THE BELL SYSTEM
TECHNICAL JOURNAL

Volume 47 November 1968 Number 9

Influence of Bulk and Surface Properties on Image Sensing Silicon Diode Arrays

By T. M. BUCK, H. C. CASEY, Jr., J. V. DALTON, and M. YAMIN

(Manuscript received March 7, 1968)

Silicon diode arrays for use as the electron-beam accessed target in camera tubes for the **Picturephone**® *visual telephone set have been fabricated and their properties evaluated. These targets offer significant advantages over the antimony trisulfide target commonly used in vidicon-type tubes. But there are certain potential limitations which must be dealt with in developing a silicon target. Three of its critical requirements are adequate sensitivity to visible light, low dark current, and junction uniformity and freedom from defects across at least 300,000 diodes per square centimeter. Sensitivity to visible light is expressed here by the efficiency for conversion of incident photons to electrons in the read-out circuit. Conversion efficiencies exceeding 50 percent in the visible region have been achieved by oxidizing or by diffusing phosphorus into the light-receiving surface to reduce the surface-recombination velocity. Diode leakage currents of* $\leqq 1 \times 10^{-13}$ *A per diode are required, and are obtained for target voltages up to about 5 to 7 V. Surface generated current dominates in the 8-*μ *diameter diodes of the array, but this component of current can be reduced substantially by use of (100) surfaces or by hydrogen annealing. Visible defects in a picture can result from leaky diodes or oxide pinholes which cause bright spots, and diodes covered by oxide which cause dark spots. Our best targets show a video display with only a few defects; processing must be improved to eliminate defects completely.*

I. INTRODUCTION

A television camera tube with a silicon diode array target has been reported recently by Crowell, Gordon and their co-workers.[1-3] A target of the general type used in this tube was first proposed in 1951 by Reynolds,[4] later discussed by Heijne,[5] and recently analyzed by Wendland.[6] It is similar, but not identical in operation, to the evaporated-film photoconductive target, typically antimony trisulfide, which is commonly used in vidicon TV camera tubes.[7]

A vidicon-type tube is of interest for use in the *Picturephone*® visual telephone station set because it is the least expensive and smallest camera tube that has the required sensitivity and resolution. The silicon target has certain potential advantages over the evaporated-film photoconductive target.

This paper describes the effect of bulk and surface properties on the performance of the silicon target. The properties that dominate the conversion of incident photons to electrons in the external circuit and the diode leakage current are analyzed in detail. These analyses, together with the relevant processing techniques and resulting behavior, have been combined into a description of silicon diode arrays for image sensing.

The operation of the silicon target is illustrated in Fig. 1. A scanning electron beam charges the diode-array side of the silicon target down to cathode (ground) potential while the n-region is held a few volts above ground. This puts reverse bias on the diodes. Light shining on the other side of the target and absorbed in the n-region generates holes, some of which diffuse to the diodes and reduce the negative charge on the p-regions. This reduction establishes a stored charge pattern. The scanning electron beam returning to the site of the diode deposits more negative charge, an amount proportional to the light intensity. The recharging current constitutes the video signal. Leakage of the charge pattern established by the light is prevented by the rectifying p-n junction rather than by high bulk resistivity as in the case of antimony trisulfide. The usual time between scans of the electron beam at a given diode site is 1/30 second.

The silicon target has several advantages over evaporated-film photoconductive targets such as antimony trisulfide:

(*i*) It does not show aging effects (burn-in) by intense light to which it might be exposed accidentally, or by the electron beam. The absence of burn-in by the electron beam permits electronic zooming.

(*ii*) Its photoconductive lag is negligible.

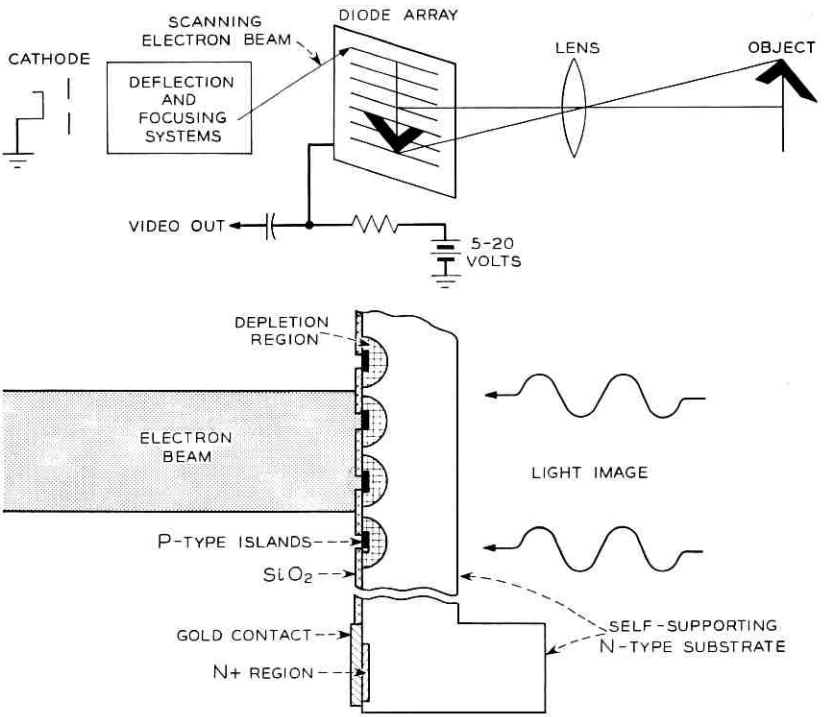


Fig. 1 — Schematic diagrams of silicon target, illustrating principle of operation.

(*iii*) It does not deteriorate on heating to temperatures sufficiently high (≈350°C) for good tube processing required for long life.

There are, however, certain potential problems and limitations which must be considered in developing a successful silicon diode array target for a camera tube. Three of these, related to materials and process factors, are:

(*i*) The target must have adequate sensitivity to visible light. We express sensitivity as conversion efficiency η_c which is defined as the ratio of electrons that flow in the external circuit to the number of incident photons. For photons in the 0.45 to 0.8-μ wavelength range, a value of $\eta_c > 20$ percent would be satisfactory.

(*ii*) The total dark current should be less than 50×10^{-9} A, which means diode leakage current for each of the approximately 1/2 million diodes must be $\leqq 1 \times 10^{-13}$ A, so that only a negligible amount of charge will leak off between scans of the electron beam. For camera

tube applications total diode capacitance is restricted to a rather narrow range. Preliminary results indicate that a substrate resistivity of 10 ohm-cm, which yields 2000 pF/cm^2 at 10 V reverse bias, is close to the optimum.[2]

(*iii*) The whole array must have uniform properties and be free of defects which can cause bright spots or dark spots in the display-tube picture.

Tube performance is, of course, the ultimate test of a good target, but for studies of efficiency and diode leakage it was convenient to make measurements outside the tube. We demonstrate the relationship between these measurements and the actual tube performance.

II. TARGET STRUCTURE

The target, as illustrated in Fig. 2, is a thin disk of n-type silicon with an array of p-n diodes on one side. These are the sensing ele-

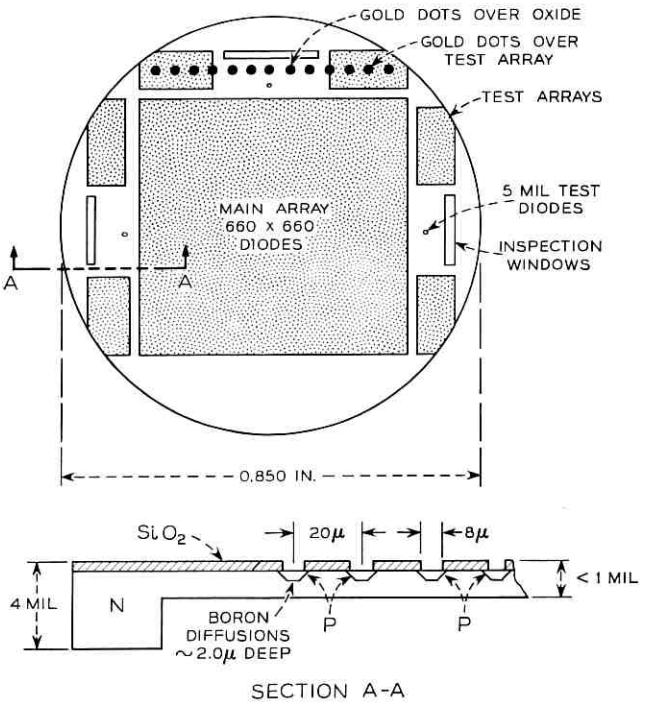


Fig 2 — Details of target structure (660 × 660 array).

ments. Silicon dioxide covers the n-type silicon between the diodes to keep the electron beam from landing there. The oxide also protects the junction edges and reduces surface leakage. In a typical design the target is 0.850 inch in diameter, 0.5 to 1 mil thick in the light sensing region, and has a 4-mil-thick rim for support. The diodes are 8 μ in diameter and are on 20 μ centers in an array of 660 $\times$ 660.

Outside the main array are 5 mil p-n diodes which can be probed for measurements at various stages of processing. There are also gold dots over the oxide for MOS measurements and gold dots over test arrays of 8 μ diodes for reverse current measurements which simulate dark current measurements in the tube. Ohmic contact to the n-region is made through an annular n^+ region which is in the thick ring on the light-receiving side of the target. The dimensions just given are representative of many targets which have been made, although larger and denser arrays are also being studied.

Planar technology is used in the fabrication with the following deviations from typical device processing:

(*i*) One-step diffusions are used, with no drive-in.

(*ii*) There is no postdiffusion reoxidation and therefore no second photoresist step or reregistration.

(*iii*) The phosphorus diffusion for ohmic contact is the last high temperature step, for reasons discussed near the end of Section 3.3.

An additional processing step, the deposition of a semi-insulating film over the diode side, is usually performed before the target is mounted in a tube. The purpose of this film is to dissipate charge, deposited by the electron beam, from the target area between the diodes. Several films have been developed for this purpose by Crowell and Labuda.[2] Targets described in this paper did not have such films except in cases where tube measurements are mentioned.

III. CONVERSION EFFICIENCY

3.1 *Calculation of Conversion Efficiency*

If a silicon diode array target is to replace the antimony trisulfide target in a vidicon-type camera tube, its sensitivity should approximate or exceed that of antimony trisulfide targets. We describe sensitivity in terms of the conversion efficiency η_c which is defined as the ratio of the number of electrons that flow in the external circuit to

the number of incident photons. Efficiency of antimony trisulfide targets is typically 20 percent at 5500 Å and falls off toward both ends of the visible region. For silicon targets the conversion efficiency will depend on the sample thickness, the wavelength of the light and several properties of the semiconductor. In order to identify the important parameters and their effects on the conversion efficiency, the steady-state, short-circuit current in a one-dimensional model of a single p-n junction has been calculated.

The most important relation governing distribution of the optically generated carriers in the semiconductor is the continuity equation. The continuity equation formulation for the transport of carriers has been rigorously treated by van Roosbroeck.[8] From that treatment the steady-state, small-signal differential equation for the minority carrier density in excess of the equilibrium concentration at zero total current and electric field in the one dimension x is

$$D_p \frac{d^2p(x)}{dx^2} - \frac{p(x)}{\tau_p} = -G(x), \tag{1}$$

where $p(x)$ is the excess minority carrier density, D_p the hole diffusion coefficient, τ_p the hole lifetime, and $G(x)$ the net carrier generation rate.

At one boundary, the surface at $x = 0$, the hole flux as determined by the surface-recombination velocity S must equal the diffusive flux:

$$j(0) = qSp(0) = q\, D_p \frac{dp}{dx}\bigg|_{x=0}. \tag{2}$$

The other boundary condition refers to the edge of the junction space-charge region located at depth $x = d$ from the illuminated surface. For the short-circuit condition, the excess hole density at the junction edge is zero,

$$p(d) = 0, \tag{3}$$

and the short-circuit current density i_{sc} is

$$i_{sc} = -q\, D_p \frac{dp}{dx}\bigg|_{x=d}. \tag{4}$$

When the generation rate for carriers is governed by Lambert's law of photon absorption, the net generation rate may be written as

$$G(x) = \frac{(1 - R)N}{A}\, \alpha \exp(-\alpha x), \tag{5}$$

where R is the reflectivity, N the number of incident photons per unit time, A the cross-sectional area, and α the optical absorption coefficient. The conversion efficiency, neglecting absorption in the junction space-charge region, becomes

$$\eta_c = \frac{i_{sc/q}}{N/A} = \frac{-D_p \left.\frac{dp}{dx}\right|_{x=d}}{N/A}. \tag{6}$$

The conversion efficiency may be obtained from equation (1) with the conditions expressed by equations (2) through (5) and the definition of equation (6). Since the generation rate $G(x)$ has a dependence on wavelength through both the reflectivity and absorption coefficient, η_c will depend on wavelength. As Section 3.3 describes in detail, difficulty has been encountered in reconciling the experimental and calculated η_c for illuminated surfaces that have been etched and aged in air. The experimental η_c for the etched surface is always significantly less than the calculated η_c, even with very high S, when a significant portion of the carriers are generated near the illuminated surface.

This situation is similar to Wittry and Kyser's[9, 10] experiences with cathodoluminescence in GaAs. Their cathodoluminescence intensity was less than could be explained by a high surface recombination alone. They assumed that minority carriers generated between the surface and a depth δ are not effective in producing recombination radiation. Similarly, in the present work it was found necessary to modify the generation rate given by equation (5) to account for a "dead layer" at the surface in order to obtain agreement between the calculated and experimental η_c.

The effect of a "dead layer" on the conversion efficiency has been introduced into the analysis by assuming that the carriers generated within a distance δ of the illuminated surface cannot diffuse to the junction and be collected. For a solid in which all the incident photons, less those lost by reflection, are absorbed in creating hole-electron pairs, the number of carriers per unit area generated between the surface and $x = \delta$ is found by integrating $G(x)\,dx$ between the limits 0 and δ. Thus, there will be $(1-R)N[1 - \exp(-\alpha\delta)]$ carriers generated within a distance δ of the surface. The number of holes that may be collected in the absence of a "dead layer" or surface and bulk recombination is equal to the number of absorbed photons and is simply $(1-R)N$. If the carriers generated between $x = 0$

and δ are lost and not available for collection, then the carriers available for collection in the absence of other losses is simply the difference of these two quantities, $(1-R)N \exp(-\alpha\delta)$. Therefore, a generation rate of

$$G'(x) = \left[\frac{(1-R)N}{A}\,\alpha \exp(-\alpha x)\right] \exp(-\alpha\delta), \tag{7}$$

when integrated from 0 to ∞, will give the same number of collectable carriers as $[(1-R)N/A]\ \alpha \exp(-\alpha x)$ integrated from δ to ∞: that is,

$$\int_0^\infty G'(x)\,dx = \int_\delta^\infty G(x)\,dx.$$

This formalism, which has been found useful in representing the effect of the surface space-charge region on optically generated carriers, permits retaining the field-free continuity equation and the representation of surface recombination by S. The conversion efficiency for the generation rate given by equation (7) is

$$\eta_c = \frac{\exp(-\alpha\delta)(1-R)\alpha L_p}{\alpha^2 L_p^2 - 1}\left\{\left[\frac{(S + \alpha L_p^2/\tau_p)\ \text{sech}\ (d/L_p)}{(S \tanh (d/L_p) + L_p/\tau_p)}\right] - \left[\alpha L_p + \tanh(d/L_p) + \frac{S\ \text{sech}^2\ (d/L_p)}{(S \tanh (d/L_p) + L_p/\tau_p)}\right] \exp(-\alpha d)\right\}, \tag{8}$$

where the minority carrier diffusion length L_p is given by $L_p = (D_p\tau_p)^{1/2}$. The efficiency η_c will be expressed in percent. To further describe the "dead layer" and delineate the role of bulk and surface recombination on the target sensitivity, it is necessary to compare the experimental η_c variation as a function of wavelength with η_c calculated from equation (8).

3.2 *Experimental Procedure for Efficiency Measurements*

To evaluate the properties of the silicon diode arrays, it was convenient to make measurements outside the tube. Targets for this purpose had large diodes and had received the same processing as the 8-μ diode arrays except that the semi-insulating film was omitted. Figure 3 illustrates the structure used. The pattern with diodes from 5 to 40 mils in diameter was used so that the effect of diode diameter on η_c for a given light-spot size could be determined. Because equation (1) applies only to a one-dimensional problem, one-dimensional experimental conditions must be achieved. These conditions were

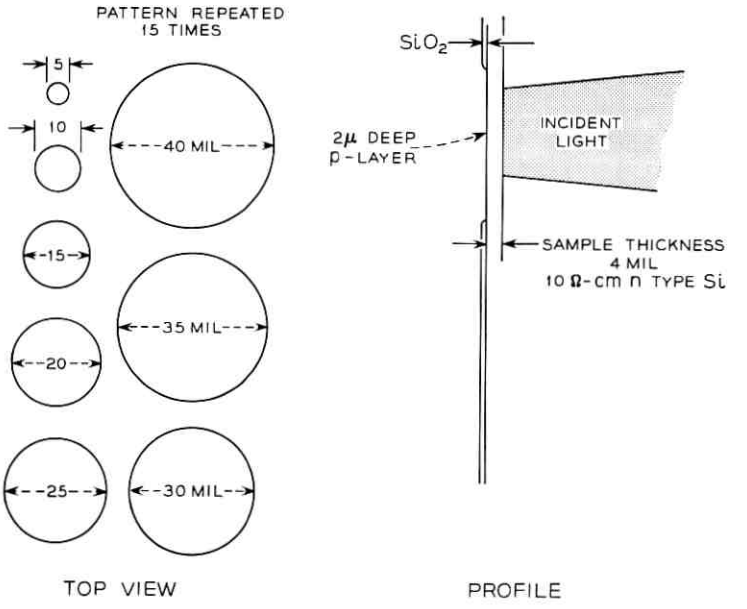


Fig. 3 — Array of large diodes for studies of conversion efficiency and diode leakage.

taken to apply whenever further increase in diode diameter did not increase the conversion efficiency, which occurred for diameters greater than 30 mils. The η_c measurements reported here were taken with the 40-mil diameter diodes.

The experimental arrangement for measuring the conversion efficiency is illustrated in Fig. 4. Emission from the tungsten 250 W quartz-iodine lamp was filtered by a Corning filter C. S. 1-69 (heat-absorbing glass) to reduce the infrared intensity. Plane mirror M_1 and spherical mirror M_2 focused the light onto the entrance slit of the Perkin–Elmer model 99 single-prism spectrometer. The light was chopped at the spectrometer entrance at 37.5 Hz. An Optics Technology band-pass filter was used at each measurement wavelength to prevent light of undesired wavelengths from being transmitted through the spectrometer. The light from the spectrometer exit slit could be directed to the sample or to a calibrated thermocouple by the movable mirror. The spherical mirror M_3 focused the radiation onto the thermocouple whose output was measured with the Princeton Applied Research model HR-8 lock-in amplifier.

The spot size of approximately 4 × 12 mils on the sample was obtained

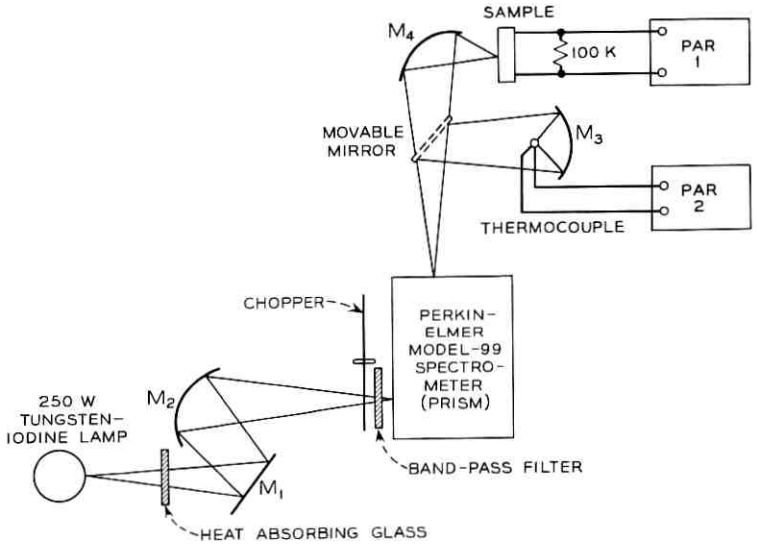


Fig. 4 — Experimental arrangement for measurements of conversion efficiency.

by reducing the height of the spectrometer slits and using M_4, a 90° ellipsoidal mirror with a 6 : 1 reduction. The short-circuit current was determined by measuring the voltage across a 100 KΩ resistor connected between the n and p sides of the diode. To maintain junction short-circuit conditions, the light intensity was kept low enough so that the junction voltage was $\leqq 0.1\ kT/q (\leqq 0.002$ V). Linearity with light intensity was confirmed with neutral density filters and absence of significant leakage currents could be demonstrated by linear variation of junction voltage with load resistance.

In order to make quantitative comparison of the experimental and calculated η_c the absolute photon flux must be known. For this purpose, a silicon solar cell was calibrated by comparison with several calibrated thermopiles. Then the calibrated solar cell was placed in the position of the sample and the spectrometer thermocouple was calibrated. The absolute photon flux can be assigned an uncertainty of ±10 percent.

3.3 *Experimental and Calculated Conversion Efficiency*

In this section the control of target efficiency by process variations is discussed. Experimental efficiency data are compared with calculated curves, and this permits determination of the parameters L_p, S, and δ. Consideration of equation (8) shows that the best discrimination between L_p, S, and δ is obtained when d is two to four times L_p.

Preliminary measurements indicated that a 4-mil target thickness was suitable. The wavelength dependence of the absorption coefficient α was taken from the data of Dash and Newman,[11] and the reflectivity R for an etched silicon surface was taken from Philipp and Taft's data.[12] The hole diffusion coefficient D_p was assigned a value of 10 cm^2 per second. Given these experimental dependences of α and R, the wavelength becomes the independent variable for equation (8), and the variation of η_c with wavelength is determined by the physical parameters d, L_p, S, and δ.

Initially, consider the data in Fig. 5 for the etched surface aged in air (the dots). The calculated η_c curve for a generation rate expressed by equation (5) (no "dead layer") is shown by the dashed line. The diffusion length L_p has been assigned a value of 50 μ to produce agreement between the calculated and experimental η_c in the long wavelength region. The use of the generation rate of equation (5) is equivalent to $\delta = 0$ in equation (8) for η_c. Even with the maximum value of surface-recombination velocity $S_{max} = (kT/2\pi m)^{\frac{1}{2}} \approx 10^7$ cm per second, the calculated η_c does not decrease as rapidly at short wavelengths as the experimental η_c. The discrepancy between the calculated and experimental η_c for wavelengths in the visible region leads to the "dead layer" concept.

Fitting the data in Fig. 5 with a finite δ in equation (8) does not lead to unique values of S and δ. However, their values are limited to a reasonably narrow range. For example, by plotting the experimental data with the calculated η_c (lower curve, Fig. 5) it is not possible to discriminate between values of $S = 10^7$ cm per second, $\delta = 0.8$ μ, and $S = 6 \times 10^4$ cm per second, $\delta = 1.8$ μ. The larger δ value was obtained with a nonlinear least-squares technique described by Marquardt.[13] The high surface-recombination velocity agrees with earlier studies of Buck and McKim[14] and Harten[15] in which it was shown that S is normally very high on an etched silicon surface. Harten's[15] measurement technique was similar to the one described here. To resolve this ambiguity in S and δ, experimental data at wavelengths less than 0.45 μ are necessary. Because the efficiency of the diode is rapidly decreasing and the intensity of the light source is also becoming smaller, measurements in this wavelength range are not presently possible. The significant point is that the "dead layer" thickness is approximately a micron and S is very high.

In order to gain insight into the significance of the "dead layer," the surface potential was determined by surface conductivity meas-

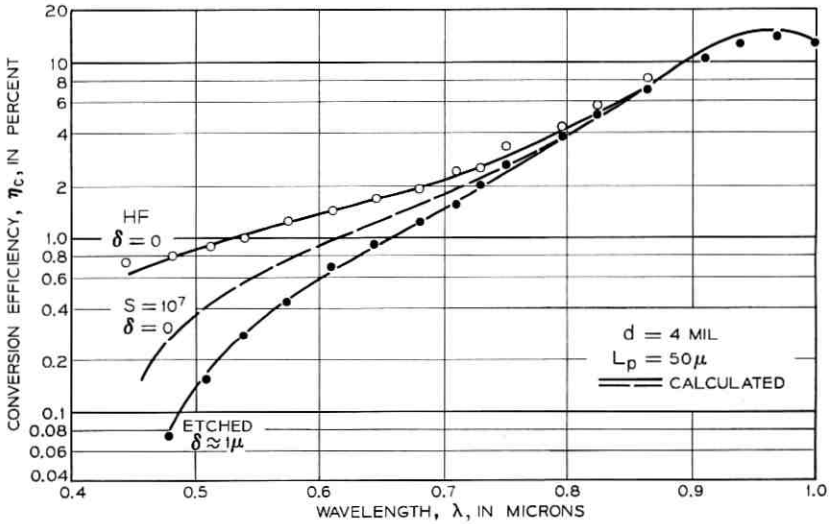


Fig. 5 — Conversion efficiency as function of wavelength for etched and HF-treated surfaces. Target is 4 mils thick.

urements[14] on silicon slices from the same 10 ohm-cm n-type crystal. The surface was found to be in a condition of depletion with the potential varying from −0.25 V at about ½ h after etching to −0.6 V (energy bands bent upward) after several hours of aging in air. The variation of potential within the surface depletion region was obtained from Poisson's equation[16] for a donor concentration of 6.0×10^{14} cm^{-3}. Because the electric field in the surface depletion region goes to zero very slowly it is difficult to define a depletion-region depth. If, however, the potential from Poisson's equation is approximated by a simple parabolic potential of the form

$$V(x) = (V_s/x_s^2)(x - x_s)^2, \tag{9}$$

where V_s is the surface potential, then the depletion layer thickness may be approximated by x_s. The quantity x_s is determined by a reasonable fit from $x \cong 0.1x_s$ to $x \cong 0.8x_s$. For a surface potential of −0.25 V, x_s is 0.8 μ and for −0.60 V, x_s is about 1.1 μ. Therefore, the depletion layer thickness is about the same as the "dead layer."

Because the depth of the "dead layer" and the surface depletion region are about the same, it is reasonable to attribute the "dead layer" to the surface depletion region. This assumption suggests that

the "dead layer" results from the built-in field at the surface whose direction is such as to oppose the diffusion of holes to the junction. To test this hypothesis, reversal of the surface field should eliminate the "dead layer" because the field for holes would be in the same direction as the diffusion toward the junction. It has been shown[14, 15] that a hydrofluoric acid treatment bends the energy bands downward at a silicon surface temporarily. In the present work a surface potential of +0.2 V was determined after such treatment. The upper curve in Fig. 5 shows that the experimental data for an HF-treated surface may be fitted with the same L_p, but no "dead layer" is required. Aging in air causes a shift in surface potential back to a depletion condition and a response that requires a "dead layer" correction.

Although a 4-mil target thickness is useful in efficiency studies for discriminating among the critical parameters, a target for a camera tube should be $\leqq 1$ mil thick for adequate resolution. For a 1 mil thickness Fig. 6 shows two sets of experimental data together with calculated curves which illustrate the importance of S and δ. Consider first the experimental data for the etched surface (the dots). The L_p of 32 μ was obtained from the η_c measurement at a thickness of 4 mils before

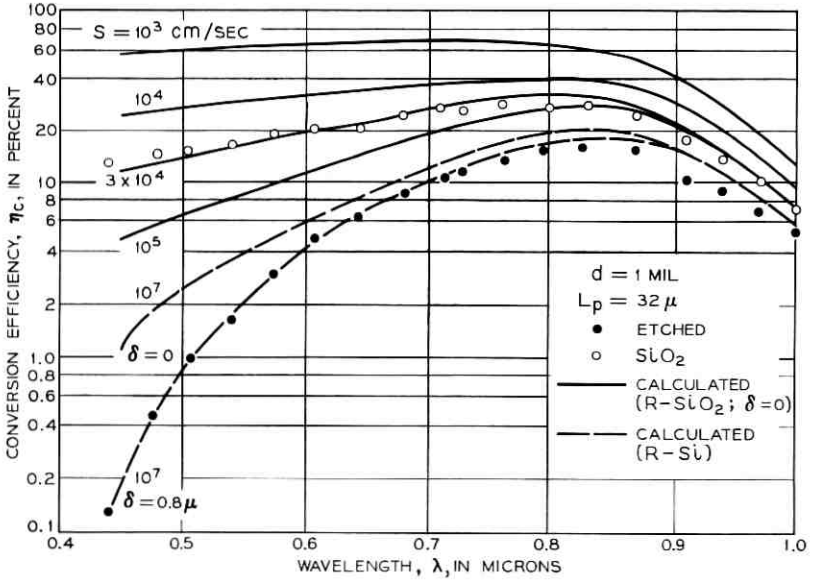


Fig. 6 — Influence of surface-recombination velocity on efficiency for 1-mil thick target.

the target was thinned to 1 mil. The values of $S = 10^7$ cm per second and $\delta = 0.8\ \mu$ were previously shown to be representative of an etched surface.

The dashed lines are calculated for $S = 10^7$ cm per second and $\delta = 0$ or $0.8\ \mu$ to illustrate the effect of the "dead layer" at this thickness. It may be seen that the efficiency in the visible region drops from 10 percent at $0.7\ \mu$ to 1 percent for high S and no "dead layer," but drops to 0.1 percent for the $0.8\ \mu$ "dead layer." Calculations based on equation (8) show that when $d \leqq L_p$, further increase in L_p does not improve efficiency very much. This also means that efficiency is insensitive to small lateral variations in d which may occur during etching. However, S and δ have a strong influence on efficiency and these parameters must be substantially decreased. The HF treatment and other chemical treatments[14,15] on the etched surface which reduce S and δ are too unstable for long life in a camera tube.

A more permanent improvement in efficiency can be made by high temperature oxidation of the surface. Experimental efficiency data taken after a light steam oxidation (900°C for 10 minutes) are also shown in Fig. 6 (circles). Oxidation raised efficiency above 10 percent at the blue end of the visible region. Values of $S = 3 \times 10^4$ cm per second and $\delta = 0$ yield a calculated efficiency curve which fits the experimental points. Values of S for steam oxide with no further treatment have been as low as 10^3 cm per second which gives an efficiency of 50 to 60 percent. The 800 Å oxide also serves as an antireflection coating which contributes an additional slight increase in efficiency.

Experimentally determined reflectivity of the oxidized surface was used in equation (8) to obtain the solid curves shown in Fig. 6. No "dead layer" correction was needed for the oxidized surface. This is attributed to the fact that the oxidized surface is more n-type than the bulk; the energy bands are bent downward at the surface, eliminating the depletion layer. The HF soak eliminated the "dead layer" temporarily for the same reason, that is, it bends the bands down at the surface.

It is assumed that oxidation reduces S both by reducing the density of recombination centers and by shifting surface potential, although the data to confirm this are not complete. Surface potential is shifted from about -0.6 V, the previously mentioned depletion-layer condition for an etched surface, to $+0.2$ V (bands bent downward). Fast state density is not known for the etched surface but is 3×10^{11} cm^{-2} eV^{-1} for the oxidized surface. The S values of 10^3 to 3×10^4 agree

reasonably well with values for oxidized surfaces reported by Rosier[17] but not with the 5 to 10 cm per second values reported by Grove and Fitzgerald.[18] However, approximate agreement with the Grove and Fitzgerald values is obtained for an oxide which has been through boron and phosphorus diffusions, as discussed in Section 4.2.

Oxidation can thus provide satisfactory efficiency, but there are two rather serious objections to oxidation for this application:

(*i*) The oxidation must be done after the target is otherwise complete, and at this stage it frequently causes an increase in diode leakage current.

(*ii*) Vacuum bake-out of the tube and target at 350°C, which is desirable for good tube processing, may increase the interface state density and the recombination velocity.

Although both of these problems could be overcome by improved control of the oxide, an alternative procedure was found which obviated the final oxidation step for low S. Figure 7 shows the improvement in efficiency caused by a phosphorus diffusion on the light-receiving surface of a 4-mil thick target (circles). The phosphorus was diffused at 925°C for 10 min in PBr_3 vapor, yielding a depth of about 0.4 μ. This reduced S to a nominal value of 50 cm per second and the diffusion length was increased to 52 μ, yielding an efficiency of about 20 percent. The efficiency has been 50 to 60 percent on 1-mil thick targets.

Equation (8) is rather insensitive to surface-recombination velocity for $S \leqq 200$ cm per second, and 50 cm per second is only given as an approximate value. The phosphorus data in Fig. 7 were obtained with the phosphate glass on the surface. Removal of the glass did not change the response. When the diffused phosphorus layer was removed, the efficiency dropped to the original level for an etched surface, except for an upward displacement at long wavelengths resulting from the improved bulk lifetime. In some cases L_p has been increased to 100 μ ($\tau = 10$ μsec). The effect on S is evidently due to the built-in field of the diffused phosphorus layer which repels holes from the surface and causes a low recombination velocity. The principle of surface doping to reduce S was proposed by Moore and Webster[19] but we are not aware that a demonstration of it has been published. The slight droop in the phosphorus curve at about 0.5 μ, requiring a "dead layer" correction of 0.1 μ, seems to result from diffusion damage. At present the best conditions that have been found to minimize this effect are 850°C for 30 minutes.

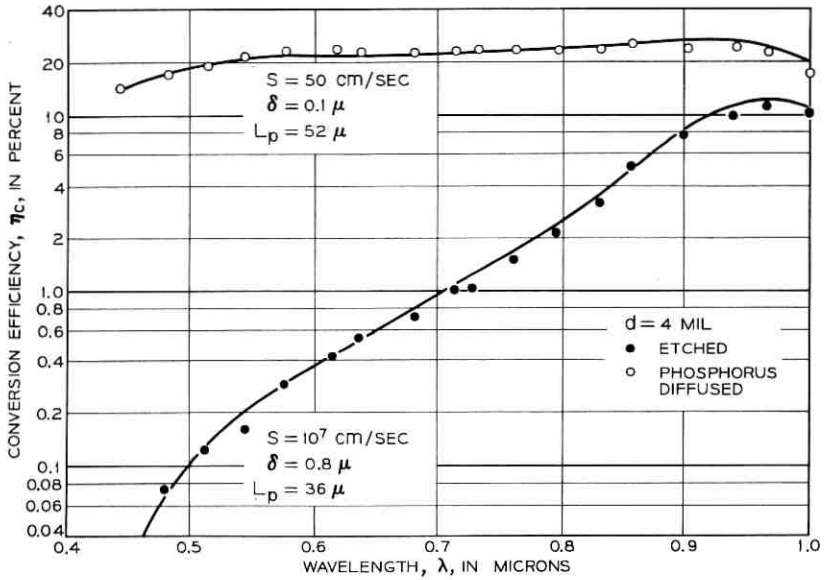


Fig. 7 — Effect of phosphorus diffusion on efficiency. Lower curve shows condition with etched surface before diffusion.

These results were all obtained on targets with large diodes, outside the tube. Results for a silicon target in a tube are shown in Fig. 8 where it is compared with a standard vidicon, and with the ideal case of unity conversion efficiency. Efficiency is expressed in units commonly used for vidicons, microamps of output current per microwatt of incident radiation. The curve for unity efficiency slopes upward because at longer wavelengths there are more photons per second per microwatt of radiation, with each photon capable of exciting an electron-hole pair. Efficiency of the silicon target is more than twice that of the vidicon in the middle of the visible region, and the sensitive range is much broader. The conversion efficiency exceeds 50 percent at a wavelength of 0.7 μ.

The diffused layer of phosphorus provides a very stable reduction in S which is unaffected by vacuum bake-out or deposition of antireflection coatings. Furthermore, the phosphorus treatment does not harm the diode characteristics; instead, it improves them as discussed in the next section. The phosphate glass must be removed from the p-type islands by a brief etch, but this requires no remasking.

IV. DARK CURRENT

4.1 *Bulk and Surface Generated Current*

Low dark current is another important requirement for this target. Current, in the absence of illumination, should not exceed 1×10^{-13} A per diode at 5 to 10 V reverse bias, or about 50×10^{-9} A for the whole array. Currents slightly above this range reduce the dynamic range of the camera tube or the available picture contrast. Currents five times greater prevent integration of the incident light flux over a full television scan period. The most readily observable effect of excessive dark current (more than 100×10^{-9} A) is "whiting out" of the picture on the display tube.

The dark current required for good performance is considerably lower than for most silicon devices. The reverse current in a target array can be separated into the two general categories of bulk generated current and surface generated current. An estimate of the bulk current can be obtained from the expression by Sah, Noyce, and Shockley[20] for current generated in the space-charge region:

$$I = qAwn_i\left\{2(\tau_p\tau_n)^{\frac{1}{2}} \cosh\left[\frac{E_t - E_i}{kT} + \tfrac{1}{2}\ln\left(\frac{\tau_p}{\tau_n}\right)\right]\right\}^{-1}. \qquad (10)$$

In equation (10) w is the depletion width, n_i the intrinsic carrier concentration, τ_p and τ_n the hole and electron minority carrier lifetimes on their respective sides of the junction and $E_t - E_i$ the energy dif-

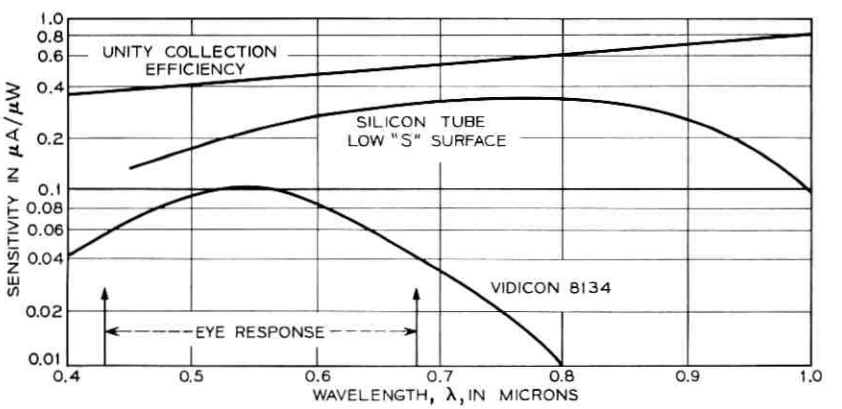


Fig. 8 — Comparison of 1 mil silicon target with antimony trisulfide target.

ference between an assumed single recombination-generation level and the intrinsic Fermi level. For an 8 μ diode at 10 V bias with the common simplifying assignment of $E_t = E_i$ and $\tau_p = 0.1$ μsec $> \tau_n$ as lifetime values representative of diffused structures, the current would be predicted by equation (10) as approximately 3×10^{-12} A. Since this estimate of bulk current alone exceeds the permissible dark current, diode leakage current was regarded as a potential source of difficulty for this device.

Of several diffusion conditions tried, those which yielded the lowest bulk generated reverse currents were 1140°C for 20 minutes with a BBr_3 source. These conditions gave a sheet resistance of 5 to 6 ohms per square and a junction depth of 2 μ. A subsequent phosphorus diffusion, which is required for ohmic contact and low S on the light-receiving surface, is an important part of the process. As a result of this treatment, the diffusion length was typically improved by a factor of three, while the reverse current, as measured on the 5-mil test diodes on the actual targets, was reduced by an order of magnitude to a median value of 10^{-12} A. In addition, the exponent in $I \propto V^n$ was reduced from $\sim$1.0 to $\sim$0.5.

The improvement in both the diffusion length and dark current by the phosphorus diffusion is presumably caused by a gettering action by the phosphate glass on impurities, such as gold and copper. A gettering effect has been suggested in several reports in which phosphorus treatments have improved silicon diode reverse characteristics[21-23] or minority carrier lifetime.[24, 25] The improvements in reverse current already described are quite similar to those reported by Ing and his coworkers about p^+n diodes of 0.9Ω-cm material.[22] In our work, neutron-activation analysis showed that a boron diffusion increased the gold concentration from approximately 4×10^{12} cm^{-3} to about 2×10^{13} cm^{-3}, while the phosphorus diffusion reduced it again to 4×10^{12} cm^{-3}. Insufficient sensitivity obscured any similar effect on copper if it was present. Cleaning the substrate with nitric acid or aqua regia before diffusion gave better post-boron I–V characteristics than did cleaning treatments without a strong oxidizing acid. The phosphorus treatment then caused further improvement.

The above process was developed using as control information the results of efficiency and dark current measurements on 5-mil test diodes available on target arrays, and also on the graduated-diameter diodes shown in Fig. 3. In addition to their function as control specimens, the graduated-diameter diodes were used to establish the rela-

tive importance of bulk and surface currents. This designation is made by assuming the total current to be a linear combination of an area and perimeter component:

$$I_r = I_{\mathfrak{a}} + I_{\mathfrak{p}} = \mathfrak{a}D^2 + \mathfrak{p}D, \tag{11}$$

where D is the diode diameter. It was possible to estimate $\mathfrak{a}$ and $\mathfrak{p}$ graphically from a plot of I_r/D vs D for the diodes of graduated diameter, and then calculate the $I_{\mathfrak{a}}$ and $I_{\mathfrak{p}}$ components for a diode of a particular diameter. In a typical case for a 5 mil diode, $I_{\mathfrak{a}}$ was 4×10^{-13} A and $I_{\mathfrak{p}}$ was 8×10^{-13} A. This corresponds to a bulk current density of 3.15×10^{-9} A per cm^2, and a perimeter current density of 2×10^{-11} A per cm. The relative importance of $I_{\mathfrak{p}}$ becomes greater as the diode size decreases. Extrapolating to an 8 μ diode, $I_{\mathfrak{a}} = 1.5 \times 10^{-15}$ A and $I_{\mathfrak{p}} = 5 \times 10^{-14}$ A. Total dark currents of targets in tubes have been 5 to 50 $\times 10^{-9}$ A at 5 to 10 V or 2 to 20 $\times 10^{-14}$ A per diode, thus falling closer to the perimeter dependent limit.

The larger area diodes, which minimize the contribution of the surface current, permit assignment of $E_t - E_i$ in equation (10). In Fig. 9 the I–V characteristic of a typical 25 mil diode is shown with curves predicted by the Sah-Noyce-Shockley theory for different values of $E_t - E_i$. The reverse current density at 10 V for this diode was 6×10^{-9} A per cm^2. The lifetime τ_p of 5 μsec was obtained from conversion efficiency measurements. Also, τ_p was taken as much greater than τ_n and the argument of the cosh was greater than unity. The depletion-region width in equation (10) was obtained from experimental capacitance-voltage data and the expression $w = \epsilon A/C$. It may be seen in Fig. 9 that the reverse current calculated for the given 5 μsec lifetime and the single recombination center at about an $E_t - E_i$ of 0.08 to 0.1 eV matches the experimental I–V characteristic.

A quantity called the effective lifetime τ_{eff}, which is the lifetime obtained from equation (10) on the assumptions that $E_t = E_i$ and $\tau_{\text{eff}} = \tau_p = \tau_n$, may be used as a figure of merit to compare low leakage diodes of different resistivities. For our diodes, typical values of $\tau_{\text{eff}} = 100$ μsec were obtained. Ing and his co-workers obtained τ_{eff} between 10 and 40 μsec for their gettered diodes just described,[22] and Sah cited a τ_{eff} of 28 μsec for a high lifetime diode.[26]

Other measurements on our large area diodes also suggest the reverse current is dominated by bulk generation current within the space-charge region. Inversion layer surface leakage is not observed

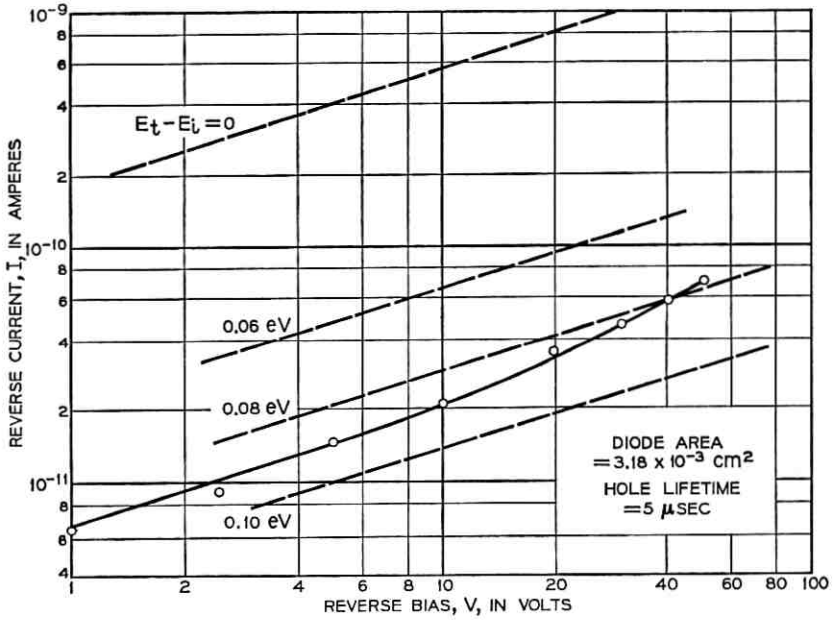


Fig. 9 — Reverse characteristic of 25 mil diode (circles) compared with characteristics predicted by Sah-Noyce-Shockley theory for different energy levels of recombination centers (dashed lines).

in these diodes, nor is it expected, since charge at the interface is positive and induces an n-type accumulation layer. The activation energy of reverse leakage current has been measured as 0.5 to 0.6 eV in the range 0 to 30°C. Such an activation energy is compatible with recombination-generation current rather than inversion layer (channel) leakage or with bulk diffusion current.[27] In addition, the forward I–V measurements are characterized by $1 < m < 1.3$ in the expression $I = I_o \exp(qV/mkT)$. Sah has found such a value of m for recombination-generation current.[26]

The foregoing observations were on relatively *large* diodes measured with no field applied across the passivating oxide. These results are all reasonably consistent with a model of reverse leakage current dominated by generation in the space-charge region of the metallurgical junction, and the currents are satisfactorily low when proper diffusion conditions are used.

4.2 *Leakage Induced by Electron-Beam Charging of the Oxide*

The satisfactory behavior observed for large diodes is a necessary condition if good arrays of small diodes are to be obtained. How-

ever, it may not be sufficient if the field induced by electron beam charging of the oxide leads to high surface-generated currents. In fact, the reverse characteristics of 8 μ diodes in target arrays, with a field across the oxide, are more complicated, although currents are still low enough at 5 to 7 V.

This section shows how surface generation current is influenced by a field across the oxide on the junction side of the array. At present, the problem of achieving low dark current in silicon target arrays is to reduce this type of leakage current. However, if improvement ten times better than present results is achieved, bulk generated reverse current will again become a problem.

The electron beam charging of the oxide can be simulated outside the tube by evaporating gold dots over both the oxide and p-regions and applying a negative bias. This gold dot structure then represents, in most respects, the situation in the camera tube, where the electron beam falls on both the diodes and the oxide. In Fig. 10 the experimental I–V curves are compared with the maximum allowable current of 50×10^{-9} A and the characteristic obtained for bulk generated current only. The bulk generated current curve is based on the assumption that the entire 1×10^{-12} A of reverse current for a 5-mil test diode is bulk current.

The top curve (data points given by $\triangle$) of Fig. 10 is a reverse characteristic measured by applying a negative bias to a 25-mil gold dot which covered 790 of the 8 μ diodes in a test array as illustrated in Fig. 2. The current was scaled up to 435,600 diodes.

The curve shown by the circles is a dark current characteristic measured in a tube. This curve shows behavior similar to the gold dot characteristic. However, the tube characteristics sometimes do not flatten out, for reasons not yet fully understood, although a few leaky diodes in the array are responsible in some cases. It may be seen that for these two curves the current rises steeply with increasing voltage and then changes slope abruptly between 6 and 10 V. Following the model of Grove and Fitzgerald,[18] this behavior can be described as current generated by interface states. This current increases, and finally saturates as a depletion layer is induced under the oxide. In specimens with resistivity more than about 8 Ω-cm, this depletion layer may result from the merging under the oxide of the space-charge regions of the 8 μ diodes, which are less than 12 μ apart edge to edge.

However, curves like those of Fig. 10 are also observed on targets with substrate resistivity as low as 0.1Ω-cm. In this case, the deple-

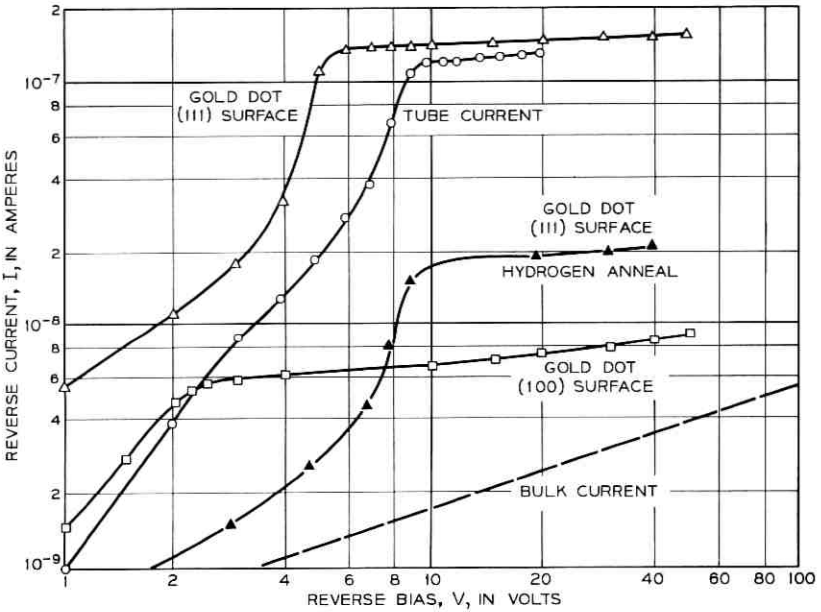


Fig. 10 — Total diode reverse current as a function of bias for target arrays. Current measured by contacting gold dots over diodes and oxide or as dark current in tube. Gold dot measurements cover 790 diodes and are extrapolated to 435,600 diodes. Dashed line would result from bulk space-charge generation alone. 5×10^{-8} amperes is the approximate upper limit for good performance.

tion is induced by a field normal to the semiconductor surface resulting from the negative bias on the gold dot, as in the gate controlled diodes of Grove and Fitzgerald.[18] In either case, the entire silicon surface between diodes is in a depleted condition. Every surface recombination-generation center on that part of the target surface not actually occupied by a diode contributes to the reverse current.

After depletion is established, the subsequent slower increase in reverse current beyond the discontinuity in Fig. 10 is caused by deepening of the space-charge regions under the oxide and around the metallurgical junctions. No decrease in current occurs at higher voltage across the oxide, as observed by Grove and Fitzgerald,[18] since in our case the potential on the oxide equals that on the p-region. Under these circumstances, inversion cannot occur to isolate the surface states from the depletion region.[18] Grove and Fitzgerald relate surface generation current to the surface-recombination velocity s_0 by the equation:

$$I = qn_i s_0 . \tag{12}$$

Using the current at the discontinuity s_o is estimated as 24 cm per second.

The density D_{ss} and electron capture cross section σ_s of interface states on a similar oxide, which had been through the same diffusion process, were measured by the MIS conductance technique of Nicollian and Goetzberger.[28] D_{ss} was 10^{11} cm^{-2} eV^{-1} near the center of the energy gap and σ_s was 2×10^{-16} cm^2. Using Grove and Fitzgerald's[18] definition of surface-recombination velocity for a depleted surface:

$$s_0 = \sigma s v_{th} \pi k T D_{ss} , \tag{13}$$

s_o is 16 cm per second with v_{th} as 10^7 cm per second. This value is in reasonable agreement with the value obtained from the generation current. In addition, conversion-efficiency measurements on a similar oxide, that is, one exposed to diffusion conditions, yielded surface-recombination values of about 50 cm per second. Notice, however, that the diode leakage and efficiency pertain to different values of surface potential.

The current at the discontinuity in gold dot I–V curves has been reduced by two methods. The upper curve in Fig. 10 for a (111) surface can be lowered to the solid triangle curve by hydrogen annealing at 500°C. The decrease in current by an order of magnitude at the discontinuity is assumed to result from a reduction in fast state density expected from this treatment.[29–31] A (100) silicon surface without hydrogen anneal produced the curve shown by squares. The current at the discontinuity has been lowered to 5.5×10^{-9} A at 2.5 V, presumably because of a reduction in interface state density.[29, 32] This current corresponds to an s_o of 1.6 cm per second. Thus, significant decreases in the leakage current have been made by these simple changes in processing and further improvement can reasonably be expected.

V. DEFECTS

A third very important requirement of any camera-tube target is freedom from defects. For the silicon target this means near perfection in an array of nearly ½ million diodes and the demands on planar technology are obviously severe. Leaky diodes, for example, can cause bright spot defects, while diodes which are covered and cannot be contacted by the electron beam, cause dark spot defects. Pinholes in the passivating oxide may cause bright spots by allowing the electron beam to contact the substrate directly. Certain dark

features have been identified with dislocation arrays revealed by etch pits in a neighboring slice of the crystal.

Not all defects can be explained at present. Figure 11 shows some which have been identified. They were intentionally introduced during the target processing. The picture at the top was taken with the camera tube viewing a transparency illuminated by tungsten light. Below it are photomicrographs taken of small areas of the target after it was removed from the tube. The dark area in the display corresponds to a spot in the array where oxide holes are missing because high spots in the oxide lifted the mask and allowed exposure of the photoresist over that area. The bright spot on the left corresponds to a large hole, revealed in the photomicrographs, which was etched in the oxide because contamination on the mask prevented exposure of the photoresist and, therefore, a hole was etched in the oxide. Boron diffused into that entire area and produced a large leaky diode. In ad-

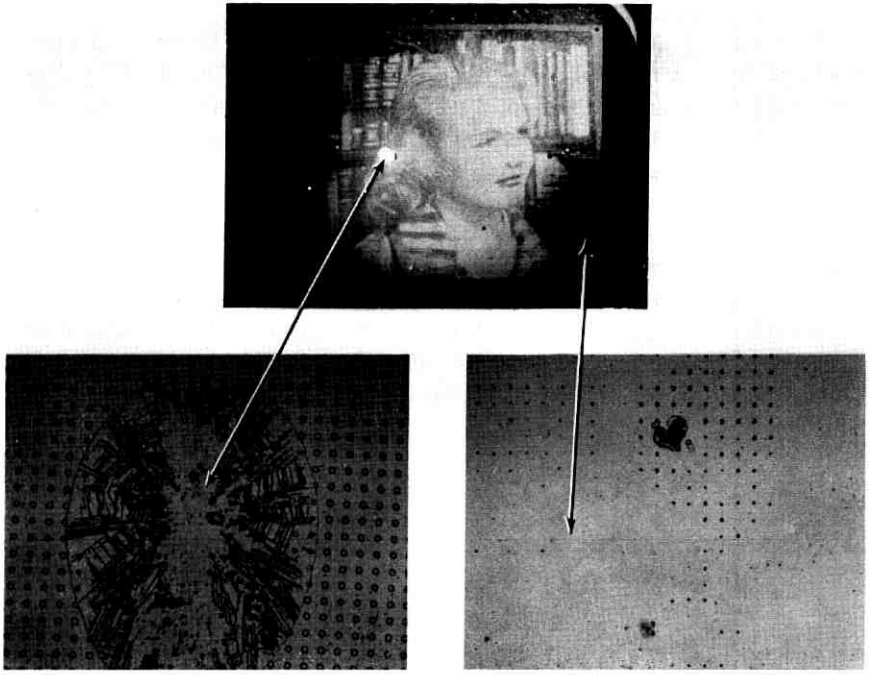


Fig. 11 — Example of gross defects in target. Top: picture from TV monitor. Lower right: photomicrograph of target showing spot where oxide holes are missing, corresponding to dark spot in display. Lower left: photomicrograph showing large hole in oxide which corresponds to bright spot in picture.

Fig. 12 — Picture taken with relatively defect-free target.

dition to defects associated with planar technology steps of target fabrication, bright spots can also be introduced by the deposition of the semi-insulating film for dissipating charge from the area between diodes.

The defects of Fig. 11 were introduced deliberately. Figure 12 shows a picture with only a few small, unintentional, defects visible. We have not yet made a target entirely free of defects. Substantially greater improvement in the defect situation has been made by the group at Bell Telephone Laboratories, Reading, Pa.[33]

VI. SUMMARY AND CONCLUSIONS

The status of the three important factors efficiency, diode leakage current, and defects in the array may be summarized as follows.

(*i*) Satisfactory conversion efficiency has been achieved. Surface recombination velocity (S) at the illuminated surface is the dominant

parameter controlling efficiency, and it should be $\leqq 10^3$ cm per second. Sufficiently low S can be obtained by wet chemical treatments or by oxidation in steam, but the most reliable treatment has been a phosphorus diffusion. This provides a built-in field which repels minority holes from the surface. S values of ≈ 50 cm per second, and efficiency of 40 to 60 percent throughout the visible region, have been obtained with this treatment. Adequate diffusion length, 30 to 100 μ (τ = 1 to 10 μsec) has also been obtained.

(*ii*) Diode leakage current is low enough ($\leqq 1 \times 10^{-13}$ A per diode or $\leqq 50 \times 10^{-9}$ A total dark current) for satisfactory operation at low target voltage ($\leqq 7$ V). Bulk generated current of $< 1 \times 10^{-8}$ A per cm^2 at 10 V is observed in large diodes (5 to 40 mil diameter). This would yield 4×10^{-15} A for an 8 μ diode. However, surface generation complicates the behavior when a field is applied across the passivating oxide as under a gold dot or in electron beam scanning. This causes an initial steep rise in current followed, usually, by an abrupt decrease in slope at 6 to 10 V. The current at which this occurs has been lowered substantially by use of (100) instead of (111) silicon slices and by hydrogen annealing.

(*iii*) Present technology produces targets which are reasonably defect-free but processing must be improved to eliminate defects completely. Leaky diodes or groups of diodes and oxide pinholes cause bright spots, while diodes which are covered and cannot be contacted by the electron beam cause dark spots. A decrease in average dark current should reduce the ability to observe fluctuations from diode to diode.

The emphasis on these three factors is not intended to imply that they are the only critical problems. Two others, discharging of the passivating oxide and resolution, have been studied by Crowell and Labuda.[2]

ACKNOWLEDGMENTS

We are indebted to R. H. Kaiser for taking the efficiency data, to R. Lieberman for help in studies of chemical processing and defects, to E. H. Nicollian for MOS conductance measurements, and M. H. Crowell and E. F. Labuda for useful discussions and data from their studies of tube behavior. We also wish to acknowledge the helpful interest of E. I. Gordon throughout the program.

REFERENCES

1. Crowell, M. H., Buck, T. M., Labuda, E. F., Dalton, J. V., and Walsh, E. J., "A Camera Tube with a Silicon Diode Array Target," B.S.T.J. *46,* No. 4 (February 1967), pp. 491–495.
2. Crowell, M. H. and Labuda, E. F., unpublished work.
3. Gordon, E. I., "A 'Solid-State' Electron Tube for the *PICTUREPHONE*® Set," Bell Laboratories Record *45,* No. 6 (June 1967), pp. 174–180.
4. Reynolds, F. W., "Solid State Light Sensitive Storage Device," U. S. Patent 3,011,089, issued November 28, 1961.
5. Heijne, L., "Photoconductive Properties of Lead-Oxide Layers," Philips Res. Rep. Supplement No. 4 (1961), pp. 149–154.
6. Wendland, P. H., "A Charge-Storage Diode Vidicon Camera Tube," IEEE Trans. Elect. Devices *ED-14,* No. 6 (June 1967), pp. 285–291.
7. Weimer, P. K., Forgue, J. V., and Goodrich, R. R., "The Vidicon Photoconductive Camera Tube," Electronics *23,* (May 1950), pp. 70–73.
8. van Roosbroeck, W., "The Transport of Added Current Carriers in a Homogeneous Semiconductor," Phys. Rev. *91,* (July 1953), pp. 282–289.
9. Wittry, D. B. and Kyser, D. F., "Measurement of Diffusion Lengths in Direct-Gap Semiconductors by Electron-Beam Excitation," J. Appl. Phys. *38,* No. 1 (January 1967), pp. 375–382.
10. Wittry, D. B. and Kyser, D. F., "Surface Recombination Velocities and Diffusion Lengths in GaAs," Proc. International Conf. on Physics of Semiconductors, Kyoto, 1966; J. Phys. Soc. Japan *21* Suppl., 1966, pp. 312–316.
11. Dash, W. C. and Newman, R., "Intrinsic Optical Absorption in Single-Crystal Germanium and Silicon at 77°K and 300°K," Phys. Rev. *99,* No. 4 (August 15, 1955), pp. 1151–1155.
12. Philipp, H. R. P., and Taft, E. A., "Optical Constants of Silicon in the Region 1 to 10 eV," Phys. Rev. *120,* No. 1 (October 1960), pp. 37–38.
13. Marquardt, D. W., "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," J. Soc. Industrial Appl. Math. *11,* No. 2 (June 1968), pp. 431–441. The implementation of Marquardt's method used in this work is embodied in a program written by W. A. Burnette and C. S. Roberts of Bell Telephone Laboratories.
14. Buck, T. M. and McKim, F. S., "Effects of Certain Chemical Treatments and Ambient Atmospheres on Surface Properties of Silicon," J. Electrochem. Soc. *105,* No. 12 (December 1958), pp. 709–714.
15. Harten, H. U., "Surface Recombination of Silicon," Philips Res. Rep. *14,* (August 1959), pp. 346–360.
16. Many, A., Goldstein, Y., and Grover, N. B., *Semiconductor Surfaces* Amsterdam: North Holland Publishing Co., 1965, p. 138.
17. Rosier, L. L., "Surface State and Surface Recombination Velocity Characteristics of Si-SiO_2 Interfaces," IEEE Trans. Elec. Devices, *Ed-13,* No. 2 (February 1966), pp. 260–268.
18. Grove, A. S. and Fitzgerald, D. J., "Surface Effects on p-n Junctions: Characteristics of Surface Space-Charge Regions Under Non-Equilibrium Conditions" Solid-State Elec., *9,* No. 8 (August 1966), pp. 783–806.
19. Moore, A. R. and Webster, W. M., "The Effective Surface Recombination of a Germanium Surface with a Floating Barrier," Proc. I.R.E. *43,* No. 4 (April 1955), pp. 427–435.
20. Sah, C. T., Noyce, R. H. and Shockley, W. "Carrier Generation and Recombination in P-N Junctions and P-N Junction Characteristics," Proc. I.R.E. *45,* No. 9 (September 1957), pp. 1228–1243.
21. Goetzberger, A. and Shockley, W., "Metal Precipitates in Silicon P-N Junctions," J. Appl. Phys. *31,* No. 10 (October 1960), pp. 1821–1824.
22. Ing. S. W., Morrison, R. E., Alt, L. L., and Aldrich, R. W., "Gettering of Metallic Impurities from Planar Silicon Diodes," J. Electrochem. Soc. *110,* No. 6 (June 1963), pp. 533–537.

