shown that the thickness (d_1) of the strip line film has little effect on the characteristic impedance and delay time. However, the characteristic impedance changes proportionally with dielectric thickness (h) while the delay time τ_c decreases nonlinearly by only 20 percent for an order of magnitude change in h.

III. A SUPERCONDUCTING STRIP LINE WITH PERIODIC STRUCTURE

The cross film cryotron consists of a control strip and a gate strip crossing and perpendicular to each other. In a memory circuit or in a tree-type selective circuit, a single control strip usually crosses many gate strips. At each intersection there exists coupling between the control and gate strips to form a periodic structure. The characteristic impedance and propagation constant of the control line are functions of the periodic loading. (Refer to Fig. 4.)

The control line and gate line are assumed to be terminated with their respective characteristic impedances Z_c and Z_g. The control line is also assumed to be uniform without discontinuity except for the

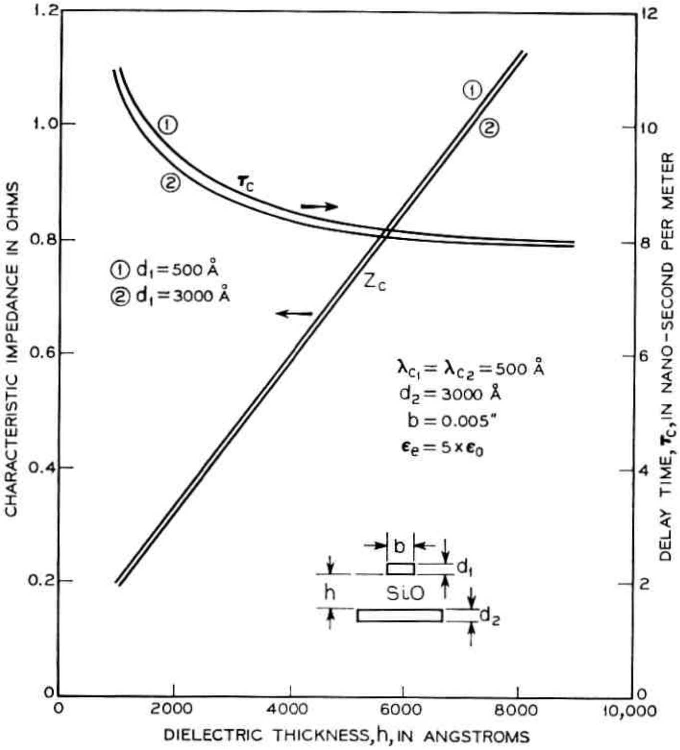


Fig. 3 — Relation between characteristic impedance delay time, and dielectric thickness for superconducting thin film strip line.

periodic loading of the couplings to the gates. The equivalent circuit is shown in Fig. 5.

In Fig. 5, γ_c is the propagation constant of the control line, Y_c is the coupling admittance between gate and control line. The characteristic impedance and propagation constant of this periodically-loaded line is as follows (see Appendix A):

$$Z_0 = Z_c\left[1 - \frac{YZ_c}{\sinh 2\gamma_c l + YZ_c \cosh^2 \gamma_c l}\right]^{\frac{1}{2}} \tag{27}$$

$$\gamma_0 = \cosh^{-1} [\cosh 2\gamma_c l + YZ_c/2 \sinh 2\gamma_c l], \tag{28}$$

where

$$Y = \frac{Y_c}{1 + \frac{1}{2}Y_c Z_g} \qquad l = \tfrac{1}{2}(d_g + W_g).$$

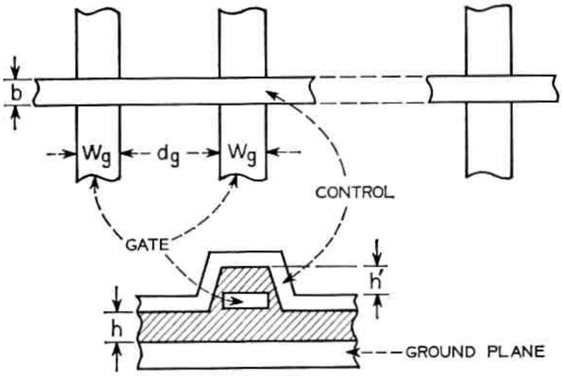


Fig. 4 — Periodic gate crossing of a cryotron circuit.

If the ratio of the control line width b to the distance h_g between gate and controls lines are large $(b/h) > 10'$ this capacitance to a very good approximation is[8]

$$C = \epsilon_e \frac{A}{h'}, \tag{29a}$$

where A is the intersection area of the control and gate lines, and ϵ_e is the permittivity of the insulation material.

Since the magnetic fields in control and gate lines is assumed orthogonal to each other; therefore, there is no magnetic coupling, and

$$Y_c = j\omega C.$$

If we want to take into account of the dielectric loss of the insulation material, Y_c becomes

$$Y_c = j\omega\epsilon_e \frac{A}{h'}\left(1 - j\frac{\sigma_e}{\omega\epsilon_e}\right), \tag{29b}$$

where σ_e is the conductivity of the insulation material. For SiO, $\sigma_e/\omega\epsilon_e$ is approximately 10^{-2}.[7] Hence, this term can be neglected in this case.

By use of this result, we find

$$Y = \frac{j\omega C}{1 + \frac{1}{2}j\omega CZ_g}. \tag{30}$$

For a typical cryotron, the width of the gate is approximately 20 milli-inches and the width of control line is 5 milli-inches. The insulation material SiO has a relative permittivity of 5. If h' is 5000 Å, then

$$C = \frac{1}{36\pi} \times 10^{-9} \times 5 \frac{20 \times 5 \times 2.54^2 \times 10^{-10}}{5 \times 10^{-7}}.$$

$$C \approx 5.7 \times 10^{-12} \text{ farads.}$$

If the gate is terminated by its characteristic impedance, then Z_g is approximately one ohm for an ordinary cryotron. Hence,

$$\omega C Z_g \approx 6 \times 10^{-3}$$

at a frequency of 100 MHz. Therefore, (30) can be simplified as

$$Y \approx j\omega C(1 - \tfrac{1}{2} j\omega C Z_g). \tag{31}$$

We have shown in Section II that the propagation constant of a superconducting strip line is an imaginary number; therefore, we set

$$\gamma_c = j\beta_c .$$

Hence,

$$Z_0 = Z_c \left[1 - \frac{\omega C Z_c (1 - \frac{1}{2} j\omega C Z_g)}{\sin 2\beta_c l + \omega C Z_c (1 - j\frac{1}{2}\omega C Z_g) \cos^2 \beta_c l} \right]^{\frac{1}{2}}. \tag{32}$$

For typical cryotron circuits, the spacing between gates (d_g) is about equal to the gate width (W_g) for maximum compactness. At a frequency of 10^8 Hz,

$$\beta_c l \approx 10^{-3} \text{ radians.}$$

Hence,

$$\sin 2\, \beta_c l \approx 2\, \beta_c l$$

$$\cos^2 2\, \beta_c l \approx 1.$$

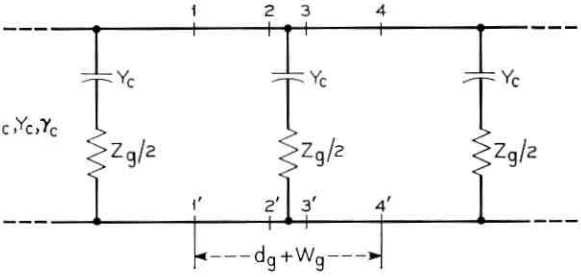


Fig. 5 — Equivalent circuit of a periodic loaded cryotron.

Neglecting the $(\omega C)^2 Z_g Z_c$ term, we find

$$Z_0 = Z_c \left[1 - \frac{1}{1 + \dfrac{2\beta_c l}{\omega C Z_c}} \right]^{\frac{1}{2}}. \tag{33}$$

By use of (25), (26), and (29a), we find

$$\frac{2\beta_c l}{\omega C Z_c} = \frac{h'(d_g + W_g)}{hW_g}. \tag{34}$$

If $h' = h$, then

$$Z_0 = Z_c \left(\frac{W_g + d_g}{2W_g + d_g} \right)^{\frac{1}{2}}. \tag{35a}$$

Setting

$$d_g = KW_g ,$$

$$Z_0 = Z_c \left(\frac{1 + K}{2 + K} \right)^{\frac{1}{2}}. \tag{35b}$$

For maximum package density, the gates are placed as close as possible, and thus, d_g and consequently K are made as small as possible. However, to avoid interference between adjacent gates, it is usual to set the distance between gates at least equal to their width. For this condition, $Z_0 = 0.815\, Z_c$. As K becomes larger, Z_0 approaches Z_c in value.

In the next step, the propagation constant of the periodic-loaded line is determined.

First substituting the following condition in (28):

$$\gamma_c = j\beta_c$$

$$\sinh 2\gamma_c l \approx 2j\beta_c l$$

$$\cosh 2\gamma_c l \approx 1 - \frac{(2\beta_c l)^2}{2}$$

and set

$$\gamma_0 = \alpha_0 + j\beta_0 .$$

Equation (28) can be rewritten as

$$\cosh (\alpha_0 + j\beta_0) = 1 - \frac{(2\beta_c l)^2}{2} - \omega C Z_c \beta_c l + j\tfrac{1}{2}(\omega C)^2 Z_c Z_g \beta_c l.$$

It is noticed that the real part on the right side of this equation is very close to but less than 1. The imaginary part is very small. Hence, we conclude that α_0 and β_0 must be a small quantity. The real and imaginary parts of this equation become, respectively,

$$\cosh \alpha_0 \cos \beta_0 = 1 - \frac{(2\beta_c l)^2}{2} - \omega C Z_c \beta_c l \tag{36a}$$

$$\sinh \alpha_0 \sin \beta_0 = \tfrac{1}{2}(\omega C)^2 Z_c Z_g \beta_c l. \tag{36b}$$

Using the approximate relationship

$$\cosh \alpha_0 \approx 1 + \frac{\alpha_0^2}{2}$$

$$\cos \beta_0 \approx 1 - \frac{\beta_0^2}{2}$$

$$\sinh \alpha_0 \approx \alpha_0$$

$$\sin \beta_0 \approx \beta_0$$

$$Z_c \approx Z_g$$

and defining the following constants [see (34)]:

$$\frac{\omega C Z_c}{2\beta_c l} = \frac{W_g}{d_g + W_g} = R \leqq 1 \tag{37}$$

$$2\beta_c l = \theta_c ,$$

where θ_c is the phase shift between gate crossings along the control line without loading, (36a) and (36b) can be rewritten, respectively, as

$$\left(1 + \frac{\alpha_0^2}{2}\right)\left(1 - \frac{\beta_0^2}{2}\right) = 1 - \frac{\theta_c^2}{2} - \tfrac{1}{2}R\theta_c^2$$

$$\alpha_0 \beta_0 = \tfrac{1}{4}R^2 \theta_c^3 .$$

Neglecting the $\alpha_0^2\beta_0^2$ term, we find

$$\beta_0^2 - \alpha_0^2 = \theta_c^2 + R\theta_c^2$$

$$\alpha_0 \beta_0 = \tfrac{1}{4}R^2 \theta_c^3 .$$

Hence,

$$\beta_0 = \frac{1}{\sqrt{2}}\,\theta_c\left\{(1 + R)\left[1 \pm \left(1 + \frac{\frac{1}{4}(R^2\theta_c)^2}{(1 + R)^2}\right)^{\frac{1}{2}}\right]\right\}^{\frac{1}{2}}.$$

Since

$$\tfrac{1}{4}(R^2\theta_c)^2 \ll (1 + R)^2,$$

β_0 becomes

$$\beta_0 = \theta_c(1 + R)^{\frac{1}{2}}\left(1 + \frac{1}{16}\frac{R^4\theta_c^2}{(1 + R)^2}\right)^{\frac{1}{2}}. \tag{38}$$

Further neglecting $R^4\theta_c^2/(1 + R)^2$, then we obtain

$$\beta_0 = \theta_c(1 + R)^{\frac{1}{2}}$$

$$\alpha_0 = \tfrac{1}{4}\theta_c^2\,\frac{R^2}{(1 + R)^{\frac{3}{2}}}.$$

Replacing R by (37), we obtain

$$\beta_0 = \theta_c\left(\frac{d_g + 2W_g}{d_g + W_g}\right)^{\frac{1}{2}} \tag{39a}$$

$$\alpha_0 = \tfrac{1}{4}\theta_c^2\,\frac{W_g^2}{(d_g + W_g)^{\frac{3}{2}}(d_g + 2W_g)^{\frac{1}{2}}}. \tag{39b}$$

Using the relation $d = KW_g$, then

$$\beta_0 = \theta_c\left(\frac{2 + K}{1 + K}\right)^{\frac{1}{2}} = \text{phase constant} \tag{40a}$$

$$\alpha_0 = \tfrac{1}{4}\theta_c^2\,\frac{1}{(K + 1)^{\frac{3}{2}}(K + 2)^{\frac{1}{2}}} = \text{attenuation constant.} \tag{40b}$$

Equations (40a) and (40b) are phase constant and attenuation per period.* If there are n gates crossing in one meter length of control line and if they are equally spaced, then the phase constant and attenuation per meter is

$$\beta = \beta_c(d_g + W_g)n\left(\frac{2 + K}{1 + K}\right)^{\frac{1}{2}}\ \text{rad/m}$$

$$\alpha = \tfrac{1}{4}\beta_c^2(d_g + W_g)^2 n\,\frac{1}{(1 + K)^{\frac{3}{2}}(2 + K)^{\frac{1}{2}}}\ \text{neper/m.}$$

Replacing β_c by τ_c and realizing that $n = 1/d_g + W_g$, we find

$$\beta = \tau_c\omega\left(\frac{2 + K}{1 + K}\right)^{\frac{1}{2}}\ \text{radians/m} \tag{41a}$$

$$\alpha = \tfrac{1}{4}\tau_c^2\omega^2 W_g\,\frac{1}{(K + 1)^{\frac{1}{2}}(K + 2)^{\frac{1}{2}}}\ \text{neper/m.} \tag{41b}$$

* One period is the distance $(d_g + W_g)$ between cryotron gate crossing of the strip line.

The delay time per meter of the loaded line is

$$\tau_0 = \tau_c\left(\frac{2 + K}{1 + K}\right)^{\frac{1}{2}} \text{sec/m}. \tag{42}$$

For a typical cryotron having control width 0.005 inch and gate width $W_g = 0.02$ inch, SiO as dielectric of a thickness 5000 Å, and $K = 1$, then its attenuation at 10^8 Hz becomes

$$\alpha = 2.56 \times 10^{-3} \text{ neper/m}$$

or

$$= 2.12 \times 10^{-2} \text{ dB/m}.$$

With this configuration, on a 3″ × 3″ substrate from one side to the other side of the substrate, 75 cryotron can be laid down. The attenuation will be only about 5×10^{-4} dB. This is extremely small. For larger K, this attenuation will still be less. Hence, for our purpose, it can be neglected. The delay time for this particular example is approximately 10 nanosecond per meter. For the same substrate, the time for a pulse to travel from one side to the other side of the substrate is approximately one nanosecond.

Fig. 6 shows the relation between characteristic impedance Z_0, delay time τ_0 and the ratio (K) of gate separation (d_g) to gate width (W_g). From Fig. 6 it is shown that the characteristic impedance Z_0 of loaded line is less than the characteristic impedance Z_c of an unloaded line at smaller K (at $K = 1$, $Z_0 \approx 0.82\ Z_c$). As K becomes larger, ($K > 10$) Z_0/Z_c ratio approaches unity. The delay time τ_0 per meter of a loaded line is larger than the delay time of an unloaded line (at $K = 1$ $\tau_0 = 1.22\ \tau_c$). However, their ratio also approaches unity as K becomes larger.

IV. CONCLUSION

Section II derives the characteristic impedance and propagation constant for typical thin film superconducting strip lines used for interconnecting cryotron elements. The characteristic impedance is shown to be a real number with negligible frequency dependence. The propagation constant is shown to be directly proportional to frequency and hence the group and phase velocity is identical. It is found that the transmission performance of these lines are, for practical purposes, independent of film thickness when in excess of 500 Å. Fig. 3 shows the characteristic impedance (Z_c) and delay time τ_c for a film strip

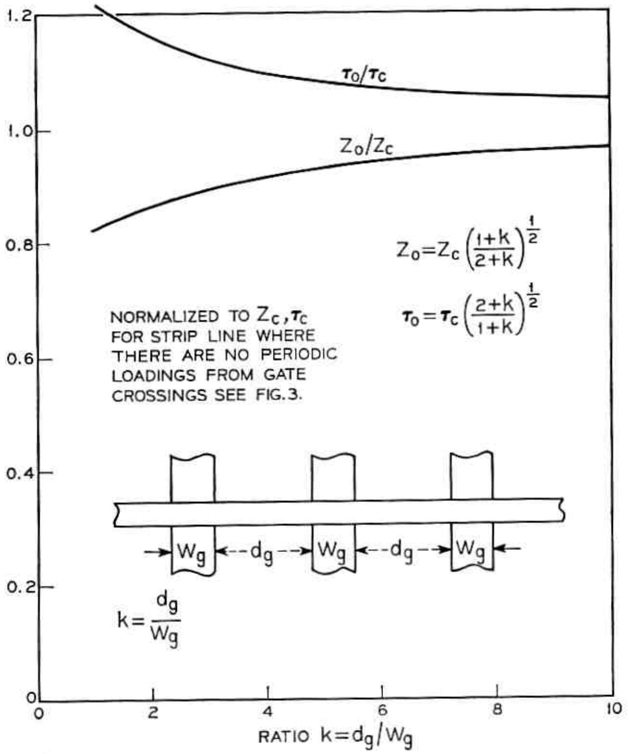


Fig. 6 — Relation between characteristic impedance (Z_0) delay ( ), and ratio (k) of gates separation (d) and gate width (W_g) of superconducting strip line with periodic gate crossing.

of 5 milli-inches width on SiO dielectric. For other widths, Z_c is inversely proportional to width and τ_c is independent of width. For example, at a dielectric thickness of 5000 Å and film width of 5 milli-inches, $Z_c = 0.725$ ohms and $\tau_c = 8.2 \times 10^{-9}$ sec/meter.

Section III derives the characteristic impedance (Z_0) and propagation constant for thin film superconducting strip line used as a common control which crosses a series of cryotron gates periodically spaced. This is related to the transmission characteristics (Z_c, τ_c) for the strip line if they did not cross cryotron gates. Fig. 6 shows this relation with plots of characteristic impedance and delay time per meter versus the ratio (K) of distance between gate crossing with gate width. For high packing density $K = 1$, it is found that the characteristic impedance is reduced almost 20 percent due to the loading of the gate crossings.

The delay time τ_0 is increased 20 percent. The attenuation constant (α_0) is extremely small, hence it can be neglected for practical purposes.

APPENDIX A

Referring to Fig. 5 from point 1–1′ to point 2–2′ the control line has characteristic impedance Z_c and propagation constant γ_c. The voltage and current equations in terms of the Z_c and γ_c are[9,10,11,12]

$$\begin{vmatrix} V_1 \\ I_1 \end{vmatrix} = \begin{vmatrix} \cosh \gamma_c l & Z_c \sinh \gamma_c l \\ \frac{1}{Z_c} \sinh \gamma_c l & \cosh \gamma_c l \end{vmatrix} \begin{vmatrix} V_2 \\ I_2 \end{vmatrix}, \tag{43}$$

where

$$l = \tfrac{1}{2}(d_g + W_g).$$

From point 2–2′ to point 3–3′, the voltage and current relations are

$$\begin{vmatrix} V_2 \\ I_2 \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ Y & 1 \end{vmatrix} \begin{vmatrix} V_3 \\ I_3 \end{vmatrix}, \tag{44}$$

where

$$Y = \frac{Y_c}{1 + \frac{1}{2} Y_c Z_g}.$$

If the gate lines are terminated in their characteristic impedance then Z_g is the characteristic impedance and a real number independent of frequency. Otherwise, Z_g might be a complex and be frequency dependent.

Similarly, from point 3–3′ to 4–4′ we have

$$\begin{vmatrix} V_3 \\ I_3 \end{vmatrix} = \begin{vmatrix} \cosh \gamma_c l & Z_c \sinh \gamma_c l \\ \frac{1}{Z_c} \sinh \gamma_c l & \cosh \gamma_c l \end{vmatrix} \begin{vmatrix} V_4 \\ I_4 \end{vmatrix}. \tag{45}$$

Hence, for each period, we have

$$\begin{vmatrix} V_1 \\ I_1 \end{vmatrix} = \begin{vmatrix} A & B \\ C & D \end{vmatrix} \begin{vmatrix} V_4 \\ I_4 \end{vmatrix},$$

where

$$A = \cosh 2\gamma_c l + YZ_c/2 \sinh 2\gamma_c l,$$

$$A = D,$$

and

$$B = Z_c \sinh 2\gamma_c l + YZ_c^2 \sinh^2 \gamma_c l,$$

$$C = \frac{1}{Z_c} \sinh 2\gamma_c l + Y \cosh^2 \gamma_c l.$$

Since $A = D$ this is a symmetrical circuit, accordingly its characteristic impedance Z_0 and propagation constant γ_0 are

$$\begin{aligned} Z_0 &= \sqrt{B/C} \\ &= Z_c\left[1 - \frac{YZ_c}{\sinh 2\gamma_c l + YZ_c \cosh^2 \gamma_c l}\right]^{\frac{1}{2}} \end{aligned} \tag{46}$$

$$\begin{aligned} \gamma_0 &= \cosh^{-1} A \\ &= \cosh^{-1} [\cosh 2\gamma_c l + YZ_c/2 \sinh 2\gamma_c l]. \end{aligned} \tag{47}$$

LIST OF SYMBOLS USED

l^e = Inductance
C = Capacitance
g = Conductance
$W_o d_1 b,\ h,\ d_o$ = Geometric Parameters
μ_e = Permeability of dielectric material
ϵ_e = Permittivity of dielectric material
σ_e = Conductivity of dielectric material
μ_c = Permeability of superconductor
ϵ_c = Permittivity of superconductor
σ_c = Residual normal conductivity of superconductor
J = Current Density
E = Electric Field
H = Magnetic Field
ω = Angular velocity
λ_c = Penetration depth of superconductor
δ = Classic skin-effect depth
Z = Impedance
Z_c , Z_0 = Characteristic Impedance
γ_c = Progapation constant
θ_c = Phase constant
Λ_e = Wave length
τ = Delay time

REFERENCES

1. King, Ronald W. P., *Transmission-Line Theory,* McGraw-Hill Book Co., Inc., New York, 1955.
2. Swihart, James C., Field Solution for a Thin-Film Superconducting Strip Transmission Line, J. Appl. Phys., *32,* No. 2, March, 1961, pp. 461–469.
3. London, Fritz, *Super Fluids Microscopic Theory of Superconductivity, 1,* Dover Publication, Inc., New York.
4. Ramo and Whinnery, *Fields and Waves in Modern Radio,* John Wiley and Sons, p. 237 to 239.
5. AD 272 769, "A Compendium of the Properties of Materials at Low Temperature, (Phase II)," December 1961.
6. *Ibid.* (4) p. 309.
7. Bruce, J. H. and Balmer, J. R., Ultrathin Dielectric Film, Electrotechnology, December, 1960, p. 157.
8. Palmer, H. B., Capacitance of a Parallel-Plate Capacitor, Elec. Engrg., *56,* March, 1937, pp. 363–366.
9. Gullemin, *Communication Network,* Vol. II, Chapter 4, McGraw-Hill Book Co., Inc., New York, 1953.
10. McQuillian, J. D. R., The Design Problems of a Megabit Storage Matrix of Use in a High-Speed Computer, IRE Trans. Electron. Computer, June, 1962, pp. 390–404.
11. Vowels, R. E., Matrix Methods in the Solution of Ladder Networks, J. IEE, *95,* January, 1948, pp. 40–49.
12. Lines, A. W., Nicoll, G. R., and Woodward, A. M., Some Properties of Waveguides with Periodic Structure, Proc. IEE, Part II *97,* 1950 pp. 263–276.

Scattering Relations in Lossless Varactor Frequency Multipliers

By C. DRAGONE and V. K. PRABHU

(Manuscript received May 22, 1967)

In recent years, the use of varactor diodes for harmonic generation has become increasingly widespread. Varactor harmonic generators come under the general class of pumped nonlinear systems, which are networks driven periodically by a pump or a local oscillator at a frequency ω_0 and its harmonics. For such systems, a general method has been presented in this paper to obtain the scattering parameters which relate the small-signal fluctuations present at various points in the system. In particular, the scattering parameters of lossless abrupt-junction varactor harmonic generators of order 2^n, 3^s, and 2^n3^s with minimum number of idlers have been obtained. It has been shown for these multipliers that there is no amplitude-to-phase or phase-to-amplitude conversion if fluctuations are in the vicinity of the carriers. With minor modifications this theory can be extended to the study of lossy varactor harmonic generators.

I. INTRODUCTION

The carrier voltages and currents present in a varactor frequency multiplier are perturbed by small amplitude and phase fluctuations due to a variety of causes, such as noise, synchronizing signals, etc. In some applications, these fluctuations may be due to modulations purposely applied to the carriers. An example of such applications is that in which a frequency modulated signal is multiplied in frequency to increase its modulation index. It is the purpose of this paper to study how these perturbations propagate in the circuit of a multiplier. In other words, this paper considers the problem of determining the small-signal behavior—a problem which is of basic importance in understanding the problem of stability and noise performance in high efficiency varactor multipliers.‡

‡ See Ref. 1. The problem of stability is also treated in a subsequent paper.[2] Part of the results obtained in this paper represent generalizations of some of the results presented in Refs. 1, 3, and 4.

In the earlier part of this paper, a general method has been presented in order to obtain the scattering parameters of pumped nonlinear systems which are networks driven periodically by a pump or a local oscillator at a frequency¶ ω_0 and its harmonics. Harmonic generators discussed in this paper come under this class of systems. Some of the formalisms usually used to describe the fluctuations in these systems are also briefly reviewed.

In the second part of this paper we discuss varactor multipliers in which the diode is not overdriven and is of the abrupt-junction type. The equivalent circuit of this type of multiplier consists of an ordinary linear, passive, and time-invariant circuit connected to the time-varying component of the elastance $S(t)$ of the varactor.[5] In general, it is shown that a complete solution of the small-signal behavior of such a circuit requires that $S(t)$ be known. On the other hand, it is well known that certain properties of the small-signal behavior of a multiplier do not depend at all on the particular form of $S(t)$. For instance, a general and well-known property of a multiplier of order N is that slow fluctuations in the phase of the input drive produce N times as large fluctuations in the phase of the output signal. One of the main results of this paper is that, under certain general conditions, many other properties of the multiplier are related in a simple way only to the order of multiplication N. All the small-signal characteristics of a multiplier that are of practical interest can, therefore, be readily determined without having to calculate $S(t)$.

Specifically, we consider a lossless multiplier of order $N = 2^n3^s = 2, 3, 4$ etc., which is tuned at all carrier frequencies§ and has the least number of idlers. Then, if the various small-signal fluctuations of such a multiplier are properly normalized with respect to the corresponding carriers, one finds that the small-signal terminal behavior of the elastance $S(t)$ is completely determined by N only. It is important to point out that this is exactly true only for $\omega \ll \omega_0$, where ω is the frequency of the fluctuations and ω_0 is the carrier frequency of the drive. If this inequality is not satisfied, then the small-signal behavior will also depend on ω. A consequence of these results is that the AM and PM scattering parameters of the multiplier of order $N = 2^n3^s$ considered in this paper only depend on n and s, in the vicinity of the carriers. They are given, respectively, by the two matrices

¶ In this paper the word frequency has been used exclusively for the angular frequency of a sinusoidal signal. If f is the frequency of a signal in Hz its angular frequency ω is given by $\omega = 2\pi f$ in radians/second.

§ Tuning of idlers, and input, and output circuits usually gives near optimum efficiency.[5,6,7,8]

$$\begin{bmatrix} \frac{1}{3} - \frac{(-1)^n}{3} 2^{-n} & (-1)^n 2^{-n} \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & (-1)^n 3^{-s} \\ 2^n 3^s & \frac{1}{3} - \frac{(-1)^n}{3} 2^n \end{bmatrix}.$$

It is important to point out that for the above multiplier it has been assumed in deriving the results that the bias circuit is properly designed so that there are no low-frequency fluctuations of the average capacitance of the varactor diode.[6] This assumption leads to the result that there is no amplitude-to-phase and phase-to-amplitude conversion if $\omega/\omega_0 \ll 1$.

Several other results are also presented in this paper. For instance, it is shown that, if the number of idlers is minimum, then an abrupt-junction varactor multiplier of order $N = N_1 \times N_2 \times \ldots \times N_n$ is equivalent to a cascade of n multipliers of order $N_1, N_2, \ldots, N_n$. If the varactor is not overdriven, this property furnishes the basic equivalent circuit for studying the properties of most of the higher-order multipliers encountered in practice ($N = 4, 6, 8$, etc.).

Finally, it is important to point out that techniques presented in this paper are applicable to the derivation of scattering parameters of multipliers, of any order, with any arbitrary configuration of idlers, and using a varactor diode having arbitrary capacitance variation and drive level. We only assume that the elastance $S(t)$ of the diode used in the multiplier has a Fourier series.‡

II. SOME CONSIDERATIONS OF PERIODICALLY DRIVEN NONLINEAR SYSTEMS

As mentioned earlier in this paper, frequency multipliers come under the general class of nonlinear systems driven by a strong periodic carrier. It is our interest to study in this paper how small perturbations on the periodic driving of such systems are propagated, and to this end we shall give a brief introduction§ of a circuit theory which enables us to relate the perturbations at different parts of the system. The perturbations or fluctuations that we would like to analyze could be caused by desired or undesired modulation, noise, hum, or synchronizing signals. The origin of these sources of fluctuations is not relevant to our development of this theory.

Let us consider a nonlinear system. It is our assumption that the

‡ The conditions under which a periodic time function $x(t)$ has a Fourier series are well-known; and can be found in any book on Fourier series. See, for example, Ref. 9.

§ See Ref. 10 for a more detailed account of this theory.

large-signal voltages and currents at various parts within the system are, by design, periodic with some frequency ω_0. Thus, the voltage at some specific point within the network or across one of its terminal pairs, $v(t)$, is of the form

$$v(t) = \sum_{k=-\infty}^{\infty} V_k \exp(jk\omega_0 t), \tag{1}$$

where the V_k's are half-amplitude¶ Fourier coefficients, with $V_{-k} = V_k^*$. However, the actual voltage $v(t)$ may deviate from (1) because of fluctuations present in the system. Thus,

$$v(t) = \sum_{k=-\infty}^{\infty} V_k \exp(jk\omega_0 t) + \delta v(t), \tag{2}$$

where $\delta v(t)$ is small compared to $v(t)$ in (1). The circuit theory that we shall use in the rest of this paper is one which describes perturbations $\delta v(t)$ and relates them to similar perturbations of voltages and currents in other parts of the system. The perturbations are assumed to be small and they are at frequencies close to the carriers.||

The carrier voltage at some particular point in the system is of the form

$$V_k \exp(jk\omega_0 t) + V_k^* \exp(-jk\omega_0 t), \tag{3}$$

where V_k has some phase angle φ_{vk}. The actual voltage $v_k(t)$ in the vicinity of this carrier deviates from (3) because of the perturbation $\delta v_k(t)$;

$$v_k(t) = V_k \exp(jk\omega_0 t) + V_k^* \exp(-jk\omega_0 t) + \delta v_k(t). \tag{4}$$

Similar expressions can be written for currents and voltages at various places in the network. The various voltages like $v(t)$ obey Kirchhoff's voltage law, and various currents $i(t)$ defined at various points in the network obey Kirchhoff's current law. Furthermore, the carrier voltages and currents at various points in the network obey these Kirchhoff's laws, leading us to conclude that the perturbations like $\delta v(t)$ and $\delta i(t)$ also obey them.

Let us now assume that the perturbation $\delta v_k(t)$ contain frequencies that are located in a band of width $2\omega_c$ centered about frequency $k\omega_0$ where $2\omega_c < \omega_0$.[10] We can write[10] $v_k(t)$ as‡

¶ Note the use of half-amplitudes, rather than amplitudes or rms values.

|| The large signal voltage or current present in the system at frequency $\pm k\omega_0$ will be referred to from hereon as the carrier voltage or current at that frequency. In frequency multipliers carriers are at different frequencies at different parts of the system.

‡ In writing (6), it is assumed that $| v_{pk}(t) |/| V_k | \ll 1$ for all t.

$$v_k(t) = 2 \text{ Re } [| V_k | + v_{ak}(t) - jv_{pk}(t)] \exp [j(k\omega_0 t + \varphi_{vk})] \tag{5}$$

$$\approx 2 \text{ Re } [| V_k | + v_{ak}(t)] \exp \left\{ j \left[k\omega_0 t + \varphi_{vk} - \frac{v_{pk}(t)}{| V_k |} \right] \right\}, \tag{6}$$

where $v_{ak}(t)$ and $v_{pk}(t)$ are slowly varying functions of time. The voltage $v_{ak}(t)$ can be interpreted, since it is small, as a perturbation on the amplitude $| V_k |$ of the carrier. Similarly, voltage $v_{pk}(t)$, because of (6), can be interpreted as a perturbation on the phase $k\omega_0 t + \varphi_{vk}$ of the carrier. We shall refer to $v_{ak}(t)$ as amplitude (AM) fluctuations and to $v_{pk}(t)$ as phase (PM) fluctuations. Similar AM and PM fluctuations can be defined at various points in the system.

If these AM and PM fluctuations are sinusoidal, we have§

$$v_{ak}(t) = V_{ak} \exp (j\omega t) + V_{ak}^* \exp (-j\omega t) \tag{7}$$

and

$$v_{pk}(t) = V_{pk} \exp (j\omega t) + V_{pk}^* \exp (-j\omega t). \tag{8}$$

The actual voltage $v_k(t)$ is then given by

$$v_k(t) = 2 \text{ Re } [| V_k | + (V_{ak} - jV_{pk}) \exp (j\omega t) \\ + (V_{ak}^* - jV_{pk}^*) \exp (-j\omega t)] \exp [j(k\omega_0 t + \varphi_{vk})] \tag{9}$$

$$= 2 \text{ Re } \{ V_k \exp (jk\omega_0 t) \\ + V_{\alpha k} \exp [j(k\omega_0 + \omega)t] + V_{\beta k} \exp [j(-k\omega_0 + \omega)t] \}, \tag{10}$$

where $V_{\alpha k}$, $V_{\beta k}$, V_{ak}, and V_{pk} are related. The relation is

$$\begin{bmatrix} V_{ak} \\ V_{pk} \end{bmatrix} = \lambda_{vk} \begin{bmatrix} V_{\alpha k} \\ V_{\beta k} \end{bmatrix}, \tag{11}$$

where the matrix λ_{vk} is a function of only the carrier phase angle φ_{vk}. The matrix λ_{vk} can be represented as

$$\lambda_{vk} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ j & -j \end{bmatrix} \begin{bmatrix} \exp (-j\varphi_{vk}) & 0 \\ 0 & \exp (j\varphi_{vk}) \end{bmatrix}. \tag{12}$$

Equation (10) shows explicitly the three frequencies $k\omega_0$, $k\omega_0 + \omega$, and $-k\omega_0 + \omega$. The two sidebands here are both higher in frequency than $k\omega_0$, and $-k\omega_0$, respectively, and therefore, these representations are referred to as upper sideband $(\alpha - \beta)$ representations. We will use them along with the representations of the form (9) in the rest

§ Since the fluctuations $v_{ak}(t)$ and $v_{pk}(t)$ are band limited around *dc*, ω must be less than ω_c in magnitude, where $2\omega_c < \omega_0$.

of this work. Their mutual relation is given in (11). Because of (10) we shall refer to ω as the fluctuation difference frequency.

Let us now consider a pumped nonlinear system exchanging power at the carrier frequencies $\pm\omega_0$, $\pm 2\omega_0$, $\cdots$, $\pm n\omega_0$. A nonlinear system exchanging power at a number of frequencies can be considered as a multiport multifrequency system as shown in Fig. 1. In Fig. 1 the system exchanges power at n carrier frequencies and it is assumed, without loss of generality, that no two ports exchange power at the same carrier frequency. Let the perturbation voltage and current at port k be denoted by $\delta v_k(t)$ and $\delta i_k(t)$, respectively. Since δv's are small, they must be linearly related.‡ Hence, there is a relation which relates δv_k to δi's of the form

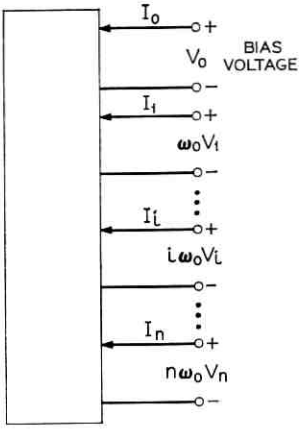


Fig. 1 — Pumped nonlinear system exchanging power at n carrier frequencies.

$$\delta v_k(t) = \sum_{j=1}^{n} \int_{-\infty}^{\infty} h_{kj}(t, \tau)\, \delta i_j(t - \tau)\, d\tau, \tag{13}$$

where $h_{kj}(t, \tau)$'s are functions of time t, as well as of time difference τ. Since the driving is periodic, if δi's were applied one period later, δv_k would be the same, except that it would be delayed by one period. This argument leads to the conclusion that $h_{kj}(t, \tau)$'s are periodic functions in t, with period $T_0 = 2\pi/\omega_0$ and can be expressed in a Fourier series of the form

‡ This is because only first-order terms in δv's and δi's are retained. Higher-order terms are assumed to be negligible even when first-order terms vanish.

$$h_{kj}(t, \tau) = \sum_{l=-\infty}^{\infty} (h_{kj})_l(\tau) \exp (jl\omega_0 t), \tag{14}$$

where $(h_{kj})_l(\tau)$ is a function of τ. Upon substituting (14) in (13), we find

$$\delta v_k(t) = \sum_{j=1}^{n} \sum_{l=-\infty}^{\infty} \exp (jl\omega_0 t) \int_{-\infty}^{\infty} (h_{kj})_l(\tau)\ \delta i_j(t - \tau)\ d\tau. \tag{15}$$

If $\delta v_k(t)$ is represented in the $\alpha - \beta$ form,

$$\delta v_k(t) = 2 \operatorname{Re} \{ V_{\alpha k} \exp [j(k\omega_0 + \omega)t] + V_{\beta k} \exp [j(-k\omega_0 + \omega)t]\}, \tag{16}$$

we find[‡]

$$\begin{bmatrix} V_{\alpha k} \\ V_{\beta k} \end{bmatrix} = \begin{bmatrix} Z_{\alpha k \alpha 1}\ Z_{\alpha k \beta 1} & \cdots & Z_{\alpha k \alpha k}\ Z_{\alpha k \beta k} & \cdots & Z_{\alpha k \alpha n}\ Z_{\alpha k \beta n} \\ Z_{\beta k \alpha 1}\ Z_{\beta k \beta 1} & \cdots & Z_{\beta k \alpha k}\ Z_{\beta k \beta k} & \cdots & Z_{\beta k \alpha n}\ Z_{\beta k \beta n} \end{bmatrix} \mathbf{I}, \tag{17}$$

where[§]

$$\mathbf{I} = \{I_{\alpha 1}\ , I_{\beta 1}\ , \cdots , I_{\alpha k}\ , I_{\beta k}\ , \cdots , I_{\alpha n}\ , I_{\beta n}\}. \tag{18}$$

III. SMALL-SIGNAL ANALYSIS OF PUMPED NONLINEAR SYSTEMS

For the nonlinear systems that we shall consider in this paper, we shall assume that the total voltage $v(t)$ across the nonlinear element is related to the current $i(t)$ through it by the equation[¶]

$$v(t) = F\{i(t)\}, \tag{19}$$

where $F\{i(t)\}$ is a single-valued functional of $i(t)$.

Assuming that there are carrier currents flowing in the system at frequencies $\pm i\omega_0$, $0 \leqq i \leqq n$, the spot frequency terminal behavior of this system at a difference frequency ω is given according to (17) by an equation of the form[||]

[‡] Essentially, we are discussing impedance formalism here which relates voltages to currents through an impedance matrix. Several other kinds of formalisms like scattering matrix representation or chain matrix representation can also be used to relate other desired sets of variables.

[§] A column matrix **a** is written in the form $\{a_1, a_2, \ldots , a_n\}$, the curly braces being used to identify it as a column matrix.

[¶] If there are any physical sources of fluctuations (such as noise sources) in the pumped nonlinear system, (20) is to be suitably modified. For the discussion of the case in which noise sources may be present in the pumped nonlinear system, see Ref. 11.

[||] It must be pointed out that this equation only relates the small-signal fluctuations present in the system and not the carrier voltages and currents.

$$\begin{bmatrix} V_{\alpha 0} \\ V_{\alpha 1} \\ V_{\beta 1} \\ \vdots \\ V_{\alpha i} \\ V_{\beta i} \\ \vdots \\ V_{\alpha n} \\ V_{\beta n} \end{bmatrix} = \begin{bmatrix} Z_{\alpha 0 \alpha 0} & Z_{\alpha 0 \alpha 1} & Z_{\alpha 0 \beta 1} & \cdots & Z_{\alpha 0 \alpha i} & Z_{\alpha 0 \beta i} & \cdots & Z_{\alpha 0 \alpha n} & Z_{\alpha 0 \beta n} \\ Z_{\alpha 1 \alpha 0} & Z_{\alpha 1 \alpha 1} & Z_{\alpha 1 \beta 1} & \cdots & Z_{\alpha 1 \alpha i} & Z_{\alpha 1 \beta i} & \cdots & Z_{\alpha 1 \alpha n} & Z_{\alpha 1 \beta n} \\ Z_{\beta 1 \alpha 0} & Z_{\beta 1 \alpha 1} & Z_{\beta 1 \beta 1} & \cdots & Z_{\beta 1 \alpha i} & Z_{\beta 1 \beta i} & \cdots & Z_{\beta 1 \alpha n} & Z_{\beta 1 \beta n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots & \vdots \\ Z_{\alpha i \alpha 0} & Z_{\alpha i \alpha 1} & Z_{\alpha i \beta 1} & \cdots & Z_{\alpha i \alpha i} & Z_{\alpha i \beta i} & \cdots & Z_{\alpha i \alpha n} & Z_{\alpha i \beta n} \\ Z_{\beta i \alpha 0} & Z_{\beta i \alpha 1} & Z_{\beta i \beta 1} & \cdots & Z_{\beta i \alpha i} & Z_{\beta i \beta i} & \cdots & Z_{\beta i \alpha n} & Z_{\beta i \beta n} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots & \vdots \\ Z_{\alpha n \alpha 0} & Z_{\alpha n \alpha 1} & Z_{\alpha n \beta 1} & \cdots & Z_{\alpha n \alpha i} & Z_{\alpha n \beta i} & \cdots & Z_{\alpha n \alpha n} & Z_{\alpha n \beta n} \\ Z_{\beta n \alpha 0} & Z_{\beta n \alpha 1} & Z_{\beta n \beta 1} & \cdots & Z_{\beta n \alpha i} & Z_{\beta n \beta i} & \cdots & Z_{\beta n \alpha n} & Z_{\beta n \beta n} \end{bmatrix} \begin{bmatrix} I_{\alpha 0} \\ I_{\alpha 1} \\ I_{\beta 1} \\ \vdots \\ I_{\alpha i} \\ I_{\beta i} \\ \vdots \\ I_{\alpha n} \\ I_{\beta n} \end{bmatrix}, \qquad (20)$$

where $V_{\alpha i}$ and $V_{\beta i}$ are the terminal voltages at frequencies $j\omega_0 + \omega$ and $-j\omega_0 + \omega$, respectively; and $I_{\alpha i}$ and $I_{\beta i}$ are the corresponding terminal currents. We would like to note here that $V_{\alpha 0} = V_{\beta 0}$ is the small-signal terminal voltage at the frequency ω. We shall, for brevity, write (20) as

$$(\mathbf{V}_{\alpha-\beta})_n = (\underline{Z}_{\alpha-\beta})_n(\mathbf{I}_{\alpha-\beta})_n \, . \qquad (21)$$

Let us now specifically consider a varactor diode which is pumped at a frequency ω_0 and its harmonics. The varactor model that we shall use is shown in Fig. 2. It is a variable capacitance in series with a constant resistance R_s .‡ The instantaneous varactor voltage $v(t)$ can be written as some function f of the charge, plus the drop across the series resistance R_s :

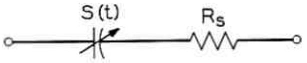


Fig. 2 — Varactor model.

$$v(t) = f[q(t)] + R_s i(t), \qquad (22)$$

where

$$q(t) = \int_{-\infty}^{t} i(t)\, dt. \qquad (23)$$

For such a varactor, we can make use of the small-signal equations given in Ref. 5 in order to obtain the impedance matrix $\underline{Z}_{\alpha-\beta}$ in (21). If the elastance $S(t)$ of the varactor diode can be written in a Fourier series of the form

$$S(t) = \sum_{k=-\infty}^{\infty} S_k \exp(jk\omega_0 t), \qquad (24)$$

‡ Mainly we shall be concerned with varactor diodes which are lossless in the succeeding sections of this paper. For a lossless varactor diode $R_s = 0$.

the matrix $(\underline{Z}_{\alpha-\beta})_n$ can be represented as§

$$(\underline{Z}_{\alpha-\beta})_n = \begin{bmatrix} \frac{S_0}{j\omega} & \frac{S_1^*}{j(\omega_0+\omega)} & \frac{S_1}{j(-\omega_0+\omega)} & \cdots & \frac{S_i^*}{j(i\omega_0+\omega)} & \frac{S_i}{j(-i\omega_0+\omega)} & \cdots & \frac{S_n^*}{j(n\omega_0+\omega)} & \frac{S_n}{j(-n\omega_0+\omega)} \\ \frac{S_1}{j\omega} & \frac{S_0}{j(\omega_0+\omega)} & \frac{S_2}{j(-\omega_0+\omega)} & \cdots & \frac{S_{i-1}^*}{j(i\omega_0+\omega)} & \frac{S_{i+1}}{j(-i\omega_0+\omega)} & \cdots & \frac{S_{n-1}^*}{j(n\omega_0+\omega)} & \frac{S_{n+1}}{j(-n\omega_0+\omega)} \\ \frac{S_1^*}{j\omega} & \frac{S_2^*}{j(\omega_0+\omega)} & \frac{S_0}{j(-\omega_0+\omega)} & \cdots & \frac{S_{i+1}^*}{j(i\omega_0+\omega)} & \frac{S_{i-1}}{j(-i\omega_0+\omega)} & \cdots & \frac{S_{n+1}^*}{j(n\omega_0+\omega)} & \frac{S_{n-1}}{j(-n\omega_0+\omega)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots & \vdots \\ \frac{S_i}{j\omega} & \frac{S_{i-1}}{j(\omega_0+\omega)} & \frac{S_{i+1}}{j(-\omega_0+\omega)} & \cdots & \frac{S_0}{j(i\omega_0+\omega)} & \frac{S_{2i}}{j(-i\omega_0+\omega)} & \cdots & \frac{S_{n-i}^*}{j(n\omega_0+\omega)} & \frac{S_{n+i}}{j(-n\omega_0+\omega)} \\ \frac{S_i^*}{j\omega} & \frac{S_{i+1}^*}{j(\omega_0+\omega)} & \frac{S_{i-1}^*}{j(-\omega_0+\omega)} & \cdots & \frac{S_{2i}^*}{j(i\omega_0+\omega)} & \frac{S_0}{j(-i\omega_0+\omega)} & \cdots & \frac{S_{n+i}^*}{j(n\omega_0+\omega)} & \frac{S_{n-i}}{j(-n\omega_0+\omega)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{S_n}{j\omega} & \frac{S_{n-1}}{j(\omega_0+\omega)} & \frac{S_{n+1}}{j(-\omega_0+\omega)} & \cdots & \frac{S_{n-i}}{j(i\omega_0+\omega)} & \frac{S_{n+i}}{j(-i\omega_0+\omega)} & \cdots & \frac{S_0}{j(n\omega_0+\omega)} & \frac{S_{2n}}{j(-n\omega_0+\omega)} \\ \frac{S_n^*}{j\omega} & \frac{S_{n+1}^*}{j(\omega_0+\omega)} & \frac{S_{n-1}^*}{j(-\omega_0+\omega)} & \cdots & \frac{S_{n+i}^*}{j(i\omega_0+\omega)} & \frac{S_{n-i}^*}{j(-i\omega_0+\omega)} & \cdots & \frac{S_{2n}^*}{j(n\omega_0+\omega)} & \frac{S_0}{j(-n\omega_0+\omega)} \end{bmatrix}. \quad (25)$$

§ We have put $R_s = 0$ in order to obtain (25). If $R_s \neq 0$, $(\underline{Z}_{\alpha-\beta})_n = \{(\underline{Z}_{\alpha-\beta})_n$ in (25)$\} + R_s\underline{1}_{2n+1}$. $\underline{1}_n$ is the unit matrix of order n.

Equation (25) shows that the impedance matrix $(Z_{\alpha-\beta})_n$ always exists for a pumped varactor diode as long as the elastance $S(t)$ is expressible in the form (24).

Let us now assume that the input carrier frequency is $l\omega_0$, $1 \leqq l \leqq n$; and that the output carrier frequency is $s\omega_0$, $1 \leqq s \leqq n$ (see Fig. 3). Let us also assume that the terminal constraints at other carrier frequencies are such that

$$\mathbf{V}' = -Z'_{\alpha-\beta}\mathbf{I}', \tag{26}$$

where $\mathbf{V}'$ is an $\alpha - \beta$ terminal voltage column matrix given by

$$\mathbf{V}' = \begin{bmatrix} V_{\alpha 0} \\ V_{\alpha 1} \\ V_{\beta 1} \\ \vdots \\ V_{\alpha(l-1)} \\ V_{\beta(l-1)} \\ V_{\alpha(l+1)} \\ V_{\beta(l+1)} \\ \vdots \\ V_{\alpha(s-1)} \\ V_{\beta(s-1)} \\ V_{\alpha(s+1)} \\ V_{\beta(s+1)} \\ \vdots \\ V_{\alpha n} \\ V_{\beta n} \end{bmatrix}. \tag{27}$$

$\mathbf{I}'$ is the corresponding terminal current column matrix. $Z'_{\alpha-\beta}$ is the impedance matrix determined by the terminal constraints imposed by the external circuits on the system. These terminal constraints at all carrier frequencies excluding $l\omega_0$ and $s\omega_0$ are assumed to be known. Even though the currents flowing in the varactor are not limited by the diode in the range of available frequencies, it is assumed that the external circuits are such as to offer an infinite impedance at any frequency very far from the carrier frequencies present in the multiplier. This enables us to consider the multiplier as a finite port multifrequency system.

It may now be seen that by using (25) and (26) we can obtain a relation between $V_{\alpha l}$, $V_{\beta l}$, $I_{\alpha l}$, $I_{\beta l}$, $V_{\alpha s}$, $V_{\beta s}$, $I_{\alpha s}$, and $I_{\beta s}$. In particular we can write‡

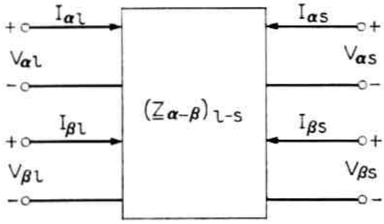


Fig. 3 — Small-signal terminal behavior of pumped nonlinear twoport.

$$\begin{bmatrix} V_{\alpha l} \\ V_{\beta l} \\ V_{\alpha s} \\ V_{\beta s} \end{bmatrix} = \begin{bmatrix} Z''_{\alpha l \alpha l} & Z''_{\alpha l \beta l} & Z''_{\alpha l \alpha s} & Z''_{\alpha l \beta s} \\ Z''_{\beta l \alpha l} & Z''_{\beta l \beta l} & Z''_{\beta l \alpha s} & Z''_{\beta l \beta s} \\ Z''_{\alpha s \alpha l} & Z''_{\alpha s \beta l} & Z''_{\alpha s \alpha s} & Z''_{\alpha s \beta s} \\ Z''_{\beta s \alpha l} & Z''_{\beta s \beta l} & Z''_{\beta s \alpha s} & Z''_{\beta s \beta s} \end{bmatrix} \begin{bmatrix} I_{\alpha l} \\ I_{\beta l} \\ I_{\alpha s} \\ I_{\beta s} \end{bmatrix} \tag{28}$$

or

$$(\mathbf{V}_{\alpha-\beta})_{l-s} = (\underline{Z}_{\alpha-\beta})_{l-s}(\mathbf{I}_{\alpha-\beta})_{l-s} . \tag{29}$$

Equation (29) relates the small-signal fluctuations existing at input and output terminals of a pumped varactor diode. In case one is interested in relating the AM and PM fluctuations at the input and output terminals of a pumped varactor diode, we make use of (11). If φ_{vl}, φ_{vs}, φ_{il}, and φ_{is} are the phase angles of carrier voltages and currents at the input and output of a pumped varactor diode we get the following equation which relates the different fluctuations:

‡ In certain cases it is possible that the matrix $(\underline{Z}_{\alpha-\beta})_{l-s}$ does not exist. Even though $(\underline{Z}_{\alpha-\beta})_{l-s}$ may not exist, in most cases of practical interest, we can always find a relation between the terminal voltages and currents at the sideband frequencies in the vicinity of input and output carriers. This will be shown to be true in the case of a tripler which is discussed elsewhere in this paper. However, the matrix $(\underline{Z}_{\alpha-\beta})_n$ always exists for a pumped varactor diode.

$$\begin{bmatrix} V_{al} \\ V_{pl} \\ V_{as} \\ V_{ps} \end{bmatrix} = (\underline{Z}_{a-p})_{l-s} \begin{bmatrix} I_{al} \\ I_{pl} \\ I_{as} \\ I_{ps} \end{bmatrix}, \tag{30}$$

where

$$(\underline{Z}_{a-p})_{l-s} = (\lambda_v)_{l-s}(\underline{Z}_{\alpha-\beta})_{l-s}\{(\lambda_i)_{l-s}\}^{-1}, \tag{31}$$

$$(\lambda_v)_{l-s} = \left[\begin{array}{c|c} \lambda_{vl} & 0 \\ \hline 0 & \lambda_{vs} \end{array}\right], \tag{32}$$

and

$$(\lambda_i)_{l-s} = \left[\begin{array}{c|c} \lambda_{il} & 0 \\ \hline 0 & \lambda_{is} \end{array}\right]. \tag{33}$$

The matrices λ's are given as in (12).

Once we have obtained the impedance matrix representation for the pumped varactor diode other kinds of representations like scattering matrix representation or chain matrix representation could be derived for any specific application or convenience. Mutual relations between these representations are given in Refs. 11 and 12, and we shall not discuss them in this paper. Scattering matrix representation of lossless abrupt-junction varactor multipliers is extensively treated in later sections of this paper.

IV. SCATTERING PARAMETERS FOR PUMPED NONLINEAR SYSTEMS

The total voltage $v(t)$ in the vicinity of a carrier at frequency $\pm k\omega_0$ can be represented as in (5) or (6). $v_{ak}(t)$ can be interpreted as a small perturbation on the amplitude $| V_k |$ of the carrier, and $v_{pk}(t)$ as a perturbation on the phase $k\omega_0 t + \varphi_{vk}$ of the carrier. Since the device acts as a time-variant linear device to the fluctuations and since superposition holds, $v_{ak}(t)$ and $v_{pk}(t)$ can be represented as in (7) and (8).

The voltage AM and PM modulation indexes at the carrier frequency $k\omega_0$ may, therefore, be defined as

$$m_{vk} = \frac{V_{ak}}{| V_k |} \tag{34}$$

and

$$\theta_{vk} = \frac{V_{pk}}{| V_k |}. \tag{35}$$

The AM and PM indexes at the input and output of a pumped nonlinear system are, according to (30), related by an equation of the form

$$\begin{bmatrix} m_{vl} \\ \theta_{vl} \\ m_{vs} \\ \theta_{vs} \end{bmatrix} = (\underline{Z}_{m-\theta})_{l-s} \begin{bmatrix} m_{il} \\ \theta_{il} \\ m_{is} \\ \theta_{is} \end{bmatrix}, \tag{36}$$

where

$$(\underline{Z}_{m-\theta})_{l-s} = \begin{bmatrix} |V_l|^{-1} & & & 0 \\ & |V_l|^{-1} & & \\ & & |V_s|^{-1} & \\ 0 & & & |V_s|^{-1} \end{bmatrix} (\underline{Z}_{a-p})_{l-s} \cdot \begin{bmatrix} |I_l| & & & 0 \\ & |I_l| & & \\ & & |I_s| & \\ 0 & & & |I_s| \end{bmatrix}. \tag{37}$$

It is assumed that carrier voltages at frequencies $l\omega_0$ and $s\omega_0$ are nonzero.

Until now we have exclusively used the impedance formalism to describe the properties of the pumped nonlinear system at the sideband frequencies. The choice of an appropriate formalism is particularly important in theoretical studies where important properties of the system may be obscured by complicated equations. The scattering parameters of a system are a set of quantities which can describe the performance of the system under any specified terminating conditions, just as the impedance (or admittance) quantities can, but while the scattering coefficients may not be particularly convenient for short or open-circuit computations, they may be applied in a relatively simple fashion when the network is terminated in a prescribed load impedance. Since we will be mainly interested in studying proper terminations for the system in order to realize certain desirable characteristics, scattering matrix formulation to describe AM and PM fluctuations in pumped nonlinear systems seems to be the most desirable.[1,3,4,13] Equation (36) relates the AM and PM indexes or normalized voltages and currents at the input and output of a pumped nonlinear system.‡ The incident

‡ Most of these concepts can be extended in a straightforward fashion if the pumped nonlinear system has more than two accessible ports.

and reflected AM and PM indexes (see Fig. 4) can, therefore, be written[13] as

$$(m_i)_j = \tfrac{1}{2}(m_{vj} + m_{ij}), \qquad j = l, s, \tag{38}$$

$$(m_r)_j = \tfrac{1}{2}(m_{vj} - m_{ij}), \qquad j = l, s, \tag{39}$$

$$(\theta_i)_j = \tfrac{1}{2}(\theta_{vj} + \theta_{ij}), \qquad j = l, s, \tag{40}$$

and

$$(\theta_r)_j = \tfrac{1}{2}(\theta_{vj} - \theta_{ij}), \qquad j = l, s. \tag{41}$$

Using (37) through (41), we can now obtain the following scattering matrix representation for a pumped nonlinear system:

$$\begin{bmatrix} (m_r)_l \\ (m_r)_s \\ (\theta_r)_l \\ (\theta_r)_s \end{bmatrix} = \left[\begin{array}{c:c} \underline{S}_{aa} & \underline{S}_{ap} \\ \hdashline \underline{S}_{pa} & \underline{S}_{pp} \end{array}\right] \begin{bmatrix} (m_i)_l \\ (m_i)_s \\ (\theta_i)_l \\ (\theta_i)_s \end{bmatrix}. \tag{42}$$

The relation between scattering matrix in (42) and impedance matrix in (37) is easily derived.[13] In our case, this is given by

$$(\underline{S})_{l-s} = \underline{1}_4 - 2\{\underline{1}_4 + (\underline{Z}'_{m-\theta})_{l-s}\}^{-1}, \tag{43}$$

where $\underline{1}_4$ is the unit matrix of order 4, and where $\underline{Z}'_{m-\theta}$ is related to $\underline{Z}_{m-\theta}$ in (37) by

$$(\underline{Z}'_{m-\theta})_{l-s} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} (\underline{Z}_{m-\theta})_{l-s} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{44}$$

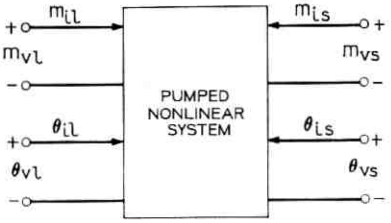


Fig. 4 — Representation of AM and PM fluctuations in a pumped nonlinear system.

The scattering matrix which relates the small-signal fluctuations of a pumped nonlinear system is, therefore, given by (43). We assume that the matrix $\underline{1}_4 + (\underline{Z}'_{m-\theta})_{l-s}$ is nonsingular.

We would like to point out here that the matrices $\underline{S}_{aa}$, $\underline{S}_{ap}$, $\underline{S}_{pa}$, and $\underline{S}_{pp}$ are all square matrices of second order. For reasons which are evident from (42), the matrix $\underline{S}_{aa}$ will be referred to as AM scattering matrix, $\underline{S}_{ap}$ as the AM-PM scattering matrix, $\underline{S}_{pa}$ as the PM-AM scattering matrix, and $\underline{S}_{pp}$ as the PM scattering matrix.

V. SCATTERING MATRICES OF NOMINALLY DRIVEN LOSSLESS ABRUPT-JUNCTION VARACTOR FREQUENCY MULTIPLIERS‡

The theory developed in the preceding sections will be utilized from hereon in order to obtain the scattering parameters of nominally driven lossless abrupt-junction varactor frequency multipliers. The elastance $S(t)$ of the varactor diode as it is pumped is assumed to be given by§

$$S(t) = \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} S_k \exp(jk\omega_0 t). \tag{45}$$

In this section we shall first obtain the scattering matrix of a varactor doubler whose input and output circuits are tuned. In the later part of this section the scattering parameters of a tripler, whose input, output, and idler circuits are tuned, are also derived. The discussion of the scattering parameters of multipliers of higher order is postponed to later sections of this paper. For all the multipliers considered in this paper it is assumed that the bias circuit is properly designed so that there are no currents flowing at the sideband frequencies ±ω.[6] Even though the currents flowing in the varactor are themselves not limited by the diode in the range of available frequencies we assume that the external circuits connected to the diode are such that they allow currents to flow in the varactor if and only if the frequency spectrum of these currents is in the vicinity of input, output, and idler carrier frequencies. This enables us to consider the multiplier as a finite port multifrequency system.

‡ See also Refs. 3 and 4 for alternate derivation of some of these results.

§ The average elastance S_0 of the varactor diode can always be included with the external circuit. The assumption that $S_0 = 0$ made in this section does not, therefore, involve any loss of generality.

5.1 *Scattering Parameters of a Doubler.*

In a doubler, the only nonzero elastance coefficients are $S_{\pm 1}$, and $S_{\pm 2}$. The impedance matrix $(Z_{\alpha-\beta})_2$ in (25) is represented as (see Fig. 5)

$$\begin{bmatrix} V_{\alpha 1} \\ V_{\beta 1} \\ V_{\alpha 2} \\ V_{\beta 2} \end{bmatrix} = \begin{bmatrix} 0 & \dfrac{S_2}{j(-\omega_0 + \omega)} & \dfrac{S_1^*}{j(2\omega_0 + \omega)} & 0 \\ \dfrac{S_2^*}{j(\omega_0 + \omega)} & 0 & 0 & \dfrac{S_1}{j(-2\omega_0 + \omega)} \\ \dfrac{S_1}{j(\omega_0 + \omega)} & 0 & 0 & 0 \\ 0 & \dfrac{S_1^*}{j(-\omega_0 + \omega)} & 0 & 0 \end{bmatrix} \begin{bmatrix} I_{\alpha 1} \\ I_{\beta 1} \\ I_{\alpha 2} \\ I_{\beta 2} \end{bmatrix}. \qquad (46)$$

We shall now assume that input and output circuits are tuned which usually gives near optimum efficiency for a doubler.[5,7,8] We also assume that

$$\frac{\omega}{\omega_0} \ll 1. \qquad (47)$$

With these two assumptions, we can write the following matrix equation for a doubler:

$$\begin{bmatrix} V_{\alpha 1} \\ V_{\beta 1} \\ V_{\alpha 2} \\ V_{\beta 2} \end{bmatrix} = \begin{bmatrix} 0 & \dfrac{|S_2|}{\omega_0} & \dfrac{|S_1|}{2\omega_0} & 0 \\ \dfrac{|S_2|}{\omega_0} & 0 & 0 & \dfrac{|S_1|}{2\omega_0} \\ -\dfrac{|S_1|}{\omega_0} & 0 & 0 & 0 \\ 0 & -\dfrac{|S_1|}{\omega_0} & 0 & 0 \end{bmatrix} \begin{bmatrix} I_{\alpha 1} \\ I_{\beta 1} \\ I_{\alpha 2} \\ I_{\beta 2} \end{bmatrix}. \qquad (48)$$

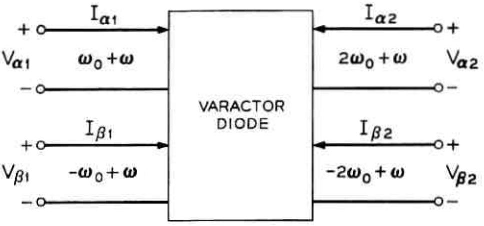


Fig. 5 — Small-signal behavior of a varactor doubler.

The phase angles‡ of carrier voltages and currents are given by[5]

$$\varphi_{V1} = 0, \tag{49}$$

$$\varphi_{I1} = 0, \tag{50}$$

$$\varphi_{V2} = \pi, \tag{51}$$

and

$$\varphi_{I2} = \pi. \tag{52}$$

From (31) the impedance matrix $(Z_{a-p})_2$ is, therefore, written as§

$$(Z_{a-p})_2 = \begin{bmatrix} \frac{|S_2|}{\omega_0} & 0 & -\frac{|S_1|}{2\omega_0} & 0 \\ 0 & -\frac{|S_2|}{\omega_0} & 0 & -\frac{|S_1|}{2\omega_0} \\ \frac{|S_1|}{\omega_0} & 0 & 0 & 0 \\ 0 & \frac{|S_1|}{\omega_0} & 0 & 0 \end{bmatrix}. \tag{53}$$

Equation (53) clearly shows that in a doubler which is properly tuned and whose bias circuit is properly designed, there is no amplitude-to-phase and phase-to-amplitude conversion.[6]

It is shown in Appendix A that

$$\frac{|V_1|}{|I_1|} = \frac{|S_2|}{\omega_0}, \tag{54}$$

$$\frac{|V_2|}{|I_2|} = \frac{|S_1|^2}{4|S_2|\omega_0}, \tag{55}$$

$$\frac{|V_1|}{|I_2|} = \frac{|S_1|}{2\omega_0}, \tag{56}$$

‡ Without loss of generality the phase angle of input carrier voltage at frequency ω_0 is assumed to be zero. φ_{I2} is the phase angle of the current through the load connected to the doubler.

§ The matrix $(Z)_{1-s}$ will be written as $(Z)_s$ in case this does not lead to any ambiguity.

and

$$\frac{|V_2|}{|I_1|} = \frac{|S_1|}{2\omega_0}. \tag{57}$$

According to (37) and (53), the matrix $(\underline{Z}_{m-\theta})_2$ is represented as

$$(\underline{Z}_{m-\theta})_2 = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & -1 \\ 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \end{bmatrix}. \tag{58}$$

The scattering matrix of a doubler, whose input and output circuits are tuned is, therefore, according to (43)

$$(\underline{S})_2 = \left[\begin{array}{cc:cc} \frac{1}{2} & -\frac{1}{2} & \multicolumn{2}{c}{\multirow{2}{*}{0}} \\ 1 & 0 & & \\ \hdashline \multicolumn{2}{c:}{\multirow{2}{*}{0}} & 0 & -1 \\ & & 2 & 1 \end{array}\right]; \tag{59}$$

$$(\underline{S}_{aa})_2 = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ 1 & 0 \end{bmatrix}, \tag{60}$$

$$(\underline{S}_{ap}) = \underline{0}, \tag{61}$$

$$(\underline{S}_{pa}) = \underline{0}, \tag{62}$$

and

$$(\underline{S}_{pp}) = \begin{bmatrix} 0 & -1 \\ 2 & 1 \end{bmatrix}. \tag{63}$$

5.2 *Scattering Parameters of a Tripler.*

In a tripler, carrier currents flow at the frequencies $\pm\omega_0$, $\pm 2\omega_0$, and $\pm 3\omega_0$. The impedance matrix $(\underline{Z}_{\alpha-\beta})_3$ is given as

$$\begin{bmatrix} V_{\alpha 1} \\ V_{\beta 1} \\ V_{\alpha 2} \\ V_{\beta 2} \\ V_{\alpha 3} \\ V_{\beta 3} \end{bmatrix} = \begin{bmatrix} 0 & \dfrac{S_2}{j(-\omega_0 + \omega)} & \dfrac{S_1^*}{j(2\omega_0 + \omega)} & \dfrac{S_3}{j(-2\omega_0 + \omega)} & \dfrac{S_2^*}{j(3\omega_0 + \omega)} & 0 \\ \dfrac{S_2^*}{j(\omega_0 + \omega)} & 0 & \dfrac{S_3^*}{j(2\omega_0 + \omega)} & \dfrac{S_1}{j(-2\omega_0 + \omega)} & 0 & \dfrac{S_2}{j(-3\omega_0 + \omega)} \\ \dfrac{S_1}{j(\omega_0 + \omega)} & \dfrac{S_3}{j(-\omega_0 + \omega)} & 0 & 0 & \dfrac{S_1^*}{j(3\omega_0 + \omega)} & 0 \\ \dfrac{S_3^*}{j(\omega_0 + \omega)} & \dfrac{S_1^*}{j(-\omega_0 + \omega)} & 0 & 0 & 0 & \dfrac{S_1}{j(-3\omega_0 + \omega)} \\ \dfrac{S_2}{j(\omega_0 + \omega)} & 0 & \dfrac{S_1}{j(2\omega_0 + \omega)} & 0 & 0 & 0 \\ 0 & \dfrac{S_2^*}{j(-\omega_0 + \omega)} & 0 & \dfrac{S_1^*}{j(-2\omega_0 + \omega)} & 0 & 0 \end{bmatrix} \begin{bmatrix} I_{\alpha 1} \\ I_{\beta 1} \\ I_{\alpha 2} \\ I_{\beta 2} \\ I_{\alpha 3} \\ I_{\beta 3} \end{bmatrix}. \qquad (64)$$

If we now assume that

$$\frac{\omega}{\omega_0} \ll 1 \tag{65}$$

and also that input, output, and idler circuits are tuned and that[5]

$$\varphi_{V1} = 0, \tag{66}$$

$$\varphi_{I1} = 0, \tag{67}$$

$$\varphi_{V2} = 0, \tag{68}$$

$$\varphi_{I2} = 0, \tag{69}$$

$$\varphi_{V3} = \pi, \tag{70}$$

and

$$\varphi_{I3} = \pi, \tag{71}$$

we can show that the impedance matrix $(Z_{a-p})_3$ does not exist for a tripler.‡ It is also shown in Appendix A that

$$| S_3 | = | S_1 |/2. \tag{72}$$

Equations (64)-(72) show that§

$$\frac{I_{a1}}{I_{a3}} = -\frac{2}{3}, \tag{73}$$

$$\frac{V_{a1}}{V_{a3}} = \frac{3}{2}, \tag{74}$$

$$\frac{I_{p1}}{I_{p3}} = -\frac{2}{9}, \tag{75}$$

and

$$\frac{V_{p1}}{V_{p3}} = \frac{1}{2}. \tag{76}$$

Equations (73) through (76) show that even though the amplitude-phase impedance matrix may not exist for a pumped nonlinear system (such as

‡ The termination at the idler port just tunes out the average elastance of the varactor diode at the carrier frequency $2\omega_0$.

§ We note that a tripler behaves like an ideal transformer of ratio 3/2 to the amplitude components of the fluctuations. Higher order terms in ω/ω_0 are assumed to be negligible, even when first-order terms vanish. The frequency-dependence usually introduced by the external idler termination, therefore, does not appear in the scattering matrix of the tripler.

a tripler) we are able to find a relation between the different terminal variables. These relations are sufficient to obtain the scattering parameters of the network.[13] We immediately note from (73) through (76), that there is no amplitude-to-phase or phase-to-amplitude conversion in a tripler. Accordingly,

$$(\underline{S}_{ap})_3 = \underline{0} \tag{77}$$

and

$$(\underline{S}_{pa})_3 = \underline{0}. \tag{78}$$

It is shown in Appendix A that

$$\frac{|V_1|}{|V_3|} = \frac{3}{2} \tag{79}$$

and

$$\frac{|I_1|}{|I_3|} = \frac{2}{3}. \tag{80}$$

From (34), (35), (73) through (76), (79), and (80), we can show that

$$(\underline{S}_{aa})_3 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{81}$$

and

$$(\underline{S}_{pp})_3 = \begin{bmatrix} 0 & 3^{-1} \\ 3 & 0 \end{bmatrix}. \tag{82}$$

Equation (81) could have been written down by noting that a tripler behaves like an ideal transformer of ratio $\frac{3}{2}$ to the amplitude components of its fluctuations.

The scattering parameters of a tripler, whose input, output, and idler circuits are tuned, are therefore, given by

$$(\underline{S})_3 = \left[\begin{array}{cc|cc} 0 & 1 & \multicolumn{2}{c}{\multirow{2}{*}{0}} \\ 1 & 0 & & \\ \hline \multicolumn{2}{c|}{\multirow{2}{*}{0}} & 0 & 3^{-1} \\ & & 3 & 0 \end{array}\right]. \tag{83}$$

In order to obtain scattering parameters of multipliers of higher order with the least number of idlers we shall show that a multiplier of order 2^n is equivalent to a cascade of n doublers, a multiplier of order 3^s is equivalent to a cascade of s triplers, and that a multiplier of order $2^n 3^s$ is equivalent to a cascade of n doublers and s triplers.

VI. EQUIVALENCE OF A MULTIPLIER OF ORDER $N = N_1 \times N_2$ TO A CASCADE OF TWO MULTIPLIERS OF ORDER N_1 AND N_2‡

In this section it is shown that, if the idler configuration of a multiplier of order $N = N_1 \times N_2$ satisfies certain conditions, then the multiplier can be represented as a cascade connection of two multipliers of order N_1 and N_2. In this and in the following two sections, no restriction is placed on the type of input, output, and idler circuits. Therefore the following discussion also applies to the case of a multiplier which is lossy and not tuned.

Consider an abrupt-junction varactor multiplier of order $N = N_1 \times N_2$. Let B denote the set of all positive and negative integers which are equal in magnitude to the orders of the various harmonics present in the varactor current. Furthermore, let B_1 indicate the subset of B which consists of the elements of magnitude equal to or less than N_1, and let B_2 denote the subset of B which consists of the elements of magnitude equal to or greater than N_1. In this section it will be shown that, if B satisfies the following condition:

$$B \text{ is such that, if } (r, s, h) \text{ is a subset of } B \text{ and if } r + s + h = 0,$$

$$\text{then either } (r, s, h) \subset B_1 \text{ or } (r, s, h) \subset B_2, \tag{84}$$

then the multiplier is equivalent to a cascade connection of two multipliers of order N_1 and N_2. Notice that an abrupt-junction multiplier of order $N = N_1 \times N_2$ which satisfies (84) must have an idler at the harmonic $N_1\omega_0$. In fact, for such a multiplier this idler is necessary in order to produce harmonics of order higher than $N_1\omega_0$.[5]

Consider then a multiplier of order $N_1 \times N_2$ which satisfies (84) and let it be represented by the very general equivalent circuit shown in Fig. 6. The generator $v_g(t)$ is sinusoidal and is of frequency ω_0. $Z(\omega)$, the impedance of the external multiplier circuit as seen from the nonlinear part of the capacitance of the varactor, is assumed to be finite only in the vicinity of the input, output, and idler frequencies. Since $Z(\omega)$ includes the average elastance and the series resistance of the varactor, the nonlinear capacitor of Fig. 6 has a q-v characteristic of the type: $v = Aq^2$, in which A is a constant multiplier. Consider now

‡ The results of Sections VI, VII, and VIII represent extensions of an earlier result, the equivalence demonstrated in Ref. 1 for the case $N = 2^n$.

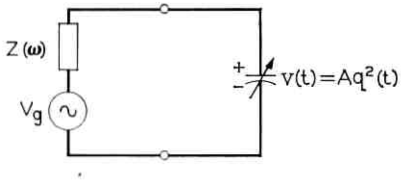


Fig. 6 — Varactor harmonic generator.

the circuit shown in Fig. 7. It will be shown that this circuit provides an alternative and complete representation of the multiplier of order $N_1 \times N_2$. The two nonlinear capacitors of Fig. 7 and that of Fig. 6 are assumed to have the same q-v characteristics. The three networks F_1, F_2, F_3 are ideal filters which have zero impedances at the carrier frequencies which satisfy respectively the relations $\omega < N_1\omega_0$, $\omega = N_1\omega_0$, $\omega > N_1\omega_0$, and also at their sidebands. Furthermore, at frequencies different from these, they have infinite impedance.

Before beginning the demonstration of the equivalence of the two circuits of Figs. 6, and 7, it may be profitable to examine briefly the behavior of the circuit of Fig. 7. The circuit of Fig. 7 represents the cascade connection of two multipliers of order N_1 and N_2. More precisely, consider the network connected on the left side of the first capacitor. For $\omega < N_1\omega_0$, it is equivalent to the network connected to the capacitor of Fig. 6. Therefore, it pumps at $\omega = \omega_0$ the first capacitor of Fig. 7 and, in addition, it provides the proper idler terminations for the generation of the harmonic $N_1\omega_0$. A current component at this harmonic is therefore generated by the first capacitor and it flows in the second loop shown in Fig. 7. The second capacitor is thus pumped at $\omega = N_1\omega_0$ by this current. Note that the network connected to its right provides the proper idler terminations for the generation of the

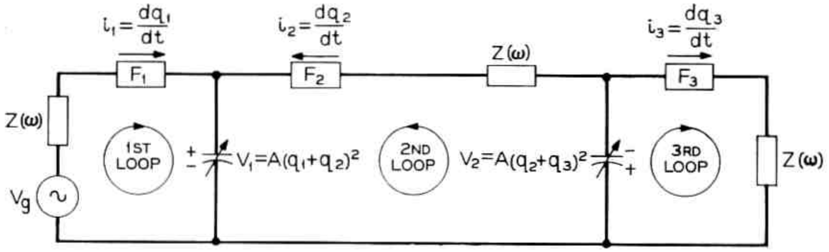


Fig. 7 — Equivalence of a multiplier of order N_1N_2 to a cascade of multipliers of orders N_1 and N_2.

output harmonic $N_1N_2\omega_0$. Therefore, the second capacitor delivers power at this harmonic to the network on its right.

Proof: First, consider the circuit of Fig. 6. The nonlinear capacitor has a q-v characteristic of the type: $v = Aq^2$. Thus, V_{-h}, the complex amplitude of $v(t)$ at the frequency $\omega = -h\omega_0$, is related to the various complex amplitudes of $q(t)$ through the relation

$$V_{-h} = A \sum_{\substack{(r,s) \subset B \\ r+s+h=0}} Q_r Q_s \,. \tag{85}$$

By introducing the constraint given by the linear circuit at $\omega = -h\omega_0$, one obtains

$$V_{g,-h} + jh\omega_0 Z(\omega) Q_{-h} = A \sum_{\substack{(r,s) \subset B \\ r+s+h=0}} Q_r Q_s \,, \qquad h \varepsilon B, \tag{86}$$

where $V_{g,-h}$ is the complex amplitude of $v_g(t)$, and is zero for $|h| \neq 1$.

Relations (86) give the equilibrium equations of the circuit of Fig. 6 and they determine the various charge amplitudes Q_1, Q_2, etc. Notice that in the summation of the righthand side of (86) one has $r+s+h=0$ and $(r, s, h) \subset B$. Therefore, from (84) one obtains the following three cases: if $|h| < N_1$, then $(r, s, h) \subset B_1$; if $|h| = N_1$, then, depending on the values of r, s, either $(r, s, h) \subset B_1$ or $(r, s, h) \subset B_2$; if $|h| > N_1$, then $(r, s, h) \subset B_2$. Accordingly, (86) can be written as

$$V_{g,-h} + jh\omega_0 Z Q_{-h} = A \sum_{\substack{(r,s) \subset B_1 \\ r+s+h=0}} Q_r Q_s \,, \quad \text{if} \quad |h| < N_1 \,. \tag{87}$$

$$jh\omega_0 Z Q_{-h} = A \{ \sum_{\substack{(r,s) \subset B_1 \\ r+s+h=0}} Q_r Q_s + \sum_{\substack{(r,s) \subset B_2 \\ r+s+h=0}} Q_r Q_s \}, \quad \text{if} \quad |h| = N_1 \tag{88}$$

$$jh\omega_0 Z Q_{-h} = A \sum_{\substack{(r,s) \subset B_2 \\ r+s+h=0}} Q_r Q_s \,, \quad \text{if} \quad |h| > N_1 \,. \tag{89}$$

Let now the circuit of Fig. 7 be examined. Consider the charges $q_1(t)$, $q_2(t)$ and $q_3(t)$ flowing through the three filters F_1, F_2, and F_3. Notice that $q_1(t) + q_2(t)$ is the total charge of the first capacitor, and that $q_2(t) + q_3(t)$ is the total charge of the second capacitor. Because of the characteristics of the three filters F_1, F_2, F_3, $q_1(t) + q_2(t)$ contains (all and) only the frequencies $r\omega_0$ for which $r \varepsilon B_1$. Similarly, $q_2(t)+q_3(t)$ contains only the frequencies for which $r \varepsilon B_2$. Now consider the total charge

$$q'(t) = q_1(t) + q_2(t) + q_3(t) \tag{90}$$

and let the symbol ( )′ distinguish the complex amplitudes of $q'(t)$ from those of $q(t)$. It will be shown that $q(t) = q'(t)$; more precisely, it will be shown that the two circuits of Figs. 6 and 7 have the same equilibrium equations at the carriers.

First notice that, for $\omega = -h\omega_0$, the complex amplitude of the voltage of the first varactor of Fig. 7 is

$$A \sum_{\substack{(r,s)\subset B_1\\ r+s+h=0}} Q'_r Q'_s \tag{91}$$

and that of the second varactor is

$$A \sum_{\substack{(r,s)\subset B_2\\ r+s+h=0}} Q'_r Q'_s \,. \tag{92}$$

Next, notice that the equilibrium equations of the circuit of Fig. 7 for $|\omega| < N_1\omega_0$ are obtained by applying Kirchhoff's law to the first loop of Fig. 7. Similarly, for $|\omega| = N_1\omega_0$, one considers the second loop and, for $|\omega| > N_1\omega_0$, one considers the third loop. Therefore, by taking into account (91) and (92), one obtains that the equilibrium equations of the circuit of Fig. 7 are given by (87), (88), and (89), with Q_r, Q_s replaced by Q'_r, Q'_s. Therefore, $q'(t) = q(t)$.

The preceding demonstration has shown that the two circuits of Figs. 6 and 7 are equivalent at the carrier frequencies. At the various sideband frequencies, the equivalence is demonstrated in very much the same way. Since the elastance coefficients of the two circuits are equal, one finds that the sets of small-signal equations of the two circuits are equal.

VII. DISCUSSION OF THE TWO PARTICULAR CASES $N = N_1 \times 2$ AND $N = N_1 \times 3$

In this section the two particular cases $N_2 = 2$ and $N_2 = 3$ will be examined. More precisely, it will be shown that in these two cases condition (84) becomes:

If $N_2 = 2$, the two highest harmonics present in the varactor current are $N_1\omega_0$ and $2N_1\omega_0$. (93)

If $N_2 = 3$, the three highest harmonics are $N_1\omega_0$, $2N_1\omega_0$, $3N_1\omega_0$. (94)

The demonstrations are very much the same in the two cases and therefore, only the case $N_2 = 2$ will be considered.

Proof: Consider the case $N_2 = 2$ and suppose that (93) is satisfied. Then, consider the three sets B, B_1, B_2 defined in the preceding section. From (93) one has $B_2 = (-2N_1, -N_1, N_1, 2N_1)$.

Now, consider a subset (r, s, h) of B and suppose that $r + s + h = 0$.

First, notice that (r, s, h) cannot belong to both B_1 and B_2 because this would give $| r | = | s | = | h | = N_1$, which violates the hypothesis $r + s + h = 0$. It is therefore, sufficient to prove that if (r, s, h) does not belong to B_1, then it must belong to B_2.

Suppose therefore, that (r, s, h) does not belong to B_1. Then one of the three elements r, s, h has magnitude equal to $2N_1$ and, since $r + s + h = 0$, the remaining two elements have magnitude equal to N_1. Therefore, $(r, s, h) \subset B_2$.

The conclusion is that, if $(r, s, h) \subset B$ and $r + s + h = 0$, then either $(r, s, h) \subset B_1$ or $(r, s, h) \subset B_2$. This concludes the demonstration.

VIII. EQUIVALENCE OF A MULTIPLIER OF ORDER 2^n3^s TO A CHAIN OF DOUBLERS AND TRIPLERS

Consider an abrupt-junction varactor multiplier of order $N = 2^n3^s$ which has the least number of idlers. In this section it will be shown that this multiplier is equivalent to a chain of doublers and triplers. The order in which the various doublers and triplers are connected depends on the particular idler configuration. This will be clarified by the following demonstration which shows how to derive the equivalent chain of multipliers.

Proof: Since the multiplier has the least number of idlers, there are two cases:[5] either the highest idler frequency is $N\omega_0/2$, or the two highest idler frequencies are $N\omega_0/3$, $2N\omega_0/3$. In both cases, the results of the preceding sections are applicable and therefore, the multiplier can be represented by means of a cascade of two multipliers of order N_1 and N_2. Note that N_2 is 2 in the first case and 3 in the second case. Note, furthermore, that if $n + s = 2$, then either $N_1 = 2$ or $N_1 = 3$, and therefore the demonstration would end at this point.

If $n + s > 2$, on the other hand, then $N_1 > 3$ and the decomposition of the multiplier of order 2^n3^s into two multipliers of lower order can evidently be continued by the decomposition of the first of the two multipliers, the multiplier of order N_1. If this process is carried as far as possible, the final structure will be a chain of doublers and triplers.

It is important to point out that the results of this and the preceding section can be generalized in the following way:

> An abrupt-junction varactor multiplier of order $N = N_1 \times N_2 \times \cdots \times N_n$ which has the least number of idlers can be represented by a cascade of n multipliers of order N_1, N_2, etc., each with minimum number of idlers. (95)

In fact, if $n = 2$, then (95) follows directly from the equivalence demonstrated in Section VI because it can be shown that a multiplier which has the least number of idlers satisfies (84).

If $n > 2$, then the multiplier can be decomposed into multipliers of lower order as it has been done for the particular case $N = 2^n 3^s$.

IX. SCATTERING RELATIONS FOR HIGHER-ORDER LOSSLESS ABRUPT-JUNCTION VARACTOR FREQUENCY MULTIPLIERS

The scattering matrices of lossless abrupt-junction varactor frequency doubler and tripler are derived in Section V. Multipliers of order 2^n, 3^s, and $2^n 3^s$ with least number of idlers‡ are treated in this section.

9.1 *Multipliers of Order 2^n with Least Number of Idlers*

A lossless abrupt-junction varactor frequency multiplier of order 2^n with least number of idlers has been shown to be completely equivalent to a chain of n doublers. We shall assume in this section that the input, output, and all idler circuits are tuned, and that these idler terminations are lossless. The idlers are at frequencies $2^r \omega_0$, $1 \leqq r \leqq (n - 1)$. The equivalence of a multiplier of order 2^n to a chain of doublers can be utilized in getting the scattering relations for multipliers of order 2^n when $n > 1$. The scattering relations when $n = 1$ are given in (59).

We can show that a multiplier of order 2^n with least number of idlers has the following scattering matrix:

$$(\underline{S})_{2^n} = \left[\begin{array}{cc:cc} \left\{\frac{1}{3} - \frac{(-1)^n}{3} 2^{-n}\right\} & (-1)^n 2^{-n} & \multicolumn{2}{c}{0} \\ 1 & 0 & & \\ \hdashline \multicolumn{2}{c:}{} & 0 & (-1)^n \\ \multicolumn{2}{c:}{0} & 2^n & \left\{\frac{1}{3} - \frac{(-1)^n}{3} 2^n\right\} \end{array}\right]. \tag{96}$$

9.2 *Multipliers of Order 3^s with Least Number of Idlers*

Multiplier of order 3^s with least number of idlers has the idler currents flowing at frequencies $2\omega_0$, $3\omega_0$, $6\omega_0$, $9\omega_0$, $\cdots$, $3^{i-1}\omega_0$, $3^i - 3^{i-1}\omega_0$, $3^i\omega_0$, $\cdots$, $3^n - 3^{n-1}\omega_0$. We have shown that such a multiplier is completely equivalent to a chain of s triplers.

‡ It is assumed that all these idler circuits are tuned and that the idler terminations are lossless. It is also assumed that input and output circuits are tuned.

The scattering parameters of a tripler are shown to be

$$(\underline{S})_3 = \left[\begin{array}{cc|cc} 0 & 1 & \multicolumn{2}{c}{\multirow{2}{*}{0}} \\ 1 & 0 & & \\ \hline \multicolumn{2}{c|}{\multirow{2}{*}{0}} & 0 & 3^{-1} \\ & & 3 & 0 \end{array}\right]. \tag{97}$$

If s such triplers are cascaded, we obtain a multiplier of order 3^s with minimum number of idlers. The scattering parameters of a cascade of s triplers can be shown to be given by‡

$$(\underline{S})_{3^s} = \left[\begin{array}{cc|cc} 0 & 1 & \multicolumn{2}{c}{\multirow{2}{*}{0}} \\ 1 & 0 & & \\ \hline \multicolumn{2}{c|}{\multirow{2}{*}{0}} & 0 & 3^{-s} \\ & & 3^s & 0 \end{array}\right]. \tag{98}$$

9.3 *Multipliers of Order 2^n3^s with Least Number of Idlers*

It has been shown that a lossless abrupt-junction harmonic generator of order 2^n3^s, n and s integers, with least number of idlers is completely equivalent to a cascade of n doublers and s triplers connected in proper order depending upon the idler configuration.

The AM and PM transfer scattering matrices[13] of a doubler and a tripler can be shown to be

$$(\underline{T}_{aa})_2 = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{bmatrix}, \tag{99}$$

$$(\underline{T}_{pp})_2 = \begin{bmatrix} -1 & 0 \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}, \tag{100}$$

$$(\underline{T}_{aa})_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \tag{101}$$

and

$$(\underline{T}_{pp})_3 = \begin{bmatrix} 3^{-1} & 0 \\ 0 & 3^{-1} \end{bmatrix}. \tag{102}$$

‡ We have assumed that input, output, and all idler circuits are tuned, and lossless.

Since AM and PM transfer scattering matrices of a tripler are scalar matrices,[14] values of $(\underset{\sim}{T}_{aa})_{2^n3^s}$ and $(\underset{\sim}{T}_{pp})_{2^n3^s}$ are independent of positions of the triplers in the cascaded multiplier.‡ This shows that

$$\begin{aligned}(\underset{\sim}{T}_{aa})_{2^n3^s} &= (\underset{\sim}{T}_{aa})_{2^n} \\ &= \begin{bmatrix} (-1)^n2^{-n} & \frac{1}{3} - \frac{(-1)^n}{3}\,2^{-n} \\ 0 & 1 \end{bmatrix}\end{aligned} \tag{103}$$

and

$$\begin{aligned}(\underset{\sim}{T}_{pp})_{2^n3^s} &= 3^{-s}(\underset{\sim}{T}_{pp})_{2^n} \\ &= \begin{bmatrix} (-1)^n & 0 \\ \frac{(-1)^n}{3} - \frac{1}{3}\,2^{-n} & 2^{-n} \end{bmatrix} 3^{-s}.\end{aligned} \tag{104}$$

Since we also know that there is no AM to PM and PM to AM conversion in both a doubler and in a tripler it follows that

$$(\underset{\sim}{S}_{ap})_{2^n3^s} = (\underset{\sim}{S}_{pa})_{2^n3^s} = 0. \tag{105}$$

Using (103) through (105), we conclude that the scattering parameters of a multiplier of order 2^n3^s with minimum number of idlers are given by

$$(\underset{\sim}{S})_{2^n3^s} = \left[\begin{array}{cc|cc} \frac{1}{3} - \frac{(-1)^n2^{-n}}{3} & (-1)^n2^{-n} & \multicolumn{2}{c}{0} \\ 1 & 0 & & \\ \hline & & 0 & (-1)^n3^{-s} \\ \multicolumn{2}{c|}{0} & 2^n3^s & \frac{1}{3} - \frac{(-1)^n2^n}{3} \end{array}\right]; \tag{106}$$

or

$$(\underset{\sim}{S}_{aa})_{2^n3^s} = \begin{bmatrix} \frac{1}{3} - \frac{(-1)^n}{3}\,2^{-n} & (-1)^n2^{-n} \\ 1 & 0 \end{bmatrix} \tag{107}$$

and

$$(\underset{\sim}{S}_{pp})_{2^n3^s} = \begin{bmatrix} 0 & (-1)^n3^{-s} \\ 2^n3^s & \frac{1}{3} - \frac{(-1)^n}{3}\,2^n \end{bmatrix}. \tag{108}$$

‡ Matrix product $\underset{\sim}{A}\underset{\sim}{B}$ is not, in general, commutative.[14]

Equations (106) and (107) show that the scattering parameters of a multiplier of order 2^n3^s are independent of the idler configuration of the multiplier. For example, 1-2-3-6-12 and 1-2-4-6-12 multipliers have the same scattering matrix. This result arises because of the scalar character of AM and PM transfer scattering matrices of a tripler and is not true in general.‡

X. RESULTS AND CONCLUSIONS

A general method to obtain the scattering parameters of a pumped nonlinear system when the system is subjected to small band limited fluctuations has been presented.

For a lossless abrupt-junction varactor frequency multiplier of order 2^n which has minimum number of idlers and whose input, output, and idler circuits are tuned, it is shown that the scattering matrix $\underset{\sim}{S}$ is given by

$$(\underset{\sim}{S})_{2^n} = \left[\begin{array}{cc:cc} \frac{1}{3} - \frac{(-1)^n}{3} 2^{-n} & (-1)^n 2^{-n} & \multicolumn{2}{c}{0} \\ 1 & 0 & & \\ \hdashline & & 0 & (-1)^n \\ \multicolumn{2}{c:}{0} & 2^n & \frac{1}{3} - \frac{(-1)^n}{3} 2^n \end{array}\right]. \tag{96}$$

Such a multiplier has also been shown to be completely equivalent to a cascade of n doublers.

For a lossless abrupt-junction varactor harmonic generator of order 3^s with minimum number of idlers and whose input, output, and idler circuits are all tuned it is shown that the scattering matrix $\underset{\sim}{S}$ can be represented as

$$(\underset{\sim}{S})_{3^s} = \left[\begin{array}{cc:cc} 0 & 1 & \multicolumn{2}{c}{0} \\ 1 & 0 & & \\ \hdashline & & 0 & 3^{-s} \\ \multicolumn{2}{c:}{0} & 3^s & 0 \end{array}\right]. \tag{98}$$

A multiplier of order 3^s has been shown to be equivalent to a cascade of s triplers.

However, for a lossless abrupt-junction varactor multiplier of order 2^n3^s with minimum number of idlers it has been shown that this multi-

‡ The transfer scattering matrix $(T)_{N_1N_2}$ is not, in general, equal to $(T)_{N_2N_1}$.

plier is equivalent to a cascade of n doublers and s triplers, and that the scattering matrix $\underline{S}$ can be written as

$$(\underline{S})_{2^n3^s} = \left[\begin{array}{cc:cc} \frac{1}{3} - \frac{(-1)^n}{3}2^{-n} & (-1)^n2^{-n} & \multicolumn{2}{c}{0} \\ 1 & 0 & & \\ \hdashline & & 0 & (-1)^n3^{-s} \\ \multicolumn{2}{c:}{0} & 2^n3^s & \frac{1}{3} - \frac{(-1)^n}{3}2^n \end{array}\right]. \tag{106}$$

For lossless abrupt-junction varactor multipliers of order 2^n, 3^s, and 2^n3^s, n and s integers, with minimum number of idlers, one of the general results is also that if $\omega/\omega_0 \ll 1$, there is no amplitude to phase or phase to amplitude conversion or equivalently

$$\underline{S}_{ap} = \underline{S}_{pa} = 0. \tag{109}$$

The scattering matrices, of lossless abrupt-junction varactor multipliers of order different from those treated in this paper can be obtained by straightforward application of the methods presented in this paper. We, however, feel that most of the lossless abrupt-junction varactor multipliers commonly encountered in practice are covered in this paper. If the junction characteristic of the varactor diode is far from being abrupt or if the junction is overdriven, the same general methods can be applied in order to get the general scattering matrix which relates the fluctuations at different parts of the system. If the bias circuit is poorly designed so that there are currents flowing in the system at frequencies $\pm\omega$, the techniques developed in this paper are still applicable.

At present very little is known about the stability of driven systems like harmonic generators. The results derived in this paper can be made use of in studying the stability of such systems and, in particular, in obtaining the restrictions imposed by the condition of stability on the available circuit configurations. This theory also enables us to derive an expression for the output signal of a pumped nonlinear system having noise sources at several locations in the circuit. A complete analysis of the noise performance of the systems like harmonic generators can be carried out once we know the general scattering parameters of the system. All these and other related results are reserved for a future publication.

APPENDIX *A*

Large-Signal Analysis of Abrupt-Junction Varactor Doubler and Tripler

The large-signal equations of a varactor harmonic generator are given in Ref. 5. The varactor considered in this paper is a lossless varactor whose average elastance S_0 is considered to be a part of the external circuit for the sake of convenience. Let *S(t)* be the elastance of the varactor as pumped. For an abrupt-junction varactor diode we also note[5] that

$$\frac{jk\omega_0 S_k}{I_k} = \frac{j(k-1)\omega_0 S_{k-1}}{I_{k-1}} = \cdots = \frac{j2\omega_0 S_2}{I_2} = \frac{j\omega_0 S_1}{I_1}, \quad k \text{ an integer.} \tag{110}$$

The large-signal equations for a doubler can be written as

$$V_1 = \frac{S_2 I_1^* + S_1^* I_2}{j\omega_0} \tag{111}$$

and

$$V_2 = \frac{S_1 I_1}{j2\omega_0}. \tag{112}$$

It can be shown[5] that the time origin can be chosen so that I_1, and I_2 are both real. From (110) through (112), we can now write

$$\left| \frac{V_1}{I_1} \right| = \frac{| S_2 |}{\omega_0}, \tag{113}$$

$$\left| \frac{V_2}{I_2} \right| = \frac{| S_1 |^2}{4 \, | S_2 | \, \omega_0}, \tag{114}$$

$$\left| \frac{V_1}{I_2} \right| = \frac{| S_1 |}{2\omega_0}, \tag{115}$$

and

$$\left| \frac{V_2}{I_1} \right| = \frac{| S_1 |}{2\omega_0}. \tag{116}$$

The large-signal equations for a tripler can be written as

$$V_1 = \frac{S_3 I_2^* + S_2 I_1^* + S_1^* I_2 + S_2^* I_3}{j\omega_0}, \tag{117}$$

$$V_2 = \frac{S_3 I_1^* + S_1 I_1 + S_1^* I_3}{j2\omega_0}, \tag{118}$$

and

$$V_3 = \frac{S_1 I_2 + S_2 I_1}{j3\omega_0}. \tag{119}$$

It can again be shown[5] that if we choose I_1 to be real and positive I_2 and I_3 are real. Let us assume that the idler termination is tuned and is lossless. From (118) we can write

$$| S_3 | = | S_1 |/2. \tag{120}$$

According to (110), (117), (119), and (120), we also have

$$\frac{| I_1 |}{| I_3 |} = \frac{2}{3} \tag{121}$$

and

$$\left| \frac{V_1}{V_3} \right| = \frac{3}{2}. \tag{122}$$

REFERENCES

1. Dragone, C., Phase and Amplitude Modulation in High-Efficiency Varactor Frequency Multipliers—General Scattering Properties, B.S.T.J., *46,* No. 4, April, 1967, pp. 775–796.
2. Prabhu, V. K., Stability Considerations in Lossless Varactor Frequency Multipliers, to be published.
3. Dragone, C., Phase and Amplitude Modulation in High Efficiency Varactor Frequency Multipliers of Order $N = 2^n$—Stability and Noise, B.S.T.J., *46,* No. 4, April, 1967, pp. 797–834.
4. Dragone, C., AM and PM Scattering Properties of a Lossless Multiplier of Order $N = 2^n$, Proc. IEEE, *54,* No. 12; December, 1966.
5. Penfield, Jr., P. and Rafuse, R. P., *Varactor Applications,* The M.I.T. Press, Cambridge, Mass.; 1962.
6. Prabhu, V., Noise Performance of Abrupt-Junction Varactor Frequency Multipliers, Proc. IEEE, *54,* No. 2, February, 1966, pp. 285–287.
7. Davis, J. A., The Forward-Driven Varactor Frequency Doubler, S.M. Thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass., 1963.
8. Burckhardt, C. B., Analysis of Varactor Frequency Multipliers for Arbitrary Capacitance Variation and Drive Level, B.S.T.J., *44,* No. 4, April, 1965, pp. 675–692.
9. Hardy, G. H. and Rogosinski, W. W., *Fourier Series,* Cambridge University Press, London, U.K., 1950.
10. Penfield, Jr., P., Circuit Theory of Periodically Driven Nonlinear Systems, Proc. IEEE, *54,* No. 2, February, 1966, pp. 266–280.
11. Prabhu, V. K., Representation of Noise Sources in Pumped Nonlinear Systems, to be published.
12. Haus, H. A. and Adler, R. B., *Circuit Theory of Linear Noisy Networks,* The Technology Press, Cambridge, Mass., and John Wiley and Sons, Inc., New York, N. Y., 1959.
13. Carlin, H. J. and Giordano, A. B., *Network Theory,* Prentice-Hall, Inc., Englewood Cliffs, N. J., 1964.
14. Hohn, F. E., *Elementary Matrix Algebra,* The MacMillan Co., New York, N. Y., 1963.

Effect of Redirectors, Refocusers, and Mode Filters on Light Transmission Through Aberrated and Misaligned Lenses

By E. A. J. MARCATILI

(Manuscript received June 16, 1967)

The field distortion of a beam propagating through a sequence of identical, misaligned and slightly aberrated lenses is calculated as a perturbation of the Gaussian beam that would propagate in the absence of aberration.

It is found that most of the converted power goes to the first and second modes. They produce deflection and spot-size change of the ideal beam, respectively. The power coupled to modes higher than the second deform the Gaussian profile.

In general, the mode conversion per unit length of guide can be reduced by making the spot size small and by avoiding in-phase coupling at every lens. This last condition is achieved by choosing the period of oscillation of the beam different from an integer number of lens spacings.

Before the beam becomes too distorted, the converted modes must be eliminated. Power in the first and second modes can be reconverted losslessly to the fundamental Gaussian beam by means of servoloops that redirect and refocus the beam. If refocusers are not used, the power in the second mode, as well as the power in the higher-order modes must be absorbed in mode filters such as irises.

For lenses with fourth-order aberration such that at a beam half-width distance from the center the focal length departs δ percent from ideal, the following typical results are obtained:

In a guide in which the distance between the beam and guide axes is a constant plus a sinusoid, the converted power is proportional to δ^2, to the fourth power of the amplitude of the sinusoid and to the square of the number of lenses, but is roughly independent of the curvature of the guide axis.

On the other hand, in a guide in which the distance between the beam and guide axes is a constant plus a random quantity the converted power is proportional to δ^2, to the square of the guide curvature, to the mean square of the random deviation, and to the number of lenses.

For δ = 1 percent, a 1 percent power conversion to the second mode occurs in typical examples, after a few tens of lenses, and the order of magnitude of mode conversion is 0.001 dB/lens. Most of that power is in the second mode and can be recovered with refocusers.

I. INTRODUCTION

Sequences of widely separated glass lenses[1] or periscopic mirrors,[2] as well as sequences of low loss closely spaced gas lenses,[3, 4] have been proposed as beam waveguides for long distance optical transmission.

The theory describing the wave and ray propagations through a sequence of misaligned, thin, perfect lenses is known, and those results are applicable even if aberrations are present, provided that the number of lenses is small. Nevertheless, when that number is large, the cumulative effect of lens imperfections must be included.

Before introducing aberrations, though, let us briefly review what is already known about wave and geometric optics in a beam waveguide, assuming throughout the two-dimensional problem instead of the more realistic and complex, but not more enlightening, three-dimensional one.

A paraxial Gaussian beam launched in a periodic sequence of identical thin, aberration-free, but misaligned lenses[5,6,7,8] conserves throughout the Gaussian transverse field distribution. The spot size depends on the initial conditions, the focal length f of the lenses and their spacing L, but is independent of the lens alignment and does not grow with the number of lenses. The geometry of the beam axis, on the other hand, depends also on f, L, and the initial conditions, but more important, it depends on the alignment of the lenses. In general, the beam axis oscillates about the guide axis and the amplitude of the oscillations increases with the number of lenses. For a given set of lenses through which a beam is to be guided, there are then alignment tolerances which must be satisfied in order to prevent the beam from hitting the edges of the lenses. Those tolerances can be drastically alleviated by using redirectors,[9] that is, servoloops that realign the beam axis with the guide axis.

Nevertheless, if the lenses have aberrations, the beam does not conserve the Gaussian profile, but deforms itself[10,11,12,13] as it travels along the guide, the definition of the beam axis then becomes fuzzy, the redirectors become less and less effective, and eventually the grossly distorted beam hits the edges of the lenses.

Gloge[14] has found the effects of random aberrations such as those which occur both in glass lenses and in the controlled atmosphere between them. Because of the randomness of the aberrations, the beam distortion is independent of the beam trajectory. On the other hand, if all the lenses have the same aberration such as in gas lenses, the beam deformation is strongly dependent on the relation between the beam and the guide axes.

This paper gives an estimate of the beam deformation as a function of systematic aberrations, lens misalignments, and presence or absence of redirectors. It also suggests ways of preventing the beam deterioration together with their price tags.

II. WAVE TRANSMISSION THROUGH SLIGHTLY ABERRATED AND MISALIGNED LENSES

We begin reviewing the wave transmission through ideal lenses and afterward the lenses are slightly perturbed and the mode conversion is calculated.

The wave transmission through a sequence of ideal, thin, equidistant and misaligned lenses as those shown in Fig. 1 is known.[5,6,7,8]

The guide is completely defined by the focal length f of the lenses, their separation L and the radius of curvature R_n of the guide axis at every lens.

The beam axis is characterized by its distance s_n to the guide axis at the nth lens. If the beam is launched through the center of the first lens, it is known[9] that s_n is related to L, f, and R_n by

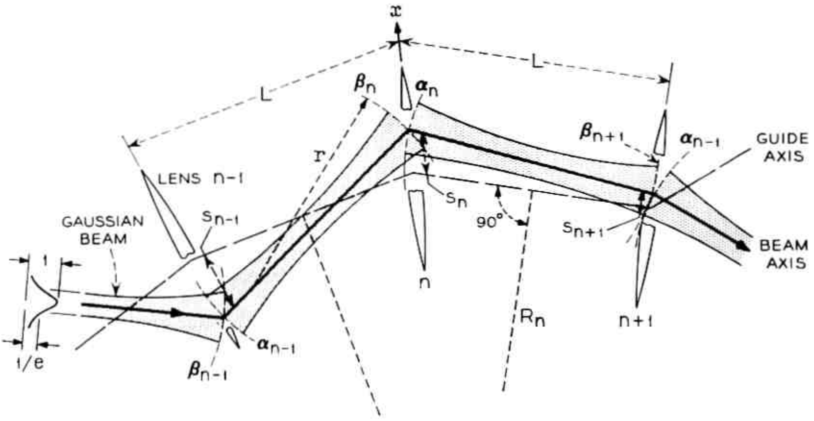


Fig. 1 — Beam transmitted through misaligned ideal lenses.

$$s_n = \frac{L^2}{\sin\theta} \sum_{m=1}^{n-1} \frac{\sin(n - m)\theta}{R_m}, \tag{1}$$

where

$$\cos\theta = 1 - \frac{L}{2f}. \tag{2}$$

The field distribution at every equiphase surface is Gaussian, and its width varies along the beam. Nevertheless, if properly launched, the beam maintains the same half-width w and the same radius of curvature r of the wavefront at every lens. As depicted in Fig. 1, the beam between surfaces α_{n-1} and β_n is the same for all n.

Proper beam launching is achieved if, at the first lens, the radius of curvature of the wavefront is

$$r = 2f, \tag{3}$$

and if the beam half-width w is related to the wavelength λ and the guide parameters in the following manner:

$$w = \sqrt{\frac{\lambda L}{\pi \sin\theta}} = \sqrt{\frac{2\lambda}{\pi} f \tan\frac{\theta}{2}}. \tag{4}$$

Assuming the lenses to be two-dimensional, then the Gaussian beam is also two-dimensional and the electric field measured along the circles α_n or β_n is

$$E = D_0(2\xi) = e^{-\xi^2} \tag{5}$$

and

$$\xi = \frac{x}{w}. \tag{6}$$

Strictly speaking, the normalized length ξ measured along the surfaces α_n or β_n does not coincide with x/w, but if the beam is paraxial, the discrepancy is negligible.

If a higher mode is launched with the same wavefront curvature $1/r$ and the same width w, the field distribution at every surface α_n, β_n is described by the parabolic cylinder function[15]

$$D_p(2\xi) = e^{-\xi^2} He_p(2\xi) \tag{7}$$

in which $He_p(2\xi)$ is the Hermite polynomial of order p. Between the surfaces α_n and β_{n+1}, the phaseshift of the pth mode is $p\theta$ radians smaller than the phaseshift of the fundamental Gaussian mode.

Now let us calculate the phaseshift in passing from the equiphase surface β_n to the equiphase surface α_n, Fig. 2. The length $\overline{AB}$ of a typical ray path between those two wavefronts is

$$\overline{AB} = x\frac{s_n}{f} + \frac{x^2}{r}.$$

Calling

$$\sigma_n = \frac{s_n}{w} \tag{8}$$

and using (3), (4), and (8), we obtain

$$\overline{AB} = \frac{\lambda}{\pi}(2\sigma_n\xi + \xi^2)\tan\frac{\theta}{2}.$$

The total phaseshift in passing from surface β_n to α_n is

$$\varphi_n(\sigma_n + \xi) = 2(2\sigma_n\xi + \xi^2)\tan\frac{\theta}{2} + \Phi_n(\sigma_n + \xi). \tag{9}$$

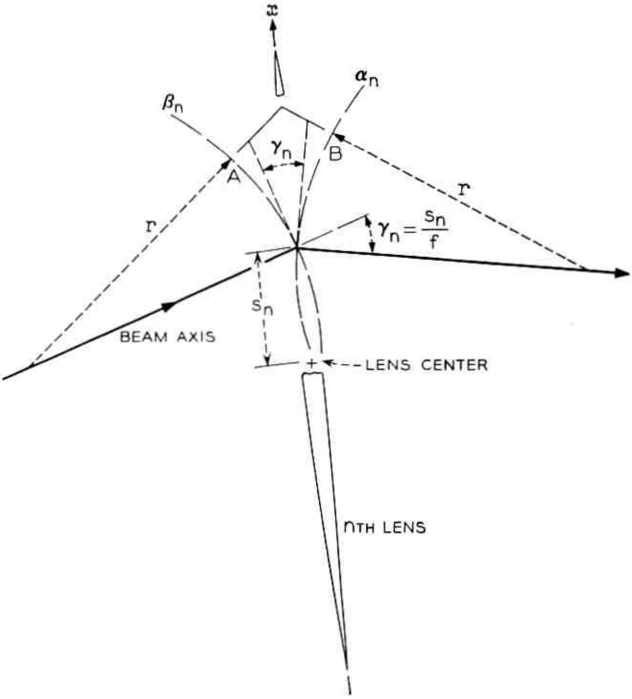


Fig. 2 — Phase front change through a lens.

It is made essentially of two terms. The first is the phase due to the path $\overline{AB}$, the second, $\Phi_n(\sigma_n + \xi)$, is the phase contributed by the lens. Only if the lens is ideal the surface α_n will be an equiphase, and $\varphi_n(\sigma_n + \xi)$ is a constant. The phaseshift through a perfect lens is then

$$\Phi_{n\ \text{ideal}}(\sigma_n + \xi) = \text{const.} - 2(2\sigma_n\xi + \xi^2) \tan \frac{\theta}{2} \tag{10}$$

and since the constant introduces only an uninteresting uniform phaseshift, we will call it zero throughout.

On the other hand, if the lens is not perfect, $\varphi_n(\sigma_n + \xi)$ is not a constant and the field on the surfaces α_n and β_n can no longer be described by a single mode, but rather by a superposition of modes as those given in (7). In general then, the field on the surface α_n is

$$E(\alpha_n) = \sum_{q=0}^{\infty} a_{qn} D_q(2\xi). \tag{11}$$

The amplitude a_{qn} of each mode has been calculated in (57) under the assumptions of small lens distortions and small beam departure from ideal, that is,

$$| \varphi_n(\sigma_n + \xi) | \ll \pi \tag{12}$$

and

$$| a_{qn} | \ll 1 \quad \text{for} \quad q > 0. \tag{13}$$

That amplitude is

$$a_{qn} = \sum_{\nu=1}^{n} c_{oq\nu} e^{iq(\nu-n)\theta} \tag{14}$$

in which $c_{oq\nu}$ is the coupling coefficient between the fundamental and the qth mode at the νth lens, and its value, derived in (58), is

$$c_{oq\nu} = \begin{cases} 1 & \text{if} \quad q = 0, \\ \dfrac{i}{2^q q!} \displaystyle\sum_{\mu=0}^{\infty} \frac{1}{2^{3\mu}\mu!} \frac{\partial^{q+2\mu}\varphi_\nu(\sigma_\nu)}{\partial\sigma_\nu{}^{q+2\mu}} & \text{if} \quad q > 0. \end{cases} \tag{15}$$

Within the approximations involved, the fundamental mode has amplitude one throughout. The amplitude a_{qn} of the qth mode immediately after the nth lens, (14), is made up of n terms. All of them have simple physical meaning. Consider the νth term. The fundamental mode couples $c_{oq\nu}$ into the qth mode of the νth lens and this travels up to the nth lens without further conversion; its phaseshift with respect to the fundamental mode is $q(\nu - n)\theta$.

Since $c_{oo\nu} = 1$, the reconversion into the fundamental mode is not calculated explicitly. Nevertheless, it is automatically taken into account when the power in the higher order modes is ascertained.

The amplitude of the coupled mode a_{qn} can be maintained small by techniques well known from coupled waves theory and which will be used later on:

(*i*) Selecting the phase at the coupling points to provide destructive interference.

(*ii*) Dissipating the power in the unwanted mode.

(*iii*) Providing mode transformers capable of changing unwanted modes into the fundamental one.

III. RECOVERY OF THE FUNDAMENTAL MODE

Physical interpretations of the deformation of a beam traveling through aberrated lenses and ways of preventing that deterioration are considered next.

The field (11) after the *n*th lens is made essentially by the fundamental mode slightly modified by higher-order modes. Neglecting powers of a_{qn} bigger than one, because of (13), and grouping the first three terms,

$$E(\alpha_n) = (1 - a_{2n}) D_0[2\xi(1 - 2a_{2n}) - 2a_{1n}] + \sum_{q=3}^{\infty} a_{qn} D_q(2\xi). \qquad (16)$$

The first term is a Gaussian beam different from the ideal one. Its axis is at a distance

$$c_n = \operatorname{Re} a_{1n} \qquad (17)$$

from the beam axis of Fig. 1, and its half-width is

$$w_n = 1 + \operatorname{Re} 2a_{2n} . \qquad (18)$$

Both dimensions are normalized to the ideal beam half-width w.

Furthermore, the angle between the two axes is

$$\theta_n = \frac{\lambda}{\pi w} \operatorname{Im} a_{1n} \qquad (19)$$

and the radius of curvature of the wavefront results

$$r_n = 2f \left[1 - \frac{4}{\tan \dfrac{\theta}{2}} \operatorname{Im} a_{2n} \right] \cdot \qquad (20)$$

Expressions (17) through (20) are valid as long as

$$\begin{aligned} | a_{1n} | &\ll 1 \\ | a_{2n} | &\ll 1. \end{aligned} \tag{21}$$

These inequalities can be satisfied for an arbitrarily large number of lenses by periodically realigning the beam and changing its width to the proper size.

The realignment of the beam can be made with redirectors.[9] A feedback servoloop as shown in Fig. 3 senses the position of the beam with photosensors p_1 and p_2 which are centered, for example, on the axis of the pipe in which the lenses are housed.

The difference signal from the sensors is amplified in A and used to displace lens $n - 1$ laterally until the beam axis passes through the center of the photosensors. In general, at every servoloop the beam's axis will pass through the center of the sensor.

The beam size and the curvature of the wavefront can also be adjusted with servoloops which we will call refocusers. The principle of operation is shown in Fig. 4. The beam is aligned with three lenses and three photosensors p_1, p_2, and p_3, are placed at distances from the axis such that an ideal beam would produce equal signals. The difference signal between p_1 and p_2 is amplified in A and controls the focal length of lens $n - 1$, while the difference signal between p_2 and p_3 is amplified in B and used to change the focal length of the nth lens. Once these differences are small, the three signals from the photosensors are practically identical and the beam coincides with the ideal one.

Beam size correction is also possible changing the distance between lenses instead of their focal length.

Obviously, if the lenses are three-dimensional instead of the two-

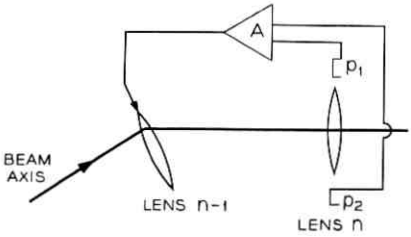


Fig. 3 — Redirector:servoloop for beam realignment.

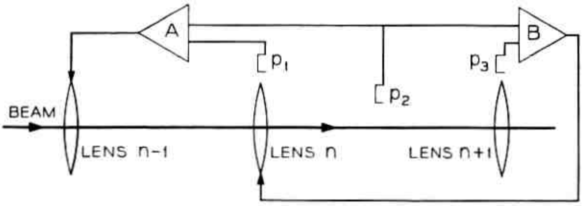


Fig. 4 — Refocuser:servoloop for beam-width adjustment.

dimensional considered above, similar servoloops must be used in two perpendicular directions.

For gas lenses of the tubular type,[3] the beam deflection and focusing may be achieved by dividing the tube in four sectors (Fig. 5) and controlling the temperature T of each of them.[16] If $T_1 = T_2 = T_3 = T_4$, the lens focuses only, but if $T_2 = T_3 = T_4 < T_1$, the lens focuses and deflects the beam downward. If $T_1 = T_3 > T_2 = T_4$, the focusing in the vertical direction is stronger than in the horizontal direction.

For focusing devices such as periscopic mirrors[2] (Fig. 6), the deflection of the beam may be achieved by rotating one or both mirrors around perpendicular axes x and y. As suggested by R. Kompfner, beam refocusing may be obtained by mechanically deforming the mirrors in the two perpendicular directions.

The beam losses in the process of beam refocusing and redirection are due to the interception of the beam by the photosensors, and, in principle at least, they can be made very small indeed. These devices then operate on the idea of reconverting higher unwanted modes into the fundamental.

Unfortunately, it is not simple to make a mode converter capable of taking care of the modes higher than the second contained in the summation in (16). For them and also for the second mode, if refocusers are not used, S. E. Miller suggested another technique which

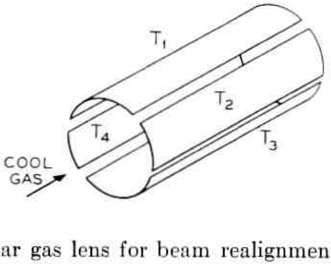


Fig. 5 — Tubular gas lens for beam realignment and refocusing.

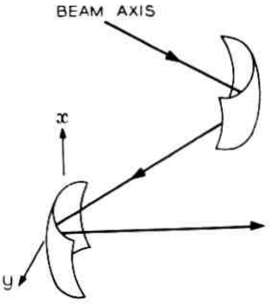


Fig. 6 — Periscopic mirrors.

is essentially lossy, but may be simpler to implement. It consists of using mode filters, perhaps irises aligned with the centers of the redirectors' photosensors.

What are the powers involved in these filtering schemes? At the nth lens, the power in the second mode normalized to that in the fundamental one is

$$P_n^{(1)} = 2 \mid a_{2n} \mid^2. \tag{22}$$

If a_{2n} is real, the radius of curvature of the wavefront (20) coincides with the ideal one $2f$, while the half beam-width departure from ideal results from (18),

$$\mid w_n - 1 \mid = \sqrt{2P_n^{(1)}}. \tag{23}$$

At the same lens, the power carried by the higher-order modes is

$$P_n^{(2)} = \sum_{q=3}^{\infty} \mid a_{qn} \mid^2 q!. \tag{24}$$

There is another filtering scheme which consists of using at every lens redirectors and filters capable of absorbing the second- and higher-order modes. The power absorbed by the filters in the n first lenses normalized to the power in the fundamental mode is

$$P_n^{(3)} = \sum_{\nu=1}^{n} \sum_{q=2}^{\infty} q! \mid c_{oq\nu}(\sigma_\nu) \mid^2. \tag{25}$$

In the following section, several examples are considered and a comparison is made between the powers $P_n^{(1)}$, $P_n^{(2)}$, and $P_n^{(3)}$ to find the most efficient way of avoiding beam deterioration.

IV. EXAMPLES

Let us assume that all the lenses are imperfect but identical, and that the aberration is of fourth order. The phaseshift introduced by the aberration

$$\varphi_\nu(\sigma_\nu) = \delta\sigma_\nu^4 \tag{26}$$

means physically that at a beam half-width w from the center of the lenses, the phaseshift due to aberration is δ radians, while the ideal phaseshift (10) is $-2 \tan \theta/2$. Another physical interpretation is provided by the focal length

$$f(\sigma_\nu) = f\left(1 + \frac{\delta}{\tan \theta/2}\,\sigma_\nu^2\right) \tag{27}$$

calculated from (10). At a distance w from the center, the focal length is $\delta f/\tan \theta/2$ longer than ideal. For gas lenses,[3] a typical value for δ is 0.01.[17]

Then, the coupling coefficients between the fundamental and the higher-order modes at the νth lens (15) are

$$\begin{aligned}
c_{o2\nu} &= i\tfrac{3}{8}\delta(1 + 4\sigma_\nu^2)\\
c_{o3\nu} &= i\,\frac{\delta}{2}\,\sigma_\nu\\
c_{o4\nu} &= i\,\frac{\delta}{16}\\
c_{oq\nu} &= 0 \quad \text{for} \quad q > 4
\end{aligned} \tag{28}$$

and the amplitudes of the different modes after the nth lens (14) turn out to be

$$\begin{aligned}
a_{2n} &= i\tfrac{3}{8}\delta \sum_{\nu=1}^{n} (1 + 4\sigma_\nu^2)e^{i2(\nu-n)\theta}\\
a_{3n} &= i\,\frac{\delta}{2} \sum_{\nu=1}^{n} \sigma_\nu e^{i3(\nu-n)\theta}\\
a_{4n} &= \frac{i\delta}{16} \sum_{\nu=1}^{n} e^{i4(\nu-n)\theta}\\
a_{qn} &= 0 \quad \text{for} \quad q > 4.
\end{aligned} \tag{29}$$

To assign values to σ_ν we consider two examples.

4.1 *Beam Guide with Curved Axis*

The lens centers are on a circle of radius R and the beam axis intersects the νth lens at a distance

$$\sigma_\nu = h_0 + h_1 \cos \nu\theta \tag{30}$$

from its center. This means that the beam axis oscillates sinusoidally with amplitude h_1 about a circle of radius $R + h_0 w$.

The constant h_1 depends only in the beam launching conditions, while h_0 is related to the other parameters of the guide[18] by

$$h_0 = \frac{L^2}{4wR \sin^2 \theta/2} = \frac{Lf}{wR}. \tag{31}$$

Substituting (29) in (22) and (24), as well as (26) in (25), assuming θ to be of the order of $\pi/2$ and neglecting terms that do not grow with n, the following powers are derived:

$$\begin{aligned} P_n^{(1)} &= \tfrac{9}{2}\delta^2 h_1^2 \left[h_0 \left| \frac{\sin 3/2n\theta}{\sin 3/2\theta} \right| + \frac{h_1}{4}\left(n + \left| \frac{\sin 2n\theta}{\sin 2\theta} \right| \right) \right]^2 \\ P_n^{(2)} &= \tfrac{3}{2}\delta^2 \left[\left(h_0 \frac{\sin 3/2n\theta}{\sin 3/2\theta} \right)^2 + \left(\frac{h_1^2}{4} + \frac{1}{16} \right) \frac{\sin^2 2n\theta}{\sin^2 2\theta} \right] \\ P_n^{(3)} &= \tfrac{3}{8}\delta^2 \left[1 + 10\left(h_0^2 + \frac{h_1^2}{2} \right) + 12h_0^4 + 36h_0^2h_1^2 \right. \\ &\qquad \left. + \tfrac{3}{2}h_1^4\left(3 + \left| \frac{\sin 2n\theta}{\sin 2\theta} \right| \right) + 12h_0h_1^3 \left| \frac{\sin 3/2n\theta}{\sin 3/2\theta} \right| \right] n. \end{aligned} \tag{32}$$

To minimize these quantities, one must choose the distance L between lenses in such a way that θ is different enough from $\pi/2$ and $2\pi/3$ as to satisfy the inequalities

$$\left| \theta - \frac{\pi}{2} \right| \gg \frac{\pi}{2n}$$

and

$$\left| \theta - \frac{2\pi}{3} \right| \gg \frac{2\pi}{3n}. \tag{33}$$

This choice prevents the systematic in-phase coupling of higher-order modes at every lens. This result can be extended to guides of identical lenses with any aberration. The separation L must be chosen in such a way that the period of oscillation of the beam about the axis does not coincide with an integer number of lens spacings.

If (33) are satisfied, the powers of (32) become

$$P_n^{(1)} = \tfrac{9}{32}\delta^2 h_1^4 n^2 \tag{34}$$

$$P_n^{(2)} = 0 \tag{35}$$

$$P_n^{(3)} = \tfrac{3}{8}\delta^2\left[1 + 10\left(h_0^2 + \frac{h_1^2}{2}\right) + 12h_0^4 + 36h_0^2h_1^2 + \tfrac{9}{2}h_1^4\right]n. \tag{36}$$

Since $P_n^{(2)} = 0$, there is no build-up of power in the third and fourth modes. The beam maintains a Gaussian profile and can be refocused without any power loss.

The power in the second mode grows proportionally to the square of the number of lenses and to the fourth power of the amplitude of the beam axis oscillations, but is independent of h_0 and consequently independent of the radius of curvature R of the guide.

If absorbing mode filters are used, one observes that, while $P_n^{(1)}$ grows proportionally to n^2, (34), $P_n^{(3)}$ grows only proportionally to n. There is cross-over at a number of lenses n_0 for which $P_{n_0}^{(1)} = P_{n_0}^{(3)}$. It is

$$n_0 = \frac{4}{3h_1^4}\left[1 + 10\left(h_0^2 + \frac{h_1^2}{2}\right) + 12h_0^4 + 36h_0^2h_1^2 + \tfrac{9}{2}h_1^4\right]\cdot \tag{37}$$

If $n < n_0$ it is less power consuming to have one filter every n lenses. If $n > n_0$ it is better to use filters at every lens.

For

$$h_0^2 = 1,^*$$

$$h_1^2 = 1, \quad \text{and}$$

$$\delta = 0.01,$$

we calculate from $P_n^{(1)}$ in (34) that the power converted to the second mode is 1 percent after 19 lenses.

Furthermore, one mode filter at the 19th lens, or filters at every lens, would dissipate, respectively,

$$P_{19}^{(1)} = 0.01 \qquad \text{equivalent to 0.0023 dB/lens}$$

and

$$P_{19}^{(3)} = 0.049 \qquad \text{equivalent to 0.011 dB/lens.}$$

* This value $h_0 = 1$ is derived from (31) using the following typical quantities:

$$L = 0.7 \text{ m}$$
$$R = 1 \text{ km}$$
$$w = 0.5 \text{ mm}$$
$$\theta = \pi/3.$$

It is roughly five times less power consuming to use one filter every 19 lenses. This occurs because the conversions at successive lenses have enough phaseshift to interfere destructively and reduce the converted power level from 0.011 dB/lens to 0.0023 dB/lens.

Given the lenses with aberrations and a length of guide D, is there less mode conversion crowding the lenses or keeping them far apart? To answer this question, (34) is rewritten substituting for n the ratio D/L,

$$P_{D/L}^{(1)} = \left(\frac{9}{32}\frac{\delta^2 h_1^4}{\tan^2 \theta/2} D^2\right)\frac{\tan^2 \theta/2}{L^2}. \tag{38}$$

Because of the normalizations (8), (27), (30) to the ideal beam size w, the parenthesis is a constant and $P_{D/L}^{(1)}$ is minimized by making $\tan^2 \theta/2/L^2$ as small as possible. From (2)

$$\frac{\tan^2 \theta/2}{L^2} = \frac{1}{L^2(4f/L - 1)}.$$

This expression and consequently the power $P_{D/L}^{(1)}$ is independent of the wavelength λ and can be minimized by choosing the separation of the lenses as close to confocal as possible without violating the inequality (33).

Following a similar line of thought one deduces from (32) that the power $P_n^{(2)}$, in modes higher than the second, is reduced by choosing λ as small as possible. This result is illustrated next via a computer experiment[10] that goes beyond the limits of applicability of the perturbation analysis developed in this paper.

Consider a sequence of aberrated, aligned, two-dimensional lenses of width $2a$, spacing L, and focal length

$$f = f_0\left[1 + 0.02\left(\frac{s}{a}\right)^2\right]. \tag{39}$$

A Gaussian beam of half-width w, which is the correct spot size for a sequence of ideal lenses of focal length $f = f_0$, enters parallel to the guide axis at a distance $a/3$. Assuming

$$L = 2f_0 \quad \text{(confocality)}$$

and

$$\frac{w}{a} = \frac{1}{3},$$

the distorted power beam profiles at the 167th and 168th lenses are illustrated in Fig. 7(a).

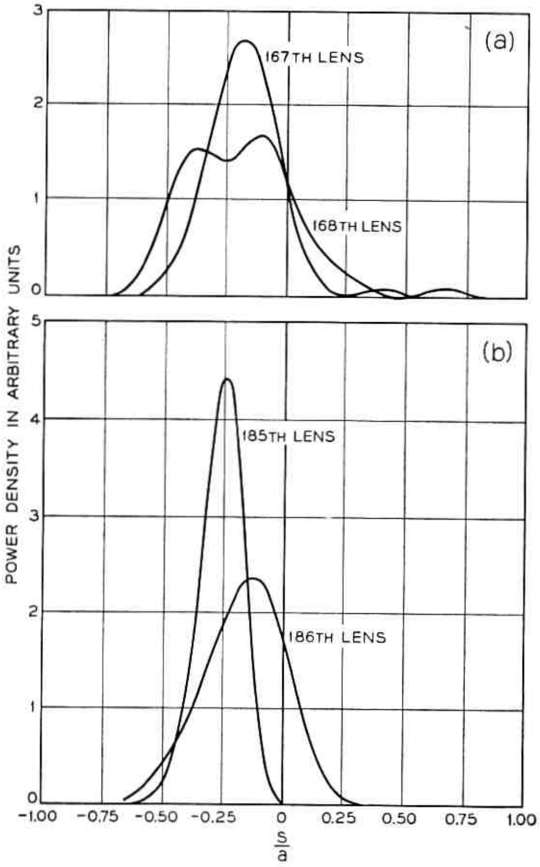


Fig. 7 — Power beam profile after many lenses of focal length $f = f_0[1 + 0.02(s/a)^2]$. (a) $L = 2f_0$ (confocal lenses); $w/a = \frac{1}{3}$. (b) $L = 1.8f_0$ (10% off confocal); $w/a = 1/3\sqrt{2}$ (shorter wavelength).

Conversion to distorting high-order modes is substantially reduced by avoiding confocality and by reducing the beam width. For example, assuming

$$L = 1.8f_0$$

and

$$\frac{w}{a} = \frac{1}{\sqrt{2}\,3}$$

the power beam profiles after the same length of guide, Fig. 7(b), are still close to Gaussian.

4.2 *Beam Axis Randomly Dispaced from a Circle*

This beam axis occurs, for example, in the following beam guide. Assume a metallic tube in which the lenses are housed and whose axis is a circle of radius R. With each lens there is a redirector rigidly connected with the tube. To keep the beam away from the wall, the photosensors' centers should coincide with the tube axis, but they don't because of alignment tolerances. At the nth photosensor, their separation is a Gaussian random length d_n, Fig. 8, which we have normalized to the beam half-width w.

The beam axis forced to pass through the center of every photosensor will also depart from the tube axis d_n.

From Fig. 8, one finds that d_n, R, L, and f are related to the normalized distance σ_n between beam axis and lens center by the expression

$$\sigma_n = 2\,d_n - d_{n-1} - d_{n+1} + h_0 \tag{40}$$

in which h_0 is once more the constant defined in (31).

Since d_n is a Gaussian random variable, it follows from (35) with obvious nomenclature

$$\langle\sigma\rangle = h_0 \tag{41}$$

$$\langle\sigma^2\rangle = 6\langle d^2\rangle + h_0^2 \tag{42}$$

$$\langle\sigma^4\rangle = 108\langle d^2\rangle^2 + 36h_0^2\langle d^2\rangle + h_0^4\,. \tag{43}$$

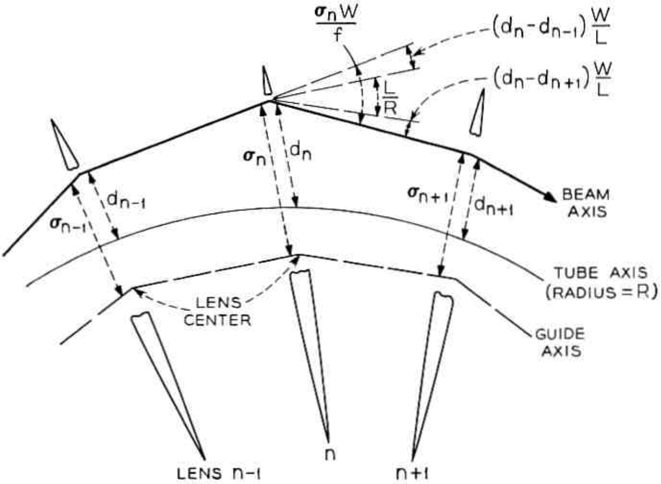


Fig. 8 — Beam axis at random distance d_n from a circle of radius R. $\sigma_n = 2d_n - d_{n-1} - d_{n+1} + h_0$; $h_0 = Lf/wR$.

Now we calculate from (22), (24), and (29) the expected power in the second mode and in the higher-order modes at the nth lens

$$\langle P_n^{(1)} \rangle = \tfrac{9}{2}\delta^2(\langle \sigma^4 \rangle - \langle \sigma^2 \rangle^2)n \tag{44}$$

$$\langle P_n^{(2)} \rangle = \tfrac{3}{2}\delta^2\left[(\langle \sigma^2 \rangle - \langle \sigma \rangle^2)n + \frac{1}{16}\left(\frac{\sin 2n\theta}{\sin 2\theta}\right)^2 + \left(h_0 \frac{\sin 3/2n\theta}{\sin 3/2\theta}\right)^2\right]. \tag{45}$$

The expected power to be absorbed by mode filters at every lens is deduced from (25) and (28). It is

$$\langle P_n^{(3)} \rangle = \tfrac{3}{8}\delta^2(1 + 10\langle \sigma^2 \rangle + 12\langle \sigma^4 \rangle)n. \tag{46}$$

More explicit results are obtained substituting in the last three expressions the averages $\langle \sigma \rangle$, $\langle \sigma^2 \rangle$, and $\langle \sigma^4 \rangle$ with their equivalents in (41) through (43):

$$\langle P_n^{(1)} \rangle = 108\,\delta^2\langle d^2 \rangle(h_0^2 + 3\langle d^2 \rangle)n \tag{47}$$

$$\langle P_n^{(2)} \rangle = 9\,\delta^2\langle d^2 \rangle n + \frac{3}{32}\left(\delta \frac{\sin 2n\theta}{\sin 2\theta}\right)^2 + \frac{3}{2}\left(\delta h_0 \frac{\sin 3/2n\theta}{\sin 3/2\theta}\right)^2 \tag{48}$$

$$\langle P_n^{(3)} \rangle = \tfrac{3}{8}\delta^2[1 + 10h_0^2 + 12h_0^4 + 12\langle d^2 \rangle(36h_0^2 + 5) + 1296\langle d^2 \rangle^2]n. \tag{49}$$

The power in the second mode grows proportionally to the number of lenses, and if $h_0^2 \gg 3\langle d^2 \rangle$, is proportional to the mean square displacement and also proportional to h_0^2.

To prevent build-up of $\langle P_n^{(2)} \rangle$ proportionally to n^2, it is necessary to avoid choosing $\theta = 2\pi/3$ or $\theta = \pi/2$.

In general, $\langle P_n^{(2)} \rangle \ll \langle P_n^{(3)} \rangle$; therefore, it is less power consuming to use beam refocuser and mode filters after several lenses and not at every lens.

For

$$h_0^2 = 1$$

$$\delta = 0.01$$

$$\sqrt{\langle d^2 \rangle} = 0.1$$

an expected power conversion to the second mode of 1 percent occurs after 90 lenses. At that lens (47), (48), and (49) become

$$\langle P_{90}^{(1)} \rangle = 0.01 \qquad \text{equivalent to } 0.00048 \text{ dB/lens}$$

$$\langle P_{90}^{(2)} \rangle = 0.0008 \qquad \text{equivalent to } 0.00004 \text{ dB/lens}$$

$$\langle P_{90}^{(3)} \rangle = 0.095 \qquad \text{equivalent to } 0.0047 \text{ dB/lens.}$$

If only mode filters every 90 lenses are used, the loss is $\approx$0.00052 dB/lens. If beam refocuser and mode filters are used every 90 lenses, the loss is 12 times smaller, 0.00004 dB/lens.

Differently from the previous example, the conversion per unit length is reduced by choosing both the separation L between lenses and the wavelength λ as short as possible.

V. CONCLUSIONS

A beam transmitted through few tens of identical misaligned and aberrated lenses becomes distorted due to coupling to unwanted higher-order modes. Unless the beam is restored to ideal shape, the distortion continues until power is lost through the edges of the lenses.

In general, mode conversion per unit length of guide is minimized but not eliminated: (*i*) by choosing the distance between lenses such that the period of oscillation of the beam does not coincide with an integer number of lens spacings; (*ii*) by reducing the spot size, that is, by using short lens spacing and wavelength.

Most of the converted power goes to the first and second modes. They change the beam path and the beam size, respectively. In principle, both can be corrected with negligible loss by means of servomechanisms which redirect and refocus the beam.

Power converted to higher modes than the second distort the wavefront and must be absorbed by mode filters such as irises, for example.

For lenses with fourth-order aberration such that at a beam half-width from the center the focal length is 1 percent shorter than on axis, a 1 percent power conversion to the second mode occurs after few tens of lenses. Mode filters every few tens of lenses restore the original beam with losses of the order of 0.001 dB/lens. If refocusers and filters are used simultaneously, the second-order mode power is recovered and the losses are reduced by one order of magnitude.

Mode filters at every lens are, in general, lossier.

Long distance transmission through aberrated lenses such as our present form gas lenses seems possible, but it hinges heavily on our ability to build efficient and reliable redirectors, refocusers, and mode filters.

APPENDIX

Field in a Sequence of Distorted Lenses

The field on the surface α_n (Fig. 1) is made of a superposition of normal modes

$$E(\alpha_n) = \sum_{q=0}^{\infty} a_{qn} D_q(2\xi). \tag{50}$$

This field is related to the field $E(\beta_n)$ on the surface β_n by the phaseshift $\varphi_n(\sigma_n+\xi)$ given in (9). Thus,

$$E(\alpha_n) = E(\beta_n) \exp\,[i\varphi_n(\sigma_n + \xi)]. \tag{51}$$

Furthermore, the field $E(\beta_n)$ on the surface β_n is related to that on the surface α_{n-1} through the phaseshift of every mode,

$$E(\beta_n) = \sum_{q=0}^{\infty} a_{qn-1} \exp\,(-iq\theta) D_q(2\xi). \tag{52}$$

Substituting (50) and (52) in (51), we obtain

$$\sum_{q=0}^{\infty} a_{qn} D_q(2\xi) = \exp\,[i\varphi_n(\sigma_n + \xi)] \sum_{q=0}^{\infty} a_{qn-1} D_q(2\xi) \exp\,(iq\theta) \tag{53}$$

and because of the orthogonality property of the parabolic cylinder function

$$a_{qn} = \sum_{p=0}^{\infty} a_{pn-1} c_{pqn} \exp\,(-ip\theta), \tag{54}$$

where

$$c_{pqn} = \frac{\displaystyle\int_{-\infty}^{\infty} \exp\,[i\varphi_n(\sigma_n + \xi)] D_p(2\xi) D_q(2\xi)\, d\xi}{\displaystyle\int_{-\infty}^{\infty} D_q^2(2\xi)\, d\xi} \tag{55}$$

is the coupling coefficient between the pth and qth mode at the nth lens.

We are interested only in small lens distortions and small beam departure from ideal; therefore,

$$\begin{gathered} |\,\varphi_n(\sigma_n + \xi)\,| \ll \pi \\ a_{qn} \ll 1 \quad \text{for} \quad q > 0. \end{gathered} \tag{56}$$

Accordingly, keeping only first-order perturbation terms and expanding $\varphi_n(\sigma_n+\xi)$ in Taylor's series, we obtain for (54) and (55)

$$a_{qn} = \sum_{\nu=1}^{n} c_{oq\nu} \exp\,[iq(\nu - n)\theta] \tag{57}$$

and

$$c_{oq\nu} = \begin{cases} 1 & \text{if} \quad q = 0, \\ \dfrac{i}{2^q q!} \displaystyle\sum_{\mu=0}^{\infty} \frac{1}{2^{3\mu}\mu!} \frac{\partial^{q+2\mu}\varphi_\nu(\sigma_\nu)}{\partial\sigma_\nu^{q+2\mu}} & \text{if} \quad q > 0. \end{cases} \tag{58}$$

REFERENCES

1. Goubau, G. and Schwering, F., On the Guided Propagation of Electromagnetic Wave Beams, Trans. IRE, *AP-9,* May, 1961, p. 248.
2. Kompfner, R., Optical Reflecting System for Redirection of Energy, Patent No. 3,224,330, issued Dec. 21, 1965.
3. Marcuse, D. and Miller, S. E., Analysis of a Tubular Gas Lens, B.S.T.J., *43,* July, 1964, pp. 1759–1782.
4. Berreman, D. W., A Lens or Light Guide Using Convectively Distorted Thermal Gradients in Gases, B.S.T.J., *43,* July, 1964, pp. 1469–1475.
5. Tien, P. K., Gordon, J. P., and Whinnery, J. R., Focusing of a Light Beam of Gaussian Field Distribution in Continuous and Periodic Lenslike Media, Proc. IEEE, *53,* February, 1965, pp. 129–136.
6. Rowe, H., unpublished work.
7. Hirano, J. and Fukatsu, Y., Stability of a Light Beam in a Beam Waveguide, Proc. IEEE, *52,* November, 1964, pp. 1284–1292.
8. Marcuse, D., Statistical Treatment of Light-Ray Propagation in Beam-Waveguides, B.S.T.J., *44,* November, 1965, pp. 2065–2081.
9. Marcatili, E. A. J., Ray Propagation in Beam-Waveguides with Redirectors, B.S.T.J., *45,* January, 1966, pp. 105–116.
10. Marcuse, D., Deformation of Fields Propagating through Gas Lenses, B.S.T.J., *45,* October, 1966, pp. 1345–1368.
11. Marcatili, E. A. J., Off-Axis Wave-Optics Transmission in a Lens-Like Medium with Aberration, B.S.T.J., *46,* January, 1967, pp. 149–167.
12. Miller, S. E., Light Propagation in Generalized Lenslike Media, B.S.T.J., *44,* November, 1965, pp. 2017–2064.
13. Gordon, J. P., Optics of General Guiding Media, B.S.T.J., *45,* February, 1966, pp. 321–332.
14. Gloge, D., Mode Conversion Loss in a Gas-Filled Underground Lens-Waveguide, to be published.
15. Magnus, W. and Oberhittinger, F., *Formulas and Theorems for the Functions of Mathematical Physics,* Chelsea Publ. Co., New York, 1954, p. 93.
16. Chinnock, E. L., A Gas Lens-Beam Deflector, to be published; Electronic Control for the Gas Lens-Beam Deflector, to be published.
17. Kaiser, P., Experiments with a Beam Waveguide Made of Tubular Gas Lenses, to be published.
18. Kogelnik, H., Imaging of Optical Modes—Resonators with Internal Lenses, B.S.T.J., *44,* March, 1965, pp. 455–495.