23. Lawrence, J. E., "Metallographic Analysis of Gettered Silicon," presented at AIME Conf. on Preparation and Properties of Elec. Materials, New York, August 28–30, 1967.
24. Waldner, M. and Sivo, L., "Lifetime Preservation in Diffused Silicon," J. Electrochem. Soc. *107*, No. 4 (April 1960), pp. 298–301.
25. Murray, L. A. and Kressel, H., "Improvement of Minority Carrier Lifetime in Silicon Diodes," Electrochem. Technology, *5*, No. 7–8 (July-August 1967), pp. 406–407.
26. Sah, C. T., "Effect of Surface Recombination and Channel on P-N Junction and Transistor Characteristics," IRE Trans. Elec. Devices, *ED-9*, (January 1962), pp. 94–107.
27. Bergh, A. A. and Bartholomew, C. Y., Jr., "The Effect of Heat Treatments on Low Current Gain with Various Ambients and Contamination," Late News Paper presented at Electrochem. Soc. Meeting, Philadelphia, Pa., October 14, 1966.
28. Nicollian, E. H. and Goetzberger, A., "The Si-SiO_2 Interface—Electrical Properties as Determined by the Metal-Insulator-Silicon Conductance Technique," B.S.T.J. *46*, No. 6 (July-August 1967), pp. 1055–1133.
29. Hofstein, S. R., "Stabilization of MOS Devices," Solid-State Elec., *10*, No. 7 (July 1967), pp. 657–670.
30. Balk, P., "Effects of Hydrogen Annealing on Silicon Surfaces," Extended Abstr., Electrochem. Soc. Meeting, May 9–13, 1965, pp. 237–240.
31. Schmidt, R., unpublished work.
32. Gray, P. V. and Brown, D. M., "Density of SiO_2-Si Interface States," Appl. Phys. Letters *8*, No. 2 (January 1966), pp. 31–33.
33. Batdorf, R. L., Beadle, W. E., Mathews, J. R., unpublished work.

A Charge Storage Target for Electron Image Sensing

By EUGENE I. GORDON and MERTON H. CROWELL

(Manuscript received April 12, 1968)

A charge storage target consisting of a dense array of silicon photodiodes has been described as the image-sensing element in a vidicon type of camera tube for the **Picturephone**® *station set. The target stores a spatially distributed charge pattern corresponding to an optical image in the form of a partial discharge of the reverse-bias voltage of the diodes. The discharge results from leakage current associated with hole-electron pairs created in the silicon substrate by incident photons during the raster interval. Recharging of the diodes to the full reverse-bias voltage along a prescribed raster by the scanning, low energy, electron beam creates the desired video signal.*

This paper describes creation of the hole-electron pairs in the silicon substrate by impinging high energy electrons. Since these electrons, incident from the side opposite the diode array, create a multiplicity of pairs, charge gain results. As in photon sensing, the discreteness of the array allows preservation of detail in the spatial distribution of impinging electrons. Measurements of charge gain as a function of electron energy and target resolution are presented.

Applications in scan conversion, low light level TV, X-ray image intensification, and electron microscopy are indicated.

I. INTRODUCTION

The subject of this paper is the use and properties of a self-supporting silicon wafer containing an array of about one-half million diodes in an area of 12.5 millimeters on a side. See Fig. 1. The thickness of the substrate under the diode array is in the range 10-25 microns depending on the application. The wafer perimeter which is considerably thicker, provides increased physical strength. As an image sensing target in a vidicon type of camera tube,[1] developed for the *Picturephone*® station set,[2] it converts incoming photons that are absorbed in the n-type conductivity substrate into hole-electron pairs.

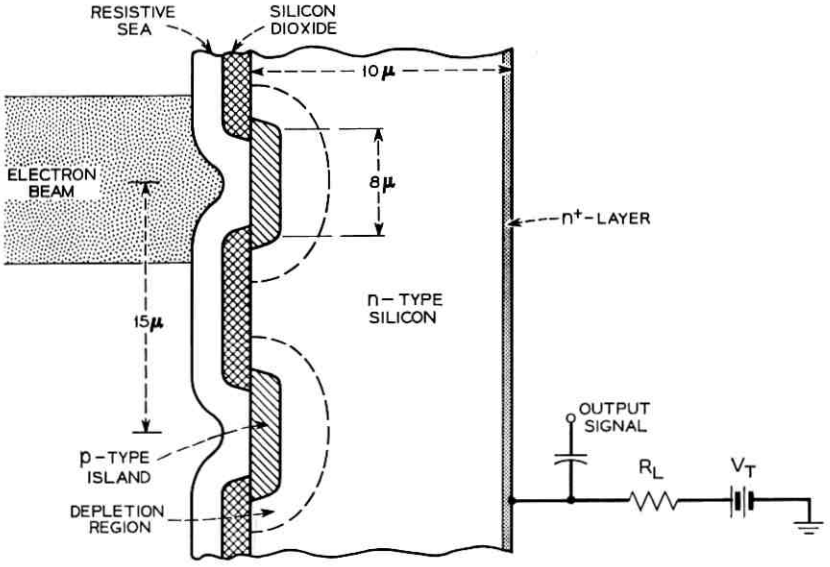


Fig. 1 — Cross section of active target area illustrating the substrate, diodes, oxide mask, resistive sea, and scanning electron beam.

Except for those holes created within a surface layer, a few thousand angstroms thick opposite the diodes, there is virtually unity probability that any minority carrier hole can diffuse to the depletion region of one of the normally reverse-biased diodes. The hole is swept across the depletion region to the diode p-type conductivity region and contributes to the total leakage current of the diode. The totality of holes reaching the diode during a raster interval, partially discharging the diode, constitute a stored charge proportional to the integrated local intensity of the photon flux. Recharging of the diode by a scanning, low energy electron beam creates a current in an external circuit which constitutes the desired video signal. The recharged diode is primed for integration of the hole flux during the next interval by the same process. Figure 2 illustrates the performance on one such target illuminated by a conventional TV test pattern and scanned in a 525-line raster with a frame interval of 1/30 second. Other forms of radiation will create hole-electron pairs in a silicon substrate. Figure 2 therefore indicates the potential performance for imaging these as well.

In this paper the radiation of interest is energetic electrons. An

energetic electron, impinging on silicon will create 286 hole-electron pairs per kiloelectron volt of kinetic energy.[3] When the holes can diffuse to the nearest diodes with high probability, the resulting charge exceeds the charge incident on the target and amplification results. Thus the target, in conjunction with scanning, video processing, and display is an electron image transducer. It has the potential of being useful in any one of a large number of systems or devices in which it is desired to convert spatial intensity variations in incident radiation into a visible image. For example, image intensifiers transduce an optical image into an equivalent electron image by absorbing the incident light on a large area photocathode. The resulting low energy electron image is refocused at high energy onto a second plane by an appropriate electron-optical system. A phosphor screen transducer placed in this plane produces an intensified optical image. The electron image in an electron microscope similarly is viewed by a phosphor screen transducer. The addition of a transducer for X-rays to light

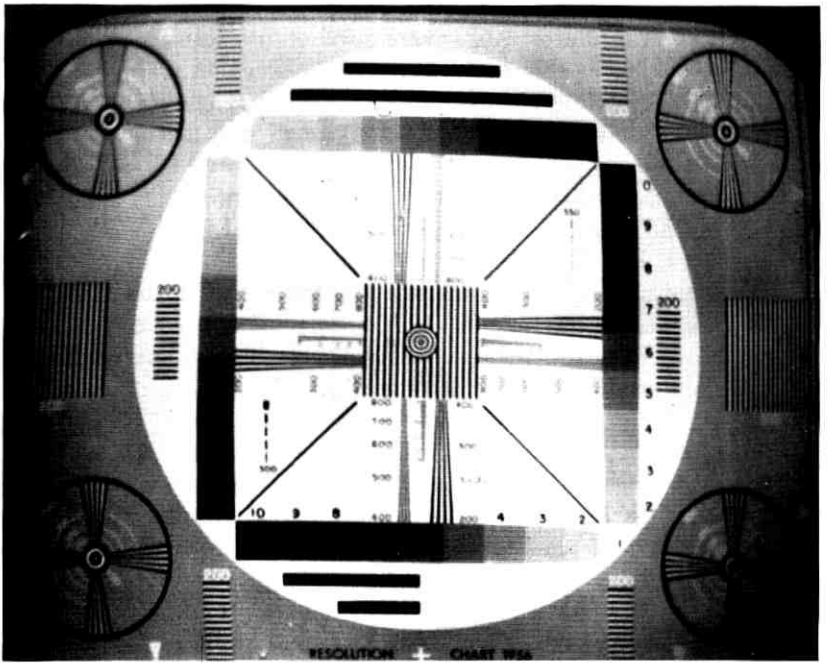


Fig. 2 — Monitor display illustrating performance of diode array target in a camera tube.

at the front end of an image intensifier allows X-ray image intensification. The transducer would be a phosphor screen photon-coupled to the photo-cathode of the image intensifier.

In all of these applications it is becoming common practice to view the output phosphor screen with a closed circuit TV system. This allows observation of a magnified, bright image with the ability to perform video processing and to produce a permanent record on video tape. In these cases, especially when a direct visual output is not really necessary, a single stage electron-to-video transducer with charge amplification replacing the output phosphor screen, lens, and pickup tube would effect a considerable simplification in the system with savings in volume and cost and would afford the possibility of better performance.

Indeed in a class of pickup tubes exemplified by the secondary election conduction (SEC) camera tube,[4] the fundamental image sensor is a photocathode, and the electron image is focused at high energy onto a charge storage target with charge gain. The target is scanned in a vidicon fashion to produce the video output. In this case the target is an insulator, KCl, evaporated onto a thin metallic backplate. The charge in this case is created in a region of high electric field, and the resulting electrons are swept out. The remaining positive charge is immobile and constitutes the storage mechanism. Such a target is similarly an electron-to-video transducer with charge gain and has been used in the applications suggested above.

In a second class of applications exemplified by double beam storage tubes, Fig. 3, the input electrons are produced by a writing gun and form an amplified, stored charge pattern. The stored pattern could represent, for example, a video display, as in a scan conversion device,[5] an oscilloscope trace for highspeed, nonrepetitive events[6] or a closed, nonintersecting path for variable delay of analog or digital signals.[7]

The double-beam device of Fig. 3 has been chosen as the vehicle for study of the target imaging characteristics under electron bombardment. The target has also been studied under conditions which allow multiple readout of the stored charge. In what follows the double-beam device will be referred to as a scan converter.

II. THE SCAN CONVERTER

By way of introduction and for comparison, it is worthwhile to review some aspects of the target optimized for use in a camera tube.

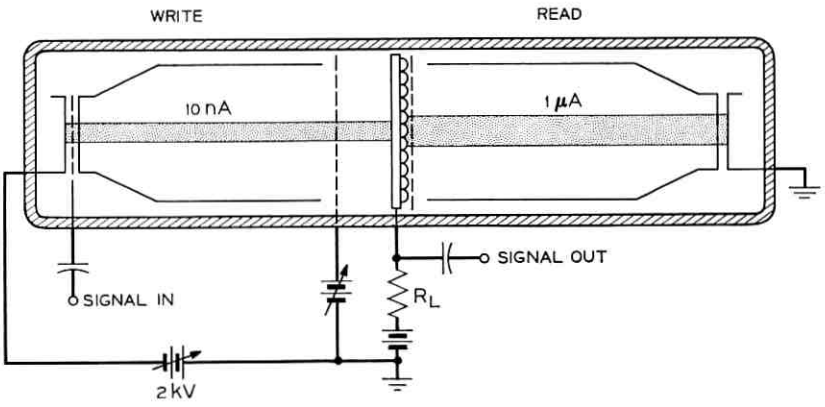


Fig. 3 — Scan converter. Notice back-to-back gun structure.

The array of reverse-biased diodes stores positive charge, created in the form of minority carrier holes by photons incident on the n-type silicon substrate from the side opposite the diode array. The holes diffuse to the diode depletion region and are swept across to the p-region or island of the diode. The stored charge is manifest as a partial discharging of the diodes from their full reverse-bias voltage, which equals the fixed potential of the target substrate, V_T, relative to the potential of the electron beam cathode. The scanning, low energy electron beam, landing on the exposed p-islands, periodically charges them toward cathode potential. Full recharging brings the potential of the p-islands down to cathode potential at which point the electrons can no longer land on the surface. This reestablishes the full reverse bias across the diodes.

Since the substrate potential is held fixed, the p-islands of partially discharged diodes exhibit a positive potential variation on the surface facing the electron beam. These islands are charged back to cathode potential on the next pass of the scanning electron beam. The recharging current constitutes the desired video signal and is proportional to the number of holes collected by the diodes at the position of the scanning beam. Since the number of holes stored by these diodes is proportional to the number of photons incident during the preceding frame period, the video current measures the integrated light intensity at the position of the diodes. The use of a discrete diode array preserves the spatial integrity of the incident light pattern to the extent that lateral diffusion of the holes is negligible and spatial

frequencies with periods comparable to or smaller than the diode spacing are not important.

In the design of the camera tube target the proper choice of diode capacity is important as can be seen by the following example. Suppose that the peak video signal is set at 200 nanoamperes and the scanned area has 670,000 diodes. The scan time is approximately 1/30 second. The maximum charge restored to the diodes during scanning is 2×10^{-7} amperes $\times$ 1/30 second $= 6.7 \times 10^{-9}$ coulombs or 10^{-14} coulombs per diode. The voltage swing of a p-island should never be more than about 5 volts since the beam may be pulled, producing landing errors for greater values. Thus the minimum required diode capacity is approximately 10^{-14} coulombs/5 volts $= 2 \times 10^{-15}$ farads.

Under the conditions specified, the scanning beam current required to recharge the diode to a major fraction of the full reverse-bias voltage (during the submicrosecond interval the beam is incident on the diode) is 1 to 2 μamperes. Beam currents below this value lead to image lag resulting from incomplete recharging of the diode and reduced video signal levels. If the diode capacity is doubled relative to the minimum value, keeping everything else fixed and neglecting the dependence of capacitance upon reverse-bias voltage, the voltage swing of the diode is halved.

The beam current required to recharge the diode to the same extent as in the previous case is increased significantly, possibly more than a factor of two, because the beam landing efficiency is a strong function of the landing energy and is significantly reduced if smaller voltage swings are used. (The beam landing efficiency is defined as the ratio of the surface charging current to the incident current. It is less than unity because of secondary emission and elastic reflection of electrons.) Large beam currents are not desirable and in general not practical; hence the diode capacity must be critically controlled.

For the diode geometry used in the camera tube the silicon resistivity to achieve the appropriate capacitance range is about 10 Ω-cm. The optimum capacity may be achieved by adjusting the potential of the target substrate which varies the full reverse-bias voltage.

A major requirement on the diode performance is the ability to sustain the reverse bias for an interval that is long compared with the scanning interval. With a diode dark current of 10^{-13} amperes, the time for the diode reverse bias to decay to less than half its original value without recharging is about one second, which is sig-

nificantly greater than the usual 1/30 second recharging interval. The ability to hold the diode leakage current well below 10^{-13} amperes at room temperature over the necessary range of target voltage represents one of the major accomplishments of the target development program for the camera tube. Leakage current for the full array is usually well below 50 nanoamperes. A major requirement on the substrate is that volume and surface recombination rates of minority carrier holes be reduced to the point that a large fraction of the holes created by incident radiation can reach the diode depletion region.

The scan converter device is based on the charge storage and electron beam readout properties of the target. The ability to create hole-electron pairs in the target substrate by bombardment with energetic electrons forms the basis of the writing function. Writing is accomplished with a CRT type of electron beam, current modulated with the incoming video signal and incident on the side opposite the diode array (Fig. 3). Each incident electron creates a multiplicity of hole-electron pairs, some of which discharge the diodes, creating a pattern of stored charge just as in the camera tube. The charge stored in the diode array actually can be greater than the charge deposited by the incident writing beam, requiring however that the number of hole-electron pairs created per incident electron times the probability of collection for the hole, be greater than unity. The ratio of stored charge to incident charge will be called the charge gain.

Despite a possible difference in scanning rates, in equilibrium the current level of the video signal generated by the reading beam will be larger than the current in the writing beam by just the charge gain factor. (Application of the concept of charge conservation will indicate the validity of the statement.) On the other hand, the reading beam current is required to be greater than the video signal because the beam landing efficiency is substantially less than unity. Hence the writing beam will usually have much lower currents than the reading beam. In addition, the writing beam electrons will land with energies in the kiloelectron-volt range while the reading beam electrons will land with energies in the range 0–5 electron-volts. The result is that the writing beam may be much more finely focused than the reading beam. Since the penetration range of the writing beam electrons in the silicon substrate is normally under one micron, the resolution of the scan converter should be essential identical to that of the camera tube for very short wavelength light which is absorbed close to the surface.

Figure 3 indicates that some of the writing beam current could return to the writing beam cathode through the target resistor, creating an undesirable video signal. In actual fact, the potential of the mesh immediately preceding the target is adjusted to such a value that the effective secondary emission coefficient of the surface is almost exactly unity. This balance is achieved almost instantaneously so there is virtually no net writing beam current to the target (but rather the current is to the mesh) and virtually no crosstalk. In case the precise secondary emission balance cannot be achieved uniformly over the target, the secondary emission reduction in writing beam current to the target coupled with the charge gain for the desired video signal brings the crosstalk signal down to a tolerably low value.

The writing beam is scanned at a rate appropriate to the incoming video signal. On the other hand, the reading beam may be scanned at any rate desirable. Scanning the reading beam more rapidly than the writing beam produces a multiplicity of time-compressed frames as would be required for slow scan TV.

There is a technique for achieving multiple readouts while preserving the full gray scale. During the early stages of the development of the camera tube, it became clear that it was difficult, if not impossible, to obtain a stable video response unless the silicon dioxide (which covers the area between diodes and prevents beam electrons from landing on the substrate as well as protecting the diode junctions) was provided with a charge leakage path of some kind. Without the charge leakage path it was not possible to control the surface potential of the oxide and a great many deleterious effects resulted.

The scheme adopted to provide the leakage path has come to be known as "the resistive sea" and is simply the formation of a thin resistive film over the entire array. The sheet resistance of the film is chosen to allow surface leakage with a charging time constant of order one second which allows control without causing loss of resolution through lateral spreading. Additional constraints on the dielectric relaxation time and thickness of the film are required to allow complete charging of the diodes to avoid lag. These same parameters can be optimized to allow multiple readout.

Consider Fig. 4 which illustrates the target with a resistive sea and some lumped circuit equivalents for the various parameters of interest. The diode has a capacity C_d. A pulsed current source i_d accounts for the partial discharging of the diode when the writing

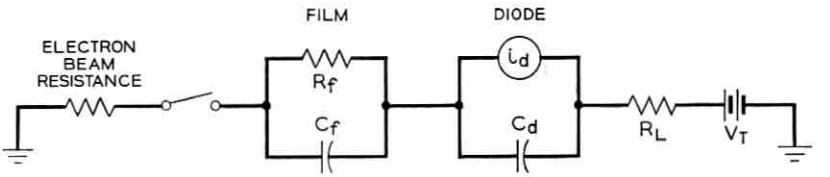


Fig. 4 — Schematic of lumped circuit equivalent for diode and resistive sea. $R_F C_F = \rho\epsilon < \tau_F = 1/30$; $C_d/C_F \ll 1$ low lag; $C_d/C_F \gg 1$ for multiple readouts.

beam is incident on the substrate adjacent to the diode. Assuming a frame compression ratio of N is desired, the video signal must be created N times for each scan of the writing beam. To maintain adequate signal-to-noise ratio the video output current must be at a level comparable with that achieved in the camera tube. This requires N times as much charge storage; hence the diode capacity must be about N times larger than the equivalent capacity of the diode in the camera tube. The most direct way of achieving the increased capacity without changing the diode geometry is the use of N^2 times higher conductivity in the n-type silicon substrate than might be desirable for conventional camera use.

The thickness and resistivity of the resistive sea is arranged to allow negligible leakage from one diode to the next during the 1/30 second between scans of the reading beam. Thus the lateral or spreading resistance of the film will be ignored as well as the shunt capacity. This considerably simplifies the discussion. The leakage resistance R_f and film capacity C_f must have an RC time constant much less than the 1/30 second between successive scans of the beam, yet long compared with the approximately 10^{-7} second or less that the reading beam is incident on the diode. For this particular geometry the RC time constant is about equal to the dielectric relaxation time constant, $\rho\epsilon\epsilon_o$, of the film material (ρ is the resistivity, ϵ the dielectric constant and ϵ_o the permeability of free space). Assuming $\epsilon \approx 10$ and choosing $\rho\epsilon\epsilon_o \approx 5 \times 10^{-3}$ second yields $\rho \approx 6 \times 10^9$ Ω-cm.

The ratio of diode capacity to film capacity C_d/C_f should be about equal to $N - 1$, for reasons which will become clear shortly. Thus the film capacity should be about the same as that of the camera tube diode, 2×10^{-15} farads, requiring a film thickness of about 2 microns over the 8 micron diameter diodes. The sheet resistance of the film is about 3×10^{13} ohms per square which for a 1/30 second frame is adequate to control surface charging without reducing resolution.

The substrate potential is held at V_T, the target bias supply voltage. In the absence of stored signal the reverse-bias voltage of the diode equals V_T, the voltage across the film is zero, and the surface potential of the film equals the cathode potential, that is, zero. Reading beam electrons incident on the surface have a landing efficiency of essentially zero and no further negative charging of the surface can occur.

Suppose now that writing beam electrons are incident on the substrate, producing hole-electron pairs and that a fixed fraction of these with total charge Q_o diffuse to the diode, discharging the diode by a voltage $\Delta V_o \approx Q_o/C_d$. The charge Q_o will be referred to as the signal charge and ΔV_o as the signal voltage. It is assumed for the purpose of discussion that $\Delta V_o \ll V_T$ so that the diode capacity C_d is constant and has a value appropriate to V_T. In practice the film surface has negligible capacitance to ground so that no displacement current need flow through the film capacitance when the interface potential rises and the voltage across the film therefore remains at zero. Thus the film surface is brought to a potential ΔV_o. The maximum value of ΔV_o is about 5 volts to avoid beam bending as in the camera tube.

When the reading beam comes to the diode, the surface of the film is charged down essentially to cathode potential. Thus the series combination of capacitors C_f and C_d is recharged by an amount ΔV_o, requiring that the reading electron beam place a charge on the film surface

$$\Delta Q_1 = \Delta V_o/(1/C_d + 1/C_f)$$

$$= Q_o/(1 + C_d/C_f).$$

(The reading beam current is set at a value high enough to provide the charge, ΔQ_1, during the short reading time interval. During this interval, conduction current through the film is negligible compared with the displacement current.) The charge ΔQ_1 flows through the target resistor R_L producing an output voltage proportional to the signal charge Q_o. The signal charge stored in the diode capacitance, originally Q_o, is reduced by the amount ΔQ_1 to a value

$$Q_1 = Q_o/(1 + C_f/C_d).$$

The voltage across the diode, originally $V_T - \Delta V_o$, is now $V_T - \Delta V_o/(1 + C_f/C_d)$. The voltage across the film is $\Delta V_o/(1 + C_f/C_d)$ which, because of the short RC time constant of the film, decays to zero before the next return of the reading beam. The surface potential therefore achieves the value

$$\Delta V_1 = \Delta V_o/(1 + C_f/C_d)$$

as compared with to the original value ΔV_o immediately preceding the first read. The process is then repeated. At the nth reading after the original signal charge Q_o was established by the writing beam, the output voltage is proportional to

$$\Delta Q_n = Q_o/(1 + C_d/C_f)^n$$

and the film surface potential at the instant preceding the nth read is

$$\Delta V_n = \Delta V_o/(1 + C_f/C_d)^{n-1}.$$

Since by design C_f/C_d is uniform over the target, the relative gray scale is preserved, the output signal is linearly proportional to the input signal and the output signal decays from one read to the next in a well-defined exponential fashion. The exponential time constant is

$$\tau = \tau_f(1 + C_d/C_f)$$

in which τ_f is the reading frame time. Thus for a compression ratio of N an appropriate value might be $C_d/C_f = N - 1$ which implies that the signal decays to $1/e$ in one writing frame time. For this case $\Delta Q_1 = Q_o/N$ which establishes that the signal level is the same as that in the camera tube. Since $\Delta Q_1/\Delta V_o = C_f(N - 1)/N$ this establishes the correctness of the choice of C_f about equal to the capacity of the camera tube diode (at least for $N \gg 1$).

Incidentally, in the camera tube and in many other applications, it is desirable to read virtually all of the signal stored in the diode on the first read (that is, one wants $N = 1$). This is accomplished by making C_d/C_f as small as possible. Thus the film thickness in a low lag target should always be under 0.1 micron as compared with 2 microns in the scan converter example above.

III. EXPERIMENTAL RESULTS

Figure 5 shows the scan converter. The tube is two one-inch vidicon-type guns facing opposite sides of the diode array target. The reading gun has a close-spaced decelerating mesh, as required for good resolution, while the writing gun mesh has been spaced back about one inch. This space permits light to be directed onto the target for measurement of the collection efficiency for holes generated by photons as in a camera tube.

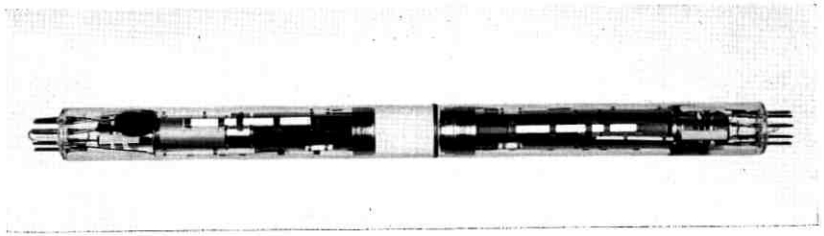


Fig. 5 — Scan converter. The diameter of the glass envelope is one inch. The large spacing allows optical measurements to be made on the target.

Most of the targets used to date are identical to those used in the camera tube; hence they have not been used for multiple readout. However, the validity of the capacitive subdivision of the available signal and the multiple read capability has been established with at least one target.

The resolution of the scan converter target is measured as follows. Both the writing gun and the reading gun are normally operated with magnetic deflection and focus. The writing beam current is modulated sinusoidally in time at frequencies locked to the horizontal scan rate producing a fixed sinusoidal charge pattern on the array. Feed-through is eliminated by adjusting the potential of the writing beam mesh as described earlier. The reading beam scanning rate is locked to that of the writing beam to avoid fluctuations in the relative number of reads per write. The reading beam scans over the fixed charge pattern producing a sinusoidal output signal. The measured peak-to-peak amplitude of the signal normalized to the value measured at low spatial frequencies is called the contrast ratio or modulation transfer function (MTF).

The MTF as a function of spatial frequency in cycles per inch of target is shown in Fig. 6 for a target with a substrate thickness of 20 microns and a diode spacing of 20 microns. Notice that the MTF is 55 per cent at 300 cycles per inch or 12 lines pairs per mm. The falloff may be attributed to four sources: (*i*) writing beam size, (*ii*) reading beam size, (*iii*) finite number of diodes, and (*iv*) lateral diffusion of holes. For the particular target the first two are least significant since under magnetic focus the reading beam is capable of resolving individual diodes and the writing beam has even greater resolving power. The third source of falloff may be appreciated by noticing that the linear diode density is 50 per mm, which means that there are about four diodes per spatial period at 12 cycles per

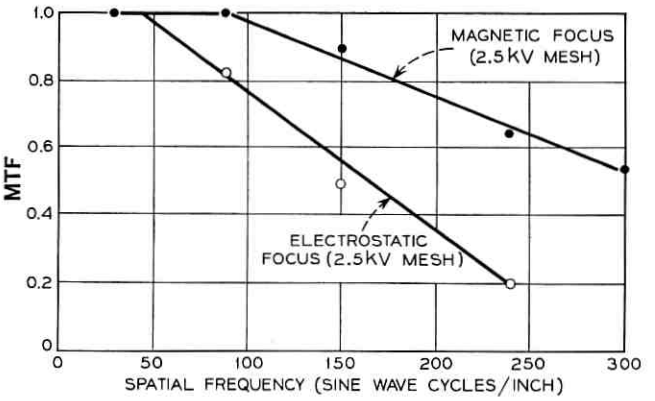


Fig. 6 — Measured modulation transfer function for the scan converter target.

mm. This is barely enough and a part of the falloff may be attributed to this fact.

Most of the falloff results from lateral diffusion of the holes which under the conditions of the measurement are produced within 1/10 micron of the substrate surface. Therefore, those that are collected must diffuse a distance of about 15-20 microns perpendicular to the surface. Since the lateral diffusion distance can be of the same order or greater, any detail requiring a spatial frequency of greater than 25 cycles per mm is effectively destroyed. The precise nature of the MTF falloff for the target depends on the target thickness, the diode geometry, and the volume recombination length and surface recombination velocity for holes. We plan a detailed discussion for a future paper.

The MTF at high spatial frequencies can be increased considerably by using thinner targets and increasing the diode density. Figure 2 illustrates a target with a substrate thickness of 10 microns and diode spacing of 15 microns corresponding to 67 diodes per mm. For this target the MTF is 100 per cent out to 280 cycles per inch and falls to 50 per cent at well over 400 cycles per inch. For this target the reading beam contributes substantially to the falloff in resolution.

The effective charge gain, of course, is a function of spatial frequency and its relative dependence on spatial frequency is the MTF shown in Fig. 6. However, the absolute charge gain for uniform storage patterns is a parameter of importance. Large values of gain are not really required or desirable for the scan converter; values of order 10 are useful. Values of order 10^3 or greater are desirable for some of the other applications.

The charge gain is measured in the following way. The unmodulated writing beam is scanned over some small area of the substrate at a current level sufficiently low that the resulting holes do not completely discharge the diodes. This can be checked by varying the writing beam current. The writing beam mesh is held negative relative to the substrate to suppress true low energy secondary electrons but not elastically reflected primaries. The redistributed secondary electrons, which land at much reduced energies compared with the primary beam, produce few, if any, hole-electron pairs. With the reading beam turned off, the current in the target lead measures the incident current in the writing beam that penetrates into the target. The charge gain in this case is effectively zero since the diodes are discharged and the hole-electron pairs are forced to recombine. The reading beam is turned on and scanned over an area which includes the area scanned by the writing beam. The increase in time average current measured in the target lead measures the arrival rate at the scanned diodes of all the holes generated in the substrate. The thermal part of the hole generation (the diode dark current) is determined by turning off the writing beam. The net current divided by the writing beam current penetrating the target is called the charge gain. A preliminary discussion of the expected results is appropriate at this point.

The charge gain should be describable by the expression

$$G(V) = \int_0^L \eta(x) \frac{dP(x, V)}{dx} dx \tag{1}$$

in which $\eta(x)$ is the probability that the hole, created at a distance x from the surface upon which the electrons are incident, will reach the diode space charge region and be collected, $dx(dP/dx)$ is the number of hole electron pairs created between x and $x + dx$ for an electron incident normally with kinetic energy V electron volts, and L is the substrate thickness. The function

$$P(V) = \int_0^\infty \frac{dP(x, V)}{dx} dx \tag{2}$$

which defines the total pair production per incident electron is given by

$$P(V) \approx V/3.5 \tag{3}$$

(corresponding to the fact that it takes on the average 3.5 eV to create one hole-electron pair). Writing

$$G(V) = P(V)\left[\int_0^L \eta(x)\frac{dP(x, V)}{dx}\,dx \Big/ \int_0^\infty \frac{dP(x, V)}{dx}\,dx\right] = P(V)\mathfrak{n}(V) \tag{4}$$

combines the effects of a variety of different phenomena into the function $\mathfrak{n}(V)$ which will be called the effective collection efficiency.

A number of different target fabrication procedures have been studied to optimize the effective collection efficiency. The best to date is similar to that used for optimizing the sensitivity of the target for visible light and consists of a thin n^+ layer formed on the writing beam side of the target. The n^+ layer is formed by a shallow phosphorous diffusion into the n-type conductivity substrate. The effect of this layer is discussed a few paragraphs further on. The measured collection efficiency $\mathfrak{n}(V) = 3.5G(V)/V$ as a function of electron energy is shown in Fig. 7. The collection efficiency approaches 0.5 for electron energies of order 10 KeV but falls well below 0.01 for energies below 2 KeV. Indeed in the energy range under 2 KeV the measured effective collection efficiency of a target for which the phosphorous diffusion was eliminated (the surface was merely etched) was higher at a constant value of 0.016.

An understanding of the measurements requires a knowledge of $\eta(x)$. However, a theoretical evaluation of $\mathfrak{n}(V)$ is complicated by the fact that $dP(x, V)/dx$ is not negligible very close to the surface and $\eta(x)$ near the surface is strongly dependent on the surface properties of the silicon crystal. Aside from the surface complication $\eta(x)$ may be accurately

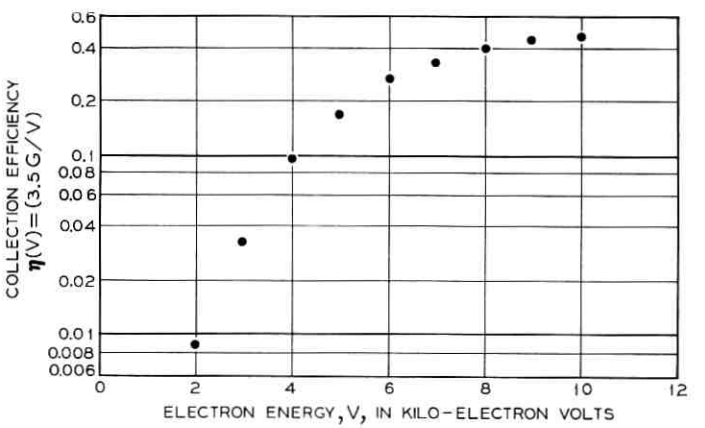


Fig. 7 — Measured effective collection efficiency for holes created by electrons of incident electron energy V. The effective collection efficiency $\mathfrak{n}(V)$ is determined from the measured charge gain $G(V)$ by the relation $\eta(V) = 3.5G(V)/V$.

evaluated on the basis of a model in which only the bulk recombination lifetime τ, the surface recombination velocity S at $x = 0$, the diffusion coefficient D, all for the minority carrier hole, and the substrate thickness L are relevant,

$$\eta(x) = \frac{\sinh(x/l) + (D/lS)\cosh(x/l)}{\sinh(L/l) + (D/lS)\cosh(L/l)}$$
$$l = (D\tau)^{\frac{1}{2}}, \tag{5}$$

which follows from a one-dimensional Green's function solution of the diffusion equation for excess holes. In fact (1) might be considered to be the normalized Green's function solution for hole current through the plane at $x = L$.

Quite generally $\eta(x)$ increases with increasing recombination length l and the target fabrication is optimized to make l as large as possible. Typically $l \gg L$ and it is appropriate to make this assumption. Thus (5) becomes

$$\eta(x) \approx \frac{1 + Sx/D}{1 + SL/D}. \tag{6}$$

For an etched silicon surface $S \approx 10^6$cm per second while $D = 10$ cm^2 per second. Thus for $L = 10$ microns, which represents a minimum practical value, $SL/D \approx 10^2$. It can be appreciated that $\eta(x)$ will be quite small for $x \ll L$ unless S is substantially reduced. Thus for low beam energies corresponding to small penetration depths $\mathfrak{n}(V) \approx \eta(0) = (1 + SL/D)^{-1} \approx 10^{-2}$ which is consistent with measurements on etched targets. As mentioned above, the most relevant technique among those that have been tried to reduce S for this application is a shallow phosphorous diffusion into the surface upon which the electrons are incident. This produces an n^+ layer which repels holes diffusing toward the surface resulting in an effective value of $S \approx 10^3$cm per second. As a result the x-dependence in (6) is relatively small. Unfortunately the phosphorous diffusion drastically increases the recombination rate of holes in the n^+ layer and the layer can be characterized as dead. As a result $\eta(x)$ is not well known for very small x.

The uncertainty in interpretation of the experimental results introduced by the dead layer makes it desirable to study also the collection efficiency for holes produced by incident photons. For this case the initial distribution of holes created by the photons is accurately known. A corresponding effective collection efficiency function $\mathfrak{n}(\lambda)$ can be defined for pair production by photons of wavelength λ,

$$\eta(\lambda) = \frac{\int_0^L \eta(x)\alpha e^{-\alpha x}\,dx}{\int_0^\infty \alpha e^{-\alpha x}\,dx} = \int_0^L \eta(x)\alpha e^{-\alpha x}\,dx \tag{7}$$

in which $\alpha(\lambda)$ is the absorption coefficient for photons of wavelength λ. This formulation is not strictly correct for $\alpha L < 1$ since light penetrating to the surface at $x = L$ may be reflected back into the substrate. Also the distance L is not well defined because of the diodes. If it is assumed that within the dead layer of thickness δ, $\eta(x) = 0$ and beyond the dead layer $\eta(x)$ is given by the equation in (6) with x measured from the edge of the dead layer, then for $\alpha L \gg 1$

$$\eta(\lambda) \approx \int_\delta^\infty \eta(x - \delta)\alpha e^{-\alpha x}\,dx$$

$$\eta(\lambda) \approx \left[\frac{1 + S/\alpha D}{1 + SL/D}\right] \exp - \alpha\delta. \tag{8}$$

The measured collection efficiency as a function of wavelength (for the same target used for obtaining the data of Fig. 7) is shown in Fig. 8. The data were obtained by admitting light onto the target through the glass wall of the tube envelope. The data are corrected for Fresnel reflection losses at the glass surfaces and from the silicon. Curves of (8) with $S = 1.1 \times 10^4$ cm per second, $D = 10$ cm^2 per second, $L = 20$

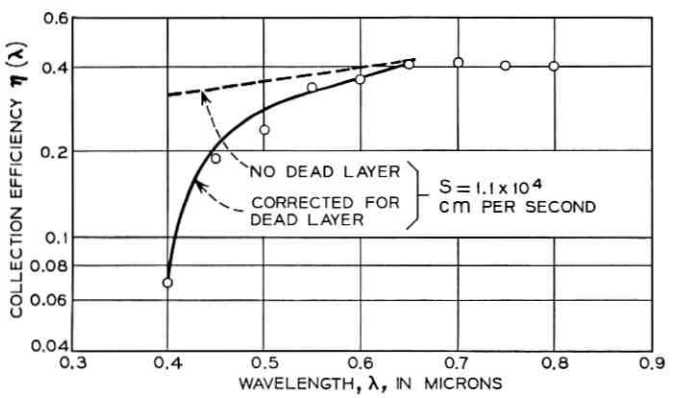


Fig. 8 — The measured collection efficiency for holes created by photons of wavelength λ. The solid and dashed curves were computed from equation (8).

microns and $\delta = 0$ and 0.21 microns are also shown. The calculation is not carried beyond $\lambda = 0.65$ microns since for longer wavelengths the target does not absorb all of the light transmitted past the surface. In fact, the slight falloff in collection efficiency above 0.7 microns may be attributed to this. Notice the good agreement with the curve corresponding to $\delta = 0.21$ microns indicating the approximate validity of the simple model of the dead layer. The dead layer thickness corresponds roughly to the phosphorous diffusion depth.

In the range $0.4 < \lambda < 0.65$ microns the absorption depth for photons[8] corresponds very roughly to the penetration depth for electrons in the energy range $4 < V < 10$ KeV. The penetration depth increases monotonically with λ or V to a maximum value of a few microns. Although the distribution of created holes for any value of λ is not the same as the distribution for any value of V, it is not surprising that the range of measured values of $\mathbf{n}(\lambda)$ and $\mathbf{n}(V)$ are quite similar. In either case a reduction of S to 10^3 cm per second, which is the more typical value observed in camera tubes, would increase the collection efficiency to close to unity over most of the range. A shallower phosphorous diffusion would also improve matters for low V or λ.

To date the only feature of the target which causes some concern about its future applicability is a slight burn-in or aging phenomena associated with the writing beam. It shows up as a decrease in charge gain over very heavily scanned areas. The rate of aging increases with writing beam current. It is not yet known whether the aging effect saturates, or whether it even occurs at all for low writing beam currents typical of most applications. The aging may account for the larger values of S observed in these targets as compared with camera tube targets.[9]

IV. CONCLUSION

A charge storage target for low energy scanning beam readout has been described with respect to its ability to produce a video representation of an electron image. Measurements of resolution and charge gain have been described. The target has general application in devices requiring an electron-image-to-video transducer and in scan conversion devices.

V. ACKNOWLEDGMENT

The authors are greatly indebted to E. J. Zimany, Jr., who has performed all of the experimental measurements and designed and con-

structed the test racks, as well as E. J. Walsh and R. P. Haynes who executed the mechanical design of the tube. The tubes were vacuum processed by E. J. Zimany, Sr. J. V. Dalton supplied the silicon diode array targets and N. C. Wittwer fabricated the resistive seas. E. F. Labuda participated in informative discussions.

REFERENCES

1. Crowell, M. H., Buck, T. M., Labuda, E. F., Dalton, J. V., and Walsh, E. J., "A Camera Tube with a Silico Diode Array Target," B.S.T.J., *46,* No. 2 (February 1967), pp. 491–495.
2. Gordon, E. I., "A 'Solid-State' Electron Tube for the *Picturephone*® Set," Bell Laboratories Record, *45,* No. 6 (June 1967), pp. 174–179.
3. Miyazaki, E., Maeda, H., and Miyaji, K., *Advances in Electron Physics,* Vol. 22A, New York: Academic Press, 1966, pp. 331–339. This work is relevant and provides a number of earlier references. In all previous measurements significant charge gain is observed only when the hole-electron pairs are created in the uniform depletion region of a large area diode.
4. Goetze, G. W., *Advances in Electronics and Electron Physics,* Vol. 22A, New York: Academic Press, 1966, pp. 219–223.
5. Crowell, M. H. and Gordon, E. I., 1967 International Electron Devices Meeting, "A Television Scan Converter Tube Using a Silicon Diode Array Target," Paper No. 10.2, Washington, D. C., October 1967.
6. E. D. Niper, unpublished work.
7. Danielson, W. E., unpublished work.
8. Runyan, W. R., *Silicon Semiconductor Technology,* New York: McGraw-Hill, 1965, pp. 187–201.
9. An increase in the surface recombination velocity of silicon surfaces covered with thin layers of oxide has been reported by E. H. Snow and D. J. Fitzgerald, "Radiation Study on MOS Structures," Fairchild Semiconduction, Scientific Report No. 4, AFCRL-68-0045.

Hologram Heterodyne Scanners

By L. H. ENLOE, W. C. JAKES, JR., and C. B. RUBINSTEIN

(Manuscript received April 1, 1968)

Several techniques are proposed in this paper which use a scanning coherent light beam to produce an electrical signal which corresponds to a scanned hologram. The hologram itself is not formed at the transmitting end of the system as a physical entity, rather the modulated electrical carrier frequency corresponding to the spatial carrier frequency of the hologram is generated by heterodyning. An advantage of the hologram heterodyne scanner is that it reduces the spatial resolution required of the camera tube in a holographic television system by at least a factor of four.

I. INTRODUCTION

In principle, it is possible to conceive of a holographic television system, but a practical system hinges on removing a number of formidable roadblocks. Among these is transducing the holographic information into a relatively narrowband electrical signal by means of a camera that has limited spatial resolution. This problem is shared by both a three-dimensional and a two-dimensional holographic system, but with present technology the 3-D system certainly presents many more problems. The two-dimensional system becomes much more tractable if the camera resolution problem is overcome. It is to this problem and our proposed solution that this paper addresses itself.

Consideration of this problem is not new, as evidenced by the early outline of the bandwidth requirements of a holographic television system by E. N. Leith and others in 1965.[1] There have also been reports of experiments involving the transmission via television of a Fresnel type of hologram in which the original object was a transparency.[2, 3] The difficulties encountered in these transmissions illustrate the crux of the problem.

A hologram consists of low spatial frequency information (dictated by the spatial information content of the object) which has been modulated onto a high spatial carrier frequency derived from the

interference of the reference and object beams. Conventionally, a television camera converts a spatial display, which in this case is a hologram, into an electrical signal by scanning it with an electron beam. However, the camera has a spatial frequency response that is low-pass in nature and limited in extent. Hence, the spatial carrier frequency rests out on the skirt of the passband, and the positive and negative sidebands are thus treated differently, producing distortion. We have so far lacked a feasible technique which would allow the effective use of the available bandwidth. That is to say, we need a technique which would allow the hologram information to be low-pass in nature until after being processed by the photosensitive surface of whatever transducer is used. The method we propose, which we call the heterodyne scanner, does exactly this. As a consequence, the required spatial resolution is reduced by at least a factor of four compared with the transmission of a conventional off-axis reference beam hologram.

The technique we propose envisions a scanning coherent light beam to produce an electrical signal which corresponds to a scanned hologram. In contrast with conventional methods, the hologram itself is not formed at the transmitting end of the system as a physical entity; instead the modulated electrical carrier frequency corresponding to the spatial carrier frequency of the hologram is generated by heterodyning.

These heterodyne scanners have important potential advantages over conventional methods of scanning. First, the nuisance terms corresponding to the direct beam in the reconstruction of a conventional hologram are not transduced. This halves the scanning beam aperture's resolution requirements. Second, the necessity for resolving the spatial carrier frequency of the equivalent hologram with the scanning beam is circumvented. This halves again the scanning beam resolution requirements, thus reducing resolution requirements to a quarter of the original requirements.

Let us briefly review the formation of a conventional hologram in order to point out the terms involved in the formation and reconstruction. We will then be in a better position to discuss the desired reduction and the means for accomplishing it.

II. CONVENTIONAL HOLOGRAM

Figure 1 depicts a typical holographic situation, showing an object beam

$$E_o(x, y) = A_o(x, y)\epsilon^{j[\omega_o t + \phi_o(x,y)]}$$

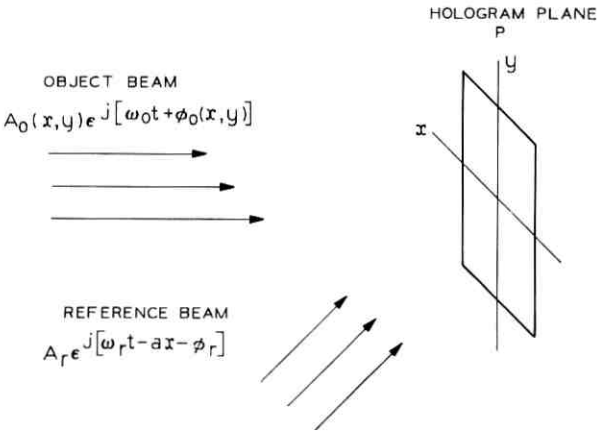


Fig. 1 — Formation of a conventional hologram. Object and reference beams impinge on hologram plane P.

and a reference beam

$$E_r(x, y) = A_r e^{j[\omega_r t - ax - \phi_r]}$$

impinging upon a hologram plane P. The intensity is given by

$$I = A_r^2 + A_o(x, y)^2 + 2A_r A_o(x, y) \cos [(\omega_o - \omega_r)t + \phi_o(x, y) + ax + \phi_r]. \quad (1)$$

For the formation of a conventional hologram the two beams must be at the same frequency, that is*, $\omega_r = \omega_o$, so that equation 1 becomes

$$I = A_r^2 + A_o(x, y)^2 + 2A_r A_o(x, y) \cos [\phi_o(x, y) + ax + \phi_r]. \quad (2)$$

During reconstruction, the first two terms of equation 2 form the direct beam and are unimportant except that they tend to obscure the reconstructed image which comes from the third term. Notice that the third term is a spatial carrier wave which is amplitude and phase modulated. If the spatial bandwidth of the wave to be reconstructed, $A_o(x, y)e^{j\phi_o(x,y)}$, is 2 W, then the bandwidth of the direct beam, $A_r^2 + A_o(x, y)^2$, and the bandwidth of the desired term, $A_r A_o(x, y) \cos [\phi_0(x, y) + ax + \phi_r]$, will each be $4W$, as shown in Figure 2. Thus if angular overlap between the direct beam and the desired reconstructed wavefront is to be avoided, we require that the spatial carrier frequency a of the desired third term in equation 2 be at least $3W$. This results in a total spatial bandwidth of $8W$.

* There are exceptions. See Ref. 4.

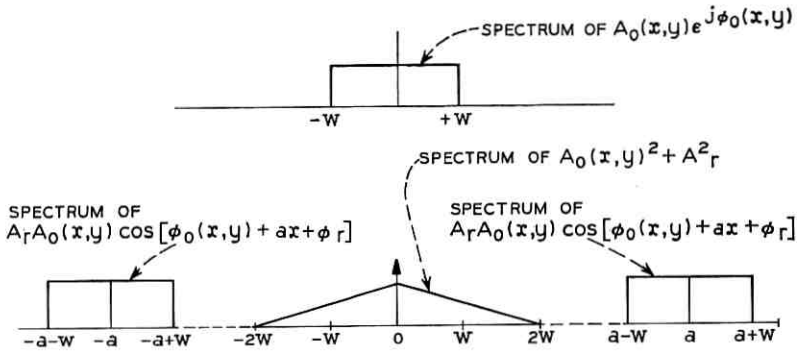


Fig. 2 — Spatial spectra of a conventional hologram signal. This representation is only intended to convey the limits of the spatial spectra and not the explicit form of the function.

III. ELIMINATING TERMS CORRESPONDING TO THE DIRECT BEAM

Since the third term in equation 2 is all that is necessary to reconstruct a hologram, it is desirable to eliminate the generation of the direct beam. The spatial carrier frequency a could then be reduced from its present minimum value of $3W$ to W, reducing the minimum total spatial bandwidth which must be transduced from $8W$ to $4W$. We are able to accomplish this rather easily with the heterodyne techniques discussed in this paper.

If equation 2 represents the intensity formed on a television camera tube, the camera output current will be proportional to the intensity being scanned by the electron beam. For an x-direction scan, the amplitude and phase modulated spatial carrier frequency is converted to a correspondingly modulated electrical signal. Notice that precisely the same electrical signal can be obtained by using a photomultiplier or other large area photodetector with a pinhole aperture placed just in front of the photosurface, the pinhole being scanned over the photosurface in a raster-like fashion in the same manner that the electron beam scans the photosurface in a camera tube. The output current from the photodetector would be proportional to the intensity of the light sampled by the pinhole.

A more desirable transducer is obtained if one shrinks the reference beam into a small pencil beam of pinhole dimensions which is then scanned over the photosurface. The pinhole aperture is now superfluous and can be removed, thereby gaining a distinct advantage. The intensity of the light on the photosurface contributed by the

object beam acting alone and the reference beam acting alone is now constant as a function of time. Thus, the nuisance terms corresponding to the direct beam in the reconstructed hologram, that is, the first two terms of equation 2, are not transduced as time varying currents. The only time varying term is produced by the interaction of the pencil reference beam with the object beam as it scans out the raster. This time varying signal is proportional to the desired third term in equation 2 which represents the useful information. Thus, the carrier frequency (dictated by the angle between the reference and object beams) can be reduced from $3W$ to W, which cuts in half the resolution requirements of the aperture of the scanning beam. (See Fig. 2).

IV. ELIMINATING THE SPATIAL CARRIER FREQUENCY

In the situation discussed in the previous section, spectrum overlap in the electrical signal is prevented by adjusting the angle between the object and reference beams to provide a sufficiently large spatial carrier frequency. As a consequence, the aperture of the scanning beam must resolve the highest spatial sideband frequency associated with this spatial carrier frequency.

As an alternative to generating the electrical carrier frequency by scanning a corresponding spatial carrier frequency, it is possible and indeed advantageous to generate the electrical carrier by heterodyning the object beam with the reference beam. This can be done in such a manner that it is unnecessary for the scanning aperture to resolve an equivalent modulated spatial carrier frequency. It must resolve only those spatial frequencies present in the complex field of the object beam itself, that is, it must resolve only a spatial bandwidth of $2W$. This realizes a subsequent reduction in the resolution requirements of the scanning beam by a factor of two over and above the reduction discussed in Section III.

For convenience, let us rewrite equation 1,

$$I = A_r^2 + A_0(x, y)^2 + 2A_rA_0(x, y) \cos [(\omega_0 - \omega_r)t + \phi_0(x, y) + ax + \phi_r], \quad (1)$$

which represents the intensity of the light incident on the photodetector when the object beam and reference beams are at different frequencies ω_0 and ω_r, respectively. Conceptually, these two frequencies can be sidebands produced by modulation, or can be two phase-locked modes of

the same laser, or produced by two phase-locked lasers, to name just a few methods.

In the present situation, the scanning pencil-like reference beam has an amplitude which may be written as a delta function, $A_r = \delta(x - ut)\delta(y - vt)$, where u and v are the horizontal and vertical scanning velocities, respectively. When this is substituted into equation 1 and integrated over the surface of the photodetector to find the output current, we obtain for the time varying component:

$$i(t) = 2A_0(ut, vt) \cos [(\omega_0 - \omega_r + au)t + \phi_0(ut, vt) + \phi_r].$$

We see that the frequency difference $\omega_0 - \omega_r$ has been added to the electrical carrier frequency au produced by scanning the spatial carrier frequency a. Thus, we may use a spatial carrier frequency of zero and avoid spectrum overlap (which produces beam overlap problems at reconstruction) by controlling the frequency difference between the reference and object beams. This means that the aperture of the scanning beam need resolve only those spatial frequencies present in the complex field of the object beam. For example, in Fig. 2 it need resolve only the spatial spectrum of $A_0(x, y)\epsilon^{j\phi_0(x,y)}$. Thus, the total spatial bandwidth required has been reduced by a factor of four.

V. ADDITIONAL CONSIDERATIONS

It might be appropriate to speculate on other aspects of the hologram heterodyne scanner. One might consider other means of implementing this technique. For example, one could replace the large area photodetector with a small area or point detector such as a photodiode. Rather than deflecting the pencil-like reference beam relative to a fixed object beam, the object beam is deflected relative to a fixed pencil reference beam. This allows one to reduce the size of the photodetector from that equal to the size of the equivalent hologram to that equal to a point. The price paid for this simplification is the difficulty associated with deflecting the information-bearing object beam without introducing excessive distortion. This will, of course, be more difficult than deflecting a pencil-like reference beam.

Another aspect of the use of the hologram heterodyne scanner is its potential sensitivity. Heterodyning is a well known technique for converting a weak light signal into an electrical current whose signal-to-noise ratio is determined by the fundamental quantum nature of light.[5] The heterodyne scanner should have potentially the best sensi-

tivity obtainable by any technique which does not use charge storage.

Without reported experimental verification of the hologram heterodyne scanner, it would not be especially fruitful to spend much time at this point speculating on possible display devices. There are a number of possible approaches to this problem but they are admittedly speculative. These include the Swiss Eidophor System[6] as a direct recording medium for the holographic electrical signal, or the use of photochromic materials for recording the output of a laser beam modulated by the holographic electrical signal and raster scanned over the photochromic surface. Both these methods have been referred to by E. N. Leith and his colleagues.

VI. COMMENTS AND COMPARISONS

Let us compare the hologram heterodyne scanner to the method proposed by C. B. Burckhardt and E. T. Doherty.[7] An extension of their technique* reduces the spatial resolution required of the camera tube by the same factor of four as the hologram heterodyne scanner, but even in its improved version it still requires a 50 percent greater electrical information rate than the scanner. However, the technique can be implemented with present technology.

A technique has been reported by L. H. Lin which involves spatial frequency sampling to reduce the information content of the hologram.[8] This is accomplished by an iterative Fourier transform technique. It is possible to augment Lin's technique with the hologram heterodyne scanner to increase the bandwidth saving. The translation of the hologram peculiar to Lin's technique can be carried out at the receiver. K. Haines and D. B. Brumm have also reported on a technique which can be used to reduce the information to be transduced and is compatible with the hologram heterodyne scanner.[9]

The hologram heterodyne scanner is a general technique that would apply to three-dimensional objects or two-dimensional transparencies. However, this does not mean that the implementation of this technique presents the same difficulties in each case. The spot size of the scanning reference beam and the speed of scan are directly influenced by the type of object to be transduced. The requirements are more stringent for the three-dimensional object if reasonable parallax is to be observed. Advances are currently being made in laser scanner technology which will help alleviate one aspect of the problem. A. B.

* The technique presented in Ref. 7 has been extended by C. B. Burckhardt to apply to TV transmission but this specific aspect is unpublished.

Larsen of Bell Telephone Laboratories is conducting experiments implementing the hologram heterodyne scanner technique. Many of the concepts of this paper have been improved and extended by him and will be reported in the near future.

VII. ACKNOWLEDGMENTS

We wish to thank R. C. Brainard, C. C. Cutler, and A. B. Larsen for helpful discussion and comments.

REFERENCES

1. Leith, E. N., Upatnieks, J., Hildebrand, B. P., and Haines, K., "Requirements for a Wavefront Reconstruction Television Facsimile System," J. SMPTE, *74*, No. 10 (October 1965), pp. 893–986.
2. Enloe, L. H., Murphy, J. A., and Rubinstein, C. B., "Hologram Transmission Via Television," B.S.T.J., *45*, No. 2 (February 1966), pp. 335–339.
3. Klimenko, I. S., and Rukman, G. I., "Regeneration of the Wavefront with the Aid of Holograms Transmitted Over a Television Channel," Zhurnal Tekhnicheskoi Fiziki, *37* (August 1967), pp. 1532–1534.
4. Denisyuk, Yu, N. and Staseliko, D. I., "The Possibility of Obtaining Holograms by Using Reference Beams the Wavelength of Which Differs from the Wavelength of the Radiation Scattered by the Object," Doklady Akademii Nauk SSSR, *176*, No. 6 (1967), pp. 1274–1275.
5. Oliver, B. M., "Signal-to-Noise Ratios in Photoelectric Mixing," Proc. IRE, *49*, No. 12 (December 1961), pp. 1960–1961.
6. Baumann, E., "The Fischer Large-Screen Projection System (Eidophor)," J. SMPTE, *60*, No. 4 (April 1953), pp. 344–356.
7. Burckhardt, C. B. and Doherty, E. T., "Formation of Carrier-Frequency Hologram with an On-Axis Reference Beam," Applied Optics, *7*, No. 6 (June 1968), pp. 1191–2.
8. Lin, L. H., "A Method of Hologram Information Reduction by Spatial Frequency Sampling," Applied Optics, *7*, No. 3 (March 1968), pp. 545–548.
9. Haines, K. and Brumm, D. B., "A Technique for Bandwidth Reduction in Holographic Systems," Proc. IEEE, *55*, No. 8 (August 1967), pp. 1512–1513; and "Holographic Data Reduction," Applied Optics, *7*, No. 6 (June 1968), pp. 1185–89.

Analysis of Thermal and Shot Noise in Pumped Resistive Diodes

By CORRADO DRAGONE

(Manuscript received April 8, 1968)

This paper discusses certain important aspects of the noise behavior of a pumped resistive diode containing shot and thermal noise sources. The derivation of the following result has a central role in the discussion. It is shown that the noise behavior of a pumped diode which does not contain 1/f *noise sources can be derived in a very simple way from Nyquist's theorem. This follows from the fact that the small-signal terminal behavior of such a diode can always be represented, in the frequency range of practical interest, by means of a connection of two linear and time-invariant networks of which one is noiseless and the other is dissipative, contains only thermal noise sources and is held at a uniform temperature.*

I. INTRODUCTION

The process of frequency conversion and its applications are well known and are extensively treated in the literature.[1–22] This paper considers the special case of a resistive diode frequency converter. An important limitation on the minimum noise figure of such a frequency converter is imposed by the noise generated by the diode, and it is the main purpose of this paper to study the properties of this noise.

Until a few years ago, much of the noise generated by the diode was $1/f$ noise. Therefore, since very little was known about this type of noise, the early theories of frequency converters using positive resistance diodes paid little attention to the noise performance, and somewhat later theories accounted for noise only in a very approximate way. However, as the semiconductor craft has developed, $1/f$ noise has been subject to considerable reduction and, even though its exact mechanism has not yet been completely established, in present diodes it appears to be important only at very low frequencies.[23] Therefore, the study of shot and thermal noise in pumped diodes is of great practical importance.

Strutt showed the method of treating shot and thermal noise in a pumped diode many years ago.[6] Since that time the method has been applied to tunnel diode frequency converters by a number of authors[16–20] However, in the case of a frequency converter using a positive resistance diode, it is normally believed that, in order to calculate its noise figure, a detailed analysis of the noise behavior of the diode is not necessary.[10, 12–14]

Consider a positive resistance diode which does not contain frequency dependent noise sources. That is, assume that, for any fixed voltage v applied to its terminals, the small-signal terminal behavior of the diode is equivalent to that of an ordinary resistor held at a uniform temperature T. The conductance g of this resistor is equal to the differential conductance of the diode and T is the so-called equivalent noise temperature of the diode. Since in a frequency converter the diode is pumped periodically by the pump, v varies with time. Therefore, g and T also vary with time, because they both depend upon v, and one can write $g = g(t)$ and $T = T(t)$.

Normally it is convenient to represent the small-signal terminal behavior of the diode by means of a linear and time-invariant network with several separate terminal pairs, one for each frequency of interest. A study of the noise behavior of this equivalent network generally requires that the self- and cross-power spectral densities of its short-circuit terminal currents, or of its open-circuit terminal voltages, be determined. Normally, however, the difficulty in determining the statistics of these noise terms is overcome by making the assumption that the equivalent network may be treated as an ordinary time-invariant dissipative system which contains only thermal noise sources and is held at a uniform temperature T_x. T_x is normally assumed to be equal to a certain time average of $T(t)$.