Communications and Radar Receiver Gains for Minimum Average Cost of Excluding Randomly Fluctuating Signals in Random Noise

By STEPHEN S. RAPPAPORT

(Manuscript received October 25, 1966)

The problem of automatic gain control is approached from a statistical point of view. A simple generic equation is found whose solution yields the required receiver gain or attenuation for minimum average cost of excluding (from the receiver's limited dynamic range) randomly fluctuating signals in random noise. A canonical phase-incoherent link is considered and the resulting transcendental equation is solved using an iterative technique. The analysis and the results obtained apply to both linear and nonlinear incoherent receivers including those of the logarithmic or lin-log type and to a range of fluctuation models including Rician, Rayleigh, and nonfluctuating cases. It is shown that the optimum receiver gain is relatively insensitive to the ratio of costs of saturation at the upper and lower dynamic range bounds, differing at most by about 3 dB from the optimum for the equal cost (minimum exclusion probability) case for typical parameters. The effect of noise introduced by the gain adjustment cascade itself is discussed.

The results, presented in concise normalized form, are applicable to a wide range of signal, noise, and channel conditions and have important implications for communications through fading channels as well as for radar observation of fluctuating targets.

I. INTRODUCTION

Since the ability of both communications and radar receivers to perform satisfactorily can be seriously degraded when the signal amplitude does not lie within the dynamic range of the receiver, the setting of receiver gain to minimize or prevent saturation at the upper and lower dynamic range bounds is an important problem. The problem arises in various forms. In simple receivers, the gain might be fixed to optimize performance for nominal signal and noise parameters. More complicated receivers can adjust the gain automatically by any of several methods. The most common AGC circuits, for example, use a time averaged baseband signal as an indication of signal strength. Another possibility is to have the gain adjusted on command from a digital computer. This latter configuration has important implications for communications terminals which can use sophisticated techniques for estimation of signal and noise parameters as well as for certain radars which must observe from look-to-look radar targets of different cross-section which have been illuminated by various transmitted waveforms. The analysis presented here does not depend on the particular configuration and is applicable to both linear and nonlinear receivers including those of the logarithmic and lin-log type. The application to a nonlinear receiver can be accomplished by referring the overall dynamic range of the signal processing chain to a point before the nonlinearity.

In a recent correspondence, Ward[1] determined the placement of dynamic range bounds to minimize the probability of excluding a Rayleigh distributed signal. This was extended by Rappaport[2] who determined dynamic range bounds for minimum probability of excluding a signal from the dynamic range of incoherent radar or communications receivers. The viewpoint taken there[2] considered randomness due *either* to background noise *or* target fluctuations (channel fading). This paper considers several further generalizations of the problem. The case in which the randomness is due to both causes together is treated. In addition, the criterion for optimization is taken as the minimum average cost of exclusion. The required receiver gains as well as the optimum dynamic range bounds are determined.

The present paper proceeds from the specific to the general. That is, first the determination of optimum dynamic range bounds for minimum exclusion probability with non-fluctuating target (no channel fading) is presented.

The criterion is then generalized from minimum exclusion probability to minimum average exclusion cost; the former being a special case of the latter. Finally, dynamic range bounds and receiver gains for minimum average exclusion cost in an environment of fluctuating targets or channel fading is determined. It is assumed throughout that the signal, noise, and fluctuation parameters are known to the receiver. By letting the parameters involved assume certain values the relations for the fluctuating case reduce to the nonfluctuating case. Hence, the general treatment presented here includes either criteria and either the case of fluctuating or nonfluctuating SNR. Rician and Rayleigh SNR fluctuations are considered.

Fig. 1 shows a model for an incoherent radar or communications link. The blocks labeled K_1, K_2, and K_3 represent variable gain devices (perhaps, variable attenuation pads) whose total gain $K = K_1K_2K_3$ is to be adjusted so that the random signal appearing at (E) is in some sense confined to a specified range. The model used for the propagation medium and/or target is described in Section IV. Extension of the explicit results obtained here to an important class of nonlinear receivers by conceptually including a zero-memory nonlinear device between points (D) and (E) will be described subsequently. The other blocks in the figure require no further explanation. The figure is presented so that the reader can obtain a clear understanding of where in the signal processing chain various quantities arising in the following analysis are being determined. However, the analysis applies to incoherent signal processing links in general and is not constrained, for example, by the number of components or IF frequencies that may be used.

The receiver structure shown in Fig. 1 may be used for recovering the envelope of a transmitted sinusoidal signal or it may represent an optimum incoherent receiver for the detection of finite duration signals of known form in a background of Gaussian noise. The probability density function (pdf) of interest in the former case is that of the voltage appearing at the input to the video processing

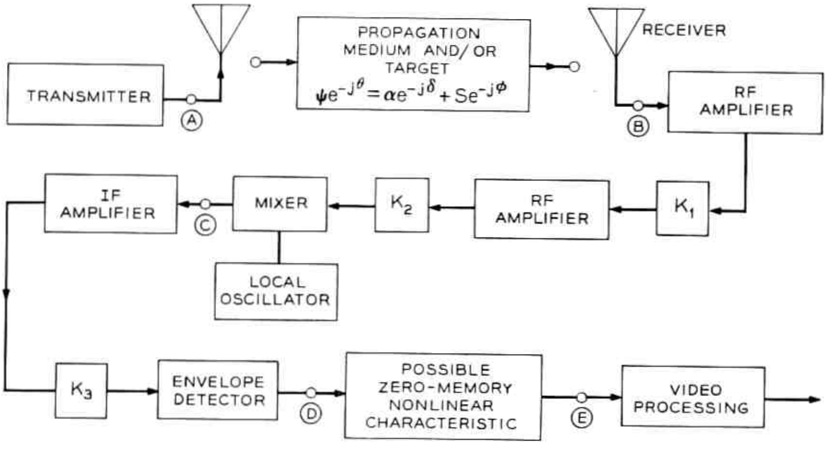

TRANSMITTER
A
PROPAGATION MEDIUM AND/OR TARGET
$\psi e^{-j\theta} = \alpha e^{-j\delta} + Se^{-j\phi}$
RECEIVER
B
RF AMPLIFIER
K1
RF AMPLIFIER
K2
MIXER
C
IF AMPLIFIER
LOCAL OSCILLATOR
K3
ENVELOPE DETECTOR
D
POSSIBLE ZERO-MEMORY NONLINEAR CHARACTERISTIC
E
VIDEO PROCESSING

Fig. 1 — Model for an incoherent radar or communications link.

[point (E) in Fig. 1] while in the latter case the pdf of concern is that of the voltage at the same point in the receiver at the decision time only. Either of these cases gives rise to a Rician pdf. Differences between the two are reflected only in the definition of a suitable SNR.[3, 4]

The analysis presented here can be easily extended to determine optimum gain settings for an important class of nonlinear incoherent receivers; namely, those in which the nonlinearity can be represented by a memoryless nonlinear device acting on the envelope of the received signal. For example, logarithmic or lin-log receivers can be represented by a logarithmic or lin-log characteristic inserted between points (D) and (E) regardless of whether the actual nonlinear device is a video or IF amplifier. One needs only to first refer the dynamic range and equivalent limiting voltages from the output of the nonlinear characteristic [point (E)] to the corresponding values at the input to the characteristic [point (D)] and then to determine the gain setting by considering only the linear portion of the receiver. In what follows it will be assumed that this first step has been taken if necessary and only the linear incoherent receiver will be treated explicitly.

II. DYNAMIC RANGE BOUNDS FOR MINIMUM EXCLUSION PROBABILITY WITH NONFLUCTUATING SNR

For nonfluctuating SNR the voltage gain of the radar or communications link is fixed. It is convenient to assume (without loss of generality) that the voltage gain ψ, of the propagation medium and/or target [i.e., the portion of the link from (A) to (B) in Fig. 1] is unity. In the more general case of fluctuating SNR, the voltage gain of this portion of the link will be treated as a random quantity.

The optimum placement of dynamic range bounds for incoherent receivers is determined by the pdf of the envelope detected signal, v, which appears at point (D) in Fig. 1. Let σ be a normalization parameter and define

R = normalized envelope of received signal = v/σ
a = lower normalized bound of dynamic range
ad = upper normalized bound of dynamic range
$D = 20 \log_{10} d$ = dynamic range in dB,

where these quantities are referred to point (D) in Fig. 1. If $p_\gamma\ (R)$ denotes the pdf of the normalized envelope, the corresponding exclusion probability is

$$P_e(a, d) = 1 - \int_a^{ad} p_\gamma(R)\, dR. \tag{1}$$

To minimize $P_e(a, d)$ with respect to a for fixed dynamic range d, (1) is differentiated with respect to a and this derivative is set to zero. This yields the necessary optimization condition

$$p_\gamma(a) = dp_\gamma(ad) \tag{2}$$

which must be solved for a. Consider an optimum incoherent receiver for detection of signals of known form in Gaussian noise. Let $2\sigma^2$ be the mean square value of the signal voltage envelope, v, when only noise is present. In this case, the pdf of the normalized signal envelope is

$$p_o(R) = R \exp\,[-R^2/2]. \tag{3}$$

The optimization condition (2) for this case leads to

$$A^2 = \frac{2 \ln (d)}{d^2 - 1} \triangleq A_o^2 \tag{4}$$

in which $A \triangleq a/\sqrt{2}$ is the optimum normalized lower dynamic range bound. When signal-plus-noise is present the probability density of the normalized envelope has a Rician distribution[3,4]

$$p_\gamma(R) = R \exp\,[-(R^2 + \gamma^2)/2]I_o(\gamma R) \tag{5}$$

in which

$I_o(x)$ = modified Bessel function of first kind and order zero
γ = voltage signal-to-noise ratio for $\psi = 1$.

In this case condition (2) gives the following transcendental equation which must be solved for the optimum normalized lower dynamic range bound $A = a/\sqrt{2}$:[2]

$$A^2 = A_o^2 + \frac{1}{(d^2 - 1)}\,[\ln I_o(A\, d\gamma\sqrt{2}) - \ln I_o(A\gamma\sqrt{2})]. \tag{6}$$

The minimum exclusion probability becomes

$$P_e(a, d) = 1 - Q(\gamma, A\sqrt{2}) + Q(\gamma, Ad\sqrt{2}) \tag{7}$$

in which $Q(\alpha, \beta)$ is the tabulated[5] Q-function defined by

$$Q(\alpha, \beta) = \int_\beta^\infty \xi \exp\,[-(\xi^2 + \alpha^2)/2]I_o(\alpha\xi)\, d\xi. \tag{8}$$

Solutions to (6) for various γ and d are presented in Ref. 2 along with minimum exclusion probabilities for this case. The solutions can be obtained from Fig. 2 with γ used in place of $\tilde{\gamma}$ and A in place of $\tilde{A}$.

III. DYNAMIC RANGE BOUNDS FOR MINIMUM AVERAGE EXCLUSION COST WITH NONFLUCTUATING SNR

In certain situations it may not be desirable to use dynamic range bounds which minimize the exclusion probability. It may be reasonable to favor saturation at one dynamic range bound to the other. In the case of a radar, for example, the signal is invisible if it falls beneath the lower dynamic range bound, while if the receiver saturates at the upper dynamic range bound the presence of the signal would be detected although its information content would be corrupted by the limiting. In such cases, a more reasonable criterion might be to minimize the average cost of excluding the signal from the receiver's dynamic range.

Suppose that when the signal falls below the lower bound a loss, c_1, is incurred, while saturation at the upper bound causes a loss, c_2

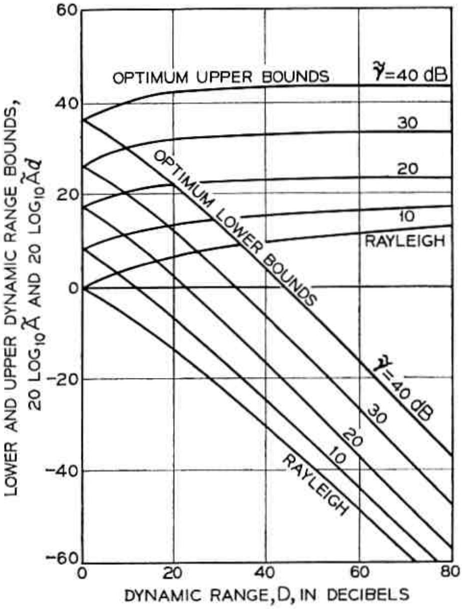


Fig. 2 — Optimum dynamic range centering for $\nu_{dB} = 0$.

(c_1, $c_2 > 0$). The expected or average cost of excluding the signal from the dynamic range is

$$L = c_1 \int_0^a p_\gamma(R)\, dR + c_2 \int_{ad}^{\infty} p_\gamma(R)\, dR. \tag{9}$$

It is convenient to divide (9) by c_1 to obtain the normalized average exclusion cost

$$l = \int_0^a p_\gamma(R)\, dR + \nu \int_{ad}^{\infty} p_\gamma(R)\, dR \tag{10}$$

in which ν is the cost ratio, $\nu \triangleq c_2/c_1$. It is seen that for $\nu = 1$ the normalized average exclusion cost, l, reduces to the exclusion probability. In order to minimize the average exclusion cost, the derivative of (10) with respect to a is set to zero. One then finds that the optimum lower normalized dynamic range bound a must satisfy

$$p_\gamma(a) = \nu\, d p_\gamma(ad), \tag{11}$$

which reduces to (2) for $\nu = 1$ as it should. For the incoherent receiver substitution of (5) in (11) leads to the following transcendental equation for the optimum normalized lower dynamic range bound*

$$A^2 = A_c^2 + A_o^2 + \frac{1}{(d^2 - 1)} [\ln I_o(A\, d\gamma \sqrt{2}) - \ln I_o(A\gamma\sqrt{2})] \tag{12}$$

in which by definition

$$A_c^2 = \frac{\ln \nu}{d^2 - 1}. \tag{13}$$

It is noted that it is entirely possible for c_2 to be less than c_1 making A_c^2 negative. However, the sum $A_c^2 + A_o^2$ is positive if νd^2 is greater than unity. Using (12) it is seen that for $\gamma = 0$, i.e., Rayleigh distributed envelope, the optimum normalized lower bound can be determined explicitly from

$$A_R^2 = A_c^2 + A_o^2. \tag{14}$$

When the cost ratio, ν, is unity (14) reduces to (4) as expected. It is convenient to measure the cost ratio in dB using

$$\nu_{\text{dB}} = 20 \log_{10} \nu. \tag{15}$$

* The desired lower dynamic range bound is the positive real root of (12).

Thus, positive values of ν_{dB} imply $c_2 > c_1$, while negative values imply $c_2 < c_1$. $\nu_{\text{dB}} = 0$ is the case of minimum exclusion probability. If $\nu_{\text{dB}} = -2D$, then A_R^2 is 0 and $A^2 = 0$ solves (12) for any γ. For the foregoing formulation to be meaningful, A^2 must be positive. Hence, (4), (12), (13), and (14) require that $\nu_{\text{dB}} > -2D$. If this constraint is not satisfied, then the average exclusion cost, l, is not stationary with respect to A. For given γ, d, and ν it is generally necessary to solve (12) for A. This equation is of the general form $x = f(x)$. A proposed scheme to find the solution is to iterate $x_{i+1} = f(x_i)$ beginning with an approximate solution x_o. It can be shown that this scheme will converge if the magnitude of the derivative of the RHS, $| f'(x) |$, is less than unity in the neighborhood of the solution, x_s. Moreover the convergence is faster as $| f'(x_s) |$ is closer to zero. In order to speed convergence an *extrapolated* iteration scheme can be used by introducing another parameter, β. Consider the equation

$$x = f(x) - \beta[x - f(x)]. \tag{16}$$

Provided $\beta \neq -1$ the solution to this equation is the same as that of $x = f(x)$. If one could choose

$$\beta = \frac{f'(x_s)}{1 - f'(x_s)} \qquad f'(x_s) \neq 1, \tag{17}$$

the derivative of the RHS of (16) would be zero at x_s. However, since x_s is not known at the outset the approximate solutions are substituted for x_s in (17) to speed convergence.

Using this approach (12) can be solved to any desired accuracy with the aid of the iteration formulas

$$A_{i+1}^2 = F_i - \beta_i(A_i^2 - F_i) \qquad \beta_i \neq -1 \tag{18}$$

$$F_i = A_c^2 + A_o^2 + \frac{1}{(d^2 - 1)} [\ln I_o(A_i d\gamma\sqrt{2}) - \ln I_o(A_i\gamma\sqrt{2})] \tag{19}$$

$$G_i = \frac{\gamma\sqrt{2}}{2(d^2 - 1)A_i}\left[d\,\frac{I_1(A_i d\gamma\sqrt{2})}{I_o(A_i d\gamma\sqrt{2})} - \frac{I_1(A_i\gamma\sqrt{2})}{I_o(A_i\gamma\sqrt{2})}\right] \tag{20}$$

$$\beta_i = G_i/(1 - G_i) \qquad G_i \neq 1 \tag{21}$$

in which $I_n(x)$ denotes the modified Bessel function of the first kind and nth order. One may begin with small values of γ, $i = 0$, A_o^2 given by (4) and A_c^2 given by (13). The iteration is stopped when $| A_{i+1} - A_i |$ is less than the allowable error. Equation (12) was solved by this method for various values of ν_{dB}, D(dB), and γ(dB).

For $d \gg 1$ an approximate solution to (12) can be found explicitly.

Neglecting the second term in brackets in comparison with the first and taking $I_o(x) \approx e^x$ reduces (12) to a quadratic equation in Ad which has the solution $Ad \approx (\gamma/\sqrt{2}) + \sqrt{(\gamma^2/2) + \ln \nu\, d^2}$. Thus, the optimum normalized upper bounds (as shown in Fig. 2 for $\nu = 1$) continue to increase slowly as D increases.

The solutions obtained using (18) to (21) show that the optimum lower bounds for a wide range of cost ratios do not differ appreciably from those for $\nu_{\mathrm{dB}} = 0$. The difference (in dB) in optimum lower bounds for values of $\nu_{\mathrm{dB}} = \pm 25$ and $\nu_{\mathrm{dB}} = 0$, and for $\nu_{\mathrm{dB}} = \pm 50$ and $\nu_{\mathrm{dB}} = 0$ for various values of $\tilde{\gamma}$ and d are shown in Fig. 3(a) and (b). (For non-fluctuating SNR take $\tilde{\gamma}$ in the figures as γ and $\tilde{A}$ as A.) It can be seen that for any given values of ν and d, the largest difference is for $\tilde{\gamma} = 0$. This maximum deviation can be determined explicitly. From (14)

$$A_R^2 = A_o^2(1 + A_c^2/A_o^2). \tag{22}$$

Equations (4), (13), and (15) then yield

$$20 \log_{10} A_R - 20 \log_{10} A_o = 10 \log_{10}\left(1 + \frac{\nu_{\mathrm{dB}}}{2D}\right). \tag{23}$$

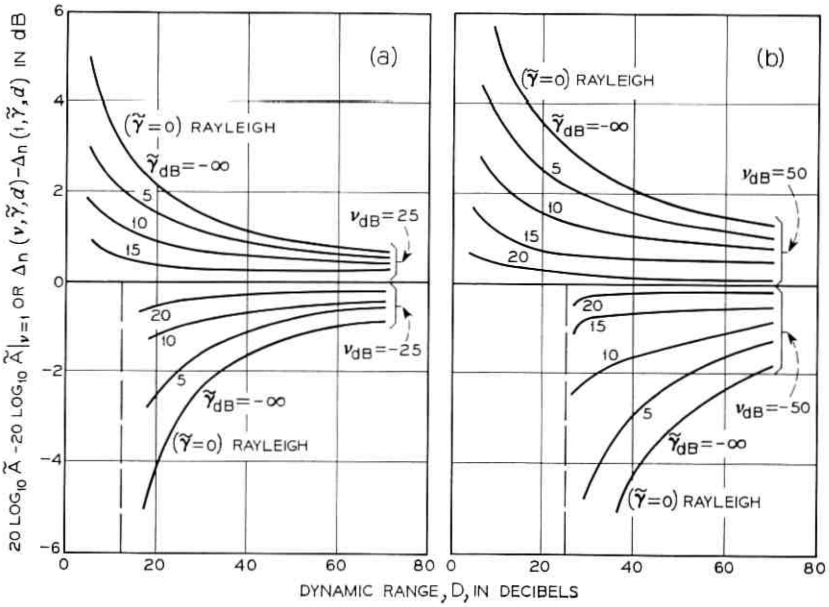


Fig. 3. — Increase in optimum normalized dynamic range bounds or in required receiver attenuation due to nonunity cost ratio. (a) $\nu_{\mathrm{dB}} = -25{,}25$; (b) $\nu_{\mathrm{dB}} = -50{,}50$.

Thus, the maximum difference in optimum dynamic range bounds is determined by the ratio of the cost ratio in dB to the dynamic range in dB. A plot of (23) is shown in Fig. 4.

The fact that the optimum bound is relatively insensitive to cost ratio at least for large D and large $\tilde{\gamma}$ is an important one since exact assessment of the costs c_1 and c_2 is difficult or impossible. However, this analysis shows that for large $\tilde{\gamma}$ and D an optimum solution for $\nu_{\text{dB}} = 0$ is nearly optimum for $-D \leqq \nu_{\text{dB}} \leqq 2D$. For $\tilde{\gamma} = 0$ the optimum dynamic range bounds are most sensitive to cost ratio but in this region differ only by about 3 dB from the optimum for $\nu = 1$.

When the optimum normalized lower bound, A, is determined, the normalized minimum average cost of excluding the signal is given in this case by

$$l = 1 - Q(\gamma, A\sqrt{2}) + \nu Q(\gamma, Ad\sqrt{2}). \tag{24}$$

These minimum average exclusion costs are shown in Fig. 5, in which the parameter $\tilde{\gamma}$ is to be taken as γ. For $\nu_{\text{dB}} = 0$, (24) becomes the minimum exclusion probability (7).

IV. A MODEL FOR TARGET FLUCTUATION AND FADING CHANNELS

Thus far this paper has considered the case where the SNR at the receiver is fixed. However, in the case of radar the target cross section

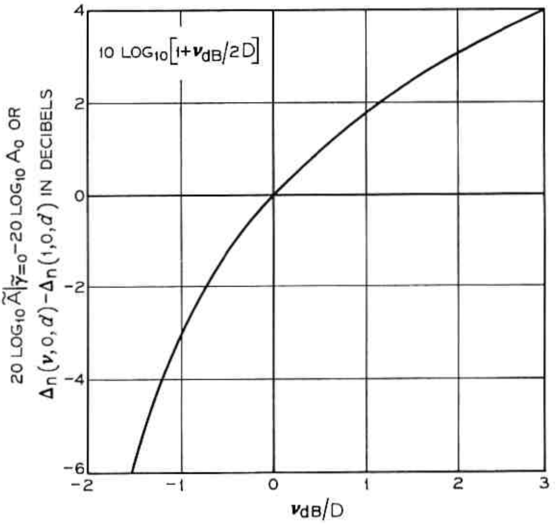


Fig. 4 — Maximum change in optimum normalized dynamic range bounds or in required receiver attenuation due to nonunity cost ratio.

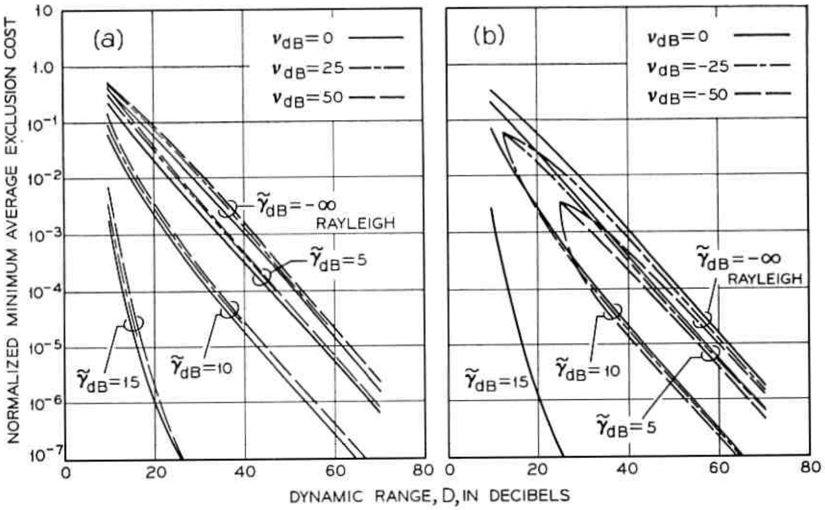


Fig. 5 — Minimum average exclusion costs.

presented to the receiver generally varies, while in communications links the phenomenon of fading generally causes fluctuation in the received SNR. Consider the case in which the SNR fluctuation is slow so that it is essentially fixed for the duration of a given signal but will fluctuate over longer time intervals.

Following Turin[6] it is assumed that the medium from the transmitter to the receiver can be characterized as propagating two components, a fixed or specular component and a completely random or scatter component. Thus corresponding to a transmitted signal Re $\{s(t) \exp(j\omega_c t)\}$ the reciever's IF signal [point (C) in Fig. 1] with $K \triangleq K_1K_2K_3 = 1$ is given by Re $\{x(t)\}$ where

$$x(t) = s(t - \tau)[\alpha \exp(-j\delta) + S \exp(-j\varphi)] \cdot \exp(j\omega_o t) + n(t) \exp(j\omega_o t). \tag{25}$$

In (25) α and δ are fixed while S and φ are independent variates; S having a Rayleigh pdf with mean square $2\mu^2$ and φ a uniformly distributed phase over an interval of 2π. ω_c and ω_o denote the angular frequencies of the transmitted carrier and the receiver intermediate frequency, respectively. $n(t)$ is a narrowband Gaussian noise process. It can be shown[3, 6] that the joint distribution of the resultant amplitude, ψ, and phase, θ, of the sum of the fixed vector (α, δ) and the random vector (S, φ) is

$$p(\psi, \theta) = \frac{\psi}{2\pi\mu^2} \exp\left[- \frac{\psi^2 + \alpha^2 - 2\alpha\psi \cos(\theta - \delta)}{2\mu^2} \right]$$

$$\text{for} \quad 0 \leqq \psi \qquad 0 \leqq \theta - \delta \leqq 2\pi. \tag{26}$$

In (26) α^2 can be regarded as the strength of the fixed component while μ^2 is proportional to the strength of the scattered component. The quantity $\beta^2 \triangleq \alpha^2/\mu^2$ is twice the ratio of the energy received via the specular component to that received via the scatter component. The variates ψ and θ are, respectively, the instantaneous voltage gain and phase shift of the path from the transmitter to receiver and δ is the average phase shift of the path. Note that (26) is just the two dimensional Gaussian distribution in polar form. The pdf of ψ is found by integrating (26) over the range of θ giving[3,6]

$$p(\psi) = \frac{\psi}{\mu^2} \exp\left[- \frac{\psi^2 + \alpha^2}{2\mu^2} \right] I_o\left(\frac{\alpha\psi}{\mu^2}\right) \quad \text{for} \quad \psi \geqq 0 \tag{27}$$

$$= 0 \text{ elsewhere.}$$

Letting $r = \psi/\mu$, (27) becomes*

$$p_\beta(r) = r \exp\left[-(r^2 + \beta^2)/2\right] I_o(\beta r). \tag{28}$$

The voltage gain of the propagation medium and/or target [from (A) to (B) in Fig. 1] is $\psi = \mu r$. The model above is an adequate representation of propagation conditions which are encountered on ionospheric and tropospheric radio links.[6] The pdf (27) is sufficiently general since as β approaches zero (no specular component) (27) becomes the Rayleigh distribution with parameter μ while if β approaches infinity (presence of specular component only (27) may first be approximated by a Gaussian pdf of mean α and variance μ^2 and in the limit by a delta function, $\delta(\psi - \alpha)$ corresponding to the case of no SNR fluctuation. Radar target fluctuations have been described by Rayleigh statistics[7] a special case of the above model ($\beta = 0$). In this case, μ^2 is proportional to the average target cross-section. It is reasonable to expect that radar targets which can be modeled as a single large reflector plus a large number of independent scatterers will yield signal returns of the form (25). For this more general Rician fluctuating target $\mu^2(1+\beta^2/2)$ is proportional to the average target cross-section.

* Note that (28) is a pdf of the same form as (5) as it must be since either equation is the probability density of the magnitude of the sum of a constant vector and a Gaussian vector.

V. GENERAL CASE: DYNAMIC RANGE BOUNDS FOR MINIMUM AVERAGE EXCLUSION COST WITH FLUCTUATING SNR

The phenomena of radar target fluctuation and channel fading are evidenced by fluctuating SNR in the receiver. To account for these fluctuations, γ in (10) must be weighted by a random voltage gain μr, where μ is a parameter and r is a random variable whose pdf $\hat{p}_\beta(r)$ determines the form of the SNR fluctuation. The normalized average exclusion cost can then be obtained from (10) giving

$$\tilde{l} = \int_0^a \tilde{p}_{\mu\gamma}(R)\,dR + \nu \int_{ad}^\infty \tilde{p}_{\mu\gamma}(R)\,dR, \tag{29}$$

where

$$\tilde{p}_{\mu\gamma}(R) = \int_0^\infty \hat{p}_\beta(r) p_{r\mu\gamma}(R)\,dr. \tag{30}$$

In (30), γ is the voltage SNR at the receiver for unity channel gain, i.e., for $\psi = r\mu = 1$. The product $r\mu\gamma$ appearing in the integrand is the voltage SNR at the receiver for a particular realization of the random gain $\psi = r\mu$; that is, the "instantaneous" voltage SNR at the receiver.

The condition for minimization of the average exclusion cost (29) becomes

$$\tilde{p}_{\mu\gamma}(a) = \nu\, d\tilde{p}_{\mu\gamma}(ad). \tag{31}$$

Consider phase incoherent reception of signals in Gaussian noise with fading or target fluctuations described by the probability law (28), i.e., $\hat{p}_\beta(r) = p_\beta(r)$. In this case the integral appearing in (30) becomes

$$\tilde{p}_{\mu\gamma}(R) = \int_0^\infty Rt \exp\,[-(t^2 + \beta^2)/2] I_o(\beta t) \cdot \exp\,[-(\mu^2\gamma^2 t^2 + R^2)/2] I_o(\gamma R t)\,dt \tag{32}$$

which can be evaluated using an identity in Watson* giving

$$\tilde{p}_{\mu\gamma}(R) = \frac{R}{1 + \mu^2\gamma^2} \exp\left[-\frac{(R^2 + \mu^2\gamma^2\beta^2)}{2(1 + \mu^2\gamma^2)}\right] I_o\left(\frac{\mu\gamma\beta R}{1 + \mu^2\gamma^2}\right). \tag{33}$$

To evaluate the average exclusion cost (29) one needs to integrate (32) or (33) with respect to R from some number η to ∞. This integral of (33) can be easily evaluated using the definition (8).

* See Ref. 8, p. 395. Take Watson's $a = i\mu\gamma R$, $b = i\beta$, $\nu = 0$, $p^2 = (1 + \mu^2\gamma^2)/2$.

Performing the integration with respect to R in (32) then establishes the result†

$$\begin{aligned}\int_\eta^\infty p_{\mu\gamma}(R)\,dR &= \int_0^\infty t \exp\,[-(t^2+\beta^2)/2]I_o(\beta t)Q(\mu\gamma t,\,\eta)\,dt \\ &= Q\left(\frac{\mu\gamma\beta}{\sqrt{1+\mu^2\gamma^2}},\,\frac{\eta}{\sqrt{1+\mu^2\gamma^2}}\right).\end{aligned} \tag{34}$$

Using (34) in (29) yields

$$\begin{aligned}\bar{l} = 1 - Q&\left(\frac{\mu\gamma\beta}{\sqrt{1+\mu^2\gamma^2}},\,\frac{a}{\sqrt{1+\mu^2\gamma^2}}\right) \\ &+ \nu Q\left(\frac{\mu\gamma\beta}{\sqrt{1+\mu^2\gamma^2}},\,\frac{ad}{\sqrt{1+\mu^2\gamma^2}}\right)\end{aligned} \tag{35}$$

in which a is the optimum normalized lower dynamic range bound obtained as a solution to (31). By substituting (33) in (31) and manipulating the result it can be shown that the optimum a must satisfy

$$\begin{aligned}\frac{a^2}{2(1+\mu^2\gamma^2)} = A_c^2 + A_o^2 + &\frac{1}{(d^2-1)} \\ &\cdot\left[\ln I_o\left(\frac{\mu\gamma\beta ad}{1+\mu^2\gamma^2}\right) - \ln I_o\left(\frac{\mu\gamma\beta a}{1+\mu^2\gamma^2}\right)\right].\end{aligned} \tag{36}$$

Let

$$\tilde{A} = \frac{a}{\sqrt{2}\,\sqrt{1+\mu^2\gamma^2}} \tag{37}$$

and

$$\tilde{\gamma} = \frac{\mu\gamma\beta}{\sqrt{1+\mu^2\gamma^2}}. \tag{38}$$

From (33) it can be seen that if β is zero the mean square value of R is $2(1+\mu^2\gamma^2)$. Thus $\tilde{A}$ is the optimum lower dynamic range bound normalized to the rms voltage that would appear at the output of the envelope detector [point (D) in Fig. 1], if the specular component were zero. Since $\mu\beta = \alpha$ the quantity $\tilde{\gamma}$ in (38) is twice the ratio of the rms voltage that would appear at (D) when only the noiseless specular com-

† The integral to the right of the first equality in (34) would appear if the average cost for the nonfluctuating case is determined first as in (24) and then this cost is averaged over the random fluctuations of ψ.

ponent of (25) is present, to the rms voltage that would appear if only the scatter and noise components of (25) were present. The mean square voltage at the output of the envelope detector when specular, scatter, and noise components are present is $2\sigma^2[1 + \mu^2\gamma^2 + (\alpha^2\gamma^2)/2]$. Using (37) and (38) in (36) gives

$$\tilde{A}^2 = A_c^2 + A_o^2 + \frac{1}{(d^2 - 1)} [\ln I_o(\tilde{\gamma}\tilde{A}\, d\sqrt{2}) - \ln I_o(\tilde{\gamma}\tilde{A}\sqrt{2})] \tag{39}$$

and (35) becomes

$$\bar{l} = 1 - Q(\tilde{\gamma},\, \tilde{A}\sqrt{2}) + \nu Q(\tilde{\gamma},\, \tilde{A}\, d\sqrt{2}). \tag{40}$$

Equations (39) and (40) are of the same form as (12) and (24), respectively, with γ replaced by $\tilde{\gamma}$ and A by $\tilde{A}$. These results show that the optimum dynamic range bounds and performance curves obtained previously for nonfluctuating SNR can be used directly for the more general case of fluctuating SNR by merely changing the variables via (37) and (38). Therefore, although there are two additional parameters in the fluctuating case it is not necessary to increase the number of curves to describe performance. For $\nu_{\text{dB}} = 0$ the criterion reduces to the minimum average exclusion probability as in the case of nonfluctuating SNR.

VI. RECEIVER GAIN REQUIRED FOR THE GENERAL CASE

The optimum gain or attenuation required for insertion in the signal processing chain at a point preceding the components which limit the dynamic range can now be calculated. Let c be the lowest voltage at which the signal processing chain can operate satisfactorily, referred to the output of the envelope detector.* Optimum dynamic range utilization requires that the signal be multiplied by a factor K such that the scaled lower normalized dynamic range bound is equal to the voltage c, when normalized to the same base. That is,

$$Ka = c/\sigma. \tag{41}$$

Substituting from (37) for a and transposing, (41) gives

$$K(\sigma\sqrt{2}/c)\sqrt{1 + \mu^2\gamma^2} = \tilde{A}^{-1} \tag{42}$$

in which $\tilde{A}$ is the solution to (39). Denote the LHS of (42) as K_n, the normalized voltage gain, and let Δ_n be the normalized required attenuation in dB. Then

$$\Delta_n = -20 \log_{10} K_n = 20 \log_{10} \tilde{A} \tag{43}$$

* Point (D) in Fig. 1.

which expresses the normalized required attenuation in dB as a function of the optimum normalized lower dynamic range bound. Since $\tilde{A}$ is the solution to (39) Δ_n depends only on ν, $\tilde{\gamma}$, and d. It is fortunate that the normalized results can be expressed in terms of only a few parameters since this permits a concise description of optimum performance for many signal, noise and channel conditions. Optimum normalized attenuation required for the case of minimum exclusion probability ($\nu_{\text{dB}} = 0$) is shown plotted in Fig. 6. From (43) the actual required attenuation in dB can be obtained. Let $\Delta = -20 \log_{10} K$ denote the actual required attenuation in dB. Then (42) and (43) yield

$$\Delta = \Delta_n(\nu, \tilde{\gamma}, d) + 20 \log_{10} (\sigma \sqrt{2}/c) + 10 \log_{10} (1 + \mu^2 \gamma^2) \qquad (44)$$

in which the functional dependence of Δ_n is shown explicitly.

From (43) and (44) it can be seen that the difference in optimum receiver attenuations is the same as the difference in optimum normalized dynamic range bounds. Hence, Fig. 3(a) and (b) also show the differences

$$\Delta_n(\nu, \tilde{\gamma}, d) - \Delta_n(1, \tilde{\gamma}, d) \qquad (45)$$

for values of $\nu_{\text{dB}} = \pm 25, \pm 50$, and various values of $\tilde{\gamma}$. For given ν, $\tilde{\gamma}$, and d one can, therefore, determine $\Delta_n(\nu, \tilde{\gamma}, d)$ by finding $\Delta_n(1, \tilde{\gamma}, d)$

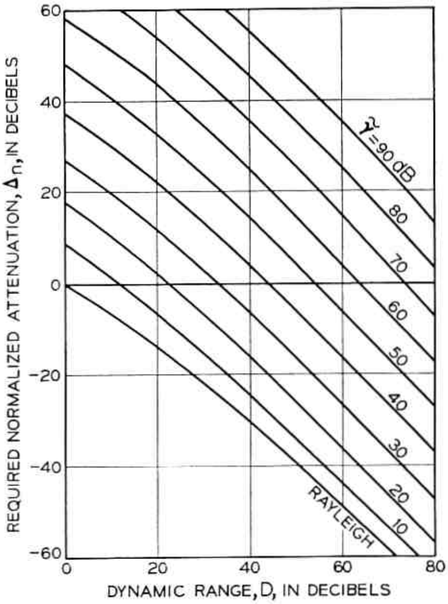


Fig. 6 — Required attenuation for minimum exclusion probability. $\nu_{\text{dB}} = 0$.

in Fig. 6 and adding the difference (45) found in Fig. 3. Fig. 4 is a plot of the differences

$$\Delta_n(\nu, 0, d) - \Delta_n(1, 0, d), \tag{46}$$

that is, a plot of (45) for zero SNR. For given ν and d these differences are of the same sign as (45) but are always larger in magnitude. Hence, Fig. 4 shows the maximum change in optimum receiver attenuation due to a nonunity cost ratio.

The definitions of the parameters appearing in (44) are summarized by the following list:

ν = cost ratio
σ^2 = noise power with no fluctuation
γ = voltage signal-to-noise ratio for unity propagation and/or target gain (i.e., $\psi = 1$)
μ^2 = strength of scatter component of the propagation medium
β^2 = twice the ratio of strength of specular component of the medium to that of the scatter component
$\mu^2(1 + \beta^2/2)$ = for the case of Rician fluctuating radar targets this quantity is proportional to the average target cross-section over all target fluctuations. $\beta = 0$ corresponds to the case of Rayleigh fluctuating targets
d = dynamic range of receiver

$$\begin{aligned}\tilde{\gamma} &= \mu\gamma\beta/\sqrt{1 + \mu^2\gamma^2} \\ &= 2 \times \frac{\text{rms voltage at } (D) \text{ for noiseless specular component only}}{\text{rms voltage at } (D) \text{ for scatter and noise components only}}.\end{aligned}$$

In the general analysis presented here, which includes fluctuating or nonfluctuating SNR and the criteria minimum average exclusion cost or minimum exclusion probability, special cases which may arise in various applications are represented when the parameters take on particular values. Some special cases are shown in Table I. The entries in the table are for either criterion.

TABLE I — CONSTRAINTS ON PARAMETERS FOR SPECIAL CASES

Constraints on parameters	Type of fluctuation or fading	Type of envelope detected signal
$\mu > 0, \beta > 0, \ \gamma > 0$	Rician	Rician
$\mu > 0, \beta = 0, \ \gamma > 0$	Rayleigh	Rayleigh
$\mu \to 0, \beta \to \infty, \mu\beta = \alpha, \gamma > 0$	none	Rician
$\mu\beta = \alpha, \gamma = 0$	none	Rayleigh

It is noted that for μ, β, and γ greater than zero one has the general case of Rician SNR fluctuation and Rician envelope detected signal.

When β is zero the medium from transmitter to receiver does not propagate any specular component and the envelope of the received signal has a Rayleigh pdf independent of the other parameters. Setting β to zero in (33) shows that the envelope of the received signal in this case has a mean square of $2\sigma^2(1 + \mu^2\gamma^2)$. When γ is zero only noise at the receiver is demodulated again giving rise to a Rayleigh distributed envelope but of mean square $2\sigma^2$ independent of the other parameters. In each of these two cases (i.e., $\beta = 0$ and $\gamma = 0$), the optimum *normalized* lower bound is found from (14), $\tilde{A}_R = \sqrt{A_c^2 + A_o^2}$. Since the minimum average exclusion cost (40) depends on the value of $\tilde{A}$ the minimum costs are equal for these two cases. However, it can be seen from (44) that the *actual* optimum lower bounds or *actual* required attenuations for these cases differ. This is because the quantity, $\sigma\sqrt{1 + \mu^2\gamma^2}$, to which the received signal voltage envelope is normalized is different in these instances. Note that in the former case ($\beta = 0$) the required receiver attenuation (44) is affected by the randomness of the scatter component while in the latter case ($\gamma = 0$) it is not. This can also be seen from (25). When β is zero there is no specular component and the received signal (25) depends upon the scatter component while if γ is zero the entire first term can be omitted and the received signal consists of only noise.

When μ goes to zero and β approaches infinity such that $\mu\beta = \alpha$ (a constant), the medium from transmitter to receiver propagates only a specular component with a voltage gain of α. In this case there is no SNR fluctuation (nonfluctuating case) and the envelope of the detected signal is Rician if $\gamma > 0$ and Rayleigh if $\gamma = 0$. Letting $\mu = 0$ and $\mu\beta = \alpha$ in (33) and (38) shows that for this case the SNR at the reciever is $\alpha\gamma$, a result which is clear from (25) if the scatter component of the medium is deleted. There is no essential loss in generality in this case if α is taken as unity. With $\mu = 0$ and $\alpha \triangleq \mu\beta = 1$ in (33) that equation reduces to the pdf considered in Sections II and III.

For $\gamma > 0$, $\mu > 0$, β finite, the optimum gain settings for the fluctuating and nonfluctuating cases differ and the minimum average exclusion cost (probability) for the fluctuating case will be greater for the same values of γ, d, ν, and α.

VII. EFFECT OF NOISE INTRODUCED BY THE GAIN ADJUSTMENT CASCADE

In the foregoing discussion, the attenuation or gain required for optimum dynamic range utilization has been idealized as a multiplicative parameter. These results apply when the noise introduced by the gain adjustment cascade itself is virtually independent of its gain. This condition is often realized in practice. When this condition is not satisfied some modification is necessary. The phenomenon can be represented by using an equivalent noise source at the point in the signal processing chain where the dynamic range calculations are being made. The quantities σ and γ used previously must be replaced by equivalent σ_e and γ_e, respectively. Let $\mathcal{F}_o(\mathcal{G})$ be the operating noise figure[9] of the cascade when it is set for an available gain of $\mathcal{G}$ (dB). Then the equivalent noise power when the gain is $\mathcal{G}$, is

$$\sigma_e^2(\mathcal{G}) = \frac{\sigma^2 \mathcal{F}_o(\mathcal{G})}{\mathcal{F}_o(\mathcal{G}_o)} \tag{47}$$

in which $\mathcal{G}_o$ is the gain for which the SNR is γ and the noise power is σ^2. The equivalent SNR is determined by

$$\gamma_e^2(\mathcal{G}) = \gamma^2 \frac{\mathcal{F}_o(\mathcal{G}_o)}{\mathcal{F}_o(\mathcal{G})}. \tag{48}$$

In the case where the noise depends upon the gain, K, ($\mathcal{G} = 20 \log_{10} K$), both the quantities γ and a in (29) or (35) depend upon gain. Hence, the optimization condition must be found by differentiating (29) or (35) with respect to K (or $\mathcal{G}$) rather than a and setting that derivative to zero. However, this condition is generally too complicated to be useful and it is usually better to evaluate (29) or (35) for various $\mathcal{G}$ to determine the optimizing value. For the general case of the incoherent receiver the normalization for R in (29) is with respect to σ_e rather than σ. In addition γ_e and σ_e must be used. Define

$$a_e = c/[K\sigma_e(\mathcal{G})] \tag{49}$$

and

$$\tilde{A}_e(\mathcal{G}) = \frac{a_e}{\sqrt{2}\,\sqrt{1 + \mu^2\gamma_e^2}}. \tag{50}$$

Then

$$\begin{aligned} \tilde{A}_e(\mathcal{G}) &= \tilde{A}\left[\frac{\mathcal{F}_o(\mathcal{G}_o)}{\mathcal{F}_o(\mathcal{G})}\right]^{\frac{1}{2}}\left[\frac{1 + \mu^2\gamma^2}{1 + \mu^2\gamma_e^2}\right]^{\frac{1}{2}} \\ &= \frac{c\cdot 10^{(-\mathcal{G}/20)}}{\sigma\sqrt{2}\,\sqrt{1 + \mu^2\gamma_e^2}}\left[\frac{\mathcal{F}_o(\mathcal{G}_o)}{\mathcal{F}_o(\mathcal{G})}\right]^{\frac{1}{2}}. \end{aligned} \tag{51}$$

From (38), (47), and (48) one can find

$$\tilde{\gamma}_e(\mathcal{G}) = \frac{\mu\beta\gamma_e(\mathcal{G})}{[1 + \mu^2\gamma_e^2(\mathcal{G})]^{\frac{1}{2}}}. \tag{52}$$

In order to find the optimum receiver gain (47), (48), (51), and (52) are used in conjunction with

$$\bar{l} = 1 - Q[\tilde{\gamma}_e(\mathcal{G}), \sqrt{2}\, \tilde{A}_e(\mathcal{G})] + \nu Q[\tilde{\gamma}_e(\mathcal{G}), d\sqrt{2}\, \tilde{A}_e(\mathcal{G})] \tag{53}$$

which must be minimized with respect to $\mathcal{G}$. It is easiest to use a numerical method which requires only successive evaluation of (53) for various values of $\mathcal{G}$, as opposed to methods requiring analytical evaluation of the derivative of (53). In the important case where only a finite number of gain settings $\mathcal{G}_i$, $(i = 1, 2, \cdots N)$ are possible, minimization of (53) is easy requiring at most N evaluations for any given set of parameters. Likewise when $-\bar{l}$ is a unimodal function of $\mathcal{G}$ any of various search methods can be used.[10]

VIII. SUMMARY AND COMMENTS

This paper considers the general problem of determining optimum receiver gains for radar and communications receivers. Dynamic range bounds and receiver gains are determined which yield minimum average cost of excluding fluctuating signals in noise. The analysis is general enough to include minimum exclusion probability as a special case as well as a range of fluctuation models including Rician, Rayleigh, and nonfluctuating cases. The analysis is applicable to both linear and nonlinear receivers and has important implications for certain radar processors and communications terminals which can use sophisticated techniques for signal and noise parameter estimation. The results are presented in a concise normalized form making them applicable to a wide range of signal, noise, and channel conditions. It is shown that the optimum receiver gain is relatively insensitive to cost ratio for $-D \leqq \nu_{\mathrm{dB}} \leqq 2D$ differing at most by about 3 dB from the optimum gain for $\nu = 1$. The effect of noise introduced by the gain adjustment cascade is discussed.

The analysis presented assumes that certain signal, noise, and channel parameters are known to the receiver. In practice the receiver would be required to estimate these parameters. When these estimates are good, performance of the system will approach that described here. An extension of this work is to study both the optimization problem and the deterioration in performance when the parameters are not known to the receiver. Optimum dynamic range utilization for various coherent and partially coherent receivers can also be studied.

IX. ACKNOWLEDGMENT

The author is grateful to Ira Jacobs, S. O. Rice, and W. D. Rummler for commenting on the manuscript.

REFERENCES

1. Ward, H. R., Dynamic Range Centering for Minimum Probability of Excluding a Rayleigh Distributed Signal, Proc. IEEE, *54*, January, 1966, pp. 59–60.
2. Rappaport, S. S., On Optimum Dynamic Range Centering, Proc. IEEE, *54*, August, 1966, pp. 1067–68.
3. Rice, S. O., Mathematical Analysis of Random Noise, in *Selected Papers on Noise and Stochastic Processes,* ed. by N. Wax, Dover Press, New York, 1954, p. 239.
4. Helstrom, C. W., *Statistical Theory of Signal Detection,* Pergamon Press, New York, 1960, pp. 152–154.
5. Marcum, J. I., *Table of Q-Functions,* Report RM-339, Rand Corporation, Santa Monica, California, 1 January 1950.
6. Turin, G. L., Error Probabilities for Binary Symmetric Ideal Reception through Nonselective Slow Fading and Noise, Proc. IEEE, *46,* September, 1958, pp. 1603–1619.
7. Swerling, P., Probability of Detection for Fluctuating Targets, IRE Trans. Info. Theory, *IT-6,* April, 1960, pp. 269–308.
8. Watson, G. N., *A Treatise on the Theory of Bessel Functions,* Cambridge University Press, Cambridge, 1962.
9. Davenport, Jr., W. B. and Root, W. L., *An Introduction to the Theory of Random Signals and Noise,* McGraw-Hill Book Co., New York, 1958, pp. 207–213.
10. Wilde, D. J., *Optimum Seeking Methods,* Prentice-Hall, Englewood Cliffs, New Jersey, 1964, pp. 10–36.

Floating-Point-Roundoff Accumulation in Digital-Filter Realizations

By I. W. SANDBERG

(Manuscript received June 20, 1967)

In this paper, several results are presented concerning the effects of roundoff in the floating-point realization of a general discrete filter governed ideally by a stable difference equation of the form

$$w_n = \sum_{k=0}^{M} b_k x_{n-k} - \sum_{k=1}^{N} a_k w_{n-k} , \qquad n \geqq N \tag{1}$$

in which $\{w_n\}$ and $\{x_n\}$ are output and input sequences, respectively.

In particular, for a large class of filters it is proved that there is a function $f(K)$ with $f(K) \to 0$ as $K \to \infty$ and a constant c, both dependent on the b_k, the a_k, the order in which the products on the right side of (1) are summed in the machine, and t, the number of bits allotted to the mantissa, such that

$$\langle e \rangle_K \leqq c \langle y \rangle_K + f(K)$$

for all $K \geqq N$, in which, with $\{y_n\}$ the computed output sequence of the realized filter,

$$\langle y \rangle_K = \left(\frac{1}{K+1} \sum_{n=0}^{K} | \, y_n \, |^2 \right)^{\frac{1}{2}}$$

and

$$\langle e \rangle_K = \left(\frac{1}{K+1} \sum_{n=0}^{K} | \, w_n - y_n \, |^2 \right)^{\frac{1}{2}}.$$

Bounds on $f(K)$ and c are given that are not difficult to evaluate, and which, in many realistic cases, are informative. For example, for the second-order bandpass filter:

$$w_n = x_n - a_1 w_{n-1} - a_2 w_{n-2} , \qquad n \geqq 2 \tag{2}$$

with a_1 and a_2 chosen so that its poles are at approximately ± 45° and at distance approximately (but not less than) 0.001 from the unit circle, *we find that c, an upper bound on the "asymptotic output error-to-signal ratio", is not greater than 0.58 × 10^{-4}, assuming that t = 27, that the terms on the right side of (2) are summed in the machine in the order indicated (from right to left), and that the x_n in (2) are machine numbers. If the x_n are not machine numbers, and hence must be quantized before processing, then c ≦ 0.76 × 10^{-4}.*

In addition to error bounds, an inequality is derived which, if satisfied, rules out certain types of generally undesirable behavior such as self-sustained output limit cycles due to roundoff effects. This inequality is satisfied for the example described above.

I. INTRODUCTION

The difference equation

$$w_n = \sum_{k=0}^{M} b_k x_{n-k} - \sum_{k=1}^{N} a_k w_{n-k}, \qquad n \geqq N \tag{1}$$

with $M \leqq N$ defines the behavior of a general time-invariant discrete filter which acts on an input sequence $x_0, x_1, x_2, \cdots$ to produce an output sequence $w_N, w_{N+1}, w_{N+2}, \cdots$ that depends on the starting values $w_0, w_1, \cdots, w_{N-1}$.

There is a vast literature concerned with techniques for designing discrete filters [i.e., for determining the a_k and the b_k in (1)] to meet specifications of various types (see, for example, Refs. 1, 2, and 3), and a good deal of material is available on the subject of roundoff effects in fixed-point realizations of discrete filters (see, for instance, Refs. 4 and 5). In this paper, we derive some bounds on a meaningful measure of the overall effect of roundoff errors for discrete filters realized as digital filters on a machine employing floating-point arithmetic operations. This type of realization, as opposed to the fixed-point kind, is of particular importance in connection with, for example, digital computer simulations of systems, as a result of the large dynamic range afforded by the floating-point mode.

There are basic differences concerning fixed-point and floating-point error estimation problems which stem from the fact that the modulus of every individual arithmetic error in the fixed-point mode is bounded by a constant determined by the machine, whereas the maximum modulus of the error in forming, for example, the floating-point sum of two floating-point numbers is proportional to the magnitude of the true sum. For this reason, the approach* presented here, as well as the

* The approach can be extended in several different directions. For example, it can be used to obtain statistical error estimates based on the assumption that each roundoff error is an independent random variable.

character of the results, are quite different from those of earlier writers concerned with fixed-point realizations.

In addition to error bounds, an inequality is derived which, if satisfied, rules out certain types of generally undesirable behavior such as self-sustained output limit cycles due to roundoff effects.

II. ASSUMPTIONS AND RESULTS

2.1 *Assumptions*

It is assumed that:

(*i*) each machine number q is equal to sgn (q) a 2^b in which the exponent b is an integer, and a, the mantissa, is a t-bit number contained in $[\frac{1}{2}, 1]$ or $[\frac{1}{2}, 1] \cup \{0\}$;

(*ii*) the range of values of b is adequate to ensure that all computed numbers lie within the permissible range;

(*iii*) the machine operations of addition and multiplication are performed in accordance with standard rounding conventions* (described, for example, by Wilkinson[6]); and

(*iv*) the coefficients a_k and b_k in (1) are machine numbers.†

2.2 *Results: x_n Machine Numbers*

It is assumed throughout Section 2.2 that the x_n of (1) are floating-point machine numbers.

If the discrete filter (1) is realized on a floating-point machine, then

$$y_n = fl\left(\sum_{k=0}^{M} b_k x_{n-k} - \sum_{k=1}^{N} a_k y_{n-k} \right), \qquad n \geqq N \tag{2}$$

in which the y_n are approximations to the infinite precision numbers w_n, and $fl(\Sigma - \Sigma)$ denotes the machine number corresponding to $(\Sigma - \Sigma)$ with the understanding that the floating-point numbers corresponding to the products $b_k x_{n-k}$ and $a_k y_{n-k}$ are to be machine-added in some specified order.

Let

$$D(z) \triangleq 1 + \sum_{k=1}^{N} a_k z^{-k}, \tag{3}$$

* That is, conventions for which the first two equations of Section III are satisfied.

† It is certainly true that *preliminary* design considerations may lead to coefficients that are not machine numbers, and one may then be interested also in the overall effect of approximating the coefficients by machine numbers. That problem also can be treated with the approach used here.

let

$$\langle q \rangle_K \triangleq \left(\frac{1}{K+1} \sum_{k=0}^{K} | q_k |^2 \right)^{\frac{1}{2}}$$

for every sequence $\{q_k\}$ and all $K \geqq 0$, and let e_n denote the nth error $(y_n - w_n)$ for $n \geqq 0$.

Our first result (all proofs are given in Section III) is as follows. If $D(z) \neq 0$ for $| z | \geqq 1$ [i.e., if the discrete filter (1) is stable], then

$$\begin{aligned}\langle e \rangle_K \leqq & \max_{0 \leqq \omega \leqq 2\pi} | D(e^{i\omega})^{-1} | \left(\frac{1}{K+1} \sum_{n=0}^{N-1} | \eta_n |^2 \right)^{\frac{1}{2}} \\ & + 2^{-t} \left(\sum_{k=0}^{M} | b_k | \beta_k \right) \max_{0 \leqq \omega \leqq 2\pi} | D(e^{i\omega})^{-1} | \left(\frac{1}{K+1} \sum_{n=N-M}^{K} | x_n |^2 \right)^{\frac{1}{2}} \\ & + 2^{-t} \left(\sum_{k=1}^{N} | a_k | \alpha_k \right) \max_{0 \leqq \omega \leqq 2\pi} | D(e^{i\omega})^{-1} | \langle y \rangle_K \end{aligned} \tag{4}$$

for all $K \geqq N$, in which, with $y_n = w_n = 0$ for $n < 0$,

$$\eta_n = \sum_{k=0}^{N} a_k (y_{n-k} - w_{n-k}) \qquad n = 0, 1, 2, \cdots, (N-1)$$

and the α_k and β_k are easily evaluated nonnegative numbers which depend on the order in which the products in (2) are summed.

Since the first term on the right side of (4), which arises as a result of the possibility of differences in the starting values, approaches zero as $K \to \infty$, we see that, after a reasonable number of evaluations of the successive y_n, $\langle e \rangle_K$ is bounded essentially by a constant times the root-mean-squared value of the input sequence, plus another constant times the root-mean-squared value of the output sequence.

In order to determine the α_k and β_k, we draw a signal-flow graph that indicates the ordering of the operations that would be used to compute

$$fl\left(\sum_{k=0}^{M} b_k x_{n-k} - \sum_{k=1}^{N} a_k y_{n-k} \right) \tag{5}$$

if x_n and y_n were unity for all n. This graph is to contain an input node with input b_k' for each $b_k \neq 0$, an input node with input a_k' for each $a_k \neq 0$, no other input nodes, and a single output node θ which is associated with

$$\sum_{k=0}^{M} b_k' - \sum_{k=1}^{N} a_k' .$$

All other nodes represent an addition or subtraction of two signals to produce a third signal. Exactly one branch is connected to each of the input nodes and to the output node. We assign the value ρ to all of the branch transmissions with the exception of those branches, if any, which terminate on an input b_k' or a_k' for which b_k or a_k, respectively, is equal to unity. These branches are assigned unity transmission. Then, by inspection, we evaluate the signal at θ, which must clearly be of the form

$$\sum_{k=0}^{M} b_k' \rho^{\varphi_\beta(k)} + \sum_{k=1}^{N} a_k' \rho^{\varphi_\alpha(k)} \tag{6}$$

in which $\varphi_\beta(k)$ and $\varphi_\alpha(k)$ are positive-integer valued functions. In terms of these functions*

$$\beta_k = (1.06)\varphi_\beta(k)$$

$$\alpha_k = (1.06)\varphi_\alpha(k).$$

For example, if the right side of (2) is computed as the floating-point difference of the machine sums

$$fl(b_0 x_n + b_1 x_{n-1} + \cdots + b_M x_{n-M})$$

and

$$fl(a_1 y_{n-1} + a_2 y_{n-2} + \cdots + a_N y_{n-N}),$$

each obtained by performing machine summations in the order indicated (from left to right), if all of the b_k and a_k are nonzero and not unity, and if $M \geqq 1$ and $N \geqq 2$, then the relevant flow graph is shown in Fig. 1, from which it follows that

$$\beta_0 = (1.06)(M + 2)$$

$$\beta_1 = (1.06)(M + 2)$$

$$\beta_k = (1.06)(3 + M - k); \qquad k = 2, 3, \cdots, M$$

$$\alpha_1 = (1.06)(N + 1)$$

$$\alpha_2 = (1.06)(N + 1)$$

$$\alpha_k = (1.06)(3 + N - k); \qquad k = 3, 4, \cdots, N.$$

The bound (4), although revealing, requires a knowledge of both $\langle x \rangle_K$ and $\langle y \rangle_K$ and is, therefore, not as explicit as we would like.

* We are assuming here only that $\max_k |\varphi_\beta(k)| 2^{-t} < 0.1$ and $\max_k |\varphi_\alpha(k)| 2^{-t} < 0.1$. Also if $\varphi_\beta(k) = 1$, then we can take $\beta_k = 1$, and similarly for $\varphi_\alpha(k)$.

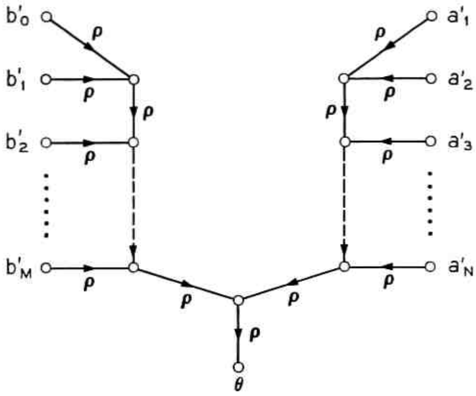


Fig. 1 — Flow graph for the example.

For the important case in which $b_0 \neq 0$ and $N(z) \triangleq \sum_{k=0}^{M} b_k z^{-k} \neq 0$ for $|z| \geqq 1$ (i.e., for the minimum-phase filter case) we prove that if the filter (1) is stable and if

$$\min_{0 \leqq \omega \leqq 2\pi} |N(e^{i\omega})| > 2^{-t} \sum_{k=0}^{M} |b_k| \beta_k \,, \tag{7}$$

then there exists a constant c, independent of K, and a function $f(K)$ with the property that $f(K) \to 0$ as $K \to \infty$ such that

$$\langle e \rangle_K \leqq c \langle y \rangle_K + f(K) \tag{8}$$

for all $K \geqq N$. Moreover, it is proved that

$$c \leqq 2^{-t} \max_{0 \leqq \omega \leqq 2\pi} |D(e^{i\omega})^{-1}| \left\{ \sum_{k=1}^{N} |a_k| \alpha_k + \sum_{k=0}^{M} |b_k| \beta_k \right.$$
$$\left. \cdot \frac{\max_{\omega} |D(e^{i\omega})/N(e^{i\omega})| + \max_{\omega} |N(e^{i\omega})^{-1}| \, 2^{-t} \sum_{k=1}^{N} |a_k| \alpha_k}{1 - 2^{-t} \sum_{k=0}^{M} |b_k| \beta_k \max_{\omega} |N(e^{i\omega})^{-1}|} \right\} \tag{9}$$

and

$$f(K) \leqq \max_{\omega} |D(e^{i\omega})^{-1}| \left(\frac{1}{K+1} \sum_{n=0}^{N-1} |\eta_n|^2 \right)^{\frac{1}{2}} + \max_{\omega} |D(e^{i\omega})^{-1}| \, 2^{-t}$$

$$\cdot\left(\sum_{k=0}^{M} | b_k | \beta_k\right) \frac{\max | N(e^{i\omega})^{-1} | \left(\frac{1}{K+1} \sum_{n=0}^{N-1} | q_n |^2\right)^{\frac{1}{2}}}{1 - 2^{-t} \sum_{k=0}^{M} | b_k | \beta_k \max_{\omega} | N(e^{i\omega})^{-1} |} \tag{10}$$

for all $K \geqq N$, in which, with $a_0 = 1$ and $x_n = y_n = 0$ for $n < 0$,

$$q_n = \sum_{k=0}^{N} a_k y_{n-k} - \sum_{k=0}^{M} b_k x_{n-k}$$

for $n = 0, 1, 2, \cdots, (N - 1)$.

Since $\langle y \rangle_K$ is the root-mean-squared value of the *computed output*, and since $f(K) \to 0$ fairly rapidly as $K \to \infty$, we may interpret the smallest value of c for which (8) is satisfied (for all input sequences) as an "output error-to-signal ratio" of the realized digital filter. Note that the bound (9) on c is not difficult to evaluate.

2.2.1. *Stability in the Presence of Roundoff*

If roundoff effects are ignored, it is well known that the discrete filter is stable in several different senses of the word if $D(z) \neq 0$ for $| z | \geqq 1$. In Section III it is proved that, with roundoff effects taken into account, the digital filter is stable in the sense that there is a constant c_1 and a function $f_1(K)$, with $f_1(K)$ independent of the values of x_n for $n \geqq N$ and $f_1(K) \to 0$ as $K \to \infty$, such that

$$\langle y \rangle_K \leqq c_1 \langle x \rangle_K + f_1(K) \tag{11}$$

for all $K \geqq N$, provided that $D(z) \neq 0$ for $| z | \geqq 1$, and

$$\min_{\omega} | D(e^{i\omega}) | > 2^{-t} \sum_{k=1}^{N} | a_k | \alpha_k . \tag{12}$$

Roughly speaking, inequality (12) is satisfied if the damping of the infinite precision counterpart of the digital filter is sufficiently large relative to the number of bits allotted to the mantissa. Stability in the sense of (11) rules out, for example, the possibility, due to roundoff effects, of a limit-cycle response to a zero input sequence or to an input sequence $\{x_n\}$ that approaches zero as $n \to \infty$.*

* There are simple examples which illustrate that instability may result with $D(z) \neq 0$ for $| z | \geqq 1$ if (12) is not satisfied. For instance, suppose that each machine number is represented in the form $(-m_0 2^0 + m_1 2^{-1} + m_2 2^{-2} + \cdots + m_t 2^{-t}) 2^b$ with the m_j zeros or ones, and $t > 1$. Let

$$w_n = (1 - 2^{-t}) w_{n-1} + (1 - 2^{-t}) 2^{-t} w_{n-2} \text{ for } n \geqq 2, \text{ with } w_0 = w_1 = -1.$$

Then $fl[(1 - 2^{-t}) w_1] = -(1 - 2^{-t})$, $fl[(1 - 2^{-t}) 2^{-t} w_0] = -(1 - 2^{-t}) 2^{-t}$, and $fl[-(1 - 2^{-t}) - (1 - 2^{-t}) 2^{-t}] = -1$, which shows that the computed approximation y_n to w_n satisfies $y_n = -1$ for *all* $n \geqq 0$. This example is a slight modification of one suggested by S. Darlington.

2.3 *A Result Concerning the Overall Effect of Input Quantization Errors*

In many applications the sequence $\{x_n\}$ of (1) is obtained by quantizing an input sequence $\{\bar{x}_n\}$ [i.e., by replacing each $\bar{x}_n$ with the machine number (or one of the possibly two machine numbers) of closest value]. The infinite precision response $\bar{w}_N$, $\bar{w}_{N+1}$, $\cdots$ to the sequence $\{\bar{x}_N\}$ satisfies

$$\bar{w}_n = \sum_{k=0}^{M} b_k\bar{x}_{n-k} - \sum_{k=1}^{N} a_k\bar{w}_{n-k}, \qquad n \geqq N \tag{13}$$

with $\bar{w}_0$, $\bar{w}_1$, $\cdots$, $\bar{w}_{N-1}$ some set of starting values. Let w_N, w_{N+1}, $\cdots$ be defined by (1) with $w_n = \bar{w}_n$ for $n = 0, 1, 2, \cdots, (N - 1)$. It is clear that $\langle y - \bar{w}\rangle_K$, the root-mean-squared value of the difference of the computed output and the infinite precision response to $\{\bar{x}_n\}$, satisfies

$$\langle y - \bar{w}\rangle_K \leqq \langle y - w\rangle_K + \langle w - \bar{w}\rangle_K . \tag{14}$$

Bounds on the first term on the right side of (14) are given in Section 2.2. In Section III it is proved that if both $N(z)$ and $D(z)$ have no zeros on or outside the unit circle, $b_0 \neq 0$, and

$$\min_{\omega} | N(e^{i\omega}) | > 2^{-t} \sum_{k=0}^{M} | b_k | \beta_k ,$$

then* there is a constant c_2 and a function $f_2(K)$ such that $f_2(K) \rightarrow 0$ as $K \rightarrow \infty$, and

$$\langle w - \bar{w}\rangle_K \leqq c_2\langle y\rangle_K + f_2(K) \tag{15}$$

for all $K \geqq N$. It is proved also that

$$c_2 \leqq 2^{-t} \sum_{k=0}^{M} | b_k | \max_{\omega} | D(e^{i\omega})^{-1} |$$

$$\cdot \frac{\max_{\omega} | D(e^{i\omega})/N(e^{i\omega}) | + \max_{\omega} | N(e^{i\omega})^{-1} | 2^{-t} \sum_{k=1}^{N} | a_k | \alpha_k}{1 - 2^{-t} \sum_{k=0}^{M} | b_k | \beta_k \max_{\omega} | N(e^{i\omega})^{-1} |} . \tag{16}$$

2.4 *A Realistic Example*

For the ideally stable second-order bandpass filter

$$w_n = x_n - a_1w_{n-1} - a_2w_{n-2}, \qquad n \geqq 2$$

[*] It is assumed here that the range of values assigned to the mantissa includes the number zero.

with poles in the z-plane at angles $\approx \pm 45°$ and at distance ≈ 0.001 (but not less than 0.001) from the unit circle, we have $a_1 \approx -1.41$, $a_2 \approx 1$, and $\min_\omega \mid D(e^{i\omega})^{-1} \mid \approx (0.00141)^{-1}$. We assume that the operations are performed as indicated in Fig. 2, so that $\beta_0 = 1$, $\alpha_1 = 3(1.06)$, and $\alpha_2 = 3(1.06)$. Assuming that $t = 27$, we find that c our bound on the "asymptotic output error-to-signal ratio," ignoring input quantization effects, is approximately 0.584×10^{-4}. For this problem, our bound on c_2 is approximately 0.18×10^{-4}. Thus, even taking into account input quantization effects, the error-to-signal ratio is not more than 0.764×10^{-4}. Finally, a simple calculation shows that this filter is stable in the presence of roundoff, in the sense of inequality (11).

III. PROOFS

3.1 *Derivation of Inequality (4)*

If a and b are floating-point machine numbers, then the floating-point product and sum $fl(ab)$ and $fl(a + b)$, respectively, satisfy[6]

$$fl(ab) = ab(1 + \epsilon)$$

$$fl(a + b) = (a + b)(1 + \delta)$$

with $\mid \epsilon \mid \leqq 2^{-t}$ and $\mid \delta \mid \leqq 2^{-t}$. Thus,

$$fl\left(\sum_{k=0}^{M} b_k x_{n-k} - \sum_{k=1}^{N} a_k y_{n-k} \right)$$

is equal to the value of the output signal θ of the flow graph described in Section II with

(*i*) $b'_k = b_k x_{n-k}$

$a'_k = a_k y_{n-k}$

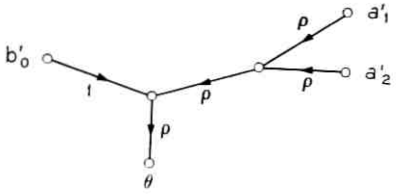


Fig. 2 — Flow graph for the second-order band-pass filter.

and

(*ii*) each of the branch transmissions of the form: $(1 + \epsilon)$ with $|\epsilon| \leqq 2^{-t}$ (recall that in certain special cases ϵ is *taken* to be zero), or $-(1 + \epsilon)$ with $|\epsilon| \leqq 2^{-t}$. Therefore,

$$fl\left(\sum_{k=0}^{M} b_k x_{n-k} - \sum_{k=1}^{N} a_k y_{n-k}\right)$$

is equal to

$$\sum_{k=0}^{M} b_k x_{n-k} q_k - \sum_{k=1}^{N} a_k y_{n-k} r_k$$

in which

$$(1 - 2^{-t})^{\varphi\beta(k)} \leqq q_k \leqq (1 + 2^{-t})^{\varphi\beta(k)} \tag{17}$$

and

$$(1 - 2^{-t})^{\varphi\alpha(k)} \leqq r_k \leqq (1 + 2^{-t})^{\varphi\alpha(k)}. \tag{18}$$

Inequalities (17) and (18) imply[6]

$$1 - (1.06)\varphi_\beta(k)2^{-t} \leqq q_k \leqq 1 + (1.06)\varphi_\beta(k)2^{-t}$$

$$1 - (1.06)\varphi_\alpha(k)2^{-t} \leqq r_k \leqq 1 + (1.06)\varphi_\alpha(k)2^{-t}$$

provided that $2^{-t} \max_k \varphi_\beta(k) < 0.1$ and $2^{-t} \max_k \varphi_\alpha(k) < 0.1$.

Thus, for $n \geqq N$

$$\begin{aligned} y_n &= fl\left(\sum_{k=0}^{M} b_k x_{n-k} - \sum_{k=1}^{N} a_k y_{n-k}\right) \\ &= \sum_{k=0}^{M} b_k x_{n-k} - \sum_{k=1}^{N} a_k y_{n-k} + \eta_n \end{aligned} \tag{19}$$

with

$$|\eta_n| \leqq 2^{-t} \sum_{k=0}^{M} |b_k| \cdot |x_{n-k}| \beta_k + 2^{-t} \sum_{k=1}^{N} |a_k| \cdot |y_{n-k}| \alpha_k \tag{20}$$

and

$$\beta_k = (1.06)\varphi_\beta(k), \qquad \alpha_k = (1.06)\varphi_\alpha(k).$$

Using (1) and (19),

$$\sum_{k=0}^{N} a_k e_{n-k} = \eta_n , \qquad n \geqq 0$$

in which, with $y_n = w_n = 0$ for $n < 0$,

$$\eta_n = \sum_{k=0}^{N} a_k(y_{n-k} - w_{n-k})$$

for $n = 0, 1, \cdots, (N - 1)$. By Propositions 1 and 2 (see Sections 3.5 and 3.6)

$$\langle e \rangle_K \leqq \max_{0 \leqq \omega \leqq 2\pi} | D(e^{i\omega})^{-1} | \langle \eta \rangle_K , \qquad K \geqq 0. \tag{21}$$

By Proposition 3 (Section 3.7), inequality (20), and Minkowski's inequality

$$\langle \eta \rangle_K \leqq \left(\frac{1}{K+1} \sum_{n=0}^{N-1} | \eta_n |^2 \right)^{\frac{1}{2}}$$
$$+ 2^{-t} \sum_{k=0}^{M} | b_k | \beta_k \left(\frac{1}{K+1} \sum_{n=-N-M}^{K} | x_n |^2 \right)^{\frac{1}{2}} + 2^{-t} \sum_{k=1}^{N} | a_k | \alpha_k \langle y \rangle_K \tag{22}$$

for all $K \geqq N$. This proves inequality (4).

3.2 *Inequality (8)*

Here we assume that both $D(z)$ and $N(z)$ are zero free for $| z | \geqq 1$, that $b_0 \neq 0$, and that

$$\min_{0 \leqq \omega \leqq 2\pi} | N(e^{i\omega}) | > 2^{-t} \sum_{k=0}^{M} | b_k | \beta_k . \tag{23}$$

From (19), we have, with $a_0 \triangleq 1$,

$$\sum_{k=0}^{N} a_k y_{n-k} = \sum_{k=0}^{M} b_k x_{n-k} + q_n , \qquad n \geqq 0, \tag{24}$$

where

$$\begin{aligned} q_n &= \eta_n , \qquad n \geqq N \\ &= \sum_{k=0}^{N} a_k y_{n-k} - \sum_{k=0}^{M} b_k x_{n-k} , \qquad n = 0, 1, 2, \cdots, (N-1) \end{aligned}$$

with $x_n = y_n = 0$ for $n < 0$. Therefore, by Propositions 1 and 2,

$$\langle x \rangle_K \leqq \max_{\omega} | D(e^{i\omega})/N(e^{i\omega}) | \langle y \rangle_K + \max_{\omega} | N(e^{i\omega})^{-1} | \langle q \rangle_K , \quad K \geqq 0. \tag{25}$$

Using Proposition 3, Minkowski's inequality, and (20),

$$\langle q \rangle_K \leqq \left(\frac{1}{K+1} \sum_{n=0}^{N-1} | q_n |^2 \right)^{\frac{1}{2}} + 2^{-t} \sum_{k=0}^{M} | b_k | \beta_k \langle x \rangle_K$$
$$+ 2^{-t} \sum_{k=1}^{N} | a_k | \alpha_k \langle y \rangle_K , \qquad K \geqq N. \tag{26}$$

Therefore,

$$\langle x\rangle_K \leqq \frac{\max\limits_{\omega} \mid D(e^{i\omega})/N(e^{i\omega}) \mid + \max\limits_{\omega} \mid N(e^{i\omega})^{-1} \mid 2^{-t} \sum\limits_{k=1}^{N} \mid a_k \mid \alpha_k}{1 - 2^{-t} \sum\limits_{k=0}^{M} \mid b_k \mid \beta_k \max\limits_{\omega} \mid N(e^{i\omega})^{-1} \mid} \langle y\rangle_K + \frac{\max\limits_{\omega} \mid N(e^{i\omega})^{-1} \mid \left(\frac{1}{K+1} \sum\limits_{n=0}^{N-1} \mid q_n \mid^2\right)^{\frac{1}{2}}}{1 - 2^{-t} \sum\limits_{k=0}^{M} \mid b_k \mid \beta_k \max\limits_{\omega} \mid N(e^{i\omega})^{-1} \mid} \tag{27}$$

for all $K \geqq N$, which together with (21) and (22) yields

$$\begin{aligned}\langle e\rangle_K \leqq 2^{-t} \max_{0 \leqq \omega \leqq 2\pi} \mid D(e^{i\omega})^{-1} \mid &\left\{ \sum_{k=1}^{N} \mid a_k \mid \alpha_k + \sum_{k=0}^{M} \mid b_k \mid \beta_k \right. \\ &\left. \cdot \frac{\max\limits_{\omega} \mid D(e^{i\omega})/N(e^{i\omega}) \mid + \max\limits_{\omega} \mid N(e^{i\omega})^{-1} \mid 2^{-t} \sum\limits_{k=1}^{N} \mid a_k \mid \alpha_k}{1 - 2^{-t} \sum\limits_{k=0}^{M} \mid b_k \mid \beta_k \max\limits_{\omega} \mid N(e^{i\omega})^{-1} \mid} \right\} \langle y\rangle_K \\ &+ \max_{\omega} \mid D(e^{i\omega})^{-1} \mid \left(\frac{1}{K+1} \sum_{n=0}^{N-1} \mid \eta_n \mid^2\right)^{\frac{1}{2}} + \max_{\omega} \mid D(e^{i\omega})^{-1} \mid 2^{-t} \\ &\cdot \sum_{k=0}^{M} \mid b_k \mid \beta_k \frac{\max\limits_{\omega} \mid N(e^{i\omega})^{-1} \mid \left(\frac{1}{K+1} \sum\limits_{n=0}^{N-1} \mid q_n \mid^2\right)^{\frac{1}{2}}}{1 - 2^{-t} \sum\limits_{k=0}^{M} \mid b_k \mid \beta_k \max\limits_{\omega} \mid N(e^{i\omega})^{-1} \mid}.\end{aligned} \tag{28}$$

This proves that there exists a constant c and a function $f(K)$ with the property that $f(K) \to 0$ as $K \to \infty$ such that (8) is satisfied for all $K \geqq N$, and of course it also proves that c and $f(K)$ are bounded as stated in Section 2.2.

3.3 *Proof of (11) Under the Conditions Stated*

From (24) and Propositions 1 and 2,

$$\langle y\rangle_K \leqq \max_{\omega} \mid N(e^{i\omega})/D(e^{i\omega}) \mid \langle x\rangle_K + \max_{\omega} \mid D(e^{i\omega})^{-1} \mid \langle q\rangle_K ,$$

and using (26)

$$\langle y\rangle_K \leqq \frac{\max_\omega \mid N(e^{i\omega})/D(e^{i\omega}) \mid + \max_\omega \mid D(e^{i\omega})^{-1} \mid 2^{-t} \sum_{k=0}^{M} \mid b_k \mid \beta_k}{1 - 2^{-t} \sum_{k=1}^{N} \mid a_k \mid \alpha_k \max_\omega \mid D(e^{i\omega})^{-1} \mid} \langle x\rangle_K$$

$$+ \frac{\max_\omega \mid D(e^{i\omega})^{-1} \mid \left(\dfrac{1}{K+1} \sum_{n=0}^{N-1} \mid q_n \mid^2\right)^{\frac{1}{2}}}{1 - 2^{-t} \sum_{k=1}^{N} \mid a_k \mid \alpha_k \max_\omega \mid D(e^{i\omega})^{-1} \mid},$$

which completes the proof.

3.4 *Derivation of Inequalities (15) and (16)*

We have, from (1) and (13),

$$\sum_{k=0}^{N} a_k(w_{n-k} - \bar{w}_{n-k}) = \xi_n\,, \qquad n \geqq 0 \tag{29}$$

in which $a_0 \triangleq 1$,

$$\xi_n = \sum_{k=0}^{M} b_k(x_{n-k} - \bar{x}_{n-k}), \qquad n \geqq N$$

and

$$\xi_n = 0, \qquad n = 0, \qquad 2, \cdots, (N-1).$$

Since $\bar{x}_n = \operatorname{sgn}(\bar{x}_n)h2^b$ for some integer b and some $h \in [\frac{1}{2}, 1]$ (assuming that $\bar{x}_n \neq 0$), the magnitude of the error in approximating $\bar{x}_n$ by the closest machine number $x_n = \operatorname{sgn}(\bar{x}_n)a2^b$ is at most $\frac{1}{2}2^{-t}2^b = \frac{1}{2}2^{-t}a^{-1} \mid x_n' \mid \leqq 2^{-t} \mid x_n \mid$. Therefore, for $n \geqq N$

$$\mid \xi_n \mid \leqq 2^{-t} \sum_{k=0}^{M} \mid b_k \mid \cdot \mid x_{n-k} \mid,$$

and by Propositions 1, 2, and 3

$$\langle w - \bar{w}\rangle_K \leqq \max_\omega \mid D(e^{i\omega})^{-1} \mid 2^{-t}$$

$$\cdot \sum_{k=0}^{M} \mid b_k \mid \left(\frac{1}{K+1} \sum_{n=N-M}^{K} \mid x_n \mid^2\right)^{\frac{1}{2}}, \qquad K \geqq N. \tag{30}$$

From (30) and (27)

$$\langle w - \bar{w}\rangle_K \leqq 2^{-t} \sum_{k=0}^{M} \mid b_k \mid \max_\omega \mid D(e^{i\omega})^{-1} \mid$$

$$\cdot \frac{\max_{\omega} \mid D(e^{i\omega})/N(e^{i\omega}) \mid + \max_{\omega} \mid N(e^{i\omega})^{-1} \mid 2^{-t} \sum_{k=1}^{N} \mid a_k \mid \alpha_k}{1 - 2^{-t} \sum_{k=0}^{M} \mid b_k \mid \beta_k \max_{\omega} \mid N(e^{i\omega})^{-1} \mid} \langle y \rangle_K$$

$$+ \frac{2^{-t} \sum_{k=0}^{M} \mid b_k \mid \max_{\omega} \mid D(e^{i\omega})^{-1} \mid \max_{\omega} \mid N(e^{i\omega})^{-1} \mid \left(\frac{1}{K+1} \sum_{n=0}^{N-1} \mid q_n \mid^2 \right)^{\frac{1}{2}}}{1 - 2^{-t} \sum_{k=0}^{M} \mid b_k \mid \beta_k \max_{\omega} \mid N(e^{i\omega})^{-1} \mid}$$

for all $K \geqq N$, provided that $N(z) \neq 0$ for $\mid z \mid \geqq 1$, $b_0 \neq 0$, and

$$\min_{\omega} \mid N(e^{i\omega}) \mid > 2^{-t} \sum_{k=0}^{M} \mid b_k \mid \beta_k .$$

This completes the derivation.

3.5 *Proposition 1:*

If

$$\sum_{l=0}^{L} c_l r_{n-l} = \sum_{l=0}^{L'} d_l s_{n-l} + f_n , \qquad n \geqq 0$$

with: $r_n = s_n = 0$ for $n < 0$, $c_0 \neq 0$, and $\sum_{l=0}^{L} c_l z^{-l} \neq 0$ for $\mid z \mid \geqq 1$, then

$$r_n = \sum_{k=0}^{n} u_{n-k} s_k + \sum_{k=0}^{n} v_{n-k} f_k , \qquad n \geqq 0$$

in which

$$\sum_{n=0}^{\infty} \mid u_n \mid < \infty, \qquad \sum_{n=0}^{\infty} \mid v_n \mid < \infty,$$

$$\sum_{n=0}^{\infty} u_n e^{-in\omega} = \sum_{l=0}^{L'} d_l e^{-il\omega} \Big/ \sum_{l=0}^{L} c_l e^{-il\omega},$$

and

$$\sum_{n=0}^{\infty} v_n e^{-in\omega} = 1 \Big/ \sum_{l=0}^{L} c_l e^{-il\omega}$$

for $0 \leqq \omega \leqq 2\pi$.

Proof:[*]

[*] The proof of this result, although rather trivial, is included because the writer knows of no reference where it is proved without the assumption that the sequences $\{s_n\}$ and $\{f_n\}$ are z-transformable.

Let $M > 0$, and let

$$\begin{aligned} \hat{s}_n &= s_n \quad \text{for} \quad n \leqq M \\ &= 0 \quad \text{for} \quad n > M \\ \hat{f}_n &= f_n \quad \text{for} \quad n \leqq M \\ &= 0 \quad \text{for} \quad n > M. \end{aligned}$$

Then $r_n = \hat{r}_n$ for $n \leqq M$, with

$$\sum_{l=0}^{L} c_l \hat{r}_{n-l} = \sum_{l=0}^{L'} d_l \hat{s}_{n-l} + \hat{f}_n , \qquad n \geqq 0$$

and with $\{\hat{r}_n\}$, $\{\hat{s}_n\}$, and $\{\hat{f}_n\}$ z-transformable. Therefore, we have

$$\hat{R}(z) = \left(\sum_{l=0}^{L'} d_l z^{-l}\right)\left(\sum_{l=0}^{L} c_l z^{-l}\right)^{-1} \hat{S}(z) + \left(\sum_{l=0}^{L} c_l z^{-l}\right)^{-1} \hat{F}(z)$$

in which

$$\begin{aligned} \hat{R}(z) &= \sum_{n=0}^{\infty} \hat{r}_n z^{-n} \\ \hat{S}(z) &= \sum_{n=0}^{M} s_n z^{-n} \\ \hat{F}(z) &= \sum_{n=0}^{M} f_n z^{-n}. \end{aligned}$$

Thus,

$$\hat{r}_n = \sum_{k=0}^{n} u_{n-k} \hat{s}_k + \sum_{k=0}^{n} v_{n-k} \hat{f}_k , \qquad n \geqq 0$$

and hence

$$r_n = \sum_{k=0}^{n} u_{n-k} s_k + \sum_{k=0}^{n} v_{n-k} f_k , \tag{31}$$

for $n = 0, 1, \cdots, M$. However, since M is arbitrary, (31) is satisfied for all $n \geqq 0$. This proves Proposition 1.

3.6 *Proposition 2:*

If

$$f_n = \sum_{l=0}^{n} c_{n-l} g_l , \qquad n \geqq 0$$

with $\sum_{l=0}^{\infty} | c_l | < \infty$, then

$$\langle f \rangle_K \leqq \max_{0 \leqq \omega \leqq 2\pi} \left| \sum_{l=0}^{\infty} c_l e^{-il\omega} \right| \langle g \rangle_K$$

for all $K \geqq 0$.

Proof:

$$\begin{aligned} \sum_{n=0}^{K} | f_n |^2 &= \frac{1}{2\pi} \int_0^{2\pi} \left| \sum_{n=0}^{K} e^{-in\omega} \sum_{l=0}^{n} c_{n-l} g_l \right|^2 d\omega \\ &= \frac{1}{2\pi} \int_0^{2\pi} \left| \sum_{n=0}^{K} e^{-in\omega} \sum_{l=0}^{n} c_{n-l} \hat{g}_l \right|^2 d\omega \end{aligned}$$

in which

$$\begin{aligned} \hat{g}_l &= g_l , & l &= 0, 1, \cdots , K \\ &= 0, & l &> K. \end{aligned}$$

Thus,

$$\begin{aligned} \sum_{n=0}^{K} | f_n |^2 &\leqq \frac{1}{2\pi} \int_0^{2\pi} \left| \sum_{n=0}^{\infty} e^{-in\omega} \sum_{l=0}^{n} c_{n-l} \hat{g}_l \right|^2 d\omega \\ &\leqq \frac{1}{2\pi} \int_0^{2\pi} \left| \sum_{l=0}^{\infty} c_l e^{-il\omega} \sum_{n=0}^{\infty} e^{-in\omega} \hat{g}_n \right|^2 d\omega \\ &\leqq \max_{\omega} \left| \sum_{l=0}^{\infty} c_l e^{-il\omega} \right|^2 \frac{1}{2\pi} \int_0^{2\pi} \left| \sum_{n=0}^{\infty} e^{-in\omega} \hat{g}_n \right|^2 d\omega \\ &\leqq \max_{\omega} \left| \sum_{l=0}^{\infty} c_l e^{-il\omega} \right|^2 \sum_{n=0}^{K} | g_n |^2, \end{aligned}$$

which proves Proposition 2.

3.7 *Proposition 3:*

If

$$| f_n | \leqq \sum_{l=0}^{L} | g_l | \cdot | h_{n-l} |, \qquad n \geqq N$$

with $L \leqq N$, then

$$\left(\sum_{n=N}^{K} | f_n |^2 \right)^{\frac{1}{2}} \leqq \left(\sum_{l=0}^{L} | g_l | \right) \left(\sum_{n=N-L}^{K} | h_n |^2 \right)^{\frac{1}{2}}$$

for all $K \geqq N$.

Proof:

$$\sum_{n=N}^{K} | f_n |^2 \leqq \sum_{n=N}^{K} \left| \sum_{l=0}^{L} | g_l |\cdot| h_{n-l} | \right|^2$$
$$\leqq \sum_{n=N}^{K} \left| \sum_{l=0}^{L} | g_l |^{\frac{1}{2}} | g_l |^{\frac{1}{2}} | \hat{h}_{n-l} | \right|^2$$

in which

$$\begin{aligned} \hat{h}_n &= h_n \qquad n = 0, 1, 2, \cdots, K \\ &= 0 \qquad n > K. \end{aligned}$$

Therefore, by the Schwarz inequality,

$$\sum_{n=N}^{K} | f_n |^2 \leqq \sum_{n=N}^{K} \sum_{l=0}^{L} | g_l | \sum_{l=0}^{L} | g_l |\cdot| \hat{h}_{n-l} |^2$$
$$\leqq \sum_{l=0}^{L} | g_l | \sum_{l=0}^{L} \left(| g_l | \sum_{n=N}^{K} | \hat{h}_{n-l} |^2 \right)$$
$$\leqq \sum_{l=0}^{L} | g_l | \sum_{l=0}^{L} \left(| g_l | \sum_{m=N-l}^{K-l} | \hat{h}_m |^2 \right)$$
$$\leqq \left(\sum_{l=0}^{L} | g_l | \right)^2 \sum_{m=N-L}^{K} | h_m |^2.$$

This completes the proof.

IV. ACKNOWLEDGMENT

The writer is indebted to his colleague J. F. Kaiser for emphasizing the need for analytical results relating to problems of the type discussed here.

REFERENCES

1. Blackman, R. B. and Tukey, J. W., *The Measurement of Power Spectra from the Point of View of Communication Engineering,* Dover Press, 1959.
2. Blackman, R. B., *Linear Data-Smoothing and Prediction in Theory and Practice,* Addison-Wesley, Reading, Massachusetts, 1965.
3. Kaiser, J. F., Digital Filters, Chapter 7 of *System Analysis by Digital Computer,* edited by F. F. Kuo and J. F. Kaiser, John Wiley & Sons, Inc., New York, 1966.
4. Bennett, W. R., Spectra of Quantized Signals, B.S.T.J., *27,* July, 1948, pp. 446–472.
5. Knowles, J. B. and Edwards, R., Effects of a Finite-Word-Length Computer in a Sampled-Data Feedback System, Proc. IEE (London), *112,* June, 1965, pp. 1197–1207.
6. Wilkinson, J. H., *Rounding Errors in Algebraic Processes,* Prentice-Hall, Englewood Cliffs, New Jersey, 1963.

An Approach to a Unified Theory of Automata

By J. E. HOPCROFT* AND J. D. ULLMAN

(Manuscript received April 7, 1967)

A model of an automaton, called a balloon automaton is proposed, It consists of a finite control, which may be deterministic or nondeterministic, an input tape which may be one way or two way, and an abstract, infinite memory, called the balloon, which can enter any of a countable number of states. There is assumed to be a recursive function which manipulates the state of the balloon, and another which passes a finite amount of information from the balloon to the finite control.

A subset of the balloon automata is considered a closed class if it obeys two very simple closure properties. Certain closed classes recognize exactly the languages recognized by such familiar automata as the pushdown automaton or stack automaton. Unfortunately, no closed class recognizes the sets accepted by linear bounded automata or the time and tape complexity classes of Turing machines.

It is shown that many of the usual closure properties of languages accepted by the pushdown automaton, stack automaton, etc., hold for an arbitrary closed class of balloon automata. For example, the languages accepted by a closed class of one-way, nondeterministic balloon automata are closed under concatenation. Of special interest is the fact that a closed class of two-way deterministic balloon automata is closed under inverse g.s.m. mappings. This fact is not obvious, and was not known for all of the types of automata which form closed classes of balloon automata.

It should be emphasized that the purpose of this paper is not to propose another "model of a computer." Rather, we are proposing a method of proving the standard theorems about existing and future models. Hopefully, when a model is proposed in the future, one will simply show it equivalent to a closed class of balloon automata, and have many of the closure properties automatically proven.

* Currently at Cornell University, Ithaca, N. Y.

I. INTRODUCTION

In the past, and especially recently, people have been examining various species of automata, perhaps as models of the compiling and translating processes, or for the insights they lend to computation. A partial list includes the Turing machine,[1] pushdown automaton,[2, 3, 4] deterministic pushdown automaton,[5] counter machine,[6, 7] stack automation, in all its forms, two-way,[8] one-way,[9, 10, 11] nonerasing,[12] deterministic and nondeterministic, the nested stack automaton,[13] and the time[14, 15] and tape[16, 17, 18] bounded Turing machines. This list is not meant to be a complete survey of past writings, and more can be expected in the future.

Many of the properties of each of the automaton classes mentioned are the same. For example, one would expect the set of languages accepted by each class to be closed under intersection with a regular set. Our plan is to propose a model of an automaton abstracting the common features of most of the models mentioned. We will define a *class* of automata to be a subset of the set of all such automata if it satisfies certain simple and physically meaningful closure properties. Then, from these closure properties, we will derive many of the common closure theorems which have been proven for the specific types of automata mentioned, and which, presumably, would be proven for future types.

The basic model is shown in Fig. 1. It consists of a two-way *input tape,* with end markers, a *finite control,* and an infinite storage of unspecified structure, called the *balloon.*

We assume that the states of the balloon are represented by the positive integers. A move of the automaton is a three-stage process. First, a recursive function is used to get a finite amount of informa-

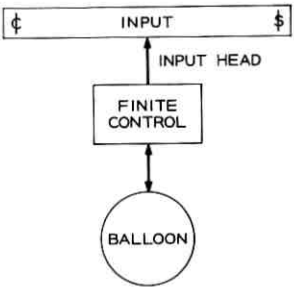


Fig. 1 — Balloon automaton.

tion from the balloon. Typically, this information is analogous to the symbol scanned by the storage head of an automaton with a tape memory. Second, based on the information from the balloon, the state of the finite control, and the symbol scanned by the input head, a new state of finite control and a direction of input head motion is determined. Third, based on the new state of finite control, and the current state of the balloon, a recursive function determines the next state of the balloon. Certain states of the finite control are *final states*. If the input causes the automaton to enter a final state, the input is *accepted*.

A subset of the set of balloon automata is called a *closed class*, or simply a *class*, if:

(*i*) It contains the finite automata.

(*ii*) If two automata are in the class, a third in the class can be found by associating in any way, the recursive functions getting information from the balloon and determining the next state of the balloon.

The latter condition is vague, but will be made formal.

Most, but not all, of the types of automata mentioned can be interpreted as classes under our definition. It seems that a type of automaton is a class if its definition involves only the ways in which the infinite storage may be locally manipulated. Sets such as the time and tape complexity classes of Turing machines do not form classes. With special emphasis, the linear-bounded automata unfortuately do not form a class in our formulation. Note that single changes in the next state of finite control function for a Turing machine may cause it to use much more time or tape than did the original machine, so condition (*ii*) would not be satisfied. Some of the automata, all two-way deterministic, which do form classes are:

(*i*) Pushdown automaton.
(*ii*) Stack automaton.
(*iii*) Nonerasing stack automaton.
(*iv*) Nested Stack automaton.
(*v*) Single counter machine.
(*vi*) Finite automaton.
(*vii*) Turing machine.

Our model shall be modified to treat nondeterministic and one-way input devices later in the paper. We have chosen two-way deterministic devices to treat first because, with one exception, the theorems involved are quite straightforward.

II. THE TWO-WAY DETERMINISTIC BALLOON AUTOMATON

A balloon automaton consists of:

(*i*) A finite, nonempty set of *states*, S.
(*ii*) A finite set of *input symbols*, I, which includes ¢ and \$, the *left* and *right end-markers* of the input, respectively.
(*iii*) A set of *balloon states*, which is always the positive integers, denoted by Z.
(*iv*) A finite, nonempty set of integers, M, known as the *balloon information.*
(*v*) A total recursive function, h, from Z to M, known as the *balloon information function.*
(*vi*) A function g, with finite domain, $S \times I \times M$ and finite range $S \times \{-1, 0, +1\}$. We will also allow φ, the null set, in the range of g. We call g the *finite control function.*
(*vii*) A partial recursive function, f, from $S \times Z$ to Z, known as the *balloon control function.*
(*viii*) A subset, F, of S, called the *final states.*
(*ix*) A state q_0 in S, the *start state.* To simplify matters later, we will here assume that the start state is not a final state. The balloon automaton is denoted $(S, I, M, f, g, h, q_0, F)$.

We denote a *configuration* of the automaton $A = (S, I, M, f, g, h, q_0, F)$ by (q, w, j, i), where:

(*i*) q is a state of the finite control, in S.
(*ii*) w is in I^*. More specifically, $w = ¢a_1a_2 \cdots a_n\$$, $n \geqq 0$, where for $1 \leqq k \leqq n$, a_k is in $I - \{¢, \$\}$. Thus, ¢ marks the left end and \$ the right end. We call n the *length* of w. Endmarkers do not contribute to the length.
(*iii*) j is an integer between 0 and $n + 1$, denoting the position of the *input head* of A.
(*iv*) i is a positive integer, the state of the balloon.

As previously mentioned, a *move* of A is a three-stage process. Let (q_1, w, j_1, i_1) be a configuration of A, and the j_1th symbol of w be a. Let w, exclusive of endmarkers, consist of n symbols. We call ¢ the 0th symbol, \$ the $n + 1$st, and number the non-endmarker symbols from 1 to n from the left. Suppose $h(i_1) = m$. Then, find $g(q_1, a, m)$.

If it is φ, no move is possible. Suppose $g(q_1, a, m) = (q_2, d)$, where q_2 is in S and $d = -1, 0$, or $+1$. Then, compute, if possible, $f(q_2, i_1)$. Let it be i_2. If $j_2 = j_1 + d$ lies in the range 0 to $n + 1$, we say that a move is possible, and the next configuration is (q_2, w, j_2, i_2).

Note that $f(q_2, i_1)$ does not necessarily have a value. In that case, there is no move possible.

Intuitively, to make a move of A, we get what information we can from the balloon by calculating $h(i_1)$. Then, using g, we find the new state of finite control and direction of motion of the input head. Finally, using f, with the new state of finite control, we find the new balloon state.

If, from configuration (q_1, w, j_1, i_1), the next configuration of A is (q_2, w, j_2, i_2), we say: $(q_1, w, j_1, i_1) \vert\overline{_A} (q_2, w, j_2, i_2)$. If A can go from configuration (q_1, w, j_1, i_1) to configuration (q_2, w, j_2, i_2) by some number of moves, including zero moves, we say: $(q_1, w, j_1, i_1) \vert\overset{*}{\overline{_A}} (q_2, w, j_2, i_2)$.

Notation: We will, for a balloon control function f and state q in S, often use $f_q(i)$ for $f(q, i)$. Also define $\alpha^{(0)}$ to be the function from Z to Z such that $\alpha^{(0)}(i) = i$ for all i. Let $\alpha^{(j)}$, for integer $j \geqq 1$, be the function that takes i to j for all i in Z.

If $A = (S, I, M, f, g, h, q_0, F)$ is a balloon automaton, let the *tapes accepted by* A, denoted $T(A)$, be the set of w such that

$$(q_0, w, 0, 1) \vert\overset{*}{\overline{_A}} (q, w, j, i)$$

for some q in F, input head position, j, and balloon state, i. That is, starting in the start state with the input head at the left endmarker and the balloon in state 1, w must cause A to enter an accepting state.

Note that if g determines, in some configuration, that A enters state p, and p is an accepting state, but for the state of the balloon, i, $f_p(i)$ is not defined, then A has no next move, hence does not accept.

Let C be a subset of the set of all balloon automata. We say C is a *closed class*, hereafter shortened to *class*, if it satisfies the following two conditions:

I. $(S, I, M, f, g, h, q_0, F)$ is in C for any finite sets, S, I, $F \subseteq S$, q_0 in S, and arbitrary mapping g from $S \times I \times M$ to $(S \times \{-1, 0, +1\}) \cup \{\varphi\}$. We restrict h to be $\alpha^{(j)}$ for some $j \geqq 1$ and $M = \{j\}$. Also, for each q in S, f_q is $\alpha^{(k)}$ for some $k \geqq 0$.

II. Let $(S_1, I_1, M_1, f_1, g_1, h_1, q_1, F_1)$ and $(S_2, I_2, M_2, f_2, g_2, h_2, q_2, F_2)$ be in C. Then $(S_3, I_3, M_3, f_3, g_3, h_3, q_3, F_3)$ is in C if;

(*i*) S_3 and I_3 are arbitrary finite sets.
(*ii*) M_3 is the range of h_3.
(*iii*) q_3 is in S_3.
(*iv*) $F_3 \subseteq S_3$.
(*v*) g_3 is an arbitrary mapping from

$$S_3 \times I_3 \times M_3 \quad \text{to} \quad (S_3 \times \{-1, 0, +1\}) \cup \{\varphi\}.$$

(*vi*) For each q in S_3, $(f_3)_q$ is $(f_1)_p$ or $(f_2)_p$ for some p in S_1 or S_2, respectively.†
(*vii*) h_3 is a total recursive function such that if $h_3(i_1) \neq h_3(i_2)$ then either $h_1(i_1) \neq h_1(i_2)$ or $h_2(i_1) \neq h_2(i_2)$.

Intuitively, assumption (*i*) causes each of the regular sets to be accepted by some automaton in the class. Note that the function h is such that no information can be obtained from the balloon.

Assumption II insures that balloon control functions can be used interchangeably. The function associated with some state may be associated with none, one, or many states of a new automaton.

The information obtainable from h_3 is no more than the information obtainable from the combination of h_1 and h_2.

If C is a class of automata, then the set of languages which can be recognized by some automaton in C is called a *closed class of languages,* or simply a *class of languages.*

It should be clear that to every class, C, there corresponds a set of allowable balloon information functions, H_C. That is, a function, h, is in H_C if and only if it is the balloon information function for some automaton, A, in C. Likewise, there is a set of functions, F_C, which is the set of allowable balloon control functions restricted to a single state. That is, f is in F_C if and only if for some automaton A, in C, with balloon control function f_1, $f(i) = f_1(q, i)$ for some fixed state q of A.

Note that $\alpha^{(i)}$ is in H_C for all $i \geqq 1$, and $\alpha^{(i)}$ for $i \geqq 0$ is in F_C, or any class C. We can use the following obvious result:

Lemma 1: Lethbein H_C *and* $f_1, f_2, \cdots, f_s$ *bein* F_C. *Let* $S = \{q_1, q_2, \cdots, q_s\}$, *$I$ be an arbitrary set of inputs including ¢ and \$, M the range of h, g an arbitrary map from $S \times I \times M$ to $(S \times \{-1, 0, +1\}) \cup \{\varphi\}$,* and *$F \subseteq S$. Then $(S, I, M, f, g, h, q_k, F)$ is in C for any q_k in S, and f defined by $f(q_i, i) = f_i(i)$ for all i.*

Proof: Let $B_0 = (S_0, I_0, M, d_0, g_0, h, p_0, F_0)$ be an automaton in C

† Recall $(f_3)_q$ is by definition the function such that $(f_3)_q(i) = f_3(q,i)$ for all i. Likewise $(f_1)_p$ and $(f_2)_p$.

with balloon information function h, and $A_i = (S_i, I_i, M_i, d_i, g_i, h_i, p_i, F_i)$ be automata in C such that for each i, $1 \leqq i \leqq s$, there is a state, r_i, in S_i, such that $d_i(r_i, j) = f_i(j)$ for all j.

For $1 \leqq i \leqq s$, define B_i from B_{i-1} and A_i according to rule II. Let $B_i = (S, I, M, e_i, g, h, q_k, F)$, where $(e_i)_{q_j} = f_j$ if $j \leqq i$, and $(e_i)_{q_j} = f$, if $j > i$. Surely, $e_s = f$, so B_s is our desired balloon automaton.

Lemma 2: Let $A = (S, I, M, f, g, h, q_0, F)$ *be an automaton in Class* C. *Let* $A_1 = (S_1, I_1, M, f_1, g_1, h, q_1, F_1)$ *be such that for every* p *in* S_1, $(f_1)_p$ *is either* $\alpha^{(i)}$ *for some* $i \geqq 0$ *or* f_q *for some* q *in* S. *Then* A_1 *is in class* C.

Proof: All f_q, for q in S are in F_C, and h is in H_C. Also, $\alpha^{(i)}$ is in F_C for all $i \geqq 0$ by rule (I). A_1 is in class C by Lemma 1.

We should comment that it is quite natural to force $\alpha^{(0)}$ to be in F_C for any class, C. Intuitively, the consequence is that an automaton may do computation in its finite control without affecting the infinite portion of storage. We also force $\alpha^{(i)}$, for $i \geqq 1$ to be in F_C. These mappings enable us to reset the infinite memory to any given state. Their use will be apparent, but their justification is not so clear. We only observe that for any of the seven types of automata mentioned, suitable modifications, which do not change the power of the devices, can be made, so that a device can reset itself to a given state.

For example, a Turing machine can surely erase its tape and print any given tape string thereon. Of course, it takes more than one move to do so, but this fact should not concern us. Even a nonerasing stack automaton can print a dummy "end of stack" marker at the top of stack to simulate an erasure of the stack.

Example: Let us indicate how to interpret a two way deterministic pushdown automaton as a class of balloon automata. We will not give a formal definition here. Most readers should be familiar with the concept of an automaton with pushdown storage, usually taken to be nondeterministic, with a one-way input. The two-way, deterministic variety is defined formally in Ref. 4.

Informally, the infinite storage is a pushdown tape, of which the automaton can at any time read only the top symbol. The pushdown tape can be altered by erasing the top symbol, or by adding a symbol to the top of the list.† The pushdown automaton has a finite control, input tape and input head, similar to these portions of a balloon automaton.

† The model of Ref. 4 allows one to add any finite number of symbols, but this mode is equivalent to adding one at a time.

We shall not formally prove that there is a closed class of balloon automata accepting exactly the sets accepted by two-way, deterministic pushdown automata. We shall merely give the sets H_C and F_C of balloon information and balloon control functions, and indicate how they reflect the pushdown structure of storage. We shall also indicate how any balloon automaton in the class can be simulated by a two-way, deterministic pushdown automaton.

To begin, we shall assign the usual Gödel numbering to pushdown tapes. That is, let the allowable pushdown symbols be $Z_1, Z_2, \cdots, Z_m$. Represent the pushdown list $Z_{i_1}Z_{i_2} \cdots Z_{i_k}$ by $2^{i_1}3^{i_2}5^{i_3} \cdots [\pi(k)]^{i_k}$. Here $\pi(i)$ stands for the ith prime. ($\pi(1) = 2$, $\pi(2) = 3$, $\pi(3) = 5$, etc.). Define $\mu(i)$, for $i \neq 1$, to be the number of the largest prime dividing i. and define $\kappa(i)$ to be the number of times $\pi(\mu(i))$ divides i. Let $\mu(1) = 0$: $\kappa(1)$ also is 0. For example, $\mu(75) = 3$, because the third prime, 5, is the largest prime dividing 75. $\kappa(75) = 2$, since 5 divides 75 twice.

Define F to be a set of recursive functions given by:

(*i*) $\alpha^{(i)}$, for all $i \geqq 0$ is in F.

(*ii*) For any integer, d, the function f, defined by $f(i) = i[\pi(\mu(i) + 1)]^d$ $i \geqq 1$, is in F. Note that $f(i)$ finds the prime above the largest prime dividing i, and multiplies i by that prime, raised to the power d.

(*iii*) The function f, given by $f(1)$ is undefined, $f(i) = i/[\pi(\mu(i))]^{\kappa(i)}$, $i > 1$, is in F. This function divides i by the largest prime dividing i, as many times as it divides i.

The set H_C includes $\alpha^{(i)}$ for $i \geqq 1$. H_C also includes any total recursive function h if there is an integer d such that $h(i) \neq h(j)$ only if $\kappa(i) \neq \kappa(j)$, and at least one of $\kappa(i)$ and $\kappa(j)$ is equal to or less than d.

Let a given pushdown automaton, P, have m pushdown symbols, $Z_1, Z_2, \cdots, Z_m$. We will find a balloon automaton, A, whose balloon information function is in H, and whose balloon control function for any given state is found in F. The balloon information function, h, will have $h(i) \neq h(j)$ if $\kappa(i) \neq \kappa(j)$ for $\kappa(i)$ and $\kappa(j)$ each $\leqq m$. According to the Gödel numbering of pushdown tapes we mentioned, $h(i)$ will always indicate the top pushdown symbol of the tape numbered i, provided tape i involves symbols $Z_1, Z_2, \cdots, Z_m$ only.

Based on the top pushdown symbol, the state of P's finite control (which is carried in the finite control of A), and the symbol scanned by A's input head, A can move its input head, and change state according to what P would do. A may then have to adjust its balloon state to simulate a change in P's pushdown store. If P does nothing to the pushdown store, the function $\alpha^{(0)}$ serves. If P erases the top symbol, the function f_i in F by rule (*iii*) must be used. If P prints Z_j on top of the pushdown list, the function f_j in F by rule (*ii*), with $d = j$ suffices.

There is a subset, C, of balloon automata defined by placing an automaton in C exactly if its balloon information function is in H and its balloon control function, restricted to any particular state, is in F. We claim that C is a closed class. Surely every balloon automaton defined by rule I of the definition is in C.

In rule II, we have two automata, A_1 and A_2, in C, and must show that a third automaton, A_3, constructed from A_1 and A_2 is also in C. Certainly, the balloon control functions of A_3 are in F. Let h_1 and h_2 be the balloon information functions of A_1 and A_2, respectively. Assume h_1 and h_2 are in H. Let h_3 be the information function of A_3. Suppose $h_3(i) \neq h_3(j)$. Then either $h_1(i) \neq h_1(j)$ or $h_2(i) \neq h_2(j)$, by rule II. In either case, $\kappa(i) \neq \kappa(j)$. Also, since h_1 and h_2 are in H, we can find an integer, d, such that one of $\kappa(i)$ and $\kappa(j)$ is $\leqq d$. Thus, h_3 is in H.

Now we must show that any balloon automaton in C can be simulated by a two-way pushdown automaton. The details of simulating the finite control and input head of the balloon automaton can be left to the reader. We shall only discuss how the balloon can be simulated.

Let $A = (S, I, M, f, g, h, q_0, F)$ be in class C. Some f_q, for q in S, may multiply the ballon state, i, by a prime raised to some power, d. Note that this prime cannot divide i. Let d_1 be the maximum such d. Some f_q may be $\alpha^{(j)}$ for $j \geqq 1$. Now, let d be the maximum number of times a prime divides j, and let d_2 be the maximum such d. Finally, let d_3 be max (d_1, d_2).

The pushdown automaton, P, simulating A, will have $d_3 + 2$ pushdown symbols, $X, Z_0, Z_1, \cdots, Z_{d_3}$. X will mark the bottom of the pushdown list. For some k, each integer, i, can be expressed in prime factors as $[\pi(1)]^{i_1}[\pi(2)]^{i_2} \cdots [\pi(k)]^{i_k}$, where each i_j, $1 \leqq j \leqq k$, lies between 0 and d_3, but $i_k \neq 0$. Then i will be represented by pushdown tape $XZ_{i_1}Z_{i_2} \cdots Z_{i_k}$. It should be clear that if $h(i) \neq h(j)$, then the tapes representing i and j have different top (rightmost) symbols.

Suppose A uses a balloon control function that is in F according to rule (*iii*). Then P erases the top pushdown symbol. P must also erase from the top, any occurrences of Z_0. Suppose A uses a balloon control function that is in F by rule (*ii*), with some particular value of d. Surely $1 \leqq d \leqq d_3$. P must print Z_d on the top of its pushdown list. Finally, if A uses balloon control function $\alpha^{(i)}$, $i \geqq 1$, P erases its tape down to X, then prints $Z_{i_1}Z_{i_2} \cdots Z_{i_k}$ on its stack, where $i = [\pi(1)]^{i_1}[\pi(2)]^{i_2} \cdots [\pi(k)]^{i_k}$. Note that by definition of d_3, we must have $i_j \leqq d_3$ for all j.

From the way F is defined, it is easy to show that for any automaton with balloon control functions chosen from F, there is some d_3, chosen as above, such that if the balloon can enter a state i, then no prime divides i more than d_3 times. Thus P, as above, with $d_3 + 2$ pushdown symbols, can simulate the balloon of A.

III. SOME THEOREMS ABOUT TWO-WAY DETERMINISTIC BALLOON AUTOMATA

We have spent time defining closed classes of automata. Our goal is not so much to talk about the classes themselves, but rather about the properties of the closed classes of languages that they define. Let us begin with a not unexpected result.

Theorem 1: Let $A = (S, I, M, f, g, h, q_0, F)$ *be a balloon automaton. Then* $L = T(A)$ *is a recursively enumerable set.*

Proof: We shall describe, informally, a Turing machine recognizing L. First, we have assumed f to be partial recursive and h total recursive. Hence, there is a Turing machine, T_0 which, given a block of i 1's on its single tape will halt with $h(i)$ 1's on its tape. Likewise, let $S = \{q_1, q_2, \cdots, q_s\}$. Then there are Turing machines $T_1, T_2, \cdots, T_s$ such that given i 1's on its tape, T_j will eventually halt with $f_{q_j}(i)$ 1's on its tape if $f_{q_j}(i)$ is defined, and not halt otherwise, for each j, $1 \leqq j \leqq s$.

We will now construct a Turing machine, T, recognizing L, by simulating A. T is shown in Fig. 2. It has a read only input tape with endmarkers, and two storage tapes. The first is used to store the state of the balloon of A.

The second is used for the computation of h and f. The finite control of T will store the state of A's finite control.

Initially, the input head of T is at the left endmarker. Its finite control records that A's finite control is in state q_0. Storage tape 1

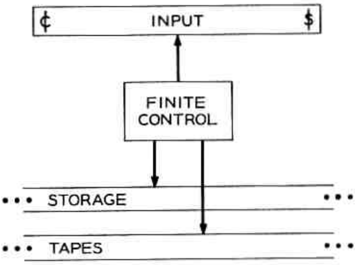


Fig. 2 — Turing machine T.

has a single 1 on it, corresponding to the initial state of A's balloon, and tape 2 is blank.

Suppose T has simulated some number of A's moves with given input. That is, T's input head is at the same position as A's would be after that number of moves. The finite control of T holds the state of A's finite control, and tape 1 holds the state of A's balloon. We will show how T simulates the next move of A, if A has a next move.

(*i*) Copy tape 1 onto tape 2.

(*ii*) Simulate T_0 on tape 2. When T_0 halts, suppose there are m 1's on tape 2 at that time.

(*iii*) Suppose T has recorded that q is the state of A's finite control. The symbol scanned by T's input head is a. Then T moves according to $g(q, a, m)$. If $g(q, a, m) = \phi$, T never completes simulation of the move of A. If $g(q, a, m) = (p, d)$, T records p as the state of A's finite control replacing q. T moves its input head in the direction indicated by d. If to do so would cause the input head to leave the input, T makes no move, but halts without accepting.

(*iv*) If T has simulated the first two stages of A's move, it again copies tape 1 onto tape 2. Let p be q_j for some j, $1 \leqq j \leqq s$. Then T simulates T_j on tape 2. If f_{q_j} is defined for the number of 1's on tape 2, T_j will eventually print on tape 2 a number of 1's equal to the new state. If not, T will not halt, hence no move of A is simulated.

(*v*) Finally, T copies tape 2 onto tape 1 and prepares to simulate another move of A. However, if the three phases of the move of A have each been successfully simulated, and p is in F, then T simulates no further moves of A, but rather, halts and accepts.

It is straightforward to see that T will simulate all moves of A, and will accept exactly when A reaches an accepting configuration.

We shall now consider three properties of closed classes of languages. These properties are that closed classes of languages are closed under reversal, intersection and inverse g.s.m. mappings. The third property is perhaps the only one in the paper that is difficult to prove.

Theorem 2: Let C be a class of automata. Let $L = T(A)$ for some $A = (S, I, M, f, g, h, q_1, F)$ in C. For any $w = ¢a_1a_2 \cdots a_n\$$ in I^*, define $w^r = ¢a_na_{n-1} \cdots a_1 \$$. Define $L^r = \{w \mid w^r \text{ is in } L\}$. Then there is an automaton, A_1, in C such that $L^r = T(A_1)$.

Proof: Let $S = \{q_1, q_2, \cdots q_s\}$. Define $S_1 = \{q_1, q_2, \cdots q_{s+1}\}$, and $A_1 = (S_1, I, M, f_1, g_1, h, q_{s+1}, F)$. We define f_1 and g_1 as follows:

(*i*) $(f_1)_q = f_q$ for q in S.

(*ii*) $(f_1)_{q_{s+1}} = \alpha^{(0)}$.

(*iii*) Suppose $g(q, a, m) = (p, d)$ for some q in S, m in M, and a in $I - \{¢, \$\}$. Then $g_1(q, a, m) = (p, \bar{d})$ where $\bar{d} = +1, 0$ or -1 as $d = -1, 0$ or $+1$, respectively.

(*iv*) Suppose $g(q, ¢, m) = (p, d)$, for q in S and m in M. Then $g_1(q, \$, m) = (p, \bar{d})$. If $g(q, \$, m) = (p, d)$, then $g_1(q, ¢, m) = (p, \bar{d})$.

(*v*) $g_1(q_{s+1}, a, m) = (q_{s+1}, +1)$ for m in M and a in $I - \{\$\}$.

(*vi*) $g_1(q_{s+1}, \$, m) = (p, \bar{d})$ for m in M, where $g(q_1, ¢, m) = (p, d)$.

(*vii*) g_1 is ϕ if not defined by (*iii*)–(*vi*).

A_1 is in class C by Lemma 2. We must show that $T(A_1) = L^r$. Let the input to A_1 be w, of length n. By rules (*ii*) and (*v*) it is seen that $(q_{s+1}, w, 0, 1) \overset{*}{\underset{A_1}{\vdash}} (q_{s+1}, w, n + 1, 1)$. From that configuration, A_1 never returns to state q_{s+1}, but simulates A with the direction of input head reversed.

That is, by rules (*i*) and (*vi*), $(q_{s+1}, w, n + 1, 1) \underset{A_1}{\vdash} (p, w, j, i)$ if and only if $(q_1, w^r, 0, 1) \underset{A}{\vdash} (p, w^r, n + 1 - j, i)$. Also, by rules (*i*), (*iii*) and (*iv*), for any q and p in S, integers i_1, i_2, j_1, j_2, with j_1 and j_2 between 0 and $n + 1$, $(q, w, j_1, i_1) \underset{A_1}{\vdash} (p, w, j_2, i_2)$ if and only if $(q, w^r, n + 1 - j_1, i_1) \underset{A}{\vdash} (p, w^r, n + 1 - j_2, i_2)$. Thus, by induction on the number of moves made by A, starting with one move,

$$(q_{s+1}, w, 0, 1) \overset{*}{\underset{A_1}{\vdash}} (p, w, j, i)$$

if and only if $(q_1, w^r, 0, 1) \overset{*}{\underset{A}{\vdash}} (p, w^r, n + 1 - j, i)$. We conclude that A_1 accepts its input, w, if and only if A accepts w^r. That is, $T(A_1) = L^r$. Note that A could not accept without making a move, since q_1 is not an accepting state.

Notation: Let h_1 and h_2 be balloon information functions, with ranges M_1 and M_2, respectively. Let M_1 have maximum element k. Define $h_1 \cdot h_2$ to be the function $[h_1 \cdot h_2](i) = h_1(i) + (k + 1)h_2(i)$. Define $M_1 \cdot M_2$ to be the range of $h_1 \cdot h_2$. We will also need the functions which are partial inverses of the $\cdot$ operator. So, we define $\sigma_1(k, j) = j$ modulo $k + 1$, and $\sigma_2(k, j) = [j/(k + 1)]$.† If k is as above, and $j = [h_1 \cdot h_2](i)$, then $\sigma_1(k, j) = h_1(i)$ and $\sigma_2(k, j) = h_2(i)$.

Note that according to the definition of closed class, if h_1 and h_2 are in H_C, then $h_1 \cdot h_2$ is in H_C for any closed class, C.

Theorem 3: If $A_1 = (S_1, I_1, M_1, f_1, g_1, h_1, q_1, F_1)$ *and* $A_2 = (S_2, I_2, M_2, f_2, g_2, h_2, q_2, F_2)$ *are automata in class* C, *then there is an automaton* A_3 *in* C *accepting* $L = T(A_1) \cap T(A_2)$.

† $[x]$ is the integer part of x.

Proof: By a simple application of Lemma 2, we can find automata accepting $T(A_1)$ and $T(A_2)$, each of whose set of input symbols is $I_1 \cup I_2$. So, we will assume that $I_1 = I_2 = I$. Likewise, from Lemma 2, we can assume S_1 and S_2 are disjoint. We construct a third automaton, $A_3 = (S_3, I, M_3, f_3, g_3, h_3, q_1, F_2)$. Here, $S_3 = S_1 \cup S_2 \cup \{q_3\}$, where q_3 is not in S_1 or S_2. Also, $M_3 = M_1 \cdot M_2$ and $h_3 = h_1 \cdot h_2$. We define f_3 and g_3 as follows:

(*i*) If q is in S_1, then $(f_3)_q = (f_1)_q$. If q is in S_2, then $(f_3)_q = (f_2)_q$.

(*ii*) $(f_3)_{q_3} = \alpha^{(1)}$.

(*iii*) Let k be the largest element in M_1, and let m be in M_3, with $m_1 = \sigma_1(k, m)$ and $m_2 = \sigma_2(k, m)$. Let a be in I. Suppose q is in S_1 but not in F_1, and $g_1(q, a, m_1) = (p, d)$. Then $g_3(q, a, m) = (p, d)$. If q is in F_1, $g_3(q, a, m) = (q_3, 0)$.

Suppose q is in S_2, instead, and $g_2(q, a, m_2) = (p, d)$. Then $g_3(q, a, m) = (p, d)$.

(*iv*) $g_3(q_3, a, m) = (q_3, -1)$, for all a in $I - \{\text{¢}\}$ and m in M_3.

(*v*) $g_3(q_3, \text{¢}, m) = (p, d)$ if $g_2(q_2, \text{¢}, m_2) = (p, d)$, where m_2 is as in (*iii*).

From rules (*i*) and (*iii*) it is clear that until A_1 enters an accepting state, A_3 enters a configuration (q, w, j, i), q in S_1, if and only if A_1 would enter that configuration. If $(q_1, w, 0, 1) \overset{*}{\underset{A_1}{\vdash}} (p, w, j, i)$, where p is in F_1, and no accepting state has been previously entered, then by rules (*i*) and (*iii*), $(q_1, w, 0, 1) \overset{*}{\underset{A_3}{\vdash}} (p, w, j, i) \underset{A_3}{\vdash} (q_3, w, j, 1)$. If w is not accepted by A_1, then A_3 will never enter state q_3.

By rules (*ii*) and (*iv*) $(q_3, w, j, 1) \overset{*}{\underset{A_3}{\vdash}} (q_3, w, 0, 1)$. By rules (*i*) and (*v*), $(q_3, w, 0, 1) \underset{A_3}{\vdash} (q, w, j, i)$ if and only if $(q_2, w, 0, 1) \underset{A_2}{\vdash} (q, w, j, i)$. From this point, A_3 simulates A_2 in a straightforward manner, entering an accepting state with w as input if and only if A_2 does. Thus, in order for A_3 to accept w, both A_1 and A_2 must accept it, and whenever these accept w, A_3 will likewise accept w. In other words, $T(A_3) = T(A_1) \cap T(A_2)$.

By part II of the definition, and Lemma 2, A_3 is in class C.

We are now going to prove a theorem on inverse g.s.m. mappings. A *generalized sequential machine* (g.s.m.) is a finite state transducer.[19] It is usually defined as a 6-tuple, $G = (K, \Sigma, \Delta, \delta, \lambda, p_0)$. K, Σ and Δ are the finite sets of *states, input symbols* and *output symbols*, respectively. δ is a mapping from $K \times \Sigma$ to K, and λ is a mapping from $K \times \Sigma$ to Δ^*. Lastly, p_0 is in K and is called the *start state*. We extend δ and λ to domain $K \times \Sigma^*$ as follows: $\delta(q, \epsilon) = q$ and $\lambda(q, \epsilon) = \epsilon$, for all q in K. For w in Σ^* and a in Σ, $\delta(q, wa) = \delta(\delta(q, w), a)$ and $\lambda(q, wa) = \lambda(q, w)\lambda(\delta(q, w), a)$. Define $G(w) = \lambda(p_0, w)$.

We can define a function γ for the g.s.m. G, as above. γ maps $K \times \Sigma^*$ to the subsets of K. If q is in K and w is in Σ^*, then

$$\gamma(q, w) = \{p \mid \delta(p, w) = q\}.$$

For w in Σ^* and a in Σ, given $\gamma(q, w)$, we can find $\gamma(q, aw)$ by: $\gamma(q, aw) = \{p \mid \text{for some } p_1 \text{ in } \gamma(q, w), \delta(p, a) = p_1\}$.

We intend to prove that if A is a balloon automaton of class C, and G is a g.s.m., then there is an automaton, A_1, in C, such that $T(A_1) = \{¢w\$ \mid \text{if } G(w) = w_1, \text{ then } ¢w_1\$ \text{ is in } T(A)\}$. We need an auxiliary definition and a lemma.

A *two-way finite automaton*[20] is a device with a two way, read only input tape and a finite control. Formally, the device is denoted $A = (K, \Sigma, \delta, p_0, F)$. K and Σ are finite sets of *states* and *input symbols,* respectively. Σ always includes ¢ and \$, the left and right endmarkers of the input, respectively. $F \subseteq K$ is the set of *final states,* and p_0, in K, is the *start state.* δ maps $K \times \Sigma$ to $K \times \{-1, +1\}$. Intuitively, if $\delta(q, a) = (p, d)$, then A, scanning a on its input, in state q, goes to state p, and moves its input head left or right, depending on whether $d = -1$ or $+1$.

We denote a *configuration* of A, with input w, by (q, w, i). We assume w can be written as $¢w_1\$$, where w_1 is in $(\Sigma - \{¢, \$\})^*$. Let w_1 consist of n symbols. The position of the input head is indicated by i. That is, $i = 0$ if the input head is scanning ¢, if $i = n + 1$, the head scans \$, and if $1 \leqq i \leqq n$, the head scans the ith symbol of w_1, counting from the left. Thus, ¢ is the zeroth symbol of w, and \$ the $n + 1$st. Of course, q is the current state of A.

Say that $(q_1, w, i_1) \mathrel{|\!\frac{}{A}} (q_2, w, i_2)$ if a is the i_1th symbol of w, $\delta(q_1, a) = (q_2, d)$ and $i_2 = i_1 + d$. However, we must have $0 \leqq i_2 \leqq n + 1$. We define the relation $\mathrel{|\!\frac{*}{A}}$ by $(q, w, i) \mathrel{|\!\frac{*}{A}} (q, w, i)$, for any configuration, (q, w, i), of A, and $(q_1, w, i_1) \mathrel{|\!\frac{*}{A}} (q_m, w, i_m)$ if there are configurations $(q_2, w, i_2), (q_3, w, i_3), \cdots, (q_{m-1}, w, i_{m-1})$ such that for $1 \leqq j < m$, $(q_j, w, i_j) \mathrel{|\!\frac{}{A}} (q_{j+1}, w, i_{j+1})$. Although we are not concerned with acceptance by two-way finite automata, (they accept the regular sets, as is well known) we will define the tapes accepted by A, denoted $T(A)$, to be $\{w \mid w \text{ in } ¢(\Sigma - \{¢, \$\})^*\$, (p_0, w, 0) \mathrel{|\!\frac{*}{A}} (p, w, i) \text{ for some } p \text{ in } F \text{ and integer, } i\}$.

Lemma 3: Let $G = (K, \Sigma, \Delta, \delta, \lambda, p_0)$ be a g.s.m., with ¢ and \$ not in Σ. Then, we can construct a two-way finite automaton,

$$A = (K_1, \Sigma \cup \{¢, \$\}, \delta_1, q_0, F)$$

with the following properties:

(i) K_1 is expressed as $K_2 \times K$. Elements of K_1 are denoted $[q, p]$ where q is in K_2, p in K.

(ii) q_1 and q_2 are particular elements of K_2.

(iii) Let $w = a_1a_2 \cdots a_n$ be in Σ^*, each a_k in Σ, $1 \leqq k \leqq n$. Suppose $\delta(p_0, a_1a_2 \cdots a_{i-1}) = p$, $i \geqq 2$. Then

$$([q_1, p], \text{¢}w\$, i) \mathrel{|\!\overset{*}{\underset{A}{-}}} ([q_2, q], \text{¢}w\$, i-1),$$

where $\delta(p_0, a_1a_2 \cdots a_{i-2}) = q$. Never is q_1 or q_2 the first component of state of A, except for the first and last configurations.

(iv) q_0 and F are irrelevant, since the lemma concerns, not the recognizing power, but the structure of two-way finite automata.

Proof: This lemma was essentially proven in Ref. 11, with direction of input head reversed. We shall, therefore, not give a formal proof, but just sketch the argument. The result in Ref. 11 did not involve the function δ of a g.s.m., but another function which had the properties needed, properties which δ has. These properties are:

(*i*) $\delta(q, w)$ is unique for q in K, w in Σ^*.

(*ii*) If γ is defined as in the definition of the g.s.m., and p_1 and p_2 are in K, $p_1 \neq p_2$, then for any w in Σ^*, $w \neq \epsilon$, $\gamma(p_1, w)$ and $\gamma(p_2, w)$ are disjoint. (For if p were in both, then $\delta(p, w) = p_1 = p_2$, violating (*i*).)

(*iii*) If p_3 is in $\gamma(p_1, w)$ and p_4 in $\gamma(p_2, w)$, and $w = w_1w_2$ with $w_2 \neq \epsilon$, then $\delta(p_3, w_1) \neq \delta(p_4, w_1)$. (For if not, let $\delta(p_3, w_1) = \delta(p_4, w_1) = p$. Then $\gamma(p_1, w_2)$ and $\gamma(p_2, w_2)$ each contain p, and $w_2 \neq \epsilon$, violating (*ii*).)

(*iv*) If p_1 and p_2 are in $\gamma(p, w)$, then $\delta(p_1, w) = \delta(p_2, w) = p$, by definition of γ.

We will now sketch the design of A. Let ¢w\$ be its input, $w = a_1a_2 \cdots a_n$, as in the statement of the lemma. Suppose the input head of A is scanning a_i, and A is in state $[q_1, p]$. Presumably,

$$\delta(p_0, a_1a_2 \cdots a_{i-1}) = p.$$

A moves its input head left, and computes $\gamma(p, a_{i-1})$. If $\gamma(p, a_{i-1})$ contains a single element, p_1, then p_1 must be $\delta(p_0, a_1a_2 \cdots a_{i-2})$. A can easily enter configuration $([q_2, p_1], w, i-1)$.

It is not possible that $\gamma(p, a_{i-1})$ is empty. Suppose $\gamma(p, a_{i-1})$ contains r elements, $r > 1$. Let these be $p_1, p_2, \cdots, p_r$. A moves left. For $j = i-2, i-3, i-4, \cdots$ it successively computes

$$\gamma(p_k, a_ja_{j+1} \cdots a_{i-2})$$

from $\gamma(p_k, a_{j+1}a_{j+2} \cdots a_{i-2})$ for $1 \leqq k \leqq r$. Unless the process terminates, in one of two ways we will describe, A then drops $\gamma(p_k, a_{j+1}a_{j+2}$

$\cdots a_{i-2})$ from memory. Given G, we can find an upper bound on r, so the amount of information stored in A's finite control is bounded.

(*i*) Suppose that for some largest j, for only one value of k, say $k = m$, is $\gamma(p_k, a_j a_{j+1} \cdots a_{i-2})$ nonempty. Then surely p_m is $\delta(p_0, a_1 a_2 \cdots a_{i-2})$. A must find its way back to position $i - 1$. Presumably, one can find k_1 and k_2 such that $\gamma(p_{k_1}, a_{j+1} a_{j+2} \cdots a_{i-2})$ and $\gamma(p_{k_2}, a_{j+1} a_{j+2} \cdots a_{i-2})$ are not empty. Choose s_1 and s_2 from these sets, respectively. A then moves right, computing $\delta(s_1, a_{j+1} a_{j+2} \cdots a_l)$ and $\delta(s_2, a_{j+1} a_{j+2} \cdots a_l)$ for $l = j + 1, j + 2, \cdots$. By comments (*iii*) and (*iv*) above, we will not have $\delta(s_1, a_{j+1} a_{j+2} \cdots a_l) = \delta(s_2, a_{j+1} a_{j+2} \cdots a_l)$ until $l = i - 1$. A is thus positioned properly, and can enter configuration

$$([q_2, p_m], w, i - 1).$$

(*ii*) Suppose that no j satisfies condition (*i*). Then A will eventually reach the left endmarker. It must be that for some m, p_0 is in $\gamma(p_m, a_1 a_2 \cdots a_{i-2})$. Thus, p_m is $\delta(p_0, a_1 a_2 \cdots a_{i-2})$. A must find its way back to position $i - 1$. So, A chooses s_1 and s_2 in $\gamma(p_{k_1}, a_1 a_2 \cdots a_{i-2})$ and $\gamma(p_{k_2}, a_1 a_2 \cdots a_{i-2})$ for some $k_1 \neq k_2$. A moves right, successively computing $\delta(s_1, a_1 a_2 \cdots a_l)$ and $\delta(s_2, a_1 a_2 \cdots a_l)$ for $l = 1, 2, \cdots$. When $\delta(s_1, a_1 a_2 \cdots a_l) = \delta(s_2, a_1 a_2 \cdots a_l)$, we must have $l = i - 1$. A easily enters configuration $[q_2, p_m], w, i - 1)$.

Theorem 4: Let $A_1 = (S_1, I_1, M, f_1, g_1, h, r_1, F_1)$ *be a balloon automaton in class C. Let* $G = (K, \Sigma, \Delta, \delta, \lambda, p_0)$ *be a g.s.m., where* $\Delta = I_1 - \{¢, \$\}$. *Then there is an automaton,* A_2 *in class C, such that*

$$T(A_2) = \{¢w\$ \mid ¢G(w)\$ \textit{ is in } T(A_1).\}.$$

$T(A_2)$ *is commonly called an inverse g.s.m. mapping of* $T(A_1)$.

Proof: Let $A = (K_1, \Sigma \cup \{¢, \$\}, \delta_1, q_0, F)$ be the two-way finite automaton constructed from G in Lemma 3. Let $A_2 = (S_2, I_2, M, f_2, g_2, h, r_2, F_2)$, where $I_2 = \Sigma \cup \{¢, \$\}$. Let $S_2 = \{[q, p, r, u, l, k] \mid q$ in K_2, p in K, r in S_1, u a string in $(I_1 - \{¢, \$\})^*$, of length at most max $(| \lambda(s, a) |$ for s in K, a in $\Sigma)$, l an integer between 0 and $| u |$, and k an integer between 1 and $8\}$.† K_2 is defined as in Lemma 3, as are its particular elements, q_1 and q_2. $r_2 = [q_2, p_0, r_1, \epsilon, 0, 1]$. F_2 is the set of all states in S_1 whose last component is 8.

We shall call the last component of states in S_2 the *pointer*. It indicates, among other things, if A_2 is simulating A, A_1 or G. The first component is part of a state of A. It is needed because A_2 may move its

† $|x|$ denotes the length of string x.

head left to simulate A_1. In that case, the routine A is needed to determine the state of G at the new position of A_2's input head. The second component of A_2's state indicates what state G would be in if it had processed whatever is to the left of A_2's input head. The third component is the state of A_1. The fourth component is the output when the input to G is the symbol currently scanned by A_2's input head. The fifth component indicates where, among the symbols of the fourth component, A_1's input head would be. In Fig. 3, the construction of A_2 is symbolically indicated.

We define f_2 by:

(*i*) $(f_2)_{[q,p,r,u,l,k]} = \alpha^{(0)}$ for $k = 3, 5, 6, 7, 8$.

(*ii*) $(f_2)_{[q,p,r,u,l,k]} = (f_1)_r$ for $k = 1, 2, 4$. For m in M, p in K, q in K_2, r in S_1 but not in F_1 and a in $I_2 - \{¢, \$\}$, we define g_2 by:

(*iii*) $g_2([q_2, p_0, r, \epsilon, 0, 1], ¢, m) = ([q_2, p_0, s, \epsilon, 0, 1], 0)$ if $g_1(r, ¢, m) = (s, 0)$. (A_2 simulates A_1, scanning and remaining at ¢ on its input.)

(*iv*) $g_2([q_2, p_0, r, \epsilon, 0, 1], ¢, m) = ([q_2, p_0, s, \epsilon, 0, \mathit{2}], +1)$ if $g_1(r, ¢, m) = (s, +1)$. (A_2 simulates A_1 moving right from ¢. The pointer is set to 2, so A_2 will next compute the output of G for the symbol it will next scan on its input.)

(*v*) $g_2([q_2, p, r, \epsilon, 0, 1], \$, m) = ([q_2, p, s, \epsilon, 0, 1], 0)$ if $g_1(r, \$, m) = (s, 0)$. ($A_2$ simulates A_1 scanning and remaining at \$.)

(*vi*) $g_2([q_2, p, r, \epsilon, 0, 1], \$, m) = ([q_1, p, s, \epsilon, 0, \mathit{4}], 0)$ if $g_1(r, \$, m) =$

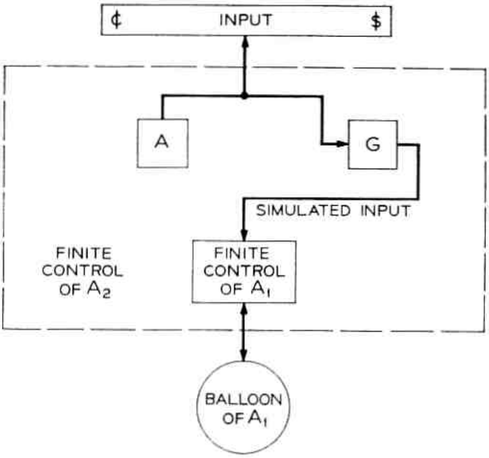


Fig. 3 — Automaton A_2.

$(s, -1)$. (A_2 simulates A_1 moving left from \$, and prepares to simulate A. The pointer is set to 4, and the first component to q_1.)

(*vii*) $g_2([q, p, r, \epsilon, 0, 4], a, m) = g_2([q, p, r, \epsilon, 0, 5], a, m) = ([q', p', r, \epsilon, 0, 5], d)$ if $\delta_1([q, p], a) = ([q', p'], d)$, for $q \neq q_2$, and a in I_2. (A_2 simulates A in A_2's first two components of state. The pointer is held at 5.)†

(*viii*) $g_2([q_2, p, r, \epsilon, 0, 5], a, m) = ([q_2, p, r, u, l, 6], 0)$ if $u = \lambda(p, a)$ and $u \neq \epsilon$. Here, $| u | = l$. (A_2 computes the output of G and prepares to simulate A_1. The pointer is set to 6.)

(*ix*) If instead, $\lambda(p, a) = \epsilon$, $g_2([q_2, p, r, \epsilon, 0, 5], a, m) = ([q_1, p, r, \epsilon, 0, 5], 0)$. ($A_2$ must simulate A again to find an input symbol that gives an output $\neq \epsilon$.)

(*x*) $g_2([q_1, p_0, r, \epsilon, 0, 4], ¢, m) = g_2([q_1, p_0, r, \epsilon, 0, 5], ¢, m)$ and is equal to $g_2([q_2, p_0, r, \epsilon, 0, 1], ¢, m)$ as defined by rules (*iii*) and (*iv*). (A_2 was prepared to begin simulating A, but found itself at the left endmarker. Note that in this case, the state of G must be p_0. A_2 immediately simulates A_1.)

(*xi*) $g_2([q_2, p, r, \epsilon, 0, 2], a, m) = g_2([q_2, p, r, \epsilon, 0, 7], a, m) = ([q_2, p, r, u, 1, 6], 0)$ if $\lambda(p, a) = u$ and $u \neq \epsilon$. (A_2 has simulated a move right of A_1's input head. It computes the output of G and prepares to simulate A_1. The pointer is set to 6, as in rule (*viii*).)

(*xii*) If instead, $\lambda(p, a) = \epsilon$, $g_2([q_2, p, r, \epsilon, 0, 2], a, m) = g_2([q_2, p, r, \epsilon, 0, 7], a, m) = ([q_2, t, r, \epsilon, 0, 7], +1)$ if $\delta(p, a) = t$. (A_2 must search right, in order to find an input symbol that does not give ϵ output when given to G.)

(*xiii*) $g_2([q_2, p, r, \epsilon, 0, 2], \$, m) = g_2([q_2, p, r, \epsilon, 0, 7], \$, m)$ and is equal to $g_2([q_2, p, r, \epsilon, 0, 1], \$, m)$ as defined by rules (*v*) and (*vi*). (A_2 was simulating a move by A_1, but encountered the right endmarker. A_2 immediately simulates another move of A_1.)

(*xiv*) Suppose $u \neq \epsilon$ and $1 \leqq l \leqq | u |$. Also, suppose $g_1(r, b, m) = (s, d)$, where b is the lth symbol of u, and $1 \leqq l + d \leqq | u |$. Then, $g_2([q_2, p, r, u, l, 1], a, m) = g_2([q_2, p, r, u, l, 6], a, m) = ([q_2, p, s, u, l + d, 1], 0)$. ($A_2$ simulates a move of A_1, where A_1 is assumed scanning the lth symbol of u.)

(*xv*) Under the assumptions of (*xiv*), if $l + d = 0$, $g_2([q_2, p, r, u, l, 1], a, m) = g_2([q_2, p, r, u, l, 6], a, m) = ([q_1, p, s, \epsilon, 0, 4], 0)$. ($A_2$ simulates A_1, but finds that A_1 moves left from u. A_2 prepares to simulate A.)

(*xvi*) Under the assumptions of (*xiv*), if $l + d > | u |$, $g_2([q_2, p, r,$

† Recall δ_1 is the next state mapping of A.

$u, l, 1], a, m) = g_2([q_2, p, r, u, l, 6], a, m) = ([q_2, t, s, \epsilon, 0, 2], +1)$, where $t = \delta(p, a)$. (A_2 simulates A_1, but finds that A_1 moves right from u. A_2 simulates the state transition of G.)