Even though no general proof has yet been given for this representation, it is widely used, mainly because it greatly simplifies the treatment of the noise performance of a frequency converter. However, it is often viewed with reservations for several reasons.[15] One very important reason is that it is generally applied to cases, in which one can easily show that it is not applicable, such as cases in which significant $1/f$ noise is generated by the diode. Another reason is that its validity is not obvious even in the limiting case where the noise power available from the diode is frequency independent and does not vary with the applied voltage. In fact, even in this limiting case, it is often considered to be not strictly valid.

However, in this paper it is shown that, besides being valid under certain limiting conditions, such a representation can also be used for formulating and interpreting in a very simple way the noise behavior of a pumped diode under quite general conditions, including a negative resistance diode.

In the following discussion it is assumed that the diode does not contain frequency-dependent noise sources, so that its small-signal terminal behavior may be completely specified by the two time-varying parameters $g(t)$ and $T(t)$. Then it is shown that, in the limiting case where T is a constant, the following theorem is true:

Theorem 1: If a pumped resistive diode is characterized by a time-invariant equivalent noise temperature T, *then its small-signal terminal behavior can be represented by means of a time-invariant equivalent network which contains only thermal noise sources and is held at a uniform temperature* $T_x = T$.

From this general theorem, which is already known to be valid under certain particular circuit conditions,[22] a number of interesting results can be derived. One important result is of course that, in a frequency converter which is bilateral and in which the noise temperature of the diode has negligible variations with time, the noise figure can be readily calculated. In fact, under these limiting conditions the noise figure can be related in a very simple way to T and to the dissipation characteristics of the circuit.[10, 12–14]

Another important result is that, also in the general case where $T(t)$ is not a constant, the terminal behavior of the diode can be readily derived from theorem 1. This is a consequence of the following general property, which follows directly from the definition of $T(t)$ and is stated as a theorem for emphasis:

Theorem 2: Consider a pumped diode characterized by the time-varying parameters $T(t)$ *and* $g(t)$. *Its short-circuit noise current* $\delta n(t)$ *is identical to that of a second diode characterized by a time-invariant temperature* T_2 *and a differential conductance* $|g(t)|\,T(t)/T_2$.

According to theorem 1, this second diode can be represented by an equivalent network held at a uniform temperature T_2. Therefore, by applying to this equivalent network the generalized form of Nyquist's theorem derived by Twiss,[24] the correlations between the various frequency components of $\delta n(t)$ can be readily determined. One finds that these correlations are simply equal to the Fourier coefficients of $g(t)$.

This property is already known to be valid under certain particular circuit conditions.[16–20]

Now, consider a linear, reciprocal, passive and time-invariant one-terminal pair network containing different elements held at different temperatures. It is well known[25] that at a given frequency ω_1 the effective noise temperature of this network can be expressed as a weighted average of the various temperatures of the lossy elements. The weighting factors in this weighted average are simply equal to the amounts of power that are dissipated by the various lossy elements when the network is connected, at its two terminals, to a generator delivering a unit amount of power at the considered frequency ω_1. This result is extended, in Section VIII, to a reciprocal and linear network containing a time-varying resistance, by introducing the concept of average temperature T_{av} of a pumped resistive diode. The significance of this parameter is best illustrated by the following example.

Suppose that one wants to calculate the noise power available from the output terminals of a frequency down-converter. It is shown that, if the frequency converter is bilateral, this power can be calculated by replacing the diode with one having the same i-v characteristic and a temperature equal to T_{av}, where T_{av} is given by the relation

$$T_{av} = \frac{\langle P(t)T(t)\rangle_{av}}{\langle P(t)\rangle_{av}}, \tag{1}$$

where $\langle\ \rangle_{av}$ indicates the time average and $P(t)$ is the instantaneous small-signal power dissipated by the differential conductance of the diode when a small-signal generator is applied to the output terminals of the frequency converter. It is important to point out that T_{av} depends, in general, both on the characteristics of the diode and on those of the circuit connected to it.

II. SMALL SIGNAL EQUATIONS OF A NOISELESS PUMPED DIODE

Let the diode current i be a nonlinear function $f(v)$ of the terminal voltage v. It is assumed that the diode is pumped by a strong periodic source at a frequency ω_o and its harmonics. Therefore v and i contain large components $v_c(t)$ and $i_c(t)$ of the type:

$$v_c(t) = \sum_{k=-\infty}^{\infty} V_k \exp jk\omega_o t \tag{2}$$

$$i_c(t) = \sum_{k=-\infty}^{\infty} I_k \exp jk\omega_o t. \tag{3}$$

It is assumed that v and i contain, in addition, small components $\delta v(t)$ and $\delta i(t)$ and it is desired to derive the relation between $\delta v(t)$ and $\delta i(t)$, for the limiting case $\delta v(t) \to 0$ and $\delta i(t) \to 0$. Thus let

$$v = v(t) = v_c(t) + \delta v(t), \tag{4}$$

$$i = i(t) = i_c(t) + \delta i(t). \tag{5}$$

The differential conductance of the diode is equal to the derivative of $f(v)$. Let it be denoted by $g_d(v)$ and let

$$g(t) = g_d[v_c(t)]. \tag{6}$$

Since $v_c(t)$ is periodic, also $g(t)$ is periodic and therefore it can be written in the form

$$g(t) = \sum_{k=-\infty}^{\infty} g_k \exp jk\omega_o t. \tag{7}$$

Since $i = f(v)$ and $g_d(v)$ is the derivative of $f(v)$, from equations 4, 5, and 6 one has:

$$\delta i(t) = g(t)\ \delta v(t) \tag{8}$$

in the limiting case $\delta v(t) \to 0$. This relation completely describes the small-signal terminal behavior of the diode, in the absence of internal noise sources.

III. SMALL SIGNAL EQUATIONS OF A NOISELESS DIODE IN THE FREQUENCY DOMAIN

From equations 7 and 8 the relations between the different frequency components of $\delta v(t)$ and $\delta i(t)$ can be readily derived[4,7]. In fact, assume that both $\delta v(t)$ and $\delta i(t)$ contain components at only the pairs of side-frequencies $k\omega_o + p$ and $k\omega_o - p$ ($| k | = 0, 1, 2$, etc.; $2p < \omega_o$). Then $\delta v(t)$ and $\delta i(t)$ can be expressed as follows:

$$\delta v(t) = 2(\mathrm{Re})\left[\sum_{k=0}^{\infty} V_{\alpha k} \exp j(p + k\omega_o)t + \sum_{k=1}^{\infty} V_{\beta k} \exp j(p - k\omega_o)t\right] \tag{9}$$

$$\delta i(t) = 2(\mathrm{Re})\left[\sum_{k=0}^{\infty} I_{\alpha k} \exp j(p + k\omega_o)t + \sum_{k=1}^{\infty} I_{\beta k} \exp j(p - k\omega_o)t\right] \tag{10}$$

and, on substituting equations 9, 10, and 7 into equation 8, one obtains the following relations between the Fourier coefficients of the various frequency components of $\delta v(t)$ and $\delta i(t)$:

$$I_{\alpha r} = \sum_{k=0}^{\infty} g_{r-k} V_{\alpha k} + \sum_{k=1}^{\infty} g_{r+k} V_{\beta k} \qquad (r = 0, 1, \text{ etc.}) \tag{11}$$

$$I_{\beta r} = \sum_{k=1}^{\infty} g_{k-r} V_{\beta k} + \sum_{k=0}^{\infty} g_{-r-k} V_{\alpha k} \qquad (r = 1, 2, \text{ etc.}) \tag{12}$$

which can be written in the form:

$$\begin{bmatrix} I_\alpha] \\ I_\beta] \end{bmatrix} = \begin{bmatrix} [G_{\alpha\alpha}] & [G_{\alpha\beta}] \\ [G_{\beta\alpha}] & [G_{\beta\beta}] \end{bmatrix} \begin{bmatrix} V_\alpha] \\ V_\beta] \end{bmatrix} \tag{13}$$

where the matrix notation is defined as follows:

$$I_\alpha] = \begin{bmatrix} I_{\alpha 0} \\ I_{\alpha 1} \\ I_{\alpha 2} \\ \vdots \\ \vdots \end{bmatrix}, \quad I_\beta] = \begin{bmatrix} I_{\beta 1} \\ I_{\beta 2} \\ I_{\beta 3} \\ \vdots \\ \vdots \end{bmatrix}, \quad V_\alpha] = \begin{bmatrix} V_{\alpha 0} \\ V_{\alpha 1} \\ V_{\alpha 2} \\ \vdots \\ \vdots \end{bmatrix}, \quad V_\beta] = \begin{bmatrix} V_{\beta 1} \\ V_{\beta 2} \\ V_{\beta 3} \\ \vdots \\ \vdots \end{bmatrix} \tag{14}$$

and the elements of the matrices $[G\alpha\alpha]$, $[G\alpha\beta]$, etc., are

$$(G_{\alpha\alpha})_{r,k} = g_{r-k} \qquad (r, k = 0, 1, \text{ etc.}) \tag{15}$$

$$(G_{\beta\beta})_{r,k} = g_{k-r} \qquad (r, k = 1, 2, \text{ etc.}) \tag{16}$$

$$(G_{\alpha\beta})_{r,k} = g_{r+k} \qquad (r, k - 1 = 0, 1, \text{ etc.}) \tag{17}$$

$$(G_{\beta\alpha})_{r,k} = g_{-r-k} \qquad (r - 1, k = 0, 1, \text{ etc.}). \tag{18}$$

Equations 13 through 18 completely specify the terminal behavior of the diode at the frequencies $p \pm k\omega_0$ in the absence of internal noise sources.

IV. SMALL SIGNAL TERMINAL BEHAVIOR OF A NOISY DIODE

Up to this point the noise generated by the diode has been ignored. In the general case of a noisy diode equation 8 has to be modified as follows:

$$\delta i(t) = g(t)\ \delta v(t) + \delta n(t) \tag{19}$$

where $\delta n(t)$ is the equivalent short-circuit noise current of the diode. Equation 19 corresponds to the equivalent circuit shown in Fig. 1 in which the spontaneous fluctuations of the diode are ascribed to a current generator of infinite internal impedence, acting in parallel to the differential conductance of the diode.

Now, consider the components of $\delta n(t)$ occurring in an infinitesimal frequency range between $\omega - (d\omega)/2$ and $\omega + (d\omega)/2$. It is convenient to account for these components by means of a single pseudosinusoid

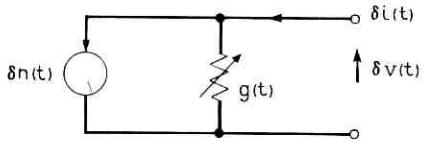


Fig. 1 — Equivalent circuit of a time-varying conductance containing noise sources.

with random complex amplitude[26-28]. Then, let $N_{\alpha k}$ and $N_{\beta r}$ ($k = 0, 1, 2$, etc.; $r = 1, 2$, etc.) be the complex Fourier amplitudes of the pseudosinusoids relative to the frequencies $p + k\omega_o$ and $p - k\omega_o$, respectively. Noise components occurring at frequencies different from these will be neglected since they have no effect on the small-signal terminal behavior of the diode at the frequencies $p \pm k\omega_o$. Then

$$\delta n(t) = 2(\mathrm{Re})\left\{\sum_{k=0}^{\infty} N_{\alpha k} \exp j(p + k\omega_o)t + \sum_{k=1}^{\infty} N_{\beta k} \exp j(p - k\omega_o)t\right\} \quad (20)$$

and from equations 13 and 19 one obtains:

$$\begin{bmatrix} I_\alpha \\ I_\beta \end{bmatrix} = \begin{bmatrix} [G_{\alpha\alpha}] & [G_{\alpha\beta}] \\ [G_{\beta\alpha}] & [G_{\beta\beta}] \end{bmatrix} \begin{bmatrix} V_\alpha \\ V_\beta \end{bmatrix} + \begin{bmatrix} N_\alpha \\ N_\beta \end{bmatrix} \quad (21)$$

where

$$N_\alpha] = \begin{bmatrix} N_{\alpha 0} \\ N_{\alpha 1} \\ \vdots \end{bmatrix} \quad \text{and} \quad N_\beta] = \begin{bmatrix} N_{\beta 1} \\ N_{\beta 2} \\ \vdots \end{bmatrix}. \quad (22)$$

Equation 21 completely specifies the small-signal terminal behavior of a pumped diode containing noise sources. Its physical interpretation is often facilitated by introducing the equivalent circuit of Fig. 2. In this equivalent circuit the diode is represented by a linear and time-invariant network in which the terminal voltages and currents occur at the same frequency. Their Fourier coefficients are equal to those of the various frequency components of $\delta v(t)$ and $\delta i(t)$.

The network of Fig. 2 is completely specified with respect to its terminal pairs by its admittance matrix

$$[G] = \begin{bmatrix} [G_{\alpha\alpha}] & [G_{\alpha\beta}] \\ [G_{\beta\alpha}] & [G_{\beta\beta}] \end{bmatrix} \quad (23)$$

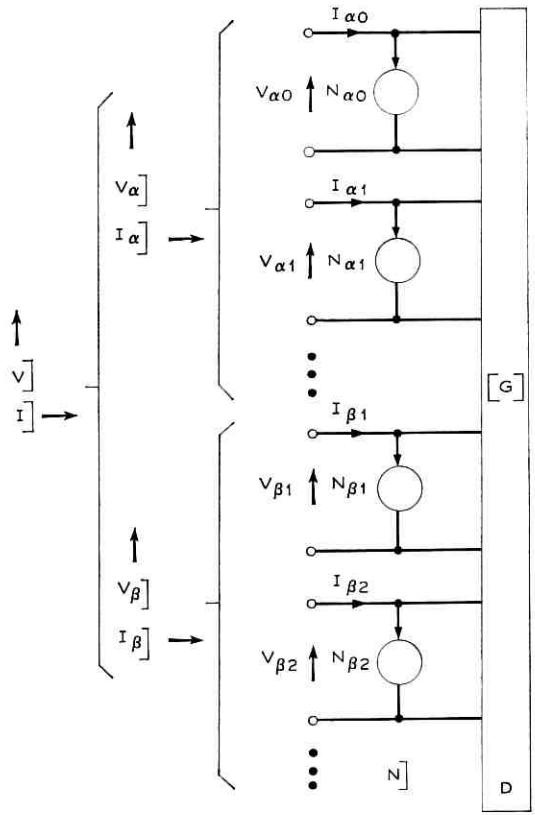


Fig. 2 — Time-invariant equivalent network of a pumped diode.

and by the noise column matrix

$$N] = \begin{bmatrix} N_\alpha] \\ N_\beta] \end{bmatrix} \tag{24}$$

which represents the complex Fourier amplitudes of its short-circuit terminal currents. The self- and cross-power spectral densities of these noise currents are

$$\frac{\langle N_{\alpha k} N^*_{\alpha r} \rangle}{df}, \quad \frac{\langle N_{\beta k} N^*_{\beta r} \rangle}{df}, \quad \text{and} \quad \frac{\langle N_{\alpha k} N^*_{\beta r} \rangle}{df} \tag{25}$$

where $\langle\ \rangle$ indicates the statistical average. They are conveniently represented by the matrix

$$\frac{1}{df}\langle N]N]^{\dagger}\rangle, \tag{26}$$

where superscript† denotes the Hermitian conjugate.

Now let the properties of the equivalent network of Fig. 2 be briefly examined. Since $g_k = g^*_{-k}$ ($k = 0, 1, 2$, etc.), from equations 15 through 18 one has that the admittance matrix $[G]$ is Hermitian, that is:

$$[G] = [G]^{\dagger}. \tag{27}$$

It is interesting to note that this condition is equivalent to the condition that

$$(IM)(V]^{\dagger}[G]V]) = 0 \tag{28}$$

for all $V]$, which requires that the total reactive power flowing into the nonlinear resistance at the various side frequencies $p \pm k\omega_0$ ($k = 0, 1$, etc.) be always zero. This property is a direct consequence of the general energy relations derived by Manley and Rowe for nonlinear resistors.[20] Because of equation 27 the average small-signal power dissipated in the admittance $g(t)$ can be expressed as

$$\langle \delta v(t)^2 g(t)\rangle_{av} = V]^{\dagger}([G] + [G]^{\dagger})V] = 2V]^{\dagger}[G]V]. \tag{29}$$

Therefore, if

$$g(t) > 0 \tag{30}$$

at all times, then $[G]$ is both Hermitian and positive definite and the equivalent network is dissipative.

Now, consider a linear and dissipative network which contains only thermal noise sources and is characterized by an admittance matrix equal to $[G]$. If such a network is held at a uniform temperature T, then the various spectral densities of its short-circuit terminal currents are simply given by the elements of the matrix

$$kT([G] + [G]^{\dagger}).$$

From this generalized form of Nyquist's Theorem, proved by Twiss,[24] and from equation 27 one has that, if condition 30 is satisfied and the matrix 26 satisfies the relation

$$\langle N]N]^{\dagger}\rangle = 2kTdf[G], \tag{31}$$

then the small-signal terminal behavior of the diode can be represented by means of an equivalent network which contains only thermal noise sources and is held at a uniform temperature T.

Of special importance is the particular case in which the circuit connected to the diode is resistive at the harmonics $2\omega_o$, $3\omega_o$, etc., of the pump frequency and, at these frequencies, does not contain generators. Under these conditions it is always possible to choose the origin of time in such a way as to make $v_c(t)$, $i_c(t)$ and $g(t)$ even functions of time.[7] In this case, since all of the coefficients g_k ($k = 0, 1$, etc.) become real,

$$g_k = g_{-k}, \tag{32}$$

and therefore $[G]$ becomes a real symmetric matrix, because of equations 15 through 18. If, in addition to equation 32, condition 30 is satisfied, then the equivalent network of Fig. 2 can be realized by means of an ordinary resistive network.[2]

Of course, in the general case where the origin of time cannot be chosen to make all the coefficients g_k real, the diode cannot be represented by a reciprocal (bilateral) network.

Condition 31 is never satisfied if $g(t)$ becomes negative for some values of t. In fact in this case $[G]$ is indefinite, while $\langle N]N]^\dagger\rangle$ is always a positive definite or semidefinite matrix. On the other hand, if

$$g(t) < 0 \tag{33}$$

for all values of t, then $[G]$ is negative definite and of special interest becomes the condition

$$\langle N]N]^\dagger\rangle = -2kT\, df[G]. \tag{34}$$

In fact, consider a pumped negative resistance diode which satisfies this condition. If a frequency converter is made from such a diode by imbedding it in a lossless network, then its noise measure, defined by Hauss and Adler,[30] is independent of the characteristics of the lossless network and is simply equal to T/T_o, where T_o is standard temperature, 290°K. Therefore, if G_e is the exchangeable gain of such a frequency converter, its noise figure F is simply equal to

$$F = 1 + T/T_o\,(1 - 1/G_e). \tag{35}$$

V. SHOT NOISE IN A PUMPED DIODE

Assume that the diode only contains shot noise sources and that in the frequency range of interest transit time effects can be neglected.[6, 11, 26, 31, 32]

Assume for the moment that the voltage v applied to the diode is time-invariant. Then $\delta n(t)$ can be treated as white noise over the

frequency range of practical interest. Therefore, if S denotes its spectral density, it can be expressed as

$$\delta n(t) = S^{\frac{1}{2}} x(t), \tag{36}$$

where $x(t)$ is white noise with unit spectral density. S will in general depend upon the voltage v applied to the diode and it is convenient to express this voltage dependence in the following form:

$$S = S_d(v) = 2kT_d(v) \mid g_d(v) \mid \tag{37}$$

where $kT_d(v)/2$ represents the exchangeable noise power at the diode terminals, per unit bandwidth. In equation 37 the occurrence of the factor 2, in place of the usual factor 4, results from the fact that here both positive and negative frequencies are considered.

Now, consider the general case where v is not a constant and let the concise notation

$$\begin{aligned} T(t) &= T_d[v_c(t)] \\ S(t) &= S_d[v_c(t)] = 2kT(t) \mid g(t) \mid \end{aligned} \tag{38}$$

be introduced. Then $\delta n(t)$ results from the superposition of statistically independent random disturbances whose probability of occurrence is proportional to the deterministic and periodic function

$$h(t) = [S(t)].^{\frac{1}{2}} \tag{39}$$

Since it is assumed that the duration of these disturbances is much smaller than the reciprocal of the highest significant frequency of $h(t)$, equation 36 is still applicable and therefore

$$\delta n(t) = h(t)x(t). \tag{40}$$

Now let consideration be restricted to the fluctuation components occurring in infinitesimal frequency intervals of width df, centered at the frequencies $p \pm k\omega_o$. Then, since $x(t)$ is white noise with unit spectral density, from equation 40 one has that $\delta n(t)$ can be expressed as follows:[6,26]

$$\delta n(t) = h(t) \sum_{s=-\infty}^{\infty} 2(df)^{\frac{1}{2}} \cos [(s\omega_o + pt) + \varphi_s] \tag{41}$$

where φ_s are statistically independent random phase angles distributed uniformly over the range $(0, 2\pi)$.

Let $h(t)$ and $S(t)$ be represented by the Fourier series

$$S(t) = \sum_{k=-\infty}^{\infty} S_k \exp jk\omega_o t \tag{42}$$

$$h(t) = \sum_{k=-\infty}^{\infty} H_k \exp jk\omega_o t. \tag{43}$$

From equation 39 one has that S_k and H_k are related through the relations:

$$S_r = \sum_{k=-\infty}^{\infty} H_k H_{r-k} \, . \tag{44}$$

Introduction of equation 43 in equation 41 gives:

$$\begin{aligned} \delta n(t) &= (df)^{\frac{1}{2}} \sum_{r=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} H_r \{\exp j[(s+r)\omega_o t + pt + \varphi_s] \\ &\qquad + \exp j[(r-s)\omega_o t - pt - \varphi_s]\} \\ &= (df)^{\frac{1}{2}} \sum_{s=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} H_{k-s} \exp j\varphi_s \exp j(k\omega_o + p)t \\ &\qquad + H_{s-k} \exp -j\varphi_s \exp -j(k\omega_o + p)t \\ &= 2(\mathrm{Re})\left\{(df)^{\frac{1}{2}} \sum_{k=-\infty}^{\infty} \sum_{s=-\infty}^{\infty} H_{k-s} \exp j\varphi_s \exp j(k\omega_o + p)t\right\}. \end{aligned} \tag{45}$$

From this last relation and from equation 20 one obtains

$$N_{\alpha k} = (df)^{\frac{1}{2}} \sum_{s=-\infty}^{\infty} H_{k-s} \exp j\varphi_s \tag{46}$$

$$N_{\beta k} = (df)^{\frac{1}{2}} \sum_{s=-\infty}^{\infty} H_{-k-s} \exp j\varphi_s \, . \tag{47}$$

Hence, since

$$\langle \exp j\varphi_s \exp -j\varphi_r \rangle = \begin{cases} 1, & r = s \\ 0, & r \neq s \end{cases},$$

from equations 46 and 47 one obtains:

$$\begin{aligned} \frac{\langle N_{\alpha k} N^*_{\alpha r} \rangle}{df} &= \sum_{s=-\infty}^{\infty} H_{k-s} H_{s-r} \\ &= S_{k-r} \end{aligned} \tag{48}$$

and

$$\frac{\langle N_{\alpha k} N^*_{\beta r} \rangle}{df} = \sum_{s=-\infty}^{\infty} H_{k-s} H_{r+s} = S_{k+r} . \tag{49}$$

Now, consider a time-varying conductance equal to $S(t)$ and let $[S]$ denote its admittance matrix. Then $[S]$ is obtained from equations 15 through 18, and 23 by formally replacing g_k and G with S_k and S throughout. Therefore the elements of $[S]$ are equal to the various Fourier coefficients S_k and, from equations 48 and 49, one obtains the final result

$$\langle N] N]^\dagger \rangle = df[S]. \tag{50}$$

VI. NOISE BEHAVIOR OF THE DIODE

The preceding section showed that if a diode contains only shot noise sources, then the self- and cross-power spectral densities of its short-circuit terminal currents are simply equal to the Fourier coefficients of $2kT(t)\ |g(t)|$, over the frequency range of practical interest. Let us examine the significance of these relations, which are valid even if the diode contains thermal noise sources.

First, consider the special case of a positive resistance diode characterized by an equivalent noise temperature $T_d(v)$ which is approximately constant over the range of voltages of interest, so that the approximation

$$T_d(v) = T = \text{constant} \tag{51}$$

can be made. In this case since equations 38 give

$$S_k = 2kTg_k , \tag{52}$$

one has

$$[S] = 2kT[G] \tag{53}$$

and therefore from equation 50 it follows that the spectral density matrix satisfies condition 31. One concludes that, if $T_d(v)$ is independent of v and condition 30 is satisfied, then the small-signal terminal behavior of the diode can be represented by a time-invariant dissipative network held at a uniform temperature T, as stated in theorem 1.

Thus, in the limiting case (51) and under the restriction $g(t) > 0$,

equation 50 can be interpreted as a direct consequence of theorem 1 and Nyquist's theorem. Now, a little reflection shows why equation 50 also is valid in the general case where $T(t)$ is not a constant and the restriction $g(t) > 0$ is removed. In fact, two diodes having the same $S(t)$ have the same short-circuit terminal currents, no matter what their differential admittances may be. Notice that from this rather obvious property theorem 2 follows at once. That is, the short-circuit terminal currents of a diode characterized by an equivalent noise temperature $T(t)$ and by a differential conductance $g(t)$ are identical to those of a diode characterized by a constant temperature T_2 and a differential conductance

$$| g(t) | [T(t)]/T_2, \tag{54}$$

where T_2 is an arbitrary temperature. It is important to point out that, even though the foregoing two diodes have the same short-circuit terminal currents, they are not equivalent since they have different conductances. On the other hand, the terminal behavior of a diode characterized by a voltage-dependent temperature $T_d(v)$ and a conductance $g_d(v)$ is equal to that of the parallel connection of the two diodes (see Fig. 3) with voltage-independent temperatures T_1 and T_2 and with the differential conductances $g_{d1}(v)$ and $g_{d2}(v)$ defined by the following equations:

$$g_{d1}(v) + g_{d2}(v) = g_d(v) \tag{55}$$

$$g_{d1}(v)T_1 + g_{d2}(v)T_2 = g_d(v)T_d(v) \tag{56}$$

where T_1 and T_2 are subject to the only condition

$$T_1 < T_d(v) < T_2 \tag{57}$$

which guarantees that $g_{d1}(v)$, $g_{d2}(v)$ and $g_d(v)$ have all the same sign. Notice that the equivalent circuit of Fig. 3 and the original diode have the same short-circuit terminal currents because of equa-

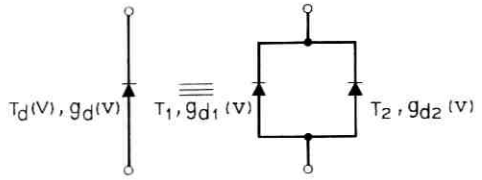


Fig. 3 — Representation of an arbitrary noisy resistive diode by means of two diodes with voltage-independent noise temperatures T_1 and T_2.

tion 56, and have the same differential conductances because of equation 55. An important feature of this equivalent circuit is that theorem 1 is applicable to both diodes and it can therefore be studied by standard techniques. Particularly interesting is the limiting case $T_1 = 0$. In fact in this case one of the two diodes becomes noiseless and the other has the time-varying conductance

$$g_2(t) = g_{d2}[v_c(t)] = g(t)[T(t)]/T_2 , \tag{58}$$

when the pump voltage $v_c(t)$ is applied to it. Hence, by comparing equation 54 with equation 58 one obtains the following result. If $g(t) > 0$ and $T_2 > T(t)$ for all values of t, then, by connecting a noiseless diode having the conductance

$$g_1(t) = g_{d1}[v_c(t)] = g(t)[T_2 - T(t)]/T_2 \tag{59}$$

in parallel with the second diode of theorem 2, one obtains a circuit completely equivalent to the original diode.

Now, consider the case where $g(t) < 0$ for all values of t and suppose that condition 51 is satisfied. Then

$$[S] = -2kT[G] \tag{60}$$

and from equation 50 one has that condition 34 is satisfied. Hence, the remarks about this possibility at the end of Section III apply. In general, where $T(t)$ is not a constant, equation 35 is not valid. However, if T_1 and T_2 are the minimum and maximum values of $T(t)$, so that

$$T_1 \leqq T(t) \leqq T_2 \tag{61}$$

then one can say that the noise performance of the diode will be bounded by the two limiting values obtained from equation 35 for the two limiting cases $T = T_1$ and $T = T_2$.

VII. TERMINAL BEHAVIOR OF THE DIODE IN THE IMPEDANCE-MATRIX REPRESENTATION

In some cases it is convenient to use the impedance-matrix representation, rather than the admittance-matrix representation, for describing the terminal behavior of the diode. Let

$$r(t) = 1/g(t) = \sum_{k=-\infty}^{\infty} r_k \exp jk\omega_o t \tag{62}$$

be the differential resistance of the diode. Then the impedance-matrix

representation of the small-signal terminal behavior of the diode can be written in the form

$$V] = [R]I] + N_v] \tag{63}$$

where the relations between the elements of the impedance-matrix $[R]$ and the coefficients r_k are identical to those between the elements of $[G]$ and g_k (see equations 15 through 18). The column matrix $N_v]$ consists of the amplitudes of the open-circuit terminal voltages of the diode. If

$$0(t) = 2kT(t) \mid r(t) \mid = \sum_{k=-\infty}^{\infty} 0_k \exp jk\omega_o t \tag{64}$$

then one has

$$\langle N_v]N_v]^\dagger \rangle = df[0] \tag{65}$$

where $[0]$ can be obtained from equations 15 through 18, and 23 by replacing G and g_k with 0 and 0_k throughout. Notice that equation 65, which is analogous to equation 50, follows from theorem 1, Nyquist's theorem[24] and the fact that two diodes having the same $0(t)$ have the same $N_v]$, no matter what their differential resistances may be.

VIII. AVERAGE TEMPERATURE OF A PUMPED DIODE

Consider a linear, reciprocal, passive and time-invariant one-terminal pair network containing different elements held at different temperatures. It is well known[25] that at a given frequency ω_1 the effective noise temperature of this network can be expressed as a weighted average of the various temperatures of the lossy elements. The weighting factors in this weighted average are simply equal to the amounts of power that are dissipated by the various lossy elements when the network is connected, at its two terminals, to a generator delivering a unit amount of power at the considered frequency ω_1. This result is extended, in this section, to a pumped diode.

The concept of average noise temperature T_{av} of a pumped diode is introduced in this section. Consideration is restricted to the case where $g(t) \geqq 0$ and condition 32 is satisfied, so that the equivalent circuit of the diode is passive and bilateral. It is shown that T_{av} depends, in general, both on the characteristics of the diode and on those of the linear and time-invariant circuit connected to it. However, if certain conditions are satisfied, then it only depends on the diode characteristics.

Consider a one-terminal-pair network N consisting of a pumped diode imbedded in a linear, time-invariant and bilateral two-terminal-pairs network N' (Fig. 4). Assume, furthermore, that the network N can exchange power at a single frequency ω_1, at its terminals. Then the average temperature T_{av} of the diode is defined in the following way:

T_{av} is such that the noise power available from the terminals of the network N does not change if the actual temperature-voltage characteristic $T_d(v)$ of the diode is replaced with a constant temperature equal to T_{av}.

Now let a small-signal generator of frequency ω_1 be connected to the terminals of the network N, and let $\delta i(t)$ and $\delta v(t)$ be the small signals produced at the diode terminals. It will be shown that

$$T_{av} = \frac{\int_0^{2\pi/\omega_1} \delta i(t)\ \delta v(t) T(t)\ dt}{\int_0^{2\pi/\omega_1} \delta i(t)\ \delta v(t)\ dt}. \tag{66}$$

Notice that this equation is equivalent to equation 1.

Proof: It is convenient to represent the circuit of Fig. 4 by means of the equivalent circuit of Fig. 5, where the network D represents the small-signal terminal behavior of the diode and each terminal pair of D exchanges power at only one frequency. Notice that in Fig. 5 the network N' has been represented by means of several separate equivalent circuits, N_1', N_2', etc., one for each frequency of interest.

The network D can be decomposed into two separate networks D_1 and D_2, each held at a uniform temperature.

In fact, let the diode of Fig. 4 be replaced by the two diodes shown in Fig. 3 and let $\delta i_1(t)$ and $\delta i_2(t)$ be the small-signal currents of the

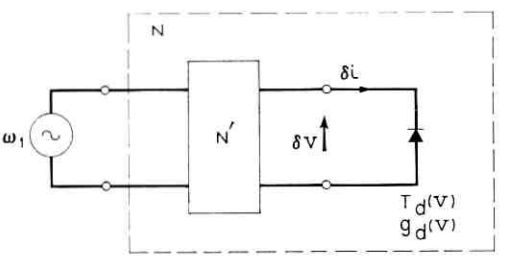


Fig. 4 — Diode imbedded in a linear, time-invariant and bilateral network N'.

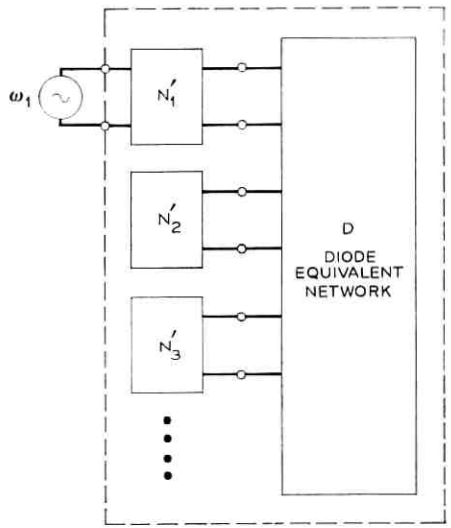


Fig. 5 — Equivalent circuit of the network N of Fig. 4.

two diodes. Then, from equation 56 one has:

$$\delta v(t)\ \delta i(t)\, T(t) = \delta v(t)\ \delta i_1(t)\, T_1 + \delta v(t)\ \delta i_2(t)\, T_2$$

which gives:

$$\frac{\langle \delta v(t)\ \delta i(t) T(t)\rangle_{av}}{\langle \delta v(t)\ \delta i(t)\rangle_{av}} = \frac{\langle \delta v(t)\ \delta i_1(t)\rangle_{av} T_1 + \langle \delta v(t)\ \delta i_2(t)\rangle_{av} T_2}{\langle \delta v(t)\ \delta i(t)\rangle_{av}}. \tag{67}$$

Since theorem 1 is applicable to both diodes of Fig. 3, it is clear that the network D of Fig. 5 can be represented by the parallel connection of two networks (D_1 and D_2) of which one is held at a uniform temperature T_1 and dissipates an average power equal to $\langle \delta v(t)\, \delta i_1(t)\rangle_{av}$, and the other is at T_2 and dissipates $\langle \delta v(t) \delta i_2(t)\rangle_{av}$.

Now, since both D_1 and D_2 are bilateral, the noise power available from the network N of Fig. 5 does not vary[25] if the temperatures of D_1 and D_2 are changed so that they become equal to

$$\frac{\langle \delta v(t)\ \delta i_1(t)\rangle_{av} T_1 + \langle \delta v(t)\ \delta i_2(t)\rangle_{av} T_2}{\langle \delta v(t)\ \delta i_1(t)\rangle_{av} + \langle \delta v(t)\ \delta i_2(t)\rangle_{av}}$$

which, together with eq. 67 gives eq. 66.

Equation 66 is of particular interest when $\omega_1 = p$, since in this case it corresponds to the example considered in the introduction. Notice that T_{av} is not, in general, a function of the diode characteristics alone.

In fact, it also depends both on the particular characteristics of the network N' in which the diode is imbedded and on the value of the frequency ω_1 at which T_{av} is defined, unless $T(t)$ is constant.

Under conditions of practical interest T always varies with time and consequently a direct application of theorem 1 is never strictly valid. Equation 66 shows, however, that if certain conditions are satisfied, then T_{av} is little affected by the particular choice of ω_1 and N', and consequently a direct application of theorem 1 may not introduce significant errors. More precisely, suppose that either

$$g_d(v) \approx 0 \qquad \text{or} \qquad \frac{1}{g_d(v)} \approx 0 \tag{68}$$

for some values of v and that

$$T_d(v) \approx T' = \text{constant} \tag{69}$$

over the range of voltages for which conditions 68 are not satisfied. Under these conditions either $T(t) \approx T'$ or $\delta v(t)\, \delta i(t) \approx 0$, for all values of t, and consequently from equation 66 one obtains $T_{av} \approx T'$. Therefore in this case, and only in this case, T_{av} can be regarded as a function of the diode characteristics alone and theorem 1 is applicable, with T replaced by T'.

An important application of the preceding result is given by an ideal Schottky barrier diode. In fact, the relations derived in Ref. 22 between the noise figure and the conversion loss of such a diode imply that its junction can be represented, under certain particular conditions, by means of an ordinary resistive network held at half the temperature T_0 of the junction.

REFERENCES

1. Strutt, M. J. O., "Mixing Valves," Wireless Eng., *12*, No. 137 (February 1935, pp. 59–64.
2. Peterson, E. and Hussey, L. W., "Equivalent Modulator Circuits," B.S.T.J., *18*, No. 1 (January 1939), pp. 32–48.
3. Herald, E. W. and Malter, L., "Some Aspects of Radio Reception at Ultra-High Frequencies," Proc. IRE, *31*, No. 10 (October 1943), pp. 575–582.
4. Peterson, L. C. and Llewellyn, F. B., "The Performance of Mixers in Terms of Linear-Network Theory," Proc. IRE, *33*, No. 7 (July 1945), pp. 458–476.
5. Herold, E. W., Bush, R. R., and Ferris, W. R., "Conversion Loss of Diode Mixers Having Image-Frequency Impedance," Proc. IRE, *33*, No. 9 (September 1945), pp. 603–609.
6. Strutt, M. J. O., "Noise Figure Reduction in Mixer Stages," Proc. IRE, *34*, No. 12 (December 1946), pp. 942–950.
7. Torrey, H. C. and Whitmer, C. A., *Crystal Rectifiers*, Radiation Laboratory Series, *15*, New York: McGraw-Hill, 1948.
8. Pound, R. V., *Microwave Mixers*, Radiation Laboratory Series, *16*, New York: McGraw-Hill, 1948.

9. Crawford, A. B., material included in G. C. Southworth's *Principles and Applications of Waveguide Transmission,* Princeton, N. J.: van Nostrand, 1950, pp. 626–636.
10. Strum, P. D., "Some Aspects of Crystal Mixer Performance," Proc. IRE, *41,* No. 7 (July 1953), pp. 876–889.
11. van der Ziel, A., *Noise,* New York: Prentice-Hall, 1954.
12. Pritchard, W. L., "Notes on a Crystal Mixer Performance," Trans. IRE, *MTT-3,* No. 11 (January 1955), pp. 37–39.
13. Strum, P. D., "A Note on Noise Temperature," IRE Trans., Microwave Theory and Technique, *MTT-4,* No. 3 (July 1956), pp. 145–151.
14. Messenger, G. C. and McCoy, C. T., "Theory and Operation of Crystal Diodes as Mixers," Proc. IRE, *45,* No. 9 (September 1957), pp. 1269–1283.
15. Houlding, N., "Mixer Crystal Noise," Proc. IRE, *46,* No. 5, Part 1 (May 1958), pp. 917–918.
16. Kim, C. S., "Tunnel Diode Converter Analysis," IRE Trans. Electron Devices, *ED-8,* No. 5 (September 1961), pp. 394–405.
17. Pucel, R. A., "Theory of the Esaky Diode Frequency Converter," Solid-State Electronics, *3,* No. 3 (November 1961), pp. 167–207.
18. Lo, Shih-Fang, "Noise in Tunnel Diode Mixers," Proc. IRE, *49,* No. 11 (November 1961), pp. 1688–1689.
19. Sterzer, F. and Presser A., "Stable Low Noise Tunnel-Diode Frequency Converters," RCA Rev., *23,* No. 1 (March 1962), pp. 3–28.
20. Barber, M. R., "A Numerical Analysis of the Tunnel-Diode Frequency Converter," IEE Trans., Microwave Theory and Technique, *MTT-13,* No. 5 (September 1965), pp. 663–665.
21. Barber, M. R. and Ryder, R. M., "Ultimate Noise Figure and Conversion Loss of the Schottky Barrier Mixer Diode," Int. Microwave Symp. Digest, May 1966, pp. 13–17.
22. Barber, M. R., "Noise Figure and Conversion Loss of the Schottky Barrier Mixer Diode," IRE Trans., Microwave Theory and Technique, *MTT-15,* No. 11 (November 1967), pp. 629–635.
23. Eng, S. T., "A New Low 1/f Noise Mixer Diode; Experiments, Theory and Performance," Solid State Elec., *8,* No. 1 (January 1965), pp. 59–77.
24. Twiss, R. Q., "Nyquist's and Thevin's Theorems Generalized for Non-reciprocal Linear Networks," J. Appl. Phys., *26,* No. 5 (May 1955), pp. 599–602.
25. Siegman, A. E., "Thermal Noise in Microwave Systems," Microwave J., *4,* No. 3 (March 1961), pp. 81–90.
26. Rice, S. O., "Mathematical Analysis of Random Noise," B.S.T.J., *23,* No. 3 (July 1944), pp. 282–332.
27. Bennett, V. R., *Electrical Noise,* New York: McGraw-Hill, 1960.
28. Rowe, H. E., *Signals and Noise in Communications Systems,* Princeton, N. J.: van Nostrand, 1965.
29. Manley, J. M. and Rowe, H. E., "Some General Properties of Nonlinear Elements, Part 1, General Energy Relations," Proc. IRE, *44,* No. 7 (July 1956), pp. 904–913.
30. Adler, R. B. and Haus, H. A., *Circuit Theory of Linear Noisy Networks,* New York: John Wiley & Sons, 1959.
31. Uhlir, Jr., A., "Shot Noise in p-n Junction Frequency Converters," B.S.T.J., *37,* No. 4 (July 1958), pp. 951–988.
32. Davenport, W. B. and Root, W. L., *Random Signals and Noise,* New York: McGraw-Hill, 1958.

The Analysis of Circular Waveguide Phased Arrays*

By N. AMITAY and VICTOR GALINDO

(Manuscript received July 9, 1968)

In this work, a planar phased array of circular waveguides arranged in an equilateral triangular grid is analyzed. The boundary value problem is first formulated rather generally in terms of a vector two dimensional integral equation for an array of elements that are arranged in a doubly periodic grid along two skewed (nonorthogonal) coordinates. Dielectric plugs, covers, and loading, as well as thin irises at the aperture, are accounted for in the formulation. Numerical solutions are obtained by using the Ritz–Galerkin method to solve the integral equation. Excellent agreement with experimental measurements using a waveguide simulator is observed. The existence of forced surface wave phenomena in equilateral triangular grid arrays and their strong dependence upon the mode of excitation is also demonstrated. These phenomena are shown to exist at isolated points in the scan coordinates. Reflection characteristics as well as the polarization characteristics of the radiation pattern are illustrated at selected planes of scan for both linear and circular polarization excitation.

I. INTRODUCTION

The requirements of modern radar and communication systems have stimulated considerable activity in the design and use of phased array antennas. To date, the design information required for their development has been obtained from the analysis of simplified array models and from experimental data. The great speed and storage capacity of present day digital computers, however, have now made it possible to solve the planar phased array boundary value problem very accurately.[1, 2]

A general formulation of the planar phased array boundary value problem may be found. A vector two dimensional integral equation

* The work reported in this paper was supported by the U. S. Army Materiel Command under contract DA-30-069-AMC-333(Y).

for the tangential aperture field (that is, the tangential field at the planar interface between the waveguides and free space) can then be derived.

In its most general form the array elements are assumed to be arranged in a doubly periodic grid along two skewed (nonorthogonal) coordinates; and dielectric loading, covers, and plugs, as well as thin irises at the aperture plane, may be accounted for in the analysis. The possibility of multimode excitation of the array has also been included. The Ritz-Galerkin method is applied to obtain a solution for circular waveguide arrays.

Numerical solutions for the reflection characteristics of dielectric-free planar arrays of circular waveguides hexagonally arranged in a conducting ground plane have been carried out. Experimental measurements have been made which compare favorably with the results. Forced surface waves[3-11] are found to occur at isolated points in the scan coordinates and can be related to certain vector and geometrical symmetries for an equilateral triangle grid array. These surface waves (or resonances) are often difficult to locate experimentally by the use of waveguide simulators[12-14] or small test arrays. The strong dependence of these forced resonances upon the mode of excitation is also demonstrated. The reflection characteristics as well as the polarization characteristics of the radiation pattern are illustrated for various combinations of linear and circular polarization excitation of the array.

II. ANALYSIS

An infinite planar array of waveguide elements, Fig. 1, is imbedded in a conducting ground plane at its interface (plane $z = 0$) with free space. The elements are arranged in a periodic grid along the skewed (nonorthogonal) coordinates s_1 and s_2. The x and s_1 axes coincide while the s_2 axis makes an angle α with respect to the x (and s_1) axis. The element location is defined by two indices (p, q) corresponding to a physical location

$$\varrho_{pq} = pb\hat{s}_1 + qd\hat{s}_2 \tag{1}$$

where $\hat{s}_1$ and $\hat{s}_2$ are unit vectors along the s_1 and s_2 axes, while b and d represent the basic periods of the two dimensional grid. A basic periodic cell,[15] the parallelogram shown in Fig. 1, is thereby defined. If the array elements are excited uniformly in amplitude with a linear

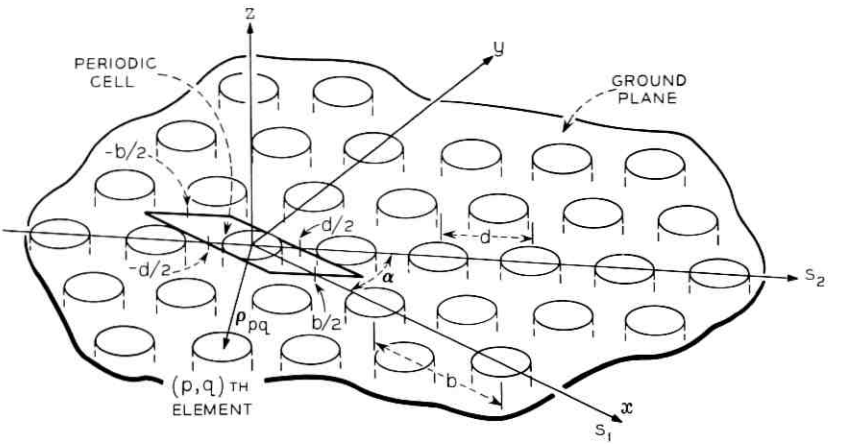


Fig. 1 — Circular waveguide array geometry.

phase taper such that the phase of the (p, q)th element is

$$\psi_{pq} = p\psi_{t_1} + q\psi_{t_2} , \tag{2}$$

then the resulting electromagnetic fields in the (p, q)th and (m, n)th periodic cells satisfy the following periodicity relationship

$$\mathbf{E}(\varrho_{pq}) = \mathbf{E}(\varrho_{mn}) \exp [i \{(m - p)\psi_{t_1} + (n - q)\psi_{t_2}\}] \tag{3}$$

where $\mathbf{E}(\varrho_{pq})$ may designate the electric or magnetic field at the (p, q)th periodic cell of the grid.* Therefore, except for a phase factor, the fields in all the cells are identical.

In order to solve the boundary value problem, the exterior (free space) fields are expressed in terms of a complete set of Floquet type solutions of Maxwell's equations $\{\Psi^p_{mn} \exp \pm iB_{mn}z\}$. These vector modes, which are functions of the steering phases ψ_{t_1} and ψ_{t_2} are derived in Appendix A. The interior $(z \leqq 0)$ fields are expressed in terms of the appropriate waveguide complete orthonormal set of vector modes $\{\Phi_i \exp \pm i\Gamma_i z\}$.† The boundary value problem is expressed in terms of an integral equation which includes the necessary continuity conditions. This equation is formulated[1] by satisfying the continuity of the transverse (to z)

* The parallelogram in Fig. 1 defines the $(p = 0, q = 0)$ cell. The (p, q)th cell is translated by pb and qd along the s_1 and s_2 axes, respectively.

† The waveguide modes are real functions and in general consist of both TE and TM modes with double subscripts. However, one can always systematically relabel these modes with a single subscript according to the increasing values of the eigenvalues.

electric and magnetic fields within a single periodic cell. As shown in Appendix B, the periodic cell consisting of the parallelogram *CDEF* of Fig. 2 can be replaced, without a loss of generality, by the parallelogram *GHIJ* (or any other periodic contiguous cell containing a complete single waveguide aperture).

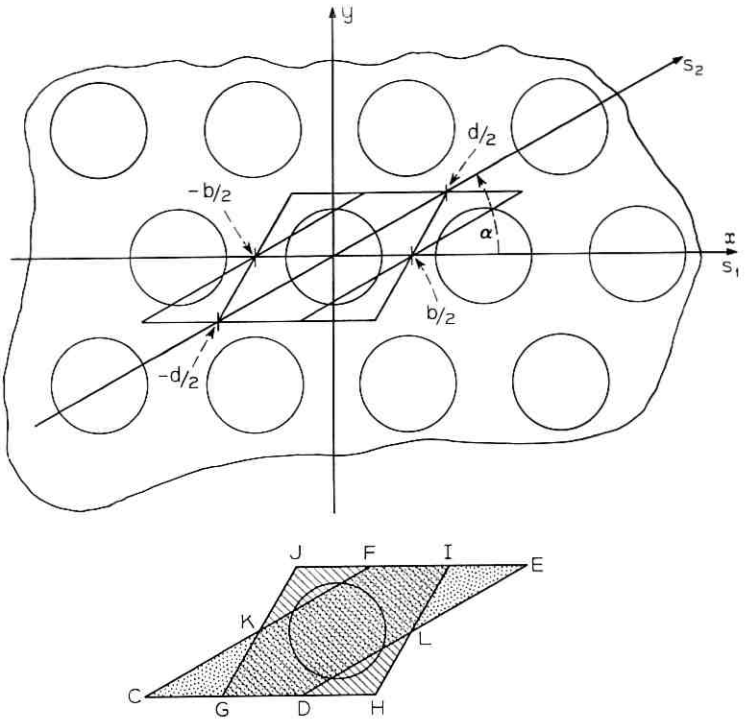


Fig. 2 — Periodic cell in skewed array geometry.

The tangential electric and magnetic fields at the array interface (**E** and **H** at $z = 0$) can be expressed in terms of a Fourier series of the complete orthonormal set of waveguide modes $\{\Phi_j\}$ for $z \leqq 0$ and by the set of Floquet type modes $\{\Psi^p_{mn}\}$ for $z \geqq 0$. Let the waveguides be excited by *any* linear combination of their propagating modes* with amplitudes $A_j (j = 1, \ldots, J$ for J propagating modes) and let the coefficients R_j represent the amplitudes of the corresponding reflected

* It is straightforward to include, if desired, *any* linear combination of both propagating and evanescent modes.

modes.* Let the coefficients D_i represent the amplitudes of the reflected evanescent modes which are generated at the aperture. Then, in terms of the waveguide modes, the electric field at $z = 0$ is given by

$$\mathbf{E}_- = \begin{cases} \sum_{i=1}^{J} (A_i + R_i)\mathbf{\Phi}_i + \sum_{i=J+1}^{\infty} D_i\mathbf{\Phi}_i & \text{(over the waveguide aperture)} \\ 0 \quad \text{(over the rest of the periodic cell).} \end{cases} \tag{4}$$

The corresponding magnetic field is

$$-\mathbf{H}_- = \begin{cases} \sum_{i=1}^{J} Y_i(A_i - R_i)\mathbf{\Phi}_i - \sum_{i=J+1}^{\infty} Y_i D_i \mathbf{\Phi}_i & \\ \qquad \text{(over the waveguide aperture)} & \\ 0 \quad \text{(over the rest of the periodic cell).} \end{cases} \tag{5}$$

where the modal admittances $\{Y_i\}$ are real for propagating modes and pure imaginary for evanescent modes. These admittances are given by[16]

$$Y_i = \frac{\Gamma_i}{\omega\mu} \quad \text{for} \quad TE \text{ modes}; \qquad Y_i = \frac{\omega\epsilon}{\Gamma_i} \quad \text{for} \quad TM \text{ modes} \tag{6}$$

for an exp $[-i\omega t]$ time convention with the Γ_i (the z propagation constant) being positive imaginary for evanscent modes. The tangential electric field at $z = 0^+$, expressed in terms of the Floquet type modes, is

$$\mathbf{E}_+ = \sum_{p=1}^{2} \sum_{(m)} \sum_{(n)} F_{mnp}\mathbf{\Psi}^p_{mn} \qquad \text{(over the periodic cell)} \tag{7}$$

where the superscript p designates TE($p = 1$) or TM($p = 2$) modes. The magnetic field is correspondingly given by

$$-\mathbf{H}_+ = \sum_{p=1}^{2} \sum_{(m)} \sum_{(n)} F_{mnp} Y'_{mnp} \mathbf{\Psi}^p_{mn} \tag{8}$$

where the modal admittances Y'_{mnp} are given by

$$Y'_{mn1} = \frac{B_{mn}}{\omega\mu}; \qquad Y'_{mn2} = \frac{\omega\epsilon}{B_{mn}}. \tag{9}$$

From the orthonormality of the sets $\{\mathbf{\Phi}_i\}$ and $\{\mathbf{\Psi}^p_{mn}\}$ it is clear that

$$\langle \mathbf{E}_- , \mathbf{\Phi}_i \rangle = \iint_A \mathbf{E}_- \cdot \mathbf{\Phi}_i \, da = \begin{cases} A_i + R_i & \text{for} \quad j \leqq J \\ D_i & j > J \end{cases} \tag{10}$$

* Actually $\{R_i\}$ can represent the reflection coefficients once $\{A_i\}$ are properly normalized.

where A is the waveguide aperture, and

$$\langle \mathbf{E}_+ , \boldsymbol{\Psi}^p_{mn} \rangle = \iint_{\text{periodic cell}} \mathbf{E}_+ \cdot (\boldsymbol{\Psi}^p_{mn})^* \, da = \iint_A \mathbf{E}_+ \cdot (\boldsymbol{\Psi}^p_{mn})^* \, da = F_{mnp} \tag{11}$$

where the symbol * denotes the complex conjugate. Notice that $\mathbf{E}_+$ vanishes on the conducting ground plane. To insure the continuity of the tangential fields across the aperture at $z = 0$, one requires

$$\mathbf{E}_+ = \mathbf{E}_- = \mathbf{E}_t \text{ over the aperture and the periodic cell} \tag{12}$$

while

$$\mathbf{H}_+ = \mathbf{H}_- = \mathbf{H}_t \text{ over the aperture } \textit{only}. \tag{13}$$

Using the various relations, (4) through (13), one obtains an integral equation for the tangential electric field $\mathbf{E}_t$ at the array interface

$$2 \sum_{j=1}^{J} A_j Y_j \boldsymbol{\Phi}_j = \sum_{j=1}^{\infty} Y_j \boldsymbol{\Phi}_j \iint_A \boldsymbol{\Phi}_j \cdot \mathbf{E}_t \, da + \sum_{p=1}^{2} \sum_{(m)} \sum_{(n)} Y'_{mnp} \boldsymbol{\Psi}^p_{mn} \iint_A (\boldsymbol{\Psi}^p_{mn})^* \cdot \mathbf{E}_t \, da. \tag{14}$$

Similarly, one can obtain[1] an integral equation for $\mathbf{H}_t$ which is defined over the entire periodic cell. Under certain conditions[17] it is possible to interchange the order of summation and integration in (14) and thereby obtain the usual form of Fredholm integral equation of the first kind.

A useful method of obtaining a solution to (14) is by the application of the Ritz-Galerkin method,[18] whereby the integral equation is reduced to a linear matrix equation. Substituting (4) in (14) for $\mathbf{E}_t$ and taking the moments of (14) with respect to the set $\{\boldsymbol{\Phi}_i\}$, while using (10), leads to the following matrix equation

$$2 \begin{bmatrix} Y_1 A_1 \\ \vdots \\ Y_J A_J \\ 0 \\ \vdots \end{bmatrix} = |K| \begin{bmatrix} A_1 + R_1 \\ \vdots \\ A_J + R_J \\ D_{J+1} \\ \vdots \end{bmatrix}, \tag{15}$$

where $|K|$ is a square matrix with the (i, q)th element given by

$$k_{iq} = Y_i \, \delta_{iq} + \sum_{p=1}^{2} \sum_{(m)} \sum_{(n)} Y'_{mnp} C^p_{mni} C^{p*}_{mnq} . \tag{16}$$

In (16) δ_{iq} is the kroenecker delta

$$\delta_{iq} = \begin{cases} 1 & \text{if } i = q \\ 0 & \text{if } i \neq q \end{cases} \tag{17}$$

and

$$C^{p}_{mni} = \iint_{A} \tilde{\mathbf{\Psi}}^{p}_{mn} \cdot \mathbf{\Phi}_i \, da \tag{18}$$

is the coupling coefficient between the designated interior and exterior modes.*

The practicality of an accurate numerical solution or approximate analytical solution often hinges upon obtaining C^{p}_{mni} in closed form. Recently the authors[19] have obtained closed form expressions of these coupling coefficients for circular as well as coaxial waveguides.

A solution for the aperture field in terms of the coefficients of the waveguide modes is given by

$$\cdot \begin{bmatrix} A_1 + R_1 \\ \vdots \\ A_J + R_J \\ D_{J+1} \\ \vdots \end{bmatrix} = 2 \mid K \mid^{-1} \begin{bmatrix} A_1 Y_1 \\ \vdots \\ A_J Y_J \\ 0 \\ \vdots \end{bmatrix}. \tag{19}$$

The solution vector can be obtained by matrix inversion or by rapidly convergent iterative methods.[5,20] A similar procedure should be followed to obtain $\mathbf{H}_t$ except that the aperture field and the moments may be taken with respect to the set $\{\mathbf{\Psi}^{p}_{mn}\}$.

Once the aperture electric field is obtained, the input impedance and radiation properties of the array, as a function of scan, are easily obtained. The reflection coefficients are obtained directly from (19), as are the amplitudes of the evanescent modes in the waveguides. The radiation pattern of a single element in the array environment, including its polarization characteristics, can also be easily obtained.[21,22]

The addition of either a dielectric sheath or plug or both to the

* Notice that other sets of functions $\{\xi_l\}$ can be used to reduce (14) by the method of moments. However, the integrability of

$$\iint_{(A)} \mathbf{\Phi}_j \cdot \xi_l \text{ and } \iint_{(A)} \mathbf{\Psi}_{mn}{}^{p} \cdot \xi_l$$

may prove difficult depending on ξ_l. The convergence properties as a function of the order of K will also be influenced by the choice of $\{\xi_l\}$.

array, Fig. 3, does not alter the functional form of the integral equation (14). As was shown by Galindo and Wu,[3, 23] only the modal admittances in (14) have to be replaced. For the case when dielectric plugs are used, Y_j is replaced by

$$Y_j \to y_j \frac{Y_j - iy_j \tan \gamma_j t_1}{y_j - iY_j \tan \gamma_j t_1} \tag{20}$$

where y_j and γ_j are the modal admittance and propagation constants, respectively, of the Φ_j mode in the dielectric and t_1 is the dielectric plug thickness. The relation between the reflection coefficients of the propagating modes in the dielectric-free region (ϵ_0) and the aperture field is given by

$$R_j = \frac{y_j \iint_A \mathbf{E}_t \cdot \mathbf{\Phi}_j \, da - A_j \exp(-i\Gamma_j t_1)[y_j \cos \gamma_j t_1 + iY_j \sin \gamma_j t_1]}{\exp(i\Gamma_j t_1)[y_j \cos \gamma_j t_1 - iY_j \sin \gamma_j t_1]}, \tag{21}$$

the phase of R_j being referred to the aperture. Similarly, for the exterior dielectric sheath covering the array, the modal admittances are replaced by

$$Y'_{mnp} \to y'_{mnp} \frac{Y'_{mnp} - iy'_{mnp} \tan \beta_{mn} t'_1}{y'_{mnp} - iY'_{mnp} \tan \beta_{mn} t'_1}, \tag{22}$$

where y'_{mnp} and β_{mn} are the modal admittance and propagation constants, respectively, of the Ψ^p_{mn} mode in the sheath, and t'_1 is the sheath thickness. The coefficient of the mode Ψ^p_{mn} in the free space region above the sheath, F_{mnp}, is related to the aperture field by the following relation:

$$F_{mnp} = \frac{\exp(-iB_{mn} t'_1) y'_{mnp}}{y'_{mnp} \cos \beta_{mn} t_1' - iY'_{mnp} \sin \beta_{mn} t'_1} \iint_{\substack{\text{periodic} \\ \text{cell}}} \mathbf{E}_t \cdot \mathbf{\Psi}^{p*}_{mn} \, da, \tag{23}$$

the phase of F_{mnp} being referred to the aperture plane.

The integral equation formulation can be extended to the case when thin metallic irises are present at the waveguide aperture for matching purposes (see Fig. 4). The integral equation for the aperture electric field, (14), is still valid except that the integral has to be defined over the effective aperture with the result that the orthogonal relations of (10) and (11) cannot be used for the Ritz-Galerkin method of solution.* The modal coupling coefficients, (18), are still integrable in closed form

* An integral equation for the magnetic field is not valid in this case because of the discontinuity of $\mathbf{H}_t$ across the iris.

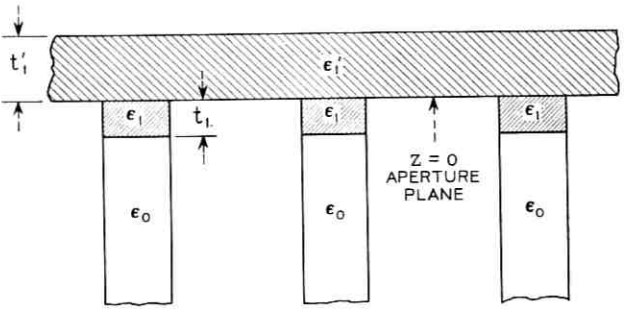


Fig. 3 — Dielectric sheath and plug geometry.

over the effective aperture if the iris is circularly symmetric. However, the matrix elements, as given by (16), do change in form when an iris is present at the aperture. As a result of the Φ_i not being orthogonal over the new effective aperture, the term $Y_i\delta_{iq}$ in (16) is replaced by an infinite sum.

Multimode excitation of waveguide antenna fields has been used for primary pattern control. Such excitations may prove useful for the reduction of mutual coupling effects.[7] They also may be used to obtain circular or elliptical polarization from two linearly polarized modes and improve the polarization characteristics of the array. In the circular waveguide array, the horizontal TE_{11} (Φ_1) and vertical TE_{11} (Φ_2) modes are degenerate (that is, they have the same z-directed propagation constant and impedance). In order to obtain a linear, elliptical or circular polarization excitation of the waveguide, one may redefine the first two modes as Φ_{1N} and Φ_{2N}:

$$\Phi_{1N} = \frac{A_1}{(|A_1|^2 + |A_2|^2)^{\frac{1}{2}}}\Phi_1 + \frac{A_2}{(|A_1|^2 + |A_2|^2)^{\frac{1}{2}}}\Phi_2 \tag{24}$$

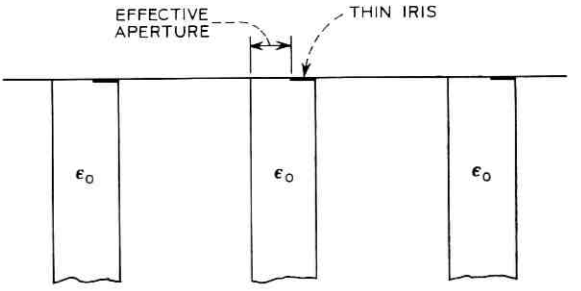


Fig. 4 — Aperture iris geometry.

$$\Phi_{2N} = -\frac{|A_2|}{(|A_1|^2 + |A_2|^2)^{\frac{1}{2}}}\Phi_1 + \frac{A_1^* |A_2|}{A_2^*(|A_1|^2 + |A_2|^2)^{\frac{1}{2}}}\Phi_2 . \quad (25)$$

This redefinition of the first two modes preserves the orthonormality and completeness of the set of waveguide modes while allowing the flexibility in adjusting the desired array excitation. The reflection coefficients which correspond to these redefined modes are R_{1N} and R_{2N}. The polarization characteristics of the radiation pattern may be determined from the θ and ϕ components of the radiated field, E_θ and E_ϕ, respectively. After proper normalization of E_θ and E_ϕ one may obtain the corresponding transmission coefficients[19]

$$T_\theta = \left(\frac{Y'_{002}}{Y_1}\right)^{\frac{1}{2}} \frac{1}{(1 - T_x^2 - T_y^2)^{\frac{1}{2}}} \iint_{\text{(aperture)}} \mathbf{E}_t \cdot (\Psi_{00}^2)^* \, da \quad (26)$$

$$T_\phi = -\left(\frac{Y'_{001}}{Y_1}\right)^{\frac{1}{2}} \iint_{\text{(aperture)}} \mathbf{E}_t \cdot (\Psi_{00}^1)^* \, da \quad (27)$$

where T_x and T_y define the scan angle directional cosines. When the first two waveguide modes are the only propagating ones and while only a single lobe propagates in the free space, equations (24) through (27) are related by the conservation of energy relation:

$$|R_{1N}|^2 + |R_{2N}|^2 + |T_\theta|^2 + |T_\phi|^2 = 1. \quad (28)$$

For more than one lobe in free space or additional propagating modes in the waveguides, (28) has to be accordingly modified.

III. NUMERICAL AND EXPERIMENTAL RESULTS

In order to obtain a numerical solution for the aperture field, the infinite dimensional matrix of equation (15) must be truncated and cast in a finite dimensional form. In other words, the electromagnetic fields will be approximated by a finite Fourier series of the waveguide and free space modes, and consequently the solution of the problem as given by (19), is finite dimensional as well.

In numerical solutions of problems of this type, various ways of ascertaining the validity and accuracy of the solution are desirable. One obvious way is to increase the number of waveguide and free space modes and check the convergence of the solution as a function of the number of modes used in truncating (15). However, for the type of kernel involved in this problem, monotonicity of the convergence is not assured. Nevertheless, convergence is an important check

since this numerical solution is variational or stationary for the impedance.[12, 24] Iterative methods and error estimates[5, 20] may also be used for checking the convergence of the solution. Special symmetries of the reflection coefficients versus the scan angles which are dictated by the array geometry and mode of excitation can serve as a semi-independent check. An important independent check in which the reflection and transmission coefficients can be measured at special scan angles with the aid of a waveguide simulator is used as well.

The numerical results will in general be presented as a function of scan angle. For convenience, however, the differential phase shifts between elements will be used as the independent variables. These quantities are ψ_x in the x-direction and ψ_y in the y-direction. Furthermore, since we limit ourselves to radial planes of scan, we introduce the quantity ψ_r. These quantities are related to the directional cosines as follows:

$$\psi_x = \frac{2\pi b}{\lambda} T_x \; ; \qquad \psi_y = \frac{2\pi d \sin \alpha}{\lambda} T_y \; ; \qquad \psi_r = (\psi_x^2 + \psi_y^2)^{\frac{1}{2}}. \tag{29}$$

The amount of computation can be reduced when one recognizes that the following symmetry in the aperture field as a function of scan exists:

$$\mathbf{E}_t(\psi_x, \psi_y) = \mathbf{E}_t(-\psi_x, -\psi_y). \tag{30}$$

Convergence tests as a function of the number of waveguide and free space modes indicated that 18 waveguide modes and 338 free space modes yield several percent (usually less than 2 percent) accuracy in the magnitude of the reflection coefficients, R_j, except near sharp changes of R_j where the position of the sharp changes is accurate to several degrees in ψ_r.

The energy conservation check, equation (28), is a necessary check but not a sufficient one in this problem as well as in other interior type boundary value problems.[17, 22, 33]

One of various special symmetry checks is depicted in Fig. 5. A square grid array in the (x, y) coordinates is excited by the vertical $\mathrm{TE}_{11}(\mathbf{\Phi}_2)$ mode. In this coordinate system the array parameters are represented in the following way

$$\left.\begin{aligned} a = \text{waveguide radius}; \qquad b = d; \qquad \alpha = 90° \\ \mathbf{\Phi}_{1N} = \mathbf{\Phi}_2 \; ; \qquad \mathbf{\Phi}_{2N} = -\mathbf{\Phi}_1 \end{aligned}\right\}. \tag{31}$$

The parameters of the same array, when viewed in the (x', y') coordi-

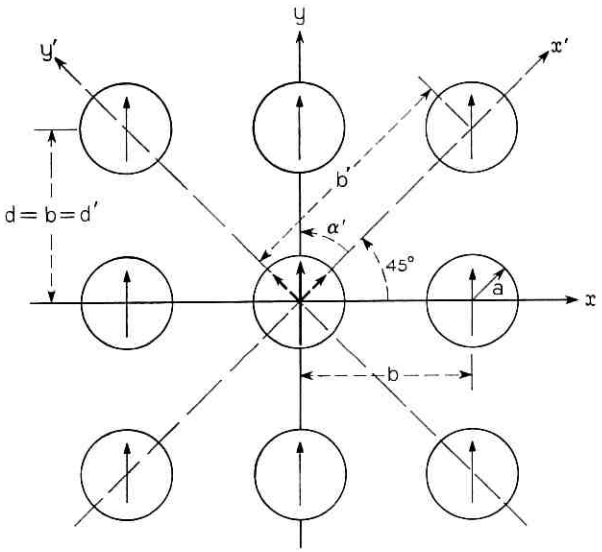


Fig. 5 — Symmetries in square grid arrays.

nate system, can be alternatively represented as

$$\left.\begin{aligned} b' &= (2)^{\frac{1}{2}}d' = (2)^{\frac{1}{2}}d; \qquad \alpha' = 45° \\ \Phi'_{1N} &= \frac{1}{(2)^{\frac{1}{2}}}\{-\Phi'_1 + \Phi'_2\}; \qquad \Phi'_{2N} = \frac{1}{(2)^{\frac{1}{2}}}\{\Phi'_1 + \Phi'_2\} \end{aligned}\right\} \tag{32}$$

where the reference to the (x', y') system is denoted by primes. The same results should be found for this array at any scan direction regardless of the representation. This offers, to a degree, a check on roundoff error. Numerically, the reflection coefficients differed only by a fraction of a percent.

An additional symmetry check is given in Fig. 6, where the magnitudes of the reflection coefficients are plotted versus ψ_r for a 45° plane of scan. R_{1N} and R_{2N} correspond to the reflection coefficients of Φ_{1N} and Φ_{2N} as defined by (31). At $\psi_r = 240°$ (shown by a vertical arrow) the main beam is grazing, while for $\psi_r > 240°$ no beam exists in real space and the total incident power is reflected and divided between the two propagating modes, Φ_{1N} and Φ_{2N}. Of special interest is the point $\psi_r = 180° \times (2)^{\frac{1}{2}} \approx 255°$. At this point $\psi_x = \psi_y = \psi_{i_1} = \psi_{i_2} = 180°$ and the array excitation is as indicated in the inset of Fig. 6. If the array is to be simulated at this scan angle, the appropriate waveguide simulator would

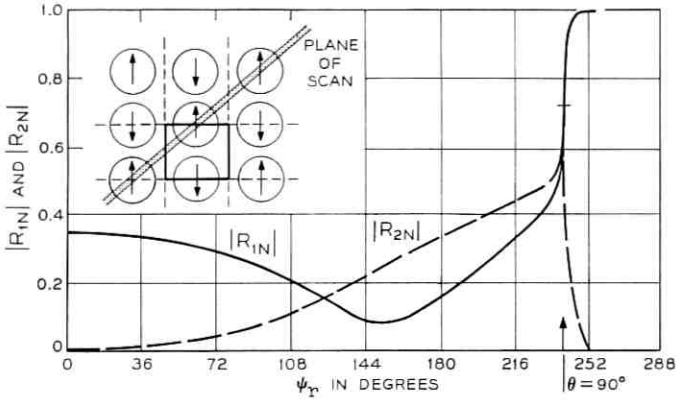


Fig. 6 — Square grid array: $|R_{1N}|$ and $|R_{2N}|$ vs ψ_r in the 45° plane of scan. $\alpha = 0.48$, $b = 1$, $d = 1$, $\lambda = 1.5$ and $\alpha = 90°$.

consist of the square waveguide (solid lines) shown in the inset. It is clear that the horizontal waveguide mode (Φ_{2N}) cannot be excited. The numerical results indeed show that $| R_{2N} | = 0$ at this scan angle.

Figure 7 shows a close agreement between experimental and numerical results for a rectangular grid array with vertical polarization excitation scanned in the H-plane. The scan angle which corresponds to steering phases $\psi_x = \psi_{t_1} = 180°$ and $\psi_y = \psi_{t_2} = 0$, can be simulated

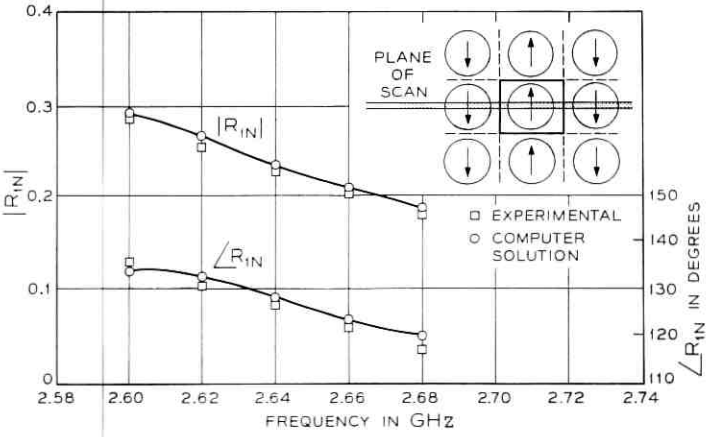


Fig. 7 — Rectangular grid array: computed and experimental results vs frequency at H-plane symmetry point. $a = 3.57$ cm., $b = 16.51$ cm., $d = 8.261$ cm., and $\alpha = 90°$.

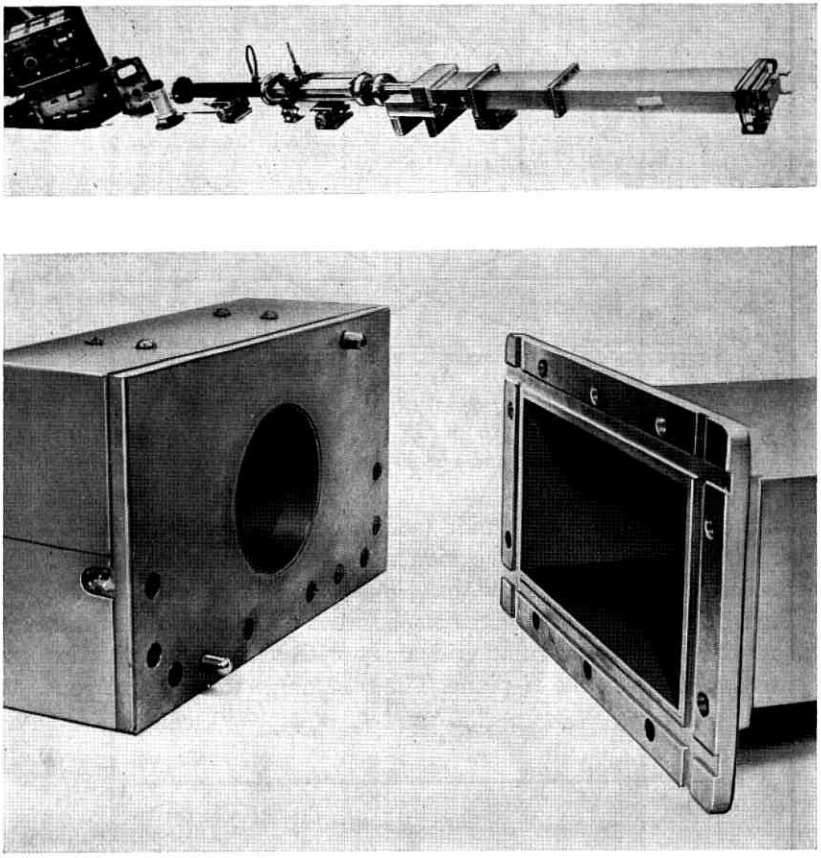


Fig. 8 — Rectangular grid array: simulator for H-plane symmetry point.

by the rectangular waveguide (solid line) in the inset. The experimental results were obtained from measurements of an abrupt junction between a circular waveguide and an L-band rectangular waveguide as shown in Fig. 8.

IV. EQUILATERAL TRIANGLE GRID ARRAYS

Let us consider the reflection and radiation characteristics of circular waveguide arrays arranged in an equilateral triangle grid, and the strong dependence of the array properties upon the mode of excitation. Grating lobe incipience or a beam at grazing is designated by a vertical arrow in the illustrations.

Figure 9 shows the reflection coefficient of an array in the E-plane scan, with vertical polarization excitation. In this plane the horizontal mode is not excited because of symmetry so that $R_{2N} \equiv 0$. As can be seen, the slope of both the magnitude and phase of R_{1N} is discontinuous and singular at a grating lobe incipience, which parallels previous observations in rectangular waveguide arrays.[26, 27] This is related to the asymptotic decay of the coupling coefficients. A forced surface wave resonance can be seen around grating lobe incipience where $|R_{1N}| = 1.0$. Notice that this forced surface wave resonance is extremely sharp and consequently may not be observed in small finite arrays.

The corresponding transmission coefficient T_θ is shown in Fig. 10. The plot of the transmission coefficient is actually the radiation pattern of a single element in the array environment and it exhibits the null which corresponds to a total reflection. Notice that the phase of the transmission coefficient will exhibit a 180° discontinuity when the magnitude has a zero. The magnitude of the reflection coefficients of the same array, in a 60° plane of scan, are shown in Fig. 11. Again,

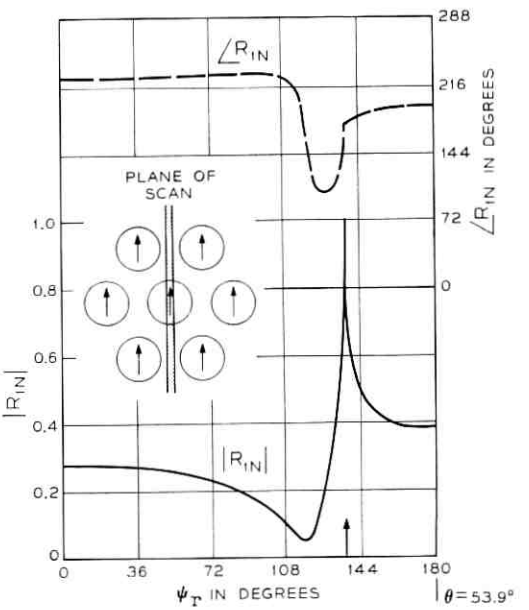


Fig. 9 — E-plane scan of R_{1N} vs ψ_T ($a = 0.48$, $b = 1$, $d = 1$, $\lambda = 1.4$ and $\alpha = 60°$).

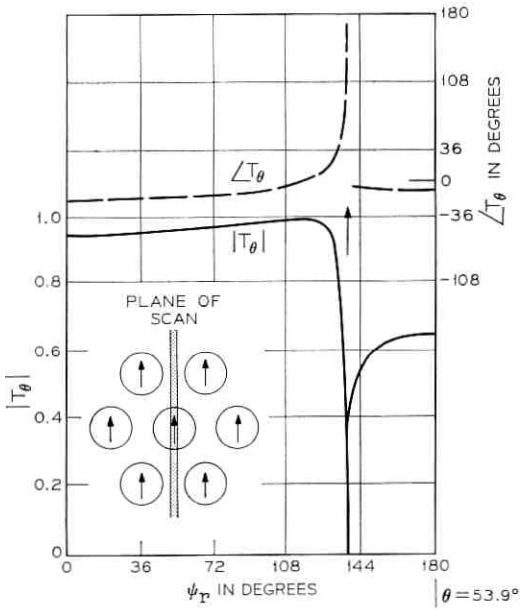


Fig. 10 — E-plane scan of T_θ vs ψ_r ($a = 0.48$, $b = 1$, $d = 1$, $\lambda = 1.4$, and $\alpha = 60°$).

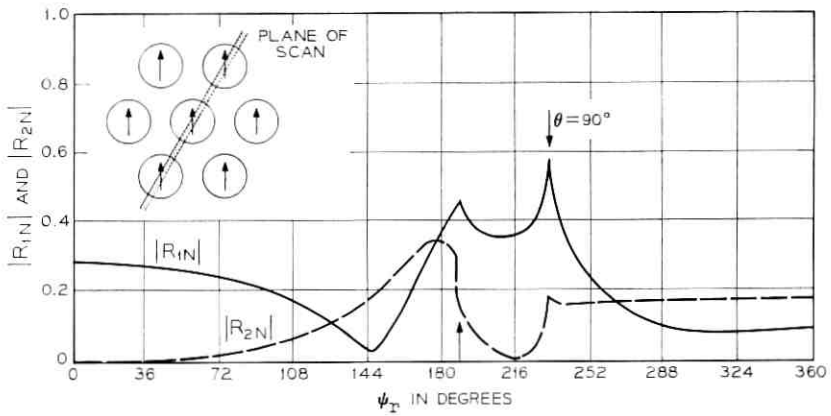


Fig. 11 — $|R_{1N}|$ and $|R_{2N}|$ vs ψ_r in the 60° plane of scan ($a = 0.48$, $b = 1$, $d = 1$, $\lambda = 1.4$, and $\alpha = 60°$).

the singular slope at grating lobe incipience can be observed. As can be seen, the distribution of the reflected power between the two modes is a function of the scan angle.

An interesting phenomenon can be observed in the corresponding transmission characteristics which are shown in Fig. 12. The ϕ trans-

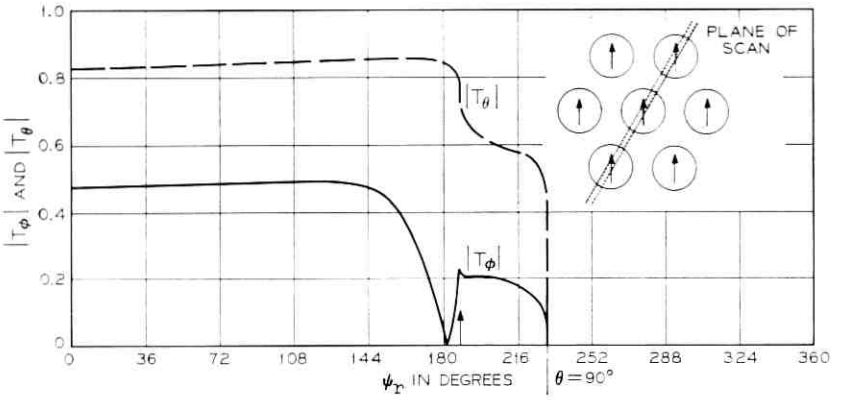


Fig. 12 — $|T_\theta|$ and $|T_\phi|$ vs ψ_r in the 60° plane of scan ($a = 0.48$, $b = 1$, $d = 1$, $\lambda = 1.4$, and $\alpha = 60°$).

mission coefficient, T_ϕ, vanishes prior to grating lobe incipience (the positions of the vertical arrows). The vanishing of T_ϕ at this scan angle can be directly related, when coupled with the vector symmetries in the array excitation and geometry, to a forced surface wave resonance. If the array excitation consists of the sum of the two modes (equal in phase and magnitude) indicated by the solid arrows in Fig. 13, then in the 60° scan plane $T_\theta \equiv 0$ by symmetry considerations and T_ϕ vanishes as shown in Fig. 12. Since the vectorial sum of the two solid arrows in Fig. 13 is the dashed arrow, zero transmission or a forced surface wave resonance will occur in the H-plane of this polarization. Figs. 14 and 15 indeed show this effect in the H-plane scan where $|R_{1N}|$ and $|T_\phi|$ attain unity and zero respectively prior to grating lobe incipience.*

Since the forced surface waves are related to the vector symmetries just mentioned, one may anticipate that the scan points at which they occur are isolated. Figure 16 indeed demonstrates that the forced resonances in the E and H planes occur at isolated points. The scan around

* The difference in the values of ψ_r at which these phenomena occur is inherent in the definition of ψ_r, equation (20).

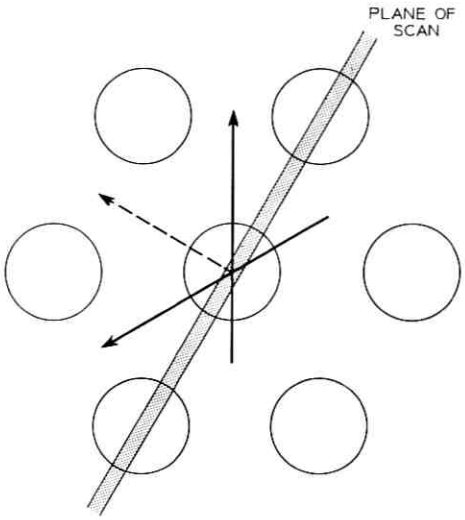


Fig. 13 — Vector symmetry relationships between transmission zeros and forced surface waves.

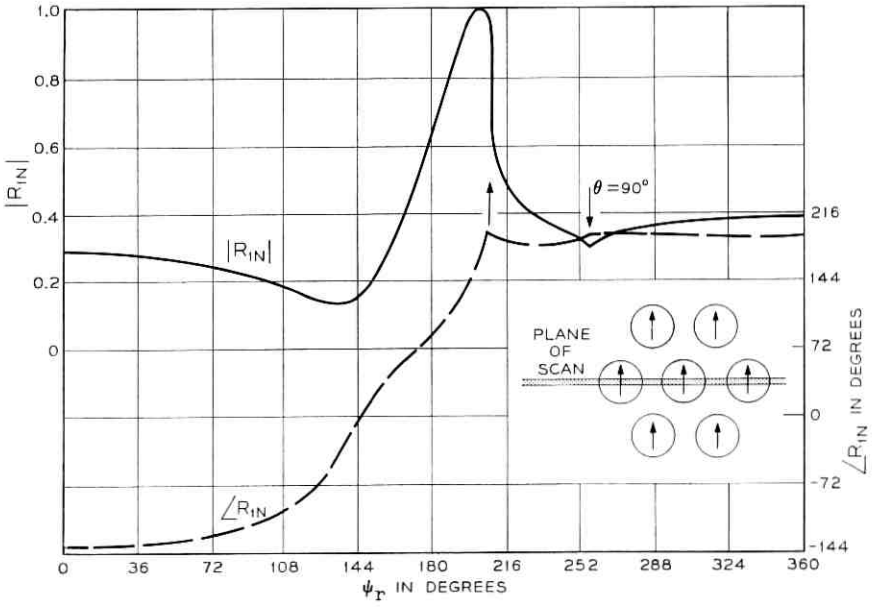


Fig. 14 — H-plane scan of R_{1N} vs ψ_r ($a = 0.48$, $b = 1$, $d = 1$, $\lambda = 1.4$, and $\alpha = 60°$)

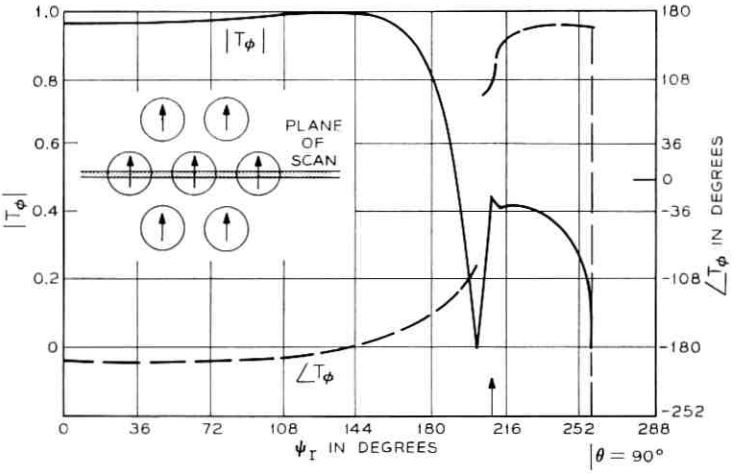


Fig. 15 — H-plane scan of T_ϕ vs ψ_T ($a = 0.48$, $b = 1$, $d = 1$, $\lambda = 1.4$, and $\alpha = 60°$).

the grating lobe circles in Fig. 17 shows that the peak of the *total* reflection $|R_T| = (|R_{1N}|^2 + |R_{2N}|^2)^{\frac{1}{2}}$ varies from unity in the H plane ($\delta = 0$), gradually decreases, and then increases again and reaches unity in the E-plane ($\delta = 90°$). The anomalous behavior near grating lobe incipience is eliminated when the polarization of the excitation is changed to horizontal, as shown in Fig. 17, indicating thus the strong dependence of the forced surface wave resonances upon the mode of excitation. Over the frequency band given by $1.3 \leqq \lambda \leqq 1.5$, qualitatively similar behavior of the radiation and reflection characteristics of the array was observed.

Figure 18 shows the reflection characteristics of the array under circular polarization excitation. In this case

$$\mathbf{\Phi}_{1N} = -\frac{1}{(2)^{\frac{1}{2}}}\{\mathbf{\Phi}_1 + i\mathbf{\Phi}_2\}; \qquad \mathbf{\Phi}_{2N} = -\frac{1}{(2)^{\frac{1}{2}}}\{\mathbf{\Phi}_1 - i\mathbf{\Phi}_2\}. \quad (33)$$

The incident mode is $\mathbf{\Phi}_{1N}$.

Again the singular slope of these curves at grating lobe incipience can be seen. The division of the reflected power between the two modes as a function of scan may be observed as well. Fig. 19 shows the polarization characteristics of the radiation pattern of an array element. The axial ratio, denoted as $A.R.$, is the ratio of the minor to major axis of the polarization ellipse while the tilt angle of the major axis, τ, is taken with respect to the ϕ axis. As indicated by

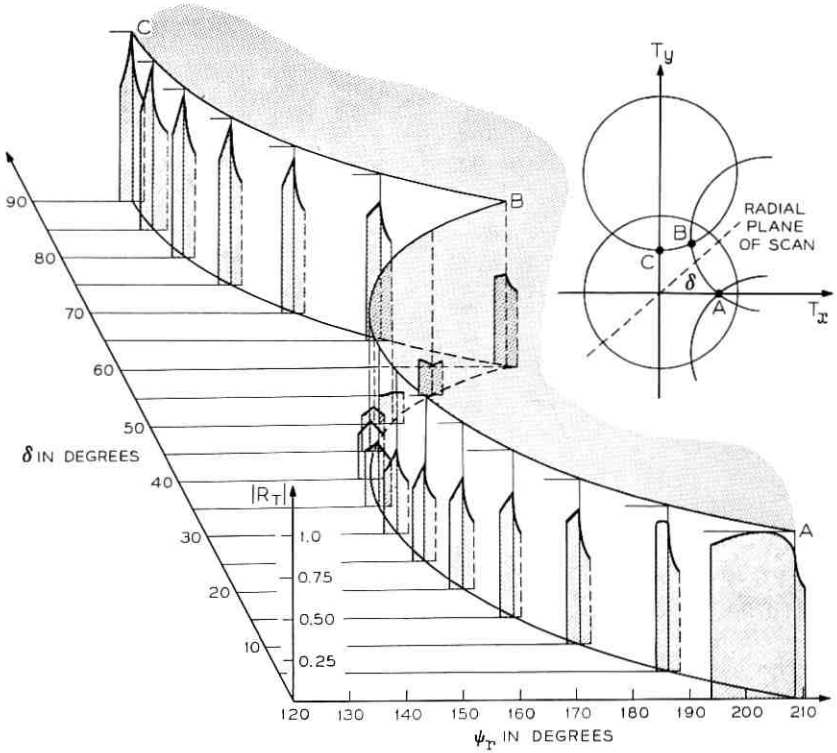


Fig. 16 — $|R_T|$ vs ψ_T in the vicinity of grating lobe incipience.

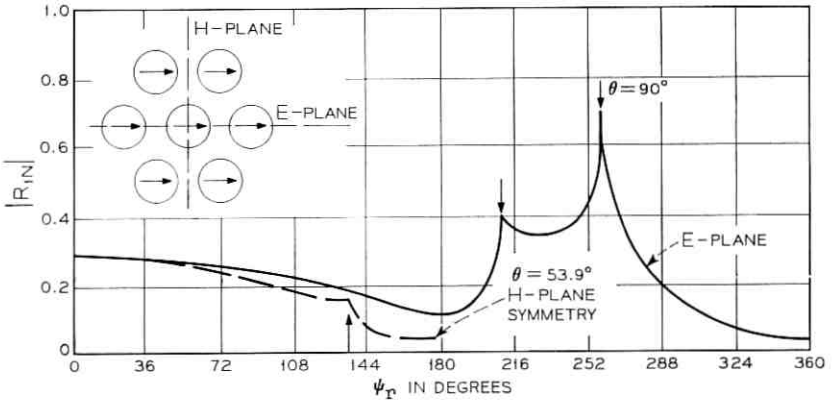


Fig. 17 — E- and H-plane scans of $|R_{1N}|$ vs ψ_T for horizontal polarization ($a = 0.48$, $b = 1$, $d = 1$, $\lambda = 1.4$, and $\alpha = 60°$).

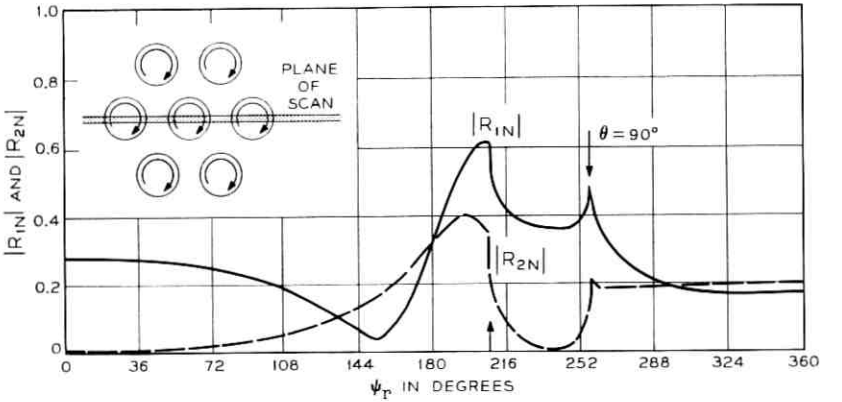


Fig. 18 — $|R_{1N}|$ and $|R_{2N}|$ vs ψ_r in the 0° plane of scan. Circularly polarized excitation ($a = 0.48$, $b = 1$, $d = 1$, $\lambda = 1.4$, and $\alpha = 60°$).

the plot, the element (or the array) far field pattern is circularly polarized around broadside, $A.R. = 1.0$. It deteriorates to linear polarization, $A.R. = 0$, at two points prior to grating lobe incipience. The linear polarization at $\psi_r = 203.5°$ with $\tau = 90°$ results from the H-plane forced surface wave resonance of Figs. 14 and 15 where T_ϕ vanishes. The null of the axial ratio at $\psi_r = 207°$ (with $\tau = 105°$)

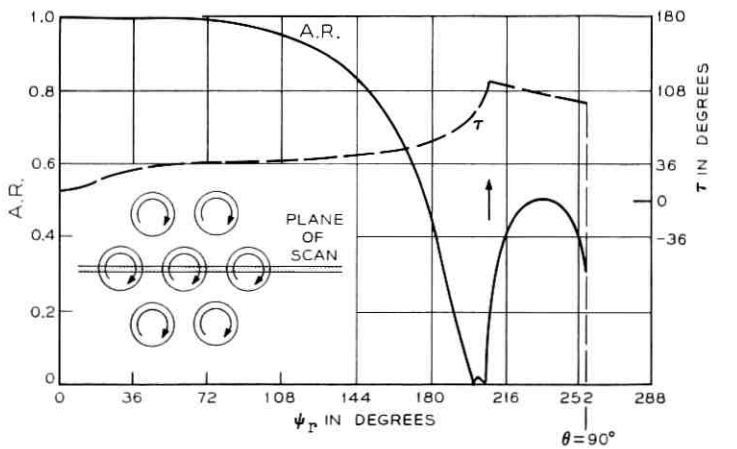


Fig. 19 — Radiated axial ratio (A.R.) and tilt angle (τ) vs ψ_r in the 0° plane of scan. Circularly polarized excitation ($a = 0.48$, $b = 1$, $d = 1$, $\lambda = 1.4$, and $\alpha = 60°$).

is caused by the difference between the phases of T_θ and T_ϕ causing the ϕ and θ components of the far field to be in phase.

The reflection and polarization characteristics in the 30° plane of scan are shown in Figs. 20 and 21, respectively. In this case the single null of the axial ratio at grating lobe incipience results from the E-plane forced surface wave resonance of Figs. 9 and 10. For planes of scan between 0° and 30° the results corresponding to one plane change gradually to those of the other plane. Around grating lobe incipience the axial ratio drops appreciably (to around 0.1) but does not reach zero. From the circular symmetry of excitation and six-fold symmetry of the array geometry, a 30° sector of scan completely specifies the array reflection and radiation characteristics.

V. CONCLUSIONS

A general formulation of the planar phased array boundary value problem has been given in terms of a vector two dimensional integral equation. The solution of this equation by the Ritz-Galerkin method closely agreed with experimental results.

Equilateral triangle phased arrays of circular waveguides were numerically analyzed. It was found that forced aperture resonances or forced surface waves, manifested by total reflection and no radiation, do exist for these arrays even in the absence of dielectric materials. These effects were observed over a 15 per cent frequency band. The forced aperture resonances occurred prior or close to grat-

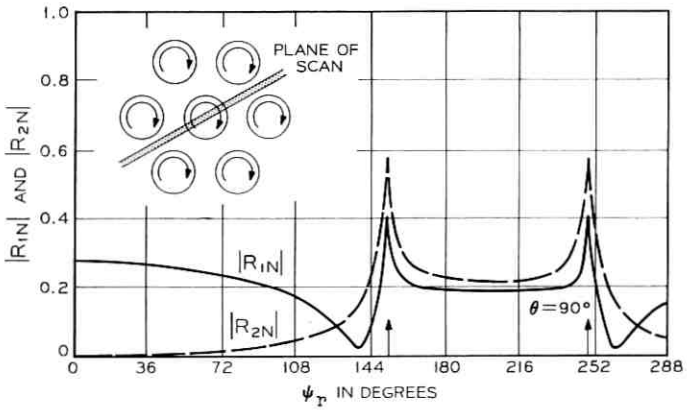


Fig. 20 — $|R_{1N}|$ and $|R_{2N}|$ vs ψ_r in the 30° plane of scan. Circularly polarized excitation ($a = 0.48$, $b = 1$, $d = 1.\lambda = 1.4$, and $\alpha = 60°$).

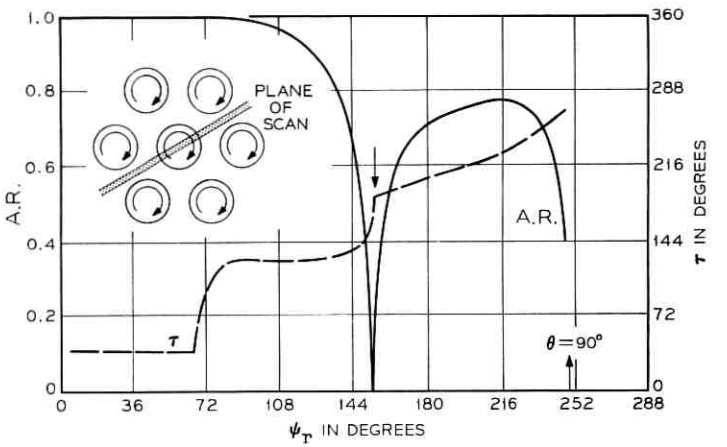


Fig. 21 — Radiated axial ratio (A.R.) and tilt angle (τ) vs ψ_T in the 30° plane of scan. Circularly polarized excitation ($a = 0.48$, $b = 1$, $d = 1$, $\lambda = 1.4$, and $\alpha = 60°$).

ing lobe incipience in the E and H plane of scan for vertical polarization excitation. These resonances were found to occur at isolated points as a function of the scan variables and are strongly influenced by the mode of excitation. The resonances vanish when the polarization of excitation changes from vertical to horizontal.

The polarization characteristics of the radiation pattern (or alternatively the radiation pattern of a single excited element in the array environment) is shown at selected planes of scan for circular polarization excitation. The degradation of the axial ratio resulting from the forced surface waves was shown. Total reflection or no transmission owing to forced aperture resonances were not observed for circular polarization excitation in the cases presented.

The analysis of coaxial waveguide arrays, as well as the incorporation of thin, circularly symmetric irises in the aperture of the waveguide element, can be carried out along lines similar to those discussed here.

The effects of dielectric loading of the array as well as dielectric covers and plugs have also been studied.[34]

VI. ACKNOWLEDGMENTS

The invaluable assistance of Mr. S. Renna and Miss A. M. Russell in the computations, and Messrs. P. E. Butzien and L. H. Hendler in the experimental measurements, are gratefully acknowledged. The

authors would like to thank Dr. C. P. Wu, Dr. J. A. Cochran, and Dr. E. R. Nagelberg for their helpful suggestions.

APPENDIX A

Floquet Type Wave Functions in Skewed Coordinates

Consider the periodic array, Figs. 1 and 2, excited with incremental phase shifts (steering phases) between adjacent cells along the s_1 and s_2 coordinates. A complete set of solutions to the scalar wave equation, each of which varies periodically, according to Floquet's theorem, along the s_1 and s_2 coordinates is

$$S_{mn} = \exp(iB_{mn}z) \exp i\left[\frac{\psi_{t_1}}{b} - \frac{2\pi m}{b}\right]s_1 \exp i\left[\frac{\psi_{t_2}}{d} - \frac{2\pi n}{d}\right]s_2 \qquad (34)$$

with the integers $m, n = -\infty, \ldots -1, 0, 1, 2, 3, \ldots, +\infty$. Equation (34) describes a wave traveling (or decaying) in the z direction with propagation constant B_{mn} (exp $-i\omega t$ time convention). The steering phases ψ_{t_1} and ψ_{t_2} are directly related to the beam pointing direction, $\hat{r}$, of a radiated plane wave with a vector propagation constant $\mathbf{k}_0 = k_0\hat{r}$, so that (34) can be rewritten as

$$S_{mn} = \exp(iB_{mn}z) \exp i\left[\mathbf{k}_0 \cdot \hat{s}_1 - \frac{2\pi m}{b}\right]s_1 \exp i\left[\mathbf{k}_0 \cdot \hat{s}_2 - \frac{2\pi n}{d}\right]s_2 . \qquad (35)$$

The free space propagation vector $\mathbf{k}_0$ can be expressed in the cartesian coordinate system as

$$\mathbf{k}_0 = k_0[T_x\hat{x} + T_y\hat{y} + T_z\hat{z}] \qquad (36)$$

where T_x, T_y and T_z are the directional cosines of $\mathbf{k}_0$ with respect to that system and $\hat{x}$, $\hat{y}$, and $\hat{z}$ are the unit vectors. The quantities

$$\mathbf{k}_0 \cdot \hat{s}_1 = \frac{\psi_{t_1}}{b} \quad \text{and} \quad \mathbf{k}_0 \cdot \hat{s}_2 = \frac{\psi_{t_2}}{d} \qquad (37)$$

are the projections of $\mathbf{k}_0$ on the reciprocal grid (lattice) coordinates t_1 and t_2, respectively.[15,28-30] The unit vectors in the t_1 and t_2 directions form a biorthogonal set with $\hat{s}_1$ and $\hat{s}_2$ (Fig. 22). To express (35) in cartesian coordinates it can easily be shown that

$$s_1 = x - y \cot \alpha, \qquad s_2 = y/\sin \alpha, \qquad (38)$$

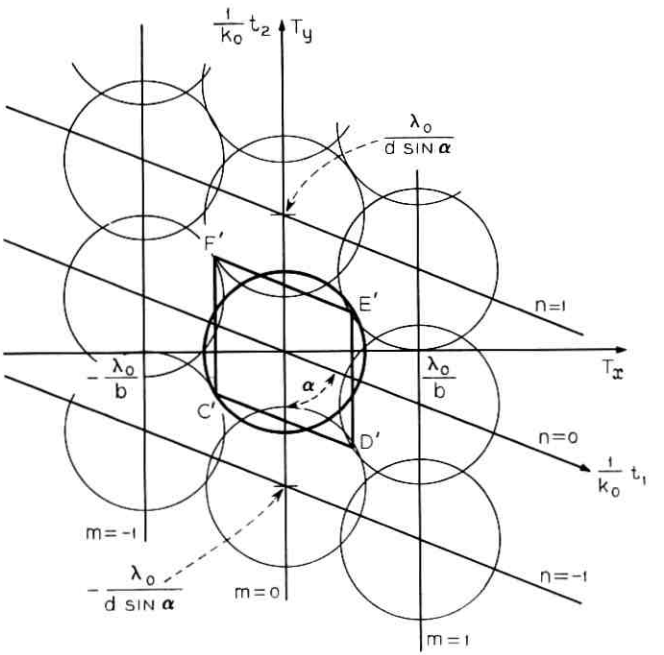


Fig. 22 — Grating lobe diagram in reciprocal lattice coordinates.

so that the substitution of (38) into (35) yields

$$S_{mn} = \exp(iB_{mn}z) \exp i\left[k_0 T_x - \frac{2\pi m}{b}\right]x \cdot \exp i\left[k_0 T_y - \left(\frac{2\pi n}{d \sin \alpha} - \frac{2\pi m}{b \tan \alpha}\right)\right]y. \quad (39)$$

The propagation constants of the (m, n)th Floquet mode along the x and y directions, k_x and k_y respectively, are

$$\begin{aligned} k_x &\equiv \mathbf{k} \cdot \hat{x} = k_0 T_x - \frac{2\pi m}{b}; \\ k_y &\equiv \mathbf{k} \cdot \hat{y} = k_0 T_y - \left(\frac{2\pi n}{d \sin \alpha} - \frac{2\pi m}{b \tan \alpha}\right). \end{aligned} \quad (40)$$

Since S_{mn} is a solution to the scalar wave equation, it can be shown that

$$B_{mn} = (k_0^2 - k_x^2 - k_y^2)^{\frac{1}{2}} = \left\{k_0^2 - \left[k_0 T_x - \frac{2\pi m}{b}\right]^2 - \left[k_0 T_y - \left(\frac{2\pi n}{d \sin \alpha} - \frac{2\pi m}{b \tan \alpha}\right)\right]^2\right\}^{\frac{1}{2}} \tag{41}$$

where the positive imaginary root holds for $(k_x^2 + k_y^2) > k_0^2$ (time convention exp $-i\omega t$).

Each mode, S_{mn}, for which B_{mn} is real corresponds to a radiated plane wave of the phased array. The plane wave with the indices $m = 0$ and $n = 0$ is identified with the main beam while m or $n \neq 0$ corresponds to a radiating grating lobe. As a function of T_x and T_y (or ψ_{t_1} and ψ_{t_2}), a given B_{mn} may become pure imaginary as it goes through a zero, as in equation (41). In such cases, the related Floquet mode, S_{mn}, becomes evanescent or nonradiating. By plotting the curves obtained by setting the $B_{mn} = 0$ as a function of T_x and T_y, one obtains a convenient diagram which illustrates these effects. Setting $B_{mn} = 0$ yields

$$\left[T_x - \frac{m\lambda_0}{b}\right]^2 + \left[T_y - \left(\frac{n\lambda_0}{d \sin \alpha} - \frac{m\lambda_0}{b \tan \alpha}\right)\right]^2 = 1, \tag{42}$$

where $\lambda_0 = 2\pi/k_0$. As a function of T_x and T_y, (42) represents a family of circles with unit radius displaced from the origin. This diagram of displaced circles constitutes the well-known grating lobe diagram, Fig. 22.[31, 32]

Notice that the steering phases, ψ_{t1} and ψ_{t2}, are related to T_x and T_y through equations (36) and (37). The parallelogram $C'D'E'F'$ of Fig. 22 corresponds to the range of steering phases

$$-\pi \leqq \psi_{t_1} \leqq \pi; \qquad -\pi \leqq \psi_{t_2} \leqq \pi \tag{43}$$

and is a periodic cell along the t_1 and t_2 coordinates.

As mentioned in Section II, it is possible to define a complete orthonormal set[16] of vector modes $\{\Psi_{mn}^{v}\}$ over the parallelogram $CDEF$, Fig. 2. The tangential electromagnetic field at the plane $z = 0^+$ can be expressed by a Fourier series of this set of modes which consists of both TE and TM modes (transverse to z). These modes, $\{\Psi_{mn}^{\mathrm{TE}}\}$ and $\{\Psi_{mn}^{\mathrm{TM}}\}$, are given by

$$\Psi_{mn}^{\mathrm{TE}} = \frac{\exp i(xk_x + yk_y)}{(bd \sin \alpha)^{\frac{1}{2}}} \left\{\frac{k_y}{k_r}\hat{x} - \frac{k_x}{k_r}\hat{y}\right\}\Bigg|_{m,n} \tag{44}$$

and

$$\Psi_{mn}^{\mathrm{TM}} = \frac{\exp i(xk_x + yk_y)}{(bd \sin \alpha)^{\frac{1}{2}}} \left\{\frac{k_x}{k_r}\hat{x} + \frac{k_y}{k_r}\hat{y}\right\}\Bigg|_{m,n} \tag{45}$$

with $k_r = (k_x^2 + k_y^2)^{\frac{1}{2}}$. The quantities k_x and k_y are functions of (m, n) as given by (40). The orthonormality of this set of vector modes is defined by the following scalar products:

$$\langle \Psi_{mn}^{\mathrm{TE}}, \Psi_{pq}^{\mathrm{TE}} \rangle = \iint_{\substack{\text{parallelogram} \\ CDEF}} \Psi_{mn}^{\mathrm{TE}} \cdot (\Psi_{pq}^{\mathrm{TE}})^* \, dx \, dy = \delta_{mn,pq} , \tag{46}$$

where

$$\delta_{mn,pq} = \begin{cases} 1 & \text{for} \quad m = p \quad \text{and} \quad n = q, \\ 0 & \text{otherwise} \end{cases} \tag{47}$$

$$\langle \Psi_{mn}^{\mathrm{TE}}, \Psi_{pq}^{\mathrm{TM}} \rangle = 0 \tag{48}$$

and

$$\langle \Psi_{mn}^{\mathrm{TM}}, \Psi_{pq}^{\mathrm{TM}} \rangle = \delta_{mn,pq} . \tag{49}$$

APPENDIX B

On the Invariance of the Scalar Product with the Shape of the Periodic Cell

The orthonormality and completeness of the set $\{\Psi_{mn}^{p}\}$ of Floquet modes, equations (44) and (45), need not be defined over a specific periodic cell such as the parallelogram $CDEF$ of Fig. 2. This fact is especially significant when a periodic cell intercepts parts of more than one circular (or other type of element) aperture. It will be shown that the orthonormality of $\{\Psi_{mn}^{p}\}$ can be preserved over a properly deformed periodic cell which contains only a single waveguide aperture.

Using (37), equation (44) can be rewritten as

$$\Psi_{mn}^{\mathrm{TE}} = \mathbf{F}(m, n) \exp i\left[\frac{\psi_{t_1} - 2\pi m}{b} s_1 + \frac{\psi_{t_2} - 2\pi n}{d} s_2\right] . \tag{50}$$

The scalar product betewen two TE modes is

$$\begin{aligned} \langle \Psi_{mn}^{\mathrm{TE}}, \Psi_{pq}^{\mathrm{TE}} \rangle &= \delta_{mn,pq} \\ &= \iint_{(CDEF)} \Psi_{mn}^{\mathrm{TE}} \cdot (\Psi_{pq}^{\mathrm{TE}})^* \, dx \, dy \\ &= \iint_{(CDEF)} \Psi_{mn}^{\mathrm{TE}} \cdot (\Psi_{pq}^{\mathrm{TE}})^* \sin \alpha \, ds_1 \, ds_2 \\ &= \mathbf{F}(m, n) \cdot \mathbf{F}(pq)^* \sin \alpha \end{aligned} \tag{51}$$

$$\cdot \iint_{(CDEF)} \exp -i\left[\frac{2\pi(m-p)}{b}\, s_1 + \frac{2\pi(n-q)}{d}\, s_2\right] ds_1\, ds_2 \,.$$

The integral over the parallelogram $CDEF$ in (51) can be divided into three (or more) integrals over the triangles CGK and LEI and the polygon $GDLIFK$ (Fig. 2):

$$\iint_{CDEF} = \iint_{CGK} + \iint_{LEI} + \iint_{GDLIFK} . \tag{52}$$

Because of the array periodicity, the triangles DHL and LEI are displaced by b with respect to the triangles CGK and KFJ, respectively, along the s_1 direction. Thus, for example, if the s_1 coordinate of the points within the triangle DHL is s_1' given by

$$s_1' = s_1 + b \qquad \text{with} \qquad ds_1' = ds_1 \tag{53}$$

then

$$\iint_{DHL} \exp -i\left[\frac{2\pi(m-p)}{b}\, s_1' + \frac{2\pi(n-q)}{d}\, s_2\right] ds_1'\, ds_2$$

$$= \exp -i2\pi(m-p) \iint_{CGK} \exp -i\left[\frac{2\pi(m-p)}{b}\, s_1 + \frac{2\pi(n-q)}{d}\, s_2\right] ds_1\, ds_2$$

$$= \iint_{CGK} \exp -i\left[\frac{2\pi(m-p)}{b}\, s_1 + \frac{2\pi(n-q)}{d}\, s_2\right] ds_1\, ds_2 \,. \tag{54}$$

Similarly

$$\iint_{LEI} \exp -i\left[\frac{2\pi(m-p)}{b}\, s_1 + \frac{2\pi(n-q)}{d}\, s_2\right] ds_1\, ds_2$$

$$= \iint_{KFJ} \exp -i\left[\frac{2\pi(m-p)}{b}\, s_1 + \frac{2\pi(n-q)}{d}\, s_2\right] us_1\, ds_2. \tag{55}$$

Thus

$$\iint_{CDEF} \exp -i\left[\frac{2\pi(m-p)}{b}\, s_1 + \frac{2\pi(n-q)}{d}\, s_2\right] ds_1\, ds_2$$

$$= \iint_{GHIJ} \exp -i\left[\frac{2\pi(m-p)}{b}\, s_1 + \frac{2\pi(n-q)}{d}\, s_2\right] ds_1\, ds_2 \tag{56}$$

and the orthonormality and completeness of the set $\{\Psi^{p}_{mn}\}$ is preserved over the new periodic cell $GHIJ$. This is the two dimensional analog of the single dimensional Fourier series whereby the functions and coefficients are independent of the initial value of the period. In fact, the two dimensional periodic cell can be deformed to any shape which contains a single waveguide provided that the area of the cell stays the same and the parts of the cell which cause the deformation are translated by b or d along the s_1 and s_2 coordinates, respectively.

REFERENCES

1. Galindo, V. and Wu, C. P., "Numerical Solutions for an Infinite Phased Array of Rectangular Waveguides with Thick Walls," IEEE Trans. Antennas and Propagation, *AP-14,* No. 2 (March 1966), pp. 149–158.
2. Wu, C. P. and Galindo, V., "Properties of a Phased Array of Rectangular Waveguides with Thin Walls," IEEE Trans. on Antennas and Propagation," *AP-14,* No. 2 (March 1966), pp. 163–173.
3. Wu, C. P. and Galindo, V., "Surface-Wave Effects on Dielectric Sheathed phased Arrays of Rectangular Waveguides," B.S.T.J., *47,* No. 1, (January 1968), pp. 117–142.
4. Wu, C. P. and Galindo, V., "Surface-Wave Effects on Phased Arrays of Rectangular Waveguides Loaded with Dielectric Plugs," IEEE Trans. Antennas and Propagation, *AP-16,* No. 3 (May 1968), pp. 358–360.
5. Amitay, N. and Galindo, V., "Application of a New Method for Approximate Solutions and Errors Estimates to Waveguide Discontinuity and Phased Array Problems," *Radio Science, 3,* No. 8 (August 1968), pp. 830–844.
6. Amitay, N., Cook, J. S., Pecina, R. G., and Wu, C. P., "On Mutual Coupling and Matching Conditions in Large Planar Phased Arrays," Proc. 1964 IEEE Antennas and Propagation Int. Symp., L. I., N. Y., September 1964, pp. 150–156.
7. Lechtreck, L. W., "Effects of Coupling Accumulation in Antenna Arrays," IEEE Trans. Antennas and Propagation, *AP-16,* No. 1 (January 1968), pp. 31–37.
8. Farrell, G. F., Jr. and Kuhn, D. H., "Mutual Coupling Effects of Triangular-Grid Arrays by Modal Analysis," IEEE Trans. Antennas and Propagation, *AP-14,* No. 5 (September 1966), pp. 652–654.
9. Diamond, B. L., "Resonance Phenomena in Waveguide Arrays," Proc. 1967 IEEE Antennas and Propagation Int. Symp., Ann Arbor, Mich., October 1967, pp. 110–115.
10. Galindo, V., "A Generalized Approach to a Solution of Aperiodic Arrays and Modulated Surfaces," IEEE Trans. Antennas and Propagation, *AP-16,* No. 4 (July 1968), pp. 424–430.
11. Collin, R. E., *Field Theory of Guided Waves,* New York: McGraw-Hill, 1960, pp. 430–441 and pp. 465–468.
12. Galindo, V. and Wu, C. P., "Integral Equations and Variational Expressions for Arbitrary Scanning of Regular Infinite Arrays," IEEE Trans. Antenas and Propagation, *AP-14,* No. 3. (May 1966), pp. 392–394.
13. Hannan, P. W., and Balfour, M. A., "Simulation of a Phased-Array Antenna in Waveguide," IEEE Trans. Antennas and Propagation, *AP-13,* No. 3 (May 1965), pp. 342–353.
14. Balfour, M. A., "Phased Array Simulators in Waveguide for Triangular Arrangement of Elements," IEEE Trans. Antennas and Propagation, *AP-13,* No. 3, (May 1965), pp. 475–476.
15. Brillouin, L., *Wave Propagation in Periodic Structures,* New York: Dover Publications, 1953, pp. 94–121.

16. *Waveguide Handbook,* ed. N. Marcuvitz, Radiation Laboratory Series., vol. 10, New York: McGraw-Hill, 1951.
17. Titchmarsh, E. C., *The Theory of Functions,* London: Oxford University Press, 1939, pp. 43–45.
18. Kantorovich, L. V. and Krylov, V. I., *Approximate Methods of Higher Analysis,* New York: Interscience Publishers, 1964.
19. Amitay, N. and Galindo, V., "On the Scalar Product of Certain Circular and Cartesian Wave Functions," IEEE Trans. Microwave Theory and Techniques, *MTT-16,* No. 4, (April 1968), pp. 265–266.
20. Cole, W. J., Nagelberg, E. R., and Nagel, C. M., "Iterative Solution of Waveguide Discontinuity Problems," B.S.T.J., *46,* No. 3 (March 1967), pp. 649–672.
21. Galindo, V. and Wu, C. P., "The Relation Between the Far-Zone Pattern of the Singly Exited Element and the Transmission Coefficient of the Principle Lobe in an Infinite Array," IEEE Trans. Antennas and Propagation, *AP-14,* No. 3 (May 1966), pp. 397–398.
22. Amitay, N. and Galindo, V., "A Note on the Radiation Characteristics and Forced Surface Wave Phenomena in Triangular Grid Circular Waveguide Phased Arrays," scheduled to be published in IEEE Trans. Antennas and Propagation, *AP-16,* No. 6 (November 1968).
23. Galindo, V. and Wu, C. P., "Dielectric Loaded and Covered Rectangular Waveguide Phased Arrays," B.S.T.J., *47,* No. 1 (January 1968), pp. 93–116.
24. Wu, C. P., "Note on Integral Equations and Variational Expressions for Arbitrary Scanning of Regular Infinite Arrays," IEEE Trans. Antennas and Propagation, *AP-16,* No. 1 (January 1968), pp. 136–138.
25. Lee, S. W., "Radiation from an Infinite Array of Parallel-Plate Waveguides with Thick Walls," IEEE Trans. Microwave Theory and Techniques, *MTT-15,* No. 6 (June 1967), pp. 364–371.
26. Galindo, V. and Wu, C. P., "Asymptotic Behavior of the Coupling Coefficients for an Infinite Array of Thin-Walled Rectangular Waveguides," IEEE Trans. Antennas and Propagation, *AP-14,* No. 2 (March 1966), pp. 248–249.
27. Galindo, V. and Wu, C. P., "On the Asymptotic Decay of Coupling for Infinite Phased Arrays," Int. Sci. Radio Union, Washington, D. C., Spring 1966.
28. Kahn, W. K., "Ideal Efficiency of a Radiating Element in an Infinite Array," IEEE Trans. Antennas and Propagation, *AP-15,* No. 4 (July 1967), pp. 534–542.
29. Borgiotti, G. V., "Impedance and Gain of a Dipole in an Infinite Periodic Phased Array," Research Triangle Institute, Durham, N. C., Technical Report TRR-25 (March 4, 1966).
30. Dufort, E. C., "Finite Scattering Matrix for an Infinite Antenna Array," Radio Science, *2,* No. 1 (January 1967), pp. 19–27.
31. Blass, J. and Rabinowitz, S. J., "Mutual Coupling in Two-Dimensional Arrays," IRE Wescon Convention Record, *1,* pt. 1 (1957), pp. 134–139.
32. von Aulock, W. H., "Properties of Phased Arrays," Proceedings IRE, *48,* No. 10 (October 1960), pp. 1715–1727.
33. Amitay, N. and Galindo, V., "On Energy Conservation and the Method of Moments in Scattering Problems," to be published.
34. Amitay, N. and Galindo, V., "Characteristics of Dielectric Covered and Loaded Circular Waveguide Phased Arrays," Proc. Int. Sci. Radio Union, Boston, Mass., September 1968.

The Effects of Strain on Electromagnetic Modes of Anisotropic Dielectric Waveguides at p-n Junctions

By JAMES McKENNA and J. A. MORRISON

(Manuscript received May 22, 1968)

A first order perturbation expansion is carried out in order to analyze the effect of small spatially uniform strains on the lowest order (even) TE and TM modes in an anisotropic dielectric waveguide. This generalizes the results of an earlier paper in which the effects of certain special cases of uniform strain were calculated. Unlike in these special cases, the perturbed modes are, in general, neither purely TE or TM, and one effect of two of the offdiagonal components of the strain is to tilt the plane of polarization and change the relative phase of the two polarizations. To first order, the modes are not exponentially attenuated. Some numerical examples are considered in order to illustrate the results. It is found that, under appropriate conditions, the effect of the small strain may be quite large in relation to its magnitude.

I. INTRODUCTION

The concept of a multilayered dielectric waveguide is central to the theory of the GaP electro-optic diode modulator.[1-9] As part of a detailed study of the properties of electro-optic diode modulators, Nelson and McKenna[4] have investigated the possible discrete modes which can propagate in a number of such waveguides and have calculated the detailed properties of the lowest order mode of each polarization.

In the fabrication of a p-n junction a certain amount of strain is always introduced. Because of the photoelastic effect[10] this strain will induce a change in the dielectric matrix describing the unstrained p-n junction. In general the strain will be spatially nonuniform, making it extremely difficult to calculate modes in such a structure. However, a knowledge of the effect of a spatially uniform strain on the mode

structure would provide insight into the effects of nonuniform strain. The effects of certain special cases of uniform strain on the modes of a simple model of a dielectric waveguide were calculated in Ref. 4. In the present paper we complete this investigation and calculate the modes in the same model dielectric waveguide when subjected to an arbitrary uniform strain. We use first order perturbation theory in a small parameter describing the magnitude of the strain.

Although the work presented in this paper was motivated by research on the theory of the electro-optic diode modulator, the results have considerable relevance to the theory of the GaAs injection laser. Here too, various dielectric waveguide models have been used to explain the light containment.[11–14] The same problems of strain exist, and the results of this paper give a qualitative picture of the effects of strain on modal structure. The effect of strain on completely different types of electro-optic light modulators have been studied by Kaminow[15] and by DiDomenico and Anderson.[16]

II. FORMULATION OF THE PROBLEM, AND RESULTS

In Ref. 4 the symmetric step model was used to study the effects of strain. This very simple model exhibits many of the main features of interest in dielectric waveguide models.