For r in F_1 and any k:

(*xvii*) $g_2([q, p, r, u, l, k], a, m) = ([q, p, r, u, l, 8], 0)$. ($A_1$ has been simulated entering an accepting state. A_2 sets the pointer to 8 and accepts.)

By rule (*xvii*) above, we see that exactly when A_2 gets to a state with third component in F_1 will it accept. It is sufficient to show that A_2 can simulate any single move of A_1 which does not start from an accepting state.

Formally, let us focus our attention on a particular input, $¢w\$$, to A_2, where w is in $(I_2 - \{¢, \$\})^*$. Let $G(w) = v$ and $|v|$ be n. For this particular w, and configuration $(r, ¢v\$, j, i)$ of A_1, we define the *inverse image* of $(r, ¢v\$, j, i)$, denoted $\text{II}(r, ¢v\$, j, i)$ as follows:

(*i*) If $j = 0$, then $([q, p_0, r, \epsilon, 0, k], ¢w\$, 0, i)$ is in $\text{II}(r, ¢v\$, 0, i)$ if either $k = 1$ and $q = q_2$ or $k = 4$ and $q = q_1$, or $k = 5$ and $q = q_1$.

(*ii*) If $j = n + 1$, then $([q_2, p, r, \epsilon, 0, k], ¢w\$, n_1 + 1, i)$ is in $\text{II}(r, ¢v\$, n + 1, i)$ if $p = \delta(p_0, w)$ and $k = 1, 2$ or 7. Here $n_1 = |w|$.

(*iii*) If $1 \leqq j \leqq n$, $([q_2, p, r, u, l, k], ¢w\$, j_1, i)$ is in $\text{II}(r, ¢v\$, j, i)$ if one can write $v = v_1uv_2$ and $w = w_1aw_2$, a in $I_2 - \{¢, \$\}$, such that the following is true:

(*a*) $\delta(p_0, w_1) = p$
(*b*) $\lambda(p_0, w_1) = v_1$
(*c*) $\lambda(p, a) = u \neq \epsilon$
(*d*) $j_1 = |w_1| + 1$
(*e*) $j = |v_1| + l$
(*f*) $k = 1$ or 6.

Intuitively, A_2's input head is scanning the symbol giving rise, when fed to G, to the symbol scanned by the input head of A_1.

We must show that if $(r, ¢v\$, j_1, i_1) \vdash_{A_1} (s, ¢v\$, j_2, i_2)$, and r is not in F_1, then if $([q, p_1, r, u_1, l_1, k_1], ¢w\$, j_3, i_1)$ is a configuration in $\text{II}(r_1, ¢v\$, j_1, i_1)$, then there is some configuration $([q', p_2, s, u_2, l_2, k_2], ¢w\$, j_4, i_2)$ in $\text{II}(s, ¢v\$, j_2, i_2)$ such that:

$$([q, p_1, r, u_1, l_1, k_1], ¢w\$, j_3, j_1) \overset{*}{\vdash}_{A_2} ([q', p_2, s, u_2, l_2, k_2], ¢w\$, j_4, i_2).$$

Case 1: $j_1 = j_2 = 0$. The result follows trivially from rules (*ii*), (*iii*) and (*x*).

Case 2: $j_1 = j_2 = n + 1$. Trivial from rules *(ii)*, *(v)* and *(xiii)*.

Case 3: $j_1 = 0$, $j_2 = 1$. By rules *(ii)*, *(iv)* and *(x)*, $([q, p_0, r, \epsilon, 0, k], ¢w\$, 0, i_1) \vdash_{A_2} ([q_2, p_0, s, \epsilon, 0, 2], ¢w\$, 1, i_2)$. By rules *(i)*, *(xi)* and *(xii)*, if $v \neq \epsilon$, $([q_2, p_0, s, \epsilon, 0, 2], ¢w\$, 1, i_2) \overset{*}{\vdash}_{A_2} ([q_2, p, s, u, 1, 6], ¢w\$, j, i_2)$, where if $w = a_1a_2 \cdots a_{n_1}$, then $\delta(p_0, a_1a_2 \cdots a_{j-1}) = p$, $\lambda(p_0, a_1a_2 \cdots a_{j-1}) = \epsilon$ and $\lambda(p, a_j) = u$. If $v = \epsilon$, by rules *(i)* and *(xii)* $([q_2, p_0, s, \epsilon, 0, 2], ¢w\$, 1, i_2) \overset{*}{\vdash}_{A_2} ([q_2, p, s, \epsilon, 0, k_1], ¢w\$, n_1 + 1, i_2)$, where $p = \delta(p_0, w)$ and $k_1 = 2$ or 7.

Case 4: $j_1 = n + 1$, $j_2 = n$. By rules *(ii)*, *(vi)* and *(xiii)*, $([q_2, p_1, r, \epsilon, 0, k], ¢w,\$ n_1 + 1, i_1) \vdash_{A_2} ([q_1, p_1, s, \epsilon, 0, 4], ¢w\$, n_1 + 1, i_2)$. If $v \neq \epsilon$, by Lemma 3, and rules *(i)*, *(vii)*, *(viii)* and *(ix)*, $([q_1, p_1, s, \epsilon, 0, 4], ¢w\$, n_1, i_2) \overset{*}{\vdash}_{A_2} ([q_2, p_2, s, u, l, 6], ¢w\$ j, i_2)$, where if $w = a_1a_2 \cdots a_{n_1}$, $\delta(p_0a_1a_2 \cdots a_{j-1}) = p_2$, $\lambda(p_0, a_1a_2 \cdots a_{j-1}) = v_1$, $v_1u = v$, and $\lambda(p_2, a_j) = u$. If $v = \epsilon$, by Lemma 3 and rules *(i)*, *(vii)* and *(ix)*, $([q_1, p_1, s, \epsilon, 0, 4], ¢w\$, n_1, i_2) \overset{*}{\vdash}_{A_2} ([q_1, p_0, s, \epsilon, 0, k], ¢w\$, 0, i_2)$, where $k = 4$ or 5.

Case 5: j_1 is not 0 or $n + 1$. Also, $l + j_2 - j_1$ lies between 1 and $| u |$, where l and u are defined in part *(iii)* of the definition of inverse image. The result is immediate from rules *(ii)* and *(xiv)*.

Case 6: j_1 is not 0 or $n + 1$, but $l = | u |$ and $j_2 = j_1 + 1$. By rules *(ii)* and *(xvi)*, $([q_2, p_1, r, u, l, k], ¢w\$, j_2, i_1) \vdash_{A_2} ([q_2, p_2, s, \epsilon, 0, 2], ¢w\$, j_3 + 1, i_2)$, where $([q_2, p_1, r, u, l, k], ¢w\$, j_3, i_1)$ is either of the inverse images of $(r, ¢v\$, j_1, i_1)$. The rest of the argument for this case is similar to that of case 3, and will be left to the reader.

Case 7: j_1 is not 0 or $n + 1$, but $l = 1$ and $j_2 = j_1 - 1$. By rules *(ii)* and *(xv)*, $([q_2, p, r, u, l, k], ¢w\$, j_3, i_1) \vdash_{A_2} ([q_2, p, s, \epsilon, 0, 4], ¢w\$, j_3, i_2)$, where the former configuration is again either of the inverse images of $(r, ¢v\$, j_1, i_1)$. The argument proceeds as in case 4.

We claim, from the above, that $([q_2, p_0, r_1, \epsilon, 0, 1], ¢w\$, 0, 1) \overset{*}{\vdash}_{A_2} ([q, p, r, u, l, k], ¢w\$, j_1, i) \vdash_{A_2} ([q, p, r, u, l, 8], ¢w\$, j_1, i)$, for some r in F_1, $k \neq 8$, if and only if $(r_1, ¢v\$, 0, 1) \overset{*}{\vdash}_{A_1} (r, ¢v\$, j, 1)$ by a sequence of moves for which A_1 never previously enters an accepting state. Here u, l, j and j_1 are related as in part *(iii)* of the definition of inverse image. Thus, $T(A_2) = \{¢w\$ \mid \text{for some } v \text{ with } ¢v\$ \text{ in } T(A_1), G(w) = v\}$. We must add that by Lemma 2, A_2 is in class C. The theorem is thus proven.

IV. OTHER TYPES OF BALLOON AUTOMATA

We have considered the two-way deterministic balloon automaton. To complete the story we should consider three other models—non-deterministic two-way balloon automata, and one-way balloon automata of the deterministic and non-deterministic varieties.

A nondeterministic device typically has the choice of a finite number of possibilities for each move. We choose to make the finite control function nondeterministic. This added capability enables us to represent the nondeterministic versions of the seven types of automata which we could represent by a deterministic balloon automaton.

A one-way balloon automaton is, quite naturally, a two-way balloon automaton, restricted so that the input head can only move right or not move at all.

We shall not repeat the definitions for each of the three new types of balloon automata, but, as a model, shall make use of the definition of two-way deterministic balloon automata.

A *two-way, nondeterministic balloon automaton* is denoted $A = (S, I, M, f, g, h, q_0, F)$ where all components are defined exactly as for the deterministic case, except that g is a mapping from $S \times I \times M$ to the subsets of $S \times \{-1, 0 + 1\}$.

A *one-way, deterministic balloon automaton* is denoted as are the two-way types, but g is a mapping from $S \times I \times M$ to $(S \times \{0, +1\}) \cup \{\varphi\}$.

A *one-way nondeterministic balloon automaton* is denoted as are the two-way types, but g is a mapping from $S \times I \times M$ to the subsets of $S \times \{0, +1\}$.

The closed classes of one way nondeterministic balloon automata are similar to the abstract families of acceptors in Ref. 21.

We shall use the abbreviations 2DBA, 2NBA, 1DBA, and 1NBA for, respectively, two-way deterministic, two-way nondeterministic, one-way deterministic and one-way nondeterministic balloon automata.

A *configuration* of any of the four types is denoted as for the 2DBA, (q, w, j, i), where, q is the state of finite control, w the input, j the input head position, and i the state of the balloon.

The possible moves of the 2NBA are determined as one would expect. One uses the balloon information function. Based on the value of that function, the input symbol at the position of the input head, and the state of finite control, one chooses a pair of next state of finite control and direction of input head, according to g. Then, based on the new state, the balloon control function is used.

Formally, if $A = (S, I, M, f, g, h, q_0, F)$ is a 2NBA, and (q_1, w, j_1, i_1)

and (q_2, w, j_2, i_2) are configurations of A, with n the length of w, then we say (q_1, w, j_1, i_1) goes to (q_2, w, j_2, i_2) by a single *move*, denoted $(q_1, w, j_1, i_1) \underset{A}{\vdash} (q_2, w, j_2, i_2)$ exactly when for some m in M, a in I, $d = -1$, 0 or $+1$, we have $h(i_1) = m$, the j_1th position of w is a, $g(q_1, a, m)$ contains (q_2, d) and $f_{q_2}(i_1) = i_2$. Also, $j_1 + d$ is between 0 and $n + 1$ and $j_2 = j_1 + d$. If (q_1, w, j_1, i_1) can go to configuration (q_2, w, j_2, i_2) by some number of moves, including 0, then we say $(q_1, w, j_1, i_1) \overset{*}{\underset{A}{\vdash}} (q_2, w, j_2, i_2)$.

The notion of move, and the relations $\vdash$ and $\overset{*}{\vdash}$ are defined for the 1DBA and 1NBA exactly as for the 2DBA and 2NBA, respectively.

A 2NBA *accepts* an input, w if for some choice of moves it enters an accepting state. Formally, define $T(A)$, for a 2NBA, $A = (S, I, M, f, g, h, q_0, F)$ to be $\{w \mid (q_0, w, 0, 1) \underset{A}{\vdash} (q, w, j, i)$ for some q in $F\}$.

For the one-way types, we require that the input head reach the right endmarker when it accepts. That is, if $A = (S, I, M, f, g, h, q_0, F)$ is a 1NBA or 1DBA, then $T(A) = \{w \mid (q_0, w, 0, 1) \overset{*}{\underset{A}{\vdash}} (q, w, n + 1, i)$ for some q in F, where n is the length of $w\}$.

The notions of *closed class* of balloon automata for the 2NBA, 1DBA and 1NBA are defined exactly as for the 2DBA.

Note that, for example, a 1DBA is not a 1NBA, although there are obvious relationships. Also, strictly speaking, a closed class of 1DBA is not a closed class of 1NBA. Both parts I and II of the definition for 1NBA would require nondeterministic finite control functions in any class of 1NBA. Analogous statements hold between 2DBA and 2NBA, 1DBA and 2DBA, 1NBA and 2NBA.

It is trivial to see that Lemmas 1 and 2 hold for the 2NBA, 1DBA and 1NBA.

A set of languages is said to be a *closed class* (or simply *class*) for the 2NBA, 2DBA, 1NBA, or 1DBA if they are exactly the languages accepted by a closed class of automata of that type.

V. TWO-WAY NONDETERMINISTIC BALLOON AUTOMATA

Theorems 2, 3 and 4, proven for the 2DBA also hold for the 2NBA. In each case, the simulation by an automaton in some class, C, of one or two other automata in C was involved. In the 2NBA case, the simulation can be nondeterministic if the simulated automata are. We will therefore omit the proofs of the three theorems for the nondeterministic case.

Likewise, Theorem 1 holds for the 2NBA. We can simulate a 2NBA by a nondeterministic Turing machine just as we simulated the 2DBA by a deterministic Turing machine. A nondeterministic Turing machine, as is well known, can be simulated by a deterministic Turing machine.

There is one additional, simple theorem we can prove for the 2NBA but not the 2DBA.

Theorem 5: If L_1 and L_2 are languages accepted by automata A_1 and A_2, respectively, in class C of 2NBA, *then there is an automaton, A_3, in C accepting $L_1 \cup L_2$.*

Proof: Let $A_1 = (S_1, I_1, M_1, f_1, g_1, h_1, q_1, F_1)$ and $A_2 = (S_2, I_2, M_2, f_2, g_2, h_2, q_2, F_2)$. As was mentioned, by Lemma 2 we can assume that $S_1 \cap S_2 = \varphi$ and $I_1 = I_2 = I$. Consider a new automaton, $A_3 = (S_3, I, M_3, f_3, g_3, h_3, q_3, F_3)$. $S_3 = S_1 \cup S_2 \cup \{q_3\}$, where q_3 is not in $S_1 \cup S_2$. $F_3 = F_1 \cup F_2$. $M_3 = M_1 \cdot M_2$ and $h_3 = h_1 \cdot h_2$.† Define f_3 by $(f_3)_{q_3} = \alpha^{(0)}$, $(f_3)_q = (f_1)_q$ if q is in S_1, and $(f_3)_q = (f_2)_q$ if q is in S_2.

Let the largest element of M_1 be k. We define g_3 as follows. For a in I and m in M_3, let $m_1 = \sigma_1(k, m)$ and $m_2 = \sigma_2(k, m)$. If q is in S_1, then $g_3(q, a, m) = g_1(q, a, m_1)$. If q is in S_2, $g_3(q, a, m) = g_2(q, a, m_2)$. Finally, $g_3(q_3, a, m) = g_1(q_1, a, m_1) \cup g_2(q_2, a, m_2)$.

It is straightforward to see that A_3 is in class C.

It should be clear that for any input, w, $(q_3, w, 0, 1) \vdash_{A_3} (q, w, j, i)$ exactly when either q is in S_1 and $(q_1, w, 0, 1) \vdash_{A_1} (q, w, j, i)$ or q is in S_2 and $(q_2, w, 0, 1) \vdash_{A_2} (q, w, j, i)$. Also, once in a state of S_1, A_3 remains in a state of S_1 and simulates A_1. Likewise, in a state of S_2, A_3 simulates A_2. Thus, by induction on the number of moves made, starting with one move, we have $(q_3, w, 0, 1) \overset{*}{\vdash}_{A_3} (q, w, j, i)$ if and only if $(q_1, w, 0, 1) \overset{*}{\vdash}_{A_1} (q, w, j, i)$ or $(q_2, w, 0, 1) \overset{*}{\vdash}_{A_2} (q, w, j, i)$. Thus, since $F_3 = F_1 \cup F_2$, and neither q_1 or q_2 may be in F_3, it follows that $T(A_3) = T(A_1) \cup T(A_2)$.

VI. ONE-WAY DETERMINISTIC BALLOON AUTOMATA

The 1DBA is the poorest of the four types in terms of the operations on languages which preserve membership in a closed class of languages for given types. Of the operations preserving membership in class for the two-way devices, only inverse g.s.m. mappings preserve membership in class for the 1DBA. The proof is along the lines of that of Theorem 4, but is simpler because the input head never has to move left. We will omit the proof.

There is one new operation which does preserve classes for the 1DBA, and, incidently, the 1NBA. This operation is intersection with

† Recall the definition of the operation "·", σ_1 and σ_2 in Section III.

a regular set. Classes for the 2NBA and 2DBA were closed under intersection of languages in the class. A simple use of part I of the definition of closed class shows that every regular set is in every closed class, so intersection with a regular set surely preserves membership in class for the 2NBA and 2DBA.

We shall give the usual formal definition of a finite automaton. See Ref. 20, for example. A *finite automaton* is a 5-tuple, $A = (K, \Sigma, \delta, q_0, F)$. K is the finite set of *states*, Σ the finite set of *input symbols*. F is a subset of K, the *final states*, and q_0, in K is the *start state*. δ is a map from $K \times \Sigma$ to K. We extend δ to domain $K \times \Sigma^*$ by $\delta(q, \epsilon) = q$ for all q in K, and $\delta(q, wa)$, for q in K, w in Σ^* and a in Σ is $\delta(\delta(q, w), a)$. Define $T(A) = \{w \mid \delta(q_0, w) \text{ is in } F\}$. The finite automata accept exactly the regular sets.

Theorem 6: Let C be a class of one-way, deterministic balloon automata. Let L be accepted by some automaton, A in C, and let R be a regular set. Then $L \cap R$ is accepted by some automaton in class C.

Proof: Let $A = (S, I, M, f, g, h, q_0, F)$ be a 1DBA. Let $R_1 = R \cap ¢(I - \{¢, \$\})^*\$$, and let $R_2 = \{w \mid w\$ \text{ is in } R_1\}$. If R is regular, then R_1 and R_2 are both regular. It is sufficient to show that there is an automaton in C accepting $L \cap R_1$. To that end, let $A_1 = (K, I, \delta, p_0, F_1)$ be a finite automaton with $T(A_1) = R_2$. Define $A_2 = (S_2, I, M, f_2, g_2, h, q_2, F_2)$ to be a 1DBA, with $S_2 = S \times K$, $q_2 = [q_0, p_0]$ and $F_2 = F \times F_1$. Define f_2 and g_2 as follows, for all q and q_1 in S, p and p_1 in K, a in I and m in M:

(*i*) Suppose $g(q, a, m) = (q_1, 0)$. Then for all p in K, $g_2([q, p], a, m) = ([q_1, p], 0)$.

(*ii*) Suppose $g(q, a, m) = (q_1, +1)$ and $\delta(p, a) = p_1$. Then $g_2([q, p], a, m) = ([q_1, p_1], +1)$.

(*iii*) $(f_2)_{[q,p]} = f_q$ for all p in K.

The states of A_2's finite control have two components. The first is a state of A and the second a state of A_1. By rules (*i*) and (*iii*), when the input head of A does not move, A_2 simulates a move of A, but does not change the state of A_1. By rules (*ii*) and (*iii*), when the input head of A moves right, A_2 simulates that move also, but adjusts the state of A_1 in the logical manner.

Formally, we can show by induction on the number of moves of A or A_2, starting with 0 moves, that $([q_0, p_0], w, 0, 1) \overset{*}{\underset{A_2}{\vdash}} ([q, p], w, j, i)$ if and only if:

(*i*) $(q_0, w, 0, 1) \overset{*}{\underset{A}{\vdash}} (q, w, j, i)$, and

(*ii*) $\delta(p_0, w_1) = p$, where w_1 is that portion of w to the left of position j.

Now a word, w, of length n, is accepted by A_2 if and only if $([q_0, p_0], w, 0, 1) \overset{*}{\underset{A_2}{\vdash}} ([q, p], w, n + 1, i)$, for some q in F, p in F_1 and any integer, i. The above is equivalent to saying that $(q_0, w, 0, 1) \overset{*}{\underset{A}{\vdash}} (q, w, n + 1, i)$ and $\delta(p_0, w_1) = p$, where $w_1\$ = w$. That is, w is in $T(A)$ and w_1 is in $T(A_1)$. But w_1 is in $R_2 = T(A_1)$ if and only if w is in R_1. Thus, $T(A_2) = L \cap R_1$. It should be clear, by Lemma 2, that A_2 is in class C.

Corollary 1: If L is a language in class C for the 1DBA, *and R is a regular set, then L − R is in class C.*

Proof: Let L be contained in I^* for some finite alphabet, I. Then $L - R = L \cap (I^* - R)$, which is in class C by Theorem 6.

Theorem 6 applies also to the 1NBA. In fact, there is an additional corollary that can be shown for the 1NBA.

Corollary 2: Let L be in class C of 1NBA, *and let R be a regular set not involving symbols ¢ or \$. Then $L \cup ¢R\$$ is in class C.*

Proof: The results is a simple extension of Theorem 6, and will be left to the reader.

VII. ONE-WAY NONDETERMINISTIC BALLOON AUTOMATA

As the 1DBA was the poorest of the four models, in terms of provable properties, the 1NBA is the richest. Theorem 4, concerning inverse g.s.m. mappings, certainly holds for the 1NBA, as do Theorem 5, Theorem 6 and its corollaries.

To begin a study of the 1NBA, we will show that with the proper definition of acceptance, endmarkers on the input are not necessary. Let $A = (S, I, M, f, g, h, q_0, F)$ be a 1NBA. We informally define $\hat{T}(A)$ as the set of strings, w, in $(I - \{¢, \$\})^*$ which cause A to leave w moving right, at the same time entering an accepting state.

We need a slightly revised notion of a configuration. Since w has no endmarkers, its length is the number of symbols comprising w. (Recall, we never counted endmarkers in determining length.) Let w be of length n. Then (q, w, j, i) is a configuration of A if q is in S, i is an integer and $1 \leq j \leq n$. The initial configuration for a 1NBA without endmarkers is $(q_0, w, 1, 1)$. For convenience, we define a configuration, (*), which is imagined to result when A is in a configuration (q, w, n, i), and the finite control function allows A, on the next move, to move its input head right and enter an accepting state.[†] There is no change in the definitions of $\vdash_{A}$ and $\overset{*}{\vdash}_{A}$. The former relates two configurations if the second is obtainable from the first by a single move, and the latter – if by some finite number of moves. Note that no configuration can result from (*).

Now, we define $\hat{T}(A)$ as $\{w \mid (q_1, w, 1, 1) \overset{*}{\vdash}_{A} (*)\}$. When talking of a 1NBA and the $\hat{T}$ definition of acceptance, we will allow the start state to be an accepting state. If so, we shall, by convention, say that ϵ is in $\hat{T}(A)$. We will endeavor to show that a language is $T(A_1)$ for some 1NBA, A_1, if and only if it is $\hat{T}(A_2)$ for A_2, a 1NBA in the same classes as A_1. The result is broken into two parts.

Theorem 7: Let $A_1 = (S_1, I, M, f_1, g_1, h, q_1, F_1)$ *be a* 1NBA *with* $L = \hat{T}(A_1)$. *Then there is another* 1NBA, $A_2 = (S_2, I, M, f_2, g_2, h, q_2, F_2)$, *such that* $¢L\$ = T(A_2)$.[‡] *Moreover, if* A_1 *is in some closed class,* C, *then* A_2 *is in* C.

Proof: Choose q_2 to be a symbol not in S_1, and let $S_2 = S_1 \cup \{q_2\}$. $F_2 = F_1$ if q_1 is not in F_1; $F_2 = F_1 \cup \{q_2\}$ otherwise. Define f_2 and g_2 as follows:

(*i*) $(f_2)_{q_2} = \alpha^{(0)}$
(*ii*) $(f_2)_q = (f_1)_q$ for q in S_1.

For all a in $I - \{¢, \$\}$ and all m in M:

(*iii*) $g_2(q, a, m) = g_1(q, a, m)$, for q in S_1.
(*iv*) $g_2(q_2, a, m) = g_1(q_1, a, m)$.
(*v*) $g_2(q, \$, m) = \varphi$, for q in S_2.
(*vi*) $g_2(q_2, ¢, m) = \{(q_2, +1)\}$.
(*vii*) $g_2(q, ¢, m) = \varphi$ for q in S_1.

Let w, of length $n \geqq 1$ be in $(I - \{¢, \$\})^*$. By rules (*i*) and (*vi*), we have $(q_2, ¢w\$, 0, 1) \vdash_{A_2} (q_2, ¢w\$, 1, 1)$. By rules (*ii*) and (*iv*), it follows that $(q_2, ¢w\$, 1, 1) \vdash_{A_2} (q, ¢w\$, j, i)$ if and only if $(q_1, w, 1, 1) \vdash_{A_1} (q, w, j, i)$. Then, by induction on the number of moves made, starting with one move, we see that $(q_2, ¢w\$, 1, 1) \overset{*}{\vdash}_{A_2} (q, ¢w\$, j, i)$ if and only if $(q_1, w, 1, 1) \overset{*}{\vdash}_{A_1} (q, w, j, i)$, for $j \leqq n$. Finally, by rules (*ii*) and (*iii*), if $(q, w, n, i) \vdash_{A} (*)$, then it must be that $g_1(q, a, m)$ con-

[†] Strictly speaking we require also that the balloon control function be defined for i and the new state if configuration (*) is to be entered.

[‡] Recall that by convention, ¢ and \$ are in every set of input symbols, even if endmarkers are not used.

tains $(p, +1)$ for some p in F_1, where $h(i) = \mathrm{m}$ and a is the nth symbol of w. Also, $f_p(i)$ is defined, so $(q, ¢w\$, n, i) \vdash_{A_2} (p, ¢w\$, n + 1, i_1)$, where $i_1 = f_p(i)$. Thus, if w is in $\hat{T}(A_1)$, then $¢w\$$ is in $T(A_2)$.

If $(q_1, ¢w\$, 0, 1) \overset{*}{\vdash}_{A_2} (p, ¢w\$, n + 1, i)$, where p is in F_1, from rule (v) we see that A_2 could not have made a move while scanning the \$. Thus, for some q in S and integer, k, $(q_1, ¢w\$, 0, 1) \overset{*}{\vdash}_{A_2} (q, ¢w\$, n, k) \vdash_{A_2} (p, ¢w\$, n + 1, i)$. From the previous paragraph, we know that $(q_1, w, 1, 1) \overset{*}{\vdash}_{A_1} (q, w, n, k)$ and $(q, w, n, k) \vdash_{A_1} (^*)$. Thus, if $¢w\$$ is in $T(A_2)$, then w is in $\hat{T}(A_1)$.

One detail remains, concerning the case $w = \epsilon$. If ϵ is in $\hat{T}(A_1)$, then q_1 is in F_1. Thus, q_2 is in F_2. By rule (vi), $(q_2, ¢\$, 0, 1) \vdash_{A_2} (q_2, ¢\$, 1, 1)$, so $¢\$$ is in $T(A_2)$. If ϵ is not in $\hat{T}(A)_1$, then q_1 is not in F_1 and q_2 is not in F_2. By rules (v) and (vi), only one move of A_2 is possible, and A_2 does not accept $¢\$$. We conclude that $T(A_2) = ¢L\$$. It is clear from Lemma 2 that A_2 is in class C.

Theorem 8: Let $A_1 = (S_1, I, M, f_1, g_1, h, q_1, F_1)$ *be a* 1NBA, *in some closed class, C, with* $L_1 = T(A_1)$. *Let* $L_2 = \{w \mid ¢w\$ \text{ is in } L_1\}$. *Then there is a* 1NBA, $A_2 = (S_2, I, M, f_2, g_2, h, [q_1, 1], F_2)$ *in class C, with* $\hat{T}(A_2) = L_2$.

Proof: We will place in S_2 all symbols of the form $[q, i]$, where q is in S_1 and $i = 1, 2, 3$, or 4. If $¢\$$ is in L_1, then $F_2 = \{[q_1, 1]\} \cup \{[q, 4] \mid q$ in $S_1\}$. If $¢\$$ is not in L_1, $F_2 = \{[q, 4] \mid q \text{ in } S_1\}$.† Define f_2 and g_2 as follows:

(i) $(f_2)_{[q,i]} = (f_1)_q$ for all q in S_1 and $i = 1, 2, 3, 4$.

For all m in M, q in S_1 and a in $I - \{¢, \$\}$:

(ii) $g_2([q, 1], a, m) = \{([p, 1], 0) \mid (p, 0) \text{ is in } g_1(q, ¢, m)\} \cup \{([p, 2], 0) \mid (p, +1) \text{ is in } g_1(q, ¢, m)\}$.

(iii) $g_2([q, 2], a, m) = \{([p, 2], d) \mid (p, d) \text{ is in } g_1(q, a, m), d = 0 \text{ or } +1\} \cup \{([p, 3], 0) \mid (p, +1) \text{ is in } g_1(q, a, m)\} \cup \{([p, 4], +1) \mid (p, +1) \text{ is in } g_1(q, a, m) \text{ and } p \text{ is in } F_1\}$.

(iv) $g_2([q, 3], a, m) = \{([p, 3], 0) \mid (p, 0) \text{ is in } g_1(q, \$, m) \text{ and } p \text{ is not in } F_1\} \cup \{([p, 4], +1) \mid (p, 0) \text{ is in } g_1(q, \$, m) \text{ and } p \text{ is in } F_1\}$.

(v) $g_2([q, 4,], a, m) = \varphi$ for any a in I, including ¢ and \$.

Intuitively, when the second component of state of A_2 is 1, A_2 imagines it is reading ¢ on its input. If the second component is 3, it imagines

† Obviously, it may not be possible to tell whether $¢\$$ is in L_1. In that case, the procedure given here can be thought of as defining two automata, one of which accepts L_2. Computation of A_2 from A_1 is not effective, but this fact will not alter our theoretical results.

it is reading \$. If the second component is 2, it uses the symbol actually scanned. A second component of 4 indicates an accepting state.

Formally, let w be in $(I - \{¢, \$\})^*$, of length $n \geqq 1$. From rules (i) and (ii), we see that $([q_1, 1], w, 1, 1) \vdash_{A_2}^{*} ([p, 1], w, 1, i)$ if and only if $(q_1, ¢w\$, 0, 1) \vdash_{A_1}^{*} (p, ¢w\$, 0, i)$. Also, $([q, 1], w, 1, i_1) \vdash_{A_2} ([p, 2], w, 1, i_2)$ if and only if $(q, ¢w\$, 0, i_1) \vdash_{A_1} (p, ¢w\$, 1, i_2)$.

Next, by rules (i) and (iii), $([p, 2], w, 1, i_1) \vdash_{A_2}^{*} ([p, 2], w, n, i_2)$ if and only if $(q, ¢w\$, 1, i_1) \vdash_{A_1}^{*} (p, ¢w\$, n, i_2)$. Also, $([q, 2], w, n, i_1) \vdash_{A_2} ([p, 3], w, n, i_2)$ if and only if $(q, ¢w\$, n, i_1) \vdash_{A_1} (p, ¢w\$, n + 1, i_2)$. In addition, $([q, 2], w, n, i_1) \vdash_{A_2} (*)$ if and only if $(q, ¢w\$, n, i_1) \vdash_{A_1} (p, ¢w\$, n + 1, i_1)$ for some p in F_1. For in the latter case, $([p, 4], +1)$ will be in $g_2(q, a, m)$, where $m = h(i_1)$ and a is the nth symbol of w. Note that $(f_2)_{[p,4]}$ is defined exactly when $(f_1)_p$ is defined.

Third, by rules (i) and (iv), $([q, 3], w, n, i_1) \vdash_{A_2}^{*} ([p, 3], w, n, i_2)$ if and only if $(q, ¢w\$, n + 1, i_1) \vdash_{A_1}^{*} (p, ¢w\$, n + 1, i_2)$ by a sequence of moves such that A_1 does not enter a state of F_1. Also, $([q, 3], w, n, i_1)$ $\vdash_{A_2} (*)$ if and only if $(q, ¢w\$, n + 1, i_1) \vdash_{A_1} (p, ¢w\$, n + 1, i_2)$ for some p in F_1.

Putting together the results above, we have that for w of length $n \geqq 1$, $([q_1, 1], w, 1, 1) \vdash_{A_2}^{*} (*)$ if and only if $(q_1, ¢w\$, 0, 1) \vdash_{A_1}^{*} (p, ¢w\$, n + 1, i)$ for some p in F_1 and integer, i. Also, $[q_1, 1]$ is in F_2 if and only if $¢\$$ is in L_1. Thus, ϵ is in $\hat{T}(A_2)$ exactly when $¢\$$ is in $T(A_1)$. We conclude $\hat{T}(A_2) = L_2$. It is again straightforward to see that A_2 is in class C.

We say a closed class of automata is *recursive* if there is an algorithm to determine if any given word is in $T(A)$, for any automaton A in the class. We have a corollary to Theorem 8.

Corollary: If C is recursive, then for L_1 and L_2 as in Theorem 8, we can effectively find an automaton, A_2, with $\hat{T}(A_2) = L_2$, from the specification for A_1.

Proof: It is sufficient to note that in this case, we can effectively determine if $¢\$$ is in L_1, hence we can effectively find A_2.

We can now prove a series of closure properties of the 1NBA.

Theorem 9: Let A_1 and A_2 be 1NBA *in some closed class, C. Let $L_1 = \hat{T}(A_1)$ and $L_2 = \hat{T}(A_2)$. Then there exists A_6 in C with $\hat{T}(A_6) = L_1L_2 = \{w \mid w = uv$ and u is in L_1, v in $L_2\}$.*

Proof: Let $A_1 = (S_1, I, M_1, f_1, g_1, h_1, q_1, F_1)$ and $A_2 = (S_2, I, M_2, f_2, g_2, h_2, q_2, F_2)$. By Lemma 2, we can assume that S_1 and S_2 are disjoint, and L_1 and L_2 are contained in $(I - \{¢, \$\})^*$. Define

$A_3 = (S_3, I, M_3, f_3, g_3, h_3, [q_1, 1], F_2)$, where $h_3 = h_1 \cdot h_2$, $M_3 = M_1 \cdot M_2$, and $S_3 = S_2 \cup \{q_3\} \cup \{[q, i] \mid q \text{ in } S_1, i = 1 \text{ or } 2\}$, where q_3 is a new symbol. $\hat{T}(A_3)$ will not be L_1L_2, but rather $\{w|w = uv, u$ and $v \neq \epsilon$, u in L_1, v in $L_2\}$. We define f_3 and g_3 as follows:

(*i*) $(f_3)_{[q,i]} = (f_1)_q$ if q is in S_1, $i = 1$ or 2.
(*ii*) $(f_3)_q = (f_2)_q$ if q is in S_2.
(*iii*) $(f_3)_{q_3} = \alpha^{(1)}$.

For all a in $I - \{¢, \$\}$ and m in M_3, with k the largest element in M_1:

(*iv*) If q is in S_1 but not in F_1, then $g_3([q, i], a, m) = \{([p, 1], 0) \mid (p, 0)$ is in $g_1(q, a, \sigma_1(k, m))\} \cup \{([p, 2], +1) \mid (p, +1)$ is in $g_1(q, a, \sigma_1(k, m))\}$, $i = 1$ or 2.

(*v*) If q is in F_1, then $g_3([q, 1], a, m)$ is defined as in (*iv*). $g_3([q, 2], a, m) = \{([p, 1], 0) \mid (p, 0)$ is in $g_1(q, a, \sigma_1(k, m))\} \cup \{([p, 2], +1) \mid (p, +1)$ is in $g_1(g, a, \sigma_1(k, m))\} \cup \{(q_3, 0)\}$.

(*vi*) $g_3(q_3, a, m) = g_2(q_2, a, \sigma_2(k, m))$.

(*vii*) $g_3(q, a, m) = g_2(q, a, \sigma_2(k, m))$ for all q in S_2.

Note that if A_3 is in a state of the form $[q, 2]$, then on its last move, its input head moved right. If in a state of the form $[q, 1]$, the input head did not move right on the previous move. When A_3 has just moved right and entered an accepting state, according to rule (*v*), it has the option of continuing to simulate A_1 or going to state q_3, resetting the balloon to state 1, and then simulating A_2.

Formally, from rules (*i*), (*iv*) and (*v*), it is straightforward to show that $([q_1, 1], w, 1, 1) \overset{*}{\underset{A_3}{\vdash}} ([p, 2], w, j, i)$ for $j \geqq 2$ and p in F, if and only if $(q_1, u, 1, 1) \overset{*}{\underset{A_1}{\vdash}} (*)$, where u is the first $j - 1$ symbols of w. Certainly, if and only if p is in F_1 does $([p, 2], w, j, i) \underset{A_3}{\vdash} (q_3, w, j, 1)$, by rules (*iii*) and (*v*). Finally, by rules (*ii*), (*vi*) and (*vii*), $(q_3, w, j, 1) \underset{A_3}{\vdash} (*)$ if and only if $(q_2, v, 1, 1) \overset{*}{\underset{A_2}{\vdash}} (*)$, where v is the jth and subsequent symbols of w.

Thus, $\hat{T}(A_3) = \{w \mid w = uv$ with $u, v \neq \epsilon$, u in L_1, v in $L_2\}$. Clearly, A_3 is in C. Suppose ϵ is in L_1. By Theorem 5, redone for the 1NBA, and Theorems 7 and 8, there exists A_4 in C with $\hat{T}(A_4) = \hat{T}(A_3) \cup L_2$. If ϵ is not in L_1, let $A_4 = A_3$. If ϵ is in L_2, there exists A_5 in C with $\hat{T}(A_5) = \hat{T}(A_4) \cup L_1$. If ϵ is not in L_2, let $A_5 = A_4$. Finally, if ϵ is in both L_1 and L_2, by Corollary 2 to Theorem 6, redone for the 1NBA, and Theorems 7 and 8, there exists A_6 in C with $\hat{T}(A_6) = \hat{T}(A_5) \cup \{\epsilon\}$. Otherwise, let $A_6 = A_5$. In any case, it should be clear that $\hat{T}(A_6) = L_1L_2$.

Corollary: If C is recursive, A_6 can be effectively found.

Proof: Immediate from the corollary to Theorem 8.

Theorem 10: Let A be a 1NBA *in class C, with $\hat{T}(A) = L$. Then there is an automaton, A_3, in C, with $\hat{T}(A_3) = L^* = \{\epsilon\} \cup L \cup LL \cup LLL \cup \cdots$.*

Proof: By Corollary 1 to Theorem 6, redone for 1NBA, and Theorems 7 and 8, there exists A_1 in C with $\hat{T}(A_1) = L_1 = L - \{\epsilon\}$. Note that $L^* = L_1^*$. Moreover, since ϵ is not in $\hat{T}(A_1)$, we can always effectively find A_1. Let $A_1 = (S_1, I, M, f_1, g_1, h, q_1, F_1)$. We will construct $A_2 = (S_2, I, M, f_2, g_2, h, q_2, F_2)$ in C, with $\hat{T}(A_2) = L_1^* - \{\epsilon\}$. Define $S_2 = \{q_2\} \cup \{[q, i] \mid q \text{ in } S_1, i = 1 \text{ or } 2\}$, where q_2 is a symbol not in S_1. Let $F_2 = \{[q, 2] \mid q \text{ in } F_1\}$. Define f_2 and g_2 as follows:

(*i*) $(f_2)_{[q,i]} = (f_1)_q$ for q in S_1, $i = 1$ or 2.
(*ii*) $(f_2)_{q_2} = \alpha^{(1)}$.
For all a in $I - \{¢, \$\}$ and m in M:
(*iii*) If q is in $S_1 - F_1$, $g_2([q, i], a, m) = \{([p, 1], 0) \mid (p, 0) \text{ is in } g_1(q, a, m)\} \cup \{([p, 2], +1) \mid (p, +1) \text{ is in } g_1(q, a, m)\}$, $i = 1$ or 2.
(*iv*) If q is in F_1, $g_2([q, 1], a, m)$ is as in rule (*iii*). $g_2([q, 2], a, m) = \{([p, 1], 0) \mid (p, 0) \text{ is in } g_1(q, a, m)\} \cup \{([p, 2], +1) \mid (p, +1) \text{ is in } g_1(q, a, m)\} \cup \{(q_2, 0)\}$.
(*v*) $g_2(q_2, a, m) = g_1(q_1, a, m)$.

The significance of 1 and 2 in the second component of state of S_2 is as in Theorem 9. A_2 simulates A_1, but when in an accepting state, just having moved its input head right, has the option of entering state q_2. Thus, it is easy to see that $(q_2, w, j_1, 1) \overset{*}{\underset{A_2}{\vdash}} ([p, 2], w, j_2, i)$, with $j_2 > j_1$ if and only if $(q_1, u, 1, 1) \overset{*}{\underset{A_1}{\vdash}} (^*)$, where u is symbols j_1 through $j_2 - 1$ of w. Exactly when p is in F_1 do we have $([p, 2], w, j_2, i) \underset{A_2}{\vdash} (q_2, w, j_2, 1)$. Thus, if and only if $(q_2, w, 1, 1) \overset{*}{\underset{A_2}{\vdash}} (^*)$, can w be written in the form $u_1 u_2 \cdots u_k$, $k \geqq 1$, where u_i, $1 \leqq i \leqq k$, is in L_1.

Thus, $\hat{T}(A_2) = L_1^* - \{\epsilon\} = L^* - \{\epsilon\}$. Surely, A_2 is in class C. By an argument used in Theorem 9, we can find A_3 in C, with $\hat{T}(A_3) = \hat{T}(A_2) \cup \{\epsilon\} = L^*$. Moreover, since ϵ is in $\hat{T}(A_3)$, we can always effectively find A_3.

Theorem 11: Let $A = (S, I, M, f, g, h, q_0, F)$ be a 1NBA *in class C. Let $G = (K, I - \{¢, \$\}, I_1 - \{¢, \$\}, \delta, \lambda, p_0)$ be a g.s.m. We assume for convenience that I_1 is a finite alphabet containing ¢ and \$. Let $L = \{w \mid ¢w\$ \text{ is in } T(A)\}$. Then there is an automaton, A_1 in C, with $T(A_1) = ¢L_1\$$ and $L_1 = G(L)$.*

Proof: Let $A_1 = (S_1, I_1, M, f_1, g_1, h, q_1, F_1)$. The proof is represented in Fig. 4. The finite control of A_1 contains a generator which nondeterministically generates symbols in $I - \{¢, \$\}$. These symbols are processed by G, and compared with the input. The input head rests on the leftmost uncompared symbol. A_1 also uses the generated symbols as inputs to A, which it simulates. A_1 accepts if A accepts while A_1 is scanning \$ on the input, with no symbols left to compare.

We define $S_1 = \{[q, p, a, u, i] \mid q$ in S, p in K, a in $I - \{\$\}$ or $a = \epsilon$, $i = 1, 2$ or 3 and u in $(I - \{¢, \$\})^*$, but $|u| \leqq \max (|\lambda(p, a)|$ for p in K, a in I.$\}$. The first component keeps track of the state of A, the second, of the state of G. The third component holds the symbol generated, and the fourth, the output of G for that symbol and the current state of G. The last component is 1 usually. It is 2 when A would have just moved its input head right, and it is 3 when A would be scanning ¢ or \$.

Define $F_1 = \{[q, p, \epsilon, \epsilon, i\}]\ q$ in F, $i = 1$ or $3\}$. Also, $q_1 = [q_0, p_0, \epsilon, \epsilon, 3]$.

Define f_1 by:

(*i*) $(f_1)_{[q,p,a,u,i]} = f_q$ if either $a \neq \epsilon$ or $i = 3$.
(*ii*) $(f_1)_{[q,p,a,u,i]} = \alpha^{(0)}$ otherwise.

For all m in M, b in $I_1 - \{¢, \$\}$, q in S and p in K, define:

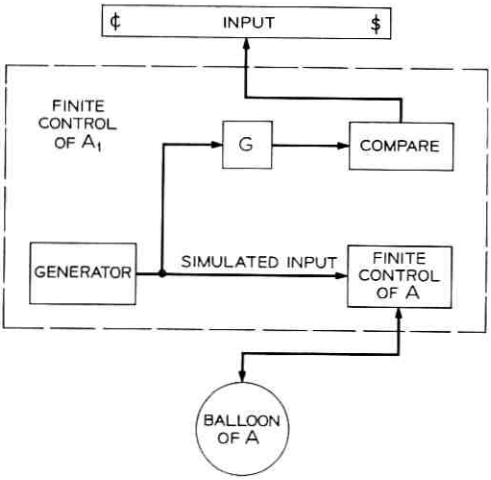


Fig. 4 — Automaton A_1.

(*iii*) $g_1([q, p_0, \epsilon, \epsilon, 3], ¢, m)$ contains $([s, p_0, \epsilon, \epsilon, 3], 0)$ if $(s, 0)$ is in $g(q, ¢, m)$.

(*iv*) $g_1([q, p_0, \epsilon, \epsilon, 3], ¢, m)$ contains $([s, p_0, ¢, \epsilon, 2], +1)$ if $(s, +1)$ is in $g(q, ¢, m)$. (A_1 simulates A with ¢ as input.)

(*v*) $g_1([q, p, \epsilon, \epsilon, 1], b, m)$ contains $\{([s, p_1, a, u, 1], 0) \mid$ for any a in $I - \{¢, \$\}$, if $(s, 0)$ is in $g(q, a, m)$, $p_1 = \delta(p, a)$, and $u = \lambda(p, a)\}$. Likewise, if $b = \$$.

(*vi*) $g_1([q, p, \epsilon, \epsilon, 1], b, m)$ contains $\{([s, p_1, a, u, 2], 0) \mid$ for a in $I - \{¢, \$\}$, if $(s, +1)$ is in $g(q, a, m)$, $p_1 = \delta(p, a)$, and $u = \lambda(p, a)\}$. Likewise, if $b = \$$. (The random generator generates symbol a, which is stored in the third component. $\lambda(p, a)$ is stored in the fourth. The new state of A, with a as input symbol is stored in the first component, and the new state of G in the second. If A would immediately move its input head right, the fifth component is 2. A 2 there tells A_1 it is finished with symbol a. Otherwise, a 1 is placed in the fifth component.)

(*vii*) $g_1([q, p, a, u, 1], b, m)$ includes $\{([s, p, a, u, 1], 0) \mid (s, 0)$ is in $g(q, a, m)\}$. Likewise, if $b = \$$.

(*viii*) $g_1([q, p, a, u, 1], b, m)$ includes $\{([s, p, a, u, 2], 0) \mid (s, +1)$ is in $g(q, a, m)\}$. Likewise, if $b = \$$. (A_1 simulates a move of A. The fifth component of A_1's state becomes 2 if the input head of A moves right.)

(*ix*) $g_1([q, p, a, u, 2], b, m) = \{([q, p, \epsilon, u, 1], 0)\}$. Likewise, if $b = \$$. (Remove a as third component and set fifth component to 1.)

(*x*) For any u, $g_1([q, p, \epsilon, bu, 1], b, m) = \{([q, p, \epsilon, u, 1], +1)\}$.

(*xi*) $g_1([q, p, \epsilon, bu, 1], b_1, m) = \varphi$ for $b_1 \neq b$. (A_1 compares its fourth component with the input.)

(*xii*) $g_1([q, p, \epsilon, \epsilon, 1], \$, m)$ contains $([s, p, \epsilon, \epsilon, 3], 0)$ if $g(q, \$, m)$ contains $(s, 0)$.

(*xiii*) $g_1([q, p, \epsilon, \epsilon, 3], \$, m)$ contains $([s, p, \epsilon, \epsilon, 3], 0)$ if $g(q, \$, m)$ contains $(s, 0)$.

We will state a series of intermediate results that follow directly from the rules given. We assume that $G(w) = x$.

I. By rules (*i*) and (*iii*): $([q_0, p_0, \epsilon, \epsilon, 3], ¢x\$, 0, 1) \overset{*}{\underset{A_1}{\vdash}} ([q, p_0, \epsilon, \epsilon, 3], ¢x\$, 0, i)$ if and only if $(q_0, ¢w\$, 0, 1) \overset{*}{\underset{A}{\vdash}} (q, ¢w\$, 0, i)$.

II. By rules (*i*), (*ii*), (*iv*), and (*ix*): $([q, p_0, \epsilon, \epsilon, 3], ¢x\$, 0, i_1) \underset{A_1}{\vdash} ([s, p_0, ¢, \epsilon, 2], ¢x\$, 1, i_2) \underset{A_1}{\vdash} ([s, p_0, \epsilon, \epsilon, 1], ¢x\$, 1, i_2)$ if and only if $(q, ¢w\$, 0, i_1) \underset{A}{\vdash} (s, ¢w\$, 1, i_2)$.

III. By rules (*i*), (*v*) and (*vii*): $([q, p, \epsilon, \epsilon, 1], ¢x\$, j, i_1) \overset{*}{\underset{A_1}{\vdash}} ([s, r, a, u, 1], ¢x\$, j, i_2)$ by a sequence of moves for which the third component of state never becomes ϵ, or the fifth $= 2$, if and only if $(q, ¢w\$, k, i_1)$

$\vdash^{*}_{A} (s, ¢w\$, k, i_2)$, where the kth symbol of w is a, $\delta(p, a) = r$ and $\lambda(p, a) = u$.

IV. From III, and rules (i), (vi) and $(viii)$: $([q, p, \epsilon, \epsilon, 1], ¢x\$, j, i_1) \vdash^{*}_{A_1} ([s, r, a, u, 2], ¢x\$, j, i_2)$, by a sequence of moves in which the third component of state never becomes ϵ if and only if $(q, ¢w\$, k, i_1) \vdash^{*}_{A} (s, ¢w\$, k + 1, i_2)$ by a sequence of moves in which the input head remains at position k until the last move is made. Here, again, the kth symbol of w is a, $\delta(p, a) = r$ and $\lambda(p, a) = u$.

V. From rule (ix), $([q, p, a, u, 2], ¢x\$, j, i) \vdash_{A_1} ([q, p, \epsilon, u, 1], ¢x\$, j, i)$ for any $a \neq \epsilon$, and if the fifth component is 2, no other move is possible.

VI. From rule (x), $([q, p, \epsilon, u, 1], ¢x\$, j_1, i) \vdash^{*}_{A_1} ([q, p, \epsilon, \epsilon, 1], ¢x\$, j_2, i)$ by a sequence of moves for which the third component of state remains ϵ, if and only if symbols j_1 through $j_2 - 1$ of x form u.

VII. Combining IV, V, and VI: $([q, p, \epsilon, \epsilon, 1], ¢x\$, j_1, i_1) \vdash^{*}_{A_1} ([s, r, \epsilon, \epsilon, 1], ¢x\$, j_2, i_2)$ by a sequence of moves in which the third component changes from ϵ to a symbol in $I - \{¢, \$\}$ back to ϵ only once, if and only if for some a in $I - \{¢, \$\}$ and u in $(I_1 - \{¢, \$\})^*$, we have

(a) $\delta(p, a) = r$;

(b) $\lambda(p, a) = u$;

(c) Symbols j_1 through $j_2 - 1$ of x are u;

(d) $(q, ¢w\$, k, i_1) \vdash^{*}_{A} (s, ¢w\$, k + 1, i_2)$ by a sequence of moves in which A's input head remains stationary until the last move, and a is the kth symbol of w.

Note that $j_1 = j_2 = n + 1$ is not prohibited.

VIII. Using I, II, and VII iterated: $([q_0, p_0, \epsilon, \epsilon, 3], ¢x\$, 0, 1) \vdash^{*}_{A_1} ([q, p, \epsilon, \epsilon, 1], ¢x\$, n + 1, i)$, where $| x | = n$, if and only if, for some w in $(I - \{¢, \$\})^*$, of length k:

(a) $(q_0, ¢w\$, 0, 1) \vdash^{*}_{A} (q, ¢w\$, k + 1, i)$ by a sequence of moves in which A's input head does not reach \$ until the last move;

(b) $\delta(p_0, w) = p$;

(c) $\lambda(p_0, w) = x$.

IX. Directly from VIII, A_1 accepts $¢x\$$ by entering a state $[q, p, \epsilon, \epsilon, 1]$, where q is in F, if and only if there is a w as in VIII such that A accepts $¢w\$$ by entering state q on the same move on which A first moves its input head to \$.

X. From rules (xii) and $(xiii)$, $([q, p, \epsilon, \epsilon, 1], ¢w\$, n + 1, i_1) \vdash^{*}_{A_1} ([s, p, \epsilon, \epsilon, 3], ¢x\$, n + 1, i_2)$, where $| x | = n$, by a sequence of moves in which the third component of state remains ϵ, if and only if $(q, ¢w\$, k + 1, i_1) \vdash^{*}_{A} (s, ¢w\$, k + 1, i_2)$, where $k = | w |$.

XI. From VIII and X, A_1 accepts $¢x\$$ by entering a state $[s, p, \epsilon, \epsilon, 3]$ if and only if A accepts $¢w\$$, where $\lambda(p_0, w) = x$, by a sequence of moves in which A enters state s while its input head remains scanning \$.

XII. Finally, from IX and XI, we have that A_1 accepts $\text{¢}x\$$ if and only if A accepts $\text{¢}w\$$, where $\lambda(p_0, w) = x$. Thus, $T(A_1) = \text{¢}G(L)\$$.

We need only add that A_1 is, by definition, in class C.

Corollary: If $L = \hat{T}(A)$ for some 1NBA, *A, in class C, and G is a g.s.m., then there is an automaton A_1 in class C, for which $\hat{T}(A_1) = G(L)$. If class C is recursive, we can effectively find A_1.*

Proof: Direct from Theorems 7 and 8.

Theorem 12: Let $L = \hat{T}(A)$ for some 1NBA, *A, in class C. Let R be a regular set. Then there is an automaton, A_3, in class C, with $\hat{T}(A_3) = L/R = \{w \mid$ for some x in R, wx is in $L\}$.*

Proof: Let $A = (S, I, M, f, g, h, q_0, F)$. We can surely find a finite automaton, $A_1 = (K, I - \{\text{¢}, \$\}, \delta, p_0, F_1)$ accepting $R \cap (I - \{\text{¢}, \$\})^*$. Intuitively, A_3 will simulate A, but will always have the additional choice of guessing that it has seen w. It then nondeterministically chooses the symbols of x, continuing to simulate A. We will construct A_2 to accept $L/R - \{\epsilon\}$. The reader can easily see how ϵ can be added to the set accepted by A_2.

Formally, let $A_2 = (S_2, I, M, f_2, g_2, h, q_0, F_2)$. $S_2 = S \cup \{[q, p, a] \mid q$ in S, p in K, a in $I - \{\text{¢}, \$\}$ or $a = \epsilon\}$. $F_2 = \{[q, p, \epsilon] \mid q$ in F, p in $F_1\}$. Define f_2 and g_2 as follows:

(*i*) $(f_2)_q = f_q$ for q in S.

(*ii*) $(f_2)_{[q,p,a]} = f_q$ for all q in S, p in K, a in $I - \{\text{¢}, \$\}$ or $a = \epsilon$.

For all b in $I - \{\text{¢}, \$\}$, m in M:

(*iii*) $g_2(q, b, m)$ contains (s, d) if $g(q, b, m)$ contains (s, d).

(*iv*) $g_2(q, b, m)$ contains $([s, p_0, \epsilon], +1)$ if $g(q, b, m)$ contains $(s, +1)$.

(*v*) $g_2(q, b, m)$ contains $([s, p, a], 0)$ if $g(q, b, m)$ contains $(s, +1)$ and $\delta(p_0, a) = p$, for any a in $I - \{\text{¢}, \$\}$.

(*vi*) $g_2([q, p, a], b, m)$ contains $([s, p, a], 0)$ if $g(q, a, m)$ contains $(s, 0)$.

(*vii*) $g_2([q, p, a], b, m)$ contains $([s, r, a_1], 0)$ for any a_1 in $I - \{\text{¢}, \$\}$ if $g(q, a, m)$ contains $(s, +1)$ and $\delta(p, a_1) = r$.

(*viii*) $g_2([q, p, a], b, m)$ contains $([s, p, \epsilon], +1)$ if $g(q, a, m)$ contains $(s, +1)$.

Note that no moves are possible if the third component of A_2's state is ϵ. From rules (*i*) and (*iii*), we see that $(q_0, w, 1, 1) \overset{*}{\vdash}_{A_2} (q, w, j, i)$ if and only if $(q_0, w, 1, 1) \overset{*}{\vdash}_{A} (q, w, j, i)$. Let w be of length n. By rules (*ii*) and (*iv*), $(q, w, n, i) \vdash_{A_2} (*)$ if and only if p_0 is in F_1 (i.e., ϵ is in R) and $(q, w, n, i) \vdash_{A} (*)$.

By rules (*ii*) and (*v*), $(q, w, n, i) \vdash_{A_2} ([s, p, a], w, n, i_2)$ if and only if $(q, wx, n, i_1) \vdash_{A} (s, wx, n + 1, i_2)$, where the first symbol of x is a and $\delta(p_0, a) = p$.† Let $| x | = k$. Then, by rules (*ii*), (*vi*), and (*vii*), $([q, p, a], w, n, i_1) \overset{*}{\vdash}_{A_2} ([s, r, a_1], w, n, i_2)$ if and only if $(q, wx, n + 1, i_1) \overset{*}{\vdash}_{A} (s, wx, n + k, i_2)$, $\delta(p, x) = r$, and x ends with a_1.

By rules (*ii*) and (*viii*), $([q, p, a], w, n, i) \vdash_{A_2} (*)$ if and only if $(q, wx, n + k, i) \vdash_{A} (*)$, where the last symbol of x is a, and x is in R.

Putting the above together, we see that $(q_0, w, 1, 1) \overset{*}{\vdash}_{A_2} (*)$ without ever entering a state of the form $[q, p, a]$, $a \neq \epsilon$ if and only if ϵ is in R and $(q_0, w, 1, 1) \overset{*}{\vdash}_{A} (*)$. Also $(q_0, w, 1, 1) \overset{*}{\vdash}_{A_2} (*)$, entering a state $[q, p, a]$, $a \neq \epsilon$ in so doing, if and only if for some x in $(I - \{¢, \$\})^*$, $(q_1, wx, 1, 1) \overset{*}{\vdash}_{A} (*)$ and x is in $R - \{\epsilon\}$. If A_2 is modified to accept ϵ, provided ϵ is in L/R, then the resulting device is A_3.

VIII. CONCLUSIONS

We have considered four types of general automata, and defined closed classes for each of these four types. We have shown certain common operations to preserve these classes, in the sense that if a language, L, is accepted by an automaton in the class, and L_1 is the result of the operation applied to L, then L_1 is accepted by some automaton in the class.

The classes model many of the common devices which have been heretofore considered in the literature, such as stack automata and counter machines. It seems as though they could be expected to model any future class of automata which are defined solely by the ways in which their infinite storage can be locally manipulated. The classes do not model such things as linear bounded automata or time/tape complexity classes of Turing machines, intuitively because such automata are defined by global restrictions on memory. (I.e., one may use "this much" memory, and no more.)

In Table I, we list the types of balloon automata and the operations considered. A check indicates that the operation preserves membership in a closed class of automata.

It is hoped that when models of automata are proposed in the future, theorists will find it efficient to show that their model is equivalent to a closed class of balloon automata. They will then have a variety of standard theorems already proven for them.

† Note, however, that the first symbol of x does not affect the operation of A at this step.

TABLE I

	2DBA	2NBA	1DBA	1NBA
Reversal	✓	✓		
Intersection	✓	✓		
g.s.m. inverse	✓	✓	✓	✓
Union		✓		✓
Intersection with regular set	✓	✓	✓	✓
Concatenation (·)				✓
Kleene closure (*)				✓
g.s.m. forward				✓
Quotient with regular set (/)				✓

IX. FUTURE PROBLEMS

There are various theorems about automata that have not been reflected in the results on balloon automata. For example, one-way deterministic pushdown automata are closed under complement. It is probably true that all common types of one-way or two-way deterministic automata are closed under complement, although proofs have not been published in all cases. Likewise, many one-way deterministic devices are closed under quotient with a regular set. Most one-way nondeterministic devices seem to be closed under reversal, and so on.

We therefore propose as an interesting and worthwhile problem, the question of putting additional restrictions on closed classes of balloon automata such that some or all of these results can be proven. Of course, the conditions must be liberal enough so that the usual automata are still modeled.

Second, it would be useful to have a model, like the balloon automaton, which could describe, as closed classes, such things as linear bounded automata and computational complexity classes. The properties of these classes deserve some treatment, and an approach similar to the one taken here might be a reasonable one.

It is hoped that the methods we have used to prove certain theorems plus the fact that we could not prove some others will shed some light on why some theorems are hard to prove, or visualize, while others are easy. Specifically, we have an indication as to why certain theorems seem easier to prove for nondeterministic devices than deterministic.

REFERENCES

1. Turing, A. M., On Computable Numbers, With an Application to the *Entscheidungsproblem,* Proc. London Math. Soc., *42,* 1936, pp. 230–265.
2. Chomsky, N., Context Free Languages and Pushdown Storage, Quarterly Progress Report No. 65, Research Laboratory of Electronics, M.I.T., Cambridge, Massachusetts, 1962.
3. Evey, R. J., The Theory and Application of Pushdown Store Machines, Math. Linguistics and Automatic Translation, Report No. NSF-10, Computation Laboratory of Harvard Univ., May, 1963.
4. Gray, J., Harrison, M., and Ibarra, O., Two-Way Pushdown Automata, Univ. of California, Berkeley Dept. of Elect. Eng. Report, 1966.
5. Ginsburg, S. and Greibach, S., Deterministic Context Free Languages, Inform. Control, *9,* No. 6, December, 1966, pp. 620–648.
6. Schutzenberger, M., Finite Counting Automata, Inform. Control, *5,* No. 2, June, 1962, pp. 91–107.
7. Minsky, M., Recursive Unsolvability of Post's Problem of 'Tag' and Other Topics in the Theory of Turing Machines, Annals of Math:, *74,* No. 3, November, 1961, pp. 437–455.
8. Ginsburg, S., Greibach, S., and Harrison, M., Stack Automata and Compiling, JACM, *14,* No. 1, January, 1967, pp. 172–201.
9. Ginsburg, S., Greibach, S., and Harrison, M., One-Way Stack Automata, JACM, *14,* No. 2, April, 1967, pp. 389–418.
10. Hopcroft, J. and Ullman J., Sets Accepted by One-Way Stack Automata are Context Sensitive. Submitted to Inform. Control.
11. Hopcroft, J. and Ullman J., Deterministic Stack Automata and the Quotient Operator, Submitted to JCSS. Also see Two Results on One-Way Stack Automata, 1967 IEEE Conf. Record Switching Auto. Theor., October, 1967.
12. Hopcroft, J. and Ullman J., Non-Erasing Stack Automata. JCSS, *1,* No. 2, August, 1967, pp. 166–186.
13. Aho, A., Indexed Grammars, — An Extension of Context Free Grammars, Princeton Univ. Ph.D. Thesis. Also 1967 IEEE Conf. Record Switching Auto. Theor., October, 1967.
14. Yamada, H., Real-Time Computation, and Functions not Real Time Computable, IRE Trans. Electron. Computer, *EC-11,* No. 6, December, 1962, pp. 753–760.
15. Hartmanis, J. and Stearns, R., On the Computational Complexity of Algorithms, Trans. Am. Math. Soc., *117,* No. 5, May, 1965.
16. Lewis, P., Stearns, R., and Hartmanis, J., Memory Bounds for the Recognition of Context Free and Context Sensitive Languages, 1965 IEEE Conf. Record Switching Circuit Theor. Logical Design, October, 1965, pp. 191–202.
17. Lewis, P., Stearns, R., and Hartmanis, J., Hierarchies of Memory Limited Computations, *Ibid.,* October 1965, pp. 179–189.
18. Hopcroft, J. and Ullman, J., Some Results on Tape Bounded Turing Machines. Submitted to JACM.
19. Ginsburg, S., Examples of Abstract Machines, IRE Trans. Electron. Computers, *EC-11,* 1962, pp. 132–135.
20. Rabin, M. and Scott, D., Finite Automata and Their Decision Problems, IBM J. Res. Devel., *3,* 1959, pp. 114–125.
21. Ginsburg, S. and Greibach, S., Abstract Families of Languages, 1967 IEEE Conf. Record Switching Auto. Theor., October, 1967.

Extensions to the Analysis of Regenerative Repeaters with Quantized Feedback

By M. K. SIMON

(Manuscript received May 23, 1967)

The functional iterative approach given by Zador for calculating the average bit error probability in a regenerative repeater with quantized feedback is extended to the vector case. For a channel with a rational fraction transfer function, the vector extension permits us at least formally to deal with the following practical conditions:

(i) The pulse transmission plan is described by an m-ary alphabet with independent digits.

(ii) Perfect and imperfect low-frequency tail cancellation cases are considered.

(iii) High-frequency signal shaping and its interaction with the predominantly low-frequency tail are taken into account.

Expressions for error probability on the kth digit are derived in terms of the kth vector iterate of a known function. The restriction to independent noise samples is also removed. The resulting expression for kth bit error probability is then derived from an operational iteration procedure which acts on the k + 1 dimensional joint distribution of the noise samples.

I. INTRODUCTION

In the design of digital communication links, various reasons exist for the removal of low-frequency components during or prior to transmission of a pulse train. In the case of vestigial sideband (VSB) modulation[1] of data over voice-frequency channels, the dc and low-frequency signal components are removed at the transmitter before modulation and carrier reinsertion. This is required to insure satisfactory carrier recovery at the receiver for relatively low transmitted carrier power. In the T-1 Carrier System,[2] the loss of low-frequency information results from transformer coupling of an unbalanced repeater to the balanced line. In either event, the effect of low-frequency suppression is to cause the positive impulse response of the overall equalized medium to exhibit an undershoot which gives intersymbol interference.

One means of reducing the effect of low-frequency suppression in a regenerative repeater is to feed back a signal in an attempt to cancel the long transient tail. This method of compensation has been called quantized feedback and its use dates back to the 1920's (as noted by Bennett[3]). We assume that the reader is familiar with Bennett's excellent expository paper. Until recently, analysis of the effects of quantized feedback on average bit error probability in a noisy environment has received essentially no attention. The first to examine this problem were Anderson, Gerrish, and Salz[4] who considered the polar binary case, neglecting signal shaping and assuming perfect matching of the feedback cancellation signal to the input signal tail. They have obtained results, with the aid of the computer, that have provided insight into the problem. In addition, they have exposed computational difficulties involved in grinding out numerical results for any given set of system parameters.

A more analytical approach to the basic problem is found in Zador[5] who used the theory of generalized random jump processes[6] to obtain an iterative procedure for computing error probability. Unfortunately, the class of physical systems that can be handled by Zador's approach as originally stated is quite restrictive in the following sense (see Fig. 1):

(*i*) The transmitted message sequence is composed of independent binary digits.

(*ii*) The low-frequency behavior of the channel as represented by $G(s)$ is dominated by a single pole.

(*iii*) $G(s)$ and $H(s)$ are exact complements of each other so that perfect feedback tail cancellation is achieved.

(*iv*) The time dispersion of the transmitted pulses caused by the medium, $C(s)$, with or without equalization $E(s)$ is strictly limited to two pulse intervals.

(*v*) The noise samples at the input to the threshold detectors are assumed independent.

It is our intention here to remove some of the above restrictions. In particular, we extend Zador's approach along the following lines:

(*i*) By allowing a multilevel threshold device as a regenerator, the

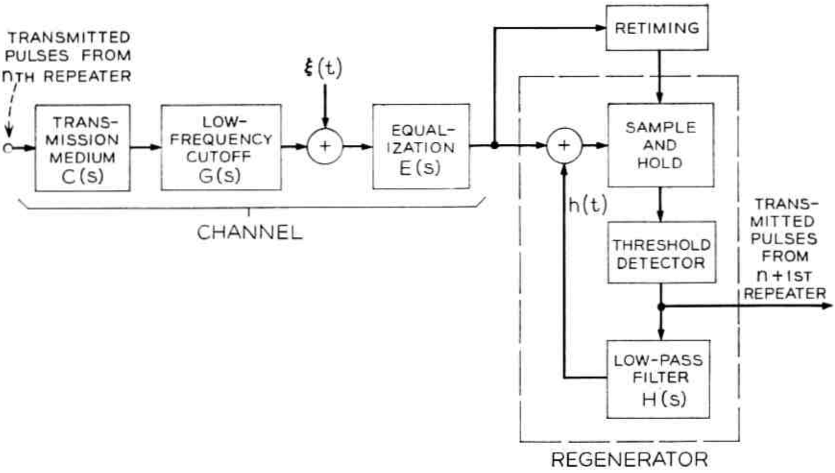

TRANSMITTED PULSES FROM nTH REPEATER
TRANS-MISSION MEDIUM C(s)
LOW-FREQUENCY CUTOFF G(s)
ξ(t)
+
EQUAL-IZATION E(s)
CHANNEL
RETIMING
+
SAMPLE AND HOLD
h(t)
THRESHOLD DETECTOR
TRANS-MITTED PULSES FROM n+1ST REPEATER
LOW-PASS FILTER H(s)
REGENERATOR

Fig. 1 — Block diagram of reconstructive repeater with quantized feedback.

allowable pulse transmission plan is extended to include m-ary alphabets with independent digits. (The ternary case is treated in detail.)

(*ii*) The high- and low-frequency behaviors of the channel may be individually characterized by rational functions. The implication of this is twofold. First, the predominantly low-frequency tail is now described by several exponentials. Secondly, the impulse response of the overall equalized medium $C(s)$ $E(s)$ is not restricted to be time-limited.

(*iii*) The restriction to perfect tail cancellation is removed to allow for imperfections in the forward and/or feedback paths.

(*iv*) The more realistic case of correlated noise samples is examined.

Extensions (*i*), (*ii*), and (*iii*) are possible only through a vector approach based on Zador's original iteration scheme. The assumption of a nonflat noise spectrum as in (*iv*) leads to an *operational* iteration procedure for calculating bit error probability. It is to be emphasized that the question of computational procedures, which even in the simple binary case was a formidable task, grows considerably in complexity with the degree of generality assumed.

The generalizations listed above will be treated one at a time so as to demonstrate individually the necessary changes in Zador's original formulation. A review of his model is given in Section II.

Section III assumes an unrestricted ternary message sequence together with the remaining restrictions as imposed by Zador. An application of the results is given for a particular high-frequency behavior of the system. The response of the high-frequency portion of the channel, $C(s)\,E(s)$, to a transmitted rectangular pulse is assumed triangular in shape and time-limited to two pulse intervals.

Section IV derives the general expression for error probability when the overall channel, $Y(s) = C(s)\,E(s)\,G(s)$ is assumed to be characterized by a rational function. The feedback network, $H(s)$, is designed to cancel only the low-frequency poles, i.e., those of $G(s)$. The special case of a binary input format is treated in detail.

Section V modifies the results of Section III by including the case of imperfect match of the $G(s)$ and $H(s)$ characteristics.

Section VI begins with Zador's original assumptions on the signaling format, channel, and feedback network characteristics, but removes the restriction of independent noise samples. An expression for kth bit error probability is derived from an *operational* iteration procedure which acts on the $k + 1$ dimensional joint distribution of the noise samples. The analogy between this scheme and the functional iteration proposed by Zador for the uncorrelated noise case is demonstrated.

II. REVIEW OF ZADOR'S MODEL

We begin with a brief review of Zador's mathematical assumptions and emphasize their physical significance. Consider once again the repeater-to-repeater transmission link illustrated in Fig. 1. The output of the nth repeater at time rT is a binary rectangular pulse* $d_r p(t-rT)$ where

$$\begin{aligned} p(t) &= p_0 \qquad |\, t \,| \leqq t_0 \\ &= 0, \qquad |\, t \,| > t_0 \end{aligned}$$

$d_r = \pm 1$, and $1/T$ is the pulse rate of the system. Zador does not explicitly describe the high-frequency behavior of the system. The class of channels that satisfies his underlying assumptions is discussed below. Let the response of $C(s)E(s)$ to the pulse $p(t)$, denoted by $z(t)$, be time limited to $2T$, and zero at its end points. It is understood that in practice these conditions are usually met only approximately. Then, by passing $z(t)$ through a single pole high-pass filter, $G(s)$, the part of the resulting

* Zador assumes ±1 impulses as repeater output. As we shall see, in the sampled systems we consider, this modification has no effect on the ensuing analysis.

response, $g(t)$, for $t \geqq 2T$ is dominated by a single exponential. If $s(t)$ is sampled at time $t = T$ and held until $t = T + t_0$, then by employing an ideal slicer element as a threshold detector a unit rectangular output pulse, $b(t)$ is regenerated. Furthermore, by passing this pulse through $H(s)$, the response tail of $s(t)$ for $t \geqq 2T$ may be exactly cancelled in the absence of noise and circuit imperfections. These observations are illustrated in Fig. 2 for a triangular pulse shape $z(t)$. The response of $H(s)$ to the regenerator output pulse is denoted by $h(t)$. Turning now to a sample notation, let g_k, h_k, and b_k represent the values of $g(t)$, $h(t)$, and $b(t)$, respectively, at time $(k + 1)T$, $k = 0, 1, 2$. Then, from Fig. 2, it is obvious that the following conditions must hold, in general, independent of the waveshape of $z(t)$ within the $2T$ interval:

$$(i) \qquad g_0 > 0, \qquad h_0 = 0$$

$$(ii) \qquad h_i + g_i = 0 \qquad i = 1, 2, \cdots .$$

$$(iii) \qquad g_i = rg_{i-1} \qquad i \geqq 2$$

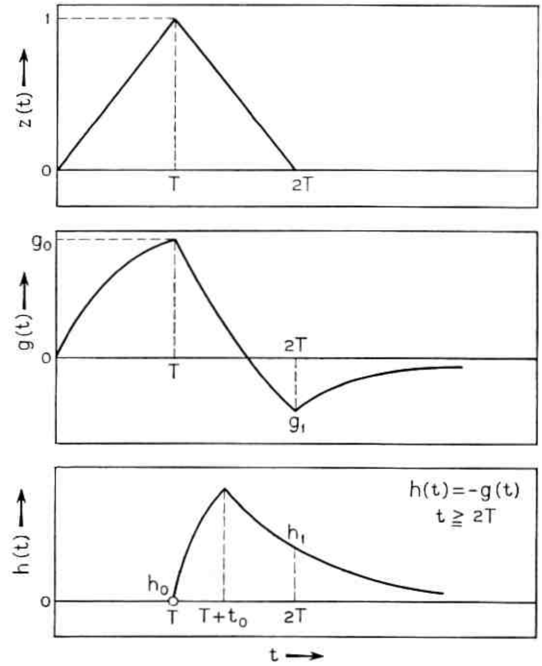


Fig. 2 — System pulse responses.

where r is related to the single pole, α, of $G(s)$ by $r = e^{-\alpha T}$. Condition (*iii*) is clear upon noting that the response of a single pole highpass filter to a time limited signal of width $2T$ has a single exponential response for values of $t \geqq 2T$. Statements (*i*) to (*iii*) as above are identical with Zador's restrictions on the system as reported in Ref. 5.* The shape of $z(t)$ is solely used in determining the two dependent quantities g_0 and g_1. For a triangular $z(t)$ waveshape of unity height (Fig. 2) and $G(s) = s/s + \alpha$,

$$g_0 = \frac{1}{\alpha T}[1 - e^{-\alpha T}]$$

$$g_1 = -\frac{1}{\alpha T}[1 - e^{-\alpha T}]^2 .$$

III. TERNARY PULSE TRANSMISSION

When considering a ternary system, the only essential modification of the model suggested by Zador is an ideal slicer with positive and negative pulse detection thresholds set at $+a_0$ and $-a_1$, respectively.

Letting s_k denote the total reshaped input at the $k + 1$th timing instant, and c_k the feedback voltage at the same instant in time as before, the slicing operation is described by

$$\begin{aligned} b_k &= 1 \quad \text{if} \quad s_k + n_k + c_k \geqq a_0 \\ &= 0 \quad \text{if} \quad -a_1 < s_k + n_k + c_k < a_0 \\ &= -1 \quad \text{if} \quad s_k + n_k + c_k \leqq -a_1 , \end{aligned}$$

where

$$s_k = \sum_{i=0}^{k} g_{k-i} d_i \qquad k = 0, 1, \cdots$$

$$c_k = \sum_{i=0}^{k} h_{k-i} b_i \qquad k = 0, 1, \cdots$$

and n_k is a sample from a stationary noise process $n(t)$ having a fixed but arbitrary distribution function $N(x)$, and independent samples. The process $n(t)$ is actually the result of passing the additive white noise process in the system, $\xi(t)$, through $E(s)$. We assume, however, that the correlation between noise samples introduced by the above is small and can be ignored as a first approximation. When this as-

* Note that Zador also requires $g_i <$ for $i \geqq 1$. This restriction is not necessary although it is often true.

sumption is invalid, the method discussed in Section VI must be used.

It is of prime interest to examine the conditions under which the system will operate error-free in the absence of noise. For $0 < a_0$, $a_1 \leqq g_0$,

$$s_0 + c_0 = g_0 d_0$$

thus if,

$$\begin{aligned} a_0 &= 1, & s_0 + c_0 &= g_0, & b_0 &= 1 \\ &= 0, & s_0 + c_0 &= 0, & b_0 &= 0 \\ &= -1, & s_0 + c_0 &= -g_0, & b_0 &= -1, \end{aligned}$$

$$\text{or} \quad b_0 = d_0 .$$

Continuing, in this way $k = 1, 2, \cdots, k - 1$,

$$\begin{aligned} s_k + c_k &= g_0 d_k + \sum_{i=0}^{k-1} (g_{k-i} + h_{k-i}) d_i \\ &= g_0 d_k . \end{aligned}$$

Thus, if $b_m = d_m$ for $m = 0, 1, \cdots, k - 1$, then $b_k = d_k$ and the system operates error-free in the absence of noise.

For the more general case when noise is present,

$$s_k + c_k = \sum_{\substack{i=0 \\ b_i \neq d_i}}^{k-1} g_{k-1}(d_i - b_i) + g_0 d_k = x_k + g_0 d_k ,$$

where x_k represents the cumulative effect of any and all errors prior to time k.

Letting p and q denote the *à priori* probabilities of a plus one and minus one, respectively, the probability of error on the kth digit $p(k)$ can be written as

$$\begin{aligned} p(k) = p \text{ Prob } \{n_k + x_k \leqq a_0 - g_0\} + q \text{ Prob } \{n_k + x_k > -a_1 + g_0\} \\ + (1 - p - q) \text{ Prob } \{n_k + x_k \geqq a_0 ; n_k + x_k \leqq -a_1\}. \end{aligned}$$

The independence of n_k and x_k allows $p(k)$ to be expressed in terms of the noise distribution function $N(x)$ and the distribution function of x_k, $F_k(x)$ as follows:

$$p(k) = p \int_{-\infty}^{\infty} N(a_0 - g_0 - x)\, dF_k(x)$$

$$+ q \int_{-\infty}^{\infty} [1 - N(g_0 - a_1 - x)]\, dF_k(x)$$

$$+ (1 - p - q) \int_{-\infty}^{\infty} [N(-a_1 - x) + 1 - N(a_0 - x)]\, dF_k(x).$$

For the case of a zero mean symmetrical noise distribution and equal *à priori* probabilities for all input symbols (i.e., $p = q = (1 - p - q)$ it is easy to show that the optimum threshold settings are $\pm g_0/2$ with

$$p(k) = (1 - p) \int_{-\infty}^{\infty} [1 - N(g_0/2 - x) + N(-g_0/2 - x)]\, dF_k(x).$$

It now remains to show that the sequence of random variables $x_0, x_1, \cdots$ are representative of a random jump process studied in Ref. 6 and thus $p(k)$ can be expressed as the kth iterate of a known function evaluated at x_0 with a finite limit as $k \to \infty$.

Consider,

$$x_{k+1} = \sum_{i=0}^{k} g_{k+1-i}(d_i - b_i)$$

$$= g_1(d_k - b_k) + \sum_{i=0}^{k-1} r g_{k-i}(d_i - b_i)$$

$$x_{k+1} = g_1(d_k - b_k) + r x_k .$$

There are five possible transition states each of which takes place with probability depending on the value of x_k.

If $d_k = 1, \ b_k = -1$, then $x_{k+1} = r x_k + 2g_1$ with probability $p_1(x_k)$.

If $d_k = 1, \ b_k = 0$

or , then $x_{k+1} = r x_k + g_1$ with probability $p_2(x_k)$.

$d_k = 0, \ b_k = -1$

If $d_k = b_k$, then $x_{k+1} = r x_k$ with probability $p_3(x_k)$.

If $d_k = -1, b_k = 0$

or , then $x_{k+1} = r x_k - g_1$ with probability $p_4(x_k)$.

$d_k = 0, \ b_k = 1$

If $d_k = -1, b_k = 1$, then $x_{k+1} = r x_k - 2g_1$ with probability $p_5(x_k)$.

The transition probabilities $p_n(x_k)$, $n = 1, 2, \cdots, 5$ are defined by

$$
\begin{aligned}
p_1(x_k) &= pN(-a_1 - g_0 - x_k) \\
p_2(x_k) &= p[N(a_0 - g_0 - x_k) - N(-a_1 - g_0 - x_k)] \\
&\qquad\qquad + (1 - p - q)[N(-a_1 - x_k)] \\
p_3(x_k) &= 1 - p_1(x_k) - p_2(x_k) - p_4(x_k) - p_5(x_k) \\
p_4(x_k) &= q[N(a_0 + g_0 - x_k) - N(-a_1 + g_0 - x_k)] \\
&\qquad\qquad + (1 - p - q)[1 - N(a_0 - x_k)] \\
p_5(x_k) &= q[1 - N(a_0 + g_0 - x_k)].
\end{aligned}
$$

Note,

$$p(k) = \int_{-\infty}^{\infty} [p_1(x) + p_2(x) + p_4(x) + p_5(x)]\, dF_k(x).$$

Defining
$$
\begin{aligned}
U^1[f(x)] = {} & p_1(x)f(rx + 2g_1) + p_2(x)f(rx + g_1) + p_3(x)f(rx) \\
& + p_4(x)f(rx - g_1) + p_5(x)f(rx - 2g_1)
\end{aligned}
$$

and denoting the kth iterate of $U^1[f(x)]$ by $U^k[f(x)]$,

$$p(k) = U^k[p_1(x) + p_2(x) + p_4(x) + p_5(x)]\, |_{x=x_0=0}\,.$$

If $A(x)$ is the limiting distribution of $F_k(x)$, then

$$
\begin{aligned}
\lim_{k\to\infty} p(k) &= \int_{-\infty}^{\infty} [p_1(x) + p_2(x) + p_4(x) + p_5(x)]\, dA(x) \\
&= \lim_{k\to\infty} U^k[p_1(x) + p_2(x) + p_4(x) + p_5(x)]\, |_{x=x_0=0}\,.
\end{aligned}
$$

A few remarks are now presented to indicate the obvious extension to the m-level (m-ary) pulse transmission scheme. A random jump process with $2m - 1$ transition states will result requiring an iteration function $U^1[f(x)]$ having $2m - 1$ terms. It should be indicated that computationally the amount of computer storage or operations required to evaluate $p(k)$ is of the order $(2m - 1)^k$.

IV. RATIONAL FUNCTION APPROXIMATIONS OF THE CHANNEL AND FEEDBACK NETWORKS

As the subtitle indicates, we are interested here in studying the repeater error performance under the assumption of a rational function approximation to the channel and feedback networks. This generalizes the assumptions of Section III in that (*i*) the tail of the pulse response, $g(t)$, is no longer described by a single exponential, and (*ii*) the high-frequency behavior of the channel allows its time response to exceed two pulse intervals. To isolate these effects, however, perfect feedback tail cancellation is still assumed and we return to a binary message format.

It is convenient to represent the output rectangular pulses of the nth repeater as the impulse response of a filter $F(s) = (1/s)[1 - e^{-st_0}]$ where t_0 is the pulse width. Including this filter in the forward path of Fig. 1, the overall channel link between repeaters, $T(s) = F(s)C(s)$ $E(s)G(s)$, is assumed to be characterized by a rational function as follows:

$$T(s) = G_0 \frac{s^M}{\prod_{i=1}^{M} (s + \alpha_i)} \times \frac{P(s)}{\prod_{i=1}^{N} (s + \beta_i)}$$

with its associated impulse response

$$g(t) = \sum_{i=1}^{M} A_i e^{-\alpha_i t} + \sum_{i=1}^{N} B_i e^{-\beta_i t}.$$

Note, the impulse response of $T(s)$ is the same as the rectangular pulse response of $Y(s) = C(s)\ E(s)\ G(s)$ and is thus denoted as before by $g(t)$. All poles are assumed to be simple, but in general may be complex. The terminology used henceforth will refer to the set $\{\alpha_i\}$, $i = 1, 2, \cdots, M$ as *low-frequency poles* and the set $\{\beta_i\}$, $i = 1, 2, \cdots, N$ as *high-frequency poles*. The inference here is that the β_i's are predominantly responsible for signal shaping and the α_i's determine the low-frequency cutoff of the channel.

A low-pass quantized feedback path $H(s)$ is proposed which in the absence of noise would provide perfect low-frequency tail cancellation at all sampling instants beyond the input pulse peak (the effect of imperfect low-frequency compensation will be discussed in Section V).* Thus, if

$$H(s) = H_0 \frac{N(s)}{\prod_{i=1}^{M} (s + \alpha_i)} e^{-s\tau_0},$$

where τ_0 represents the physical delay in the feedback path beyond the occurrence of the input pulse peak at $t = t_{\max}$, then, the response to a

* It is to be emphasized at this point that all of the following is easily generalized in terms of MacColl's conception of quantized feedback[7] wherein restoration of both low- and high-frequency signal components is attempted.

positive regenerator output pulse at $t = t_{\max}$ would be

$$h(t) = \sum_{i=1}^{M} D_i e^{-\alpha_i(t-t_{\max}-\tau_0-t_0)} = \sum_{i=1}^{M} E_i e^{-\alpha_i t} \quad \text{for} \quad t \geqq t_{\max} + \tau_0 + t_0 \,.$$

Ideally, for perfect low-frequency tail compensation, we desire

$$\sum_{i=1}^{M} E_i e^{-\alpha_i t} + \sum_{i=1}^{M} A_i e^{-\alpha_i t} = 0$$

at all instants $t_{\max} + nT$, $n = 1, 2, \cdots$, where T is the uniform sampling period.

Letting h_k and g_k represent the values of the pulse responses $h(t)$ and $g(t)$, respectively, at the kth sample point the above statements may be expressed in brief as follows:

$$(i) \qquad h_0 = 0 \qquad g_0 > 0$$

$$(ii) \qquad h_i + g_i = \sum_{n=1}^{N} e_{i,n} \qquad i \neq 0$$

$$e_{i,n} = z_n e_{i-1,n} \begin{cases} n = 1, 2, \cdots, N \\ i \geqq 2 \end{cases}$$

$$e_{0,n} = 0$$

$$z_n = e^{-\beta_n T}$$

$$(iii) \qquad h_i = \sum_{n=1}^{M} h_{i,n}$$

$$h_{i,n} = r_n h_{i-1,n} \begin{cases} n = 1, 2, \cdots, M \\ i \geqq 2 \end{cases}$$

$$r_n = e^{-\alpha_n T}.$$

The term $e_{i,n}$ represents the residual intersymbol interference at the ith timing instant due to the nth high-frequency pole. To simplify what is to follow and at the same time allow a better comparison with the previous work of Zador, we introduce the following vector notation:

$$R = \begin{bmatrix} r_1 & 0 & 0 \cdots\cdots 0 \\ 0 & r_2 \cdots\cdots & \cdots 0 \\ & & \vdots \\ 0 & & \vdots \\ \vdots & & \vdots \\ 0 \cdots & \cdots\cdots & \cdots r_M \end{bmatrix}; \qquad Z = \begin{bmatrix} z_1 & 0 & 0 \cdots\cdots 0 \\ 0 & z_2 \cdots\cdots & \cdots 0 \\ & & \vdots \\ 0 & & \vdots \\ \vdots & & \vdots \\ 0 \cdots & \cdots\cdots & \cdots z_N \end{bmatrix}$$

$$H = \begin{bmatrix} h_{0,1} & h_{1,1} & h_{2,1} & h_{3,1} & \cdots & h_{k,1} \\ h_{0,2} & h_{1,2} & h_{2,2} & h_{3,2} & \cdots & h_{k,2} \\ \vdots & \vdots & \vdots & \vdots & & \\ h_{0,M} & h_{1,M} & h_{2,M} & h_{3,M} & \cdots & h_{k,M} \end{bmatrix}$$

$$G = \begin{bmatrix} g_{0,1} & g_{1,1} & g_{2,1} & g_{3,1} & \cdots & g_{k,1} \\ g_{0,2} & g_{1,2} & g_{2,2} & g_{3,2} & \cdots & g_{k,2} \\ \vdots & \vdots & \vdots & \vdots & & \\ g_{0,M+N} & g_{1,M+N} & g_{2,M+N} & g_{3,M+N} & \cdots & g_{k,M+N} \end{bmatrix}$$

$$E = \begin{bmatrix} 0 & e_{1,1} & e_{2,1} & e_{3,1} & \cdots & e_{k,1} \\ 0 & e_{1,2} & e_{2,2} & e_{3,2} & \cdots & e_{k,2} \\ \vdots & \vdots & \vdots & \vdots & & \\ 0 & e_{1,N} & e_{2,N} & e_{3,N} & \cdots & e_{k,N} \end{bmatrix}.$$

Using H as an example, the ith row written as a column vector is denoted by $\mathbf{h}^{\mathrm{i}}$ and the ith column by the vector $\mathbf{h}_{\mathrm{i}}$. Also any vector written not in bold face is by definition the scalar representing the sum of its elements (e.g., $h^i = \sum_{p=1}^{k} h_{p,i}$). Finally, we denote the column vector obtained by summing all rows of H (i.e., whose ith component is h_i) by $\mathbf{h}$. All of the above statements are equally applied to the matrices R, Z, G, and E.

In terms of the above, (*i*), (*ii*), and (*iii*) may now be rewritten as:

(*i*) $\quad h_0 = 0, \qquad g_0 > 0$

(*ii*) $\quad \mathbf{h} + \mathbf{g} = \mathbf{e}$

$$\mathbf{e}_{\mathrm{i}} = Z\mathbf{e}_{\mathrm{i}-1}, \qquad \mathrm{i} \geqq 2$$

(*iii*) $\quad \mathbf{h}_{\mathrm{i}} = R\mathbf{h}_{\mathrm{i}-1}, \qquad \mathrm{i} \geqq 2.$

Some further interpretation of the above statements in terms of the actual system operation might prove helpful at this point. (*i*) indicates a positive input pulse peak ($g_0 > 0$) and a delay in the feedback path ($h_0 = 0$). Statements (*ii*) indicate that perfect feedback tail cancellation is achieved at the sample points starting with the second except for the effects of the high-frequency poles ($\beta_1, \beta_2, \cdots, \beta_N$). In contrast to the previous sections, we do not assume that the high-frequency components of the response have died out before the occurrence of the next input pulse peak. Feedback cancellation of only the channel low-frequency poles is described by statements (*iii*).

For the binary message case, the threshold detector box of Fig. 1 reduces to a simple ideal slicer element operating between +1 and −1 levels. The input sequence $\{d_i\}$ is a random train of +1 and −1 impulses represented by the vector **d** with elements d_i .

The total reshaped input at the $k + 1$th timing instant, s_k , and the feedback voltage at the same instant, c_k , are described by,

$$\begin{aligned} s_k &= (\mathbf{d} * \mathbf{g})_k \qquad k = 0, 1, 2, \cdots, \\ c_k &= (\mathbf{b} * \mathbf{h})_k \end{aligned}$$

where the kth element of **b**, $b_k = \operatorname{sgn} \{s_k + c_k + n_k\}$ is the kth regenerator output digit. The notation $(\mathbf{a} * \mathbf{b})_k$ represents the convolution of two $k + 1$ dimensional vectors **a** and **b** (i.e., $\sum_{i=0}^{k} a_i b_{k-i}$).

Considering first operation in the absence of noise, we see by inspection $b_0 = d_0$. (This tacitly assumes that no intersymbol interference due to precursors is present.) Proceeding as in Zador,[5] if $b_m = d_m$ for $m = 0, 1, \cdots, k - 1$, then

$$s_k + c_k = (\mathbf{d} * \mathbf{e})_k = g_0 d_k + (\mathbf{d} * \mathbf{e})_{k-1} .$$

From this, one concludes that if

$$g_0 > \sum_{i=1}^{k-1} | e_i |,$$

then $b_k = d_k$ and the eye is open. The system will therefore operate error-free in the absence of noise for *any* length input sequence if

$$g_0 > \sum_{i=1}^{\infty} | e_i |.$$

If all N high-frequency poles $(\beta_1, \beta_2, \cdots, \beta_N)$ have positive residues, the above criterion reduces to

$$g_0 > \sum_{n=1}^{N} \frac{e_{1,n}}{1 - z_n}.$$

The above implies that the eye is open if the total high-frequency contribution at all sample points beyond the first is smaller than the pulse peak.

More specifically, the values of $e_{1,n}$ and z_n may be related to the allowable amount of degradation of the eye. That is, for any eye which is X percent closed.

$$\sum_{n=1}^{N} \frac{e_{1,n}}{1 - z_n} = \frac{X}{100} g_0 .$$

Turning now to the more realistic situation in the presence of noise

$$s_k + c_k = g_0 d_k + (\mathbf{d} * \mathbf{g})_{k-1} + (\mathbf{b} * \mathbf{h})_{k-1} .$$

Consider subdividing the vector $\mathbf{d}$ into two parts $\mathbf{d}'$ and $\mathbf{d}''$ in such a way as to separate the input digits into two classes corresponding to $b_i = d_i$ and $b_i \neq d_i$ respectively. That is,

$$\begin{aligned} d_i' &= d_i ; & d_i'' &= 0 & \{i; b_i = d_i\} \\ &= 0 & &= d_i & \{i; b_i \neq d_i\}. \end{aligned}$$

(Obviously $\mathbf{d} = \mathbf{d}' + \mathbf{d}''$.)
Then, using (*ii*),

$$\mathbf{u}_k = (\mathbf{d} * E^T)_{k-1} ; \mathbf{v}_k = -2(\mathbf{d} * H^T)_{k-1}$$

$$\begin{aligned} s_k + c_k &= g_0 d_k + u_k + v_k \\ &= g_0 d_k + x_k , \end{aligned}$$

where $(\mathbf{d} * G)_{k-1}$ is a vector whose ith component is the convolution of $\mathbf{d}$ with the ith column of G. Again omission of the bold face notation indicates summation over all the components and T is the transpose operator.

The first term in x_k denoted by u_k represents intersymbol interference due to residual high-frequency tail components irrespective of previous decisions. The second term v_k again represents the cumulative effect of any and all errors prior to time k.

The expression for error probability on the kth digit is identical to that given by Zador, namely,

$$p(k) = p \int_{-\infty}^{\infty} N(-g_0 - x) \, dF_k(x) + q \int_{-\infty}^{\infty} [1 - N(g_0 - x)] \, dF_k(x).$$

The only difference being the nature of the distribution function $F_k(x)$.

The recursive properties of the intersymbol interference x_k are now examined.

$$x_{k+1} = (\mathbf{d} * E^T)_k - 2(\mathbf{d}'' * H^T)_k .$$

If $b_k \neq d_k$, then

$$\begin{aligned} x_{k+1} &= (e_1 - 2h_1)d_k + \mathbf{z}^T(\mathbf{d} * E^T)_{k-1} - 2\mathbf{r}^T(\mathbf{d}'' * H^T)_{k-1} \\ &= (e_1 - 2h_1)d_k + \mathbf{z}^T\mathbf{u_k} + \mathbf{r}^T\mathbf{v_k} . \end{aligned}$$

If $b_k = d_k$, then

$$\begin{aligned} x_{k+1} &= e_1 d_k + \mathbf{z}^T(\mathbf{d} * E^T)_{k-1} - 2\mathbf{r}^T(\mathbf{d}'' * H^T)_{k-1} \\ &= e_1 d_k + \mathbf{z}^T\mathbf{u_k} + \mathbf{r}^T\mathbf{v_k} . \end{aligned}$$

Letting $\boldsymbol{a} = -2\mathbf{h}_1$, the intersymbol interference sequence x_0, x_1, x_2, $\cdots$ may be expressed as a random jump process,[5,6] with the following transition states:

If $d_k = 1 \neq b_k$, then with probability $p_1(x_k)$

$$x_{k+1} = \mathbf{z}^T\mathbf{u_k} + e_1 + \mathbf{r}^T\mathbf{v_k} + a.$$

If $d_k = 1 = b_k$, then with probability $p_2(x_k)$

$$x_{k+1} = \mathbf{z}^T\mathbf{u_k} + e_1 + \mathbf{r}^T\mathbf{v_k} .$$

If $d_k = 1 \neq b_k$, then with probability $p_3(x_k)$

$$x_{k+1} = \mathbf{z}^T\mathbf{u_k} - e_1 + \mathbf{r}^T\mathbf{v_k} - a.$$

If $d_k = -1 = b_k$, then with probability $p_4(x_k)$

$$x_{k+1} = \mathbf{z}^T\mathbf{u_k} - e_1 + \mathbf{r}^T\mathbf{v_k} ,$$

where

$$\begin{aligned} p_1(x_k) &= pN(-g_0 - x_k) \\ p_2(x_k) &= p[1 - N(-g_0 - x_k)] \\ p_3(x_k) &= q[1 - N(g_0 - x_k)] \\ p_4(x_k) &= q[N(g_0 - x_k)]. \end{aligned}$$

In the above, $N(x)$ is the distribution function of the stationary noise process, and p and q are the *a priori* probabilities of a plus one and minus one, respectively. In terms of the above elementary probability density functions, the error probability on the kth digit may be expressed as:

$$p(k) = \int_{-\infty}^{\infty} [p_1(x) + p_3(x)]\, dF_k(x).$$

We propose a vector extension of Zador's procedure, namely; an $M + N$ dimensional iteration scheme in which *each* of $M + N$ variables is replaced by a linear transformation on itself during each iteration. To elucidate the meaning of $M + N$ dimensional iteration and at the same time recall some of our earlier vector notation, the first-order iteration function U^1f is written in summation notation as:

$$\begin{aligned} U^1f(\boldsymbol{\theta}, \boldsymbol{\varphi}) = {} & p_1\left(\sum_{n=1}^{N} \theta_n + \sum_{m=1}^{M} \varphi_m\right) \cdot f\left[\sum_{n=1}^{N} (z_n\theta_n + e_{1,n}) + \sum_{m=1}^{M} (r_m\varphi_m + a_m)\right] \\ & + p_2\left(\sum_{n=1}^{N} \theta_n + \sum_{m=1}^{M} \varphi_m\right) \cdot f\left[\sum_{n=1}^{N} (z_n\theta_n + e_{1,n}) + \sum_{m=1}^{M} (r_m\varphi_m)\right] \\ & + p_3\left(\sum_{n=1}^{N} \theta_n + \sum_{m=1}^{M} \varphi_m\right) \cdot f\left[\sum_{n=1}^{N} (z_n\theta_n - e_{1,n}) + \sum_{m=1}^{M} (r_m\varphi_m - a_m)\right] \\ & + p_4\left(\sum_{n=1}^{N} \theta_n + \sum_{m=1}^{M} \varphi_m\right) \cdot f\left[\sum_{n=1}^{N} (z_n\theta_n - e_{1,n}) + \sum_{m=1}^{M} (r_m\varphi_m)\right]. \end{aligned}$$

It follows that the probability of error on the kth digit is

$$p(k) = U^k[p_1 + p_3] \,|_{\boldsymbol{\theta},\varphi=0}$$

where U^k is the kth $M + N$ dimensional iterate of U^1. The convergence of $p(k)$ in the limit as $k \to \infty$ has not been examined for an $M + N$ dimensional branching process. From Zador's work on one-dimensional branching processes[6] we may conjecture that absolute system stability (i.e., all poles in left-half plane) implies convergence in the multidimensional case.

Although the notation in the foregoing analysis appears formidable (quite an understatement) the procedure and its usage are straightforward (at least analytically) for a particular example. At the expense of being redundant, we once again point out that even in simple cases, numerical results are hard to come by.

V. IMPERFECT LOW-FREQUENCY TAIL CANCELLATION

It is relatively simple at this point to include the effect of imperfect low-frequency cancellation in the results of Section IV. As an example, such a phenomenon might be caused by a delay of amount τ in the feedback path. Defining an L matrix by

$$L = \begin{bmatrix} 0 & l_{1,1} & l_{2,1} & l_{3,1} & \cdots & l_{k,1} \\ 0 & l_{1,2} & l_{2,2} & l_{3,2} & \cdots & l_{k,2} \\ \vdots & \vdots & \vdots & \vdots & & \\ 0 & l_{1,M} & l_{2,M} & l_{3,M} & \cdots & l_{k,M} \end{bmatrix},$$

where

$$l_{1,m} = h_{1,m}\left[1 - \exp\left(-\frac{\tau}{T}\log_e \frac{1}{r_m}\right)\right] \qquad m = 1, 2, \cdots, M,$$

statement (*ii*) of Section IV may be modified as follows:

$$\begin{aligned} (ii) \qquad \mathbf{h} + \mathbf{g} &= \mathbf{e} + \mathbf{1} \\ \mathbf{l}_{\mathrm{i}} &= R l_{\mathrm{i}-1} \qquad \mathbf{i} \geqq 2 \\ \mathbf{e}_{\mathrm{i}} &= Z e_{\mathrm{i}-1} \,. \end{aligned}$$

The effect of this on the recursion relationship for x_k is as follows: If $b_k \neq d_k$, then

$$x_{k+1} = (e_1 + l_1 - 2h_1)d_k + \mathbf{z}^T\mathbf{u}_{\mathrm{k}} + \mathbf{r}^T\mathbf{v}_{\mathrm{k}} + \mathbf{r}^T\boldsymbol{\omega}_{\mathrm{k}}.$$

If $b_k = d_k$, then

$$x_{k+1} = (e_1 + l_1)d_k + \mathbf{z}^T\mathbf{u}_{\mathrm{k}} + \mathbf{r}^T\mathbf{v}_{\mathrm{k}} + \mathbf{r}^T\boldsymbol{\omega}_{\mathrm{k}} \,,$$

where $\boldsymbol{\omega}_{\mathrm{k}} = (\mathbf{d} * L^T)_{k-1}$.

If $d_k = 1 \neq b_k$, then with probability $p_1(x_k)$

$$x_{k+1} = \mathbf{z}^T\mathbf{u}_{\mathrm{k}} + e_1 + \mathbf{r}^T\mathbf{v}_{\mathrm{k}} + a + \mathbf{r}^T\boldsymbol{\omega}_{\mathrm{k}} + l_1.$$

If $d_k = 1 = b_k$, then with probability $p_2(x_k)$

$$x_{k+1} = \mathbf{z}^T\mathbf{u}_{\mathrm{k}} + e_1 + \mathbf{r}^T\mathbf{v}_{\mathrm{k}} + \mathbf{r}^T\boldsymbol{\omega}_{\mathrm{k}} + l_1.$$

If $d_k = 1 \neq b_k$, then with probability $p_3(x_k)$

$$x_{k+1} = \mathbf{z}^T\mathbf{u}_{\mathrm{k}} - e_1 + \mathbf{r}^T\mathbf{v}_{\mathrm{k}} - a + \mathbf{r}^T\boldsymbol{\omega}_{\mathrm{k}} - l_1.$$

If $d_k = -1 = b_k$, then with probability $p_4(x_k)$

$$x_{k+1} = \mathbf{z}^T\mathbf{u}_{\mathrm{k}} - e_1 + \mathbf{r}^T\mathbf{v}_{\mathrm{k}} + \mathbf{r}^T\boldsymbol{\omega}_{\mathrm{k}} - l_1 \,,$$

where $p_1(x_k)$, $p_2(x_k)$, $p_3(x_k)$, and $p_4(x_k)$ are still defined as in Section IV.

The *k*th bit probability of error is now evaluated by a $2M + N$ dimensional iteration scheme where the first-order iteration function $U^1 f$ is written as

$$\begin{aligned} U^1 f(\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\gamma}) = p_1&\left(\sum_{n=1}^{N} \theta_n + \sum_{m=1}^{M} \varphi_m + \sum_{m=1}^{M} \gamma_m\right) \\ &\cdot f\left[\sum_{n=1}^{N} (z_n\theta_n + e_{1,n}) + \sum_{m=1}^{M} (r_m\varphi_m + a_m) + \sum_{m=1}^{M} (r_m\gamma_m + l_{1,m})\right] \end{aligned}$$

$$+ p_2\left(\sum_{n=1}^{N} \theta_n + \sum_{m=1}^{M} \varphi_m + \sum_{m=1}^{M} \gamma_m\right)$$

$$\cdot f\left[\sum_{n=1}^{N} (z_n\theta_n + e_{1,n}) + \sum_{m=1}^{M} (r_m\varphi_m) + \sum_{m=1}^{M} (r_m\gamma_m + l_{1,m})\right]$$

$$+ p_3\left(\sum_{n=1}^{N} \theta_n + \sum_{m=1}^{M} \varphi_m + \sum_{m=1}^{M} \gamma_m\right)$$

$$\cdot f\left[\sum_{n=1}^{N} (z_n\theta_n - e_{1,n}) + \sum_{m=1}^{M} (r_m\varphi_m - a_m) + \sum_{m=1}^{M} (r_m\gamma_m - l_{1,m})\right]$$

$$+ p_4\left(\sum_{n=1}^{N} \theta_m + \sum_{m=1}^{M} \varphi_m + \sum_{m=1}^{M} \gamma_m\right)$$

$$\cdot f\left[\sum_{n=1}^{N} (z_n\theta_n - e_{1,n}) + \sum_{m=1}^{M} (r_m\varphi_m) + \sum_{m=1}^{M} (r_m\gamma_m - l_{1,m})\right].$$

It once again follows that the probability of error on the kth digit is

$$p(k) = U^k[p_1 + p_3] \,|_{\boldsymbol{\theta},\boldsymbol{\varphi},\boldsymbol{\gamma}=0} ,$$

where U^k is the kth $2M + N$ dimensional iterate of U^1.

To reward the reader for his patience up to this point, we will at least demonstrate that the general expression for $p(k)$ given above reduces to Zador's result for the single low-frequency pole, perfect cancellation case. The assumption of no high-frequency signal shaping and perfect cancellation imply that $\boldsymbol{\theta}$, $\mathbf{e}_1$, and $\boldsymbol{\gamma}$, $\mathbf{l}_1$ are, respectively, zero. Furthermore, a single low-frequency pole results in r_1, φ_1, and a_1 being the only nonzero components of $\mathbf{r}$, $\boldsymbol{\varphi}$, and $\mathbf{a}$, respectively. Under these conditions,

$$p(k) = U^k[p_1 + p_3] \,|_{\varphi_1=0} ,$$

where

$$U^1f = p_1(\varphi)f(r_1\varphi_1 + a_1) + p_2(\varphi_1)f(r_1\varphi_1)$$
$$+ p_3(\varphi_1)f(r_1\varphi_1 - a_1) + p_4(\varphi_1)f(r_1\varphi_1)$$

which is identical to Zador's result upon combining $p_2(\varphi_1)$ and $p_4(\varphi_1)$.

VI. THE EFFECT OF NOISE CORRELATION

In this part, the emphasis is placed upon removing the restriction of uncorrelated noise while at the same time arranging the results in a form which allows easy comparison with the uncorrelated case. The approach to be followed is the reformulation of Zador's work into an *operational* iteration procedure which acts on the joint distribution of the noise samples. The details are presented for the simple binary case with perfect feedback cancellation considered by Zador. With sufficient patience, extension to the more general situations covered in the foregoing sections can be accomplished, but that is not done here.

To review, the operation of the simplified system may be described by the equation

$$b_k = \operatorname{sgn} \{n_k + g_0 d_k + x_k\}, \tag{1}$$

where

$$x_k = 2 \sum_{\substack{i=0 \\ b_i = -d_i}}^{k-1} g_{k-i} d_i \tag{2}$$

represents the intersymbol interference accumulated at time t_k as a result of errors ($d_i \neq b_i$) prior to that time.

The system output b_k is in error when

$$\begin{aligned} n_k + x_k &< -g_0 \quad \text{and} \quad d_k = 1 \\ n_k + x_k &> g_0 \quad\;\; \text{and} \quad d_k = -1. \end{aligned} \tag{3}$$

Since the noise samples are not assumed to be independent, the random variables n_k and x_k are not independent. Hence, the distribution of the effective noise $n_k + x_k$ is not simply the convolution of the distributions of n_k and x_k. Instead, the expression for error probability $p(k) =$ prob $\{b_k \neq d_k\}$ must be written as

$$\begin{aligned} p(k) = p \int_{-\infty}^{\infty} \int_{-\infty}^{-g_0 - x_k} & m_2(n_k\ , x_k)\, dn_k\, dx_k \\ & + q \int_{-\infty}^{\infty} \int_{g_0 - x_k}^{\infty} m_2(n_k\ , x_k)\, dn_k\, dx_k\ , \end{aligned} \tag{4}$$

where $m_2(n_k, x_k)$ is the joint density function of n_k and x_k and p and q are the *a priori* probabilities of a $+1$ and -1, respectively.

A careful examination of the branching process described in Refs. 5 and 6 for the uncorrelated case shows that a similar process governs in the correlated noise case. Define the integral operators $p_1(x)$, $p_2(x)$, $p_3(x)$ by

$$\begin{aligned} p_1(x) &= p \int_{-\infty}^{-g_0 - x} \\ p_2(x) &= \int_{-\infty}^{\infty} - p \int_{-\infty}^{-g_0 - x} - q \int_{g_0 - x}^{\infty} \\ p_3(x) &= q \int_{g_0 - x}^{\infty} . \end{aligned} \tag{5}$$

Note that the action of each of these three operators on a single dimension Gaussian density function results in the three transition probabilities defined by (13) of Ref. 5. If $f(x)$ is defined analogously (but operationally) as

$$f(x) = p_1(x) + p_3(x) = p \int_{-\infty}^{-g_0-x} + q \int_{g_0-x}^{\infty} , \tag{6}$$

then the first-order iterative *operator* $Uf(x)$ is expressed as in Zador, namely:

$$Uf(x) = p_1(x)f(rx - a) + p_2(x)f(rx) + p_3(x)f(rx + a) \tag{7}$$

with $a = -2g_1$. We note that after separating $f(x)$ into its two components parts, each term of (7) represents a double integration and thus (7) has meaning only when applied to a second-order density function. Proceeding as in Zador, the error probability on the k+1th digit in a random input sequence is expressed as the kth iterate of the operator $Uf(x)$ acting on the k+1 dimensional joint density function of the noise process v_{k+1} $(\gamma_1, \gamma_2, \cdots, \gamma_{k+1})$ evaluated at $x = 0$, i.e.,

$$p(k) = U^k f(x)[v_k(\gamma_1 , \gamma_2 , \cdots , \gamma_{k+1})] \mid_{x=0} .\text{*} \tag{8}$$

The meaning of iteration for the operators defined here is the same as in Zador's functional case. As an example, we write out $p(1)$ in detail:

$$\begin{aligned} p(1) = Uf(x)[v_2(\gamma_1 , \gamma_2)] \mid_{x=0} &= p^2 \int_{-\infty}^{-g_0} \int_{-\infty}^{-g_0+a} v_2(\gamma_1 , \gamma_2)\, d\gamma_1\, d\gamma_2 \\ &+ pq \int_{-\infty}^{-g_0} \int_{g_0+a}^{\infty} v_2(\gamma_1 , \gamma_2)\, d\gamma_1\, d\gamma_2 + p \int_{-\infty}^{\infty} \int_{-\infty}^{-g_0} v_2(\gamma_1 , \gamma_2)\, d\gamma_1\, d\gamma_2 \\ &+ q \int_{-\infty}^{\infty} \int_{g_0}^{\infty} v_2(\gamma_1 , \gamma_2)\, d\gamma_1\, d\gamma_2 - p^2 \int_{-\infty}^{-g_0} \int_{-\infty}^{-g_0} v_2(\gamma_1 , \gamma_2)\, d\gamma_1\, d\gamma_2 \\ &- pq \int_{-\infty}^{-g_0} \int_{g_0}^{\infty} v_2(\gamma_1 , \gamma_2)\, d\gamma_1\, d\gamma_2 - pq \int_{g_0}^{\infty} \int_{-\infty}^{-g_0} v_2(\gamma_1 , \gamma_2)\, d\gamma_1\, d\gamma_2 \\ &- q^2 \int_{g_0}^{\infty} \int_{g_0}^{\infty} v_2(\gamma_1 , \gamma_2)\, d\gamma_1\, d\gamma_2 + pq \int_{g_0}^{\infty} \int_{-\infty}^{-g_0-a} v_2(\gamma_1 , \gamma_2)\, d\gamma_1\, d\gamma_2 \\ &+ q^2 \int_{g_0}^{\infty} \int_{g_0-a}^{\infty} v_2(\gamma_1 , \gamma_2)\, d\gamma_1\, d\gamma_2 . \end{aligned} \tag{9}$$

* The convergence of the *operational* iteration procedure defined by (7) and (8) has not yet been proven. Nonetheless, we proceed with our results.

The above expression for $p(1)$ can be simplified for a symmetric density function v_2. It is further emphasized that the arguments of each term in the operator as defined by (7) determine the limits on the integrals in (9) (i.e., the region of integration).

VII. CONCLUSIONS

The analysis presented in this paper might in a broad sense be described as vector and operational extensions of the work of Zador. In addition to simply considering a vector of low-frequency poles, however, the vector approach has enabled us to remove certain other restrictions from the basic regenerator problem such as lack of high-frequency signal shaping and perfect tail cancellation. Although, the question of convergence of the operational iteration scheme for correlated noise samples remains as yet unanswered, the formulation itself, is of interest. Little has been suggested for solving the exact computational problem. A future paper will discuss some useful approximations to cases of relatively low dimensionality. This will generalize results given in Ref. 8.

VIII. ACKNOWLEDGMENT

The author wishes to express his thanks to M. R. Aaron for his criticism of the manuscript. Much insight into the problem was obtained through many fruitful discussions with him.

REFERENCES

1. Becker, F. K., Davey, J. R., and Saltzberg, B. R., *AM Vestigial Sideband Data Transmission Set Using Synchronous Detection,* AIEE Trans., Part I, Commun. and Electron., No. 60, May, 1962, pp. 97–101.
2. Aaron, M. R., *PCM Transmission in the Exchange Plant,* B.S.T.J., *41,* January, 1962, pp. 99–141.
3. Bennett, W. R., *Synthesis of Active Networks,* 1955 Proc. Symp. Mod. Net. Synth., pp. 45–61.
4. Anderson, R. R., Gerrish, A. M., and Salz, J., *Error Rates for DC Restoration in VSB Transmission of Binary Data,* unpublished work.
5. Zador, P. L., *Error Probabilities in Data System Pulse Regenerator with DC Restoration,* B.S.T.J., *45,* July, 1966, pp. 979–984.
6. Zador, P. L., *On Random Jump Processes,* unpublished work.
7. MacColl, L. A., U. S. Patent 2,056,284, issued October 6, 1936.
8. Aaron, M. R. and Simon, M. K., *Approximation of the Error Probability in a Regenerative Repeater with Quantized Feedback,* B.S.T.J., *45* December, 1966, pp. 1845–1847.

Factoring Polynomials Over Finite Fields

By E. R. BERLEKAMP

(Manuscript received May 9, 1967)

We present here an algorithm for factoring a given polynomial over $GF(q)$ into powers of irreducible polynomials. The method reduces the factorization of a polynomial of degree m over $GF(q)$ to the solution of about $m(q - 1)/q$ linear equations in as many unknowns over $GF(q)$.

There are many applications in which one wishes to factor polynomials. Some programming systems, such as Brown's ALPAK,[1] deal with polynomials and rational functions with integer coefficients. In such a context one is interested not in approximate numerical values for the real and complex roots, but rather in irreducible factors which are themselves polynomials with integer coefficients. One of the standard tricks mentioned by Johnson[2] for finding such irreducible factors is to reduce all of the coefficients of the original polynomial modulo some prime, p, and then factor the reduced polynomial over the Galois Field, $GF(p)$. If the reduced polynomial factors, one gets certain constraints on the factors of the original polynomial; if the reduced polynomial does not factor over $GF(p)$, then one may conclude that the original polynomial is irreducible over the integers. The success of this method for factoring polynomials over the integers clearly depends upon having an efficient procedure for factoring polynomials over $GF(p)$.

The problem of factoring polynomials over finite fields arises directly in Golomb's study[3] of feedback shift register sequences. In Golomb's words, this study "... has found major applications in a wide variety of technological situations, including secure, reliable and efficient communications, digital ranging and tracking systems, deterministic simulation of random processes, and computer sequencing and timing schemes." The properties of all cyclic error correcting codes, including the important Bose-Chaudhuri[4]-Hocquenghem[5] codes, depend on the factors of their generator polynomials in some finite field. Such codes have been studied extensively by Peterson[6] and MacWilliams.[7] Recent advances in decoding techniques by Berlekamp[8] make these codes even more attractive from the practical standpoint.

We present here an algorithm for factoring a given polynomial,

$$f(z) = \sum_{k=0}^{m} f_k z^k, \qquad f_i \,\varepsilon\, GF(q),$$

into powers of irreducible polynomials.

First, we construct the $m \times m$ matrix Q over $GF(q)$, whose ith row represents $z^{q(i-1)}$ reduced modulo $f(z)$. Specifically,

$$z^{qi} \equiv \sum_{k=0}^{m-1} Q_{i+1,k+1} z^k \bmod f(z).$$

The Q matrix may be computed with a shift register wired to multiply by z mod $f(z)$. The register is started at 1, which is the first row of Q. After q shifts, the register contains the second row of Q; after q more shifts, it contains the third row of Q, $\cdots$, etc. After $q(m - 1)$ shifts, it contains the last row of Q.

Given any polynomial $g(z)$ of degree $<m$ over $GF(q)$, $g(z) = \sum_{i=0}^{m-1} g_i z^i$, we may compute the residue of $(g(z))^q$ mod $f(z)$ by multiplying the row vector $[g_0, g_1, \cdots, g_{m-1}]$ by the Q matrix. This follows from the observation that

$$(g(z))^q = g(z^q) = \sum_{i=0}^{m-1} g_i z^{qi} = \sum_{i=0}^{m-1} \left(\sum_{k=0}^{m-1} g_i Q_{i+1,k+1} z^k \right)$$
$$= \sum_{k=0}^{m-1} \left(\sum_{i=0}^{m-1} g_i Q_{i+1,k+1} \right) z^k.$$

Similarly, we could compute $(g(z))^q - g(z)$ mod $f(z)$ by multiplying the row vector $[g_0, g_1, \cdots, g_{m-1}]$ by the matrix $(Q - I)$, where I is the $m \times m$ identity matrix over $GF(q)$.

Second, we find a set of row vectors which span the null space of $(Q - I)$. This may be done by appropriate column operations on the matrix $(Q - I)$.[8] Each such row vector in the null space of $(Q - I)$ represents a polynomial $g(z)$ which satisfies the equation $(g(z))^q - g(z) \equiv 0 \bmod f(z)$, and conversely, each $g(z)$ which satisfies this equation is represented by a row vector in the null space of $(Q - I)$.

Third, we select any of the polynomials $g(z)$ found in the second step, and apply Euclid's algorithm to determine the greatest common divisor of $f(z)$ and $g(z) - s$ for each $s \varepsilon GF(q)$.* We then have the factorization

$$f(z) = \prod_{s \varepsilon GF(q)} (\text{g.c.d.}\ (f(z),\ g(z) - s)).$$

Remark: If $g(z)$ is a scalar, then this factorization degenerates into

$$f(z) = \text{g.c.d.}\ (f(z), 0) \prod_{s \neq 0} \text{g.c.d.}\ (b(z), s)$$
$$= f(z) \prod_{s \neq 0} 1.$$

However, if $g(z)$ has positive degree, then the factorization is nontrivial.

Proof: Since $(g(z))^q - g(z) \equiv 0 \bmod f(z)$, $f(z)$ divides $(g(z))^q - g(z) = \prod_{s \varepsilon GF(q)}(g(z)) - s$. Therefore, $f(z)$ also divides

$$\prod\nolimits_{s \varepsilon GF(q)}(\text{g.c.d.}\ (f(z),\ g(z) - s)).$$

On the other hand, g.c.d. $(f(z),\ g(z) - s)$ divides $f(z)$. If $s \neq t$, and $s,\ t \varepsilon GF(q)$, then $g(z) - s$ and $g(z) - t$ are relatively prime, as are g.c.d. $(f(z),\ g(z) - s)$, and g.c.d. $(f(z),\ g(z) - t)$. Therefore,

$$\prod\nolimits_{s \varepsilon GF(q)}(\text{g.c.d.}\ (f(z),\ g(z) - s))$$

divides $f(z)$. Assuming both polynomials to be monic, they must be equal since each divides the other. Q.E.D.

Example I: Let $f(z)$ be the binary polynomial 1110001110001, or $f(z) = 1 + z + z^2 + z^6 + z^7 + z^8 + x^{12}$. The successive powers of z are

100000000000
010000000000
001000000000
000100000000
000010000000
000001000000
000000100000
000000010000
000000001000
000000000100
000000000010
000000000001

111000111000
011100011100
001110001110
000111000111
111011011011
100101010101
101010010010
010101001001
110010011100
011001001110
001100100111

[*] In practice, there is no need to perform all of Euclid's Algorithm q separate times to determine all of the g.c.d.'s. A short cut will be seen in the example.

so

$$Q = \begin{matrix} 100000000000 \\ 001000000000 \\ 000010000000 \\ 000000100000 \\ 000000001000 \\ 000000000010 \\ 111000111000 \\ 001110001110 \\ 111011011011 \\ 101010010010 \\ 110010011100 \\ 001100100111 \end{matrix} \quad \text{and} \quad Q - I = \begin{matrix} 000000000000 \\ 011000000000 \\ 001010000000 \\ 000100100000 \\ 000010001000 \\ 000001000010 \\ 111000011000 \\ 001110011110 \\ 111011010011 \\ 101010010110 \\ 110010011110 \\ 001100100110 \end{matrix}.$$

If we number the columns of $Q - I$ from 0 to 11, then the upper right quarter of the $Q - I$ matrix may be zeroed if we add the 3rd column to the 6th column, the 1st, 2nd, and 4th columns to the 8th column, and the 5th column to the 10th column. The lower right quarter of the $Q - I$ matrix then becomes

$$R = \begin{matrix} 011000 \\ 111110 \\ 011001 \\ 010110 \\ 011110 \\ 001110 \end{matrix}.$$

The equation $[g_6, g_7, \cdots, g_{11}]R = 0$ is found to have solutions $[g_6, g_7, \cdots, g_{11}] = [A, 0, 0, A, 0, A]$ where $A = 0$ or 1. The first six coordinants of g are then readily found from the equation $g(Q - I) = 0$, with solutions $g = [B, A, 0, A, A, 0, A, 0, 0, A, 0, A]$; $A, B \in GF(2)$. Finally, we apply Euclid's algorithm to $f(z) = 1110001110001$ and $g(z) = s10110100101$. By letting $t = s + 1$, and leaving s as an indeterminate, we may effectively find the g.c.d. of 111000111001 and 010110100101 with the same computation that computes the g.c.d. of 111000111001 and 110110100101:

$$\begin{array}{r}
1110001110001 \\
\underline{s10110100101} \\
1\,t001110101 \\
\underline{s10110100101} \\
s0\,t11101 \\
\underline{1\,t001110101} \\
1\,t0\,s1\,s01 \\
\underline{s0\,t11101}. \\
t\,t\,t\,t0\,t
\end{array}$$

If $t = 0$, the g.c.d. is 10011101; if $s = 0$, the g.c.d. of 1110001110001 and 010110100101 is equal to the g.c.d. of 111101 and 11001001, which is 111101. Both 111101 and 10011101 are irreducible and the factorization is complete:

$$\begin{aligned}(1 + z + z^2 + z^6 + z^7 + z^8 + z^{12})& \\ = (1 + z + z^2 + z^3 + z^5)&(1 + z^3 + z^4 + z^5 + z^7) \quad \text{over } GF(2).\end{aligned}$$

In general, suppose $f(z) = \prod_i (p^{(i)}(z))^{e_i}$, where each $p^{(i)}(z)$ is an irreducible polynomial over $GF(q)$. Then $f(z)$ divides

$$\prod_{s \in GF(q)} (g(z) - s)$$

if each $(p^{(i)}(z))^{e_i}$ divides $g(z) - s_i$ for some $s_i \in GF(q)$. On the other hand, given any set of scalars $s_1, s_2, \cdots, s_n \in GF(q)$, then the Chinese remainder theorem guarantees the existence of a unique $g(z) \bmod f(z)$ such that $g(z) \equiv s_i \bmod (p^{(i)}(z))^{e_i}$ for all i. Since there are q^n choices of $s_1, s_2, \cdots, s_n$, there are exactly q^n solutions of the equation $(g(z))^q - g(z) \equiv 0 \bmod f(z)$. Therefore,

The number of distinct irreducible factors of $f(z)$ is equal to the dimension of the null space of $(Q - I)$.

In particular, the polynomial $f(z)$ is the power of an irreducible polynomial iff the null space of $(Q - I)$ has dimension 1. In this case, the only solutions of $(g(z))^q - g(z) \equiv 0 \bmod f(z)$ are scalars in $GF(q)$, and the null space of $Q - I$ contains only vectors of the form $[s, 0, 0, \cdots, 0]$. If the null space of $Q - I$ has dimension n, it has a basis consisting of n monic polynomials: $g^{(1)}(z), g^{(2)}(z), \cdots, g^{(n)}(z)$. Without loss of generality, we may assume that $g^{(n)}(z) = 1$ and that the other $n - 1$ basis polynomials have positive degree.

When we apply Euclid's algorithm to $f(z)$ and $g^{(1)}(z) - s$, we obtain a factorization of $f(z)$. If this gives fewer than n factors of $f(z)$, then we may compute the g.c.d. of $g^{(2)}(z) - s$ and each known factor of $f(z)$. By this process, we may continue to refine the factorization of $f(z)$. The following argument shows that this process must eventually yield all n irreducible-powers which are factors of $f(z)$.

Let C be the $n \times n$ matrix over $GF(q)$ defined by the equations $g^{(j)}(z) \equiv C_{i,j} \bmod (p^{(i)}(z))^{e_i}$. Then C must be nonsingular. For if $\sum_j A_j C_{i,j} = 0$ for all i, then $\sum_j A_j g^{(j)}(z) \equiv 0 \bmod (p^{(i)}(z))^{e_i}$ for all i, whence $\sum_j A_j g^{(j)}(z) = 0$, contradicting the linear independence of $g^{(1)}(z), g^{(2)}(z), \cdots, g^{(n)}(z)$. When we apply Euclid's algorithm to $f(z)$ and $g^{(j)}(z) - s$, we obtain a factorization of $f(z)$ into as many different factors as there are distinct elements in the jth row of C. The irreducible-powers $(p^{(i)}(z))^{e_i}$ and $(p^{(k)}(z))^{e_k}$ are separated iff $C_{i,j} \neq C_{k,j}$. Since C is nonsingular, for every i and k there exists some j such that $C_{i,j} \neq C_{k,j}$. Thus, any two irreducible-power factors of $f(z)$ will be separated by some $g^{(i)}(z)$.

The factorization of any power of an irreducible polynomial is readily accomplished by applying Euclid's algorithm to the polynomial and its derivative.

We conclude with another example.

Example II: Following a suggestion of R. L. Graham, we let $f(z) = z^n - 1$ over $GF(q)$, where n and q are relatively prime. Then $Q_{i+1,j+1} = 1$ if $qi \equiv j \bmod n$. Specifically, suppose $n = 15$ and $q = 2$. Then

$$Q = \begin{array}{l}
100000000000000\\
001000000000000\\
000010000000000\\
000000100000000\\
000000001000000\\
000000000010000\\
000000000000100\\
000000000000001\\
010000000000000\\
000100000000000\\
000001000000000\\
000000010000000\\
000000000100000\\
000000000001000\\
000000000000010
\end{array}
\qquad
Q - I = \begin{array}{lr}
000000000000000 & 0\\
011000000000000 & 1\\
001010000000000 & 2\\
000100100000000 & 3\\
000010001000000 & 4\\
000001000010000 & 5\\
000000100000100 & 6\\
000000010000001 & 7.\\
010000001000000 & 8\\
000100000100000 & 9\\
000001000010000 & 10\\
000000010001000 & 11\\
000000000100100 & 12\\
000000000001010 & 13\\
000000000000011 & 14
\end{array}$$

By suitably permuting the rows and columns, we can bring $Q - I$ into the form

$$\begin{array}{c|c|c|c|cr}
0 & 0000 & 0000 & 0000 & 00 & 0\\ \hline
0 & 1100 & 0000 & 0000 & 00 & 1\\
0 & 0110 & 0000 & 0000 & 00 & 2\\
0 & 0011 & 0000 & 0000 & 00 & 4\\
0 & 1001 & 0000 & 0000 & 00 & 8\\ \hline
0 & 0000 & 1100 & 0000 & 00 & 7\\
0 & 0000 & 0110 & 0000 & 00 & 14\\
0 & 0000 & 0011 & 0000 & 00 & 13\\
0 & 0000 & 1001 & 0000 & 00 & 11\\ \hline
0 & 0000 & 0000 & 1100 & 00 & 3\\
0 & 0000 & 0000 & 0110 & 00 & 6\\
0 & 0000 & 0000 & 0011 & 00 & 12\\
0 & 0000 & 0000 & 1001 & 00 & 9\\ \hline
0 & 0000 & 0000 & 0000 & 11 & 5\\
0 & 0000 & 0000 & 0000 & 11 & 10
\end{array}$$

A basis for the null space of $Q - I$ is seen to be

$$\begin{aligned} g^{(1)}(z) &= z + z^2 + z^4 + z^8 \\ g^{(2)}(z) &= z^7 + z^{14} + z^{13} + z^{11} \\ g^{(3)}(z) &= z^3 + z^6 + z^{12} + z^9 \\ g^{(4)}(z) &= z^5 + z^{10}. \end{aligned}$$

In general, if $f(z) = z^n - 1$ over $GF(q)$, then we may choose

$$g(z) = \sum_{k \epsilon C} z^k,$$

where C is any set of numbers which is closed under multiplication by q mod n. Each such polynomial $g(z)$ has some nontrivial factor in common with $z^n - 1$.

REFERENCES

1. Brown, W. S., ALPAK System for Numerical Algebra on a Digital Computer-I:—Polynomials in Several Variables and Truncated Power Series with Polynomial Coefficients, B.S.T.J., *42*, September, 1963, pp. 2081–2119.
2. Johnson, S. C., Tricks for Improving Kronecker's Polynomial Factoring Algorithm, unpublished work.
3. Golomb, S. W., *Shift Register Sequences*, Holden-Day, Inc., 1967.
4. Bose, R. C. and Ray-Chaudhuri, D. K., On a Class of Error-Correcting Binary Group Codes, Inform. Control, *3*, 1960, pp. 68–79.
5. Hocquenghem, A., Codes Correcteurs D'Erreurs, Chiffres, *2*, 1959, pp. 147–156.
6. Peterson, W. W., *Error-Correcting Codes*, MIT Press-Wiley, New York, 1961.
7. MacWilliams, J., The Structure and Properties of Binary Cyclic Alphabets, B.S.T.J., *44*, February 1965, pp. 303–332.
8. Berlekamp, E. R., *Algebraic Coding Theory*, McGraw-Hill Book Company, Inc., New York, 1968.

The Enumeration of Information Symbols in BCH Codes

By E. R. BERLEKAMP

(Manuscript received May 9, 1967)

This paper presents certain formulas for $I(q, n, d)$, the number of information symbols in the q-ary Bose-Chaudhuri-Hocquenghem code of block length $n = q^m - 1$ and designed distance d. By appropriate manipulations on the m-digit q-ary representation of d, we derive a simple linear recurrence for a sequence whose mth term is the number of information symbols in the BCH code.

In addition to an exact solution of all finite cases, we obtain exact asymptotic results, as n and d go to infinity while their ratio n/d remains fixed. In this limit, the number of information symbols increases as n^s. Specifically, we show that for fixed u, $0 \leqq u \leqq 1$,

$$\lim_{m\to\infty} q^{-ms} I(q, q^m - 1, uq^m) = 1,$$

where s is a singular function of u. The function $s(u)$ is continuous and monotonic nonincreasing; it has derivative zero almost everywhere. Yet $s(0) = 1$ and $s(1) = 0$. For $q = 2$, $s(u)$ is plotted in Fig. 1.

Any cyclic code of block length n over $GF(q)$ may be defined by its generator polynomial, $g(x)$, which is some factor of $x^n - 1$ over $GF(q)$, or by its check polynomial, $h(x) = (x^n - 1)/g(x)$. The number of check digits in the code is given by the degree of $g(x)$; the number of information digits, by the degree of $h(x)$. We assume that n and q are relatively prime. It is most convenient to work in a particular extension field of $GF(q)$, namely $GF(q^m)$, where m is the multiplicative order of q mod n. In this field, $x^n - 1$ factors into distinct linear factors:

$$x^n - 1 = \prod_{i=1}^{n} (x - \alpha^i).$$

Here α is any primitive nth root of unity in $GF(q^m)$; $\alpha^n = 1$. From the factorization $x^n - 1 = g(x)h(x)$, we see that every power of α is a root

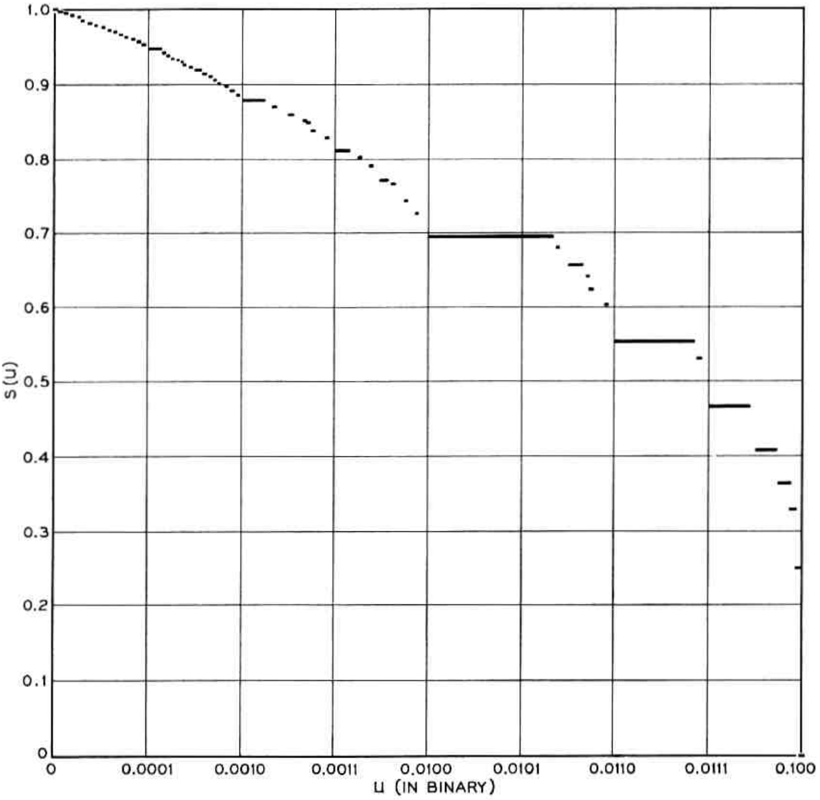


Fig. 1 — Graph of $s(u)$ vs. u.

of either $g(x)$ or of $h(x)$, but not both. Thus, a cyclic code partitions the powers of α into two sets: those powers which are roots of the generator polynomial, and those powers which are roots of the check polynomial. If $g(x)$ and $h(x)$ were permitted to have coefficients in $GF(q^m)$, then any partition of the powers of α would define a cyclic code. However, the coefficients of $g(x)$ and $h(x)$ must lie in the ground field $GF(q)$. Consequently, if α^i is a root of $g(x)$, then so are the conjugates of α^i, namely α^{iq}, α^{iq^2}, α^{iq^3}, $\cdots$. Conversely, if all conjugates of roots of $g(x)$ are also roots of $g(x)$, and all conjugates of roots of $h(x)$ are also roots of $h(x)$, then all of the coefficients of the polynomials $g(x)$ and $h(x)$ lie in $GF(q)$.

The previous remarks hold for all cyclic codes.

A q-ary BCH code of block length n over $GF(q)$ may be defined as the cyclic code whose generator's roots include only $\alpha, \alpha^2, \cdots, \alpha^{d-1}$ and their conjugates. This code is capable of correcting any combination of less than $d/2$ errors; (*cf.* Berlekamp[1]) the minimum Hamming distance of this code is at least d. For this reason, d is called the designed distance of the code.

The first result on the number of information symbols in BCH codes is the following lemma:

Classical Lemma I: Let $I(q, n, d)$ be the number of information symbols in the q-ary BCH code of block length n and designed distance d.

Define $\lceil i \rceil$ by the equations

$$i \equiv \lceil i \rceil \text{ mod } n \quad \text{and} \quad 1 \leqq \lceil i \rceil \leqq n.$$

Then $I(q, n, d)$ is the number of integers j, such that $1 \leqq j \leqq n$ and $\lceil jq^k \rceil \geqq d$ for all k.

Proof: α^i is a root of the generator polynomial of the BCH code iff there exists some $k(j)$ such that $\lceil jq^k \rceil < d$. Conversely, α^i is a root of the check polynomial iff $\lceil jq^k \rceil \geqq d$ for all k. Q.E.D.

The classical lemma enables one to compute the number of information symbols in any q-ary given BCH code without doing any calculations in $GF(q)$ or its extensions. One need only enumerate certain types of residue classes mod n. In practice, this enumeration is still often tedious, particularly when n and d are large.

In order to obtain more tractable results for large n and d, we prefer to start from an alternate form of the classical lemma:

Classical Lemma II: Let $I(q, n, d)$ be the number of information symbols in the q-ary BCH code of block length n and designed distance d.

Define $\lfloor i \rfloor$ by the equations

$$i \equiv \lfloor i \rfloor \text{ mod } n \quad \text{and} \quad 0 \leqq \lfloor i \rfloor \leqq n - 1.$$

Then, $I(q, n, d)$ is the number of integers i, such that $0 \leqq i \leqq n - 1$ and $\lfloor iq^k \rfloor < n + 1 - d$ for all k.

Proof: $1 \leqq j \leqq n$ and $\lceil jq^k \rceil \geqq d$ for all k iff $0 \leqq (n - j) \leqq n - 1$ and $\lfloor (n - j)q^k \rfloor \leqq n - d$ for all k. Let $i = n - j$. Q.E.D.

In the wide sense, BCH codes may be defined over any alphabet whose order, q, is a prime power, and for any block length, n, which is relatively prime to q. In the narrow sense, however, n is required to be one less than a power of q. For narrow sense codes, the smallest extension field of $GF(q)$ which contains the nth roots of unity is $GF(n+1)$. For wide sense codes, this extension field is always larger, usually much larger. Since the decoder must perform certain computations in this extension field, narrow sense BCH codes are more easily implemented than their more general wide sense counterparts.

We shall enumerate the information symbols in narrow sense BCH codes by reducing the problem to the enumeration of certain kinds of sequences over the alphabet consisting of the integers $0, 1, 2, \cdots, q-1$, as first suggested by Mann.[2] We begin by defining the appropriate manipulations with such sequences.

We shall always use *capital letters for sequences.* We let $(Q - 1)$ denote the sequence consisting of the single letter $q - 1$. Unless otherwise stated, we allow every sequence to be either finite or infinite.

Let $V = V_1V_2V_3 \cdots$ be any finite or infinite q-ary sequence (i.e., a sequence of numbers V_i, where V_i is an integer, $0 \leqq V_i \leqq q - 1$. We let $\bar{V} = \bar{V}_1\bar{V}_2\bar{V}_3 \cdots$ denote the *complement* of V, defined by $\bar{V}_i = (q-1) - V_i$ for all i. If $W = W_1W_2 \cdots W_k$ is a finite q-ary sequence, then we may form the *cyclic shifts* of W: $W_2W_3 \cdots W_kW_1$, $W_3W_4 \cdots W_kW_1W_2$, $\cdots$. If X is a finite q-ary sequence, $X = X_1X_2 \cdots X_j$, then we may form the *concatenation* $X * V = X_1X_2X_3 \cdots X_jV_1V_2V_3 \cdots$. This concatenation may be formed whenever V is a finite or infinite q-ary sequence. If V is a finite q-ary sequence, then $V * X$ is a cyclic shift of $X * V$.

The q-ary sequence Y is said to be a *prefix* of X iff $X = Y * Z$ for some Z; Y is said to be a *suffix* of X iff $X = Z * Y$ for some Z. A prefix must be a finite (or empty) sequence; a suffix may be empty, finite, or infinite. V is a *proper prefix* of X iff $X = V * Z$, and neither V nor Z is empty. Z is a *proper suffix* of X iff $X = V * Z$ and neither V nor Z is empty. If X is a finite q-ary sequence, $X = X_1X_2 \cdots X_k$, then we may form the *iterated concatenation* of X with itself, $\dot{X} = X_1X_2 \cdots X_kX_1X_2 \cdots X_k \cdots$. In particular, $(Q \dot{-} 1)$ denotes the infinite q-ary sequence all of whose letters are $q - 1$.

We say $X < Y$ iff there exists a j such that $X_i = Y_i$ for $i = 1, 2, \cdots, j-1$, but $X_j < Y_j$. If $X \not< Y$ and $Y \not< X$, then one is a prefix of the other.

This ordering is similar to the numerical ordering of q-ary fractions, but there are important differences. For example, $\frac{1}{4} = 0.01 < 0.0101 = 5/16$, but the sequences 01 and 0101 are incomparable, because one is a prefix of the other. On the other hand, $0.0111111 \cdots = 0.1 = \frac{1}{2}$, yet $01111 \cdots < 1$. This type of example may be excluded by writing all fractions in their terminating form if they have one. We may then assert the following:

Let

$$u = \sum_i U_i q^{-i}, \qquad v = \sum_i V_i q^{-i}, \qquad U = U_1U_2U_3 \cdots$$

and $V = V_1V_2V_3 \cdots$, *and suppose that* $(Q \dot{-} 1)$ *is not a suffix of* U *or* V. *Then*

$$U < V \Rightarrow u < v$$

$$u \leqq v \Leftrightarrow \begin{cases} U < V \\ \quad \textit{or} \\ U \textit{ is a prefix of } V. \end{cases}$$

We say that X is an *immediate subordinate* of Y iff X is a finite sequence, $X = X_1X_2X_3 \cdots X_k$, and $X_1 = Y_1$, $X_2 = Y_2$, $\cdots$, $X_{k-1} = Y_{k-1}$, but $X_k < Y_k$. The sequence Y has Y_1 immediate subordinates of length 1, Y_2 immediate subordinates of length 2, Y_3 immediate subordinates of length 3, $\cdots$ Y_k immediate subordinates of length k. If the sequence Y has only a finite number of nonzeros, then we may define the *greatest immediate subordinate* of Y. If the last nonzero in the sequence $Y = Y_1Y_2 \cdots$ is Y_k, then the greatest immediate subordinate of Y is $Y_1Y_2 \cdots Y_{k-1}(Y_k - 1)$. If the sequence Y contains an infinite number of nonzeros, then Y has infinitely many immediate subordinates. All of them are less than Y itself, but none of them is the greatest immediate subordinate.

Similarly, we say that Y is an *immediate superior* of X iff $Y = Y_1Y_2Y_3 \cdots Y_k$, where $Y_1 = X_1$, $Y_2 = X_2$, $\cdots$, $Y_{k-1} = X_{k-1}$ but $Y_k > X_k$. If $X = X_1X_2 \cdots X_k$ and $X_k \neq (Q - 1)$, then the *least immediate superior* of X is $Y = Y_1Y_2 \cdots Y_k$; $Y_i = X_i$ for $i = 1, 2, \cdots, k - 1$, and $Y_k = X_k + 1$. It should be evident that the least immediate superior is among the longest immediate superiors, and the greatest immediate subordinate is among the longest immediate subordinates.

Definition: If q is any integer, U is any infinite q-ary sequence and m is any integer, we define $J(q, U, m)$ to be the number of q-ary m-tuples all of whose cyclic shifts are less than U.