The model consists of an anisotropic crystalline slab bounded by the planes $x = \pm w$, whose refractive index is raised uniformly by some constant amount, the physical origin of which is still obscure, and which is embedded in an isotropic medium of relatively lower index of refraction. The central slab represents the junction region whose anisotropy is caused by the junction field $\mathbf{E}_J$ acting through the electro-optic effect.[9] The direction of the x-axis is always taken parallel to $\mathbf{E}_J$. The isotropic medium represents the normal GaP.

The model is determined by its dielectric matrix, which in the absence of strain and for certain orientations of $\mathbf{E}_J$ with respect to the crystal axes can be diagonal in a coordinate system having its x-axis parallel to $\mathbf{E}_J$. For such orientations of $\mathbf{E}_J$, the diagonal matrix elements of the dielectric matrix $K_\alpha^{(0)}(x)$, $\alpha = 1, 2, 3$, depend only on x. (We use x, y, z for the coordinates rather than x_1, x_2, x_3.) The matrix elements in the absence of strain are then defined by the equations

$$K_\alpha^{(0)}(x) = K_\alpha , \qquad |x| < w \tag{1}$$

$$K_\alpha^{(0)}(x) = K_0 , \qquad |x| > w \tag{2}$$

where $\alpha = 1, 2, 3$ (see Fig. 1).

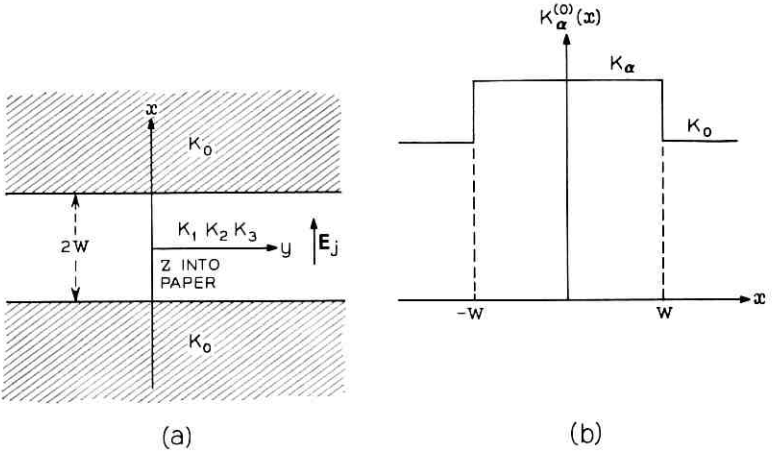


Fig. 1 — (a) The coordinate system used in the symmetric step model. (b) A graph of $K_\alpha(x)$.

There are two orientations of $\mathbf{E}_J$ of particular interest which allow us to diagonalize the dielectric matrix in the desired coordinate system. If $\mathbf{E}_J$ is in the [111] direction, then the x, y, and z axes can be taken in the [111], [$\bar{1}$10], and [$\bar{1}\bar{1}$2] directions, while if $\mathbf{E}_J$ is in the [100] direction, the x, y, and z axes can be taken in the [100], [01$\bar{1}$], and [011] directions. The set of axes determined by the unstrained model will be used in all the strain calculations and the dielectric matrix will always be referred to these axes. See Ref. 4 for further details of the model.

In the presence of a uniform strain, the dielectric matrix is in general no longer diagonal, and we can write for the dielectric matrix elements $K_{\alpha\beta}(x)$,

$$K_{\alpha\alpha}(x) = K_\alpha^{(0)}(x) + \eta S_{\alpha\alpha}, \qquad \alpha = 1, 2, 3 \tag{3}$$

$$K_{\alpha\beta}(x) = \eta S_{\alpha\beta}, \qquad \alpha \neq \beta. \tag{4}$$

The symmetric matrix $(\eta S_{\alpha\beta})$ is the contribution of the photoelastic effect[10] which we have written in this form for convenience in the perturbation analysis. The matrix elements $S_{\alpha\beta}$ are spatially constant. We assume that $n^{-2}S_{\alpha\beta}$ is of order unity, where n is the index of refraction of GaP and η is a small parameter. In Section II we express $\eta S_{\alpha\beta}$ in terms of the strain matrix and give estimates for the size of η.

We now seek solutions of the Maxwell curl equations

$$\nabla \times \mathbf{E} = -\mu_0 \dot{\mathbf{H}}, \tag{5}$$

$$\nabla \times \mathbf{H} = \epsilon_0 \mathbf{K}(x) \cdot \dot{\mathbf{E}}, \tag{6}$$

of the form

$$\mathbf{E} = \mathbf{e}(x) \exp i(\omega t - \beta k z), \tag{7}$$

$$\mathbf{H} = \mathbf{h}(x) \exp i(\omega t - \beta k z). \tag{8}$$

These solutions correspond to waves travelling in the positive z direction, where $k = \omega(\mu_0 \epsilon_0)^{\frac{1}{2}} = 2\pi/\lambda$ is the free space wave number and λ the free space wavelength of the light.

In the strain free cases ($\eta = 0$), there are both TE and TM modes and these modes can be either even or odd functions of x. At most only a finite number of modes can exist, and Ref. 4 shows that for the typical parameter values encountered in GaP diode modulators only the lowest order even TE and TM modes can exist. For that reason we confine ourselves here to solutions which in the limit of zero strain ($\eta = 0$) reduce to even modes. However, the perturbation technique used here applies equally well to solutions which in the limit $\eta = 0$ reduce to odd modes.

When $\eta \neq 0$, we seek solutions of Maxwell's equations of the form

$$e_\alpha(x) = A_\alpha \exp -kp(x - w) + B_\alpha \exp -kq(x - w), \qquad x \geqq w \tag{9}$$

$$= C_\alpha \exp kr(x + w) + D_\alpha \exp ks(x + w), \qquad x \leqq -w \tag{10}$$

$$= F_\alpha \exp ikfx + G_\alpha \exp -ikgx + L_\alpha \exp iklx + M_\alpha \exp -ikmx, \qquad |x| \leqq w \tag{11}$$

for $\alpha = 1, 2, 3$. The general solution in each region is a sum of four linearly independent solutions, but in the regions $|x| > w$, the boundary conditions at infinity eliminate two of these solutions. The expressions for $h_\alpha(x)$ can be obtained from equation (5). The various coefficients and parameters $A_\alpha \ldots, p, \ldots$ can be expanded in powers of η

$$A_\alpha = A_\alpha^{(0)} + \eta A_\alpha^{(1)} + \cdots, \tag{12}$$

$$p = p_0 + \eta p_1 + \cdots, \quad \text{and so on.} \tag{13}$$

In Section III we list the terms in these expansions of order zero and one in η, and in Section IV we outline their derivation. In this section we merely discuss some of the features of the solutions.

We refer to solutions which in the limit as $\eta \to 0$ reduce to even TE modes as "perturbed TE modes"; similarly, we refer to "perturbed

TM modes." Expressions for the unperturbed TE and TM modes ($\eta = 0$) are given in equations (25) through (32).

Although the expressions we have obtained for the coefficients and parameters are quite complicated, several features of the perturbed modes stand out. If $S_{12} \neq 0$ or $S_{23} \neq 0$, then the modes cannot be purely TE or TM. In general an important effect of the strain is to tilt the plane of polarization. This tilt is in general a function of x but not of z. Since the coefficients A_α, ... are complex in general, the relative phase between E_y in the perturbed TE mode and E_x in the perturbed TM mode is a function of x. This relative phase at $x = 0$ cannot be determined unless the method of excitation is known, since all the A_α cannot be determined, as is shown in Section IV. Ref. 4 considers the special case where $S_{12} = S_{23} = 0$, $S_{13} \neq 0$ and shows that the modes are rigorously TE or TM. That paper calculates only such parameters as β and p, not such coefficients as A_α and B_α. The parameters are expanded in two small quantities δ and Δ describing the unstrained dielectric matrix. If we expand the expressions for the parameters in this paper to first order in the same small quantities δ and Δ (to second order for β) complete agreement is obtained with the Ref. 4 results.

In the absence of strain, the surfaces of constant phase for both TE and TM modes are the planes $z =$ constant. However, in the presence of strain, the surfaces of constant phase are no longer planes, and are different for the perturbed TE and TM modes.[4]

Finally, $\beta_0 + \eta\beta_1$ is real in all cases. Thus at least to first order in η the modes experience no exponential attenuation as they propagate.

In order to get some feel for the magnitude of the effects involved, we consider several numerical examples. We first estimate the order of magnitude of η by relating it to observable phase differences. Consider a plane wave whose free space wavelength is λ propagating over a distance l in a medium of index of refraction $n + \Delta n$. The phase difference $\Delta\varphi$ which this wave would experience over the same wave if the index of refraction were n is

$$\Delta\varphi = \frac{2\pi}{\lambda} l(\Delta n). \tag{14}$$

If ηn^2 is the photoelastic contribution to the dielectric constant, then

$$\Delta n = (n^2 + \eta n^2)^{\frac{1}{2}} - n \cong \tfrac{1}{2} n\eta. \tag{15}$$

Therefore, we have

$$\eta \cong \lambda\Delta\varphi/nl\pi. \tag{16}$$

Now the upper limit of phase shifts observed[17] in GaP at $\lambda = 6328$ Å over a length $l = 0.6$ mm is about $\pi/4$. Taking[18] $n = 3.31$ this yields $\eta \cong 0.8 \times 10^{-4}$. This probably represents an extreme upper limit, and so we will assume $\eta = 10^{-6}$ in our examples. Recent X-ray measurements[19] of strain in P doped Si yield a value of η of about 10^{-6} when the concentration of the dope is $N_D \approx 10^{18}$ atoms per cubic centimeter.[9]

It can be shown that the matrix $(\eta S_{\alpha\beta})$ is approximately related to the strain matrix $(\epsilon_{\alpha\beta})$ by the equations[10]

$$\eta S_{\alpha\beta} = -n^4 \sum_{\mu,\nu=1}^{3} P_{\alpha\beta\mu\nu}\epsilon_{\mu\nu} \tag{17}$$

where n is the index of refraction and $P_{\alpha\beta\mu\nu}$ are the elasto-optical coefficients. Crystals of class $\bar{4}3m$ have only three different elasto-optic coefficients when referred to the crystal axes. (See p. 251 of Ref. 10.) For GaP these are[20]

$$P_{11} = -0.151, \qquad P_{44} = -0.074, \qquad P_{12} = -0.082. \tag{18}$$

Since[18] $n = 3.31$ for GaP, and $n^{-2}S_{\alpha\beta}$ is at most of order one, it follows that the magnitude of the strain is roughly proportional to η. In order to obtain the values of the elasto-optic coefficients in the coordinate system used in this paper, it is necessary to make a transformation of the elasto-optic tensor from its representation in the crystal axes. We will not do that here; rather we take $(\eta S_{\alpha\beta})$ as given. In Table I we define three possible strain contributions to the dielectric matrix, labelled a, b and c. Matrices a and b were chosen to demonstrate the effect of the off-diagonal elements S_{12} and S_{23}, respectively (Ref. 4 considered the effect of S_{13} alone), while c was chosen to demonstrate a possible effect when all the off-diagonal elements are nonzero.

For a GaP diode modulator we can write[2, 4]

$$K_\alpha = n^2(1 - \delta_\alpha), \qquad \alpha = 0, 1, 2, 3, \tag{19}$$

where $n = 3.31$ is the index of refraction of GaP[17] and the quantities δ_α, $\alpha = 1, 2, 3$ are functions of the applied bias voltage V. In the original

TABLE I—STRAIN CONTRIBUTION TO THE DIELECTRIC MATRIX*

Type	η	S_{11}	S_{22}	S_{33}	S_{12}	S_{23}	S_{13}
a	10^{-6}	0	0	0	10	0	0
b	10^{-6}	0	0	0	0	10	0
c	10^{-6}	7.07	7.07	7.07	7.07	7.07	7.07

* Components of the strain contribution, S_{ij}, and the magnitude parameter η.

symmetric step model δ_0 is independent of V. For $\mathbf{E}_J \,||\, [111]$ Ref. 4 showed that

$$\delta_1 = -2\delta, \qquad \delta_2 = \delta_3 = \delta, \tag{20}$$

where

$$\delta = -\bar{E}_J r_{41} n^2/(3)^{\frac{1}{2}}, \tag{21}$$

and where $\bar{E}_J$ is the (spatial) average junction field, r_{41} is the electro-optic coefficient, and n is the index of refraction. For a typical diode (diode KC46CA of Ref. 9), $\bar{E}_J$(measured in V/m) is related to the diode half width w(measured in m) and bias voltage V(measured in V) by[9]

$$\bar{E}_J = (2 - V)/(2w). \tag{22}$$

The half width can be determined by capacitance measurements[8, 9] and related to the bias voltage by

$$w(V) = 0.139 \times 10^{-6}(1 - V/1.8)^{0.380}. \tag{23}$$

Using the value $r_{41} = -0.86 \times 10^{-12}$ m/V,* we can now calculate δ_1, δ_2, and δ_3 as functions of V.

For this diode, $\delta_0 = 1.612 \times 10^{-3}$. However, it has been shown that the voltage dependence of the parameters of the symmetric step model is not correct, and the double walled waveguide much more closely describes the true voltage dependence.[4, 9] We have not used the double walled model because it is analytically complex. Instead, since the modes in the single and double walled guides are very similar in form because they both decay exponentially as functions of x outside the guide, we have used the single walled model but simulated the voltage dependence of the double walled model. This has been achieved by letting δ_0 vary with voltage. The voltage variation of δ_0 has been obtained by requiring the equality of expressions (2.33) and (3.18) in Ref. 4 for the decay constants p, and letting $w_1 = w(0)$ and $w_2 = w(V)$. This yields the relation

$$\delta_0 = (2.24 \times 10^{-10})/w. \tag{24}$$

In Table II we list these basic constants describing the unstrained diode as functions of V. Using these values, we can calculate from equations (33) through (37) the parameters of the unstrained TE

* This is the unclamped value of r_{41} given in Ref. 18. After these calculations were made it was determined that the clamped value $r_{41} = -0.97 \times 10^{-12}$ m/V should be used. However, since our results supply only qualitative information about actual diodes, we have not redone the numerical example.

TABLE II—CHARACTERISTICS FOR A TYPICAL GaP DIODE*

Bias voltage (V)	w (10^{-5} cm)	$10^4\ \delta_1$	$10^4\ \delta_2$	$10^4\ \delta_3$	$10^4\ \delta_0$
−2	1.87	−1.17	0.58	0.58	12.00
−12	3.04	−2.51	1.25	1.25	7.38
−24	3.85	−3.67	1.84	1.84	5.82

* Given as functions of the applied reverse bias voltage V. The half width of the junction is w, and the components of the unstrained dielectric matrix are $K_j = n^2(1 - \delta_j)$, $j = 0, 1, 2, 3$, where $n = 3.31$ is the index of refraction of GaP.

and TM modes for $\lambda = 6328$ Å. These values are listed in Table III. Finally, in Tables IV and V we list the parameters of the corresponding perturbed TE and TM modes respectively. The accuracy of those terms less than 10^{-4} is uncertain in case c of Tables IV and V. In Figs. 2 through 7 we plot some of the components of the perturbed TE modes correct to first order in η. In Figs. 2, 3, 5, 6, and 7 the imaginary part of the component is negligible and is neglected, while in Fig. 4 the real part is negligible with respect to the imaginary part and is neglected. In all cases the e_3 component is negligible compared to the e_1 component. We have chosen the undetermined coefficients so that at $x = 0$, $z = 0$, e_2 in the perturbed TE mode and e_1 in the perturbed TM mode have zero phase to first order in η.

This example illustrates how much tilting of the plane of polarization, or coupling of the TE and TM modes, is to be expected. The S_{12} component produces the main effect, which from Figs. 2 and 3, is a maximum tilt of the plane of polarization of 3.5°. This effect

TABLE III—UNPERTURBED MODE PARAMETERS*

Type of mode	Bias voltage (V)	β_0	p_0	f_0	l_0
TE	−2	3.308	0.0226	0.1096	0.1180†
TE	−12	3.309	0.0195	0.0796	0.1022†
TE	−24	3.309	0.0160	0.0641	0.1007†
TM	−2	3.308	0.0259	0.1089†	0.1174
TM	−12	3.309	0.0307	0.0759†	0.0994
TM	−24	3.309	0.0363	0.0552†	0.0953

* Describing the unstrained TE modes, and the parameters β_0, p_0 and l_0 describing the unstrained TM modes as functions of the applied reverse bias voltage V. The wavelength of the light is 6328 Å.

† Derived parameters l_0 for the TE modes and f_0 for the TM modes. These derived parameters appear only in first and higher order corrections to the field.

TABLE IV—PARAMETERS FOR PERTURBED TE MODES*

Type of strain	Bias voltage (V)	β_1	p_1		q_1		l_1	m_1
			$Re(p_1)$	$Im(p_1)$	$Re(q_1)$	$Im(q_1)$		
a	−2	0	221		−221		0	0
	−12	0	257		−257		0	0
	−24	0	312		−312		0	0
b	−2	0		1.51		−1.51	0	0
	−12	0		1.51		−1.51	0	0
	−24	0		1.51		−1.51	0	0
c	−2	1.07	156	-0.3×10^{-7}	−156	−2.14	2.13	-0.4×10^{-8}
	−12	1.07	181	$+0.2\times10^{-7}$	−181	−2.14	2.13	-0.4×10^{-8}
	−24	1.07	221	-0.2×10^{-7}	−221	−2.14	2.13	-0.4×10^{-8}

* For all perturbed TE modes $f_1 = g_1 = 0$.

decreases with increasing reverse bias voltage. However, it should be noticed from Figs. 4 and 5 that the coupling effect resulting from S_{23} increases with reverse bias voltage. The e_1 component is roughly proportional to η, so a doubling of the strain would double the mode coupling. Mathematically, the existence of this relatively large effect results from the largeness of the factor c given in equation (64) for perturbed TE modes and in equation (84) for perturbed TM modes. The TM modes exhibit a similar behavior.

From Tables IV and V we see that the changes in the parameters, ηp_1, ηq_1, $\eta\beta_1$, and so on, are indeed small, which gives us confidence that the perturbation treatment is reasonable.

III. FORMULAS FOR THE SOLUTIONS

To list the formulas for the coefficients and parameters, $A_a, \ldots, p, \ldots$ (which appear in the expressions (7) through (11) for the solutions in terms of the various parameters describing the symmetric step model and the strain matrix), we begin by writing down the solution for the strain free ($\eta = 0$) case for both the even TE and TM modes. When $\eta = 0$, we have for the even TE modes

$$e_1(x) = e_3(x) = 0, \qquad \text{all } x \tag{25}$$

$$e_2(x) = \cos(kf_0x), \qquad |x| \leqq w \tag{26}$$

$$= \cos(kf_0w)\exp kp_0(w - |x|), \qquad |x| \geqq w \tag{27}$$

while for the even TM modes

TABLE V—PARAMETERS FOR THE PERTURBED TM MODES

Type of strain	Bias voltage (V)	β_1	p_1		q_1		f_1	g_1	l_1	m_1
			$Re(p_1)$	$Im(p_1)$	$Re(q_1)$	$Im(q_1)$				
a	−2	0	193		−193		0	0	0	0
	−12	0	163		−163		0	0	0	0
	−24	0	138		−138		0	0	0	0
b	−2	0		1.51		−1.51	0	0	0	0
	−12	0		1.51		−1.51	0	0	0	0
	−24	0		1.51		−1.51	0	0	0	0
c	−2	1.07	136	-0.2×10^{-7}	−136	−2.13	-1.2×10^{-4}	-1.2×10^{-4}	2.13	−2.13
	−12	1.07	115	-0.2×10^{-7}	−115	−2.14	1.9×10^{-4}	1.9×10^{-4}	2.13	−2.13
	−24	1.07	97.3	-0.3×10^{-7}	−97.3	−2.14	0.72×10^{-4}	0.72×10^{-4}	2.13	−2.13

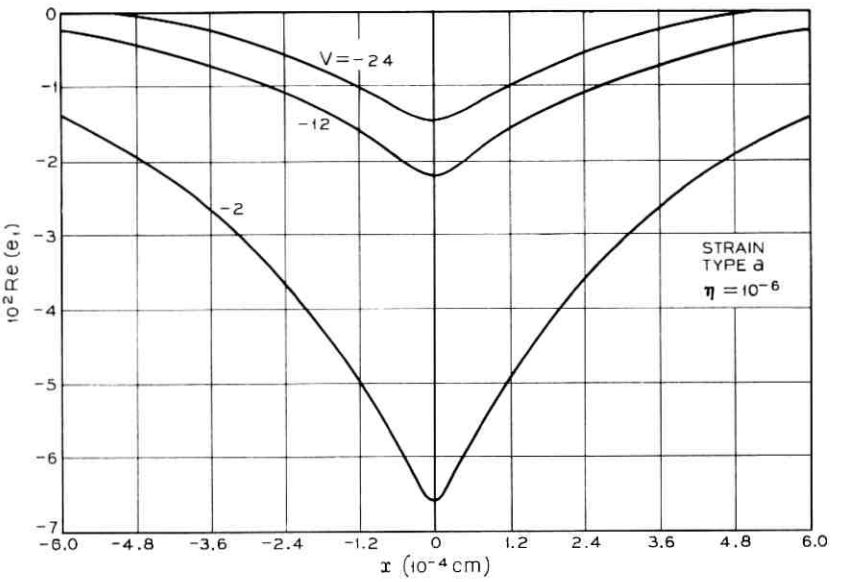


Fig. 2 — The relative amplitude of the real part of e_1 for the perturbed TE mode.

$$e_1(x) = \cos(kl_0x), \qquad |x| \leqq w \tag{28}$$

$$= \frac{K_1}{K_0}\cos(kl_0w)\exp kp_0(w - |x|), \qquad |x| \geq w \tag{29}$$

$$e_2(x) = 0, \qquad \text{all } x \tag{30}$$

$$e_3(x) = i\frac{l_0K_1}{\beta_0K_3}\sin(kl_0x), \qquad |x| \leqq w \tag{31}$$

$$= i\frac{p_0K_1}{\beta_0K_0}\cos(kl_0w)\,\mathrm{sgn}(x)\exp kp_0(w - |x|), \qquad |x| \geqq w. \tag{32}$$

The parameters in these equations are given for the TE modes by the positive roots of the system of equations for p_0, β_0, and f_0

$$p_0^2 = \beta_0^2 - K_0, \tag{33}$$

$$f_0^2 = K_2 - \beta_0^2, \tag{34}$$

$$f_0\tan(kwf_0) = p_0 \tag{35}$$

while for the TM modes the parameters are the positive roots of the system of equations for p_0, β_0, and l_0, consisting of equation (33) and

$$l_0^2 = K_3 - (K_3/K_1)\beta_0^2, \tag{36}$$

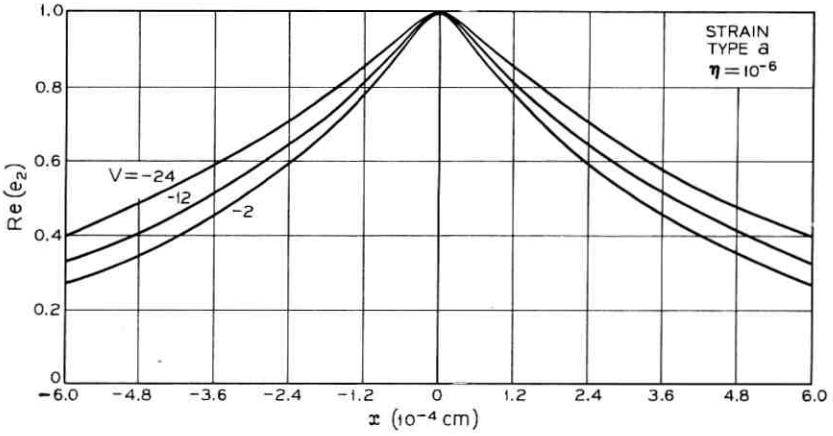


Fig. 3 — The relative amplitude of the real part of e_2 for the perturbed TE mode.

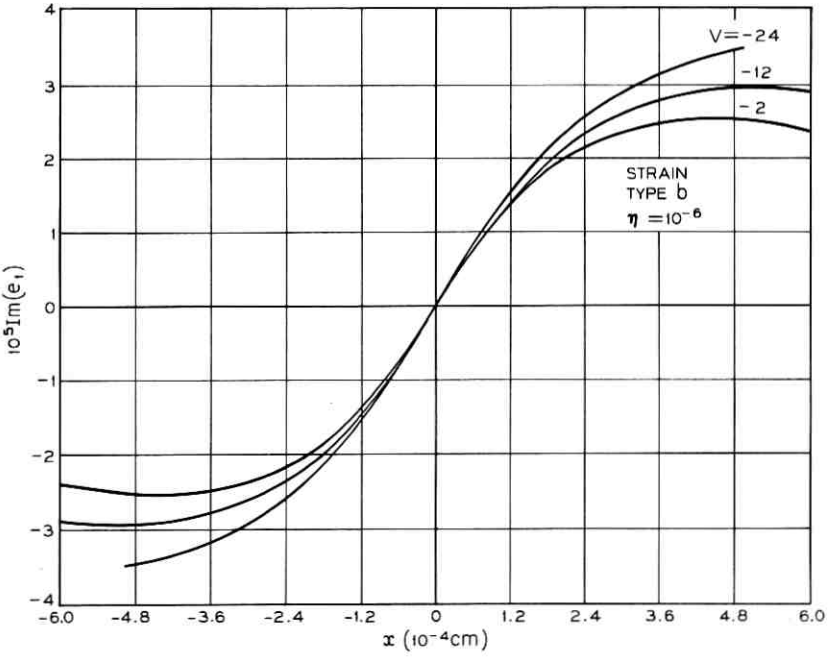


Fig. 4 — The relative amplitude of the imaginary part of e_1 for the perturbed TE mode.

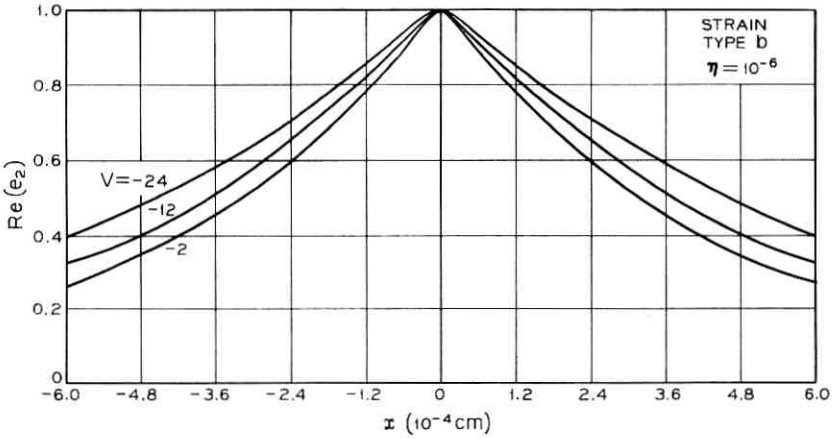


Fig. 5 — The relative amplitude of the real part of e_2 for the perturbed TE mode.

$$K_0 l_0 \tan(kwl_0) = K_3 p_0. \tag{37}$$

The expressions for $\mathbf{h}(x)$ can be obtained from equation (5).

We now turn to listing the formulas for the coefficients and parameters of the solutions for the perturbed TE and TM modes. For both the perturbed TE and TM modes we have the relations

$$\begin{aligned} \begin{matrix} p_1 \\ q_1 \end{matrix} = {} & \left(\frac{\beta_0}{p_0}\right)\beta_1 - \left(\frac{1}{4p_0K_0}\right)[K_0(S_{22} + S_{33}) + \beta_0^2(S_{11} - S_{33}) + 2ip_0\beta_0 S_{13}] \\ & \pm \left(\frac{1}{4p_0K_0}\right)\{[K_0(S_{22} - S_{33}) - \beta_0^2(S_{11} - S_{33}) - 2ip_0\beta_0 S_{13}]^2 \\ & + 4K_0[\beta_0 S_{12} + ip_0 S_{23}]^2\}^{\frac{1}{2}}, \end{aligned} \tag{38}$$

where p_1 corresponds to the "+" sign and q_1 to the "−" sign. It is also true, at least to first order in η, that

$$e_\alpha(-x) = e_\alpha(x)^*, \qquad \alpha = 1, 2, 3, \tag{39}$$

hence we only list those parameters determining the solution in $x \geqq -w$.

For the perturbed TE modes p_0, f_0, and β_0 are the positive solutions of the system of equations (33) through (35), and l_0 is then given in terms of β_0 as the positive root in equation (36). The remaining parameters are

$$q_0 = p_0 , \tag{40}$$

$$g_0 = f_0 , \tag{41}$$

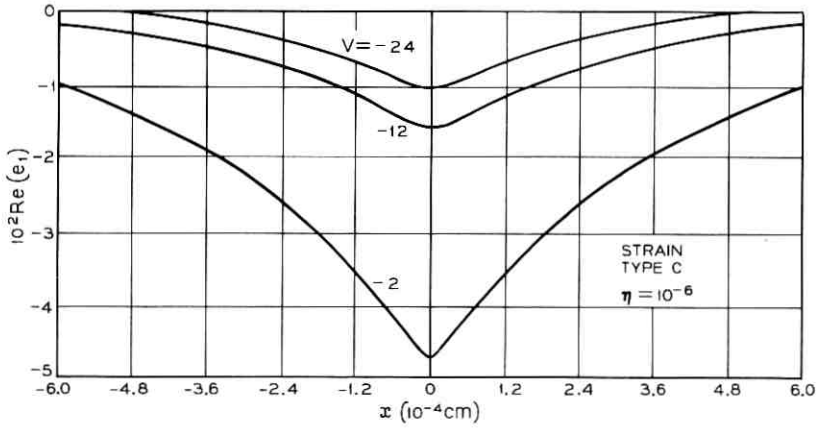


Fig. 6 — The relative amplitude of the real part of e_1 for the perturbed TE mode.

$$m_0 = l_0 , \tag{42}$$

$$\beta_1 = S_{22}/(2\beta_0). \tag{43}$$

The quantities p_1 and q_1 are now determined by equations (38) and (43). Next we have

$$f_1 = g_1 = (S_{22} - 2\beta_0\beta_1)/(2f_0), \tag{44}$$

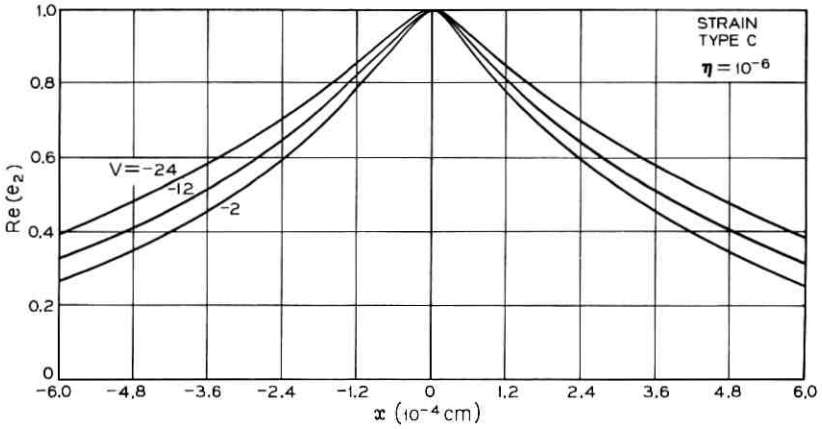


Fig. 7 — The relative amplitude of the real part of e_2 for the perturbed TE mode.

$$\begin{matrix} l_1 \\ m_1 \end{matrix} = \frac{1}{2K_1 l_0} [S_{11}(K_3 - l_0^2) - 2K_3\beta_0\beta_1 + S_{33}(K_1 - \beta_0^2) \pm 2S_{13}l_0\beta_0], \tag{45}$$

where l_1 corresponds to the "+" sign and m_1 to the "−" sign, and in equations (44) and (45) β_1 is given in equation (43). Notice that from (43) and (44), $f_1 = g_1 = 0$ for the TE case.

The expressions for the coefficients are

$$A_1^{(0)} = -\beta_0(\beta_0 S_{12} + ip_0 S_{23}) \cos(kf_0 w)/[2K_0 p_0(p_1 - q_1)], \tag{46}$$

$$A_2^{(0)} = -q_1 \cos(kf_0 w)/(p_1 - q_1), \tag{47}$$

$$A_3^{(0)} = i(p_0/\beta_0)A_1^{(0)}, \tag{48}$$

$$B_1^{(0)} = -A_1^{(0)}, \tag{49}$$

$$B_2^{(0)} = p_1 \cos(kf_0 w)/(p_1 - q_1), \tag{50}$$

$$B_3^{(0)} = i(p_0/\beta_0)B_1^{(0)}, \tag{51}$$

$$F_\alpha^{(0)} = G_\alpha^{(0)} = 0, \qquad \alpha = 1, 3 \tag{52}$$

$$F_2^{(0)} = G_2^{(0)} = \tfrac{1}{2} \tag{53}$$

$$L_\alpha^{(0)} = M_\alpha^{(0)} = 0, \qquad \alpha = 1, 2, 3 \tag{54}$$

$$\begin{matrix} F_1^{(1)} \\ G_1^{(1)} \end{matrix} = \tfrac{1}{2}c [\mp S_{23}\beta_0 f_0 - S_{12}(K_3 - f_0^2)], \tag{55}$$

$$F_2^{(1)} = G_2^{(1)}, \tag{56}$$

$$\begin{matrix} F_3^{(1)} \\ G_3^{(1)} \end{matrix} = \tfrac{1}{2}c [\mp S_{12}\beta_0 f_0 - S_{13}(K_1 - \beta_0^2)], \tag{57}$$

$$\begin{matrix} L_3^{(1)} \\ M_3^{(1)} \end{matrix} = \tfrac{1}{4}l_0 \cos(kwf_0) \{aS_{23}[2\beta_0^2(K_1 - K_0)c - 1] \pm \frac{b\beta_0}{p_0} S_{12}[2p_0^2(K_0 - K_3)c - 1]\}, \tag{58}$$

$$L_1^{(1)} = (\beta_0 K_3/l_0 K_1)L_3^{(1)}, \tag{59}$$

$$M_1^{(1)} = -(\beta_0 K_3/l_0 K_1)M_3^{(1)}, \tag{60}$$

$$L_2^{(1)} = M_2^{(1)} = 0, \tag{61}$$

where

$$a = [l_0K_0 \cos (kwl_0) + p_0K_3 \sin (kwl_0)]^{-1}, \tag{62}$$

$$b = [l_0K_0 \sin (kwl_0) - p_0K_3 \cos (kwl_0)]^{-1}, \tag{63}$$

$$c = (K_1K_3 - K_3\beta_0^2 - K_1f_0^2)^{-1}, \tag{64}$$

and p_1 and q_1 are the values appropriate to the TE modes.

Finally, we write down the three combinations

$$A_1^{(1)} + B_1^{(1)} = -i(\beta_0/p_0)[A_3^{(1)} + B_3^{(1)}] - \cos (kf_0w)[(p_0^2 - K_0)S_{12} + i\beta_0p_0S_{23}]/(2K_0p_0^2), \tag{65}$$

$$A_2^{(1)} + B_2^{(1)} = \cos (kf_0w)[F_2^{(1)} + G_2^{(1)}], \tag{66}$$

$$A_3^{(1)} + B_3^{(1)} = [F_3^{(1)} + G_3^{(1)}] \cos (kf_0w) + i[F_3^{(1)} - G_3^{(1)}] \sin (kf_0w) + [L_3^{(1)} + M_3^{(1)}] \cos (kl_0w) + i[L_3^{(1)} - M_3^{(1)}] \sin (kl_0w). \tag{67}$$

The coefficients $F_2^{(1)} = G_2^{(1)}$, and hence $A_2^{(1)} + B_2^{(1)}$, are arbitrary and correspond to an overall multiplicative constant. They can be set equal to zero with no loss in generality. We discuss this point further in Section III. Moreover, the individual coefficients $A_1^{(1)}$, $A_3^{(1)}$, $B_1^{(1)}$, and $B_3^{(1)}$ cannot be determined at this stage. However, the terms we have are sufficient to determine each component of the field up through order one in η.

For the perturbed TM modes p_0, l_0, and β_0 are the positive solutions of the system of equations (33), (36), and (37), and f_0 is given in terms of β_0 as the positive root in equation (34). The parameters q_0 and p_0 are still related by equation (40), m_0 and l_0 by equation (42), and g_0 and f_0 by equation (41). The remaining parameters are

$$\beta_1 = [1/(2\beta_0)][K_0(K_1l_0^2 + K_3p_0^2) + \zeta]^{-1} \cdot \{[S_{11}\beta_0^2/K_1][K_0K_3p_0^2 + K_1^2l_0^2 + \zeta] + [K_1S_{33}l_0^2/K_3^2][K_3(K_3 - K_0)p_0^2 + \zeta]\}, \tag{68}$$

where

$$\zeta = (kp_0w)(K_3^2p_0^2 + K_0^2l_0^2). \tag{69}$$

With the aid of (68) and (69) for β_1, p_1 and q_1 are determined by equation (38), f_1 and g_1 by equation (44), and l_1 and m_1 by equation (45).

The expressions for the coefficients are

$$A_1^{(0)} = -K_1 \cos(kwl_0)(2\beta_0\beta_1 - 2p_0p_1 - S_{22})/[2p_0K_0(p_1 - q_1)], \quad (70)$$

$$A_2^{(0)} = -K_1 \cos(kwl_0)(\beta_0 S_{12} + ip_0S_{23})/[2p_0\beta_0K_0(p_1 - q_1)], \quad (71)$$

$$B_1^{(0)} = K_1 \cos(kwl_0)(2\beta_0\beta_1 - 2p_0q_1 - S_{22})/[2p_0K_0(p_1 - q_1)], \quad (72)$$

$$B_2^{(0)} = -A_2^{(0)}, \quad (73)$$

where β_1 is given by equations (68) and (69), p_1 and q_1 are the values appropriate to the TM modes, and $A_3^{(0)}$ is related to $A_1^{(0)}$ by equation (48) and $B_3^{(0)}$ to $B_1^{(0)}$ by equation (51). Furthermore,

$$F_\alpha^{(0)} = G_\alpha^{(0)} = 0, \qquad \alpha = 1, 2, 3 \quad (74)$$

$$L_1^{(0)} = M_1^{(0)} = \tfrac{1}{2}, \quad (75)$$

$$L_2^{(0)} = M_2^{(0)} = 0, \quad (76)$$

$$L_3^{(0)} = -M_3^{(0)} = (l_0K_1)/(2\beta_0K_3), \quad (77)$$

$$F_1^{(1)} = F_3^{(1)} = G_1^{(1)} = G_3^{(1)} = 0, \quad (78)$$

$$\begin{matrix} F_2^{(1)} \\ G_2^{(1)} \end{matrix} = \cos(kwl_0)\{\beta_0 aS_{12}[K_1 + 2cp_0^2(K_0 - K_3)] \pm p_0bS_{23}[K_1 + 2c\beta_0^2(K_1 - K_0)]\}/(4p_0\beta_0K_0), \quad (79)$$

$$\begin{matrix} L_2^{(1)} \\ M_2^{(1)} \end{matrix} = -\tfrac{1}{2}c[S_{12} \pm l_0K_1S_{23}/(\beta_0K_3)] \quad (80)$$

$$A_2^{(1)} + B_2^{(1)} = [F_2^{(1)} + G_2^{(1)}] \cos(kwf_0) + i[F_2^{(1)} - G_2^{(1)}] \sin(kwf_0) + [L_2^{(1)} + M_2^{(1)}] \cos(kwl_0) + i[L_2^{(1)} - M_2^{(1)}] \sin(kwl_0), \quad (81)$$

where

$$a = [p_0 \cos(kwf_0) - f_0 \sin(kwf_0)]^{-1}, \quad (82)$$

$$b = [f_0 \cos(kwf_0) + p_0 \sin(kwf_0)]^{-1}, \quad (83)$$

$$c = (K_2 - l_0^2 - \beta_0^2)^{-1}. \quad (84)$$

Just as in the perturbed TE case, the coefficients cannot all be determined uniquely. We can with no loss of generality set

$$L_3^{(1)} = -M_3^{(1)} = 0. \quad (85)$$

Once this choice is made, we have

$$L_1^{(1)} = [S_{13} - l_0\beta_1 - l_1\beta_0 + (S_{33} - 2l_0l_1)(l_0K_1/\beta_0K_3)]/(2\beta_0l_0), \quad (86)$$

$$M_1^{(1)} = [-S_{13} - l_0\beta_1 - m_1\beta_0 + (S_{33} - 2l_0m_1)(l_0K_1/\beta_0K_3)]/(2\beta_0l_0), \quad (87)$$

$$A_3^{(1)} + B_3^{(1)} = \{-(S_{13}p_0kw)/K_0 + i[\tfrac{1}{2}kw(K_3\beta_0^2S_{11}/K_1 + K_1l_0^2S_{33}/K_3 - 2K_3\beta_0\beta_1)]/(K_3\beta_0)\} \cos(kwl_0) \quad (88)$$

$$A_1^{(1)} + B_1^{(1)} = -i(\beta_0/p_0)[A_3^{(1)} + B_3^{(1)}] + (K_1 \cos(kwl_0)/K_0\beta_0p_0) \cdot [-K_0\beta_1/p_0 - \beta_0S_{11}(p_0^2 - K_0)/(2p_0K_0) + p_0\beta_0S_{33}/(2K_0) - ip_0^2S_{13}/K_0]. \quad (89)$$

A knowledge of these terms is sufficient to determine each component of the field up through order one in η.

IV. DETAILS OF THE CALCULATIONS

In standard fashion **H** can be eliminated from equations (5) and (6) by taking the curl of equation (5) and by making use of the assumed form of the solutions, equations (7) and (8). There results the system of equations

$$i\beta\frac{de_3}{d\xi} - \beta^2e_1 + \sum_{\alpha=1}^{3} K_{1\alpha}e_\alpha = 0, \quad (90)$$

$$\frac{d^2e_2}{d\xi^2} - \beta^2e_2 + \sum_{\alpha=1}^{3} K_{2\alpha}e_\alpha = 0, \quad (91)$$

$$\frac{d^2e_3}{d\xi^2} + i\beta\frac{de_1}{d\xi} + \sum_{\alpha=1}^{3} K_{3\alpha}e_\alpha = 0, \quad (92)$$

where we have introduced the new independent variable

$$\xi = kx. \quad (93)$$

The standard boundary conditions[21] on **E** and **H** yield the conditions that e_2, e_3, $de_2/d\xi$, and $de_3/d\xi + i\beta e_1$ must be continuous at $\xi = \sigma = kw$.

The general plan of the calculation is first to consider the equations obtained by substituting into equations (90) through (92) the expressions for e_α in the various regions given by equations (9) through (11). From these equations, one can determine up through first order

in η all but one of the parameters and some of the coefficients as functions of the parameter β_1. Upon substituting these values into the boundary condition equations, a set of equations is obtained from which β_1 and some of the remaining coefficients can be determined.

Since the components of the electromagnetic field satisfy a linear, homogeneous system of equations, it follows that if $e_\alpha(x)$, $\alpha = 1, 2, 3$ is a solution set, then so is $(1 + a_1\eta + a_2\eta^2 + \ldots)e_\alpha(x)$, $\alpha = 1, 2, 3$, where the constants a_1, a_2, . . . are arbitrary. For example, if the coefficients A_α, B_α, . . . given by expansions of the form (12) represents a solution, then the coefficients given by expansions of the form

$$A_\alpha = A_\alpha^{(0)} + \eta[a_1 A_\alpha^{(0)} + A_\alpha^{(1)}] + \cdots, \tag{94}$$

with the same a_1 used in each expansion, represent another solution. Thus unless the corresponding zeroth order coefficient is zero, the first order coefficient cannot be uniquely determined. We do, however, have the arbitrary constant a_1 at our disposal. The multiplicative constant $(1 + a_1\eta + \ldots)$ can only be determined from a knowledge of the excitation of the mode.

If the assumed expressions for e_α in $\xi \geqq \sigma$ given by equation (9) are substituted into equations (90) through (92) we get the set of homogeneous, linear equations in A_α, $\alpha = 1, 2, 3$,

$$(K_0 + \eta S_{11} - \beta^2)A_1 + \eta S_{12}A_2 + (\eta S_{13} - i\beta p)A_3 = 0, \tag{95}$$

$$\eta S_{12}A_1 + (K_0 + \eta S_{22} + p^2 - \beta^2)A_2 + \eta S_{23}A_3 = 0, \tag{96}$$

$$(\eta S_{13} - i\beta p)A_1 + \eta S_{23}A_2 + (K_0 + \eta S_{33} + p^2)A_3 = 0, \tag{97}$$

plus a similar set of equations with A_α replaced by B_α and p replaced by q. The condition that these equations have a nontrivial solution, the vanishing of the determinant of coefficients, yields a relation between β and p of the form

$$D(p, \beta) = 0, \tag{98}$$

where $D(p, \beta)$ is a quartic polynomial in p and β. The second set of equations involving the B_α and q yields the same determinantal equation with p replaced by q,

$$D(q, \beta) = 0. \tag{99}$$

That is, q is a second root of the quartic. If p, q, and β are expanded in powers of η as in equation (13), equations (98) and (99) can be expanded in powers of η and the coefficients of the various powers of

η can be equated to zero. The vanishing of the lowest order term yields equation (33), which is satisfied by both p_0 and q_0, thus also yielding equation (40). The vanishing of the first order coefficients shows that p_1 and q_1 are the two roots of a quadratic which are given by equation (38). These results are independent of the TE or TM character of the mode.

Equations (95) through (97) can now be expanded in powers of η by substituting in the expansions of p, β, and A_a. The three similar equations involving q, β, and B_a can be expanded in powers of η in the same way. Because p_0 and β_0 satisfy equation (33), equation (96) vanishes to zeroth order in η, while equations (95) and (97) yield

$$(K_0 - \beta_0^2)A_1^{(0)} - i\beta_0 p_0 A_3^{(0)} = 0, \tag{100}$$

$$-i\beta_0 p_0 A_1^{(0)} + (K_0 + p_0^2)A_3^{(0)} = 0. \tag{101}$$

The determinant of this pair of homogeneous equations vanishes because equation (33) is satisfied, so a nontrivial solution exists. The quantities $B_1^{(0)}$ and $B_3^{(0)}$ satisfy the same equations. Using equation (33), it follows from equation (101) that $A_1^{(0)}$ and $A_3^{(0)}$ are related by equation (48), and $B_1^{(0)}$ and $B_3^{(0)}$ are related by equation (51).

To first order in η, equations (95) through (97) are

$$(K_0 - \beta_0^2)A_1^{(1)} - i\beta_0 p_0 A_3^{(1)} = -[(S_{11} - 2\beta_0\beta_1)A_1^{(0)} + S_{12}A_2^{(0)} + (S_{13} - ip_0\beta_1 - ip_1\beta_0)A_3^{(0)}], \tag{102}$$

$$S_{12}A_1^{(0)} + (2p_0p_1 - 2\beta_0\beta_1 + S_{22})A_2^{(0)} + S_{23}A_3^{(0)} = 0, \tag{103}$$

$$-i\beta_0 p_0 A_1^{(1)} + (K_0 + p_0^2)A_3^{(1)} = -[(S_{13} - ip_0\beta_1 - ip_1\beta_0)A_1^{(0)} + S_{23}A_2^{(0)} + (S_{33} + 2p_0p_1)A_3^{(0)}]. \tag{104}$$

With the replacement of A_a by B_a and p by q in equations (102) through (104) we obtain the first order equations satisfied by the B_a. At this stage we must differentiate between the perturbed TE and TM modes. For the perturbed TE modes we must have

$$A_1^{(0)} + B_1^{(0)} = -i(\beta_0/p_0)[A_3^{(0)} + B_3^{(0)}] = 0, \tag{105}$$

$$A_2^{(0)} + B_2^{(0)} = \cos(kf_0w), \tag{106}$$

while for the perturbed TM modes

$$A_1^{(0)} + B_1^{(0)} = -i(\beta_0/p_0)[A_3^{(0)} + B_3^{(0)}] = (K_1/K_0)\cos(kl_0w), \tag{107}$$

$$A_2^{(0)} + B_2^{(0)} = 0. \tag{108}$$

If we now add to equation (103) the equivalent equation in B_α, and make use of the fact that $A_\alpha^{(0)} + B_\alpha^{(0)}$, $\alpha = 1, 2, 3$ are prescribed for the perturbed TE modes, in equations (105) and (106), we get a new equation involving only $A_2^{(0)}$ and $B_2^{(0)}$. This equation together with equation (106) can be solved for $A_2^{(0)}$ and $B_2^{(0)}$ to yield (47) and (50). Once $A_2^{(0)}$ is known, $A_1^{(0)}$ and $A_3^{(0)}$ can be determined from equations (103) and (48) yielding (46). We get $B_1^{(0)}$ and $B_3^{(0)}$ from equation (105). In the same fashion we determine $A_\alpha^{(0)}$ and $B_\alpha^{(0)}$, $\alpha = 1, 2, 3$ for the perturbed TM modes.

Equations (102) and (104) [and the two equivalent equations in $B_1^{(1)}$ and $B_3^{(1)}$] are two inhomogeneous equations whose determinant vanishes. Thus the left side of (102) is a multiple of the left side of (104), and the equations are compatible only if the right side of (102) is the same multiple of the right side of (104). This can be shown to be the case, and so (102) and (104) provide just one relationship between $A_1^{(1)}$ and $A_3^{(1)}$. There is a corresponding relationship between $B_1^{(1)}$ and $B_3^{(1)}$.

By replacing A_α, B_α, p, q by C_α, D_α, $-r$, $-s$, respectively, in the equations so far obtained, the formulas for the region $\xi \leqq -\sigma$ are obtained. Here $-r$ and $-s$ are the remaining two roots of the quartic $D(p, \beta) = 0$.

Next, if the assumed expressions for e_α in $|\xi| < \sigma$ given by equation (11) are substituted into equations (90) through (92) we get four sets of three homogeneous, linear equations in F_α, G_α, L_α, and M_α, respectively, which hold for both the perturbed TE and TM modes. These equations are obtained from equations (95) through (97) by replacing A_α and p by F_α and $-if$, G_α and ig, L_α and $-il$, and M_α and im, respectively.

The determinental equation for each of these four sets of homogeneous equations can again be expanded in powers of η, and the coefficient of each power of η separately equated to zero. The vanishing of the zeroth order coefficients yields equations (34), (41), (36), and (42) relating f_0, g_0, l_0, and m_0 to β_0. The vanishing of the first order coefficients yields equations (44) and (45) relating f_1, g_1, l_1, and m_1 to β_1.

Each of the four sets of homogeneous equations can be expanded in powers of η, just as for the equations describing the region $\xi \geqq \sigma$. To proceed further, we must again differentiate between the perturbed TE and TM modes. For the perturbed TE modes, equations (52) through (54) must be satisfied, while for the perturbed TM modes, equations (74) through (77) must be satisfied. These values satisfy the lowest order equations identically.

For both the perturbed TE and TM modes, the first order equations can be written as

$$(K_1 - \beta_0^2)F_1^{(1)} - \beta_0 f_0 F_3^{(1)} = -S_{12}F_2^{(0)}, \tag{109}$$

$$-\beta_0 f_0 F_1^{(1)} + (K_3 - f_0^2)F_3^{(0)} = -S_{23}F_2^{(0)}, \tag{110}$$

$$(K_2 - f_0^2 - \beta_0^2)F_2^{(1)} = -S_{12}F_1^{(0)} - S_{23}F_3^{(0)} - (S_{22} - 2f_0 f_1 - 2\beta_0\beta_1)F_2^{(0)}. \tag{111}$$

For both the perturbed TE and TM modes, equation (111) vanishes and yields no information, while equations (109) and (110) have a nonzero determinant and so can be solved for $F_1^{(1)}$ and $F_3^{(1)}$ yielding the solutions given in (55), (57), and (78). If we replace F_α and f by G_α and $-g$, we get equations which can be solved for $G_1^{(1)}$ and $G_3^{(1)}$ yielding solutions given in (55), (57), and (78). The equations obtained when F_α and f are replaced by L_α and l, and M_α and $-m$, respectively, have a different character. The two equations in $L_2^{(1)}$ and $M_2^{(1)}$ corresponding to (111) do not vanish identically and can be solved for $L_2^{(1)}$ and $M_2^{(1)}$. The solutions are given in (61) and (80). The equations in $L_1^{(1)}$ and $L_3^{(1)}$, and $M_1^{(1)}$ and $M_3^{(1)}$, have a vanishing determinant. In the perturbed TE case, the equations are homogeneous and yield (59) and (60). In the perturbed TM case, the equations are nonhomogeneous but compatible, and yield the relations

$$-\beta_0 l_0 L_1^{(1)} + (K_3\beta_0^2/K_1)L_3^{(1)} = -\tfrac{1}{2}(S_{13} - l_0\beta_1 - l_1\beta_0) - \tfrac{1}{2}(S_{33} - 2l_0 l_1)(l_0 K_1/\beta_0 K_3), \tag{112}$$

$$\beta_0 l_0 M_1^{(1)} + (K_3\beta_0^2/K_1)M_3^{(1)} = -\tfrac{1}{2}(S_{13} + l_0\beta_1 + m_1\beta_0) + \tfrac{1}{2}(S_{33} - 2l_0 m_1)(l_0 K_1/\beta_0 K_3). \tag{113}$$

We finally turn to the boundary conditions at $\xi = \pm\sigma$, of which there are eight, four at each boundary. They can be grouped as follows

$$A_2 + B_2 = F_2 e^{if\sigma} + G_2 e^{-ig\sigma} + L_2 e^{il\sigma} + M_2 e^{-im\sigma}, \tag{114}$$

$$C_2 + D_2 = F_2 e^{-if\sigma} + G_2 e^{ig\sigma} + L_2 e^{-il\sigma} + M_2 e^{im\sigma}, \tag{115}$$

$$-pA_2 - qB_2 = ifF_2 e^{if\sigma} - igG_2 e^{-ig\sigma} + ilL_2 e^{il\sigma} - imM_2 e^{-im\sigma}, \tag{116}$$

$$rC_2 + sD_2 = ifF_2 e^{-if\sigma} - igG_2 e^{ig\sigma} + ilL_2 e^{-il\sigma} - imM_2 e^{im\sigma}, \tag{117}$$

$$A_3 + B_3 = F_3 e^{if\sigma} + G_3 e^{-ig\sigma} + L_3 e^{il\sigma} + M_3 e^{-im\sigma}, \tag{118}$$

$$C_3 + D_3 = F_3 e^{-if\sigma} + G_3 e^{ig\sigma} + L_3 e^{-il\sigma} + M_3 e^{im\sigma}, \tag{119}$$

$$-pA_3 - qB_3 + i\beta(A_1 + B_1) = i(fF_3 + \beta F_1)e^{if\sigma} + i(-gG_3 + \beta G_1)e^{-ig\sigma} + i(lL_3 + \beta L_1)e^{il\sigma} + i(-mM_3 + \beta M_1)e^{-im\sigma}, \quad (120)$$

$$rC_3 + sD_3 + i\beta(C_1 + D_1) = i(fF_3 + \beta F_1)e^{-if\sigma} + i(-gG_3 + \beta G_1)e^{ig\sigma} + i(lL_3 + \beta L_1)e^{-il\sigma} + i(-mM_3 + \beta M_1)e^{im\sigma}. \quad (121)$$

These equations split naturally into two groups, one group involving only the subscript 2 and the other group involving only the subscripts 1 and 3. These equations can be expanded in powers of η. The zeroth order equations are satisfied as long as (35) holds in the perturbed TE case and (37) holds in the perturbed TM case.

For the perturbed TE modes, the first order expansion of equations (114) through (117) yields four nonhomogeneous equations in $A_2^{(1)} + B_2^{(1)}$, $C_2^{(1)} + D_2^{(1)}$, $F_2^{(1)}$ and $G_2^{(1)}$. The inhomogeneous terms on the right side of these equations contain the parameter β_1. The determinant of the equations vanishes, and then the condition that they be compatible provides an equation from which β_1, given in (43), is determined. Once β_1 is determined, these equations yield (56) and (66). We can now choose the arbitrary parameter a_1—indicated in (94)—so that $F_2^{(1)} = 0$. Then from (56) and (66) $G_2^{(1)} = A_2^{(1)} + B_2^{(1)} = 0$. In addition, since β_1 is real, it can now be shown that $r_1 = p_1^*$, $s_1 = q_1^*$, $C_\alpha^{(0)} = A_\alpha^{(0)*}$, $D_\alpha^{(0)} = B_\alpha^{(0)*}$, and $C_\alpha^{(1)} + D_\alpha^{(1)} = [A_\alpha^{(1)} + B_\alpha^{(1)}]^*$, $\alpha = 1, 2, 3$, which justifies equation (39). Finally, the first order expansion of equations (118) through (121) can be combined with equations (59) and (60), equation (102), and the corresponding three equations in $B_1^{(1)}$ and $B_3^{(1)}$, $C_1^{(1)}$ and $C_3^{(1)}$, and $D_1^{(1)}$ and $D_3^{(1)}$ to form a set of equations from which $A_\alpha^{(1)} + B_\alpha^{(1)} = [C_\alpha^{(1)} + D_\alpha^{(1)}]^*$, and $M_\alpha^{(1)}$, $\alpha = 1, 3$, can be determined. These are listed in (58), (65), and (67).

For the perturbed TM modes the procedure is virtually the same, except that it is now the first order expansion of equations (118) through (121) which has a vanishing determinant. The condition that these be compatible then yields the expressions (68) and (69) for β_1. This set of equations also yields the result that

$$L_3^{(1)} + M_3^{(1)} = 0. \quad (122)$$

We can now pick the arbitrary parameter a_1 so that $L_3^{(1)} = 0$, which combined with (122) yields (85). Equations (112) and (113) now yield (86) and (87). The first order term of equation (118) then yields (88), and this result, combined with the equation obtained by adding equation (102) to the corresponding equation in B yields (89). Finally, equations (114) through (117) yield expressions (79) through (84) for $F_2^{(1)}$, $G_2^{(1)}$, $L_2^{(1)}$, $M_2^{(1)}$, $A_2^{(1)} + B_2^{(1)}$.

ACKNOWLEDGMENTS

The authors gratefully acknowledge several helpful conversations with D. F. Nelson and F. K. Reinhart, and the assistance of Miss A. Marucchi in programming the numerical calculations.

REFERENCES

1. Ashkin, A., and Gershenzon, M., "Reflection and Guiding of Light at p-n Junctions," J. Appl. Phys., *34,* No. 7 (July 1963), pp. 2116-2119.
2. Nelson, D. F., and Reinhart, F. K., "Light Modulation by the Electro-Optic Effect in Reverse-Biased GaP p-n Junctions," Appl. Phys. Letters, *5,* No. 7 (October 1964), pp. 148–150.
3. Walters, W. L., "Electro-Optic Effect in Reverse-Biased GaAs p-n Junctions," J. Appl. Phys., *37,* No. 2 (February 1966), p. 916.
4. Nelson, D. F. and McKenna, J., "Electromagnetic Modes of Anisotropic Dielectric Waveguides at p-n Junctions," J. Appl. Phys., *38,* No. 10 (September 1967), pp. 4057–4074.
5. McKenna, J., "The Excitation of Planar Dielectric Waveguides at p-n Junctions, I." B.S.T.J., *46,* No. 7 (September 1967), pp. 1491–1526.
6. Nelson, D. F., "A Proposal for Internal Electro-Optic Intensity and Frequency Modulation of Diode Lasers," IEEE J. Quantum Elec., *QE-3,* No. 12 (December 1967), pp. 667–674.
7. Oldham, W. G., and Bahraman, A., "Electro-Optic Junction Modulators," IEEE J. Quantum Elec., *QE-3,* No. 7 (July 1967), pp. 278–286.
8. Reinhart, F. K., "Reverse-Biased Gallium Phosphide Diodes as High Frequency Light Modulators," J. Appl. Phys., *39,* No. 7 (June 1968), pp. 3426–3434.
9. Reinhart, F. K., Nelson, D. F., and McKenna, J., unpublished work.
10. Nye, J. F., *Physical Properties of Crystals,* London: Oxford University Press, 1960, p. 241.
11. Yariv, A., and Leite, R. C. C., "Dielectric-Waveguide Mode of Light Propagation in p-n Junctions," Appl. Phys. Letters, *2,* No. 3 (February 1963), pp. 55–57.
12. McWhorter, A. L., "Electromagnetic Theory of the Semiconductor Junction Laser," Solid State Elec., *6,* No. 5 (September 1963), pp. 417–423.
13. Kazarinov, R. F., Konstantinov, O. V., Perel, V. I., and Éfros, A. L., "Electromagnetic Theory of the Injection Laser," Soviet Physics, Solid State, *7,* No. 5 (November 1965), pp. 1210–1217. Translated from Fizika Tverdogo Tela, *7,* No. 5 (May 1965), pp. 1506–1516.
14. Zachos, T. H., and Ripper, J. E., unpublished work.
15. Kaminow, I. P., "Strain Effects in Electro-Optic Light Modulators," Applied Optics, *3,* No. 4 (April 1964), pp. 511–515.
16. Di Domenico, M. Jr., and Anderson, L. K., "Broadband Electro-Optic Traveling-Wave Light Modulators," B.S.T.J., *42,* No. 6 (November 1963), pp. 2621–2678.
17. Reinhart, F. K., unpublished work.
18. Nelson, D. F., and Turner, E. H., "Electro-Optic and Piezoelectric Coefficients and Refractive Index of Gallium Phosphide," J. Appl. Phys., *39,* No. 7 (June 1968), pp. 3337–3343.
19. Cohen, B. G., "X-Ray Measurement of Elastic Strain and Lattice Constant of Diffused Silicon," Solid State Electron., *10,* No. 1 (January 1967), pp. 33–37.
20. Dixon, R. W., "Photoelastic Properties of Selected Materials and Their Relevance for Applications to Acoustic Light Modulators and Scanners," J. Appl. Phys., *38,* No. 13 (December 1967). pp. 5149–5153.
21. Stratton, J. A, *Electromagnetic Theory,* New York: McGraw-Hill, 1941, p 34.

A State Variable Method of Circuit Analysis Based on a Nodal Approach

By R. E. PARKIN

(Manuscript received April 12, 1968)

A method which is well suited for implementation on a digital computer is presented for the solutions of active circuits. Unlike many state variable approaches the state vector is defined as the set of voltages which exist between certain nodes and the reference node. An advantage of this approach is that degeneration in the order of complexity of the network caused by capacitance loops is handled automatically. Any type of controlled source can be specified. From the basic algorithm the circuit is specified in matrix form by inspection using standard nodal methods, and the solution is obtained by a systematic reduction of this one matrix equation. An upper bound on the order of complexity of the network is evident from the network topology or the partitioned form of the original matrix. Inductors are included in this approach by considering the equivalent gyrator-capacitor combination.

I. INTRODUCTION

State variable techniques presently being used to analyze networks require a detailed knowledge of graph theory.[1–7] Another method of state variable analysis that is based partly on a nodal approach and does not require a detailed knowledge of graph theory is very restrictive.[8] The method presented here performs a nodal analysis on a transformation of the network in which all magnetic storage elements have been replaced by gyrator-capacitor equivalents, and nothing more than a basic knowledge of graph theory nomenclature is required. The RCLMST* network can be transformed to an equivalent

---

* Resistor, capacitor, inductor, mutual inductor, source and ideal transformer.

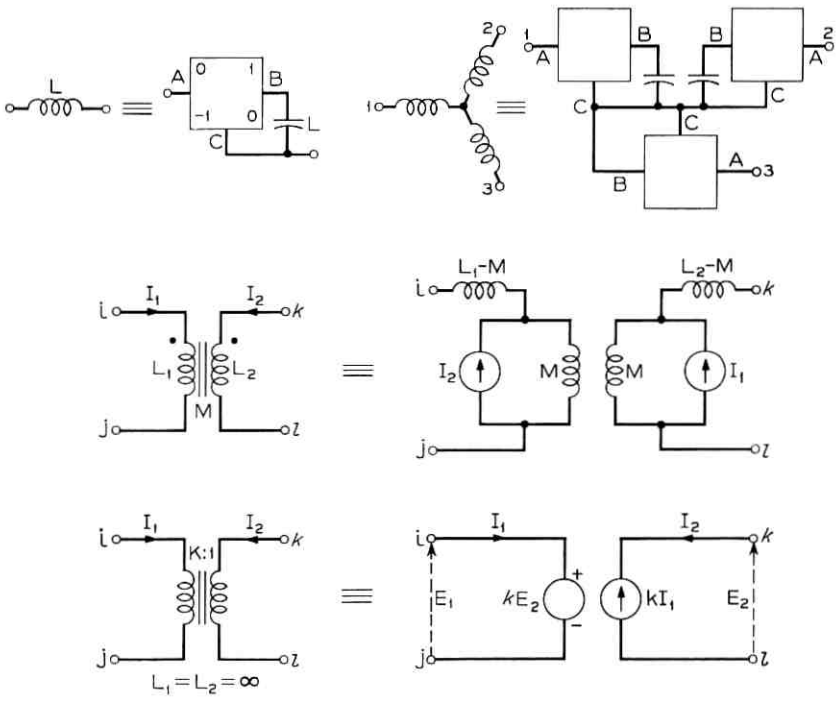


Fig. 1 — Inductor and transformer equivalents.

resistance capacitance source network using the gyrator-capacitor equivalents shown in Fig. 1. Each gyrator shown in Fig. 1 has the indefinite admittance parameters

$$\begin{bmatrix} I_A \\ I_B \\ I_C \end{bmatrix} = \begin{bmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} V_A \\ V_B \\ V_C \end{bmatrix};$$

choosing this type of gyrator enables the capacitor value in farads of the equivalent pair to be equal to the inductor value in henries.

Let the number of nodes of a transformed network be n. Using Kirchoff's current law, it can be shown that for an n-node RCS network

$$C\dot{\mathbf{V}} = \mathbf{I} - G\mathbf{V} \tag{1}$$

where $\mathbf{I}$ is an $(n-1)$th ordered column vector representing the currents

injected into the nodes, **V** is an $(n-1)$th ordered column vector representing the voltages between the nodes of the network and the reference node. If the transformed network contains l capacitors then the matrix C is an $(n-1)$th ordered symmetric matrix which contains imbedded within it l second order indefinite matrices, each having the dimensions of farads. Similarly G represents the resistors and has the dimensions of mhos, but G may be asymmetric. Node n is the common or ground node of the network; for convenience this node is always assumed to have capacitors connected to it.

The objective is to find an upper bound on the rank of the capacitance matrix C by partitioning C as described in Section II, and reducing the matrix equation (1) containing the partitioned matrix C to the rank of C; this reduction is symbolic and does not take into account degenerate cases which can occur. It is shown in Appendix B that for all conditions, for any type of circuit, an upper bound on the order of complexity of the network (rank of C) can be found from the network topology.

II. PARTITIONING OF THE CAPACITANCE MATRIX

There are basically four types of voltage source (VS), the independent VS (IVS), the voltage dependent VS (VDVS), the current dependent VS where the current is through a resistor (CDVSR), and the current dependent VS where the current is through a capacitor (CDVSC). It will be shown that the only current source (CS) which can effect the partitioning is the current dependent CS where the current is through a capacitor (CDCSC). As a result, any type of CS will be termed simply a CS, unless it is a CDCSC.

The method of partitioning makes the reduction of the matrix equation (1) to its rank a simple process. Generally only the voltage at a node to which a capacitor is connected can be a state variable node. However it is possible to choose a node to which a CDCSC or CDVSC is connected as a state variable node instead of one of the nodes of the capacitor whose current supplies the dependence, but this possibility is avoided automatically in the partitioning method presented here.

The presence of inductors and time-invariant, independent CS's forming a cut-set in the original untransformed network causes a linear dependence problem in the transformed network. In the transformed network such a cut-set appears as a capacitor tree with gyrators only connected to the end nodes of the tree as shown in Fig. 1,

and gyrators and perhaps time-invariant independent c.s.'s connected to the central node (the GCNODE); the nodes of this capacitor tree will be called the GCSET nodes.

The capacitors in the transformed network can be divided into two classes, those connected to the reference node directly or through a vs-capacitor chain (the fixed capacitors), and those not so connected (the floating capacitors). The m floating capacitor subgraphs are defined as the m unconnected subgraphs obtained from the floating capacitor plus imbedded vs graph of the transformed network.

The partitioning of the capacitance matrix will be related to the example of Appendix A in the discussion that follows. Partition the matrix C as

$$\begin{array}{c} \\ \mathbf{n1} \\ \mathbf{n2} \\ \mathbf{n3} \\ \mathbf{n4} \\ \mathbf{n5} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathbf{n1} & \mathbf{n2} & \mathbf{n3} & \mathbf{n4} & \mathbf{n5} \end{array} \\ \begin{bmatrix} C_{11} & C_{12} & C_{13} & 0 & C_{15} \\ C_{21} & C_{22} & C_{23} & 0 & C_{25} \\ C_{31} & C_{32} & C_{33} & 0 & C_{35} \\ 0 & 0 & 0 & 0 & 0 \\ C_{51} & C_{52} & C_{53} & 0 & C_{55} \end{bmatrix} \end{array}$$

where

(*i*) The nodes **n1** are all the nodes to which capacitors are connected omitting the following nodes:

(*a*) A node for each vs imbedded in a capacitor chain (these nodes are in the **n2** section), but each capacitor must be specified by at least one node.

(*b*) A node for each of the m floating capacitor subgraphs (these nodes being in the **n3** section).

(*c*) A node for each GCSET which is specified in section **n2**.

In the example in Appendix A, **n1** contains nodes $1 \rightarrow 9$.

(*ii*) The nodes **n2** represent:

(*a*) A node for each DVS imbedded in a capacitor chain.

(*b*) A node free of capacitors for each CDVSC and CDCSC free of capacitors on at least one node.

(*c*) A node for each GCSET.

In the Appendix A example **n2** contains nodes 10 and 11.

(*iii*) Section **n3** contains a node for each of the m floating capacitor subgraphs.

In the example **n3** contains nodes 12 and 13.

(*iv*) Section **n4** contains the nodes to which only resistors and CS's (but not CDCSC) are connected, including a node for each IVS, VDVS or CDCSC free of capacitors, other CDVSC's or CDCSC's on both nodes. (The other nodes of these sources are specified in section **n5**).

In the example **n4** has no entries.

(*v*) Section **n5** contains all the remaining nodes. These are:

(*a*) A node for each IVS.

(*b*) A node free of capacitors for each VDVS or CDVSR free of capacitors or CDVSC or CDCSC on at least one node.

In the example **n5** contains node 14.

The rank of the C matrix is **n1**, and **n1** = 9 for the example of Appendix A. Notice that the presence of capacitance loops in no way alters the method of partitioning.

III. REDUCTION OF THE CIRCUIT DESCRIPTION TO A MINIMAL FORM

Theorem: *An upper bound on the order of complexity of a network is the order of* **n1**.

This theorem is proved in Appendix B, where it is shown that every row in sections **n2**, **n3**, **n4**, and **n5** is linearly dependent on rows in section **n1**; the subspace spanned by sections **n2**, **n3**, **n4**, and **n5** is contained in **n1**.

The systematic reduction of equation (1) is accomplished by first eliminating section **n5** by applying the voltage restrictions caused by the VS's in section **n5**. Secondly, section **n4** is eliminated using the fact that these nodes are free of capacitors. Next, section **n3** is eliminated to correct the over specification of the floating capacitor subgraphs. Finally, the remaining dependencies of the system are caused by the DVS's imbedded in capacitive chains, the CDVSC and CDCSC free of capacitors on at least one node, and a node for each capacitive tree in which a GCSET has occurred; these dependencies are eliminated with section **n2**, yielding equation (9) of Appendix B.

Equation (9) of Appendix B can be written as

$$\dot{\mathbf{v}}\mathbf{1} = \mathbf{B} - \mathbf{A}\mathbf{v}\mathbf{1} \tag{2}$$

where

$$\mathbf{v} = \begin{bmatrix} \mathbf{v1} \\ \mathbf{v2} \\ \mathbf{v3} \\ \mathbf{v4} \\ \mathbf{v5} \end{bmatrix}$$

and **v1** is the voltage vector for nodes **n1, n2, n3,** etc.

An example of a solution based on the problem set by Pottle is given in Appendix C.

IV. CONCLUDING REMARKS

A state variable technique has been described that offers two advantages over traditional methods:

(*i*) The network can be specified completely by inspection using well known nodal techniques with little skill required, the problem then becoming one of simple matrix reduction (easily programmed for a digital computer).

(*ii*) Capacitor loops present no problem and are not even recognized as such since the partitioning and matrix reduction are unaltered if there are any capacitor loops present.

The main disadvantages are that currents must always be expressed as functions of node voltages and inductors must be replaced by gyrators and capacitors; inductor cut-sets must be recognized and the circuit redrawn before inductors are eliminated so that the cut-set encircles one node only, and this is sometimes inconvenient.

APPENDIX A

Example of Partitioning

For the example of Fig. 2(a), the transformed circuit without inductors is given in Fig. 2(b). (This is a theoretical problem and the circuit has no practical value.) This circuit is described by the equations

$$
\begin{array}{c}
\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \\ 13 \\ 14 \end{array}
\begin{bmatrix}
\begin{array}{c} 1 \\ C_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} &
\begin{array}{c} 2 \\ 0 \\ C_2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} &
\begin{array}{c} 3 \\ 0 \\ 0 \\ L_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} &
\begin{array}{c} 4 \\ 0 \\ 0 \\ 0 \\ C_6 \\ -C_6 \\ 0 \\ 0 \\ 0 \end{array} &
\begin{array}{c} 5 \\ 0 \\ 0 \\ 0 \\ -C_6 \\ C_6+C_7 \\ 0 \\ 0 \\ 0 \end{array} &
\begin{array}{c} 6 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ L_2 \\ -L_2 \\ 0 \end{array} &
\begin{array}{c} 7 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -L_2 \\ L_2+C_3 \\ -C_3 \end{array} &
\begin{array}{c} 8 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -C_3 \\ C_3+C_4 \\ +C_8 \\ -C_4 \\ 0 \\ 0 \\ -C_8 \\ 0 \end{array} &
\begin{array}{c} 9 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -C_4 \\ C_4+C_5 \\ 0 \\ 0 \\ 0 \\ -C_5 \\ 0 \end{array} &
\begin{array}{c} 10 \\ 0 \\ -C_2 \\ -L_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ C_2+L_1 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} &
\begin{array}{c} 11 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array} &
\begin{array}{c} 12 \\ 0 \\ 0 \\ 0 \\ 0 \\ -C_7 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ C_7 \\ 0 \\ 0 \end{array} &
\begin{array}{c} 13 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -C_8 \\ -C_5 \\ 0 \\ 0 \\ 0 \\ C_5+C_8 \\ 0 \end{array} &
\begin{array}{c} 14 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{array}
\end{bmatrix}
\end{array}
=
\begin{bmatrix}
\dot{v}_1 \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \dot{v}_{14}
\end{bmatrix}
$$

$$
\begin{array}{c}
\\
1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \\ 13 \\ 14
\end{array}
\begin{bmatrix}
I_1 \\
-k_2C_2(\dot{v}_{10}-\dot{v}_2) \\
0 \\
0 \\
0 \\
0 \\
0 \\
k_3i_{10} \\
-I_1 \\
k_2C_2(\dot{v}_{10}-\dot{v}_2) \\
0 \\
0 \\
I_2-k_3i_{10}
\end{bmatrix}
\begin{array}{c}
1 \\
\begin{bmatrix}
G_2 \\ 0 \\ 0 \\ -G_2 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0
\end{bmatrix}
\end{array}
\begin{array}{c}
2 \\
\begin{bmatrix}
0 \\ G_3+G_6 \\ -1 \\ -G_3 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ -G_5 \\ 0 \\ 0 \\ 0 \\ 0
\end{bmatrix}
\end{array}
\begin{array}{c}
3 \\
\begin{bmatrix}
0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0
\end{bmatrix}
\end{array}
\begin{array}{c}
4 \\
\begin{bmatrix}
-G_2 \\ -G_3 \\ 0 \\ G_2+G_3 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0
\end{bmatrix}
\end{array}
\begin{array}{c}
5 \\
\begin{bmatrix}
0 \\ 0 \\ 0 \\ 0 \\ G_4 \\ 0 \\ 0 \\ 0 \\ 0 \\ -G_4 \\ 0 \\ 0 \\ 0 \\ 0
\end{bmatrix}
\end{array}
\begin{array}{c}
6 \\
\begin{bmatrix}
0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1
\end{bmatrix}
\end{array}
\begin{array}{c}
7 \\
\begin{bmatrix}
0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1
\end{bmatrix}
\end{array}
\begin{array}{c}
8 \\
\begin{bmatrix}
0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ G_{10} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0
\end{bmatrix}
\end{array}
\begin{array}{c}
9 \\
\begin{bmatrix}
0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0
\end{bmatrix}
\end{array}
\begin{array}{c}
10 \\
\begin{bmatrix}
0 \\ -1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ G_8+G_9 \\ 0 \\ 0 \\ 0 \\ 0 \\ -G_9
\end{bmatrix}
\end{array}
\begin{array}{c}
11 \\
\begin{bmatrix}
0 \\ -G_5 \\ 0 \\ -G_4 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ G_4+G_5+G_6 \\ 0 \\ -G_6 \\ 0
\end{bmatrix}
\end{array}
\begin{array}{c}
12 \\
\begin{bmatrix}
0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ G_1 \\ 0 \\ 0
\end{bmatrix}
\end{array}
\begin{array}{c}
13 \\
\begin{bmatrix}
0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -G_6 \\ G_6+G_7
\end{bmatrix}
\end{array}
\begin{array}{c}
14 \\
\begin{bmatrix}
0 \\ 0 \\ 0 \\ 0 \\ -1 \\ 1 \\ 0 \\ 0 \\ 0 \\ -G_9 \\ 0 \\ 0 \\ 0 \\ -G_7
\end{bmatrix}
\end{array}
\begin{bmatrix}
v_1 \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ v_{14}
\end{bmatrix}
$$

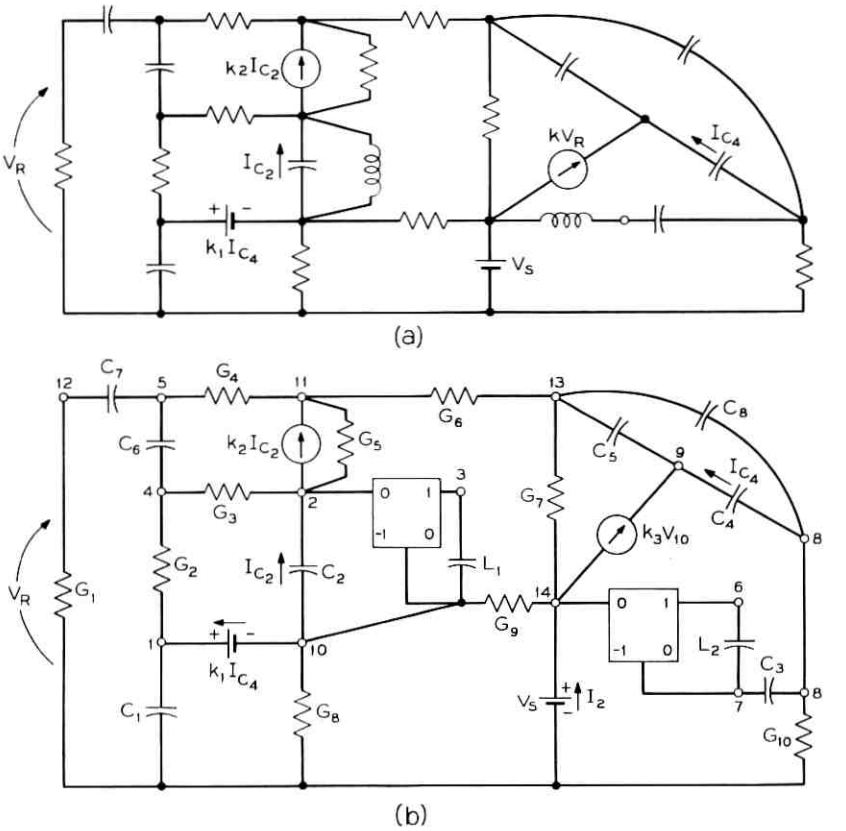


Fig. 2 — Circuit to demonstrate transformation and partitioning.

where I_1 and I_2 are the unbalance currents due to the vs's and

$$v_{14} = V_S .$$

Notice that except for degenerate cases (for example, if $C_6 = 0$), the order of complexity of this network is 9.

APPENDIX B

Matrix Reduction

Consider the partitioned form of equation (1). The section **n5** can be eliminated as follows: for an IVS of α volts connected between nodes k

and l (node k is in section **n5**, node l is not)

$$v_k = v_l + \alpha.$$

For a VDVS or CDVSR connected between nodes k and l, where the vs is dependent on the voltage vector $\mathbf{v}_m$ (each voltage of $\mathbf{v}_m$ is not in section **n5**).

$$v_k = v_l + \beta \mathbf{v}_m.$$

Thus the system can be reduced to

$$\begin{array}{c} \\ \mathbf{n1} \\ \mathbf{n2} \\ \mathbf{n3} \\ \mathbf{n4} \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{n1} & \mathbf{n2} & \mathbf{n3} & \mathbf{n4} \end{array} \\ \begin{bmatrix} C_{11} & C_{12} & C_{13} & 0 \\ C_{21} & C_{22} & C_{23} & 0 \\ C_{31} & C_{32} & C_{33} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{bmatrix} \dot{\mathbf{v}}1 \\ \dot{\mathbf{v}}2 \\ \dot{\mathbf{v}}3 \\ \dot{\mathbf{v}}4 \end{bmatrix} = \begin{bmatrix} \mathbf{i}1 \\ \mathbf{i}2 \\ \mathbf{i}3 \\ \mathbf{i}4 \end{bmatrix} - \begin{bmatrix} G_{11} & G_{12} & G_{13} & G_{14} \\ G_{21} & G_{22} & G_{23} & G_{24} \\ G_{31} & G_{32} & G_{33} & G_{34} \\ G_{41} & G_{42} & G_{43} & G_{44} \end{bmatrix} \begin{bmatrix} \mathbf{v}1 \\ \mathbf{v}2 \\ \mathbf{v}3 \\ \mathbf{v}4 \end{bmatrix}. \quad (3)$$

Nodes **n4** can be eliminated by first writing part of equation (3) as

$$\mathbf{v}4 = G_{44}^{-1} \left\{ \mathbf{i}4 - [G_{41}\ G_{42}\ G_{43}] \begin{bmatrix} \mathbf{v}1 \\ \mathbf{v}2 \\ \mathbf{v}3 \end{bmatrix} \right\} \quad (4)$$

Thus

$$\begin{aligned} \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{v}}1 \\ \dot{\mathbf{v}}2 \\ \dot{\mathbf{v}}3 \end{bmatrix} &= \begin{bmatrix} \mathbf{i}1 \\ \mathbf{i}2 \\ \mathbf{i}3 \end{bmatrix} - \begin{bmatrix} G_{11} & G_{12} & G_{13} \\ G_{21} & G_{22} & G_{23} \\ G_{31} & G_{32} & G_{33} \end{bmatrix} \begin{bmatrix} \mathbf{v}1 \\ \mathbf{v}2 \\ \mathbf{v}3 \end{bmatrix} - \begin{bmatrix} G_{14} \\ G_{24} \\ G_{34} \end{bmatrix} \mathbf{v}4 \\ &= \begin{bmatrix} \mathbf{it}1 \\ \mathbf{it}2 \\ \mathbf{it}3 \end{bmatrix} - \begin{bmatrix} G1_{11} & G1_{12} & G1_{13} \\ G1_{21} & G1_{22} & G1_{23} \\ G1_{31} & G1_{32} & G1_{33} \end{bmatrix} \begin{bmatrix} \mathbf{v}1 \\ \mathbf{v}2 \\ \mathbf{v}3 \end{bmatrix}. \end{aligned} \quad (5)$$

The matrix

$$\begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix}$$

has order **n1** + **n2** + **n3** and rank no greater than **n1** + **n2**. The **n3**

linearly dependent rows and columns can be deleted from equation (5) by adding selected rows in the section **n1** and **n2** to rows in the range **n3**. The selection is made as follows: starting with any row in the section **n3**, examine the first entry. If it is nonzero add row 1 to this row. Continue along the row, repeating if necessary, until all the entries are zero. Proceed for the other dependent rows. Equation (5) can then be written as

$$\begin{bmatrix} C_{11} & C_{12} & C_{13} \\ & & \\ C_{21} & C_{22} & C_{23} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{v}}1 \\ \dot{\mathbf{v}}2 \\ \dot{\mathbf{v}}3 \end{bmatrix} = \begin{bmatrix} \mathbf{it}1 \\ \\ \mathbf{it}2 \end{bmatrix} - \begin{bmatrix} G1_{11} & G1_{12} & G1_{13} \\ & & \\ G1_{21} & G1_{22} & G1_{23} \end{bmatrix} \begin{bmatrix} \mathbf{v}1 \\ \mathbf{v}2 \\ \mathbf{v}3 \end{bmatrix} \tag{6}$$

and

$$\mathbf{v}3 = G1'^{-1}_{33} \left\{ \mathbf{it}3' - [G1'_{31}\ G1'_{32}] \begin{bmatrix} \mathbf{v}1 \\ \mathbf{v}2 \end{bmatrix} \right\}. \tag{7}$$

It is a simple process for the reader to prove to himself that eliminating a node of a floating capacitor subgraph which is part of a GCSET as described above yields the same result as equating the algebraic sum of the voltages across the capacitors in the GCSET to zero (analogous to the algebraic sum of the currents entering the inductor cut-set node through the inductors adding up to zero).

Substituting equation (7) and its derivative into equation (6) we obtain

$$\begin{bmatrix} C2_{11} & C2_{12} \\ C2_{21} & C2_{22} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{v}}1 \\ \dot{\mathbf{v}}2 \end{bmatrix} = \begin{bmatrix} \mathbf{ip}1 \\ \mathbf{ip}2 \end{bmatrix} - \begin{bmatrix} G2_{11} & G2_{12} \\ G2_{21} & G2_{22} \end{bmatrix} \begin{bmatrix} \mathbf{v}1 \\ \mathbf{v}2 \end{bmatrix}. \tag{8}$$

The total number of restrictions have not yet been placed on the network.

(*i*) For a DVS imbedded in a capacitor chain or a CDVSC free of capacitors on one node connected between nodes k and l, where node k is specified in section **n2**,

$$v_k = v_l + \gamma \mathbf{v}_j$$

or

$$v_k = v_l + v\dot{\mathbf{v}}_j$$

where $\mathbf{v}_j$ is the set of voltages upon which the source is dependent. A particular voltage of $\mathbf{v}_j$ may be in any section **n1**, **n2**, **n3**, **n4**, or **n5**.

(*ii*) For a CDCSC connected between nodes k and l, the currents I_k and I_l injected into nodes k and l with the CDCSC removed must be modified to

$$I_k + \eta\dot{\mathbf{v}}_i$$

and

$$I_l - \eta\dot{\mathbf{v}}_i,$$

respectively, where η has the dimensions of farads.

(*iii*) For a GCSET with node j of the capacitor tree containing the GCSET specified in section **n2**, node j is eliminated as follows: examine the entries of row j of the remaining capacitance matrix. If entry $C_{j,l} \neq 0$, subtract $C_{j,l}/C_{l,l}$ times row l from row j, where $l = 1, p$; p is the order of **n1** + **n2**. Thus row j is reduced to a row of zeros.

The system can now be written as

$$[C]\dot{\mathbf{v}}\mathbf{1} = \mathbf{iF1} - [G]\mathbf{v1}. \tag{9}$$

Barring degeneracy, matrix C is nonsingular with rank **n1**.

APPENDIX C

Example of the Method

For the circuit of Fig. 3 (the example of C. Pottle[9]), nodes 1, 2, and 3 are placed in the **n1** section, and node 4 is placed in the **n5** section. Thus, by inspection

$$\begin{bmatrix} C_1 + C_4 & 0 & 0 & -C_4 \\ 0 & C_2 & 0 & 0 \\ 0 & 0 & C_3 & 0 \\ -C_4 & 0 & 0 & C_4 \end{bmatrix} \begin{bmatrix} \dot{v}_1 \\ \dot{v}_2 \\ \dot{v}_3 \\ \dot{v}_4 \end{bmatrix} = \begin{bmatrix} -2C_4(\dot{v}_4 - \dot{v}_1) - 2G_2(v_2 - v_1) \\ 0 \\ 0 \\ I_1 \end{bmatrix}$$

$$- \begin{bmatrix} G_2 + G_3 & -G_2 & -G_3 & 0 \\ -G_2 & G_1 + G_2 & 0 & -G_1 \\ -G_3 & 0 & G_3 - G_4 & 0 \\ 0 & -G_1 & 0 & G_1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}$$

where

$$v_4 = E;\ \dot{v}_4 = \dot{E}.$$

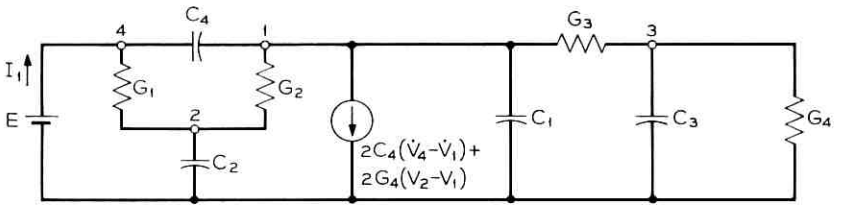


Fig. 3 — Example of Appendix C.

The derivative of the source E must be considered if a capacitor is connected to both of its nodes. Clearing out the voltage terms in the current array,

$$\begin{bmatrix} C_1 - C_4 & 0 & 0 & C_4 \\ 0 & C_2 & 0 & 0 \\ 0 & 0 & C_3 & 0 \\ -C_4 & 0 & 0 & C_4 \end{bmatrix} \begin{bmatrix} \dot{v}_1 \\ \dot{v}_2 \\ \dot{v}_3 \\ \dot{v}_4 \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ 0 \\ 0 \\ I_1 \end{bmatrix} - \begin{bmatrix} G_3 - G_2 & G_2 & -G_3 & 0 \\ -G_2 & G_1 + G_2 & 0 & -G_1 \\ -G_3 & 0 & G_3 + G_2 & 0 \\ 0 & -G_1 & 0 & G_1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}.$$

Eliminating v_4 and $\dot{v}_4$ as described in Appendix B,

$$\begin{bmatrix} C_1 - C_4 & 2C_4 & 0 \\ 0 & C_2 & 0 \\ 0 & 0 & C_3 \end{bmatrix} \begin{bmatrix} \dot{v}_1 \\ \dot{v}_2 \\ \dot{v}_3 \end{bmatrix} = \begin{bmatrix} -C_4\dot{E} \\ G_1E \\ 0 \end{bmatrix} - \begin{bmatrix} G_3 - G_2 & G_2 & -G_3 \\ -G_2 & G_1 + G_2 & 0 \\ -G_3 & 0 & G_3 + G_4 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}.$$

This is as far as we can go symbolically and as far as the method takes us. Normally all that remains is a simple inversion of the remaining capacitance matrix, but Pottle chose $C_1 = C_4$. This makes the capacitance matrix singular and so another node must be eliminated. Eliminating node 1,

$$\begin{bmatrix} C_2 + \dfrac{2C_4C_2}{G_3 - G_2} & 0 \\ \dfrac{2C_4G_3}{-G_2} & C_3 \end{bmatrix} \begin{bmatrix} \dot{v}_2 \\ \dot{v}_3 \end{bmatrix} = \begin{bmatrix} G_1E + \dfrac{C_4G_2}{G_3 - G_2}\dot{E} \\ \dfrac{C_4G_3}{G_3 - G_2}\dot{E} \end{bmatrix}$$

$$-\begin{bmatrix} G_1 + G_2 + \dfrac{G_2^2}{G_3 - G_2} & -\dfrac{G_2 G_3}{G_3 - G_2} \\ \dfrac{G_2 G_3}{G_3 - G_2} & G_3 + G_4 - \dfrac{G_3^2}{G_3 - G_2} \end{bmatrix} \begin{bmatrix} v_2 \\ v_3 \end{bmatrix}.$$

The vector

$$\begin{bmatrix} \dot{v}_2 \\ \dot{v}_3 \end{bmatrix}$$

can now be expressed explicitly.

REFERENCES

1. Bashkow, T. R., "The A Matrix, New Network Description," IRE Trans. Circuit Theory, *CT-4,* No. 3 (September 1967), pp. 117–119.
2. Bryant, P. R., "The Explicit Form of Bashkow's A Matrix," IRE Trans. Circuit Theory, *CT-9,* No. 3 (September 1962), pp. 303–306.
3. Kuh, E. S., "Stability of Linear Time-Varying Networks—The State Space Approach," IEEE Trans. Circuit Theory, *CT-12,* No. 2 (June 1965), pp. 150–157.
4. Roe, P. H., "Formulation of the State Equations for Electric and Electronic Circuit," Proc. 1963 Sixth Midwest Symp. on Circuit Theory, pp. R1–R16, University of Wisconsin, Madison, Wisc.
5. Kuh, E. S. and Rohrer, R. A., "The State-Variable Approach to Network Analysis," Proc. IEEE, *53,* No. 7 (July 1965), pp. 672–686.
6. Purslow, E. J. and Spence, R., "Order of Complexity of Active Networks," Proc. IEE, February 1967, *114,* No. 2, pp. 195–198.
7. Tow, J., "On the Order of Complexity of Linear Active Networks," Proc. IEE (London) September 1968.
8. Miller, J. A. and Newcomb, R. W., "Determination of the State-Variable Equations for Admittance Descriptions Suitable for the Computer," Proc. IEEE, *56,* No. 8 (August 1968), pp. 1372–1373.
9. Pottle, C., "State-Space Techniques for General Active Network Analysis," Chapter 7 of *System Analysis by Digital Computer,* ed. F. F. Kuo and J. F. Kaiser, New York: John Wiley and Sons, 1966.

Uniform Asymptotic Expansions for Saddle Point Integrals—Application to a Probability Distribution Occurring in Noise Theory

By STEPHEN O. RICE

(Manuscript received June 5, 1968)

The noncentral chi-square distribution occurs in noise interference problems. When the number of degrees of freedom becomes large, the middle portion of the distribution is given by the central limit theorem, and the tails by a classical saddle point expansion. Here recent work by N. Bleistein and F. Ursell on "uniform" asymptotic expansions is combined and extended to obtain an asymptotic series which apparently holds over the entire range of the distribution. General methods for expanding saddle point integrals in uniform asymptotic series are discussed. Recurrence relations are given for the coefficients in two typical cases, (i) when there are two saddle points and (ii) when there is only one saddle point but it lies near a pole or a branch point.

I. INTRODUCTION

This paper deals with the problem of obtaining asymptotic series for the complex integral

$$J = \int_{L'} t^{\lambda-1} g(t) \exp [xh(t)]\, dt \tag{1}$$

when x becomes large. Problems of this sort are quite often encountered in applied mathematics, particularly in wave propagation. The material presented here grew out of some recent work by G. H. Robertson[1] on the "Marcum Q-Function." This function, which appears in the study of radar interference, gives the distribution of the random variable (noncentral χ^2)

$$z = (1/x) \sum_{n=1}^{x} y_n^2 . \tag{2}$$

Here x is a positive integer and $y_1, y_2, \ldots, y_x$ are independent gaussian random variables with unit variances and mean values which may be different.

Mr. Robertson has devised an algorithm for computing the Q-function which may be used for a wide range of the parameters appearing in the function (that is, in the noncentral χ^2 distribution). In an earlier paper on information theory, by working with an integral of the type in equation (1), I had obtained an asymptotic (for large x) expression for the tails of the distribution.[2] However, comparison with results obtained by Robertson showed that my expression failed badly in the central part of the distribution where the central limit theorem holds.

The need for an asymptotic expansion which holds uniformly over the entire range of the distribution led to a study of the recent work on "uniform" asymptotic expansions of integrals. The first part of this paper is an exposition, plus extensions and generalizations, of some of the procedures which have been used to obtain uniform asymptotic expansions of integrals of the type in equation (1). The theory is then applied to the noncentral χ^2 distribution.

Two procedures are considered. For convenience, we call them the "Bleistein method"[3] and the "Ursell method."[4] Although these names are among the best that suggest themselves, they are not entirely satisfactory because they contain no hint of the earlier work by others, especially Olver, Chester, Friedman, and Ursell.[5, 6] Here we have recast the underlying ideas used by Bleistein and Ursell into forms better suited to our purpose.

Both methods lead to the same asymptotic series. The Bleistein method gives a compact expression for the coefficients in the expansion. However, from the few examples that have been studied, it appears that the labor required to reduce this compact expression to a computable form is at least as great as that required by the Ursell method.

Section III and Appendices A, B, and C are concerned with a preliminary change of variable in the integral J. The case, denoted by "$\lambda = 1$" for brevity, in which the exponent λ is a positive integer, is discussed in Sections IV, V and VI. This material is applied to the problem of two saddle points in Appendix E. The case in which λ is general, denoted briefly by "$\lambda \neq 1$," is discussed in Sections VII, VIII, and IX, and in the examples in Appendices F, G, and H. The results of Section IX are applied in section X to obtain the desired type of expansion for the noncentral χ^2 distribution. Useful results regarding classical saddle point

expansions are stated in Appendix D. Some of the results given in Appendix F for the general case of a saddle point near a branch point are applied in Appendix G to obtain an asymptotic series for the Poisson-Charlier polynomial, a polynomial of interest in traffic theory.

II. STATEMENT OF PROBLEM

The general problem is to obtain an asymptotic series for the integral J defined by equation (1) when x becomes large and most of the contribution to J arises from a (rather loosely defined) "critical region" around $t = 0$. The path of integration L' is supposed to start and end at $|t| = \infty$ in "valleys" in the complex t-plane where $\exp[xh(t)] \to 0$ as $|t| \to \infty$. Let the starting and ending valleys be denoted by S and E, respectively. The path L' starts in S, climbs up to and passes through the critical region, and then descends down into E.

The functions $h(t)$ and $g(t)$ are analytic in the critical region; and one or more saddle points, that is, points where $h'(t) = dh(t)/dt$ vanishes, lie in the critical region. We assume $h(0) = 0$ and that x is real and positive. If x were complex, the factor $\exp(i \arg x)$ could be included in $h(t)$.

The path L' may be deformed into a path D consisting of (*i*) paths of steepest descent which pass through some or all of the saddle points plus possibly (*ii*) loops around branch cuts and poles. The path D is independent of x. When x is extremely large, all but a negligible part of J arises from contributions of very small portions of D. If $t = 0$ is a singularity, one portion may lie close to $t = 0$. Another portion is centered on the highest (that is, largest $\exp[xh(t)]$) saddle point. If the two highest saddle points are of the same height, a portion is centered on each, and so on. Thus when x is extremely large, the asymptotic series for J may be obtained by the classical or "usual" saddle point method.

However, we may wish to compute J for values of x which, though large, are not large enough to allow J to be evaluated by the classical saddle point method. For such x's the highest saddle points and the singularity (for $\lambda \neq 1$) at $t = 0$ cannot be treated separately, that is, their interaction must be taken into account. If other saddle points of lesser height lie in the critical region, they must also be considered. This is the range of x of interest here. Our problem is to obtain the appropriate expansion of J in descending powers of x. The type of expansion we seek is shown in equation (46) for J.

This type and the type shown in equation (15) have occurred in earlier publications[3-6] and have been called "uniform" asymptotic expansions because they hold uniformly as a saddle point approaches a singularity or another saddle point.

III. CHANGE OF VARIABLE

In Appendix A it is shown that, in the critical region, $h(t)$ behaves much like a polynomial of degree $\mu + 1$ in t. Here μ is the number of saddle points in the critical region. This suggests changing the variable of integration from t to v where

$$F(v) = h(t) \tag{3}$$

and $F(v)$ is a polynomial of degree $\mu + 1$ in v. When $F(v)$ is known, solving (3) for v as a function of t gives $\mu + 1$ branches. The branch chosen for the change of variable is the one for which $dt/dv \approx c$ throughout the critical region, c being a constant. That one and only one of the $\mu + 1$ branches has this property is rendered plausible by the discussion in Appendix A.

Fortunately we do not have to solve equation (3) to obtain the asymptotic series we desire. However, for some steps we do need the values of dt/dv and higher derivatives at the saddle points. These may be obtained by repeated differentiation of (3).

$F(v)$ is not uniquely determined by $h(t)$. The factors which influence its choice are reviewed in Appendix B.

The change of variable from t to v carries the integral (1) for J into

$$J = \int_L v^{\lambda-1} f(v) \exp [xF(v)] \, dv \tag{4}$$

where

$$f(v) = g(t) \, (t/v)^{\lambda-1} t^{(1)}, \qquad t^{(1)} = dt/dv. \tag{5}$$

The path of integration L starts in the v-plane valley corresponding to valley S in the t-plane, passes through the critical region surrounding $v = 0$, then descends into the v-plane valley corresponding to valley E.

IV. THE BLEISTEIN METHOD FOR $\lambda = 1$

For the case $\lambda = 1$, the integral J becomes

$$I = \int_{L'} g(t) \exp [xh(t)]\, dt = \int_L f(v) \exp [xF(v)]\, dv, \qquad f(v) = g(t) \frac{dt}{dv} = g(t)t^{(1)}. \tag{6}$$

The Bleistein method begins by constructing a polynomial $P_0(v)$ of degree $\mu - 1$ such that $P_0(v_r) = f(v_r)$, $r = 1, 2, \ldots, \mu$ where $v_1, v_2, \ldots, v_\mu$ are the zeros, assumed simple, of $F'(v)$. By Lagrange's interpolation formula,

$$P_0(v) = \sum_{r=1}^{\mu} \frac{f(v_r)F'(v)}{(v - v_r)F''(v_r)} \tag{7}$$

where the primes denote derivatives. The polynomial may be written as

$$P_0(v) = f(v) + \frac{1}{2\pi i} \int_C \frac{f(\zeta)F'(v)\, d\zeta}{(v - \zeta)F'(\zeta)} = \frac{1}{2\pi i} \int_C f(\zeta) \left[\frac{F'(\zeta) - F'(v)}{(\zeta - v)F'(\zeta)} \right] d\zeta = \frac{1}{2\pi i} \int_C d\zeta \frac{f(\zeta)Q(\zeta, v)}{F'(\zeta)} \tag{8}$$

where $Q(\zeta, v)$ is a polynomial in v of degree $\mu - 1$,

$$Q(\zeta, v) = \frac{F'(\zeta) - F'(v)}{\zeta - v}, \tag{9}$$

and $f(v)$ has been added to remove the contribution of the pole at $\zeta = v$. The path C is taken in the counter-clockwise sense and encloses $\zeta = v$ and the zeros of $F'(\zeta)$ but no singularities of $f(\zeta)$.

The expression for $f(v)$ obtained from (8) gives

$$I = \int_L dv\, f(v) \exp [xF(v)] = \int_L dv\, P_0(v) \exp [xF(v)] + \int_L dv \exp [xF(v)] \frac{1}{2\pi i} \int_C \frac{d\zeta\, f(\zeta)F'(v)}{(\zeta - v)F'(\zeta)}. \tag{10}$$

In order to simplify interchanging the order of integration in the double integral, we cut off the tails of L in the usual fashion. The error introduced by truncation is exponentially small compared with the terms that remain. Deforming C so that it encloses the truncated L (in the sense that it encloses the point $\zeta = v$ for all v's on the truncated L), interchanging the order of integration, integrating by parts with respect to v, neglecting the contributions from the integrated portions at the ends of L, and reverting to the original order of inte-

gration carries (10) into

$$\int_L dv\, f(v) \exp [xF(v)] = \int_L dv\, P_0(v) \exp [xF(v)] + \frac{1}{x} \int_L dv \exp [xF(v)] \frac{1}{2\pi i} \int_C \frac{d\zeta\, f(\zeta)(-1)}{F'(\zeta)(\zeta - v)^2}. \tag{11}$$

Incidentally, if the contributions from the ends (say at a and b) of the truncated L were not neglected, the right side of (11) would contain the additional term

$$\left[\frac{[f(v) - P_0(v)] \exp [xF(v)]}{xF'(v)}\right]_a^b.$$

The procedure used to establish (11) can be used to show that, for any function $f_n(\zeta)$ analytic inside C, we have

$$\int_L dv\, f_n(v) \exp [xF(v)] = \int_L dv\, P_n(v) \exp [xF(v)] + \frac{1}{x} \int_L dv\, f_{n+1}(v) \exp [xF(v)] \tag{12}$$

where

$$\begin{aligned} f_{n+1}(v) &= \frac{1}{2\pi i} \int_C \frac{d\zeta\, f_n(\zeta)(-1)}{F'(\zeta)(\zeta - v)^2}, \\ P_n(v) &= \frac{1}{2\pi i} \int_C d\zeta\, f_n(\zeta) \left[\frac{Q(\zeta, v)}{F'(\zeta)}\right]. \end{aligned} \tag{13}$$

Setting $f_0(\zeta) = f(\zeta)$ and using (12) repeatedly gives

$$\begin{aligned} I &= \sum_{n=0}^{N} x^{-n} \int_L dv\, P_n(v) \exp [xF(v)] + R_N, \\ R_N &= x^{-N-1} \int_L dv\, f_{N+1}(v) \exp [xF(v)]. \end{aligned} \tag{14}$$

Since $Q(\zeta, v)$ is a polynomial of degree $\mu - 1$ in v, the same is true of $P_n(v)$ and we write

$$\begin{aligned} P_n(v) &= \sum_{l=0}^{\mu-1} p_{nl} v^l, \qquad n = 0, 1, 2, \cdots \\ I &= \sum_{l=0}^{\mu-1} U_l(x) \sum_{n=0}^{N} p_{nl} x^{-n} + R_N \end{aligned} \tag{15}$$

where

$$U_l(x) = \int_L v^l \exp [xF(v)]\, dv, \qquad l = 0, 1, \cdots, \mu - 1. \tag{16}$$

The series (15) is the type of expansion we seek. It would be desirable to have close inequalities for R_N, but none are available at the present time.

Another expression for $P_n(v)$ may be obtained from (13):

$$\begin{aligned} P_n(v) &= \frac{1}{2\pi i} \int_C \frac{d\zeta\, Q(\zeta, v)}{F'(\zeta)} \frac{1}{2\pi i} \int_{C_z} \frac{dz\, f_{n-1}(z)(-1)}{F'(z)(z - \zeta)^2} \\ &= \frac{1}{2\pi i} \int_{C_z} dz\, f_{n-1}(z) \left[\frac{1}{F'(z)} \frac{\partial}{\partial z} \frac{Q(z, v)}{F'(z)}\right] \\ &= \frac{1}{2\pi i} \int_C d\zeta\, f(\zeta) \left[\frac{1}{F'(\zeta)} \frac{\partial}{\partial \zeta}\right]^n \frac{Q(\zeta, v)}{F'(\zeta)}. \end{aligned} \tag{17}$$

In the first line C_z must enclose the point $z = \zeta$ in the z-plane in addition to the zeros of $F'(z)$. Hence initially C_z encloses C. When the order of integration is interchanged, the only singularity of the integrand in the ζ-plane lying outside C is the double pole at $\zeta = z$. Expand C until it consists of a circle of infinite radius at ∞ plus a negative loop around $\zeta = z$. The contribution of the infinite circle vanishes because the integrand is a rational function of ζ of $0(\zeta^{-3})$ at ∞. The contribution of the pole at $\zeta = z$ gives the derivative.

Notice that the coefficients p_{nl} in (15) are independent of the path L in the v-plane.

The procedure used to obtain the integral (17) for $P_n(v)$ may also be used to show that

$$f_{n+1}(v) = \frac{1}{2\pi i} \int_C d\zeta\, f(\zeta)(-1) \left[\frac{1}{F'(\zeta)} \frac{\partial}{\partial \zeta}\right]^n \left[\frac{1}{F'(\zeta)(\zeta - v)^2}\right]. \tag{18}$$

When $\mu = 1$ and $F(v) = v^2$, the polynomial $P_n(v)$ reduces to $(-1)^n f^{(2n)}(0)/(4^n n!)$ and $f_{n+1}(v)$ is equal to $[f_n(v) - vf_n'(v) - f_n(0)]/(2v^2)$.

V. COMPUTATION OF $P_n(v)$, $\lambda = 1$: BLEISTEIN METHOD

We shall regard the functions $U_l(x)$ in the series (15) for I as tabulated or easily computed. For example, when $\mu = 2$ the functions $U_0(x)$ and $U_1(x)$ may be expressed in terms of Airy functions. Then the most difficult step in applying the series is the calculation of the coefficients p_{nl}, $l = 0, 1, \ldots, \mu - 1$, of the polynomial $P_n(v)$. We

desire an expression for p_{nl} in terms of the values of the functions $g(t)$, $h(t)$ and their derivatives at the saddle points $t = t_r$, $r = 1, 2, \cdots, \mu$.

Let $t = t(v)$ denote the change of variable from t to v, and let the saddle point $v = v_r$ in the v-plane $[F'(v_r) = 0]$ correspond to t_r in the t-plane: $t_r = t(v_r)$. We shall use the notation

$$h_r^{(n)} = \left[\left(\frac{d}{dt}\right)^n h(t)\right]_{t=t_r}, \qquad g_r^{(n)} = \left[\left(\frac{d}{dt}\right)^n g(t)\right]_{t=t_r}$$
$$t_r^{(n)} = \left[\left(\frac{d}{dv}\right)^n t(v)\right]_{v=v_r}, \qquad r = 1, 2, \cdots, \mu. \tag{19}$$

When convenient, we shall write h_r for $h_r^{(0)} = h(t_r)$ and g_r for $g_r^{(0)} = g(t_r)$.

First consider the expression (7) for $P_0(v)$. As shown in Appendix B, $F(v)$ is a polynomial of degree $\mu + 1$,

$$F(v) = \sum_{j=0}^{\mu+1} A_j v^j \tag{20}$$

whose coefficients A_j may be expressed as functions of the h_r's. When this equation for $F(v)$ is used in (7), the coefficient of v^l in the resulting expression for $P_0(v)$ gives

$$p_{0l} = \sum_{j=l+2}^{\mu+1} jA_j \sum_{r=1}^{\mu} \frac{f(v_r)v_r^{j-2-l}}{F''(v_r)}, \qquad l = 0, 1, \cdots, \mu - 1. \tag{21}$$

Multiplying the right side of (21) by -1 and changing the limits of summarion for j from $l + 2$, $\mu + 1$ to 1, $l + 1$ gives another expression for p_{0l}.

Since $f(v) = g(t)t^{(1)}$, we also need an expression for $t_r^{(1)}$ in terms of $g(t)$ and $h(t)$. Differentiating $F(v) = h(t)$ twice with respect to v and using $h_r^{(1)} = 0$ leads to

$$f(v_r) = g_r t_r^{(1)}, \qquad t_r^{(1)} = [F''(v_r)/h_r^{(2)}]^{\frac{1}{2}}. \tag{22}$$

The sign of the square root is chosen to agree with the constant c in $t \approx cv$, the form assumed by the change of variable throughout the critical region.

Since the A_j's and v_r's may be expressed in terms of the h_r's, equations (21) and (22) show that p_{0l} depends only on the h_r's, g_r's and $h_r{}^{(2)}$'s.

When n is general, an expression for p_{nl} similar to (21) for p_{0l} may be obtained by expanding the derivative in the integral (17) for $P_n(v)$ in partial factions and then using the Cauchy integral theorem. For

$n = 1$ it is found that

$$p_{1l} = \sum_{j=l+2}^{\mu+1} jA_j \sum_{r=1}^{\mu} \sum_{s=1}^{\mu}{}' \frac{1}{F''(v_s)F''(v_r)} \cdot v_r^{j-2-l}\left[\frac{f(v_r) - f(v_s)}{(v_s - v_r)^2} + \frac{f'(v_r)}{v_s - v_r} + \frac{f''(v_r)}{2}\right] \tag{23}$$

where the prime on $\sum'$ denotes that the term for $s = r$ is omitted. The primes on $f(v)$ and $F(v)$ denote derivatives with respect to v.

The expression obtained for $P_n(v)$ is of the form

$$P_n(v) = \sum_{r=1}^{\mu} \sum_{m=1}^{2n+1} \alpha_{r,m}^{(n)} f^{(m-1)}(v_r)/(m-1)! \tag{24}$$

where $\alpha_{r,m}^{(n)}$ is a polynomial in v. Recurrence relations for the α's may be obtained with the help of the partial fraction expansion

$$\frac{(\zeta - v_r)^{-m}}{F'(\zeta)} = \sum_{s=1}^{\mu}{}' \frac{(v_s - v_r)^{-m}}{F''(v_s)}\left[\frac{1}{\zeta - v_s} - \sum_{q=0}^{m} \frac{(v_s - v_r)^q}{(\zeta - v_r)^{q+1}}\right]. \tag{25}$$

The relation $\sum' [1/F''(v_s)] = -1/F''(v_r)$ can be used to simplify the coefficient of $(\zeta - v_r)^{-m-1}$.

The mth derivative of $f(v)$ evaluated at v_r,

$$f^{(m)}(v_r) = \sum_{j=0}^{m} \binom{m}{j} t_r^{(j+1)}\left[\left(\frac{d}{dv}\right)^{m-j} g(t)\right]_{t=t_r} \tag{26}$$

contains derivatives of $t(v)$. They may be obtained by extending the method used to get $t_r^{(1)}$. Straightforward differentiation of $F(v) = h(t)$ with respect to v leads to

$$\begin{aligned} t_r^{(2)} &= \frac{1}{3t_r^{(1)}h_r^{(2)}}\,[F^{(3)} - h^{(3)}t^{(1)3}]_r \\ t_r^{(3)} &= \frac{1}{4t_r^{(1)}h_r^{(2)}}\,[F^{(4)} - 6h^{(3)}t^{(2)}t^{(1)2} - h^{(4)}t^{(1)4} - 3h^{(2)}t^{(2)2}]_r \end{aligned} \tag{27}$$

where the nth derivative $F^{(n)}$ of $F(v)$ is evaluated at v_r and $h^{(n)}$, $t^{(n)}$ are evaluated at t_r.

The values of $t^{(j+1)}$ for larger j's may be obtained with the help of equation (94), namely

$$F^{(n)}(v) = \sum_{k=1}^{n} h^{(k)}(t)c_{n,k} \tag{28}$$

where $c_{n,1} = t^{(n)}$, $c_{n,n} = t^{(1)n}$ and the remaining c's are given by the

recurrence relation (96). Setting $k = 1$ in (96) gives

$$c_{n,2} = nt^{(n-1)}t^{(1)} + \sum_{m=2}^{n-2} \binom{n-1}{m} t^{(n-m)}t^{(m)} \tag{29}$$

for $n \geqq 3$, the last summation being omitted when $n = 3$. The term in (28) for $k = 1$ vanishes when $v = v_r$, $t = t_r$. Substituting for $c_{n,2}$ its value given by (29) and solving for $t^{(n-1)}$ leads to the desired result when $n \geqq 3$:

$$t_r^{(n-1)} = \frac{1}{nt_r^{(1)}h_r^{(2)}} \left[F^{(n)} - \sum_{k=3}^{n} h^{(k)}c_{n,k} - h^{(2)} \sum_{m=2}^{n-2} \binom{n-1}{m} t^{(n-m)}t^{(m)} \right]_r. \tag{30}$$

The value of $F^{(n)}(v)$ is zero for $n > \mu + 1$.

VI. COMPUTATION OF $P_n(v)$, $\lambda = 1$: URSELL METHOD

The Ursell method avoids the evaluation of the derivatives of $f(v)$ which appear in equation (24) for $P_n(v)$. Instead, it makes use of classical saddle point expansions about the individual saddle points in the t and v planes.

Let μ different paths of integration, L_1', L_2', ..., L_μ' be chosen in (6) such that the chief contributions (as $x \to \infty$) along the paths corresponding to r, namely L_r' in the t-plane and its mate L_r in the v-plane, occur at the saddle points $t = t_r$ and $v = v_r$, respectively. Let the classical asymptotic expansions around t_r and v_r be

$$I_r = \int_{L'_r} g(t) \exp [xh(t)]\, dt \sim \exp [xh(t_r)] \sum_{n=0}^{\infty} \alpha_{rn} x^{-n-\frac{1}{2}} \tag{31}$$

$$[U_l(x)]_r = \int_{L_r} v^l \exp [xF(v)]\, dv \sim \exp [xF(v_r)] \sum_{m=0}^{\infty} \beta_{rlm} x^{-m-\frac{1}{2}}. \tag{32}$$

Using $h(t_r) = F(v_r)$, substituting (31) and (32) in the uniform asymptotic expansion (15) with $N = \infty$, and equating coefficients of $x^{-n-\frac{1}{2}}$ gives

$$\begin{aligned} \alpha_{rn} &= \sum_{l=0}^{\mu-1} \sum_{m=0}^{n} \beta_{rlm} p_{n-m,l} \\ &= \sum_{l=0}^{\mu-1} \beta_{rl0} p_{nl} + \sum_{l=0}^{\mu-1} \sum_{m=1}^{n} \beta_{rlm} p_{n-m,l} \end{aligned} \tag{33}$$

where the second sum in the last line is omitted when $n = 0$. The expression (17) for $P_n(v)$ shows that $P_n(v)$ remains the same, irrespective of the path of integration L, as long as $f(v)$ and $F(v)$, that is, $g(t)$ and $h(t)$, remain the same. Hence, for $n = 0$ and $r = 1, 2, \ldots, \mu$, (33) furnishes μ simultaneous linear equations which may be solved for p_{0l}, $l =$

$0, 1, \cdots, \mu - 1$. Similarly when $n = 1$, (33) determines the p_{1l}'s, and so on. It turns out (see equation 39) that $\beta_{rl0} = v_r^l \beta_{r00}$. This allows us to write the simultaneous equations in the form

$$\begin{aligned} P_0(v_r) &= \alpha_{r0}/\beta_{r00}, \qquad r = 1, 2, \cdots, \mu \\ P_n(v_r) &= \frac{\alpha_{rn}}{\beta_{r00}} - \sum_{m=1}^{n} \sum_{l=0}^{\mu-1} \left(\frac{\beta_{rlm}}{\beta_{r00}}\right) p_{n-m,l} . \end{aligned} \tag{34}$$

Expressions for α_{rn} and β_{rlm} may be obtained from the classical saddle point asymptotic expansion (103) given in Appendix D. Changing n to j in order to agree with the notation of Appendix D gives

$$\alpha_{rj} = (\pi)^{\frac{1}{2}} \left[\frac{-2}{h_r^{(2)}}\right]^{j+\frac{1}{2}} \sum_{n=0}^{2j} \frac{g_r^{(2j-n)}}{(2j-n)!} \sum_{m=0}^{n} b_{mn} \left(\tfrac{1}{2}\right)_{m+j} \tag{35}$$

where $h_r^{(n)}$, $g_r^{(n)}$ are the derivatives defined in equations (19), and $(x)_0 = 1$, $(x)_n = x(x+1) \ldots (x+n-1)$. The b_{mn}'s are computed from the recurrence relation (100), namely

$$b_{m+1,n+1} = \frac{1}{n+1} \sum_{k=1}^{n-m+1} k a_k b_{m,n-k+1} ,$$

starting with $b_{00} = 1$ and using

$$a_k = -\frac{2h_r^{(k+2)}}{(k+2)!\, h_r^{(2)}}, \qquad k = 1, 2, \cdots . \tag{36}$$

The value of $\arg [-2/h_r^{(2)}]^{\frac{1}{2}}$ is equal to $\arg (t - t_r)$ on the portion of L_r' (deformed into a path of steepest descent through t_r) leaving t_r.

Similarly,

$$\beta_{rlj} = (\pi)^{\frac{1}{2}} \left[\frac{-2}{F_r^{(2)}}\right]^{j+\frac{1}{2}} \sum_{n=0}^{2j} \frac{(-1)^{2j-n}(-l)_{2j-n} v_r^{l-2j+n}}{(2j-n)!} \sum_{m=0}^{n} b_{mn} \left(\tfrac{1}{2}\right)_{m+j} \tag{37}$$

where now the b_{mn}'s are computed from (100) with

$$a_k = \frac{-2F_r^{(k+2)}}{(k+2)!\, F_r^{(2)}}, \qquad F_r^{(n)} = \left[\left(\frac{d}{dv}\right)^n F(v)\right]_{v=v_r} . \tag{38}$$

Setting $j = 0$ in (35) and (37) gives

$$\begin{aligned} \alpha_{r0} &= (\pi)^{\frac{1}{2}} \left[\frac{-2}{h_r^{(2)}}\right]^{\frac{1}{2}} g_r^{(0)}, \qquad \beta_{rl0} = (\pi)^{\frac{1}{2}} \left[\frac{-2}{F_r^{(2)}}\right]^{\frac{1}{2}} v_r^l , \\ P_0(v_r) &= \frac{\alpha_{r0}}{\beta_{r00}} = \left[\frac{F_r^{(2)}}{h_r^{(2)}}\right]^{\frac{1}{2}} g_r^{(0)} = t_r^{(1)} g_r^{(0)} = f(v_r) \end{aligned} \tag{39}$$

where $t_r^{(1)}$ and $f(v_r)$ are the same as in equation (22). The relation

$P_0(v_r) = f(v_r)$ is the starting point for the Lagrange interpolation formula (7) in the Bleistein method.

Setting $n = 1$ in (34) and $j = 1$ in (35) and (37) leads to

$$\begin{aligned} P_1(v_r) &= \frac{\alpha_{r1}}{\beta_{r00}} - \sum_{l=0}^{\mu-1} \frac{\beta_{rl1}}{\beta_{r00}} p_{0l} \\ \frac{\alpha_{r1}}{\beta_{r00}} &= \left[\frac{-2t_r^{(1)}}{h_r^{(2)}}\right]\left\{\frac{g^{(2)}}{4} - \frac{g^{(1)}h^{(3)}}{4h^{(2)}} + g^{(0)}\left[\frac{-h^{(4)}}{16h^{(2)}} + \frac{5h^{(3)2}}{48h^{(2)2}}\right]\right\}_r \\ \frac{\beta_{rl1}}{\beta_{r00}} &= \left[\frac{-2}{F_r^{(2)}}\right]\left\{\frac{l(l-1)}{4} v^{l-2} - \frac{lv^{l-1}F^{(3)}}{4F^{(2)}} + v^l\left[\frac{-F^{(4)}}{16F^{(2)}} + \frac{5F^{(3)2}}{48F^{(2)2}}\right]\right\}_r \end{aligned} \tag{40}$$

where the subscript r on the braces indicates that the enclosed g's, h's, v's, F's have the subscript r.

VII. THE BLEISTEIN METHOD FOR $\lambda \neq 1$

Here we deal with

$$\begin{aligned} J &= \int_{L'} t^{\lambda-1} g(t) \exp[xh(t)]\, dt = \int_L v^{\lambda-1} f(v) \exp[xF(v)]\, dv \\ f(v) &= g(t)(t/v)^{\lambda-1} t^{(1)}, \qquad t^{(1)} = dt/dv. \end{aligned} \tag{41}$$

The origin is now a singularity, and its vicinity may contribute to J just as the vicinities of the saddle points do. Accordingly, we now require that the polynomial $P_0(v)$ be such that $P_0(0) = f(0)$ in addition to $P_0(v_r) = f(v_r)$, $r = 1, 2, \ldots, \mu$. Assume for the moment that $F'(0) \neq 0$, that is, that the origin is not a saddle point. Starting with Lagrange's interpolation formula and proceeding as in Section IV gives

$$\begin{aligned} P_0(v) &= \frac{f(0)vF'(v)}{vF'(0)} + \sum_{r=1}^{\mu} \frac{f(v_r)vF'(v)}{(v - v_r)v_rF''(v_r)} \\ &= f(v) + \frac{1}{2\pi i} \int_C \frac{f(\zeta)vF'(v)\, d\zeta}{(v - \zeta)\zeta F'(\zeta)} \\ &= \frac{1}{2\pi i} \int_C \frac{f(\zeta)Q(\zeta, v)\, d\zeta}{\zeta F'(\zeta)} \\ Q(\zeta, v) &= \frac{\zeta F'(\zeta) - vF'(v)}{\zeta - v} \end{aligned} \tag{42}$$

where C encloses $\zeta = v$, $\zeta = 0$, $\zeta = v_r$, $r = 1, 2, \ldots, \mu$ but no singularities of $f(\zeta)$. Here $P_0(v)$ and $Q(\zeta, v)$ are polynomials of degree μ instead of $\mu - 1$.