Lemma III: (Complemented form of Mann's Lemma)

If

$$n = q^m - 1, \qquad n + 1 - d = \sum_{i=1}^{m} U_i q^{m-i}, \qquad 0 \leqq U_i < q,$$

$U = U_1U_2 \cdots U_m$, *and* Y *is any* q*-ary sequence then*

$$I(q, n, d) = J(q, U * Y, m).$$

Proof: Lemma III reduces to Lemma II under the following correspondence: The q-ary m-tuple U corresponds to the integer $n + 1 - d$; another q-ary m-tuple $W = W_1W_2 \cdots W_m$ corresponds to the integer $w = \sum_{i=1}^{m} W_iq^{m-i}$. The first cyclic shift of W is the sequence $W_2W_3 \cdots W_mW_1$, which then corresponds to the integer

$$\sum_{i=1}^{m-1} W_iq^{m+1-i} + W_1 = qw - (q^m - 1)W_1 .$$

Modulo $n = q^m - 1$, the integer corresponding to the first cyclic shift of W is seen to be congruent to qw. Therefore, the successive cyclic shifts of an m-digit q-ary sequence W correspond to the integers $\lfloor w \rfloor$, $\lfloor wq \rfloor$, $\lfloor wq^2 \rfloor$, $\cdots$, $\lfloor wq^{m-1} \rfloor$. These integers are all $<n + 1 - d$ iff all cyclic shifts of W are $<U$, which is true iff all cyclic shifts of W are $<U * Y$, for any Y. Q.E.D.

The choice $Y = \dot{U}$ has an interesting interpretation:

$$\sum_{i=1}^{\infty} \dot{U}_iq^{-i} = \sum_{k=0}^{\infty} \sum_{i=1}^{m} U_iq^{-(i+mk)} = \left(\sum_{i=1}^{m} U_iq^{-i}\right)/(1 - q^{-m})$$
$$= \left(\sum_{i=1}^{m} U_iq^{m-i}\right)/(q^m - 1) = 1 - \frac{(d-1)}{n}.$$

Thus, the sequence $\dot{U}$ is the q-ary expansion of $1 - (d - 1)/n$. For this reason, we may investigate the behavior of $I(q, n, d)$ for large n and d with a fixed fractional error correction capability, $(d - 1)/2n$, by studying $J(q, \dot{U}, m)$ as a function of m for fixed q and $\dot{U}$.

We shall temporarily ignore the periodicity of the $\dot{U}$ sequence, and consider the function $J(q, V, m)$ for an arbitrary q-ary sequence V. We assume only that the sequence V has no terminal zeros.

From the definition of the immediate subordinates of V, it is clear that *if an* m*-digit* q*-ary sequence* W *is less than* V, *then some immediate subordinate of* V *is a prefix of* W. For if W is less than V, then there exists a k such that $W_i = V_i$ for $i = 1, 2, \cdots, k - 1$, but $W_k < V_k$, and the sequence $W_1W_2 \cdots W_k$ is a prefix of W and an immediate subordinate of V.

Now suppose that some m-digit sequence W and all of its cyclic shifts are less than V. Since W itself is less than V, some prefix of W must be an immediate subordinate of V. Are all possible immediate subordinates of V possible prefixes of W? In general, they are not, for some immediate subordinates may have suffixes which are greater than V. If $X * Y$ is an immediate subordinate of V and Y is greater than V, then $X * Y$ cannot be a prefix of W. For, if $W = X * Y * Z$, then one of the cyclic shifts of W is $Y * Z * X$ which is greater than V.

For example, consider the ternary sequence $V = 20212$. Its immediate subordinates are 0, 1, 200, 201, 2020, 20210, and 20211. The immediate subordinate 20210 has the suffix 210 which is greater than V. Therefore, if 20210 is the prefix of W, then the second cyclic left shift of W is greater than V. Similarly, V's immediate subordinate 20211 has the suffix 211, which is also greater than V.

For some sequences V, this difficulty does not arise. If V exceeds all of its own proper suffixes, then we have the following theorem:

Theorem 1: Let V be a q-ary sequence which exceeds all of its own proper suffixes. Then:

(i) No immediate subordinate of V is a proper prefix of any other immediate subordinate of V.

(ii) Every suffix of every immediate subordinate of V is a concatenation of other immediate subordinates of V.

(iii) If W and all of its cyclic shifts are less than V, then W can be uniquely decomposed into a concatenation of immediate subordinates of V, including a (possibly empty) end-around immediate subordinate. Specifically $W = W^{(1)} * W^{(2)} * \cdots * W^{(i)} * W^{(i+1)} * W^{(i+2)} * \cdots * W^{(i-1)} * W^{(i)}$; $W^{(1)}, W^{(2)}, \cdots, W^{(i-1)}$ are immediate subordinates of V; $W^{(i)} * W^{(1)} * W^{(2)} * \cdots * W^{(i)}$ is the end-around immediate subordinate. The end-around immediate subordinate has a prefix, $W^{(i)}$, which is a suffix of W, and a suffix $W^{(1)} * W^{(2)} * \cdots * W^{(i)}$ which is a prefix of W, as well as a concatenation of the shorter immediate superiors $W^{(1)}, W^{(2)}, \cdots, W^{(i)}$.

(iv) Every concatenation of immediate subordinates of V, including a (possibly empty) end-around immediate subordinate yields a sequence which has the property that all of its cyclic shifts are less than V. No such sequence of length m can exceed the maximum m-digit concatenation of immediate subordinates of V. If Y is the maximum m-digit concatenation of immediate subordinates of V, and $Y \leqq U \leqq V$, then $J(q, V, m) = J(q, U, m)$.

(v)

$$J(q, V, m) = mV_m + \sum_{k=1}^{m-1} V_k J(q, V, m-k),$$

where V_j is taken as 0 if j exceeds the length of the sequence V.

(*vi*) *Let*

$$n = q^m - 1, \qquad d = \Sigma\, D_i q^{m-i}, \qquad 0 \leqq D_i < q,$$
$$D = D_1 D_2 \cdots D_m .$$

If

$\bar{V} * (Q \dotminus 1) < D \leqq$

least m-digit concatenation of immediate superiors of V,

then,

$$I(q, n, d) = J(q, V, m).$$

Proofs:

(*i*) This property of immediate subordinates does not even depend on the suffix condition on V. From the definition of immediate subordinates, each immediate subordinate must disagree with V only in the immediate subordinate's last digit, and hence no immediate subordinate can be a prefix of any other.

(*ii*) Let us first prove the weaker assertion:

(*a*) Every proper suffix of every immediate subordinate of V has a prefix which is a shorter immediate subordinate of V.

Let S be an immediate subordinate of V, and let $S^{(2)}$ be a suffix of S. We may write $S = S^{(1)} * S^{(2)}$. Since S differs from V only in its last digit, $S^{(1)}$ is a prefix of V, and $V = S^{(1)} * V^{(2)}$. Since $S < V$, $S^{(2)} < V^{(2)}$. Since V exceeds all of its own proper suffixes, $V^{(2)} < V$. Therefore, $S^{(2)} < V$. Therefore, some prefix of $S^{(2)}$ is an immediate subordinate of V.

(*b*) If every suffix of an immediate subordinate has a prefix which is an immediate subordinate, then every suffix of every immediate subordinate is a concatenation of immediate subordinates.

For, suppose F is a suffix on an immediate subordinate, then $F = B^{(1)} * F^{(2)}$, where $B^{(1)}$ is an immediate subordinate. Since $F^{(2)}$ is a suffix of F, it is also a suffix of an immediate subordinate, and $F^{(2)} = B^{(2)} * F^{(3)}$, where $B^{(2)}$ is an immediate subordinate $\cdots$ $F = B^{(1)} * B^{(2)} * B^{(3)} * \cdots$.

(*iii*) Since $W < V$, it contains a prefix $W^{(1)}$ which is an immediate subordinate of V. After shifting this prefix around to the end, we may similarly identify $W^{(2)}, W^{(3)}, \cdots, W^{(i-1)}$, each of which is an immediate subordinate of V. The sequence $W^{(i)} * W^{(1)} * W^{(2)} * \cdots * W^{(i-1)}$ is a cyclic shift of W, and so it must have a prefix, P, which is an immediate subordinate of V. P is not a prefix of $W^{(i)}$, so $W^{(i)}$ must be a prefix of P. Suppose that $W^{(i)} * W^{(1)} * \cdots * W^{(i)}$ is a prefix of P, but that $W^{(i)} * W^{(1)} * \cdots * W^{(i+1)}$ is not a prefix of P. (This defines i.) Then $P = W^{(i)} * W^{(1)} * \cdots * W^{(i)} * S$, where the (possibly empty) sequence S is a proper prefix of $W^{(i+1)}$. Since S is a suffix of P, which is an immediate subordinate of V, S is itself a concatenation of immediate subordinates of V. But no immediate subordinate of V is a proper prefix of any other immediate subordinate of V, so S must be empty.

(*iv*) This is the converse of (*iii*). Suppose we are given the sequence $W = S^{(i)} * W^{(1)} * W^{(2)} * \cdots * W^{(i-1)} * P^{(i)}$, where $W^{(1)}, W^{(2)}, \cdots, W^{(i-1)}$ and $W^{(i)} = P^{(i)} * S^{(i)}$ are immediate subordinates of V. We must show that all cyclic shifts of W are less than V. Any cyclic shift is of the form $C = S^{(k)} * W^{(k+1)} * W^{(k+2)} * \cdots * W^{(i)} * W^{(1)} * W^{(2)} * \cdots * W^{(k+1)} * P^{(k)}$, where $W^{(k)} = P^{(k)} * S^{(k)}$. If $S^{(k)}$ is empty, C has the prefix $W^{(k+1)}$, which is an immediate subordinate of V. If $S^{(k)}$ is not empty, by (*ii*) it has a prefix which is an immediate subordinate of V, which is a prefix of C. In either case, C has a prefix which is an immediate subordinate of V. Therefore, $C < V$.

(*v*) V has V_m immediate subordinates of length m, each of which has m distinct cyclic shifts. Thus, W may be chosen as a single end-around immediate subordinate of V in mV_m ways.

If W is a concatenation of several immediate subordinates of V, $W = W^{(1)} * W^{(2)} * \cdots * W^{(i-1)} * W^{(i)}$ where $W^{(1)}, W^{(2)}, \cdots, W^{(i-1)}$ are immediate subordinates of V and $W^{(i)}$ is a (possibly empty) proper prefix of the immediate subordinate $W^{(i)} * W^{(1)} * \cdots * W^{(i)}$, then the length of W is the length of $W^{(i-1)}$ plus the length of $W^{(1)} * W^{(2)} * \cdots * W^{(i-2)} * W^{(i)}$. For each k, there are V_k choices of $W^{(i-1)}$ of length k, and $J(q, V, m - k)$ choices for $W^{(1)} * W^{(2)} * \cdots * W^{(i-2)} * W^{(i)}$.

(*vi*) *Least special case:* Suppose D is the least m-digit q-ary sequence greater than $\bar{V} * (Q \dot{-} 1)$. Letting

$$d = \Sigma D_i q^{m-i}, \qquad v = \Sigma V_i q^{m-i}, \qquad \bar{v} = \Sigma \bar{V}_i q^{m-i},$$

it is evident that $d + v = n + 1$ and $v = n + 1 - d$. According to Lemma III, $I(q, n, d) = J(q, V, m)$.

(*vi*) *Greatest special case:* Let D be the least m-digit concatenation of immediate superiors of $\bar{V}$. Complementing, $\bar{D}$ is the greatest m-digit concatenation of immediate subordinates of V. In the notation of part (*iv*), $\bar{D} = Y$. Letting $\bar{d} = \Sigma \bar{D}_i q^{m-i}$, $\bar{d} = n - d$. Letting $n + 1 - d = \Sigma U_i q^{m-i}$, $U > Y$ because $n + 1 - d > n - d$. Theorem follows from part (*iv*) and Lemma III.

(*vi*) *The general case* follows because $J(q, U, m)$ is a monotonic function of U. Q.E.D.

Example I: Let V be the binary sequence 1101. We compute

		Bose distance†		Designed distance	
m	$J(q, V, m)$	*Binary*	*Decimal*	*Binary*	*Decimal*
1	1	1	1	1	1
2	3	01	1	01	1
3	4	011	3	010	2
4	11	0011	3‡	0011‡	3
5	16	00111	7	00110	6
6	30	001101	13	001100	12
7	50	0011011	27	0011000	24
8	91	00110011	51	00110000	48
9	157	001100111	103	001100000	96
10	278	0011001101	205	0011000000	192
11	485	00110011011	411	00110000000	384
12	854	001100110011	819	001100000000	768
			⋮		⋮

Here $J(q, V, m)$ is computed by Theorem 1*v*. The designed distances are computed according to Theorem 1*vi*, using $\bar{V} = 0010$, with immediate superiors 1, 01, and 0011. $\bar{V} * (Q \dot{-} 1) = 001011111111 \cdots$.

Evidently, the binary BCH code of block length $2^{12} - 1$ and designed distance 768 is identical to the binary BCH code of block length $2^{12} - 1$ and designed distance 769 or 770 or $\cdots$ or 819. This code has 854 information symbols. This code is distinct from the binary BCH code of block length $2^{12} - 1$ and designed distance 820. This is true in general, because the least m-digit concatenation of immediate superiors of $\bar{V}$ is necessarily minimum among all of its own cyclic shifts. This "greatest designed distance" is called the *Bose distance.*

It happens that the binary BCH code of block length $2^{12} - 1$ and designed distance 768 is also distinct from the binary BCH code of block length $2^{12} - 1$ and designed distance 767, because the 12-digit binary expansion of 767 is minimum among all of its cyclic shifts. This, however, need not be true in general. For example, the binary BCH code of block length $2^4 - 1$ and designed distance 3 is not distinct from the binary BCH code of block length $2^4 - 1$ and designed distance 2, because the 4-digit binary expansion of 2 = 0010 is not minimum among its cyclic shifts; the minimum is 0001.

† Defined later in the text.

‡ This code is identical to the binary BCH code of block length 15 and designed distance 2.

In general, we would like to determine the number of information digits in the q-ary BCH code of block length $n = q^m - 1$ and designed distance $d = \Sigma D_i q^{m-i}$. The previous theorem gives us a solution to this problem if we can find a sequence V which is greater than all of its own suffixes and has the property that

$$\bar{V} * (Q \dot{-} 1) < D \leqq \text{least } m\text{-digit concatenation of immediate superiors of } \bar{V}.$$

Complementing this condition gives

$$V > \bar{D} \geqq \text{greatest } m\text{-digit concatenation of immediate subordinates of } V.$$

or

$$V > \bar{D} * (Q \dot{-} 1) > \begin{pmatrix} \text{greatest } m\text{-digit concatenation} \\ \text{of immediate subordinates of } V \end{pmatrix} * \dot{0} > \dot{X},$$

where X is the greatest immediate subordinate of V. We may assume that V has no terminal zeros, and that the length of V does not exceed the length of $\bar{D}$. Since X and V have the same length, X is a prefix of $\bar{D}$.

Since V is the least immediate superior of X, the problem of finding V is reduced to the problem of finding X, which is a prefix of $\bar{D}$. The solution is as follows:

Theorem 2: *Let X be the shortest prefix of $\bar{D}$ such that*

$$\bar{D} = X * F, \qquad F * (Q \dot{-} 1) \geqq \bar{D} * (Q \dot{-} 1),$$

and let V be the least immediate superior of X. Then

(*i*) $\bar{V} * (Q \dot{-} 1) < D \leqq$ *least m-digit concatenation of immediate superiors of $\bar{V}$.*

(*ii*) *V exceeds all of its own proper suffixes.*

Proof of (*i*):

Since X is a prefix of $\bar{D}$ and V is an immediate superior of X, V is an immediate superior of $\bar{D}$. So $V > \bar{D}$ and $V > \bar{D} * (Q \dot{-} 1)$.

Complementing gives $\bar{V} * (Q \dot{-} 1) < D * \dot{0}$, so $V * (Q \dot{-} 1) < D$.

Let $X^{(k)} = \overset{\longleftarrow\!\!\longrightarrow k \longrightarrow}{X * X * \cdots * X}$. Then $F * (Q \dot{-} 1) \geqq \bar{D} * (Q \dot{-} 1)$ is equivalent to $X^{(0)} * F * (Q \dot{-} 1) \geqq X^{(1)} * F * (Q \dot{-} 1)$. Therefore, $X * X^{(0)} * F * (Q \dot{-} 1) \geqq X * X^{(1)} * F * (Q \dot{-} 1)$ or $X^{(1)} * F * (Q \dot{-} 1) \geqq$

$X^{(2)} * F * (Q \dot{-} 1)$. By induction, $X^{(k)} * F * (Q \dot{-} 1) \geqq X^{(k+1)} * F * (Q \dot{-} 1)$ and $\bar{D} * (Q \dot{-} 1) \geqq X^{(k)} * F * (Q \dot{-} 1)$ for all k. Since this is true for arbitrarily large k, $\bar{D} * (Q \dot{-} 1) \geqq \dot{X}$. Complementing, $D * \dot{0} \leqq \dot{\bar{X}} \leqq$ any infinite concatenation of immediate superiors of $\bar{V}$. Therefore, $D \leqq$ any m-digit concatenation of immediate superiors of $\bar{V}$.

Proof of (ii):

Let $X = Y * Z * L$, where Y and Z are arbitrary (possibly empty) and L is the final digit of X. We have

$$V = Y * Z * (L + 1)$$

$$\bar{D} = Y * Z * L * F$$

$$F * (Q \dot{-} 1) \geqq Y * Z * L * F * (Q \dot{-} 1) = X * F * (Q \dot{-} 1);$$

$$X * F * (Q \dot{-} 1) = Y * Z * L * F * (Q \dot{-} 1) > Z * L * F * (Q \dot{-} 1),$$

else Y would be a shorter prefix than X which satisfied the same conditions.

No proper suffix of V can equal V, for the suffix must be shorter.

If some proper prefix of V, say $Z * (L + 1)$, (Z possibly empty) exceeds V, then

$$Z * (L + 1) > Y * Z * (L + 1) > Y * Z * L = X.$$

If $Z * L > X$, then $Z * L * F * (Q \dot{-} 1) > X * F * (Q \dot{-} 1)$, a contradiction. If $Z * L$ is a prefix of X, then $X = Z * L * G$ and from

$$X * F * (Q \dot{-} 1) > Z * L * F * (Q \dot{-} 1)$$

we have

$$Z * L * G * F * (Q \dot{-} 1) > Z * L * F * (Q \dot{-} 1)$$

$$G * F * (Q \dot{-} 1) > F * (Q \dot{-} 1) \geqq X * F * (Q \dot{-} 1).$$

Now $Z * L$ is a shorter prefix than X, a contradiction. Therefore, $Z * (L + 1) < Y * Z * (L + 1)$, i.e., V exceeds all of its own proper suffixes. Q.E.D.

Example II: Let $q = 9$, $n = 728 = 9^3 - 1$, $d = 217$. Then $D = 261$, $\bar{D} = 627$, $\bar{D} * \dot{8} = 627888 \cdots$, $X = 62$, $V = 63$, $\bar{V} = 25$.

		Bose distance	
m	*J*†	*q-ary*	*decimal*
1	6	3	3
2	42	26	24
3	270	263	219

† In this and subsequent tables we use the single later J as an abbreviation for $J(q, V, m)$.

Example III: Let $q = 2$, $n = 511$, $d = 185$. Then $D = 010111001$, $\bar{D} = 101000110$, $\bar{D} * 1 = 101000110111111 \cdots$, $X = 101000$, $V = 101001$, $\bar{V} = 010110$. Immediate superiors are 010111, 011, 1.

m	*J*	*Bose distance*	*Smaller designed distance*
1	1	1	1
2	1	11	11
3	4	011	011
4	5	0111	0111
5	6	01111	01111
6	16	010111	010111
7	22	0101111	0101110
8	29	01011111	01011100
9	49	010111011	010111000

Example IV: Let $q = 2$, $n = 511 = 2^9 - 1$, $d = 187$. Then $D = 010111011$, $\bar{D} = 101000100 = X$, $V = 101000101$, $\bar{V} = 010111010$.

m	*J*	*Bose distance*
1	1	1
2	1	11
3	4	011
4	5	0111
5	6	01111
6	10	011011
7	22	0101111
8	29	01011111
9	49	010111011

The answer agrees with Example III, although the recurrence is different. This illustrates the general nonuniqueness of V. Theorem 2 specifies one satisfactory method of finding V, but as seen from this example, this V need not be unique. The simplest recurrence rule generally arises from the shortest possible V, which corresponds to the greatest V, or the least D. This can generally be found by first reducing D insofar as permissible.

Example V: Let $n = 2^{11} - 1$, $d = 411$. Then $D = 00110011011$. We could take $\bar{D} = 11001100100 = X$ and proceed. However, we instead consider $d = 410$, $D = 00110011010$. Since D has a cyclic shift smaller than itself, the code is unchanged. But $\bar{D} = 11001100101$, $X = 1100110010$ does not look much easier, so we continue. Each prime marks the starting point of a smaller cyclic shift.

D

0011001101′0
00110011′001
00110011′000
0011′0010111
0011′0010110
⋮
0011′0010000
0011′0001111
⋮
0011′0000000
00101111111

Since 00101111111 has no cyclic shift less than itself, this designed distance is the Bose distance of a *different* BCH code. We must instead use D = 00110000000, $\bar{D}$ = 11001111111, X = 1100, V = 1101. The recurrence is given in Example I: $I(2,\ 2^{11} - 1{,}411) = 485$. This same V is obtained if we started with D = 00110011000, or any D in the region

$$00110000000 \leqq D \leqq 00110011000$$

Example VI: Let $q = 2$, $n = 2^{15} - 1$, D = 001010010100111, $\bar{D} * \dot{1}$ = 110101101011000111···, X=110101101011000, V=110101101011001, $\bar{V}$ = 001010010100110

m	J	*Bose distance*
1	1	1
2	3	01
3	4	011
4	11	0011
5	16	00111
6	36	001011
7	64	0010101
8	115	00101011
9	211	001010011
10	378	0010100111
11	694	00101001011
12	1256	001010010101
13	2276	0010100101011
14	4112	00101001010111
15	7474	001010010100111

Although the brute force method just used gives the right answer, a more devious approach proves easier. Instead of $D = 001010010100111$, let us consider $D = 001010010100101$, $X = 11010$, $V = 11011$, $\bar{V} = 00100$. This yields a different set of codes, with a much simpler recurrence:

m	J	*Bose distance*
1	1	1
2	3	01
3	4	011
4	11	0011
5	21	00101
6	36	001011
7	64	0010101
8	115	00101011
9	211	001010011
10	383	0010100101
11	694	00101001011
12	1256	001010010101
13	2276	0010100101011
14	4126	00101001010011
15	7479	001010010100101

The code with $D = 001010010100111$ has 5 less information digits than the code with $D = 001010010100101$, corresponding to the 5 distinct cyclic shifts of 001010010100101.

I. ASYMPTOTIC RESULTS

Let us define the enumerator

$$J(q, U; z) = \sum_{m=1}^{\infty} J(q, U, m) z^m.$$

Given a sequence V which is less than all of its own proper suffixes, we may also define

$$V(z) = \sum_k v_k z^k,$$

so that

$$zV'(z) = \sum_k k V_k z^k.$$

The recurrence

$$J(q, V, m) = mV_m + \sum_{k=1}^{m-1} V_k J(q, V, m - k)$$

becomes

$$J(q,\, V;z) = zV'(z) + V(z)J(q,\, V;z)$$

whose solution is

$$J(q,\, V;z) = \frac{zV'(z)}{1 - V(z)}.$$

Let ρ_1, ρ_2, $\cdots$ be the (not necessarily distinct) complex reciprocal roots of $1 - V(z)$. Then

$$1 - V(z) = \prod_i (1 - \rho_i z)$$

$$-V'(z) = -\sum_i \rho_i \prod_{j \neq i} (1 - \rho_j z)$$

$$\frac{zV'(z)}{1 - V(z)} = \sum_i \frac{\rho_i z}{1 - \rho_i z} = \sum_i \sum_{m=1}^{\infty} (\rho_i z)^m = \sum_{m=1}^{\infty} \sum_i \rho_i^m z^m.$$

Therefore,

$$J(q,\, V;z) = \sum_{m=1}^{\infty} \sum_i \rho_i^m z^m$$

so

$$J(q,\, V,\, m) = \sum_i \rho_i^m,$$

where ρ_i are the complex numbers defined by the equation

$$1 - V(z) = \prod_i (1 - \rho_i z).$$

Although this gives an explicit expression for $J(q, V, m)$, the expression depends upon the complex numbers ρ_i. For finite values of m, it is usually easier to compute $J(q, V, m)$ directly from the recurrence relation of the previous section, since these calculations involve only integers. For asymptotic results, however, the above equation is very useful.

Definition: Let $\rho = \max_i |\rho_i|$, let $s = \log_q \rho$.

Since all coefficients of the polynomial $V(z)$ must be nonnegative integers not exceeding $q - 1$, it is easily seen that the ρ_i with the maximum absolute value is real and positive, and $1 \leqq \rho \leqq q$. Clearly,

$$J(q,\, V,\, m) \approx \rho^m$$

for large m, in the sense that

$$\lim_{m\to\infty} \rho^{-m} J(q, V, m) = 1.$$

Similarly,

$$\log_q J(q, V, m) \approx m \log_q \rho = ms.$$

If

$$u = \sum_{i=1}^{\infty} U_i q^{-i},$$

and

$$\dot{X} \leqq \bar{U} \leqq V * \dot{0},$$

where V exceeds all of its own suffixes and X is the maximum subordinate of V, then

$$I(q, q^m - 1, uq^m) \approx q^{ms}.$$

In other words, if we fix the fraction $d/n = u$ and let n and d grow large, then

$$I \approx n^s$$

or, more precisely,

$$s(u) = \lim_{m\to\infty} \frac{\log_q I(q, q^m - 1, uq^m)}{m}.$$

For given q, the function $s(u)$ is a rather complicated animal. To compute it, one must first write u in q-ary. If $\bar{U}$ exceeds all of its proper suffixes, set $V = \bar{U}$; otherwise write $\bar{U} = X * F$ where X is the shortest prefix such that $\bar{U} \leqq F$. V is then taken to be the least immediate superior of X. Then, s is defined as the logarithm (base q) of the maximum reciprocal root of $1 - V(z)$.

It may easily be shown that s is a continuous, monotonic nonincreasing function of u. It may also be shown that the derivative of s with respect to u is either 0 or it is undefined. There are two kinds of points at which the derivative $s'(u)$ is undefined. First, there are the endpoints of the intervals on which $s(u)$ is constant. u is a lower endpoint of such an interval iff $\bar{U}$ is a finite sequence which exceeds all of its own proper suffixes; u is an upper endpoint of such an interval iff $\bar{U}$ is a periodic sequence, equal to some of its suffixes but not less than any others. At these endpoints, $s(u)$ is undifferentiable because it has only a right derivative or a left derivative, but not both. There is only a countable number of points of this type.

The more interesting points are those at which $s(u)$ has neither a right derivative nor a left derivative. This happens iff $\bar{U}$ is an infinite sequence which exceeds all of its own proper suffixes, and $\dot{0}$ is not a suffix of $\bar{U}$.

The set of points u such that $s(u)$ is not differentiable is uncountable, but it has measure 0. Professor T. Pitcher of the University of Southern California has also shown[3] that this set has Hausdorf dimension 1. This appears to be entirely due to the large density of these points in the vicinity of $u = 0$. In general, I conjecture that the Hausdorf dimension of the set of points in the interval $a \leqq u \leqq b$ [where $0 \leqq a$, $b \leqq 1$, $s(a) \neq s(b)$] is $s(a)$. In some sense, almost all of the nondifferentiable points in any interval seem to lie very near the leftmost cluster point of the interval.

When Mann[2] first obtained results identical to those here in the special cases $u = q^{-k}$, he also showed that ρ is the only reciprocal root of $1 - V(z)$ with magnitude greater than 1. Thus, not only is

$$I \approx \rho^m,$$

but in fact, for sufficiently large m,

$$I = \langle \rho^m \rangle,$$

where $\langle \cdot \rangle$ denotes the nearest integer to "$\cdot$". Unfortunately, this strengthened result is not true in general. For some values of u, $1 - V(z)$ has only one reciprocal root with magnitude greater than 1, but for other values of u, $1 - V(z)$ has many reciprocal roots with magnitude greater than 1. Little is known about the behavior of the smaller complex reciprocal roots of $1 - V(z)$ as a function of u, although B. F. Logan[4] has obtained a few preliminary results in this area.

II. ACTUAL DISTANCE

As one increases the designed distance, the number of information symbols in the resulting code must either remain constant or decrease. Thus,

$I(q, n, d)$ is the maximum number of information symbols in any of the q-ary BCH codes with designed distance $\geqq d$.

We must be careful to distinguish between I and $\hat{I}$, defined by

$\hat{I}(q, n, d)$ is the maximum number of information symbols in any of the q-ary BCH codes with actual distance $\geqq d$.

It is obvious that $\hat{I}(q, n, d) \geqq I(q, n, d)$.

For example, there are three binary BCH codes of block length 23, having 23, 12, and 1 information symbols. The code with 23 information digits has Bose distance = actual distance = 1, but the code with 12 information digits has Bose distance 5, actual distance 7. The code with 1 information digit has Bose distance = actual distance = 23. Therefore, $I(2, 23, 6) = I(2, 23, 7) = 1$, but $\hat{I}(2, 23, 6) = \hat{I}(2, 23, 7) = 12$. For all values of $d \neq 6$ or 7, $I(2, 23, d) = \hat{I}(2, 23, d)$.

The known cases in which $\hat{I}(q, n, d) > I(q, n, d)$ are relatively sparse. Peterson, Kasami, and Lin[5] and Berlekamp[6] have investigated this question for narrow sense binary BCH codes (where $q = 2$ and $n = 2^m - 1$). They proved that $\hat{I}(2, 2^m - 1, d) = I(2, 2^m - 1, d)$ if d divides $2^m - 1$, or if d is one less than a power of 2, or if m' divides m and $\hat{I}(2, 2^{m'} - 1, d) = I(2, 2^{m'} - 1, d) > I(2, 2^{m'} - 1, d + 1)$, or if m is sufficiently small, or if d is sufficiently small, or if d and/or m satisfy any of various other number theoretical constraints. More recently, Peterson and Lin[7] have shown that if $\hat{I}(2, 2^m - 1, d) = I(2, 2^m - 1, d) > I(2, 2^m - 1, d + 1)$, and $1 \leqq j \leqq m - d$ then $\hat{I}(2, 2^m - 1, 2^j d + 2^j - 1) = I(2, 2^m - 1, 2^j d + 2^j - 1)$. No examples are known in which $\hat{I}(2, 2^m - 1, d) > I(2, 2^m - 1, d)$, and it has been conjectured that $\hat{I}(2, 2^m - 1, d) = I(2, 2^m - 1, d)$ for all m and d.

Although this conjecture remains open, we can obtain certain results about the asymptotic behavior of $I(2, 2^m - 1, u2^m)$ from the known classes of special cases in which $\hat{I}(2, 2^m - 1, d) = I(2, 2^m - 1, d)$. We would like to define

$$\hat{s}(u) \overset{?}{=} \lim_{m \to \infty} \frac{\log_2 \hat{I}(2, 2^m - 1, u2^m)}{m}.$$

Unfortunately, however, we have no assurance that the limit exists. In order to discuss the asymptotic behavior of the best BCH codes, we define

$$\hat{s}(u) = \limsup_{m \to \infty} \frac{\log_2 \hat{I}(2, 2^m - 1, u2^m)}{m}.$$

Clearly $\hat{s}(u) \geqq s(u)$. Like $s(u)$, $\hat{s}(u)$ must be a monotonic nonincreasing function of u, because if $d' > d$, the codewords of the q-ary BCH code of distance d' are a subset of the codewords of the q-ary BCH code of distance d.

We can prove that $\hat{s}(u) = s(u)$ for certain values of u, as indicated by the following theorem:

If $u \geqq 2^{-k}$, *then* $\hat{s}(u) \leqq s(2^{-k})$

Proof: We know that if $u \geqq 2^{-k}$, then

$$\hat{I}(2, 2^{m-1}, u2^m) \leqq \hat{I}(2, 2^m - 1, 2^{m-k} - 1)$$
$$= I(2, 2^m - 1, 2^{m-k} - 1) \qquad m \geqq k$$

Hence,

$$\frac{\log \hat{I}(2, 2^m - 1, u2^m)}{m} \leqq \frac{\log I(2, 2^m - 1, 2^{m-k} - 1)}{m}.$$

So

$$\hat{s}(u) \leqq \lim_{m\to\infty} \frac{\log I(2, 2^m - 1, 2^{m-k} - 1)}{m} = s(2^{-k})$$

because $s(u)$ is continuous. Q.E.D.

This shows that $\hat{s}(u) = s(u)$ if $u = \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \cdots$. Similarly, one can show from the recent theorem of Peterson and Lin that $\hat{s}(u) = s(u)$ for certain other values of u.

We conjecture that $\hat{s}(u) = s(u)$ for all u. This is a weakened form of Peterson's conjecture that $\hat{I}(2, 2^m - 1, d) = I(2, 2^m - 1, d)$ for all m and d.

REFERENCES

1. Berlekamp, E. R., *Algebraic Coding Theory,* McGraw-Hill Book Company, Inc., New York, 1968.
2. Mann, H. P., On the Number of Information Symbols in Bose-Chaudhuri Codes, Inform. Control *5,* 1962, pp. 153–162.
3. Pitcher, T., unpublished correspondence, 1966.
4. Logan, B. F., unpublished oral communication, 1966.
5. Kasami, J., Lin, S., and Peterson, W. W., Some Results on Weight Distributions of BCH Codes, presented at PGIT Symposium at UCLA, January, 1966.
6. Berlekamp, E. R., Practical BCH Decoders, unpublished work, 1966.
7. Peterson, W. W. and Lin, S., Some New Results on Finite Fields and Their Applications to the Theory of BCH Codes, presented at Conference on Combinatorial Mathematics and Its Applications at the University of North Carolina, April, 1967.
8. Peterson, W. W., *Error Correcting Codes,* MIT Press, 1961.

Equations Governing the Electrical Behavior of an Arbitrary Piezoelectric Resonator Having *N* Electrodes*

By P. LLOYD

(Manuscript received May 24, 1967)

In a paper by J. A. Lewis (B.S.T.J., 40, 1961, pp. 1259-1280) general formulas for the electrical admittance of a piezoelectric resonator, having essentially one pair of electrodes, were derived in terms of motional parameters associated with the normal modes of vibration of the device. The logical extension of this work to a resonator with N electrodes is presented here. Expressions are given for both the admittance and impedance matrices of the resonator. These matrices are expressed in terms of motional parameters associated with, respectively, (i) the normal modes of vibration with all electrodes connected together, and (ii) all electrodes left open circuited. The electrical equivalent circuit for the 2-port characteristics of the N electrode resonator is given for two particular examples.

I. INTRODUCTION

General formulas for the electrical admittance of a piezoelectric resonator having essentially one pair of electrodes were derived by Lewis.[1] These formulas are consistent with those derived earlier for special cases such as long bars and large plates (see, for example, Mason[2]). In Lewis' work the admittance function is expanded about its poles in an infinite series. The residue at one of these poles determines the strength of the contribution of the normal mode, associated with the pole, to the overall vibrational behavior of the resonator when it is driven at a frequency close to the natural frequency of the mode. Surprisingly, the work of Lewis seems to have seen little application, as far as can be judged, except for that of Lloyd and Redwood,[3] and Byrne, et al.[4]

* Most of the work described here is based on part of the author's doctoral thesis (University of London, 1966).

With the current interest in multi-electroded resonators, such as the monolithic crystal filter,[5] it is pertinent to consider the logical extension of the work of Lewis to the case of an arbitrary resonator having N electrodes. A discussion of this problem has previously been presented by the author,[6] and also by EerNisse and Holland.[7]

Included in Section II of this paper are the basic equations governing the piezoelectric resonator, presented here for completeness.

In Section III various integral relations are derived for use in Section IV where the properties of the admittance and impedance matrices are investigated. The electrical equivalent circuits for two particular 2-port configurations of the N electrode resonator are also derived in Section IV, in order to illustrate the application of the admittance and impedance matrices.

A brief list of the principal symbols used in the text is given below.

1.1 *List of Symbols*

ρ	Mass per unit volume.
ρ_s	Mass per unit area of an electrode.
u_i	Particle displacement vector.
S_{kl}	Strain tensor.
T_{kl}	Stress tensor.
τ_i	Traction (stress vector).
ϕ	Electric scalar potential.
E_i	Electric field vector.
D_i	Electric displacement vector.
c_{ijkl}^E	Elastic stiffness tensor (measured at constant electric field).
e_{nij}	Piezoelectric constant tensor.
ϵ_{mn}^S	Dielectric constant tensor (measured at constant strain).
n_i	Unit vector normal to, and outwards from surface of body.
Φ_p	Electric potential on the pth electrode.
Q_p	Total charge on the pth electrode.
B	Volume of the body.
A	Unelectroded area of the body.
A_p	Area of the pth electrode.
ω	Angular frequency.
$\lambda \equiv \omega^2$.	

The tensor components above are referred to orthogonal Cartesian coordinate axes x_i. The comma notation is used to indicate differentiation, e.g. $D_{i,j} = \partial D_i/\partial x_j$, and the repeated index summation convention is used, e.g., $D_{i,i} = D_{1,1} + D_{2,2} + D_{3,3}$.

II. BASIC EQUATIONS OF A PIEZOELECTRIC RESONATOR

The equations describing the steady vibrations of a piezoelectric body are listed below.

The equations of motion:

$$\rho\lambda u_i + T_{ij,j} = 0. \tag{1}$$

The divergence equation of electrostatics (for an insulator):

$$D_{i,i} = 0. \tag{2}$$

The piezoelectric constitutive relations:

$$T_{ij} = c^E_{ijkl} S_{kl} - e_{nij} E_n \,, \tag{3}$$

$$D_m = e_{mkl} S_{kl} + \epsilon^S_{mn} E_n \,, \tag{4}$$

where

$$S_{kl} = \tfrac{1}{2}(u_{k,l} + u_{l,k}), \tag{5}$$

and

$$E_n = -\phi_{2_n} \,. \tag{6}$$

The symmetry relations

$$c^E_{ijkl} = c^E_{ijlk} = c^E_{jikl} = c^E_{klij} \,, \tag{7}$$

$$e_{nij} = e_{nji} \,, \tag{8}$$

$$\epsilon^S_{mn} = \epsilon^S_{nm} \,. \tag{9}$$

2.1 *Boundary Conditions*

The boundary conditions for the resonator shown in Fig. 1 will now be discussed.

On the unelectroded portion of the surface A

$$\tau_i = 0, \qquad \text{on } A, \tag{10}$$

$$D_i n_i = \epsilon_0 (E_i) \text{ ext } n_i = 0, \qquad \text{on } A, \tag{11}$$

that is, no surface tractions and zero external electric field exist normal to the surface. Note that (11) is an approximation which in practice is usually valid for materials with large values of $\epsilon^S_{nm}/\epsilon_0$. The driving electrodes are assumed to be very thin metallic conductors with infinite conductivity. Potential Φ_p and charge Q_p exist on the electrode area A_p. External electrical connections to the electrodes will not be specified at present. We pause to note, however, that $\phi = 0$ at some point ex-

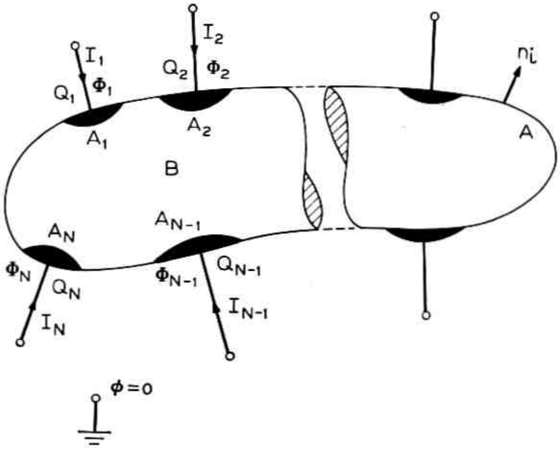


Fig. 1 — Arbitrary piezoelectric resonator with N electrodes.

ternal to the resonator. Since we have neglected the effects of the external potential distribution, this "earth" point only has significance in connection with the topography of the external electrical circuit. The latter is assumed to interact only with the currents I_p and potentials ϕ_p on the electrodes of the resonator. The mechanical properties of the electrode are assumed here to be nonexistent except for a surface mass density ρ_s . The surface of the resonator beneath an electrode is, therefore, assumed responsible only for exerting a force consistent with maintaining the acceleration of the electrode. The boundary conditions at the electrode can therefore be written as

$$\tau_i = \rho_s \lambda u_i \,, \qquad \text{on } A_p \,, \tag{12}$$

$$\phi = \text{constant}, \qquad \text{on } A_p \tag{13a}$$

and either

$$\phi = \Phi_p \qquad \text{or} \tag{13b}$$

$$\int_{A_p} D_i n_i \, dA_p = -Q_p \,. \tag{13c}$$

Note that the choice between (13b) or (13c) as a primary condition is unimportant.

III. PROPERTIES OF SOLUTIONS

As is well known, the solution of the equations reviewed in Section II for a practical case is a formidable problem, and it is often necessary to resort to some approximate method of solution. In this paper, we will not discuss the methods for solving the equations, but rather the nature of the solution assuming that it has been found.

Equations (1) through (13) can be expressed in terms of u_i and ϕ, from which it follows that

$$\rho\lambda u_i + c^E_{ijkl} u_{k,jl} + e_{nij}\phi_{,nj} = 0, \tag{14}$$

$$e_{nkj} u_{k,jn} - \epsilon^S_{nm}\phi_{,nm} = 0, \tag{15}$$

subject to the boundary conditions

$$c^E_{ijkl} u_{k,l} n_j + e_{kij}\phi_{,k} n_j = 0, \qquad \text{on } A, \tag{16}$$

$$e_{jkl} u_{k,l} n_j - \epsilon^S_{jk}\phi_{,k} n_j = 0, \qquad \text{on } A, \tag{17}$$

with

$$\phi = \text{constant on } A_p \tag{18a}$$

and either

$$\phi = \Phi_p \qquad \text{on } A_p \tag{18b}$$

or

$$\int_{A_p} D_i n_i \, dA_p = -Q_p \,, \qquad \text{on } A_p \tag{18c}$$

and

$$c^E_{ijkl} u_{k,l} n_j + e_{kij}\phi_{,k} n_j = \rho_s \lambda u_i \,, \qquad \text{on } A_p \,. \tag{19}$$

We now note that (14) through (19) become homogeneous when $\Phi_p = 0$ for all p. This latter condition represents one of the eigenvalue problems associated with Fig. 1, namely, that of mechanical vibrations possible when all electrodes are connected directly to the reference point. Other eigenvalue problems associated with Fig. 1 include those where some electrodes are open-circuited ($Q_p = 0$) and the remainder are short-circuited ($\phi_r = 0$) (i.e., connected to the reference point).

3.1 *Reciprocal Theorem*

Consider two solutions of (1) through (13) denoted, respectively, by (λ', u_i', ϕ') and $(\lambda'', u_i'', \phi'')$. The two solutions could be, for example, those associated with two different sets of forcing parameters at different frequencies.

From (1) we have

$$\int_B \rho\lambda'' u_i' u_i'' \, dB + \int_B u_i' T''_{ij,i} \, dB = 0, \tag{20}$$

and from (2)

$$\int_B \phi' D''_{i,i}\, dB = 0, \tag{21}$$

where B is volume of the body exclusive of the electrodes. Using the divergence theorem, (20) may be written

$$\int_B \rho\lambda'' u'_i u''_i\, dB - \int_B S'_{ij} T''_{ij}\, dB + \int_A u'_i T''_{ij} n_j\, dA = 0, \tag{22}$$

and (21)

$$-\int_B \phi'_{,i} D''_i\, dB + \int_A \phi' D''_i n_i\, dA = 0. \tag{23}$$

Subtracting (23) from (22), and substituting for the surface conditions given by equations (10) through (13) we have

$$\int_B \rho\lambda'' u'_i u''_i\, dB + \sum_{p=1}^{N} \int_{A_p} \rho_s \lambda'' u'_i u''_i\, dA - \sum_{p=1}^{N} \Phi'_p Q''_p = \int_B (T''_{ij} S'_{ij} - E'_i D''_i)\, dB. \tag{24}$$

Equation (24) is still valid when the primed and double-primed quantities are interchanged. Using this fact, we have

$$(\lambda'' - \lambda')\left[\int_B \rho u'_i u''_i\, dB + \sum_{p=1}^{N} \int_{A_p} \rho_s u'_i u''_i\, dA\right] - \sum_{p=1}^{N} (\Phi'_p Q''_p - \Phi''_p Q'_p) = \int_B [(T''_{ij} S'_{ij} - E'_i D''_i) - (T'_{ij} S''_{ij} - E''_i D'_i)]\, dB. \tag{25}$$

The quantity on the right-hand side is zero by virtue of the constitutive equations (3) and (4).

Equation (25) then becomes

$$(\lambda'' - \lambda') V(u'_i u''_i) = \sum_{p=1}^{N} (\Phi'_p Q''_p - \Phi''_p Q'_p), \tag{26}$$

where

$$V(u'_i u''_i) = \int_B \rho u'_i u''_i\, dB + \sum_{p=1}^{N} \int_{A_p} \rho_s u'_i u''_i\, dA. \tag{27}$$

Equation (26) is a special case of the reciprocal theorem given by Lewis[1] and discussed by Love[8] for the purely elastic case.

3.2 *Orthogonality of the Eigensolutions*

Consider two eigensolutions, $(\lambda^{(n)}, u_i^{(n)}, \phi^{(n)})$ and $(\lambda^{(m)}, u_i^{(m)}, \phi^{(m)})$ of the same homogeneous boundary problem. That is, $\Phi_p^{(m)} = 0$ if $\Phi_p^{(n)} = 0$ and $Q_r^{(m)} = 0$ if $Q_r^{(n)} = 0$. Thus, for two solutions of the same eigenset

$$\sum_{p=1}^{N} \left(\Phi_p^{(m)} Q_p^{(n)} - \Phi_p^{(n)} Q_p^{(m)} \right) = 0, \tag{28}$$

and from (26)

$$V(u_i^{(n)} u_i^{(m)}) = 0, \qquad \lambda^{(n)} \neq \lambda^{(m)}. \tag{29}$$

Thus, two solutions of the same eigenset satisfy the orthogonality condition (29). Also we have from (24), the Rayleigh quotient for the eigenvalue $\lambda^{(n)}$

$$\lambda^{(n)} = \omega_n^2 = \frac{2 \int_B H(u_i^{(n)}, \phi^{(n)})\, dB}{V(u_i^{(n)} u_i^{(n)})}, \tag{30}$$

where

$$H(u_i, \phi) = \tfrac{1}{2}(T_{ij} S_{ij} - E_i D_i). \tag{31}$$

3.3 *Expansion in Terms of Eigensolutions*

The solution to the inhomogeneous boundary value problem indicated by Fig. 1 can be expanded in terms of any of the sets of eigensolutions. These expansions are very important when electrical behavior is of prime interest. We will show here how the forced vibrational solution (λ, u_i, ϕ) may be expressed in terms of two of the possible expansions, namely: (*i*) the eigensolutions $(\lambda^{S(n)}, u_i^{S(n)}, \phi^{S(n)})$ which correspond to the normal modes of vibration of the resonator with all its electrodes connected to the reference point, and (*ii*) the eigensolutions $(\lambda^{O(n)}, u_i^{O(n)}, \phi^{O(n)})$ for the normal modes with all the electrodes open circuited.

For expansion (*i*) we set

$$u_i = u_i^{(o)} + \sum_{n=1}^{\infty} a^{(n)} u_i^{S(n)} \tag{32}$$

and

$$\phi = \phi^{(o)} + \sum_{n=1}^{\infty} a^{(n)} \phi^{S(n)}; \tag{33}$$

and for expansion (*ii*)

$$u_i = u_i^{(o)} + \sum_{n=1}^{\infty} b^{(n)} u_i^{O(n)} \tag{34}$$

and

$$\phi = \phi^{(o)} + \sum_{n=1}^{\infty} b^{(n)} \phi^{O(n)}, \tag{35}$$

where $(u_i^{(o)}, \phi^{(o)})$ is the solution to the boundary value problem of (1) through (13), as $\lambda \to 0$.

Since we have not specified the means by which the electrodes are connected to the external electric circuit we will allow the parameters Φ_p and Q_p to be of the general form

$$\Phi_p = \Phi_p^{(o)} \exp (j\omega t), \qquad Q_p = Q_p^{(o)} \exp (j\omega t), \tag{36}$$

$$\Phi_p^{(o)} = | \Phi_p^{(o)} | \exp (j\theta_p), \qquad Q_p^{(o)} = | Q_p^{(o)} | \exp (j\psi_p). \tag{37}$$

Although it is immaterial how the charges Q_p and potentials Φ_p are set up in relation to the external circuit, Q_p and Φ_p are of course not independent.

The coefficients $a^{(n)}$ in the first expansion can be found by noting that the equations of motion (1) and boundary conditions (12) require

$$\rho\lambda u_i^{(o)} = \rho \sum_{n=1}^{\infty} (\lambda^{S(n)} - \lambda) a^{(n)} u_i^{S(n)}, \qquad \text{in } B \tag{38}$$

and

$$\rho_s\lambda u_i^{(o)} = \rho_s \sum_{n=1}^{\infty} (\lambda^{S(n)} - \lambda) a^{(n)} u_i^{S(n)}, \qquad \text{on } A_p\,. \tag{39}$$

On multiplying (38) and (39) by $u_i^{S(m)}$ and carrying out the indicated integrations and adding we have:

$$\int_B \rho\lambda u_i^{(o)} u_i^{S(m)}\, dB + \sum_{p=1}^{N} \int_{A_p} \rho_s \lambda u_i^{(o)} u_i^{S(m)}\, dA_p$$
$$= \sum_{n=1}^{\infty} a^{(n)} (\lambda^{S(n)} - \lambda) \left[\int_B \rho u_i^{S(n)} u_i^{S(m)}\, dB + \sum_{p=1}^{N} \int_{A_p} \rho_s u_i^{S(n)} u_i^{S(m)}\, dA_p \right]. \tag{40}$$

We note from (27) that (40) may be written

$$\lambda V(u_i^{(o)} u_i^{S(m)}) = \sum_{n=1}^{\infty} a^{(n)} (\lambda^{S(n)} - \lambda) V(u_i^{S(n)} u_i^{S(m)}). \tag{41}$$

From the orthogonality condition (29), all terms on the right are zero except the term in $a^{(m)}$, giving

$$a^{(m)} = \frac{\lambda}{\lambda^{S(m)} - \lambda} \frac{V(u_i^{(o)} u_i^{S(m)})}{V(u_i^{S(m)} u_i^{S(m)})}. \tag{42}$$

By a similar argument we have for the coefficients in the second expansion

$$b^{(m)} = \frac{\lambda}{(\lambda^{O(m)} - \lambda)} \frac{V(u_i^{(o)} u_i^{O(m)})}{V(u_i^{O(m)} u_i^{O(m)})}. \tag{43}$$

Remembering the definition of $u_i^{(o)}$, $u_i^{S(m)}$ and $u_i^{O(m)}$ we have the following as a consequence of the reciprocal theorem (26):

$$\lambda^{S(m)} V(u_i^{(o)} u_i^{S(m)}) = \sum_{p=1}^{N} \Phi_p^{(o)} Q_p^{S(m)}, \tag{44}$$

and

$$\lambda^{O(m)} V(u_i^{(o)} u_i^{O(m)}) = -\sum_{p=1}^{N} Q_p^{(o)} \Phi_p^{O(m)}, \tag{45}$$

since

$$\Phi_p^{S(m)} = 0 \quad \text{and} \quad Q_p^{O(m)} = 0.$$

Using (44) and (45) with (42) and (43)

$$a^{(m)} = \frac{\lambda \sum_{p=1}^{N} \Phi_p^{(o)} Q_p^{S(m)}}{(\lambda^{S(m)} - \lambda)\lambda^{S(m)} V(u_i^{S(m)} u_i^{S(m)})} \tag{46}$$

$$b^{(m)} = \frac{-\lambda \sum_{p=1}^{N} Q_p^{(o)} \Phi_p^{O(m)}}{(\lambda^{O(m)} - \lambda)\lambda^{O(m)} V(u_i^{O(m)} u_i^{O(m)})}. \tag{47}$$

From (46), we see immediately that the contribution of the $S(m)$th mode (eigensolution) in the expansion (32) and (33) is dominant when $\lambda \to \lambda^{S(m)}$, if $\Phi_p^{(o)}$ is held constant with frequency. We also note that the amplitude of $a^{(m)}$ depends on the charge on the electrodes when the resonator is executing free vibrations corresponding to the $S(m)$th mode (i.e., with all electrodes short-circuited).

Equation (47) shows similarly that the contribution of the $O(m)$th mode is dominant in the expansion of (34) and (35) when $\lambda \to \lambda^{O(m)}$ if $Q_p^{(o)}$ is held constant. Also the amplitude $b^{(m)}$ depends on the potentials on the electrodes when they are open circuit with the resonator executing free vibrations corresponding to the $O(m)$th mode. When applying the expansions (*i*) and (*ii*) it should be realized that assumptions have been made concerning the completeness of the eigensets.

IV. THE ELECTRICAL ADMITTANCE AND IMPEDANCE MATRICES

4.1 *General Considerations*

The admittance matrix y_{pq} for the N-electrode piezoelectric resonator is defined by

$$I_p = \sum_{q=1}^{N} y_{pq}\Phi_q \,. \tag{48}$$

Similarly z_{pq}, the impedance matrix, is here defined by

$$\Phi_p = \sum_{q=1}^{N} z_{pq}I_q + R, \tag{49}$$

where

$$I_p = j\omega Q_p \tag{50}$$

and R is a constant depending on the external circuit configuration. The relationships (48) and (49) are postulated on the basis that the equations of the resonator are linear and that their use is restricted to steady vibrations.

We will now derive various properties of y_{pq} and z_{pq}. First we note from (13c) and the divergence theorem that

$$\sum_{p=1}^{N} I_p = j\omega \sum_{p=1}^{N} Q_p = -j\omega \sum_{p=1}^{N} \int_{A_p} D_i n_i \, dA_p = -j\omega \int_B D_{i,i} \, dB = 0. \tag{51}$$

Equation (51) is simply Kirchhoff's current law, for the conservation of charge. We note from (48) and (51) that

$$\sum_{p=1}^{N} \sum_{q=1}^{N} y_{pq}\Phi_q = 0, \tag{52}$$

and since Φ_q is arbitrary, the sum of each column of the y_{pq} matrix is zero, i.e.,

$$\sum_{p=1}^{N} y_{pq} = 0. \tag{53}$$

We will now use the reciprocal theorem to show that both y_{pq} and z_{pq} are symmetric. Consider the solutions for two sets of potentials Φ_p'

and Φ_p'' having the same frequency. Then from (26)

$$\sum_{p=1}^{N} (\Phi_p' Q_p'' - \Phi_p'' Q_p') = 0. \tag{54}$$

Using (48) and (54)

$$\sum_{p=1}^{N} \sum_{q=1}^{N} (y_{pq} \Phi_p' \Phi_q'' - y_{pq} \Phi_p'' \Phi_q') = 0, \tag{55}$$

and (49) and (54)

$$j\omega \sum_{p=1}^{N} \sum_{q=1}^{N} (z_{pq} Q_q' Q_p'' - z_{pq} Q_q'' Q_p') + R \sum_{p=1}^{N} (Q_p' - Q_p'') = 0. \tag{56}$$

Since Φ_p' and Φ_p'' are arbitrary in (55) we must have

$$y_{pq} = y_{qp} \, . \tag{57}$$

In (56) Q_p' and Q_p' are arbitrary and $\sum_{p=1}^{N} Q_p = 0$, so

$$z_{pq} = z_{qp} \, . \tag{58}$$

As a consequence of (57) and (53)

$$\sum_{q=1}^{N} y_{pq} = 0. \tag{59}$$

It has been shown in this section that the impedance and admittance matrices of an N-electrode piezoelectric resonator have properties similar in many ways to those of an N-terminal passive electrical network.[9]

4.2 *Expansions for the Electrical Parameters*

We may now make use of the eigensolution expansions of (32)–(33) and (34)–(35) to inquire into the admissible forms for y_{pq} and z_{pq} as functions of frequency.

For expansion (a)

$$Q_p = Q_p^{(o)} + \sum_{m=1}^{\infty} a^{(m)} Q_p^{S(m)} \tag{60}$$

$$\Phi_p = \Phi_p^{(o)} + \sum_{m=1}^{\infty} a^{(m)} \Phi_p^{S(m)} = \Phi_p^{(o)}, \quad \text{all } \Phi_p^{S(m)} \text{ being zero.} \tag{61}$$

Then using (46) and (60)

$$Q_p = Q_p^{(o)} + \sum_{q=1}^{N} \sum_{m=1}^{\infty} \frac{\lambda C_{pq}^{(m)} \Phi_q^{(o)}}{(\lambda^{S(m)} - \lambda)}, \tag{62}$$

where

$$C_{pq}^{(m)} = \frac{Q_p^{S(m)} Q_q^{S(m)}}{\lambda^{S(m)} V(u_i^{S(m)} u_i^{S(m)})}. \tag{63}$$

For expansion (b)

$$Q_p = Q_p^{(o)} + \sum_{m=1}^{\infty} b^{(m)} Q_p^{O(m)} = Q_p^{(o)}, \quad \text{all } Q_p^{O(m)} \text{ being zero,} \tag{64}$$

and

$$\Phi_p = \Phi_p^{(o)} + \sum_{m=1}^{\infty} b^{(m)} \Phi_p^{O(m)}. \tag{65}$$

So using (47) and (65)

$$\Phi_p = \Phi_p^{(o)} - \sum_{q=1}^{N} \sum_{m=1}^{\infty} \frac{\lambda F_{pq}^{(m)} Q_q^{(o)}}{(\lambda^{O(m)} - \lambda)}, \tag{66}$$

where

$$F_{pq}^{(m)} = \frac{\Phi_p^{O(m)} \Phi_q^{O(m)}}{\lambda^{O(m)} V(u_i^{O(m)} u_i^{O(m)})}. \tag{67}$$

We now define the charge-potential relations for the solution of the static boundary value problem ($\lambda = 0$) as follows:

$$Q_p^{(o)} = \sum_{q=1}^{N} C'_{pq} \Phi_q^{(o)} \tag{68}$$

and

$$\Phi_p^{(o)} = \sum_{q=1}^{N} F_{pq}^{(o)} Q_q^{(o)} + R. \tag{69}$$

It is assumed, from now on, that the static parameters such as C'_{pq} are such that quadratic forms like $C'_{pq}\Phi_p\Phi_q$ are positive definite. Proof of this depends on energetic considerations.

We now put

$$\lambda^{O(m)} = \omega_{Am}^2 \quad \text{and} \quad \lambda^{S(m)} = \omega_{Rm}^2, \tag{70}$$

and define

$$C_{pq}^{(o)} = C'_{pq} - \sum_{m=1}^{\infty} C_{pq}^{(m)}. \tag{71}$$

We obtain from (62) and (71) the admittance matrix of (48) in the form

$$y_{pq} = j\omega\left\{C_{pq}^{(o)} + \sum_{m=1}^{\infty} \frac{\omega_{Rm}^2 C_{pq}^{(m)}}{(\omega_{Rm}^2 - \omega^2)}\right\}. \tag{72}$$

From (66) and (69) the impedance matrix of (49) is of the form

$$z_{pq} = \frac{1}{j\omega}\left\{F_{pq}^{(o)} - \sum_{m=1}^{\infty} \frac{\omega^2 F_{pq}^{(m)}}{(\omega_{Am}^2 - \omega^2)}\right\}. \tag{73}$$

Restricting our interest for the moment to an element y_{pq} of the admittance matrix, we observe that the form of (72) is analogous in form to the admittance of the electrical network in Fig. 2. However, from (59), which is valid for all frequencies, we require

$$\sum_{q=1}^{N} C_{pq}^{(m)} = 0, \tag{74}$$

but from (63)

$$C_{pp}^{(m)} > 0, \tag{75}$$

so

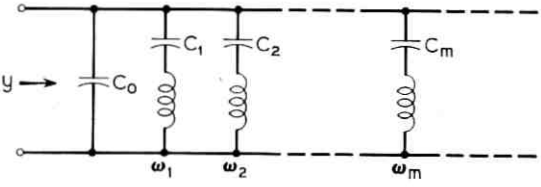
$$\sum_{q=1,\,q\neq p}^{N} C_{pq}^{(m)} < 0. \tag{76}$$

Therefore, several elements of $C_{pq}^{(m)}$ may be negative.

The form of an element of the impedance matrix (72) suggests an analogy with the circuit of Fig. 3 but, in view of the proceeding discussion, it is again probable that several of the elements $F_{pq}^{(m)}$ are negative. It should be noted, however, that $F_{pp}^{(m)} > 0$.

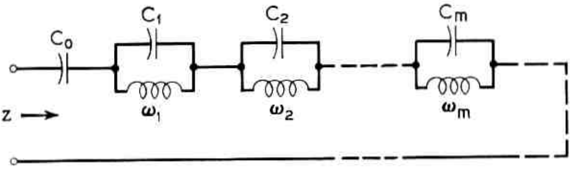
4.3 *Driving Point Functions*

From (49) the driving point impedance Z_{pq}^{D}, at the two terminals p-q, when all other terminals are left open-circuit is given by



FOR y_{pq}, $C_0 \equiv C_{pq}^{0}$, $C_m \equiv C_{pq}^{(m)}$ AND $\omega_m \equiv \omega_{Rm}$

Fig. 2 — Electrical network for admittance representation.



FOR z_{pq}, $C_0 \equiv 1/F_{pq}^{(o)}$, $C_m \equiv 1/F_{pq}^{(m)}$ AND $\omega_m \equiv \omega_{Am}$

Fig. 3 — Electrical network for impedance representation.

$$\begin{aligned} Z_{pq}^D &= (\Phi_p - \Phi_q)/I_p \\ &= z_{pp} - 2z_{pq} + z_{qq} \,. \end{aligned} \tag{77}$$

In terms of the expansions for z_{pq} given by (73), we have from (77)

$$Z_{pq}^D = \frac{1}{j\omega}\left[F^{(o)} - \sum_{m=1}^{\infty} \frac{\omega^2 F^{(m)}}{\omega_{Am}^2 - \omega^2}\right], \tag{78}$$

where

$$F^{(m)} = \frac{(\Phi_p^{O(m)} - \Phi_q^{O(m)})^2}{\lambda^{O(m)} V(u_i^{O(m)} u_i^{O(m)})} \tag{79}$$

and

$$F^{(o)} = F_{pp}^{(o)} - 2F_{pq}^{(o)} + F_{qq}^{(o)}. \tag{80}$$

So, clearly $F^{(m)} > 0$ and therefore, the analogue circuit of Fig. 3 is "physical" for Z_{pq}^D.

At first sight, it would appear that one could easily derive a driving point admittance for the *p-q port* when all other terminals are shorted. In fact, it must be found by appropriate manipulation of either y_{pq} or z_{pq} and, in general, the expression includes many elements of either matrix. We can, however, define a driving point admittance for the *p-terminal*, when all other terminals are connected to the reference point, i.e.,

$$Y_p^D = y_{pp} \,. \tag{81}$$

The analogue electrical circuit for y_{pp}, namely Fig. 2, is again physical.

4.4 *"Black Box" Matrices for a Two-Port*

Before calculating any "black box" transfer matrices it is convenient to define a transformed admittance matrix valid for the resonator and its external circuit. In Fig. 4, the terminals of the resonator are all interconnected, there being a physical component with admittance y_{pq}^{E} connected between terminals p and q. The currents flowing into the N-terminal network and resonator as a whole are

$$I'_p = \sum_{q=1}^{N} y'_{pq}\Phi_q \,, \tag{82}$$

where

$$y'_{pq} = y_{pq} - y_{pq}^{E} \,, \tag{83}$$

and

$$y_{pp}^{E} = - \sum_{q=1,\, q\neq p}^{N} y_{pq}^{E} \,. \tag{84}$$

y_{pq} is defined by (48) and $y_{pq}^{E} = y_{qp}^{E}$, as can be seen from Fig. 4. We may now form two-port networks. For the purposes of further discussion, any connections made externally to the two-port will be assumed to be consistent with

$$I'_s = -I'_p \,, \quad I'_r = -I'_q \,, \quad V_p = \Phi_p - \Phi_s \,, \quad V_q = \Phi_q - \Phi_r \,. \tag{85}$$

4.5 *Electrically Symmetric Two-Port Resonator*

Further discussion, with all y_{pq}^{E} finite for the N-terminal resonator, will not be continued. The reduction of a N-terminal network to a 2-port is discussed by Weinberg.[7]

4.5.1 *Two-Port With N-2 Terminals Shorted*

We will now consider a simple case where all terminals except p and q are connected directly to s, and an admittance y_{pq}^{E} is connected

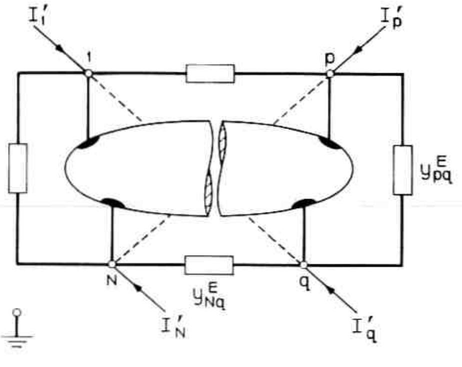


Fig. 4 — External electrical connections to the resonator.

between p and q as shown in Fig. 5. We will also assume that the construction of the resonator is such that it is electrically symmetric with respect to the ports. We have under these circumstances

$$I'_p = y'_{pp}V_p + y'_{pq}V_q \tag{86}$$

$$I'_q = y'_{pq}V_p + y'_{pp}V_q \,, \tag{87}$$

where

$$y'_{pp} = y_{pp} + y^E_{pq} \tag{88}$$

and

$$y'_{pq} = y_{pq} - y^E_{pq} \,. \tag{89}$$

We now consider the electrical lattice network of Fig. 6 as an analogue of transfer characteristics of (86) and (87). The analogue (Fig. 6) is physical if Y_a and Y_b are realizable with physical components.

Now

$$Y_a = y_{pp} - y_{pq} \,, \tag{90}$$

and

$$Y_b = y_{pp} + y_{pq} \,. \tag{91}$$

Using (72) and (63), Y_a and Y_b can be expressed in terms of the eigensolution expansion as follows:

$$Y_a = j\omega\left[(C^{(o)}_{pp} - C^{(o)}_{pq}) + \sum_{m=1}^{\infty} \frac{(Q_p^{S(m)}Q_p^{S(m)} - Q_p^{S(m)}Q_q^{S(m)})}{(\lambda^{S(m)} - \lambda)V(u_i^{S(m)}u_i^{S(m)})}\right], \tag{92}$$

$$Y_b = j\omega\left[(C^{(o)}_{pp} + C^{(o)}_{pq}) + \sum_{m=1}^{\infty} \frac{(Q_p^{S(m)}Q_p^{S(m)} + Q_p^{S(m)}Q_q^{S(m)})}{(\lambda^{S(m)} - \lambda)V(u_i^{S(m)}u_i^{S(m)})}\right]. \tag{93}$$

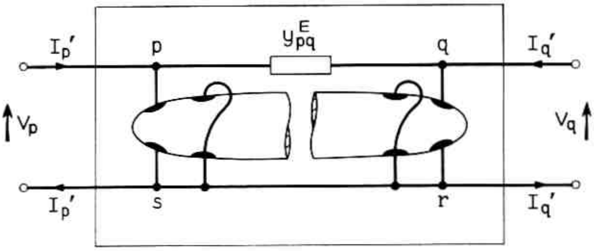


Fig. 5 — Two-port system for N-2 electrodes short-circuited.

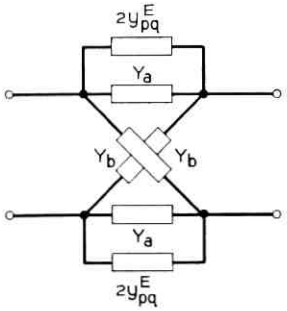


Fig. 6 — Electrical analogue of Fig. 5.

Also since we have taken $y_{pp} = y_{qq}$, then from (63), either

$$Q_p^{S(m)} = -Q_q^{S(m)} \quad \text{or} \quad Q_p^{S(m)} = Q_q^{S(m)}. \tag{94}$$

It therefore follows that the $S(m)$th eigensolution may only contribute to one of Y_a and Y_b, depending on sign of $Q_p^{S(m)}/Q_q^{S(m)}$. We also see that the electric circuit of Fig. 2 is a physical analogue for both Y_a and Y_b.

4.5.2 *Two-Port With (N-4) Terminals Open Circuit*

The symmetrical resonator with (N-4) terminals open circuit is shown in Fig. 7. We now use the z_{pq} matrix of (49), (67), and (73) to derive the impedance matrix of this two-port, again subject to the restrictions of (85). We find that

$$V_p = Z_{pp}I_p + Z_{pq}I_q \tag{95}$$

$$V_q = Z_{pq}I_p + Z_{pp}I_q, \tag{96}$$

where

$$Z_{pp} = z_{pp} + z_{ss} - 2z_{ps} \tag{97}$$

and

$$Z_{pq} = z_{pq} + z_{sr} - z_{pr} - z_{sq}. \tag{98}$$

From (82) and (67)

$$Z_{pq} = \frac{1}{j\omega}\left[G_{pq}^{(o)} - \sum_{m=1}^{\infty} \frac{\lambda G_{pq}^{(m)}}{(\lambda^{O(m)} - \lambda)}\right], \tag{99}$$

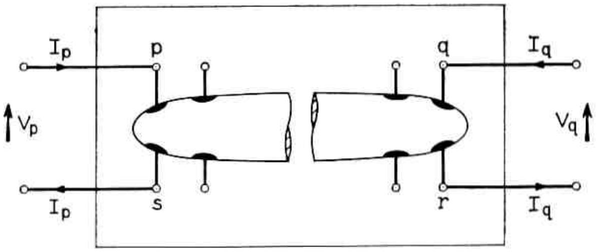


Fig. 7 — Two-port system for N-4 electrodes open-circuited.

where

$$G_{pp}^{(o)} = F_{pp}^{(o)} + F_{ss}^{(o)} - 2F_{ps}^{(o)}, \tag{100}$$

$$G_{pp}^{(m)} = \frac{(\Phi_p^{O(m)} - \Phi_s^{O(m)})^2}{\lambda^{O(m)} V(u_i^{O(m)} u_i^{O(m)})}, \tag{101}$$

$$G_{pq}^{(m)} = \frac{(\Phi_p^{O(m)} - \Phi_s^{O(m)})(\Phi_q^{O(m)} - \Phi_r^{O(m)})}{\lambda^{O(m)} V(u_i^{O(m)} u_i^{O(m)})}, \tag{102}$$

$$G_{pq}^{(o)} = F_{pq}^{(o)} + F_{sr}^{(o)} - F_{pr}^{(o)} - F_{sq}^{(o)}. \tag{103}$$

Also since the resonator has been taken to be symmetrical, i.e.,

$$Z_{pp} = Z_{qq}, \tag{104}$$

it follows, from 67, that

$$(\Phi_p^{O(m)} - \Phi_s^{O(m)}) = \pm(\Phi_q^{O(m)} - \Phi_r^{O(m)}). \tag{105}$$

If we represent the transfer equations (95) and (96) in terms of the lattice analogue of Fig. 8, subject to the restrictions of (85), we have

$$Z_a = \frac{1}{j\omega}\Bigg[(G_{pp}^{(o)} - G_{pq}^{(o)}) - \sum_{m=0}^{\infty} \frac{\lambda}{\lambda^{O(m)}} \frac{(\Phi_p^{O(m)} - \Phi_s^{O(m)})(\Phi_p^{O(m)} - \Phi_s^{O(m)} - \Phi_q^{O(m)} + \Phi_r^{O(m)})}{(\lambda^{O(m)} - \lambda) V(u_i^{O(m)} u_i^{O(m)})}\Bigg] \tag{106}$$

$$Z_b = \frac{1}{j\omega}\Bigg[(G_{pp}^{(o)} + G_{pq}^{(o)}) - \sum_{m=0}^{\infty} \frac{\lambda}{\lambda^{O(m)}} \frac{(\Phi_p^{O(m)} - \Phi_s^{O(m)})(\Phi_p^{O(m)} - \Phi_s^{O(m)} + \Phi_q^{O(m)} - \Phi_r^{O(m)})}{(\lambda^{O(m)} - \lambda) V(u_i^{O(m)} u_i^{O(m)})}\Bigg]. \tag{107}$$

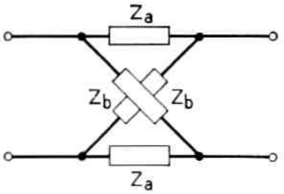


Fig. 8 — Electrical analogue of Fig. 7.