When the origin is a saddle point, $P_0(v)$ is still given by the expressions in (42) which contain integrals. In fact, we have $P_0'(0) = f'(0)$ (with primes denoting derivatives) in addition to $P_0(0) = f(0)$.

Much as in Section IV, we obtain

$$\int_L dv\, v^{\lambda-1} f_n(v) \exp [xF(v)]$$

$$= \int_L dv\, v^{\lambda-1} P_n(v) \exp [xF(v)] + \frac{1}{x} \int_L dv\, v^{\lambda-1} f_{n+1}(v) \exp [xF(v)] \tag{43}$$

where $f_0(v) = f(v)$ and

$$\begin{aligned} f_{n+1}(v) &= \frac{1}{2\pi i} \int_C \frac{d\zeta\, f_n(\zeta)(-1)}{\zeta F'(\zeta)} \frac{(\lambda\zeta - \lambda v + v)}{(\zeta - v)^2}, \\ P_n(v) &= \frac{1}{2\pi i} \int_C \frac{d\zeta\, f_n(\zeta) Q(\zeta, v)}{\zeta F'(\zeta)}. \end{aligned} \tag{44}$$

Equations (43) and (44) lead to the desired series for J:

$$\begin{aligned} J &= \sum_{n=0}^{N} x^{-n} \int_L dv\, v^{\lambda-1} P_n(v) \exp [xF(v)] + R_N, \\ R_N &= x^{-N-1} \int_L dv\, v^{\lambda-1} f_{N+1}(v) \exp [xF(v)]\, dv. \end{aligned} \tag{45}$$

When $P_n(v)$ is written out we get

$$\begin{aligned} P_n(v) &= \sum_{l=0}^{\mu} p_{nl} v^l, \qquad n = 0, 1, 2, \cdots \\ J &= \sum_{l=0}^{\mu} V_l(x) \sum_{n=0}^{N} p_{nl} x^{-n} + R_N, \\ V_l(x) &= \int_L v^{l+\lambda-1} \exp [xF(v)]\, dv, \qquad l = 0, 1, 2, \cdots, \mu. \end{aligned} \tag{46}$$

Furthermore, the recurrence relation (44) for $f_n(v)$ leads to

$$\begin{aligned} P_n(v) &= \frac{1}{2\pi i} \int_C \frac{d\zeta\, f_{n-1}(\zeta)}{\zeta F'(\zeta)} \left[\zeta^{\lambda} \frac{\partial}{\partial \zeta} \frac{\zeta^{-\lambda} Q(\zeta, v)}{F'(\zeta)} \right] \\ &= \frac{1}{2\pi i} \int_C d\zeta\, f(\zeta) \zeta^{\lambda-1} \left[\frac{1}{F'(\zeta)} \frac{\partial}{\partial \zeta} \right]^n \frac{Q(\zeta, v) \zeta^{-\lambda}}{F'(\zeta)} \\ f_{n+1}(v) &= \frac{1}{2\pi i} \int_C d\zeta\, f(\zeta)(-1) \zeta^{\lambda-1} \left[\frac{1}{F'(\zeta)} \frac{\partial}{\partial \zeta} \right]^n \frac{(\lambda\zeta - \lambda v + v) \zeta^{-\lambda}}{(\zeta - v)^2 F'(\zeta)}. \end{aligned} \tag{47}$$

When $\lambda = 1$ the formulas of this section do not reduce to those of Section IV since they contain the additional condition $P_n(0) = f(0)$. However, (46) gives the same series for J as (14) does for I because $V_\mu(x)$ can now be expressed as a linear combination of $V_0(x), \ldots, V_{\mu-1}(x)$ [which become $U_0(x), \ldots, U_{\mu-1}(x)$].

The only singularities enclosed by C in the integral (47) for $P_n(v)$ are poles at $\zeta = 0$ and at $\zeta = v_r$, $r = 1, 2, \ldots, \mu$. Evaluating the integral by Cauchy's theorem gives the coefficients in $P_n(v)$ as the sum of derivatives of $f(v)$ at $v = 0$ and at $v = v_r$. The derivatives at the saddle points may be obtained by differentiating

$$\ln f(v) = \ln g(t) + (\lambda - 1) \ln \frac{t}{v} + \ln t^{(1)} \tag{48}$$

with respect to v and using the expressions for $t_r^{(n)}$ developed in Section V. The derivatives $f^{(n)}(0)$ may be computed with the help of the series

$$t/v = t_0^{(1)} + \frac{v}{2!} t_0^{(2)} + \frac{v^2}{3!} t_0^{(3)} + \cdots \tag{49}$$

where $t_0^{(n)}$ denotes the nth derivative of t with respect to v at $v = 0$. The $t_0^{(n)}$'s may be obtained by differentiating $F(v) = h(t)$ repeatedly with respect to v and then setting $v = 0$. If $F'(0) \neq 0$,

$$\begin{aligned} t_0^{(1)} &= \frac{F_0^{(1)}}{h_0^{(1)}} \\ t_0^{(2)} &= \frac{F_0^{(2)} - h_0^{(2)} t_0^{(1)2}}{h_0^{(1)}} \end{aligned} \tag{50}$$

where the subscript 0 refers to $t = 0$ when it is on h and to $v = 0$ when it is on F. Higher order derivatives may be computed by using the results of Appendix C in much the same way as in Section V.

It may be verified that

$$\begin{aligned} f(0) &= g_0^{(0)} t_0^{(1)\lambda} \\ f^{(1)}(0) &= t_0^{(1)\lambda} \left[g_0^{(1)} t_0^{(1)} + \frac{(\lambda + 1)}{2} \frac{g_0^{(0)} t_0^{(2)}}{t_0^{(1)}} \right] \end{aligned} \tag{51}$$

where $g_0^{(n)}$ is the nth derivative of $g(t)$ with respect to t evaluated at $t = 0$.

VIII. THE URSELL METHOD FOR $\lambda \neq 1$

When the origin is not a saddle point, the $\mu + 1$ linear equations to be solved for the coefficients p_{nl}, $l = 0, 1, \ldots, \mu$ in $P_n(v)$ turn out to be

$$P_n(v_r) = \frac{\alpha_{rn}}{\beta_{r00}} - \sum_{m=1}^{n} \sum_{l=0}^{\mu} \left(\frac{\beta_{rlm}}{\beta_{r00}}\right) p_{n-m,l}\,, \qquad r = 1, 2, \cdots, \mu \tag{52}$$

$$p_{n0} = \frac{\alpha_{0n}}{\beta_{000}} - \sum_{q=1}^{n} \sum_{l+m=q} \left(\frac{\beta_{0lm}}{\beta_{000}}\right) p_{n-q,l} \tag{53}$$

where the summations are omitted when $n = 0$. The summation condition $l + m = q$ in (53) is also subject to $0 \leqq l \leqq \mu$, $0 \leqq m \leqq \infty$.

Equations (52) are given by the analysis of Section VI for the case $\lambda = 1$ when $g(t)$ is replaced by $t^{\lambda-1}g(t)$, v^l by $v^{l+\lambda-1}$, I_r by J_r, and $U_l(x)$ by $V_l(x)$. The α's and β's in the rth equation of (52) are the coefficients in the classical saddle point expansions about t_r and v_r:

$$\begin{aligned} J_r &= \int_{L_r'} t^{\lambda-1}g(t) \exp\,[xh(t)]\, dt \sim \exp\,[xh(t_r)] \sum_{n=0}^{\infty} \alpha_{rn} x^{-n-\frac{\lambda}{2}} \\ [V_l(x)]_r &= \int_{L_r} v^{l+\lambda-1} \exp\,[xF(v)]\, dv \sim \exp\,[xF(v_r)] \sum_{m=0}^{\infty} \beta_{rlm} x^{-m-\frac{l+\lambda}{2}} \end{aligned} \tag{54}$$

The α_{ri} in (52) (with j for n) is given by equation (35) for α_{ri} with $g_r^{(2i-n)}$ replaced by $\hat{g}_r^{(2i-n)}$ where

$$\begin{aligned} \hat{g}(t) &= t^{\lambda-1}g(t) = \sum_{n=0}^{\infty} \frac{(t - t_r)^n}{n!}\, \hat{g}_r^{(n)} \\ \hat{g}_r^{(n)} &= t_r^{\lambda} \sum_{k=0}^{n} \binom{n}{k} g_r^{(n-k)} (-1)^k (1 - \lambda)_k t_r^{-k-1}. \end{aligned} \tag{55}$$

The β_{rlj} in (52) is given by equation (37) for β_{rlj} with l replaced by $l + \lambda - 1$ on the right side.

Equation (53) arises from a consideration of the region around the singularity at the origin. As described in connection with equation (106) in Appendix D, let L_0' be a loop enclosing the branch cut running out from $t = 0$, and let L_0 be its mate in the v-plane. Then, as $x \to \infty$, the α's and β's in (53) are defined by

$$\begin{aligned} J_0 &= \int_{L_0'} t^{\lambda-1}g(t) \exp\,[xh(t)]\, dt \sim \sum_{n=0}^{\infty} \alpha_{0n} x^{-n-\lambda} \\ [V_l(x)]_0 &= \int_{L_0} v^{l+\lambda-1} \exp\,[xF(v)]\, dv \sim \sum_{m=0}^{\infty} \beta_{0lm} x^{-m-l-\lambda}. \end{aligned} \tag{56}$$

Substituting (56) in the uniform asymptotic expansion for J_0 given by (46) and equating coefficients of $x^{-n-\lambda}$ gives (53).

Using the asymptotic series (106) to determine α_{0n} and β_{rlm} leads to

$$\frac{\alpha_{0j}}{\beta_{000}} = t_0^{(1)\lambda} \left(\frac{-1}{h_0^{(1)}}\right)^j \sum_{n=0}^{j} \frac{g_0^{(j-n)}}{(j-n)!} \sum_{m=0}^{n} b_{mn} (\lambda)_{m+j} \tag{57}$$

where the subscript 0 refers to the origin and b_{mn} is computed from (100) with a_k given by

$$a_k = -h_0^{(k+1)}/[(k+1)!\, h_0^{(1)}]. \tag{58}$$

The value of $t_0^{(1)}$ is the value of dt/dv at $v = 0$ determined by the change of variable from t to v. Similarly,

$$\frac{\beta_{0lj}}{\beta_{000}} = \left(\frac{-1}{F_0^{(1)}}\right)^{l+j} \sum_{m=0}^{j} b_{mj}(\lambda)_{m+l+j} \tag{59}$$

where, replacing j by n, b_{mn} is computed from (100) with

$$a_k = -F_0^{(k+1)}/[(k+1)!\, F_0^{(1)}], \qquad k \geqq 1. \tag{60}$$

In Appendix F the theory which has just been developed is applied to the case of one saddle point ($\mu = 1$, $\lambda \neq 1$).

So far in this section it has been assumed that the origin is not a saddle point. Now let the μth saddle point coincide with the origin so that $F_0^{(1)}$ and $h_0^{(1)}$ vanish. The $\mu + 1$ equations determining p_{nl}, $l = 0, 1, \ldots, \mu$ are now

$$P_n(v_r) = \frac{\alpha_{rn}}{\beta_{r00}} - \sum_{m=1}^{n} \sum_{l=0}^{\mu} \left(\frac{\beta_{rlm}}{\beta_{r00}}\right) p_{n-m,l} \qquad r = 1, 2, \cdots, \mu - 1 \tag{61}$$

$$p_{n0} = \frac{\alpha_{0,2n}}{\beta_{000}} - \sum_{q=1}^{n} \sum_{m+l=2q} \left(\frac{\beta_{0lm}}{\beta_{000}}\right) p_{n-q,l} \tag{62}$$

$$p_{n1} = \frac{1}{\beta_{010}} \left[\alpha_{0,2n+1} - \beta_{001} p_{n0} - \sum_{q=1}^{n} \sum_{m+l=2q+1} \beta_{0lm} p_{n-q,l}\right] \tag{63}$$

where the summations are omitted when $n = 0$. The values of l and m occurring in the inner summations in (62) and (63) must also satisfy $0 \leqq l \leqq \mu$ and $0 \leqq m \leqq \infty$.

Equations (61) are the same as (52) except that r runs from 1 to $\mu - 1$ instead of from 1 to μ. The α's and β's in (62) and (63) are the coefficients in the asymptotic expansions

$$\begin{aligned} J_0 &= \int_{L'0} t^{\lambda-1} g(t) \exp[xh(t)]\, dt \sim \sum_{j=0}^{\infty} \alpha_{0j} x^{-(j+\lambda)/2} \\ [V_l(x)]_0 &= \int_{L0} v^{l+\lambda-1} \exp[xF(v)]\, dv \sim \sum_{m=0}^{\infty} \beta_{0lm} x^{-(m+l+\lambda)/2} \end{aligned} \tag{64}$$

where, as discussed in connection with equation (109), the paths L_0', L_0 coincide with the paths of steepest descent through $t = 0$, $v = 0$ except for indentations at those points.

Equation (62) is obtained by substituting (64) in the uniform asymptotic series for J_0 given by (46) and equating coefficients of $x^{-(2n+\lambda)/2}$. Equating coefficients of $x^{-(2n+1+\lambda)/2}$ gives equation (63).

Some results for the case of two saddle points, one of which is at the origin, are stated in Appendix H.

IX. SIMPLE POLE AT THE ORIGIN

When there is a simple pole at $t = 0$ and one saddle point in the critical region, a case discussed briefly by Bleistein,[3] we have

$$J = \int_{L'} t^{-1} \exp [xh(t)]\, dt. \tag{65}$$

In the critical region, L' is assumed to coincide with the linear path running from $\alpha - i\infty$ to $\alpha + i\infty$, $\alpha > 0$.

Let $h(t)$ be real when t is real and in the critical region. Let the saddle point t_1 lie on the real axis. As usual, $h_0 = 0$, $h_1^{(1)} = 0$; and we assume $h_1 \leqq 0$, $h_1^{(2)} > 0$. As suggested by example (*i*) of Appendix 3, we choose

$$F(v) = v^2 - 2v_1 v, \qquad v_1^2 = -h_1$$

where v_1 is real. We write

$$v_1 = \frac{t_1}{|\, t_1 \,|}(-h_1)^{\frac{1}{2}}, \qquad (-h_1)^{\frac{1}{2}} \geqq 0$$

in order to make v_1 and t have the same sign.

Equation (46) shows that the uniform asymptotic expansion for J has the form

$$J \sim V_0(x) \sum_{n=0}^{\infty} p_{n0} x^{-n} + V_1(x) \sum_{n=0}^{\infty} p_{n1} x^{-n} \tag{66}$$

where, with L parallel to, and to the right of, the imaginary v-axis,

$$V_0(x) = \int_L v^{-1} \exp [xF(v)]\, dv = i\pi[1 - \operatorname{erf} (v_1 x^{\frac{1}{2}})]$$

$$V_1(x) = \int_L \exp [xF(v)]\, dv = i(\pi/x)^{\frac{1}{2}} \exp (-xv_1^2)$$

$$\operatorname{erf} (z) = \frac{2}{\pi^{\frac{1}{2}}} \int_0^z \exp (-t^2)\, dt.$$

Putting $\lambda = 0$ in the integral (47) for $P_n(v)$ gives

$$P_n(v) = p_{n0} + p_{n1} v$$

$$= \frac{1}{2\pi i}\int_C \frac{d\zeta\, f(\zeta)}{2^n\zeta}\left(\frac{1}{\zeta - v_1}\frac{\partial}{\partial\zeta}\right)^n\left(\frac{\zeta + v - v_1}{\zeta - v_1}\right) \tag{67}$$

$$f(v) = vt^{(1)}/t, \qquad t^{(1)} = dt/dv$$

where C encloses $\zeta = 0$ and $\zeta = v_1$ but no singularities of $f(\zeta)$. Setting $v = 0$ in (67) gives

$$P_0(0) = p_{00} = f(0) = 1,$$

$$P_n(0) = p_{n0} = 0, \qquad n > 0.$$

Here the series (49) for t/v has been used to show that $f(0) = 1$. Therefore the series for J reduces to

$$J \sim i\pi\{1 - \operatorname{erf}[v_1x^{\frac{1}{2}}]\} + i(\pi/x)^{\frac{1}{2}}\exp(-xv_1^2)\sum_{n=0}^{\infty} p_{n1}x^{-n}. \tag{68}$$

Setting $v = v_1$ in (67) gives

$$p_{01} = [f(v_1) - 1]/v_1 = \frac{t_1^{(1)}}{t_1} - \frac{1}{v_1}$$

$$p_{n1} = \frac{1}{2\pi i}\int_C \frac{d\zeta\, f(\zeta)}{2^n v_1\zeta}\left(\frac{1}{\zeta - v_1}\frac{\partial}{\partial\zeta}\right)^n\left(\frac{\zeta}{\zeta - v_1}\right), \qquad n > 0. \tag{69}$$

From (22) and $F''(v) = 2$ it follows that $t_1^{(1)} = [2/h_1^{(2)}]^{\frac{1}{2}}$. The integral for p_{n1} may be evaluated in terms of the $2n$th derivative of $f(v)/v = t^{(1)}/t = (d/dv)\ln t(v)$. Thus, writing $\zeta/(\zeta - v_1)$ as $1 + v_1(\zeta - v_1)^{-1}$ and using

$$\left(\frac{1}{\zeta - v_1}\frac{\partial}{\partial\zeta}\right)^n(\zeta - v_1)^{-1} = (-1)^n 1\cdot 3\ldots(2n-1)(\zeta - v_1)^{-2n-1}, \; n > 0$$

leads to

$$p_{n1} = \frac{(-1)^n(\frac{1}{2})_n}{v_1^{2n+1}}\left\{\frac{v_1^{2n+1}}{(2n)!}\left[\left(\frac{d}{dv}\right)^{2n}\frac{t^{(1)}}{t}\right]_{v=v_1} - 1\right\}$$

The first of the p_{n1}'s required in the series (68) for J is given by equation (69) for p_{01}. The remaining ones may be obtained by using the Ursell method equation (52). Since $p_{00} = 1$ and $p_{n0} = 0$ for $n > 0$, equation (52) gives for $n > 0$

$$p_{n1} = \frac{1}{v_1\beta_{100}}\left[\alpha_{1n} - \beta_{10n} - \sum_{m=1}^{n}\beta_{11m}p_{n-m,1}\right]. \tag{70}$$

From equation (55), $\hat{g}(t) = t^{-1}$ and its nth derivative at t_1 is $\hat{g}_1^{(n)} = (-1)^n n! t_1^{-n-1}$. Replacing g by $\hat{g}$ in (35) leads to

$$\alpha_{1j} = (\pi)^{\frac{1}{2}}\left[\frac{-2}{h_1^{(2)}}\right]^{j+\frac{1}{2}} \sum_{n=0}^{j} (-1)^n t_1^{-2j+n-1} \sum_{m=0}^{n} b_{mn}(\tfrac{1}{2})_{m+j}$$

where $\arg [-2/h_1^{(2)}]^{\frac{1}{2}} = \pi/2$ and the b_{mn}'s are computed from (100) with

$$a_k = -2h_1^{(k+2)}/[(k+2)!\, h_1^{(2)}], \qquad k = 1, 2, \cdots . \tag{71}$$

Equation (54) shows that the β_{1lm}'s are the coefficients in the asymptotic expansion of $[V_l(x)]_1$, that is, in the asymptotic expansion of an integral, which has the same integrand as $V_l(x)$, taken along the path of steepest descent through the saddle point $v = v_1$. Instead of obtaining the β's by the general procedure outlined in Section VIII, we notice that the "asymptotic" series for $V_1(x)$ consists of only one term. Consequently β_{11m} is 0 when $m > 0$ and the summation in equation (70) for p_{n1} disappears. Moreover, the asymptotic series for the error function gives, when $v_1 > 0$,

$$V_0(x) \sim i(\pi)^{\frac{1}{2}} \exp(-xv_1^2) \sum_{m=0}^{\infty} (-1)^m (\tfrac{1}{2})_m (xv_1^2)^{-m-\frac{1}{2}}.$$

When $v_1 > 0$ and $x \to \infty$, $V_0(x)$ is given asymptotically by the contribution from v_1. When $v_1 < 0$, the asymptotic expression for $V_0(x)$ contains the constant term $2\pi i$, but the contribution from a path of steepest descent through v_1 is still given by the same expression as for positive v_1. Hence, irrespective of the sign of v_1,

$$\beta_{10m} = i(\pi)^{\frac{1}{2}}(-1)^m (\tfrac{1}{2})_m v_1^{-2m-1}.$$

These results enable us to write equation (70) (with j for n) as

$$p_{j1} = \frac{(-1)^j(\frac{1}{2})_j}{v_1^{2j+1}} \left\{\left[\frac{v_1 t_1^{(1)}}{t_1}\right]^{2j+1} \sum_{n=0}^{2j} (-t_1)^n \sum_{m=0}^{n} b_{mn}(j+\tfrac{1}{2})_m - 1\right\} \tag{72}$$

for $j \geq 0$. Here, to repeat,

$$v_1 = \frac{t_1}{|\, t_1 \,|}(-h_1)^{\frac{1}{2}}, \quad t_1^{(1)} = \left[\frac{2}{h_1^{(2)}}\right]^{\frac{1}{2}}, \quad \frac{v_1 t_1^{(1)}}{t_1} = \frac{1}{|\, t_1 \,|}\left[\frac{-2h_1}{h_1^{(2)}}\right]^{\frac{1}{2}},$$
$$(c)_0 = 1, \qquad (c)_n = c(c+1)\cdots(c+n-1) \tag{73}$$

and the b_{mn}'s are computed in succession from equation (100) with a_k given by (71).

The first two p_{n1}'s are

$$\begin{aligned} p_{01} &= \frac{1}{v_1}\left[\frac{v_1 t_1^{(1)}}{t_1} - 1\right] \\ p_{11} &= -\frac{1}{2v_1^3}\left\{\left[\frac{v_1 t_1^{(1)}}{t_1}\right]^3 [1 - \tfrac{3}{2}t_1 a_1 + t_1^2(\tfrac{3}{2}a_2 + \tfrac{15}{8}a_1^2)] - 1\right\} \end{aligned} \tag{74}$$

and the b_{mn}'s for p_{21} may be read from the table in Appendix D.

When $| t_1 |$ is small, the expressions which have been given for p_{n1} are essentially small differences between large numbers. If the calculation is being performed on a digital computer it may be advisable to use double precision. Expanding $h(t)$ about $t = t_1$ and then setting $t = 0$ leads to a series for $-h_1$ which may be used to obtain v_1^{-2i-1} as $[t_1^{(1)}/t_1]^{2i+1}$ times a power series in t_1 . Series of this type can be used to show that

$$\begin{aligned} p_{01} &= \frac{a_1}{2} t_1^{(1)} + O(t_1) \\ p_{11} &= -\tfrac{1}{4}t_1^{(1)3}(3a_3 + \tfrac{15}{2}a_1 a_2 + \tfrac{35}{8}a_1^3) + O(t_1) \end{aligned} \tag{75}$$

where a_k is given by (71).

X. THE NONCENTRAL χ^2 DISTRIBUTION

Let x be a positive integer and $y_1, y_2, \ldots, y_x$ be independent gaussian random variables with unit variances and respective mean values $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_x$. Let z be the (noncentral χ^2) random variable

$$z = \frac{1}{x}\sum_{n=1}^{x} y_n^2 . \tag{76}$$

It may be shown that the mean value of z is $\bar{z} = 1 + r$ and that its variance is $(2 + 4r)/x$ where

$$r = \frac{1}{x}\sum_{n=1}^{x} \bar{y}_n^2 .$$

Furthermore, from Ref. 2 (with a change of variable) the distribution function of z is

$$\text{Prob}\,[0 \leqq z \leqq s] = \frac{1}{2\pi i}\int_{\hat{c}-i\infty}^{\hat{c}+i\infty} t^{-1} \exp\,[xh(t)]\,dt \tag{77}$$

where $\hat{c} > 0$ and

$$h(t) = \tfrac{1}{2}[st - \ln(1 + t) + r(1 + t)^{-1} - r]. \tag{78}$$

The integral on the right side of (77) is seen to be equivalent to minus Marcum's Q-function (and also to an expression given by R. A. Fisher) when (77) is written as

$$\text{Prob } [0 \leqq z \leqq s] = \frac{x}{2} \int_0^s (z/r)^{(x/4)-\frac{1}{2}} \exp [-x(z + r)/2] I_{(x/2)-1}[x(rz)^{\frac{1}{2}}] dz.$$

Here I denotes a Bessel function with imaginary argument.

We are interested in computing the distribution of z when x is large. The equation $h^{(1)}(t) = 0$ gives two saddle points. However, as pointed out in Ref. 2, when x is large only the one at

$$t_1 = -1 + \frac{1 + (1 + 4rs)^{\frac{1}{2}}}{2s}$$

need be considered. The value of t_1 is real and > -1. When $s = 1 + r$, t_1 is zero; and when s increases through $1 + r$, t_1 decreases through 0.

From equation (68) the desired asymptotic expansion is

$$\text{Prob } [0 \leqq z \leqq s] \sim \tfrac{1}{2}\{1 - \text{erf } [v_1 x^{\frac{1}{2}}]\} + \tfrac{1}{2}(\pi x)^{-\frac{1}{2}} \exp (-x v_1^2) \sum_{n=0}^{\infty} p_{n1} x^{-n} \qquad (79)$$

where p_{n1} is given by (72). The quantities entering p_{n1} are

$$\begin{aligned} v_1 &= \frac{t_1}{|t_1|}(-h_1)^{\frac{1}{2}}, \qquad h_1 = h(t_1), \qquad t_1^{(1)} = [2/h_1^{(2)}]^{\frac{1}{2}} \\ h^{(n)}(t) &= \tfrac{1}{2}(-1)^n (n-1)! [nr(1 + t)^{-n-1} + (1 + t)^{-n}], \qquad n \geqq 2 \\ h_1^{(2)} &= r(1 + t_1)^{-3} + 2^{-1}(1 + t_1)^{-2} \\ a_k &= \frac{2(-1)^{k+1}}{(k + 2)(1 + t_1)^k} \left[\frac{(k + 2)r + 1 + t_1}{2r + 1 + t_1}\right], \qquad k \geqq 1. \end{aligned} \qquad (80)$$

The values of p_{01} and p_{11} may be obtained from (74) by substitution of the parameter values (80).

When $s - \bar{z} = s - 1 - r$ is small, the central limit theorem in the theory of probability states that

$$\text{Prob } [0 \leqq z \leqq s] \sim \frac{1}{2}\left\{1 + \text{erf}\left[\frac{(s - \bar{z})x^{\frac{1}{2}}}{(4 + 8r)^{\frac{1}{2}}}\right]\right\}.$$

This agrees with the approximation given by the error function term in (79) when it is noted that $t_1 \approx -(s - 1 - r)/(1 + 2r)$ and $-h_1 \approx t_1^2 h_1^{(2)}/2$ if $s - 1 - r$ is small.

The ordinary χ^2 distribution is obtained by setting $r = 0$ in the noncentral χ^2 distribution. In this case we have

$$\begin{aligned} t_1 &= \frac{1-s}{s}, \qquad t_1^{(1)} = 2/s, \\ v_1 &= \frac{1-s}{|1-s|}\left(\frac{s-1-\ln s}{2}\right)^{\frac{1}{2}}, \\ a_k &= -\frac{2(-s)^k}{k+2}, \qquad k \geqq 1. \end{aligned} \tag{81}$$

Equations (74) show that the first two coefficients p_{01}, p_{11} in the asymptotic series (79) are now

$$\begin{aligned} p_{01} &= \frac{2}{1-s} - \frac{1}{v_1}, \\ p_{11} &= -\frac{1}{2}\left\{\left(\frac{2}{1-s}\right)^3\left[s + \frac{(1-s)^2}{12}\right] - \frac{1}{v_1^3}\right\}. \end{aligned} \tag{82}$$

When s is close to its average value 1, equation (75) gives

$$\begin{aligned} p_{01} &= 2/3 + O(t_1) \\ p_{11} &= 1/135 + O(t_1). \end{aligned} \tag{83}$$

Setting $x = 2c$ and $r = 0$ gives

$$\begin{aligned} \text{Prob}\,[0 \leqq z \leqq s] &= \frac{1}{\Gamma(c)}\int_0^{cs} u^{c-1}\exp(-u)\,du \\ &= \sum_{n=c}^{\infty}\frac{(cs)^n\exp(-cs)}{n!} \end{aligned} \tag{84}$$

where c is assumed to be an integer (x even) in the last equation. These relations may be combined with the foregoing formulas to obtain asymptotic results for the incomplete gamma function and the Poisson distribution.

There is reason to believe that the asymptotic expansion (79) for Prob $[0 \leqq z \leqq s]$ may hold over the entire range $0 \leqq s \leqq \infty$. For example, consider the ordinary ($r = 0$) χ^2 distribution. In this case the first two terms in (79) give

$$\begin{aligned} \text{Prob}\,[0 \leqq z \leqq s] \sim\ & \tfrac{1}{2}\{1 - \text{erf}\,[v_1 x^{\frac{1}{2}}]\} \\ & + \tfrac{1}{2}[\pi x]^{-\frac{1}{2}}\left[\frac{2}{1-s} - \frac{1}{v_1}\right]\exp(-xv_1^2) \end{aligned} \tag{85}$$

where $x = 2c$ and v_1 is given by (81). Let c be held fixed at some large value and consider further the three cases $s \to 0$, $s \to \infty$, and $s \to 1$. In all three cases it may be verified that the leading terms in Prob $[0 \leqq z \leqq s]$ given by (85) agree with those obtained from the exact equations (84) and the asymptotic properties of the incomplete gamma function. The expressions obtained from (84) are

$$\text{Prob } [0 \leqq z \leqq s] = (2\pi c)^{-\frac{1}{2}} s^c \, [\exp c] \, [1 + O(cs) + O(c^{-1})], \qquad s \to 0$$

$$\text{Prob } [0 \leqq z \leqq s] = 1 - (2\pi c)^{-\frac{1}{2}} s^{c-1} [\exp (c - cs)][1 + O(s^{-1}) + O(c^{-1})], \qquad s \to \infty \tag{86}$$

$$\text{Prob } [0 \leqq z \leqq 1] = \tfrac{1}{2} + \tfrac{1}{3}(2\pi c)^{-\frac{1}{2}} + O(c^{-1}), \qquad s = 1.$$

APPENDIX A

The Behavior of $h(t)$ in the Critical Region

In this appendix we show that, in the critical region, $h(t)$ behaves much like a polynomial of degree $\mu + 1$, and we examine the change of variable from t to v.

First write

$$h(t) = \sum_{j=1}^{\mu+1} \frac{t^j h_0^{(j)}}{j!} + R_{\mu+1} \tag{87}$$

where $h_0^{(j)}$ is the value of $(d/dt)^j h(t)$ at $t = 0$, $h_0^{(\mu+1)}$ is not 0, and $R_{\mu+1}$ is $O(t^{\mu+2})$.

One of the distinguishing features of a polynomial in t of degree $\mu + 1$ is that when t is much larger than r, where $|t| = r$ is the smallest circle which encloses the zeros, the dominant term in the polynomial is the one containing $t^{\mu+1}$. The function $h(t)$ has a corresponding property. Suppose that the saddle points $t_1, t_2, \ldots, t_\mu$ all lie within a distance ϵ of the origin, and for the moment suppose that they may be moved towards the origin so that ϵ may be made as small as we desire. Also suppose that $h_0^{(\mu+1)} = A + O(\epsilon)$ where $A \neq 0$. We shall show that by making ϵ small enough we may find a range $\rho < |t| < \eta$ throughout which

$$h(t) \approx t^{\mu+1} h_0^{\mu+1}/(\mu + 1)! \,. \tag{88}$$

Here η is such that when $|t|$ and ϵ are less than η, the remainder $R_{\mu+1}$ in (87) is negligible in comparison with the $(\mu + 1)$ term $T_{\mu+1} = t^{\mu+1} h_0^{(\mu+1)}/(\mu + 1)!$. Once η is fixed, ρ may be chosen to be arbitrarily small, subject only to $\rho < \eta$.

In order to show that (88) holds when $\rho < |t| < \eta$, notice that by

repeated differentiation of the representation

$$h'(t) = (t - t_1)(t - t_2) \ \cdots \ (t - t_\mu)\, G(t)$$

it may be shown that $G(0) = h_0^{(\mu+1)}/\mu! + O(\epsilon)$ and that $h_0^{(j)}$ is $O[\epsilon^{\mu+1-j} h_0^{(\mu+1)}]$ for $j = 1, 2, \cdots, \mu$. Hence when $|t| > \epsilon$,

$$\sum_{j=1}^{\mu} \frac{t^j h_0^{(j)}}{j!} = O[t^\mu \epsilon\, h_0^{(\mu+1)}] = O\left(\frac{\epsilon}{t} T_{\mu+1}\right).$$

Choosing ρ to be arbitrarily small, subject only to $0 < \rho < \eta$, and then choosing ϵ so that $\epsilon/\rho \ll 1$ establishes (88).

This property of $h(t)$ suggests that some insight into the change of variable from t to v, specified by $F(v) = h(t)$, may be obtained by considering $h(t)$ to be a polynomial $\varphi(t)$ of degree $\mu + 1$. Then a natural choice of $F(v)$ is $F(v) \equiv \varphi(cv + b)$ where c and b are constants. For simplicity we take $c = 1$ and $b = 0$ so that $v_r = t_r$, $r = 1, 2, \ldots, \mu$ where v_r is the rth saddle point on the v-plane. The equation $F(v) = h(t)$ goes into $\varphi(v) = \varphi(t)$ which we write as

$$\begin{aligned}\varphi(v) - \varphi(t) &= \sum_{i=1}^{\mu+1} A_i(v^i - t^i) \\ &= (v - t)[A_1 + A_2(v + t) + A_3(v^2 + vt + t^2) + \cdots] \\ &= 0.\end{aligned}$$

The branch used in the change of variable is

$$v = t \tag{89}$$

for which $dv/dt = 1$ everywhere. The remaining μ branches, which are ignored in the change of variable, may be obtained by solving

$$A_1 + A_2(v + t) + \ldots + A_{\mu+1}(v^\mu + v^{\mu-1}t + \ldots + t^\mu) = 0 \tag{90}$$

for v as a function of t. Writing (90) as

$$G(v, t) = \frac{\varphi(v) - \varphi(t)}{v - t} = 0$$

and expanding $G(v, t)$ about $v = t_r$, $t = t_r$ shows that, near $t = t_r$, one of the remaining branches behaves like $v = t_r - (t - t_r)$. On this branch $v = t_r$ and $dv/dt = -1$ at $t = t_r$. Again, let $v = \hat{v}_s$, $s = 1, 2, \ldots, \mu - 1$, be one of the $\mu - 1$ roots of $G(v, t_r) = 0$ which is not equal to t_r. Expanding $G(v, t)$ about $v = \hat{v}_s$, $t = t_r$ shows that near $t = t_r$ the corresponding branch behaves like

$$v = \hat{v}_s + \frac{(t - t_r)^2 \varphi''(t_r)}{2\varphi'(\hat{v}_s)}$$

and that $dv/dt = 0$ at $t = t_r$.

The examination of special cases suggests that there is a one-to-one correspondence between the μ branches of (90) and the μ saddle points in the sense that dv/dt for a particular branch is equal to -1 at its corresponding saddle point and is zero at the other saddle points. Thus it appears that the branch $v = t$ [or its analogue for general $h(t)$] is the only one suitable for the change of variable throughout the entire critical region.

APPENDIX B

The Choice of $F(v)$

The polynomial $F(v)$ used in changing the variable of integration will be written as

$$F(v) = \sum_{j=0}^{\mu+1} A_j v^j. \tag{91}$$

The positions $t_1, t_2, \ldots, t_\mu$ of the saddle points and the associated values $h_r = h(t_r)$ are supposed known. We require expressions for the A_j's which are either pure numbers or depend only on the h_r's. Although one or more of the zeros $v_1, v_2, \ldots, v_\mu$ of $F'(v) = dF(v)/dv$ may appear in our final expression for $F(v)$, they will always be expressed in terms of the h_r's.

Since $F(v) = h(t)$, we have the 2μ equations

$$\begin{aligned} F(v_r) &= h_r , \\ F'(v_r) &= 0, \qquad r = 1, 2, \cdots, \mu \end{aligned} \tag{92}$$

relating the $2\mu + 2$ unknowns $v_1, v_2, \ldots, v_\mu, A_0, A_1, \ldots, A_{\mu+1}$. Consequently we have at least two arbitrary choices ($A_{\mu+1} = 0$ is forbidden). For the case $\lambda \neq 1$ we shall always require the change of variable to be such that v is 0 when $t = 0$ and thus we take $A_0 = 0$. In some instances the form of $h(t)$ aids in the choice of the A_j's. For example, when $h(t)$ is an even function of t, we can take $F(v)$ to be an even function of v.

In choosing $F(v)$ it is helpful to notice that in the critical region the change of variable takes the form $t \approx cv$, c being a constant. Consequently, from (87),

$$F(v) \approx \sum_{j=1}^{\mu+1} \frac{(cv)^j h_0^{(j)}}{j!} \tag{93}$$

and it follows that the μ zeros, v_r, of $F'(v)$ have nearly the same configuration in the v-plane (except for a possible rotation and magnification given by arg c and $|c|$, respectively) as do the zeros, t_r, of $h'(t)$. For the case $\lambda = 1$ there may also be a small displacement so that $t \approx cv$ can be written more accurately as $t \approx cv + b$, or $dt/dv \approx c$. Furthermore, from (88), we can take c to be one of the roots of

$$A_{\mu+1} = c^{\mu+1} h_0^{(\mu+1)}/(\mu + 1)! \, .$$

The following examples illustrate possible choices of $F(v)$.

(*i*) $\mu = 1$, $\lambda \neq 1$, $t_1 \neq 0$. Initially there are 4 unknowns, v_1, A_0, A_1, A_2 related by 2 equations, and $F(v)$ given by

$$F(v) = A_2 v^2 + A_1 v + A_0.$$

We take $A_0 = 0$ (because $\lambda \neq 1$) and arbitrarily choose $A_2 = 1$ (for convenience). This carries the two equations into

$$v_1^2 + A_1 v_1 = h_1$$

$$2v_1 + A_1 = 0.$$

Consequently

$$F(v) = v^2 - 2v_1 v \qquad v_1^2 = -h_1.$$

This case has been considered by Bleistein.[3]

(*ii*) $\mu = 2, \lambda = 1$. Initially there are six unknowns $v_1, v_2, A_0, A_1, A_2, A_3$ related by four equations, and $F(v)$ given by

$$F(v) = A_3 v^3 + A_2 v^2 + A_1 v + A_0.$$

We take $A_2 = 0$ in order to simplify $F'(v)$. The four equations become

$$F'(v_r) = 3A_3 v_r^2 + A_1 = 0$$

$$3F(v_r) = -A_1 v_r + 3A_1 v_r + 3A_0 = 3h_r \, , \qquad r = 1, 2.$$

It follows that $v_2 = -v_1$, $A_0 = (h_1 + h_2)/2$, $A_1 v_1 = 3(h_1 - h_2)/4$.

For the remaining choice we take A_3 to be equal to $-A_1/3$ and obtain,

$$F(v) = \tfrac{1}{4}(h_2 - h_1)(v^3 - 3v) + \tfrac{1}{2}(h_2 + h_1).$$

Another choice for A_3 is $\frac{1}{3}$ which gives

$$F(v) = \tfrac{1}{3}v^3 - v_1^2 v + \tfrac{1}{2}(h_2 + h_1)$$

$$v_1^3 = \tfrac{3}{4}(h_2 - h_1), \qquad v_2 = -v_1 \, .$$

This case has been considered by Chester, Friedman and Ursell.[6]

(*iii*) $\mu = 2$, $\lambda \neq 1$, $t_1 \neq 0$, $t_2 = 0$. Here the unknowns and $F(v)$ are the same as in example (*ii*), but because of $h_2 = h(0) = 0$ it turns out we have three arbitrary choices. Since $\lambda \neq 1$ we take $A_0 = 0$. Then both $F(v_2) = h_2$ and $F'(v_2) = 0$ are satisfied by choosing $v_2 = 0$ and $A_1 = 0$. This leaves $F(v_1) = h_1$ and $F'(v_1) = 0$ to be satisfied by the remaining three unknowns v_1, A_2, A_3. Our third choice is $A_3 = 2$. It leads to

$$F(v) = 2v^3 - 3v_1v^2, \qquad v_1^3 = -h_1 .$$

(*iv*) $\mu = 2$, $\lambda \neq 1$, t_1, t_2, $\neq 0$. This case illustrates the complication encountered for the general case when $\mu \geqq 2$. The value of A_0 must be 0 and we choose $A_3 = 1$. Then

$$F(v) = v^3 + A_2v^2 + A_1v$$

$$F'(v) = 3v^2 + 2A_2v + A_1 .$$

The last equation shows that $A_2 = -3(v_1 + v_2)/2$, $A_1 = 3v_1v_2$. Substituting in $F(v_r) = h_r$ gives

$$v_1^2(-v_1 + 3v_2)/2 = h_1$$

$$v_2^2(-v_2 + 3v_1)/2 = h_2 .$$

Setting $a = h_2/h_1$ and $\rho = v_2/v_1$ leads to

$$\rho^3 - 3\rho^2 + 3\rho a - a = 0$$

which has the three roots

$$\rho_n = 1 + (1 - a)^{\frac{1}{3}}[(1 - a^{\frac{1}{3}})^{\frac{1}{3}}\omega_n + (1 + a^{\frac{1}{3}})^{\frac{1}{3}}\omega_n^*]$$

where $\omega_1 = 1$, $\omega_2 = i^{4/3}$, $\omega_3 = i^{-4/3}$ and the star denotes "conjugate complex." When t_1 and t_2 tend to zero, one of the ρ_n's tends to t_2/t_1, and this is the value of ρ to be used. The value of v_1^3 is equal to $2h_1/(3\rho - 1)$, and we have

$$A_2 = -3v_1(1 + \rho)/2, \qquad A_1 = 3\rho v_1^2.$$

(*v*) $\mu = 3$. The general case of three saddle points may be handled by a procedure similar to that used in example (*iv*). We do not discuss this case beyond mentioning that when we set $v_2 = \rho v_1$, $v_3 = \sigma v_1$ the variable $u = (\rho - 1)/(\sigma - 1)$ must satisfy the equation

$$u^4 - 2u^3 + 2au - a = 0, \qquad a = (h_1 - h_2)/(h_1 - h_3).$$

(*vi*) $\mu = 3$, $t_3 = 0$, $h(t)$ even. Since $h(t)$ is even we start with

$$F(v) = v^4 + A_2v^2 + A_0$$

and find that

$$F(v) = v^4 - 2v_1^2v^2, \qquad v_1^4 = -h_1.$$

This case has been treated by Felsen.[7]

APPENDIX C

Derivatives of Composite Functions

A result used in Section V to compute the nth derivative, $t_r^{(n)}$, of $t(v)$ at $v = v_r$ is stated in this appendix. Let the argument u in $h(u)$ be a function $t(v)$ of v. Then

$$\left(\frac{d}{dv}\right)^n h[t(v)] = \sum_{k=1}^{n} h^{(k)} c_{n,k}, \qquad n \geqq 1 \tag{94}$$

where $h^{(k)}$ stands for $(d/du)^k h(u)$ evaluated at $u = t(v)$, and the coefficients $c_{n,k}$ are computed from the recurrence relations

$$c_{n,1} = t^{(n)}, \tag{95}$$

$$c_{n+1,k+1} = \sum_{m=k}^{n} \binom{n}{m} t^{(n+1-m)} c_{m,k}, \qquad 1 \leqq k \leqq n \tag{96}$$

in which $t^{(n)}$ denotes $(d/dv)^n t(v)$ and $\binom{n}{m}$ the binomial coefficient. Equation (96) may be proved by induction. Differentiating (94) gives

$$c_{n+1,k+1} = t^{(1)} c_{n,k} + \frac{d}{dv} c_{n,k+1}, \qquad 1 \leqq k \leqq n-1. \tag{97}$$

We assume that (96) holds when n is replaced by n-1 and use it to express $c_{n,k+1}$ as a sum. Then one of the terms in the summand for $d\,c_{n,k+1}/dv$ contains $d\,c_{m,k}/dv$. From (97), assuming $k > 1$,

$$\frac{d}{dv} c_{m,k} = c_{m+1,k} - t^{(1)} c_{m,k-1}.$$

Equation (96), with $(n-1, k-1)$ for (n, k), lets us sum the terms containing $t^{(1)} c_{m,k-1}$ with respect to m. Equation (96) follows upon combining binomial coefficients and using $c_{k,k} = t^{(1)} c_{k-1,k-1}$.

The recurrence relations may also be obtained by writing the right side of (94) as a Bell polynomial and using the recurrence relation for these polynomials.[8] Expressions for the $c_{n,k}$'s (up to $n = 8$) as polynomials in the $t^{(n)}$'s may be obtained from Riordan's table of Bell polynomials given on page 49 of Ref. 8.

APPENDIX D

Formulas for Classical Saddle Point Asymptotic Expansions

A result useful in obtaining asymptotic expansions of integrals is

$$\int_0^T \tau^{\rho-1} G(\tau) \exp [xH(\tau)]\, d\tau \sim \sum_{j=0}^{\infty} \frac{1}{\nu} \left(\frac{-1}{xH_\nu}\right)^{(\rho+j)/\nu} \sum_{n=0}^{j} G_{j-n} \sum_{m=0}^{n} b_{mn} \Gamma\left(m + \frac{\rho + j}{\nu}\right) \tag{98}$$

where $x \to \infty$, $Re\ \rho > 0$, ν is a positive integer, and

$$G(\tau) = \sum_{n=0}^{\infty} \tau^n G_n\ , \qquad H(\tau) = \sum_{n=\nu}^{\infty} \tau^n H_n\ . \tag{99}$$

The b_{mn}'s depend only on $H(\tau)$ and are computed in succession, starting with $b_{00} = 1$ and $b_{0n} = 0$ for $n \geq 1$, from

$$b_{m+1,n+1} = \frac{1}{n+1} \sum_{k=1}^{n-m+1} k a_k b_{m,n-k+1}\ . \tag{100}$$

Here $a_k = -H_{\nu+k}/H_\nu$, $k = 1, 2, \ldots$.

Special values of b_{mn} are given in Table I.

The asymptotic expansion (98) is based upon the gamma function integral

$$\int_0^\infty u^{z-1} \exp(-u^\nu)\, du = \frac{1}{\nu} \Gamma\left(\frac{z}{\nu}\right) \tag{101}$$

and the expansion

$$\exp\left[y \sum_1^\infty a_n \xi^n\right] = \sum_{n=0}^{\infty} \xi^n \sum_{m=0}^{n} b_{mn} y^m. \tag{102}$$

The recurrence relation (100) may be obtained by differentiating (102) with respect to ξ, replacing the exponential by its series, and then equating coefficients of $\xi^n y^{m+1}$.

TABLE I — SPECIAL VALUES OF b_{mn}

n	b_{0n}	b_{1n}	b_{2n}	b_{3n}	b_{4n}	b_{nn}
0	1					
1	0	a_1				
2	0	a_2	$a_1^2/2$			
3	0	a_3	$a_1 a_2$	$a_1^3/6$		
4	0	a_4	$a_1 a_3 + 2^{-1} a_2^2$	$2^{-1} a_1^2 a_2$	$a_1^4/24$	
$n > 1$	0	a_n	—	—	—	$a_1^n/n!$

For an integral, say I_r in equation (31), having a simple saddle point at $t = t_r$ and a path of integration L_r' which runs up to, through, and then down from t_r, we can use (98) with $\nu = 2$, $\rho = 1$, $\tau = t - t_r$, $G(\tau) = g(t_r + \tau)$, and $H(\tau) = h(t_r + \tau) - h(t_r)$. The contribution of the saddle point is obtained by deleting the terms in (98) for which j is odd, doubling the terms for j even, and taking arg $(-1/xH_2)^{\frac{1}{2}}$ = arg $[-2/xh_r^{(2)}]^{\frac{1}{2}}$ to be equal to arg $(t - t_r)$ on the part of L_r' just leaving t_r. The result is

$$\int_{L_r'} g(t) \exp [xh(t)]\, dt$$

$$\sim \exp [xh_r] \sum_{j=0}^{\infty} \left[\frac{-2}{xh_r^{(2)}}\right]^{j+\frac{1}{2}} \sum_{n=0}^{2j} \frac{g_r^{(2j-n)}}{(2j-n)!} \sum_{m=0}^{n} b_{mn}\Gamma(m + j + \tfrac{1}{2}) \quad (103)$$

where the derivatives of $g(t)$, $h(t)$ are defined in equation (19) and b_{mn} is computed with

$$a_k = -\frac{2h_r^{(k+2)}}{(k+2)!\, h_r^{(2)}}. \quad (104)$$

The gamma function $\Gamma(m + j + \frac{1}{2})$ may be written as $\sqrt{\pi}\ (\frac{1}{2})_{m+j}$.

For the integral J_0 given by equation (56) most of the contribution comes from the region near the branch point at $t = 0$. When $t = 0$ is not a saddle point, there is only one path of steepest descent {for exp $[xh(t)]$} leaving $t = 0$. This path may be taken to be the branch cut in the t-plane and the path of integration L_0' for J_0 may be taken to be a positive loop enclosing the cut. Then the asymptotic series for J_0 may be obtained from (98) by setting $\nu = 1$, $\rho = \lambda$, $\tau = t$, $G(\tau) = g(t)$, $H(\tau) = h(t)$ and using in place of (101) the integral

$$\int_{+\infty}^{(0+)} u^{z-1} \exp(-u)\, du = [1 - \exp(-i2\pi z)]\Gamma(z) = \frac{2\pi i \exp(-i\pi z)}{\Gamma(1-z)} \quad (105)$$

Here arg u is 0 on the part of the path of integration leaving $t = 0$. The positive real u-axis in (105) is a branch cut. The path of integration starts at $u = +\infty$ on the top side of the cut, comes in along the cut, encircles $u = 0$ in the positive direction, then runs out to $u = +\infty$ along the bottom side of the cut.

It is found that

$$J_0 = \int_{L_0'} t^{\lambda-1} g(t) \exp [xh(t)]\, dt \sim \sum_{j=0}^{\infty} \left[\frac{-1}{xh_0^{(1)}}\right]^{\lambda+j} \sum_{n=0}^{j} \frac{g_0^{(j-n)}}{(j-n)!} \cdot \sum_{m=0}^{n} b_{mn}[1 - \exp(-2\pi i\lambda)]\Gamma(m + \lambda + j) \tag{106}$$

where $\arg [-1/xh_0^{(1)}]$ is equal to $\arg t$ on the part of L_0' leaving $t = 0$, and b_{mn} is computed with

$$a_k = -\frac{h_0^{(k+1)}}{(k+1)!\, h_0^{(1)}}. \tag{107}$$

The last relation in (105) may be used to handle the case in which λ is 0 or a negative integer.

When $t = 0$ is a simple saddle point as well as a branch point, the path L_0' can be taken to coincide with the path of steepest descent through $t = 0$ except for an indentation at $t = 0$. The indentation is chosen so that a man walking in the positive direction along L_0' would have the point $t = 0$ on his left. We put $\nu = 2$, $\rho = \lambda$, $G(\tau) = g(t)$, $H(\tau) = h(t)$ in (98) and use in place of (101) the integral

$$\int_K u^{z-1} \exp(-u^2)\, du = \tfrac{1}{2}[1 - \exp(-i\pi z)]\Gamma\left(\frac{z}{2}\right) = \frac{i\pi \exp(-i\pi z/2)}{\Gamma\left(1 - \frac{z}{2}\right)} \tag{108}$$

Here K runs from $u = -\infty$ to $u = +\infty$ with a downward indentation at $u = 0$, and $\arg u$ is 0 on the part of K leaving $u = 0$. Instead of (106) we now have

$$\int_{L_0'} t^{\lambda-1} g(t) \exp [xh(t)]\, dt \sim \sum_{j=0}^{\infty} \left(\frac{-2}{xh_0^{(2)}}\right)^{(\lambda+j)/2} \sum_{n=0}^{j} \frac{g_0^{(j-n)}}{(j-n)!} \cdot \sum_{m=0}^{n} b_{mn} \tfrac{1}{2}\{1 - \exp[-i\pi(\lambda + j)]\}\Gamma\left(m + \frac{\lambda + j}{2}\right) \tag{109}$$

where $\arg [-2/xh_0^{(2)}]^{\frac{1}{2}}$ is equal to $\arg t$ on the part of L_0' leaving $t = 0$ and b_{mn} is computed with

$$a_k = -\frac{2h_0^{(k+2)}}{(k+2)!\, h_0^{(2)}}. \tag{110}$$

APPENDIX E

Two Saddle Points

In order to illustrate some of the results of Sections IV, V, and VI, we consider the case of two saddle points, $\mu = 2$. This case has been discussed by Chester, Friedman, and Ursell.[4, 6] From (15) the desired expansion is of the form

$$\int_{L'} g(t) \exp [xh(t)]\, dt \sim U_0(x) \sum_{n=0}^{\infty} p_{n0} x^{-n} + U_1(x) \sum_{n=0}^{\infty} p_{n1} x^{-n}, \qquad P_n(v) = p_{n0} + p_{n1} v \tag{111}$$

where, from example (*ii*) of Appendix B and equation (16),

$$\begin{aligned} U_l(x) &= \int_L v^l \exp [xF(v)]\, dv, \qquad l = 0, 1 \\ F(v) &= \tfrac{1}{3}v^3 - v_1^2 v + A_0 , \\ A_0 &= \tfrac{1}{2}(h_2 + h_1), \qquad v_1^3 = \tfrac{3}{4}(h_2 - h_1), \qquad v_2 = -v_1 . \end{aligned} \tag{112}$$

Arg v_1 is determined by the correspondence of v_1 with t_1 which comes with the change of variable from t to v.

Let L' and the change of variable from t to v be such that L runs in from $v = \infty \exp(-i\pi/3)$ to the critical region near $v = 0$ and then out to $\infty \exp(i\pi/3)$ (it may be necessary to reverse the direction of L'). Then

$$\begin{aligned} U_0(x) &= 2\pi i x^{-\frac{1}{3}} Ai(x^{\frac{2}{3}} v_1^2) \exp (xA_0) , \\ U_1(x) &= -2\pi i x^{-\frac{2}{3}} Ai'(x^{\frac{2}{3}} v_1^2) \exp (xA_0) \end{aligned} \tag{113}$$

where $Ai(z)$ is the Airy function and $Ai'(z)$ its derivative with respect to z.

From $F'(v) = v^2 - v_1^2$ and equation (9) it follows that $Q(\zeta, v)$ is $\zeta + v$. Consequently equation (17) for $P_n(v)$ gives

$$P_n(v) = \frac{1}{2\pi i} \int_C d\zeta\, f(\zeta) \left[\frac{1}{\zeta^2 - v_1^2} \frac{\partial}{\partial \zeta} \right]^n \frac{\zeta + v}{\zeta^2 - v_1^2}. \tag{114}$$

Here $f(v) = g(t)t^{(1)}$ in which $t^{(1)} = dt/dv$ is obtained by differentiating $F(v) = h(t)$ with respect to v. The path of integration C encloses $\zeta = \pm v_1$ but no singularities of $f(\zeta)$.

From $v_2 = -v_1$ and

$$p_{00} + v_r p_{01} = P_0(v_r) = f(v_r), \qquad r = 1, 2$$

we get

$$p_{00} = \frac{f(v_1) + f(-v_1)}{2}, \qquad p_{01} = \frac{f(v_1) - f(-v_1)}{2v_1} \tag{115}$$

which may also be obtained from equation (21) for p_{0l}.

Putting $v = v_1$, $n = 1$ in (114) and expanding the integrand in partial fractions leads to

$$P_1(v_1) = \frac{1}{8v_1^3}[f(-v_1) - f(v_1) + 2v_1 f^{(1)}(v_1) - 2v_1^2 f^{(2)}(v_1)]. \tag{116}$$

Expressions for p_{10} and p_{11} which agree with equation (23) for p_{1l} may be obtained from (116) by changing the sign of v_1 to get $P_1(-v_1)$ and using $n = 1$ in

$$\begin{aligned} p_{n0} &= \frac{P_n(v_1) + P_n(-v_1)}{2} \\ p_{n1} &= \frac{P_n(v_1) - P_n(-v_1)}{2v_1}. \end{aligned} \tag{117}$$

If we were to continue with the Bleistein method we would have to evaluate $f^{(n)}(v_r)$ by using equations (26) through (30). Instead we turn to the problem of obtaining $P_n(v_r)$ by the Ursell method. For $\mu = 2$, equations (34) become

$$\begin{aligned} P_0(v_r) &= \alpha_{r0}/\beta_{r00}, \qquad r = 1, 2 \\ P_n(v_r) &= \frac{\alpha_{rn}}{\beta_{r00}} - \sum_{m=1}^{n} \frac{\beta_{r0m}p_{n-m,0} + \beta_{r1m}p_{n-m,1}}{\beta_{r00}} \end{aligned} \tag{118}$$

where α_{rj} and β_{rlj} are given by equations (35) and (37).

Since $g(t)$ and $h(t)$ in the original integral (111) are quite general, we use equation (35) for α_{rj} as it stands. However, equation (37) for β_{rlj} simplifies considerably. This is to be expected since it gives essentially the coefficients in the asymptotic expansions of $Ai(z)$ and $Ai'(z)$. From $F_r^{(2)} = 2v_r$, $F_r^{(3)} = 2$, and $F_r^{(n)} = 0$ for $n > 3$ we have $a_1 = -1/(3v_r)$, and $a_k = 0$ for $k > 1$. It turns out that b_{mn} is 0 for $m \neq n$ and $b_{nn} = a_1^n/n!$. When l is set equal to 0 in (37), all terms vanish except the one for $n = 2j$, $m = 2j$. When $l = 1$, all terms vanish except those for $n = 2j$, $m = 2j$ and $n = 2j - 1$, $m = 2j - 1$. It is found that

$$\begin{aligned} \beta_{r00} &= \left(\frac{-\pi}{v_r}\right)^{\frac{1}{2}}, \qquad \beta_{r0j} = \beta_{r00}\left(\frac{-1}{v_r}\right)^{3j} \frac{(\frac{1}{2})_{3j}}{9^j(2j)!} \\ \beta_{r1j} &= v_r\left(\frac{1 + 6j}{1 - 6j}\right)\beta_{r0j}. \end{aligned} \tag{119}$$

Changing $P_n(v_r)$ into $P_k(v_r)$ in order to avoid confusion with the n used in Appendix D carries (118) into

$$P_k(v_r) = t_r^{(1)}\left[\frac{-2}{h_r^{(2)}}\right]^k \sum_{n=0}^{2k} \frac{g_r^{(2k-n)}}{(2k-n)!} \sum_{m=0}^{n} b_{mn}(\tfrac{1}{2})_{m+k} - \sum_{m=1}^{k} \frac{9^{-m}}{(2m)!}\left(\frac{-1}{v_r}\right)^{3m} (\tfrac{1}{2})_{3m}\left[p_{k-m,0} + v_r\left(\frac{1+6m}{1-6m}\right)p_{k-m,1}\right]. \quad (120)$$

Here $r = 1, 2$; $v_2 = -v_1$; and b_{mn} is computed from (100) with a_k given by equation (36). The value of $t_r^{(1)}$ is given by

$$t_r^{(1)} = \left(\frac{dt}{dv}\right)_{v=v_r} = \left\{\left[\frac{-2}{h_r^{(2)}}\right]^{\frac{1}{2}} \Big/ \left(\frac{-1}{v_r}\right)^{\frac{1}{2}}\right\} \quad (121)$$

where arg $t_r^{(1)}$ is calculated either from (i) the form $t \approx cv$ assumed by the change of variable in the critical region or from (ii) arg $[-2/h_r^{(2)}]^{\frac{1}{2}}$ and arg $(-1/v_r)^{\frac{1}{2}}$ being equal to arg $(t - t_r)$ and arg $(v - v_r)$, respectively, on the portions of the paths of steepest descent leaving t_r and v_r. The last summation in (120) is omitted when $k = 0$. The expression for $P_1(v_r)$ may be written with the help of equations (40).

Equation (120) was checked by using it to obtain the first few terms in the known[10] uniform asymptotic expansion for the Bessel function $H_x^{(1)}(xz)$ with $0 < z < 1$. Here $h(t) = z \sinh t - t$, the saddle points are at $\pm t_1$ ($t_1 > 0$) on the real axis, and the path of integration runs from $t = -\infty$ to $t = \infty + i\pi$. If the direction of the path of integration is reversed [so that (111) gives $-H_x^{(1)}(xz)$], the paths L' and L can be brought into correspondence by a rotation of 120°. In the approximate form $t \approx cv$ of the change of variable, arg $c = 2\pi/3$; and v_1 corresponding to t_1 is $v_1 = |-3h_1/2|^{\frac{1}{3}} \exp(-i2\pi/3)$. Furthermore, $f(v)$ turns out to be an even function, (114) gives $P_n(-v) = (-1)^n P_n(v)$, and $p_{2n,1}$, $p_{2n+1,0}$ are zero for $n = 0, 1, 2, \cdots$.