We note by virtue of (105) that the $O(m)$th eigensolution only contributes to one of Z_a and Z_b depending on the sign of $(\Phi_p^{O(m)} - \Phi_s^{O(m)}) /(\Phi_q^{O(m)} - \Phi_r^{O(m)})$. It also follows that each of Z_a and Z_b have the circuit of Fig. 3 as a physical analogue.

V. DISCUSSION AND CONCLUSIONS

It has been shown how the electrical behavior of a piezoelectric resonator with N electrodes, represented by an admittance or impedance matrix, can be determined from the eigensolutions for free vibrations of the resonator. The results for the admittance obtained by Lewis[1] are contained here as the special case $N = 2$ in (72). The impedance of the two-electrode resonator is described by (73). This result was not given by Lewis[1] since he did not consider the alternative expansion of the open circuit eigensolutions. It could be argued that since impedance is simply the reciprocal of admittance, residues of one could be found from the other. This would, however, involve rather cumbersome calculations. Furthermore, if an approximate theory is used to generate the eigensolutions, the expansions may only be valid in a small frequency range, thus making the calculation of, say, the residues of the impedance from the admittance expansion impossible.

Returning to the general case of the N-electrode resonator, the electrical behavior can be predicted in terms of the admittance or impedance matrices of (72) and (82). If external electrical components are to be connected between the N electrodes, and a two-port composite network is to be formed, its transfer characteristics can be deduced from either matrix. In the particular case of the symmetric resonator with N-2 of its electrodes connected together, the admittance matrix provides simple results, whereas the impedance matrix is useful for the case of N-4 open-circuited electrodes.

Before concluding, it might well be asked what reasons there are for preferring a motional parameter representation of the electrical characteristics over the direct method of determining the impedance or admittance matrices from the solution to the inhomogeneous boundary value problem. These are essentially twofold. Firstly, if a two-port N-electrode reasonator is to be designed to realize some transfer function or driving impedance, or so on, an appropriate synthesis procedure will usually automatically yield these motional parameters, leaving only the task of physically realizing a resonator with these parameters! Secondly, if the inhomogeneous boundary value problem is being solved, the inevitable numerical calculations are least likely to be accurate in just those ranges which are of prime interest, namely, the poles and zeros of the admittance or impedance matrices. Furthermore, considerable computing time would be lost, compared with the motional parameter method, if some transfer characteristic of the derived two-port were to be obtained for a large number of frequencies in a narrow band.

REFERENCES

1. Lewis, J. A., The Effect of Driving Electrode Shape on the Electrical Properties of Piezoelectric Crystals, B.S.T.J., *40,* September, 1961, pp. 1259–1280.
2. Mason, W. P., *Piezoelectric Crystals and Their Application to Ultrasonics,* Van Nostrand, 1950.
3. Lloyd, P. and Redwood, M., J. Acoust. Soc. Am., *39,* 1966, pp. 346–361.
4. Byrne, R. J., Lloyd, P., and Spencer, W. J., to be published.
5. Sykes, R. A. and Beaver, W. D., Proc. 20th Annual Frequency Control Symposium, Atlantic City, N.J., April, 1966.
6. Lloyd, P., Ph.D. Thesis, University of London, 1966.
7. EerNisse, E. P. and Holland, R., Proc. 21st Annual Frequency Control Symposium, Atlantic City, N.J., April, 1967.
8. Love, A. E. H., *The Mathematical Theory of Elasticity,* University Press, Cambridge, 1927, 4th Ed., pp. 278–287.
9. Weinberg, L., *Network Analysis and Synthesis,* McGraw-Hill Book Co., Inc., New York, 1962.

Properties and Device Applications of Magnetic Domains in Orthoferrites

By A. H. BOBECK

(Manuscript received August 15, 1967)

It has been shown that isolated magnetic domains in thin platelets (≈ 2 mils thick) of orthoferrites can be manipulated to perform memory, logic, and transmission functions. The purpose of this paper is to discuss the properties of orthoferrites that make them suitable for magnetic device applications and consider magnetostatic problems relevant to domain structures found to be useful. Included is a brief indication of how memory, logic, and transmission can be accomplished; however, the details will be reserved for a later paper.

The stability conditions of a cylindrical domain are discussed in detail and data is reported to support the conclusions. Of particular interest are the sizes of cylindrical domains available in the various orthoferrites. Such data has been taken on five of the fourteen possible orthoferrites and it is found that the thulium orthoferrite, $TmFeO_3$, gives the smallest stable domain diameter (2.3 mils) and $LuFeO_3$ the largest. The stability results lead to a direct method for obtaining σ_W, the domain wall energy density. For $TmFeO_3$, as an example, $\sigma_W = 2.8$ ergs/cm^2.

It is concluded that the orthoferrites are well suited indeed for device applications. Experimentally, 3 mil diameter domains have been manipulated and there is every reason to believe that operation of sub-mil domains will soon be realized.

I. INTRODUCTION

Recently, P. C. Michaelis described a technique for propagating isolated magnetic domains in thin anisotropic ferromagnetic films.[1] He obtained controlled motion along either the easy (e) or hard (h) anisotropy axis although he used distinctly different mechanisms to obtain propagation in these directions.

Michaelis' ferromagnetic films were processed to have a uniaxial anistropy (i.e., hard and easy axis) in the plane of the film. Magneti-

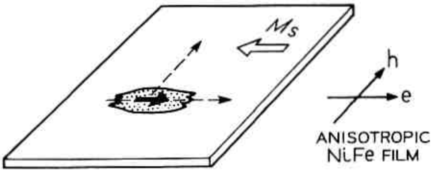


Fig. 1 — An isolated magnetic domain can be moved along the easy (e) or the hard (h) anisotropy axis.

zation lies in the plane of the film and a magnetic domain, as illustrated in Fig. 1, is seen to be an isolated reverse magnetization area bounded by a domain wall. The disparity in the propagation modes for the e and h directions is due to anisotropy inherent in the film itself. Propagation of domains along the diagonals is possible using conventional wall propagation and, in fact, Spain[2] has recently discussed such a technique.

Complete generality in the propagation (and interactions) of magnetic domains, however, demands that the magnetization be aligned *normal* to the surface of the film. Furthermore, it would be useful if the magnetic properties were isotropic (or essentially so) in the plane of the film. A cylindrical domain in such a material is drawn in Fig. 2. These conditions are met in orthoferrites as first pointed out by R. C. Sherwood. Other similar materials are magnetoplumbite, barium ferrite, and manganese bismuth.

This memorandum will describe some of the results of generating and propagating cylindrical domains. Related magnetostatic problems are introduced and discussed. It will be shown that the stability conditions for cylindrical domains lead to a method of determining wall energies and results obtained so far on orthoferrites are tabulated. Finally, a brief description is given of some of the

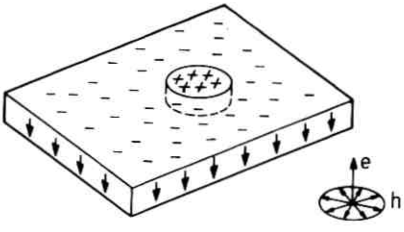


Fig. 2 — An ideal material permits magnetization normal to platelet surface and is otherwise isotropic.

device properties; however, this aspect will be treated in much more detail in a later memorandum.

II. ORTHOFERRITES

An excellent treatment of orthoferrites can be found in Ref. 3. Orthoferrites of the general formula $MFeO_3$, where M is any rare earth ion are antiferromagnetic with a weak ferromagnetism caused by a slight canting (0.5°) of the antiparallel spins. The molecular and magnetic unit cell is an orthorhombic cell of sides a, b, and c with the c side about twice the length of a or b, as illustrated in Fig. 3. The antiparallel Fe 3+ spins align along the a-axis with the c-axis exhibiting the weak ferromagnetism ($4\pi M_s \approx 100$ gauss). The lone exception is $SmFeO_3$ which has its net moment along the a-axis at room temperature. The Néel temperature for all orthoferrites is about 400°C.

The orthoferrites have a remarkable set of magnetic properties. When magnetized to saturation, fields of several thousand oersteds are needed to effect a flux reversal. This field (nucleation field H_N) is an order of magnitude greater than the magnetic moment of a typical orthoferrite. Thus, platelets having the c-axis normal to the planar surface are magnetically stable without an applied field even when fully saturated. Furthermore, once domain walls are present they can be moved with fields (wall coercivity H_c) less than one oersted. Thus,

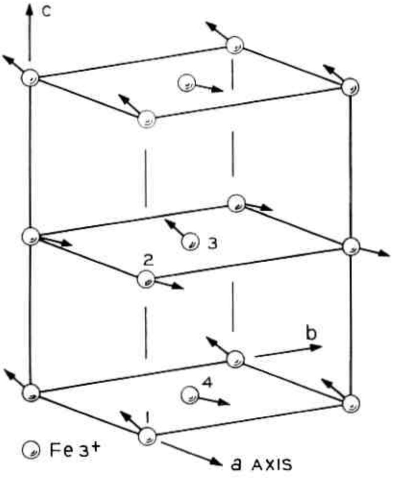


Fig. 3 — Locaton of Fe3⁺ spins in typical orthoferrite orthorhombic cell.

the orthoferrites are ideally suited for device applications which utilize materials with a reentrant *B-H* hysteresis characteristic.

Orthoferrites can be grown from a PbO flux and all of the specimens evaluated during the course of this work were so prepared by J. P. Remeika and L. J. VanUitert. Occasionally, a particular run will yield nearly perfect single crystal platelets of orthoferrite. Dimensions vary with 0.1 inch by 0.2 inch by 2 mils thick being typical. Fortunately, most of these platelets grow with the *c*-axis normal to the plane and are thus ideally suited for cylindrical domain observations.

The tendency to grow platelets is not characteristic of all orthoferrites. For this reason, it has been necessary to develop techniques for preparing platelets from larger crystals.

Orthoferrites are optically transparent, especially to the red spectrum. Magnetic domains in a thin platelet (2.3 mils thick) of $TmFeO_3$ are readily seen using the Faraday rotation of transmitted light. Note in Fig. 4 the random orderliness of the areas magnetized up (dark) and down (light). Magnetostatic and wall energies balance to determine the general shape as well as the dimensions of the domains. It is only with a great deal of reluctance that these domains yield to an inhomogeneous field.

III. MAGNETOSTATICS AND STABILITIES OF STRIPS AND CYLINDERS

Assume, as pictured in Fig. 5, that a loop of wire is placed in contact with the surface of an orthoferrite platelet. When a current

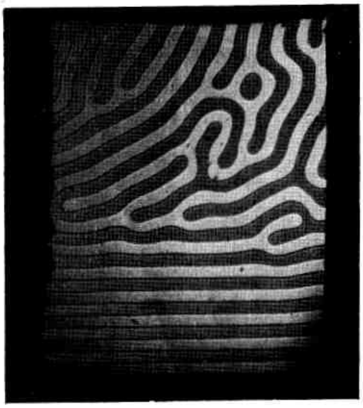


Fig. 4 — Faraday observation of magnetic domains in $TmFeO_3$. Note the isolated oval domain.

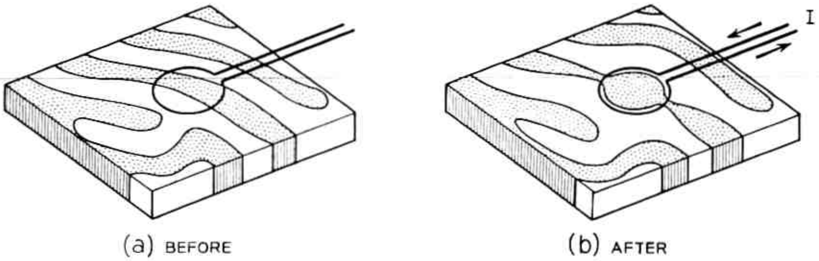


Fig. 5 — In the process of generating a cylindrical domain a current applied to a loop alters the static domain pattern of (a) to that of (b).

is applied to the loop the resulting field pattern will perturb the existing domains. Areas of magnetization seeing a favoring applied field will grow — others will shrink. To produce a cylindrical domain, note that it will be necessary to "pinch off" at several points. To maintain a cylindrical domain when the applied current, I, is removed, it is usually necessary to apply a dc bias field normal to the surface of the platelet. The conditions under which the cylindrical domain is stable will be analyzed in this section.

3.1 *Magnetic Strip Domain*

To a first approximation, the field necessary to "pinch" a strip domain to produce a cylindrical domain can be equated to the field required to compress a strip domain to zero width. The strip resists compression since a high magnetostatic energy state is generated. The magnetostatic field effective when the strip approaches zero width can be obtained by inspection and is $4\pi M_s$ (see Fig. 6). One immediately sees the significance of the low moment of the orthoferrites since the applied fields necessary to generate cylindrical domains will be directly related to the magnetic moment.

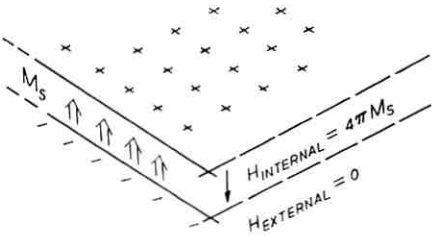


Fig. 6 — The internal magnetostatic field in a uniformly magnetized platelet is $4\pi M_s$.

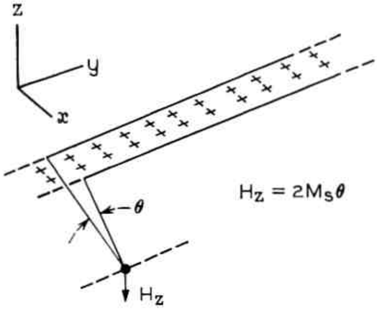


Fig. 7 — Normal magnetostatic field component generated by a strip of magnetic charge is related to the angle θ.

The relationship between the width W of a strip domain and the applied field H_s will now be discussed. (Refer to Fig. 7.) Consider a strip of magnetic charge located in the xy plane and extending to infinity in the $+y$ and $-y$ directions. The magnetic field component perpendicular to the xy plane, H_z, is directly related to the angle θ subtended by the strip and is given by $H_z = 2M_s\theta$. (A similar relationship exists for a strip carrying a uniform current.)

Consider a strip domain of width W in an orthoferrite platelet of thickness h. This case is illustrated in Fig. 8. By means of the angle relationship discussed above the z-component of field, produced by the magnetic surface charges, can be quickly obtained. For example, the field at the midpoint of either domain wall (sides of the strip) is

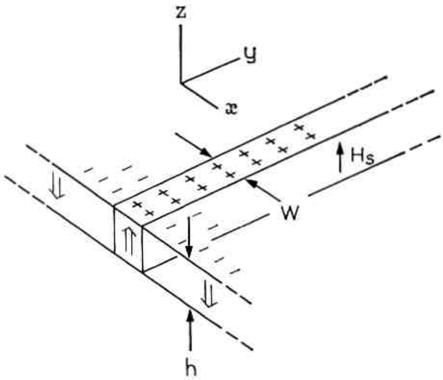


Fig. 8 — The applied field necessary to sustain a strip domain is derived in the text.

$$H(z = h/2) = 8M_s \tan^{-1} h/2W. \tag{1}$$

The polarity of this field is such as to cause the strip to widen. An applied field, equal in magnitude and opposite in sign, is, therefore, needed to maintain the strip at a width W.

More significant is the *average* field effective on the walls under the assumption that the walls are rigid (do not bulge outward). In Appendix A the desired relationship is derived and is

$$\frac{H_s}{4\pi M_s} = \frac{2}{\pi}\left[\tan^{-1}\left(\frac{h}{W}\right) - \frac{W}{2h}\ln\left(1 + \frac{h^2}{W^2}\right)\right]. \tag{2}$$

A similar expression is found in Kooy and Enz.[4] Equation (2) is plotted in Fig. 9. Note that as the strip narrows ($W \to 0$) the normalized field approaches unity as discussed previously. For a very wide strip the field effective on the wall tends to zero and the walls are stable in position without an external applied field.

Equation (2) tells us to what extent the surface charge of the strip and the sheet cancel. As W increases, the walls see cancelling fields from the surrounding magnetic charge and the wall field reduces to a low value. If the strip is driven closed, however, the quantity of nullifying charge on the strip itself goes to zero and the full internal field $4\pi M_s$ is felt. If the wall coercivity, H_c, is very low only very wide strips ($W \gg h$) will be stable in the absence of an applied field.

3.2 *Cylindrical Magnetic Domains*

In the derivation of equations which related to the performance of a strip domain, it was not necessary to include a domain wall energy

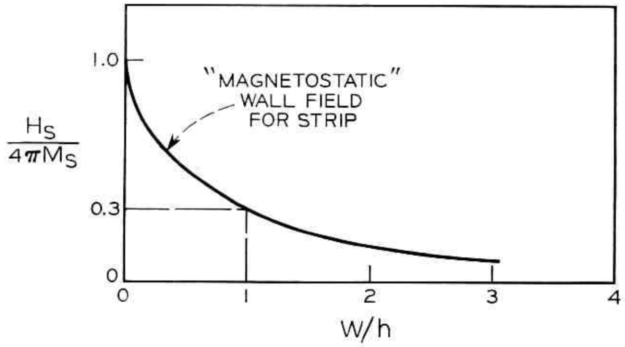


Fig. 9 — An applied field of $4\pi M_s$ is necessary to collapse a strip to zero width.

term since the walls maintained constant area. Such is not the case for a cylindrical domain.

Consider as shown in Fig. 10 a cylindrical domain of radius r in a platelet of orthoferrite of thickness h. Assume that the domain wall which defines the cylindrical domain has straight sides, i.e., the shape is neither "hour-glass" nor "barrel" like. The total energy, relative to a uniformly downward magnetized platelet can be written as

$$\xi_T(\text{total}) = \xi_W(\text{wall}) + \xi_D(\text{magnetostatic}) + \xi_H(\text{applied})$$

or,

$$\xi_T = 2\pi rh\sigma_W - \xi_D + 2M_sH_A\pi r^2h,$$

using CGS units where the domain wall energy density σ_W is in ergs/cm². The partial derivative of the energy with respect to r gives the force on the wall.

$$\frac{\partial\xi_T}{\partial r} = 2\pi h\sigma_W - \frac{\partial\xi_D}{\partial r} + 4\pi rhM_sH_A .$$

It is assumed that $\partial\sigma_W/\partial r$ is negligible. In terms of fields,

$$\underset{(I)}{\frac{\partial\xi_T/\partial r}{4\pi M_s rh}} = \underset{(II)}{\frac{\sigma_W}{2rM_s}} - \underset{(III)}{\frac{\partial\xi_D/\partial r}{4\pi M_s rh}} + \underset{(IV)}{H_A} . \tag{3}$$

Equation (3) is a stability relationship in terms of magnetic fields. It relates the net effective field on the wall (*I*) to the wall field (*II*) trying to compress the cylindrical domain, the demagnetization field (*III*) trying to expand that domain, and the applied field (*IV*). If the net field (*I*) is positive the domain will compress, if it is negative it will expand.

It is convenient to express the "wall energy" and "magnetostatic" contributions as fields. Term (*II*) can be equated to a wall field,

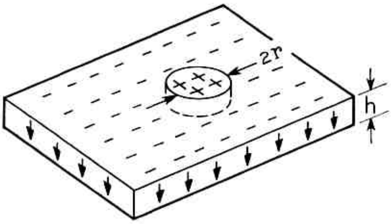


Fig. 10 — A cylindrical domain in a platelet of thickness h.

termed H_W, since it is the field contributed by the wall energy density, σ_W. So

$$H_W = \frac{\sigma_W}{2rM_s}. \tag{4}$$

Note that this field, which goes as $1/r$, is the eventual cause of the collapse inward of the smallest domains. The importance of (4) which is plotted in Fig. 11 cannot be overstressed since it points out the significant role that wall energy will play in orthoferrite devices.

A. Thiele[5] has obtained an expression in closed form for (*III*) the magnetostatic field which we shall designate as H_D. A derivation of the magnetostatic energy ξ_D is not readily obtainable. An alternative derivation, somewhat simpler but less rigorous than the technique employed by Thiele, is presented in Appendix B. The result (by either method) is

$$\frac{H_D}{4\pi M_s} = \frac{2}{\pi}\left[-\frac{2r_0}{h} + \sqrt{1 + (4r_0^2/h^2)}E(k, \pi/2)\right], \tag{5}$$

where $E(k, \pi/2)$ is the complete elliptic integral of the second kind and

$$k^2 = \frac{1}{1 + (h^2/4r_0^2)}.$$

The normalized magnetostatic field plots much (see Fig. 12) as the strip field of Section 3.1. The sense of H_D is always to attempt to expand the cylindrical domain. Equation (5) is plotted in detail in Fig. 24.

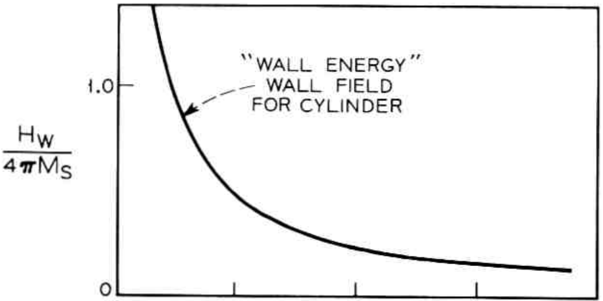


Fig. 11 — The wall energy, σ_W, generates a field of a sense to collapse a cylindrical domain.

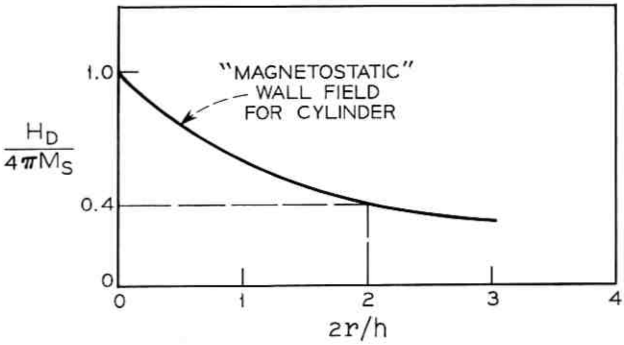


Fig. 12 — The magnetostatic energy generates a field which attempts to expand a cylindrical domain.

It is probably well to reiterate that the fields H_A, H_W, and H_D are assumed to be acting on a rigid cylindrical domain wall and H_W and H_D have no significance unless applied to the domain wall itself.

3.3 *The Stability of a Circular Domain*

We are now in a position to discuss the stability of a cylindrical domain. Three cases will be considered. They are (*i*) $r \ll h$, a very thick platelet, (*ii*) $r \gg h$, a very thin platelet, and (*iii*) $r \approx h$, a "just right" platelet. Case (*i*) will be magnetostatic energy dominated, case (*ii*) wall energy dominated, and case (*iii*) will have these energies somewhat in balance.

3.3.1 *Very Thick Platelet*

For a cylindrical domain of radius r in a platelet of thickness h, where $r \ll h$, the magnetostatic field H_D of (5) may be approximated as $4\pi M_s$. In this case the critical radius, r_a, is obtained by equating the magnetostatic field to the wall field. This is graphically done in Fig. 13. Analytically,

$$4\pi M_s = \frac{\sigma_W}{2r_a M_s}$$

and

$$r_a = \frac{\sigma_W}{8\pi M_s^2}. \tag{6}$$

For typical orthoferrites, r_a is the order of 0.5 mil. Domains of a radius greater than r_a will expand uncontrollably while those of radius

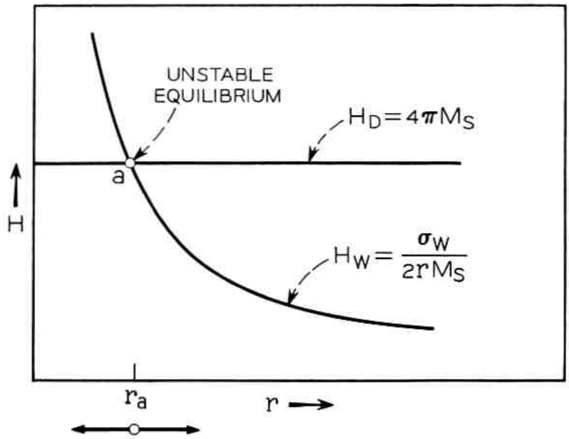


Fig. 13 — Metastable equilibrium exists with "very thick" platelets.

less than r_a will contract to oblivion. The significance of r_a is that it gives an absolute lower bound on the stable domain size. Note that the problem of initially establishing a domain of radius r_a has been carefully avoided.

3.3.2 *Very Thin Platelet*

At the other extreme are the conditions that exist in a very thin platelet, i.e., $r \gg h$ and the wall field completely dominating the magnetostatic field. Note Fig. 14. In the absence of a coercivity the criti-

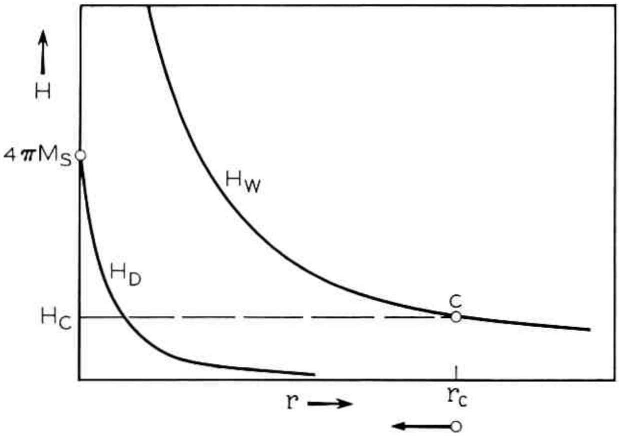


Fig. 14 — Stability condition for a "very thin" platelet.

cal radius r_c will approach infinity. With a wall coercivity H_c the minimum stable radius can be obtained from

$$H_c = \frac{\sigma_W}{2r_c M_s}$$

or (7)

$$r_c = \frac{\sigma_W}{2M_s H_c}.$$

For typical orthoferrites, r_c is 40 mils. Indeed, difficulty is experienced in generating domain patterns of any kind in very thin (less than one mil) platelets.

3.3.3 *"Just Right" Platelets*

We have seen that cylindrical domains in very thick platelets are completely unstable. Also, that in very thin platelets only excessively large domains are stable and then only if some form of wall pinning is assumed.

A stable cylindrical domain can be obtained if the thickness h of the platelet is chosen so that at the point of intersection of the H_D and H_W curves

$$\left| \frac{\partial H_D}{\partial r} \right| > \left| \frac{\partial H_W}{\partial r} \right|. \tag{8}$$

In general, a bias field (H_{bias}) is needed to secure the intersection at "b" which satisfies (8). Refer to Fig. 15 and note that an $H_W + H_{\text{bias}}$ curve

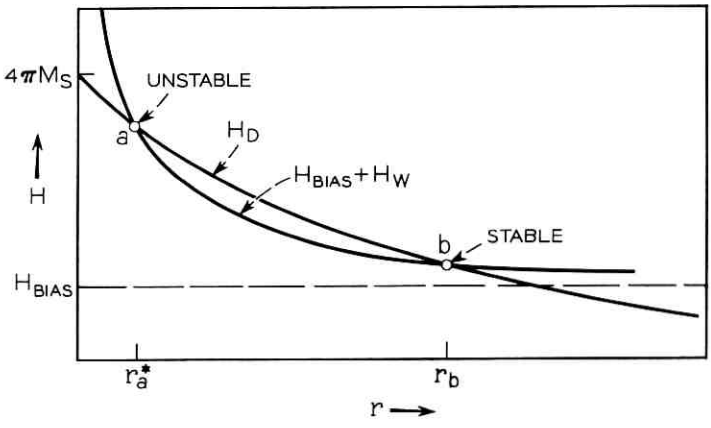


Fig. 15 — A stable cylindrical domain can be obtained if a suitable bias field is applied.

is plotted. For a radius somewhat greater than r_b the combined wall and applied fields close the domain to r_b. If the starting radius is less than r_b but greater than r_a^* the dominant magnetostatic field will open the domain to r_b. Any domain of radius less than r_a^* will collapse. By adjusting the bias field r_b can be varied somewhat. For example, in $TmFeO_3$, 2.3 mils thick, r_b can be varied from 2.8 mils to 1.2 mil as the bias field is changed from 26 Oe to 36 Oe.

As H_{bias} is increased, r_a^* and r_b approach each other and become equal when the $H_W + H_{bias}$ curve is tangential to the H_D curve. With a further increase, all cylindrical domains become unstable and collapse. This leads to a direct method of obtaining the wall energy density, σ_W.

IV. WALL ENERGY DENSITY, σ_W

The cylindrical domain radius r_b as a function of an applied field H_{bias} has been measured for a number of rare earth orthoferrites. A typical curve of domain radius vs applied field is plotted in Fig. 16. This experiment, as was pointed out above, is a direct method for obtaining σ_W the domain wall energy density.

For the 2.3-mil thick platelet of $TmFeO_3$ described previously the minimum r observed is 1.15 mil at an applied field of 36 Oe. Since $h = 2.3$ mils, $2r/h = 1$. From Fig. 24, $H_D/4\pi M_s = 0.58$. For $TmFeO_3$ $4\pi M_s = 140$ gauss so $H_D = 81$ Oe. Thus, $H_W = H_D - H_{bias} = 81 - 36 = 45$ Oe. Therefore,

$$\begin{aligned}\sigma_W &= 2rH_WM_s\\ &= 2.80 \text{ ergs/cm}^2.\end{aligned}$$

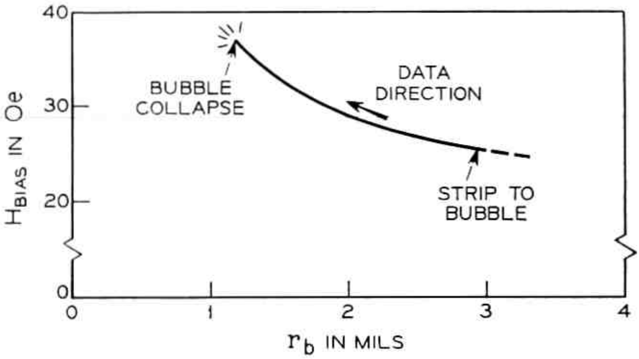


Fig. 16 — Cylindrical domain size as a function of an applied bias field.

The average results obtained on orthoferrites for which platelets were available are tabulated in Table I. Note that in addition to the measured values of $r_{\min}$ that a calculated $2\pi r_a$ (Section 3.31) is also included. Thiele[5] has shown that $2\pi r_a$ is the optimum platelet thickness. Platelets of thickness $2\pi r_a$ will sustain cylindrical domains of the minimum possible diameter and this diameter will be approximately $2\pi r_a$. Note that $TmFeO_3$ has the lowest calculated $2\pi r_a$ (as well as the lowest observed domain size), and is thus a prime contender for device applications. It is hoped that some other orthoferrite such as $YFeO_3$ which has a reported[7] σ_W of 1.0 erg/cm^2 and $4\pi M_s$ of 105 gauss may give yet smaller domains than $TmFeO_3$.

V. GENERAL COMMENTS AND OBSERVATIONS

The sequence of pictures of Fig. 17 give a pictorial display of much of the material described in the preceding sections. An "as grown" $TmFeO_3$ platelet, 2.3 mils thick was subjected first to an increasing bias field applied parallel to the c-axis (perpendicular to the planar face) and then to a decreasing field.

The platelet is demagnetized, i.e., equal areas of magnetization up and down, if no field is applied (a). The average width of a domain is the consequence of a magnetostatic and wall energy balance and as such can be used to obtain a crude estimate of the wall energy. It is an observation that for very thick or very thin samples the strip width is increased from that shown.

As the field is increased, the dark strips narrow and there is a general reshuffling of domains. At (g) one of the "dumbell" domains has collapsed into a cylindrical domain. Further increase in field finds a total of five such domains (i). At 37 Oe applied field, three of the five have collapsed (j) and the remaining two would also collapse if a few more tenths of an oersted was applied. However, at this point

TABLE I

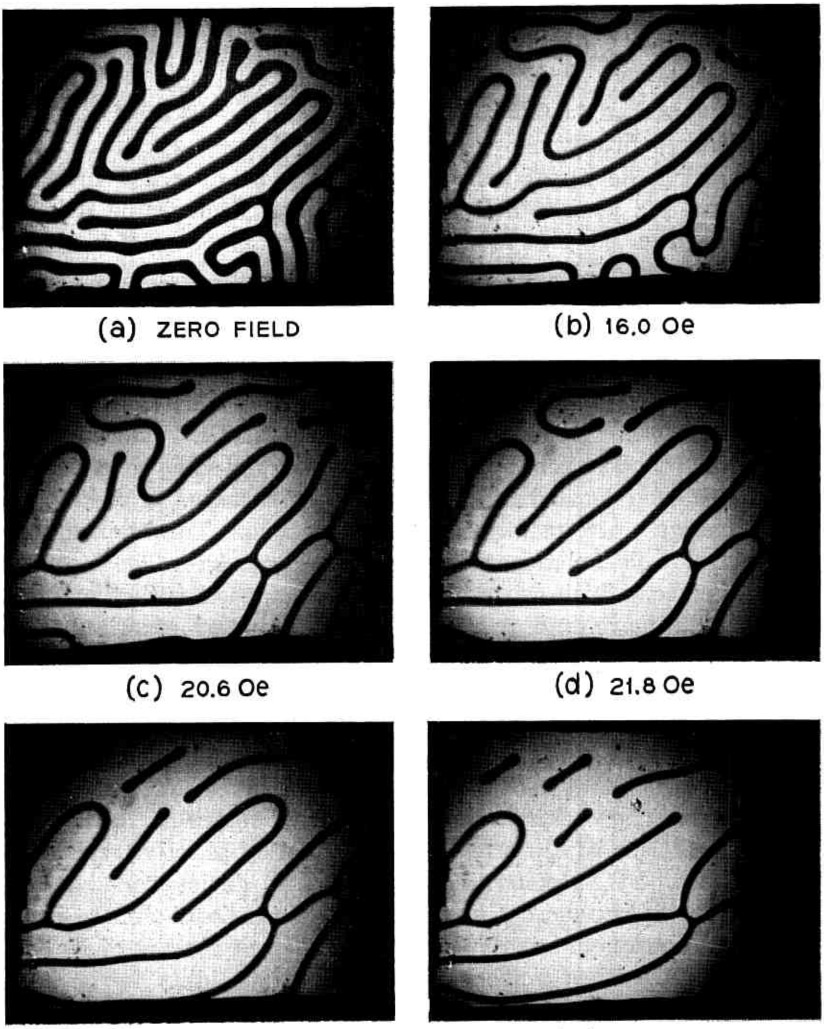
			Observed			Calculated
Orthoferrite	$4\pi M_s$[6] (gauss)	σ_W (ergs/cm^2)	$2r_{\min}$ (mils)	Field (Oe)	Thickness (mils)	$2\pi r_a$ (mils)
$HoFeO_3$	91	2.0	4.6	12	2.1	3.8
$ErFeO_3$	81	1.7	6.0	8	2.0	4.1
$TmFeO_3$	140	2.8	2.4	36	2.3	2.2
$YbFeO_3$	143	3.0	3.8	59	4.4	2.3
$LuFeO_3$	119	3.9	8.8	4	1.4	4.3

the field was slowly reduced and the two cylindrical domains grew in size. At 27 Oe they became unstable as circles and blew out into strips (*k*). The circle to strip to circle process can be repeated with a field perturbation of 1.5 oersted. Thus, the wall coercivity, H_c, is probably less than $1.5/2 = 0.75$ oersted. If the field is reduced to zero a pattern superficially like (*a*) appears.

By passing the tip of a fine magnetized wire over the surface of a demagnetized platelet it is possible to "cut" through the strip domains. Then as the bias field is applied large numbers of "bubbles" appear such as in Fig. 18. Our next problem is to look at the ways in which these "bubbles" can be manipulated to do logic and storage.

It will suffice, for the purposes of this article, to just indicate very briefly some of the operations that are possible. These are illustrated in Fig. 19. Since in the correct environment cylindrical domains have been shown to be stable, *storage* is readily available. *Transmission* of a domain from location 1 to location 2 is achieved by energizing a conductive loop located at position 2. The domain, in seeking the lowest energy state, readily moves to position 2. To obtain a complete set of logic functions an *interaction* is required. The magnetostatic repulsion which exists between domains ensures that only one domain will move into position 2 if that loop is energized. New domains can be created by *replication*. This involves literally tearing a single domain into two halves which then expand to full size. The use of the full set of operations allows the possibility of data processing applications.

It is apparent that a multi-dimensional shift register can be built using *transmission* with a three-phase drive source to achieve the desired directionality. Information is inserted by selective *replication* at the input of the register. Most of the early device work has centered on the design and operation of multi-dimensional shift registers. Initial success was obtained when a 2.2-mil thick platelet of $HoFeO_3$ was combined with a waffle-iron like high-permeability ferrite baseplate[8] The baseplate consisted of a matrix of 10-mil by 10-mil posts positioned on 15-mil centers. Single turn windings were wrapped about selected posts and series connected to form the five distinct propagation phases identified in Fig. 20(f). The waffle-iron posts served to facilitate the wiring procedure as well as to precisely define the applied field patterns. Assume as shown in (a) that domains exist in the orthoferrite platelet at the upper left and middle left Φ1 locations. The location of the domains can be ascertained by viewing the results



(a) ZERO FIELD

(b) 16.0 Oe

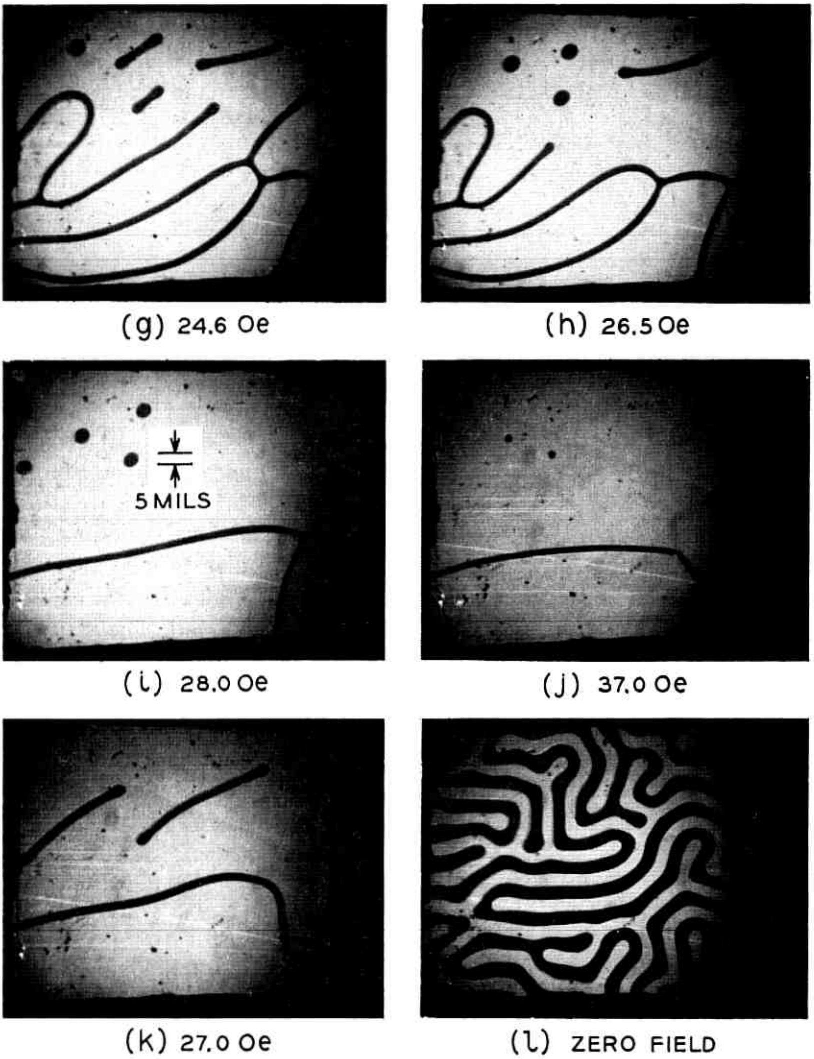
(c) 20.6 Oe

(d) 21.8 Oe

(e) 22.7 Oe

(f) 24.5 Oe

Fig. 17 — Faraday studies of a platelet $TmFeO_3$ orthoferrite, 2.3 mils thick, *c*-axis normal to the surface. Sample originally demagnetized (a). Field applied normal to the surface, first increased, then decreased. Note at (g) that the strip in the upper left-hand corner became a cylindrical domain. This, and the other



(g) 24.6 Oe

(h) 26.5 Oe

(i) 28.0 Oe

(j) 37.0 Oe

(k) 27.0 Oe

(l) ZERO FIELD

Fig. 17 — (continued)
cylindrical domains which formed, reduced in size until (j) when three of five collapsed. As the field is decreased the remaining two bubbles open into strips (k) and eventually grow to fill the entire platelet (1).

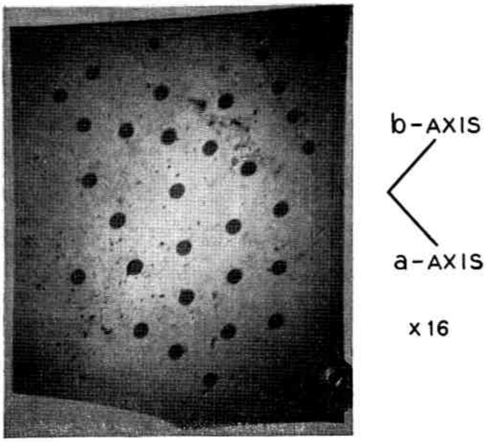


Fig. 18 — Numerous cylindrical domains produced by "cutting" strip domains with a magnetized wire.

of a magnetic colloid interaction with the domain walls defining the cylindrical domains. Application of a current pulse to Φ2 causes the pair of domains to step one post position to the right. In this manner the domain patterns of (b) through (d) are generated. The sequence 123145132154 . . . causes domains to propagate clockwise in the upper and lower loops resulting in a residual pattern (e) being generated by the colloid.

More recent work has been directed toward improving the storage density of the shift register. Operation with 3.5-mil diameter domains has been achieved and there is every indication that sub-mil domains can also be propagated. The final storage density of the device ap-

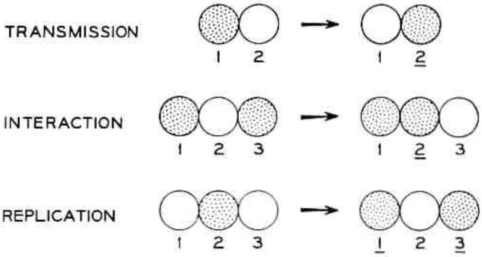


Fig. 19 — Illustration of transmission, interaction, and replication. Shaded areas represent cylindrical domains, circles the drive loops, and underlined numbers the energized loops.

pears to be limited by wiring pattern resolution rather than any magnetic property.

VI. CONCLUSIONS

A variety of experiments have been performed on orthoferrites. These include cylindrical domain stabilities, strip stabilities, magnetostatic interactions, and device applications. It has been found that the concept explained in the body of this memorandum of considering the magnetostatic and wall energies as generating equivalent fields, is useful toward a first order understanding of the phenomenon observed.

Cylindrical domain stability has been studied in detail. The wall energy density σ_W and the magnetic moment $4\pi M_s$ are seen to be significant factors limiting the minimum available diameter of a cylindrical domain. With experiments completed on five of fourteen orthoferrites the results have shown that $TmFeO_3$ has the smallest stable domain diameter, 2.3 mils. There is every reason to expect that submil domains will be realized in other orthoferrites.

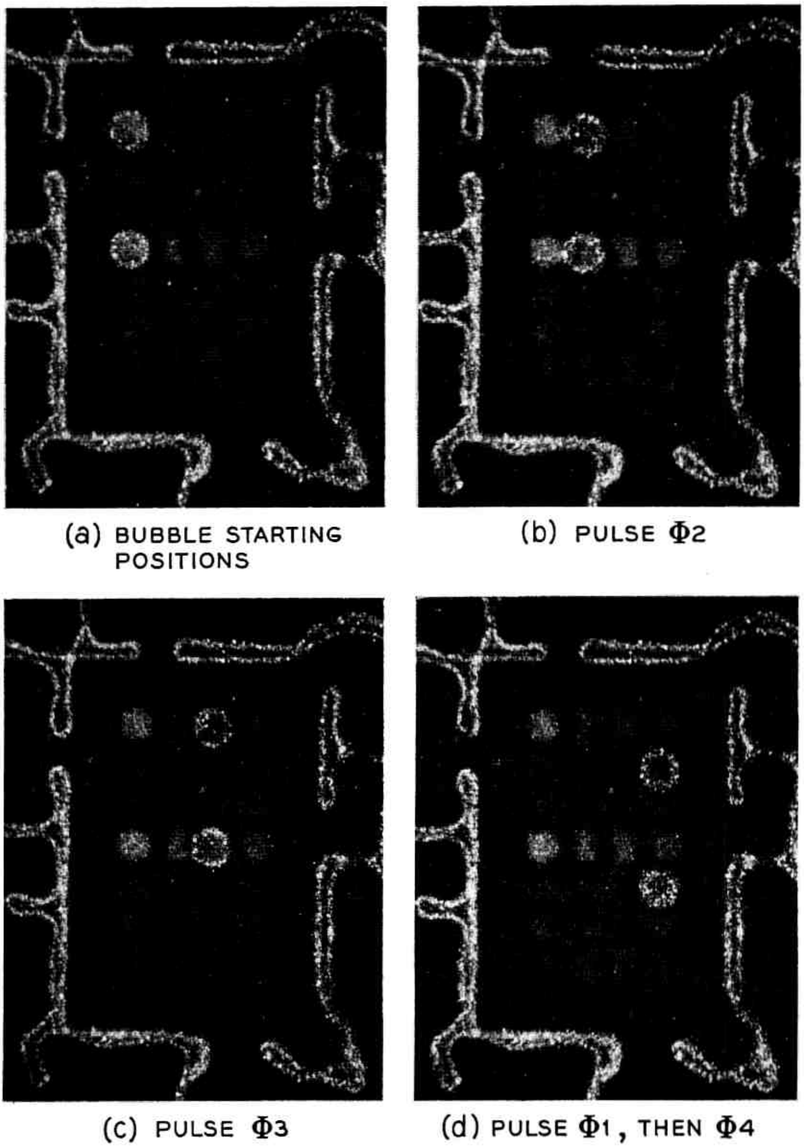
This paper has discussed idealized, elastic, domains. However, most of the early successes in manipulating domains in shift register, logic, and memory structures were accomplished with rather thin, high coercivity platelets. Further study will be required to determine the optimum blend of operating characteristics.

VII. ACKNOWLEDGMENTS

Many people have contributed directly to the orthoferrite studies. Samples were furnished by J. P. Remeika, L. G. Van Uitert, and W. H. Grodkiewicz, with inspiration by H. E. D. Scovil, W. Shockley, and U. F. Gianola. Special thanks must be given to R. C. Sherwood for initially suggesting the orthoferrites and A. Thiele for his derivations and many helpful discussions.

APPENDIX A

It is desired to derive the average z-component field $\bar{H}_z$ acting on the domain walls which define a strip of magnetization reversal of width W located in an infinitely large magnetic platelet of thickness h. See Fig. 21. Consider the domain wall located in the $X = 0$ plane. Note that the surface magnetic charge of the strip itself and that of an image strip also of width W will produce cancelling fields at any



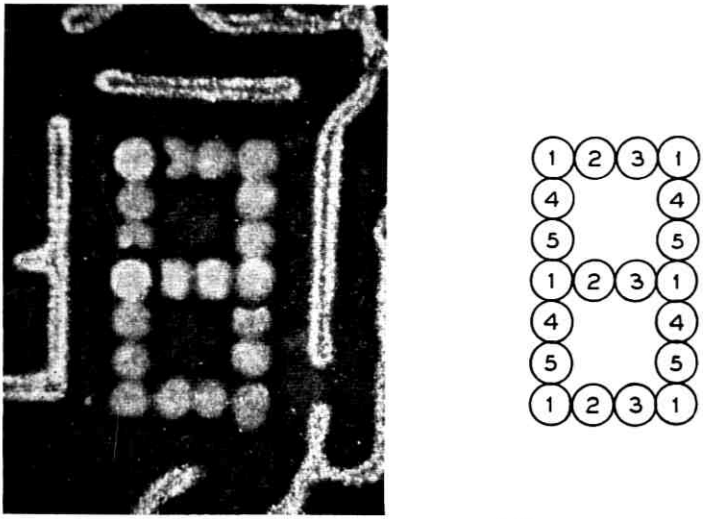
(a) BUBBLE STARTING POSITIONS

(b) PULSE Φ2

(c) PULSE Φ3

(d) PULSE Φ1, THEN Φ4

Fig. 20 — Sequence of photographs illustrating two-dimensional shifting of cylindrical magnetic domains in $HoFeO_3$ orthoferrite. Operation was obtained on a ferrite waffle-iron baseplate and observed with a 3M Colloid Viewer.



(e) RESIDUAL PATTERN IN COLLOID VIEWER AFTER CONTINUOUS 123145132154 SEQUENCE

(f) IDENTIFICATION OF DRIVE LOOP DESIGNATIONS

Fig. 20 — (continued)

"z" and thus need not be considered. The field at "z" due to the upper right-hand sheet of charge is

$$\frac{H_z}{4\pi M_s} = \frac{1}{2\pi} \tan^{-1}\left(\frac{z}{W}\right).$$

Now

$$\bar{H}_z = \frac{1}{h}\int_0^h H_z \, dz.$$

This equation assumes that the domain wall is rigid and therefore that the force acting on the wall can be averaged.
So

$$\frac{\bar{H}_z}{4\pi M_s} = \frac{1}{2\pi h}\int_0^h \tan^{-1}\left(\frac{z}{W}\right) dz$$
$$= \frac{1}{2\pi}\left[\tan^{-1}\left(\frac{h}{W}\right) - \frac{W}{2h}\ln\left(1 + \frac{h^2}{W^2}\right)\right].$$

Since four sheets of magnetic charge are acting on the wall (producing components identical in field direction and magnitude) the total

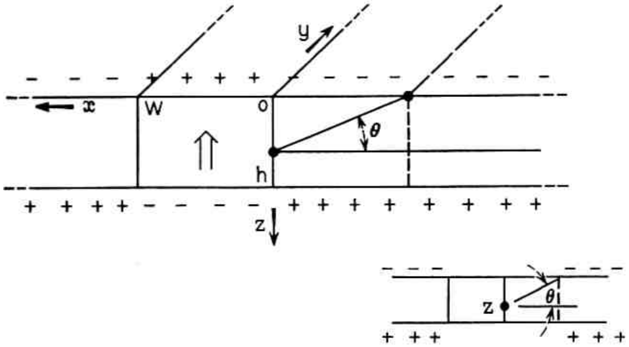


Fig. 21 — Identification of parameters used in the derivation of the strip magnetostatic field.

average field becomes

$$\frac{\bar{H}_z}{4\pi M_s} = \frac{2}{\pi}\left[\tan^{-1}\left(\frac{h}{W}\right) - \frac{W}{2h}\ln\left(1 + \frac{h^2}{W^2}\right)\right].$$

In the body of the text (Section 3.1) the z-component of field for the strip is designated H_s and the above expression is entered as (2).

The assumption that the walls defining the strip domain are rigid (and straight) is a good approximation for $W \gg h$ and poor for $W \ll h$. In the latter case, the average force acting on the walls differs significantly from the maximum forces experienced by the walls. It is expected that for $W \ll h$ the walls will bulge outward.

APPENDIX B

The calculation of the average wall field $\bar{H}_z$ produced by surface magnetic charge, for a circular magnetic domain proceeds in much the same manner as for the strip. As in the strip case the domain wall itself is assumed to be rigid. The mathematics are simplified if it is recognized that a cancellation cylinder exists as shown in Fig. 22. Only one quadrant will be considered with a factor of eight included to account for all four quadrants plus a top and bottom.

In general, as defined in Fig. 23,

$$H_z = \int \frac{M_s \cos\alpha \, dA}{r^2}$$

so

$$\frac{H_z}{8M_s} = \int_0^{\pi/2} \int_{2r_0 \sin\theta}^{\infty} \frac{\rho z \, d\rho \, d\theta}{(\rho^2 + z^2)^{\frac{3}{2}}}.$$

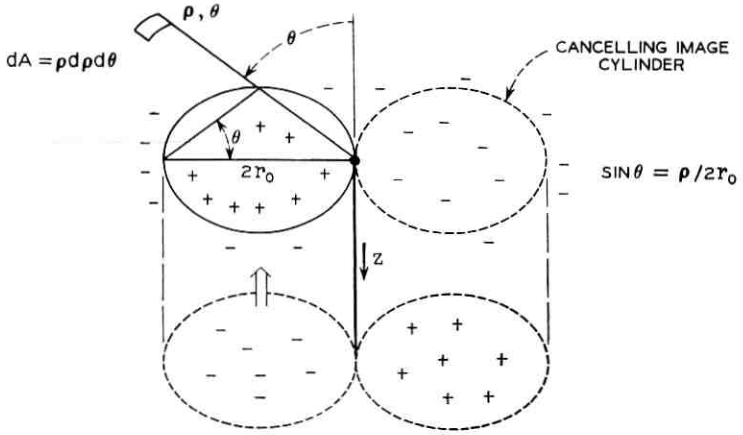


Fig. 22 — Figure useful in the derivation of the average wall field of a cylindrical domain produced by surface magnetic charge.

Using

$$\bar{H}_z = \frac{1}{h}\int H_z\,dz$$

$$\frac{\bar{H}_z h}{8M_s} = \int_0^{\pi/2}\int_{2r_0\sin\theta}^{\infty}\int_0^h \frac{z\,dz\,\rho\,d\rho\,d\theta}{(\rho^2+z^2)^{\frac{3}{2}}}$$

$$= -\int_0^{\pi/2}\int_{2r_0\sin\theta}^{\infty}\frac{\rho\,d\rho\,d\theta}{(\rho^2+z^2)^{\frac{1}{2}}}\Bigg|_0^h$$

$$= \int_0^{\pi/2}\int_{2r_0\sin\theta}^{\infty}\left[1-\frac{\rho}{(\rho^2+h^2)^{\frac{1}{2}}}\right]d\rho\,d\theta$$

$$= \int_0^{\pi/2}[\rho-(\rho^2+h^2)^{\frac{1}{2}}]\,d\theta\Bigg|_{2r_0\sin\theta}^{\infty}$$

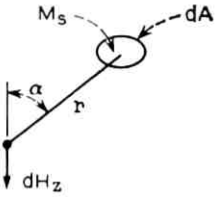


Fig. 23 — Figure showing field H_z is related to M_s and the angle α.

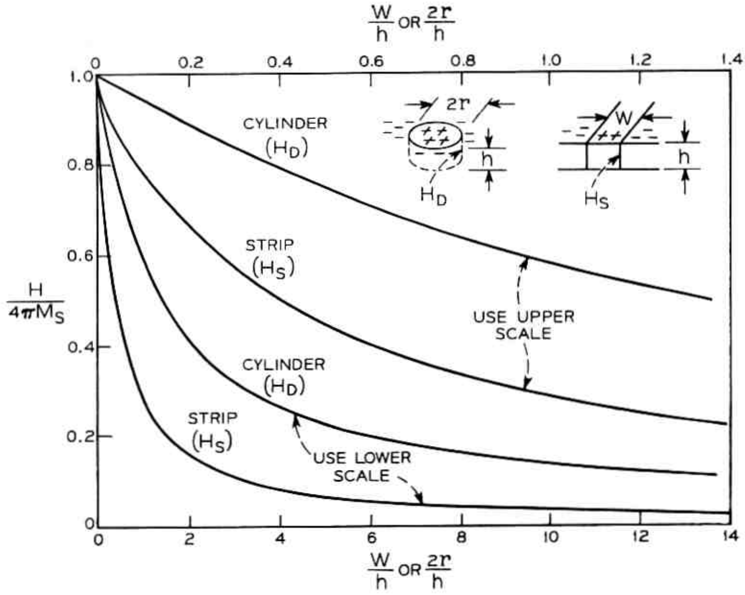


Fig. 24 — Cylinder and strip magnetostatic fields.

$$= \int_0^{\pi/2} (4r_0^2 \sin^2 \theta + h^2)^{\frac{1}{2}}\, d\theta - \int_0^{\pi/2} 2r_0 \sin \theta \, d\theta$$

$$= h \int_0^{\pi/2} \left(1 + \frac{4r_0^2}{h^2} \sin^2 \theta\right)^{\frac{1}{2}} d\theta - 2r_0 \, .$$

Letting $\sin^2\theta = 1 - \cos^2\theta$,

$$\frac{\bar{H}_z h}{8M_s} = h\sqrt{1 + \frac{4r_0^2}{h^2}} \int_0^{\pi/2} \left[1 - \frac{4r_0^2/h^2}{1 + \frac{4r_0^2}{h^2}} \cos^2 \theta\right]^{\frac{1}{2}} d\theta.$$

But

$$\int_0^{\pi/2} (1 - k^2 \cos^2 \theta)^{\frac{1}{2}}\, d\theta = \int_0^{\pi/2} (1 - k^2 \sin^2 \theta)^{\frac{1}{2}}\, d\theta.$$

So the final result can now be written as

$$\frac{\bar{H}_z}{4\pi M_s} = -\frac{4r_0}{\pi h} + \frac{2}{\pi}\sqrt{1 + \frac{4r_0^2}{h^2}} \int_0^{\pi/2} \left[1 - \left[\frac{1}{1 + \frac{h^2}{4r_0^2}}\right] \sin^2 \theta\right]^{\frac{1}{2}} d\theta.$$

The integral is the complete elliptic integral of the second kind $E(k, \pi/2)$ where

$$k^2 = \frac{1}{1 + \frac{h^2}{4r_0^2}}.$$

In the text, the magnetostatic field effective on the cylindrical domain wall is designated H_D thus the final expression, which is (5), becomes

$$\frac{H_D}{4\pi M_s} = \frac{2}{\pi}\left[-\frac{2r_0}{h} + \sqrt{1 + \frac{4r_0^2}{h^2}}\,E(k, \pi/2)\right].$$

Equation (5) and (2), representing the magnetostatic field of the cylinder and strip, respectively, are plotted in detail in Fig. 24.

REFERENCES

1. Michaelis, P. C., A New Method of Propagating Domains in Thin Ferromagnetic Films, International Congress on Magnetism, 1967.
2. Spain, R. J., et al., J. Appl. Phys., 1965, p. 1103.
3. Sherwood, R. C., Remeika, J. P., and Williams, H. J., Domain Behavior in Some Transparent Magnetic Oxides, J. Appl. Phys., *30,* February 1959, pp. 217–225.
4. Kooy, C. and Enz, U., Experimental and Theoretical Study of the Domain Configuration in Thin Layers of $BaFe_{12}O_{19}$, Philips Research Report, *15,* Feb. 1960, pp. 7–29.
5. Thiele, A., private communication.
6. Sherwood, R. C. and Van Uitert, L. G., private communication.
7. Umebayashi, H. and Ishikawa, Y., Motion of a Single Domain Wall in a Parasitic Ferromagnet $YFeO_3$, JPS of Japan, *20,* December, 1965.
8. Bobeck, A. H., The Cubic Waffle-Iron Memory, 1963 Proc. Intermag. Conference, April, 1963.

Contributors to This Issue

E. R. Berlekamp, B.S., M.S., 1962, Ph.D., 1964, Massachusetts Institute of Technology; Assistant Professor of electrical engineering at University of California, Berkeley, and consultant to Jet Propulsion Laboratory, 1964—1967; Bell Telephone Laboratories, 1967—. Mr. Berlekamp's work has been chiefly in the areas of information and coding theory and combinatorial mathematics. Member, IEEE, ACM, MAA.

Andrew H. Bobeck, B.S.E.E., 1948; M.S.E.E., 1949, Purdue University; Bell Telephone Laboratories, 1949—. Upon completion of the Laboratories' Communications Development Training Program in 1952, Mr. Bobeck initially engaged in the design of both communications and pulse transformers. Since 1953 he has designed solid-state memory and logic devices and currently supervises a group which specializes in that activity. Member, IEEE, Eta Kappa Nu, Tau Beta Pi.

C. Dragone, Laurea in E. E., 1961, Padua University (Italy); Bell Telephone Laboratories, 1961—. Mr. Dragone has been engaged in experimental and theoretical work on microwave antennas and solid-state power sources. He is currently involved in solid-state radio systems experiments at the Crawford Hill Laboratory.

V. L. Hein, B.S.M.E., 1962, University of Missouri; M.S.M.E., 1964, Lehigh University; Bell Telephone Laboratories, 1962—. Mr. Hein has been engaged in the fabrication of semiconductor encapsulations and in the thermal design of semiconductors, semiconductor encapsulations and aging systems. He is now studying in the field of solid mechanics at Lehigh University under the Laboratories Doctoral Support Plan. Member, Tau Beta Pi, Pi Tau Sigma, Pi Mu Epsilon.

John E. Hopcroft, B.S., 1961, Seattle University; M.S., 1962, Ph.D., 1964, Stanford University; Bell Telephone Laboratories, 1966—67. He has been engaged in research in the theory of automata and formal languages. He is currently at Cornell University. Member, IEEE, ACM, Pi Mu Epsilon, Sigma Xi, Tau Beta Pi.

James K. Hsiao, B.S., 1944, National Hunan University, China; M.S., 1957, Montana State College; Ph.D., 1962, Iowa State University; Bell Telephone Laboratories, 1962—1967. Mr. Hsiao worked in the Digital System Department at Whippany. Currently he is with Naval Research Lab in their Radar Division. Member, IEEE; associate member, Sigma Xi.

Peter Lloyd, B.Sc. (Eng.), 1961, Ph.D., 1966, Queen Mary College (University of London); Bell Telephone Laboratories, 1966—. Mr. Lloyd has worked for Associated Electrical Industries Ltd., in England, 1961—63. He is presently concerned with computer-aided design and analysis at piezoelectric devices and the evaluation of new piezoelectric materials. Associate Member, IEE.

E. A. J. Marcatili, Aeronautical Engineer, 1947, and E. E., 1948, University of Cordoba (Argentina); research staff, University of Cordoba, 1947—54; Bell Telephone Laboratories, 1954—. He has been engaged in theory and design of filters in multimode waveguides and in waveguide systems research. More recently he has concentrated in the study of optical transmission media. Fellow, IEEE.

Vasant K. Prabhu, B.E. (Dist.), 1962, Indian Institute of Science, Bangalore, India; S.M., 1963, Sc.D., 1966, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1966—. Mr. Prabhu is a member of the Radio Research Laboratory, and is concerned with stability and noise problems in solid-state microwave devices. His areas of interest include systems theory, network theory, noise theory, and optical communication systems. Member, IEEE, Eta Kappa Nu, Sigma Xi, Tau Beta Pi, AAAS.

S. S. Rappaport, B.E.E., 1960, Cooper Union; M.S.E.E., 1962, University of Southern California; Ph.D., 1965, New York University; Bell Telephone Laboratories, 1965—. Mr. Rappaport was first engaged in analytical studies of signal processing for radar. His current interest is in data communications theory. Member, IEEE, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.

Irwin W. Sandberg, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—. Mr. Sandberg has been concerned with analysis of military systems, synthesis and analysis of active and time-varying networks, studies of properties of nonlinear systems, and with a few problems in communication theory. His current interests are in the area of numerical analysis. Member, IEEE, SIAM, Eta Kappa Nu, Sigma Xi, Tau Beta Pi.

Marvin K. Simon, B.E.E., 1960, City College of New York; M.S.E.E., 1961, Princeton University; Bell Telephone Laboratories, 1961—1963; Ph.D., 1966, New York University; Bell Telephone 1966—. Mr. Simon's early work at Bell Telephone Laboratories dealt with station apparatus development including *Touch-Tone*® and *Picturephone*® circuit design. He is currently engaged in theoretical studies of digital transmission systems. Member, IEEE, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

Jeffrey D. Ullman, B.S., 1963, Columbia University; Ph.D., 1966, Princeton University; Bell Telephone Laboratories, 1966—. Mr. Ullman has been engaged in research in formal languages and automata theory. Member, IEEE, ACM, Tau Beta Pi, Sigma Xi.

B. S. T. J. BRIEF

Interpolation of Data With Continuous Speech Signals

By M. R. SCHROEDER and S. L. HANAUER

(Manuscript received March 1, 1967)

In some communications systems, the need arises for temporally interpolating data or signalling information during continuous speech.[1] If the required time gaps are created by simply interrupting the speech signal, severe degradation of speech quality and some loss in intelligibility results.

The reason for the degradation is twofold:

(*i*) The interruptions introduce *discontinuities* in the speech signal—two for every interruption.

(*ii*) The interruptions, unless occurring pitch synchronously, create an *inharmonic* signal.

In the following, a proposal is described which avoids discontinuities and is pitch synchronous—without the need for pitch detection. Average "off-time" ratios of 30 percent have been achieved for *continuous* speech without audible degradation. These results were obtained by computer simulation of a sampled data system. The instrumentation for a real-time analog system is simple.

The gaps created by this method occur at irregular intervals in time. Thus, for a steady flow of data or signalling information, some buffer storage and coding that distinguishes "gaps" (interpolated data) from speech is required.

In the proposed interpolation system, the speech signal $s(t)$ is divided by its envelope

$$a(t) = [s^2(t) + \hat{s}^2(t)]^{\frac{1}{2}}, \tag{1}$$

where $\hat{s}(t)$ is the Hilbert transform[2] of $s(t)$. The resulting signal is then multiplied by a modified envelope

$$\tilde{a}(t) = \{a(t) - c\}_+ \equiv \begin{cases} a(t) - c & \text{if} \quad a > c \\ 0 & \text{if} \quad a \leqq c, \end{cases} \tag{2}$$

where the function $\{\ \}_+$ equals its argument for positive arguments and is zero otherwise.

The combination of these two operations results in the desired interrupted signal

$$s_i(t) = \frac{s(t)}{a(t)} \{a(t) - c\}_+ . \tag{3}$$

The average off-time ratio depends on the magnitude of the constant c. For Gaussian signals, this ratio is given by

$$r_{\text{off}} = 1 - \exp [-c^2/\overline{a^2}]. \tag{4}$$

In one of the computer simulations, c was chosen equal to 0.5 $\bar{a}$. For a Gaussian signal, this choice corresponds to

$$c = \tfrac{1}{4}[\pi \overline{a^2}]^{\frac{1}{2}}. \tag{5}$$

Thus, the average off-time becomes

$$r_{\text{off}} = 1 - \exp [-\pi/16] = 0.18 = 18\%. \tag{6}$$

The actually observed off-time ratios for $c = 0.5\,\bar{a}$ for two test sentences were 26 percent for male speech and 16 percent for female speech. ($\bar{a}$ was obtained by averaging $a(t)$ over 20 msec with a rectangular time window.)

Fig. 1 shows microfilm outputs from a computer simulation. The constant c was chosen equal to 0.9 $\bar{a}$ in order to achieve off-time ratios near 50 percent. The first line shows the original signal $s(t)$, the second line the interrupted signal $s_i(t)$ and the third line the "switching function" $\{1 - c/a(t)\}_+$. The actual off-time ratio for the total utterance (one sentence, male speaker) was $r_{\text{off}} = 55$ percent.

The speech quality for this rather large off-time ratio was judged somewhat nasal but as intelligible as the original. It is possible that

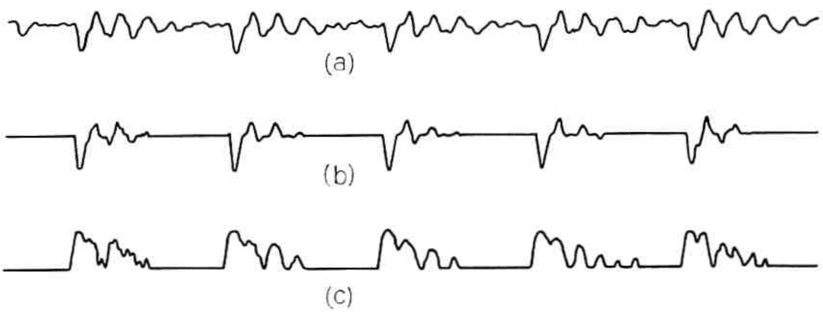


Fig. 1 — (a) Original signal $s(t)$. (b) Interrupted signal $s_i(t)$. (c) "Switching function" $\{1 - 0.9a(t)/a(t)\}_+$

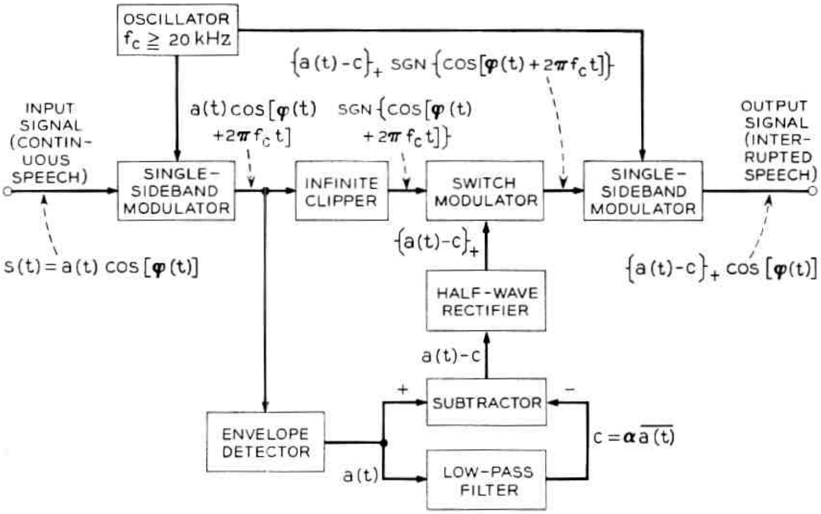

OSCILLATOR
$f_c \geqq 20$ kHz
$\{a(t)-c\}_+$ SGN $\{\cos[\varphi(t)+2\pi f_c t]\}$
INPUT SIGNAL (CONTINUOUS SPEECH)
$a(t)\cos[\varphi(t)+2\pi f_c t]$
SGN $\{\cos[\varphi(t)+2\pi f_c t]\}$
OUTPUT SIGNAL (INTERRUPTED SPEECH)
SINGLE-SIDEBAND MODULATOR
INFINITE CLIPPER
SWITCH MODULATOR
SINGLE-SIDEBAND MODULATOR
$s(t)=a(t)\cos[\varphi(t)]$
$\{a(t)-c\}_+$
$\{a(t)-c\}_+\cos[\varphi(t)]$
HALF-WAVE RECTIFIER
$a(t)-c$
+
SUBTRACTOR
−
$c=\alpha\overline{a(t)}$
ENVELOPE DETECTOR
$a(t)$
LOW-PASS FILTER

Fig. 2 — Possible implementation of a system for interpolation of data with continuous speech signals.

even better results might be achieved by "smoothing" the switching function, by filling in short gaps and by eliminating short speech bursts.

A possible implementation is shown in Fig. 2 in block diagram form. The division by $a(t)$ indicated in (3) can be effected by single-sideband modulation followed by infinite clipping. The function $\{\cdot\}_+$ corresponds to standard half-wave rectification. The multiplication by $\{a(t) - c\}_+$ can be effected by a switch-type modulator. The desired interrupted signal is obtained by a downward frequency shift of the modified-envelope single-sideband signal.

The problem of interpolating signalling information with speech arose in a new mobile communication system for trains and was brought to our attention by Mr. C. E. Paul.

REFERENCE

1. Paul, C. E., Simultaneous Noninterference Transmission of Continuous Speech and Data, unpublished work.
2. Kaplan, Wilfred, *Operational Methods for Linear Systems,* Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, 1962, pp. 395–400.