When t_1 and t_2 approach each other, $h_1^{(2)}$, $h_2^{(2)}$, v_1 and v_2 tend to zero. In this case the asymptotic behavior of the integral (111) may be determined with the help of the equation obtained by setting $\nu = 3$ in equation (98). However, if one is interested in the behavior of the coefficients p_{nl}, the following relations are useful. Putting $v_1 = 0$ in the integral (114) for $P_n(v)$ shows that in the limit

$$p_{00} = f(0), \qquad p_{01} = f'(0), \qquad p_{10} = -f^{(3)}(0)/3!, \qquad p_{11} = -2f^{(4)}(0)/4! \quad (122)$$

and so on. The derivatives $t^{(n)}$ appearing in the derivatives of $f(v) = g(t)t^{(1)}$ are now obtained by differentiating $3^{-1}v^3 = h(t)$ repeatedly

with respect to v. The leading coefficients are found to be

$$\begin{aligned} p_{00} &= g_1^{(0)} t_1^{(1)} & t_1^{(1)} &= [2/h_1^{(3)}]^{\frac{1}{3}} \\ p_{01} &= g_1^{(1)} t_1^{(1)2} + g_1^{(0)} t_1^{(2)} & t_1^{(2)} &= -h_1^{(4)} t_1^{(1)5}/12. \end{aligned} \tag{123}$$

APPENDIX F

Saddle Point Near Branch Point

Here we apply the theory of Sections VII and VIII to a case discussed by Bleistein,[3] namely, $\lambda \neq 1$ and $\mu = 1$. The paths L', L and the functions $h(t)$, $F(v)$ are assumed to be the same as in Section IX where the singularity at the origin was a simple pole instead of a branch point. In the critical region L' is parallel to, and to the right of, the imaginary t-axis. Only the case in which t_1 and v_1 are real and of the same sign will be considered. When t_1 and v_1 are positive the cut associated with the branch point is assumed to start out from the origin along the positive real axis, and then quickly bend downward to run out to $-i\infty$. When t_1 and v_1 are negative, the cut starts out along the negative axis and then bends downward to $-i\infty$.

Equation (46) and $F(v) = v^2 - 2v_1 v$ lead to

$$J \sim V_0(x) \sum_{n=0}^{\infty} p_{n0} x^{-n} + V_1(x) \sum_{n=0}^{\infty} p_{n1} x^{-n} \tag{124}$$

where, with $c > 0$ and the path of integration lying to the right of the cut,

$$V_0(x) = \int_{c-i\infty}^{c+i\infty} v^{\lambda-1} \exp [xF(v)]\, dv = i\pi x^{-\lambda/2} \sum_{n=0}^{\infty} \frac{(-2v_1 x^{\frac{1}{2}})^n}{n!\, \Gamma\left(1 - \frac{\lambda+n}{2}\right)}.$$

Replacing λ by $\lambda + 1$ gives $V_1(x)$. $V_0(x)$ and $V_1(x)$ are parabolic cylinder functions (Bleistein,[3] and pair No. 740.2 in Campbell and Foster Table[9]).

$$\begin{aligned} P_n(v) &= p_{n0} + p_{n1} v \\ &= \frac{1}{2\pi i} \int_C \frac{d\zeta\, f(\zeta) \zeta^{\lambda-1}}{2^n} \left[\frac{1}{\zeta - v_1} \frac{\partial}{\partial \zeta}\right]^n \frac{(\zeta + v - v_1)\zeta^{-\lambda}}{\zeta - v_1} \end{aligned} \tag{125}$$

where C encloses $\zeta = 0$ and $\zeta = v_1$. Setting $n = 0$ and using $f(v) = g(t)\ (t/v)^{\lambda-1} t^{(1)}$ leads to

$$\begin{aligned} P_0(0) &= p_{00} = f(0) = g_0^{(0)} t_0^{(1)\lambda}, & t_0^{(1)} &= -2v_1/h_0^{(1)} \\ P_0(v_1) &= f(v_1), & t_1^{(1)} &= [2/h_1^{(2)}]^{\frac{1}{3}}. \end{aligned}$$

Setting $n = 1$ gives

$$P_1(0) = \frac{\lambda}{2v_1^2}[f(0) - f(v_1) + v_1 f^{(1)}(0)]$$

$$P_1(v_1) = \frac{1-\lambda}{2v_1^2}[f(0) - f(v_1) + v_1 f^{(1)}(v_1)] - \frac{f^{(2)}(v_1)}{4}.$$

The values of p_{n0} and p_{n1} may be obtained by following the steps outlined in the first part of Section VIII. Equation (52) and (53) become

$$\begin{aligned} P_n(v_1) &= \frac{1}{\beta_{100}}\left[\alpha_{1n} - \sum_{m=1}^{n}(\beta_{10m}p_{n-m,0} + \beta_{11m}p_{n-m,1})\right] \\ p_{n0} &= \frac{1}{\beta_{000}}\left[\alpha_{0n} - \sum_{q=1}^{n}(\beta_{00q}p_{n-q,0} + \beta_{0,1,q-1}p_{n-q,1})\right]. \end{aligned} \tag{126}$$

The path L_1 in (54) is parallel to the imaginary axis and passes through the saddle point at $v = v_1$. The path L_0 in (56) runs up along the right side of the cut, encircles the origin in a positive direction, and then runs down to $v = -i\infty$ along the left side of the cut.

To get β_{1lj} we replace l on the right side of (37) by $l + \lambda - 1$ and notice that a_k given by (38) is 0 for the values of k used in computing b_{mn}. Consequently, the only nonvanishing b_{mn} is $b_{oo} = 1$ and (37) gives

$$\beta_{100} = i(\pi)^{\frac{1}{2}}v_1^{\lambda-1}, \qquad \frac{\beta_{1lj}}{\beta_{100}} = \frac{(-4)^{-j}}{j!}(-l-\lambda+1)_{2j}v_1^{l-2j}.$$

The expression for α_{1n} obtained by replacing $g_1^{(2j-n)}$ in (35) by $\hat{g}_1^{(2j-n)}$, where $\hat{g}(t) = t^{\lambda-1}g(t)$, contains $[-2/h_1^{(2)}]^{\frac{1}{2}}$ which may be written as $it_1^{(1)}$.

In equation (57) for α_{0j}/β_{000} we replace b_{mn} by $\hat{b}_{mn}$ to indicate that $\hat{b}_{mn}$ is computed with a_k given by (58) instead of the a_k used in computing α_{1n}/β_{100}. A still different a_k, given by (60), is used to compute β_{0lj}/β_{000}. From equation (60) all of the a_k's used to compute β_{0lj}/β_{000} are zero except $a_1 = 1/(2v_1)$. Therefore b_{mn} is zero unless $m = n$, and it follows from (59) that

$$\beta_{0lj}/\beta_{000} = (\lambda)_{l+2j}(2v_1)^{-l-2j}/j!.$$

When all of these results are used in the expression (126) for p_{no} we get, with j for n,

$$\begin{aligned} p_{j0} = (2v_1)^{-j}t_0^{(1)(\lambda+j)} &\sum_{n=0}^{j}\frac{g_0^{(j-n)}}{(j-n)!}\sum_{m=0}^{n}\hat{b}_{m,n}(\lambda)_{m+j} \\ &- \sum_{s=1}^{j}\frac{(\lambda)_{2s-1}}{s!\,(2v_1)^{2s}}[(\lambda+2s-1)p_{j-s,0} + 2v_1 s p_{j-s,1}] \end{aligned} \tag{127}$$

where $h(0) = h_0 = 0$, $h_1 < 0$, $h_1^{(2)} > 0$, $v_1 = t_1 \mid t_1 \mid^{-1}(-h_1)^{\frac{1}{2}}$, $t_0^{(1)} = -2v_1/h_0^{(1)}$, and $\hat{b}_{mn}$ is computed from (100) with a_k replaced by $\hat{a}_k$ where

$$\hat{a}_k = -h_0^{(k+1)}/[(k+1)!\, h_0^{(1)}], \qquad k \geqq 1.$$

When $j = 0$ the summation with respect to s is omitted. Similarly, $P_n(v_1)$ gives

$$p_{j1} = (-1)^j v_1^{-\lambda} t_1^{(1)(2j+1)} \sum_{n=0}^{2j} \frac{\hat{g}_1^{(2j-n)}}{(2j-n)!} \sum_{m=0}^{n} b_{mn}(\tfrac{1}{2})_{m+j} - \frac{p_{j0}}{v_1}$$
$$- \sum_{s=1}^{j} \frac{(-1)^s(1-\lambda)_{2s-1}}{s!\,(2v_1)^{2s}} \left[\frac{-\lambda + 2s}{v_1} p_{j-s,0} - \lambda p_{j-s,1} \right] \tag{128}$$

where $t_1^{(1)} = [2/h_1^{(2)}]^{\frac{1}{2}}$, $\hat{g}(t) = t^{\lambda-1}g(t)$, and b_{mn} is computed from (100) with

$$a_k = -2h_1^{(k+2)}/[(k+2)!\, h_1^{(2)}].$$

It is interesting to notice that when $\lambda = 0$, (127) and (128) give values of p_{j0} and p_{j1} which agree with those obtained in Section IX.

When the saddle point approaches the branch point, t_1, h_1, $h_0^{(1)}$, and v_1 tend to zero. In this case the integral J may be evaluated with the help of equation (109). The behavior of the coefficient p_{nl} may be studied by putting $v_1 = 0$ in the integral (125) for $P_n(v)$. It is found that

$$\begin{aligned} p_{00} &= f(0), & p_{01} &= f^{(1)}(0), \\ p_{10} &= -\lambda f^{(2)}(0)/2, & p_{11} &= -(\lambda+1)f^{(3)}(0)/12 \end{aligned} \tag{129}$$

and so on. The derivatives $t^{(n)}$ appearing in the derivatives of $f(v) = g(t)\ (t/v)^{\lambda-1}t^{(1)}$ are now obtained by differentiating $v^2 = h(t)$ repeatedly with respect to v.

For $n = 1$ and 2,

$$t_0^{(1)} = [2/h_0^{(2)}]^{\frac{1}{2}}, \qquad t_0^{(2)} = -h_0^{(3)} t_0^{(1)4}/6. \tag{130}$$

Substituting these values in equations (51) for $f(0)$ and $f^{(1)}(0)$ and using (129) gives the limiting expressions for p_{00} and p_{01}.

APPENDIX G

Poisson-Charlier Polynomial

In this appendix equations (127) and (128) for p_{j0} and p_{j1} are used to obtain an asymptotic series for the Poisson-Charlier polynomial $c_n(y, a)$ when y is $O(1)$ and both n and a are large and positive.

Multiplying the generating function

$$(1-u)^y \exp(au) = \sum_{m=0}^{\infty} \frac{a^m u^m}{m!} c_m(y, a) \tag{131}$$

by u^{-n-1} and integrating u around a small circle enclosing $u = 0$ gives a contour integral for $c_n(y, a)$. Instead of $c_n(y, a)$, we find it more convenient to deal with the polynomial (in y)

$$d_n(y, a) = \frac{a^n \exp(-a)}{n!} c_n(y, a). \tag{132}$$

Setting $x = n + 1$ and making the change of variable $t = 1 - u$ in the integral for $c_n(y, a)$ leads to

$$\begin{aligned} d_n(y, a) &= \frac{1}{2\pi i} \int_{\frac{1}{2}-i\infty}^{\frac{1}{2}+i\infty} \frac{t^y \exp(-at)}{(1-t)^x} dt, \qquad x > y + 1 \\ &= \frac{1}{2\pi i} \int_{\frac{1}{2}-i\infty}^{\frac{1}{2}+i\infty} t^y \exp[xh(t)]\, dt \\ &= \frac{J}{2\pi i} = \hat{J}, \\ h(t) &= -rt - \ln(1-t) \end{aligned} \tag{133}$$

where $r = a/x$. The ratio r is positive.

We wish to use equations (127) and (128) to compute p_{n0}, p_{n1} in the expansion obtained by dividing (124) by $2\pi i$, namely

$$d_n(y, a) = \hat{J} \sim \hat{V}_0(x) \sum_{n=0}^{\infty} p_{n0} x^{-n} + \hat{V}_1(x) \sum_{n=0}^{\infty} p_{n1} x^{-n}. \tag{134}$$

Here $\hat{V}_0(x) = V_0(x)/2\pi i$ with $\lambda = y + 1$. The function $\hat{V}_1(x)$ is obtained from $\hat{V}_0(x)$ by increasing y by 1.

We have

$$\begin{aligned} \hat{V}_0(x) &= \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} v^y \exp[x(v^2 - 2v_1 v)]\, dv, \qquad c > 0 \\ &= x^{-(y+1)/2} G[y, v_1 x^{\frac{1}{2}}] \end{aligned} \tag{135}$$

where $\arg v$ is 0 at $v = c$ and $G(y, z)$ is the parabolic cylinder function

$$\begin{aligned} G(y, z) &= \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} u^y \exp(u^2 - 2zu)\, du \\ &= 2^{-1} \sum_{n=0}^{\infty} \frac{(-2z)^n}{n!\, \Gamma\left(\frac{1-y-n}{2}\right)}. \end{aligned} \tag{136}$$

The function G is related to the function U discussed and tabulated in chapter 19 of Ref. 10 by the equation

$$G(y, z) = \frac{2^{-y/2}}{2(\pi)^{\frac{1}{2}}} \exp(-\tfrac{1}{2}z^2)\, U[-y - \tfrac{1}{2}, z2^{\frac{1}{2}}].$$

The saddle point $t = t_1$ is obtained by setting the derivative $h^{(1)}(t) = -r + (1 - t)^{-1}$ to zero; and the saddle point $v = v_1$ is given by the relation $v_1^2 = -h_1$ together with the condition that v_1 and t_1 be of the same sign:

$$t_1 = 1 - \frac{1}{r}, \qquad v_1 = \frac{t_1}{|t_1|}(r - 1 - \ln r)^{\frac{1}{2}}. \tag{137}$$

The values of the derivatives $t^{(1)} = dt/dv$ at $v = 0$ and $v = v_1$ are

$$t_0^{(1)} = -2v_1/h_0^{(1)} = \frac{2v_1}{r - 1}$$

$$t_1^{(1)} = [2/h_1^{(2)}]^{\frac{1}{2}} = \frac{2^{\frac{1}{2}}}{r}.$$

For $k > 1$ the kth derivative $h^{(k)}(t)$ is $(k - 1)!(1 - t)^{-k}$ and the coefficients used to compute $\hat{b}_{m,n}$, $b_{m,n}$ from (100) are

$$\hat{a}_k = -k!/[(k + 1)!\,(1 - r)] = 1/[(k + 1)(r - 1)]$$

and

$$a_k = -2(k + 1)!\, r^{k+2}/[(k + 2)!\, r^2] = -2r^k/(k + 2),$$

respectively.

Comparison of the integral (133) for $d_n(y, a)$ with the integral (1) for J shows that $g(t) = 1$ and $\lambda = y + 1$. Consequently $\hat{g}(t) = t^{\lambda-1}g(t)$ becomes $\hat{g}(t) = t^y$. For $k > 0$ the derivatives are $g^{(k)} = 0$ and $\hat{g}^{(k)} = y(y - 1) \cdots (y - k + 1)t^{y-k}$.

Setting $j = 0$ in (127) and (128), and using the results just obtained gives

$$p_{00} = \left(\frac{2v_1}{r - 1}\right)^{y+1}, \qquad p_{01} = \left(\frac{t_1}{v_1}\right)^y \frac{(2)^{\frac{1}{2}}}{rv_1} - \frac{p_{00}}{v_1} \tag{138}$$

for the leading coefficients in the asymptotic expansion for $d_n(y, a)$. Here t_1 and v_1 are given by (137). The next two coefficients, obtained by setting $j = 1$, reduce to

$$\begin{aligned} p_{10} &= (y + 1)(y + 2)p_{00}\left[\frac{1}{2(r - 1)^2} - \frac{1}{4v_1^2}\right] - \frac{(y + 1)p_{01}}{2v_1} \\ p_{11} &= -\left(\frac{t_1}{v_1}\right)^{y+1}\left[\frac{(2)^{\frac{1}{2}}}{r - 1}\right]^3\left[\frac{y(y - 1)}{4} - \frac{yt_1r}{2} + \frac{t_1^2r^2}{24}\right]. \end{aligned} \tag{139}$$

Some idea of the behavior of the series (134) for $d_n(y, a)$ may be gained from Table II. Equations (127) and (128) were programmed for calculation on a high speed digital computer. The table lists results for the typical case $x = 30$, $a = 25$, and $y = -5$. Here $\text{Term}_{2n} = \hat{V}_0(x)p_{n0}x^{-n}$, $\text{Term}_{2n+1} = \hat{V}_1(x)p_{n1}x^{-n}$, and $S_m = t_0 + t_1 + \cdots + t_m$. The "exact" value, 381.02, was calculated by using the recurrence relation for the Poisson-Charlier polynomials.

No study was made to decide whether the relatively large value of Term_7 results from accumulated round-off error (an accuracy of 1 part in 10^7 was used) or whether the asymptotic series actually starts its divergence around $m = 5$ or 6.

When r is near unity, v_1 and t_1 are small and the individual terms in the expressions (127) and (128) for p_{j0}, p_{j1}, become large. In this case considerable cancellation occurs, and a high degree of precision in the calculations is required to obtain accurate values of p_{j0} and p_{j1}.

An asymptotic series (nonuniform) which is useful when $r - 1$ is small may be obtained by a variation of the classical method which is sometimes used in cases of this sort. Instead of using an expansion about both $t = 0$ and $t = t_1$, which is done (in effect) in obtaining the uniform asymptotic expansion, an expansion is made only about $t = 0$. Thus, the exponent $xh(t)$ may be written as

$$xh(t) = [-x(r - 1)t + xt^2/2] + (xt^2)(t/3 + t^2/4 + \cdots).$$

Changing the variable of integration from t to $u = t(x/2)^{1/2}$ and assuming that $r - 1$ is so small that $z = (r-1)\ (x/2)^{1/2}$ is $O(1)$ gives

$$\begin{aligned} \exp\,[xh(t)] &= \exp\,[-2zu + u^2]\,\exp\,[u^2(2t/3 + 2t^2/4 + \cdots)] \\ &= \exp\,[-2zu + u^2] \sum_{n=0}^{\infty} (2/x)^{n/2} \sum_{m=0}^{n} b_{mn}u^{2m+n}. \end{aligned} \tag{140}$$

The last series is obtained from equation (102) with $\xi = t = u\ (2/x)^{1/2}$,

TABLE II — PARTIAL SUMS FOR $d_{29}(-5, 25)$

m	Term_m	S_m	m	Term_m	S_m
0	276.62	276.62	4	0.015	381.03
1	82.82	359.44	5	−0.008	381.02
2	20.19	379.63	6	0.008	381.03
3	1.39	381.01	7	−0.144	380.88

Exact $d_{29}(-5, 25) = 381.02$

$y = u^2$, and $a_n = 2/(n+2)$. The coefficient b_{mn} is computed from the a_n's and (100).

Substituting (140) in the integral (133) for $d_n(y, a)$ leads to

$$d_n(y, a) \sim \sum_{n=0}^{\infty} (2/x)^{(n+y+1)/2} \sum_{m=0}^{n} b_{mn} K_{2m+n} \tag{141}$$

where

$$K_n = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} u^{n+y} \exp(u^2 - 2zu)\, du$$
$$= G(n + y, z) = G[n + y, (r - 1)(x/2)^{\frac{1}{2}}].$$

The recurrence relation

$$K_{n+1} = zK_n - \frac{n + y}{2} K_{n-1} \tag{142}$$

permits K_{2m+n} in (141) to be expressed as a linear function of K_0 and K_1. Equation (142) is obtained by integrating the derivative

$$\frac{d}{du} u^{n+y} \exp(u^2 - 2zu) = [(n + y)u^{-1} + 2u - 2z]u^{n+y} \exp(u^2 - 2zu)$$

As $r \to 1$ the leading term, $(2/x)^{(y+1)/2} K_0$, in (141) tends to the leading term $\hat{V}_0(x) p_{00}$ in the uniform asymptotic series (134). Although (141) is much simpler than (134), it does not hold for nearly as wide a range of values of $r - 1$.

APPENDIX H

Saddle Point at Origin

Here we are concerned with the leading term when $\lambda \neq 1$ and there are two saddle points, one at $t = 0$ and the other at $t = t_1$. We assume that t_1 is real and positive, and that $h(t)$ is real on the real axis. Furthermore, we assume $h_1 < 0$, $h_0^{(2)} < 0$, $h_1^{(2)} > 0$ so that the saddle point at t_1 is lower than the one at 0, and the paths of steepest descent at 0 and t_1 are parallel to the real and imaginary axes, respectively. A cut extends from 0 to $-\infty$ along the negative real axis.

The paths of integration L' and L are taken to run in from $\infty \exp(-i\pi/3)$, cross the positive real axis in the critical region, and the run out to $\infty \exp(i\pi/3)$. Example (*iii*) of Appendix B leads us to choose

$$F(v) = 2v^3 - 3v_1 v^2, \qquad v_1^3 = -h_1 \tag{143}$$

with arg $v_1 = 0$. From equation (46), the uniform asymptotic expansion is of the form

$$J \sim V_0(x) \sum_{n=0}^{\infty} p_{n0} x^{-n} + V_1(x) \sum_{n=0}^{\infty} p_{n1} x^{-n} + V_2(x) \sum_{n=0}^{\infty} p_{n2} x^{-n} \tag{144}$$

where

$$V_l(x) = \int_L v^{l+\lambda-1} \exp [xF(v)]\, dv, \qquad l = 0, 1, 2.$$

Expanding $\exp(-3xv_1v^2)$ and integrating termwise with the help of

$$\int_L u^\rho \exp (u^3)\, du = 2\pi i / \left[3\Gamma\left(\frac{2-\rho}{3}\right)\right]$$

gives

$$V_0(x) = \frac{2\pi i}{3} (2x)^{-\lambda/3} \sum_{n=0}^{\infty} \frac{(-3v_1)^n (x/4)^{n/3}}{n!\, \Gamma\left(1 - \dfrac{\lambda + 2n}{3}\right)} \tag{145}$$

from which $V_l(x)$ may be obtained by replacing λ by $\lambda + l$. When $\lambda = 1$, $V_0(x)$ reduces to the product of an Airy function and an exponential.

Setting $n = 0$ in the integral (47) for $P_n(v)$ gives

$$\begin{aligned} P_0(v) &= p_{00} + p_{01}v + p_{02}v^2 \\ &= \frac{1}{2\pi i} \int_C f(\zeta) \frac{[\zeta^2 + \zeta(v - v_1) + (v^2 - v_1 v)]}{\zeta^2(\zeta - v_1)}\, d\zeta \\ &= f(0) + f'(0)v + [f(v_1) - f(0) - v_1 f'(0)]v^2 v_1^{-2}. \end{aligned} \tag{146}$$

The values of $t_0^{(1)}$, $t_0^{(2)}$, $t_1^{(1)}$ appearing in $f(0)$, $f'(0)$, $f(v_1)$ are

$$\begin{gathered} t_0^{(1)} = \left[\frac{-6v_1}{h_0^{(2)}}\right]^{\frac{1}{3}}, \qquad t_0^{(2)} = \frac{12 - h_0^{(3)} t_0^{(1)3}}{3h_0^{(2)} t_0^{(1)}} \\ t_1^{(1)} = \left[\frac{6v_1}{h_1^{(2)}}\right]^{\frac{1}{3}}. \end{gathered} \tag{147}$$

The derivatives $t_0^{(1)}$, $t_1^{(1)}$ are positive and nearly equal when the saddle points are close together.

When $n = 0$, the Ursell equations (61), (62), and (63) become

$$\begin{aligned} P_0(v_1) &= \alpha_{10}/\beta_{100} \\ p_{00} &= \alpha_{00}/\beta_{000} \\ p_{01} &= \frac{\alpha_{01}}{\beta_{010}} - \frac{\beta_{001}}{\beta_{010}} p_{00}\,. \end{aligned} \tag{148}$$

The values of α_{10}, β_{100} otbained from the leading terms in the asymptotic series (54) defining the α_{rn}'s and β_{rlm}'s give

$$P_0(v_1) = g_1^{(0)}(t_1/v_1)^{\lambda-1}t_1^{(1)}. \tag{149}$$

Similarly, comparing the asymptotic series (64) defining the α_{0j}'s and β_{0lm}'s with the series (109) leads to expressions which give

$$\begin{aligned} p_{00} &= g_0^{(0)}t_0^{(1)\lambda}, \\ p_{01} &= \left\{g_0^{(1)} - (\lambda+1)g_0^{(0)}\left[\frac{1}{3v_1t_0^{(1)}} + \frac{h_0^{(3)}}{6h_0^{(2)}}\right]\right\}t_0^{(1)(\lambda+1)}. \end{aligned} \tag{150}$$

The remaining coefficient, p_{02}, in $P_0(v)$ may now be obtained by combining (149) and (150). The coefficients p_{0l} give the leading part of the desired expansion (144) for J.

ACKNOWLEDGMENTS

I wish to express my appreciation for the computations made by Mr. G. H. Robertson to check my first approximations against his algorithm. I am also indebted to him for values which enabled me to check the expansion for the noncentral χ^2 distribution given in Section X. In addition, it gives me pleasure to acknowledge helpful comments made by L. A. Shepp and several other colleagues.

REFERENCES

1. Robertson, G. H., "Computation of the Noncentral Chi-Square Distribution," scheduled to be published in B.S.T.J., *48*, No. 1 (January 1969).
2. Rice, S. O., "Communication in the Presence of Noise—Probability of Error for Two Encoding Schemes," B.S.T.J., *29*, No. 1 (January 1950), pp. 60–93.
3. Bleistein, N., "Uniform Asymptotic Expansions of Integrals with Stationary Point Near Algebraic Singularity," Comm. Pure and Applied Math. *19*, No. 4 (1966), pp. 353–370.
4. Ursell, F., "Integrals with a Large Parameter. The Continuation of Uniformly Asymptotic Expansions," Proc. Cambridge Phil. Soc., *61* (1965), pp. 113–128.
5. Olver, F. W. J., "The Asymptotic Expansions of Bessel Functions of Large Order," Phil. Trans. Royal Soc., Series A, *247* (1954), pp. 328–368.
6. Chester, C., Friedman, B., and Ursell, F., "An Extension of the Method of Steepest Descents," Proc. Cambridge Phil. Soc., *53* (1957), pp. 599–611.
7. Felsen, L. B., "Radiation from a Uniaxially Anisotropic Plasma Half-Space," IEEE Trans. Antennas and Propagation, *AP-11* (1963), pp. 469–484.
8. Riordan, J., *An Introduction to Combinatorial Analysis*, New York: John Wiley and Sons, 1958.
9. Campbell, G. A., and Foster, R. M., *Fourier Integrals for Practical Applications*, New York: Van Nostrand, 1948.
10. Abramowitz, M., and Stegun, I. A., *Handbook of Mathematical Functions*, Washington, D. C.: National Bureau of Standards, 1964, equation (9.3.37).

Gain Control for Diversity Receivers

By STEPHEN S. RAPPAPORT

(Manuscript received April 18, 1968)

Previous work on optimum gain control is extended to an important class of diversity receivers used for digital data transmission through fading media and for radar. As in the single diversity case the optimum gain (which yields minimum average cost of receiver saturation) is extremely insensitive to relative costs of saturation at the upper and lower dynamic range bounds. The sensitivity to relative cost decreases as the order of diversity increases.

Optimum gain and performance characteristics are given from which dynamic range requirements for diversity receivers can be deduced.

I. INTRODUCTION

A good part of detection theory literature deals with the determination of statistically optimum or near optimum receiver structures. However, in any practical implementation of these receivers the signal processing must be performed by components of finite dynamic range. To effectively use the amplitude range of a signal processing chain it is common to scale the received signal by adjusting the receiver gain. Optimum gain settings for minimum average cost of excluding (from a receiver's finite dynamic range) the envelope of a narrowband signal plus gaussian noise were presented last year.[1] Here similar results are presented for an important class of diversity receivers used for comunications through fading media and for radar. For the single diversity ($M = 1$) case, these results reduce to those given previously.

II. PREVIOUS RESULTS

Consider the problem of determining the normalized attenuation, a to optimally scale a positive homogeneous functional, ξ, of the received signal so that the average cost, l, of excluding ξ from the receiver's dynamic range is minimized. It follows from Ref. 1 that

the average exclusion cost is given by

$$l = \int_0^a p_\omega(\xi)\, d\xi + \nu \int_{ad}^\infty p_\omega(\xi)\, d\xi \tag{1}$$

in which d denotes the dynamic range of the receiver (such that $D(\text{dB}) = 20 \log_{10} d$, $d > 1$), ν is the ratio of cost of saturation at the upper dynamic range bound to the cost of saturation at the lower, ω is a vector parameter determined by signal noise and channel conditions, and p_ω is the probability density function of ξ. When the optimizing value of a is a stationary point of l it can be found as a real positive solution to

$$p_\omega(a) = \nu\, dp_\omega(ad). \tag{2}$$

If a_s is the optimum a then the minimum average exclusion cost is

$$l = P_\omega(a_s) + \nu[1 - P_\omega(a_s\, d)] \tag{3}$$

in which P_ω is the cumulative distribution corresponding to p_ω. For $\nu = 1$, l becomes the exclusion probability. Ref. 1 considered this problem in detail for the case in which ξ represents the envelope of a narrowband signal plus gaussian noise received through a Rician fading medium. The results are based upon the solution of (2) for the case in which p_ω is the Rician[2] probability density function defined by

$$p_\gamma(\xi) = \xi \exp[-(\xi^2 + \gamma^2)/2] I_0(\gamma\xi) \qquad \gamma, \xi \geqq 0 \tag{4}$$

where γ is a suitably defined signal-to-noise ratio. ($\gamma_{\text{dB}} = 20 \log_{10}\gamma$).

III. GAIN SETTINGS FOR DIVERSITY RECEIVERS

In various diversity receivers formation of the test statistic leads to the generalized Rician probability density function given by

$$p_\omega(R) = R(R(M)^{\frac{1}{2}}/\gamma)^{M-1} \exp[-(R^2 + \gamma^2/M)/2] I_{M-1}[\gamma R/(M)^{\frac{1}{2}}] \tag{5}$$

$$R, \gamma, \geqq 0$$

$$M = 1, 2, \cdots$$

where I_K denotes the modified Bessel function of the first kind and order K and ω is the vector (γ, M). Such is the case for example in square-law combining M-fold diversity receivers for noncoherent frequency shift keyed signaling through Rician or Rayleigh (if $\gamma = 0$) fading channels, in radar receivers using post detection square law integration of M pulses,† and in partially coherent diversity reception of N-ary orthogonal

† In these cases the functional ξ is the test statistic. More generally however, the functional used for determining the optimum gain need not be actually formed in the receiver.

signals transmitted through M independent slow Rician fading channels.[3,4]

The probability density function (5) has interesting properties. It can be shown that (5) is the probability density function of the square root of the sum of squares of M independent normalized Rician variates, each variate having a probability density function of the form (4) with γ replaced by $\gamma/(M)^{\frac{1}{2}}$. For $M = 1$, (5) of course reduces to the Rician probability density function (4), and if in addition $\gamma = 0$ it becomes the Rayleigh probability density function. The density (5) can be viewed as the probability density function of the square root of a noncentral chi-square variate with $2M$ degrees of freedom and noncentrality parameter γ^2. With $\gamma = 0$ it becomes the density of the square root of a chi-square variate with $2M$ degrees of freedom.

In many practical cases γ^2 is proportional to the ratio of the total specular energy received via the M diversity branches to the sum of the scatter and noise energy received via any diversity branch (assuming that this latter sum is the same for any diversity branch). Thus γ^2/M can be thought of as the power signal-to-noise ratio per diversity branch or per pulse in the case of time diversity if, as is commonly assumed, the diversity branches are statistically independent but have identical parameters.

Since (5) arises in various applications, let us consider the canonical problem. Specifically the solutions to (2) will be obtained in which p_ω is given by (5). Letting $a = A(2)^{\frac{1}{2}}$ and $\alpha = \gamma/(M)^{\frac{1}{2}}$ leads to the following transcendental equation for A:

$$A^2 = A_c^2 + [(M+1)/2]\, A_0^2 + [1/(d^2-1)] \cdot \{\ln I_{M-1}[\alpha A\, d(2)^{\frac{1}{2}}] - \ln I_{M-1}[\alpha A(2)^{\frac{1}{2}}]\} \tag{6}$$

in which

$$A_0^2 \triangleq \frac{2 \ln d}{d^2 - 1} \tag{7}$$

determines the optimum required attenuation for the Rayleigh case with unity cost ratio ($\nu = 1$), and

$$A_c^2 \triangleq \frac{\ln \nu}{d^2 - 1}. \tag{8}$$

For the single diversity case ($M = 1$), (6) reduces to the trans-

cendental equation encountered previously.[1] One is thus led to seek an iterative solution along the same lines. Before obtaining the required iteration equations consider some properties of (6). The equation can be written in the form

$$A^2 = A_c^2 + MA_0^2 + [1/(d^2 - 1)]\{\ln [(\alpha A\ d(2)^{\frac{1}{2}})^{M-1} I_{M-1}(\alpha A\ d(2)^{\frac{1}{2}})]$$

$$- (M - 1) \ln d^2 - \ln [(\alpha A(2)^{\frac{1}{2}})^{M-1} I_{M-1}(\alpha A(2)^{\frac{1}{2}})]\}. \quad (9)$$

Combining the logarithmic terms on the right side of (9) and using the fact that $I_K(Z) \to (Z^K)/2^K K!$ as $Z \to 0$ it is seen that the quantity in braces goes to zero as $\gamma \to 0$. Hence the optimum attenuation for the chi-square case ($\gamma = 0$) is determined explicitly by $A = A_{cs}$ where

$$A_{cs}^2 \triangleq A_c^2 + MA_0^2 . \quad (10)$$

In the same manner it is seen that if $A_{cs}^2 = 0$, then $A = 0$ is a solution to (9) for any γ and d. It is easy to show that the right hand side of (9) is an even function of A having a minimum of A_{cs}^2 at $A = 0$. The left hand side of (9) is of course a standard parabola centered at the origin. These curves (*i*) do not intersect if $A_{cs}^2 < 0$, (*ii*) intersect only at $A = 0$ if $A_{cs}^2 = 0$, and (*iii*) intersect at positive (and negative) values of A if $A_{cs}^2 > 0$. Thus meaningful values of A which minimize l are stationary if and only if $A_{cs}^2 > 0$. From (7), (8), and (10) this requires $vd^{2M} > 1$, which for $M = 1$ reduces to the constraint encountered previously.[1]

The solution to (6) or (9) can be obtained using the extrapolated iteration scheme described in Ref. 1. The iteration formulas require the derivative of the right side of (9) which can be found using the identity

$$d/d\zeta\ [\zeta^n I_n(\zeta)] = \zeta^n I_{n-1}(\zeta) \qquad n = \cdots -2, -1, 0, 1, 2, \cdots . \quad (11)$$

The result is

$$f'(A^2) = \frac{\gamma(2)^{\frac{1}{2}}}{2(d^2 - 1)A}\left\{d\,\frac{I_{M-2}[A\ d\alpha(2)^{\frac{1}{2}}]}{I_{M-1}[A\ d\alpha(2)^{\frac{1}{2}}]} - \frac{I_{M-2}[A\alpha(2)^{\frac{1}{2}}]}{I_{M-1}[A\alpha(2)^{\frac{1}{2}}]}\right\} \quad (12)$$

in which f denotes the right hand side of (9) and the prime denotes differentiation with respect to the argument.

For computational purposes it is convenient to define the functions

$$\Psi_K(\zeta) \triangleq (\exp -\zeta)\ I_K(\zeta) \quad (13)$$

which are uniformly bounded on the semi-infinite interval $[0, \infty]$. For any argument ζ these functions can be readily generated by

numerical techniques using the recurrence relations and asymptotic expansions for the modified Bessel functions.

Using (12) and (13) the following iteration formulas to solve (6) are obtained.

$$
\begin{aligned}
A_{i+1}^2 &= F_i - \beta_i(A_i^2 - F_i) \qquad \beta_i \neq -1 \\
F_i &= A_c^2 + [(M+1)/2]\, A_0^2 \\
&\quad + \frac{A_i\alpha(2)^{\frac{1}{2}}}{d+1} + [1/(d^2-1)]\left[\ln \frac{\Psi_{M-1}(A_i\, d\alpha(2)^{\frac{1}{2}})}{\Psi_{M-1}(A_i\alpha(2)^{\frac{1}{2}})}\right] \\
G_i &= \frac{\alpha(2)^{\frac{1}{2}}}{2(d^2-1)A_i}\left\{d\,\frac{\Psi_{M-2}[A_i\, d\alpha(2)^{\frac{1}{2}}]}{\Psi_{M-1}[A_i\, d\alpha(2)^{\frac{1}{2}}]} - \frac{\Psi_{M-2}[A_i\alpha(2)^{\frac{1}{2}}]}{\Psi_{M-1}[A_i\alpha(2)^{\frac{1}{2}}]}\right\} \\
\beta_i &= G_i/(1-G_i) \qquad G_i \neq 1
\end{aligned}
\tag{14}
$$

The iteration is begun with $i = 1$, γ small, and $A_1^2 = A_{cs}^2$ and stopped when $|\,(A_{i+1} - A_i)/A_i\,|$ is less than the allowable error. By this method the optimum required normalized attenuation was found for various values of ν, γ, d, and M.

Inasmuch as the optimum attenuation satisfies the nonlinear equation (6) it will be helpful in interpreting results to find some useful approximations. Accordingly one notes that for $\gamma^2/M \gg 1$ and $d \gg 1$ the second term in the brackets on the right side of (6) is negligible compared with the first. Then taking $I_M(x) \approx \exp x$ leads to a quadratic equation in A whose solution

$$
A = \frac{\gamma}{(2M)^{\frac{1}{2}}\, d} + \left[\frac{\gamma^2}{2M\, d^2} + A_c^2 + \frac{(M+1)}{2}\, A_0^2\right]^{\frac{1}{2}} \tag{15}
$$

approximates the required attenuation over the range specified.

For $\gamma^2/M \ll 1$, on the other hand, one may take $I_M(x) \approx (x/2)^M/M!$ in (6). Some further approximations and manipulation lead to the result

$$
A \approx A_{cs}[1 + (\gamma^2/2M)]^{\frac{1}{2}} \qquad \gamma^2/M \ll 1 \tag{16}
$$

which is exact for $\gamma = 0$.

When the solution to (6) is found for given parameters the minimum average exclusion cost can be determined. For the probability density function (5), (3) becomes

$$
l = 1 - Q_M[\alpha,\, A(2)^{\frac{1}{2}}] + \nu Q_M[\alpha,\, Ad(2)^{\frac{1}{2}}] \tag{17}
$$

where

$$Q_M(x, y) = \int_y^\infty \zeta(\zeta/x)^{M-1} \exp\left[-(\zeta^2 + x^2)/2\right] I_{M-1}(x\zeta)\, d\zeta \qquad (18)$$

is the generalized Marcum Q-function.[3]

IV. EFFECT OF COST RATIO ON REQUIRED ATTENUATION

Since one cannot easily decide how much better or worse it is for the receiver to saturate at its upper limit than at its lower limit, the ratio, ν, is generally difficult to assess accurately. It is interesting (even fortunate!) that here as in the single diversity case,[1] the solutions obtained using (14) show that the optimum receiver attenuations for a wide range of cost ratios do not differ appreciably from those for the minimum exclusion probability case ($\nu = 1$).

For given ν, γ, d and M one may define the sensitivity, S_c, of the optimum attenuation to the cost ratio by the difference in required attenuation between the given case and the corresponding minimum exclusion probability case. Specifically, the sensitivity to cost ratio is

$$S_c(\nu, \gamma, d, M) \triangleq 20 \log_{10} A(\nu, \gamma, d, M) - 20 \log_{10} A(\nu = 1, \gamma, d, M) \qquad (19)$$

in which the functional dependence is shown explicitly. In (19) a positive value of S_c indicates an increase in required attenuation compared with the minimum exclusion probability case. It can be shown that (*i*) the sign of (19) depends only on ν (positive if $\nu > 1$, negative if $\nu < 1$), and, (*ii*) $| S_c(\nu, \gamma, d, M) | \leqq | S_c(\nu, 0, d, M) |$. Thus, one can define a maximum sensitivity

$$S_c^* \triangleq S_c(\nu, 0, d, M) \equiv 10 \log_{10} A_{cs}^2 - 10 \log_{10} MA_0^2 \qquad (20)$$

where the second equality follows from (19) and (10). Using (7) and (8), (20) can be written

$$S_c^* = 10 \log_{10} [1 + (\nu_{\text{dB}}/2MD)] \qquad (21)$$

in which

$$\nu_{\text{dB}} \triangleq 20 \log_{10} \nu \qquad (22)$$

is the cost ratio expressed in dB. S_c^* is an easily calculated bound which gives, with the correct algebraic sign, the maximum change (in dB) of the optimum required attenuation from the optimum for the minimum exclusion probability case. Figure 1 is a plot of S_c^*. It follows from (21)

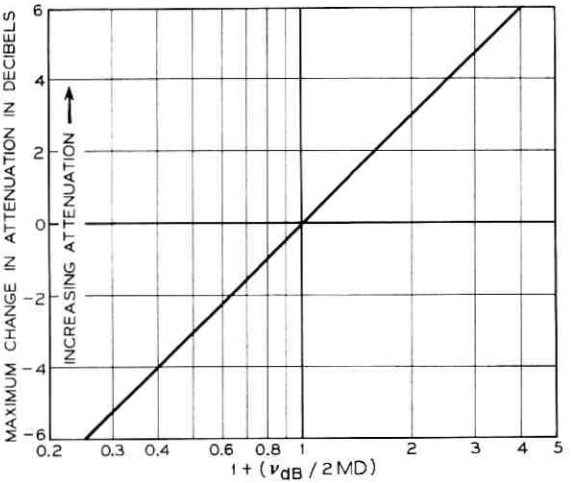


Fig. 1 — Maximum change in required receiver attenuation caused by nonunity cost ratio.

that for cost ratios in the range $-MD \leqq \nu_{dB} \leqq 2MD$, the maximum change in required receiver attenuation is less than about 3 dB. Equivalently, with γ, ν, and d fixed, the maximum sensitivity, S_c^* decreases as the order of diversity, M, increases.

Since the optimum attenuation is extremely insensitive to cost ratio for typical parameters, the minimum exclusion probability case ($\nu = 1$) is of special import among all average cost criteria of the form (1). The *numerical* results presented in this paper, therefore, include only the case $\nu = 1$, although the formulas derived apply more generally and can be used to generate numerical results in an entirely similar manner.

V. EFFECT OF DIVERSITY ON REQUIRED ATTENUATION

It is interesting to consider how the order of diversity affects the optimum required attenuation. Accordingly, in a manner analogous to (19) one can define the difference in required receiver attenuation resulting from diversity by

$$S_m(\nu, \gamma, d, M) = 20 \log_{10} A(\nu, \gamma, d, M) - 20 \log_{10} A(\nu, \gamma, d, M = 1). \quad (23)$$

Let

$$\Delta_n(\nu, \gamma, d, M) \triangleq 20 \log_{10} A(\nu, \gamma, d, M) \tag{24}$$

be the required normalized attenuation in dB. Then (23) can be written

$$S_m(\nu, \gamma, d, M) = \Delta_n(\nu, \gamma, d, M) - \Delta_n(\nu, \gamma, d, 1). \tag{25}$$

Values of $S_m(1, \gamma, d, M)$ were obtained for various γ, d, and M using (14) and (23). These are shown in Fig. 2 where all quantities except M are in dB. The optimum normalized attenuation $\Delta_n(1, \gamma, d, M)$ required for the minimum exclusion probability case can be found using (25). Specifically one finds $S_m(1, \gamma, d, M)$ from Fig. 2, and adds to it the quantity $\Delta_n(1, \gamma, d, 1)$ from Fig. 6 in Ref. 1. Notice that in Ref. 1 only the single diversity case ($M = 1$) was considered so that the functional dependence of Δ_n on M was suppressed in the notation. That is, $\Delta_n(\nu, \gamma, d, 1)$ here is identical to $\Delta_n(\nu, \gamma, d)$ in Ref. 1.

From Fig. 2 it can be seen that if γ is sufficiently small (or large), S_m is positive (or negative) so that more (or less) attenuation is required if multiple diversity is used than would be required if the same specular energy were concentrated in a single diversity branch or pulse. Also for sufficiently small (or large) γ the required attenuation increases (or decreases) as the order of diversity, M, increases. There is of course a transition region which bridges the above cases and in which, for γ fixed, the differences S_m cross one another depending on the particular values of M, and D (and, in the general case, ν). The curves for $\gamma_{dB} = 15$, for example, exhibit this behavior.

Using (10) and (16) it can be shown that for $\gamma \to 0$

$$S_m(\nu, \gamma, d, M) \approx 10 \log_{10} M + 10 \log_{10} \left[\frac{1 + (\nu_{dB}/2MD)}{1 + (\nu_{dB}/2D)}\right] + 10 \log_{10} \left[\frac{1 + (\gamma^2/2M)}{1 + (\gamma^2/2)}\right] \tag{26}$$

which is exact for $\gamma = 0$. For the minimum exclusion probability case and $\gamma = 0$ (26) yields $S_m(1, 0, d, M) = 10 \log_{10} M$ which is independent of d. Similarly it can be shown using (15) that

$$\lim_{\gamma \to \infty} S_m(\nu, \gamma, d, M) = -10 \log_{10} M \tag{27}$$

which is independent of ν and d. The differences S_m for $\gamma_{dB} = \pm\infty$ therefore appear as horizontal lines in Fig. 2. One also observes that over the range of parameters shown, the limit (27) is approached within

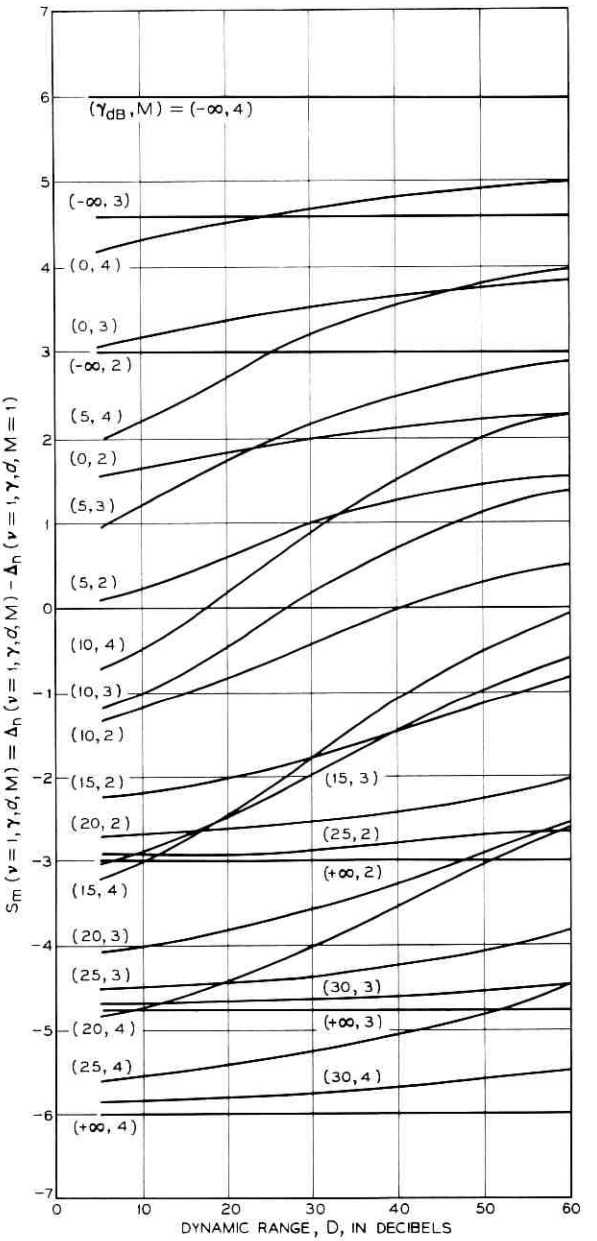


Fig. 2 — Differences in required optimum attenuation resulting from diversity. $\nu = 1$.

0.5 dB for $\gamma_{dB} = 30$. Noting that for $\gamma^\circ < \gamma < \gamma^*$, S_m is bounded by

$$S_m(\nu, \gamma^*, d, M) < S_m(\nu, \gamma, d, M) < S_m(\nu, \gamma^\circ, d, M) \qquad (28)$$

it follows from (26) through (28) that for the minimum exclusion probability case all the differences $S_m(1, \gamma, d, M)$ lie between two horizontal lines in Fig. 2 determined only by the order of diversity. $| S_m(1, \gamma, d, M) | \leqq 10 \log_{10} M$. The difference between the required optimum attenuation for dual diversity and that required for single diversity with the same total received specular energy is less than about 3 dB.

VI. EXCLUSION COSTS FOR DIVERSITY RECEIVERS

The optimum normalized attenuations obtained using the iteration equations (14) were used to obtain the minimum average exclusion costs (17) for the case $\nu = 1$. These are shown in Fig. 3(a) and for smaller values of D in Fig. 3(b). The generalized Q function (18) was evaluated by computer, using relations derived from those given by Sagon.[5]

It can be seen that for small values of γ, the smallest dynamic range D required to obtain a given exclusion probability decreases rapidly as the order of diversity is increased; the most substantial decrease is obtained in going from single to dual diversity. This trend is lessened as the available signal-to-noise ratio γ increases. As a matter of fact if γ is sufficiently large (for example, $\gamma = 20$ dB) the dynamic range required to achieve a given exclusion probability increases as M increases. However at the large values of γ where this latter effect is apparent, even modest values of D yield extremely small exclusion probabilities. Moreover on the types of channels where diversity receivers are useful one would generally encounter small values of γ.

Consider a diversity receiver operating in a small signal-to-noise ratio and let the dynamic range of the components used be such that the probability of excluding the signal at any point in the receiver is the same throughout. Then it follows from Fig. 3 and the foregoing discussion that the dynamic range required of the components used in the post-combining portions of the receiver may be considerably smaller than that required of those components used in the individual diversity branches.

VII. SUMMARY AND CONCLUSIONS

An important class of diversity receivers used for communications through fading media and for radar is considered. The required gain

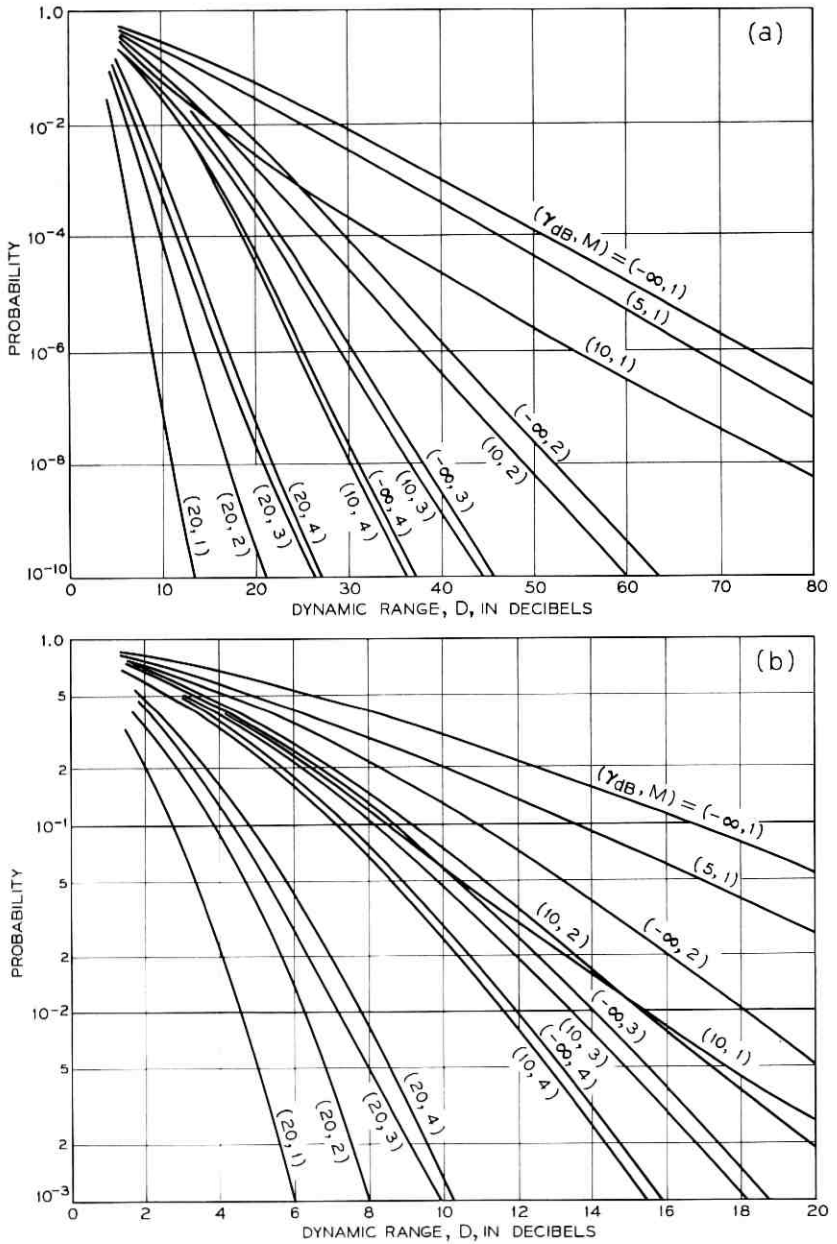


Fig. 3 — Minimum exclusion probabilities for diversity receivers.

is determined which minimizes the average cost of excluding from a finite dynamic range the signal appearing in the post-combining portions of the receiver. For the single diversity case ($M = 1$) the results reduce to those given previously.

It is shown that the required receiver gain is extremely insensitive to the relative 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 relative costs in the range $-MD \leqq \nu_{dB} \leqq 2MD$. One also finds that the sensitivity to relative cost therefore decreases as the order of diversity increases.

The difference between the required optimum receiver gain for various orders of diversity M, and that required for a single diversity receiver having the same total received specular energy is considered. Exact differences are given for the minimum exclusion probability case, and it is shown that these are less than $10 \log_{10} M$ dB independent of other parameters. Bounds on the difference are also given for non-unity cost ratio.

Performance characteristics derived show minimum exclusion probabilities obtainable as a function of dynamic range for various signal-to-noise ratios and orders of diversity. For a small signal-to-noise ratio the dynamic range required of the components used in the post-combining portions of the receiver can be considerably smaller than that required of those components in the individual diversity branches in order to achieve uniform exclusion probability throughout.

Notice that in some applications the normalization assumed in writing (4) and (5) may depend upon M and γ. This fact must be accounted for if one is calculating the *actual* required attenuation from the required *normalized* attenuation discussed in Sections IV and V. The optimum exclusion costs however, depend on the normalized attenuation and not on the normalizing factor. The results of Section VI therefore apply directly.

VIII. ACKNOWLEDGEMENT

I offer my thanks to A. Anastasio, who helped program the computer to evaluate the generalized Q-function.

REFERENCES

1. Rappaport, S. S., "Communications and Radar Receiver Gains for Minimum Average Cost of Excluding Randomly Fluctuating Signals in Random Noise," B.S.T.J., *46*, No. 8 (October 1967), pp. 1753–1773.
2. Rice, S. O., "Mathematical Analysis of Random Noise," in *Selected Papers*

on Noise and Stochastic Processes, ed. by N. Wax, New York: Dover Press, 1954, p. 239.
3. Helstrom, C. W., *Statistical Theory of Signal Detection,* New York: Pergamon Press, 1960, pp. 171–174.
4. Lindsey, W. C., "Error Probabilities for Partially Coherent Diversity Reception," IEEE Trans. Comm. Theory., *COM-14,* No. 5 (October 1966), pp. 620–625.
5. Sagon, H., "Numerical Calculation of the Incomplete Toronto Function," Proc. IEEE, *54,* No. 8 (August 1966), p. 1095.

Multimoding and its Suppression in Twisted Ring Counters

By W. BLEICKARDT

(Manuscript received March 21, 1968)

Many digital systems, such as PCM systems, data processing and data transmission systems, use twisted ring counters. Most of these twisted ring counters are subject to multimoding. This paper develops tools and methods for predicting all possible modes in twisted ring counters, and derives a general solution for suppressing the wrong modes. Suppression is accomplished by adding a few circuit connections from the output of certain stages to the input of another stage. The paper derives the number of necessary connection lines and their connection points for the various types of counters.

I. INTRODUCTION

Twisted ring counters of various types have been used for many years, and have been described in many publications.[1–5] They are designed for creating a well-defined periodic pulse pattern. But they all have one problem in common: under certain circumstances they can multimode, that is, they can create undesired patterns. Each mode of a counter creates a particular pattern. Only one of these modes is the desired one, the "correct mode;" the rest are all "wrong modes" and must be suppressed. To the knowledge of the author, none of the publications on twisted ring counters presents a rigorous treatment of the problem of multimoding, although it must have shown up in many instances and often was solved empirically.[5] The lack of a general theory on possible modes in twisted ring counters and on the prevention of undesired modes led to this investigation.

Terminology for the characterization of modes, and relations between the parameters, make it easy to find the entire set of possible modes for any twisted ring counter. There is a method for suppressing all wrong modes by adding a few circuit connections, and a general formula that indicates these additional connections for any individual

ring counter. The method for suppressing all wrong modes in any twisted ring counter is summarized in Section 5.5.

II. OPERATION OF TWISTED RING COUNTERS

A twisted ring counter consists of a shift register whose output is fed back over a twist to its input in a ringlike manner (Figs. 1, 2, and 3). An input clock keeps a certain pattern circulating around the ring. In the correct mode the stages create the desired pattern by switching on sequentially with subsequent clock pulses, and then switching off in the same sequence (part a of Figs. 1, 2, and 3).* With each clock pulse only one stage is switching. A counter with n stages creates a periodic pattern with a period of $2n$ time slots as shown in the first three figures. Some possible implementations of counter stages are shown in Fig. 4, using AND gates, NAND gates and set-reset flip-flops. Equivalent stages can be built by using OR gates and NOR gates, or any custom-designed circuit.

There are two general types of twisted ring counters: single-phase counters with one input clock line (example in Fig. 1), and double-phase counters with two input clock lines supplying interleaved pulses (examples in Figs. 2 and 3). Many of the single-phase counter stages, such as the ones shown in Fig. 4a and b, require short input clock pulses to prevent racing. The clock pulses must be shorter than the propagation delay of one stage. An example of a stage that does not require short clock pulses is shown in Fig. 4c.[4] Double-phase counters permit the use of simple gated set-rerest flip-flop stages (Fig. 4d and e) without the restriction of short clock pulses. Notice that in counters with an even number of stages (Fig. 2) the two clock phases are distributed in a different way from those in counters with an odd number of stages (Fig. 3).

The problem of multimoding arises whenever more than one mode can exist. In that case, errors can switch the counter to other (wrong) modes with undesired patterns. Such errors can be created by noise transients, aging components, marginal design, and so on. The first three figures show some examples of wrong modes. In general, the number of wrong modes possible increases with the number of stages of a counter, and is higher for single-phase counters than for double-phase counters. To design reliable circuits, one must prevent un-

* The numbers in parentheses in Figs. 1, 2, and 3 are a symbolic notation for different modes; they indicate the numbers of time slots a particular counter stage remains in one state. This notation is explained in Section III.

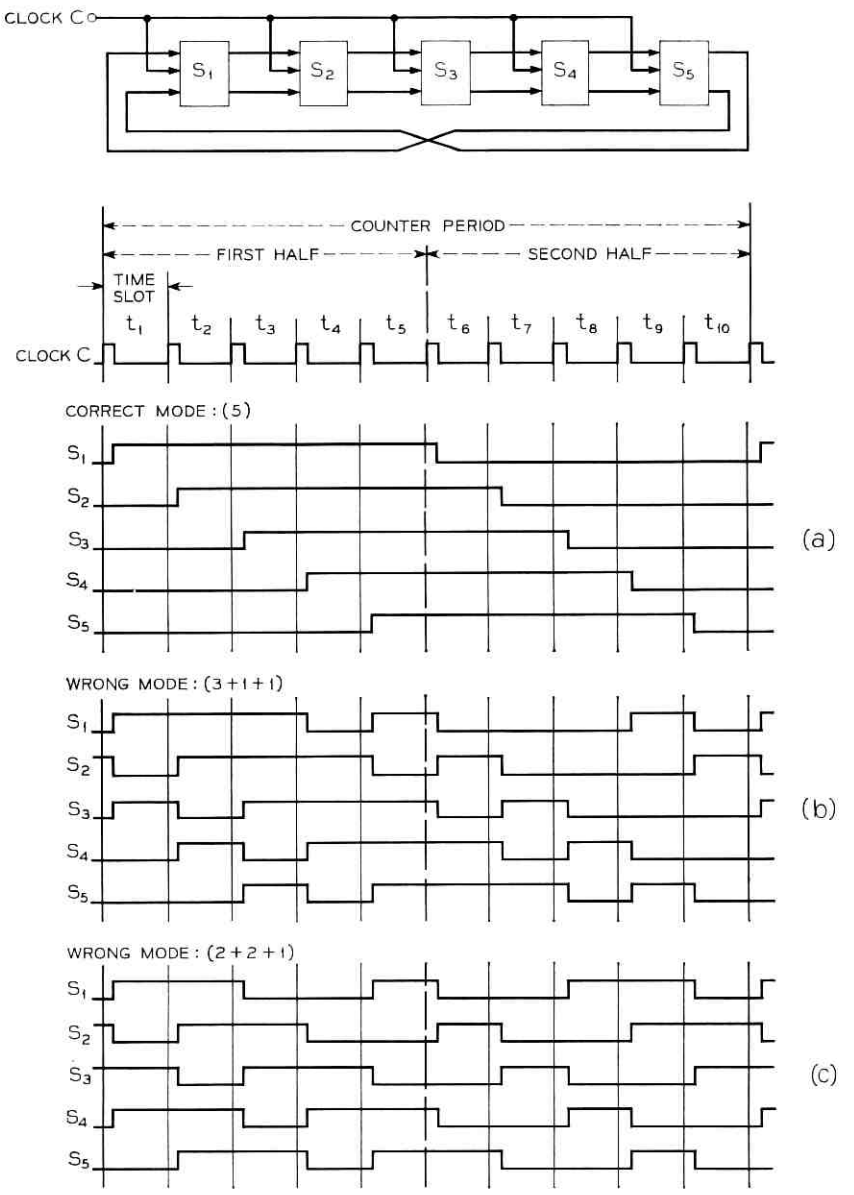


Fig. 1 — Single-phase twisted ring counter with five stages.

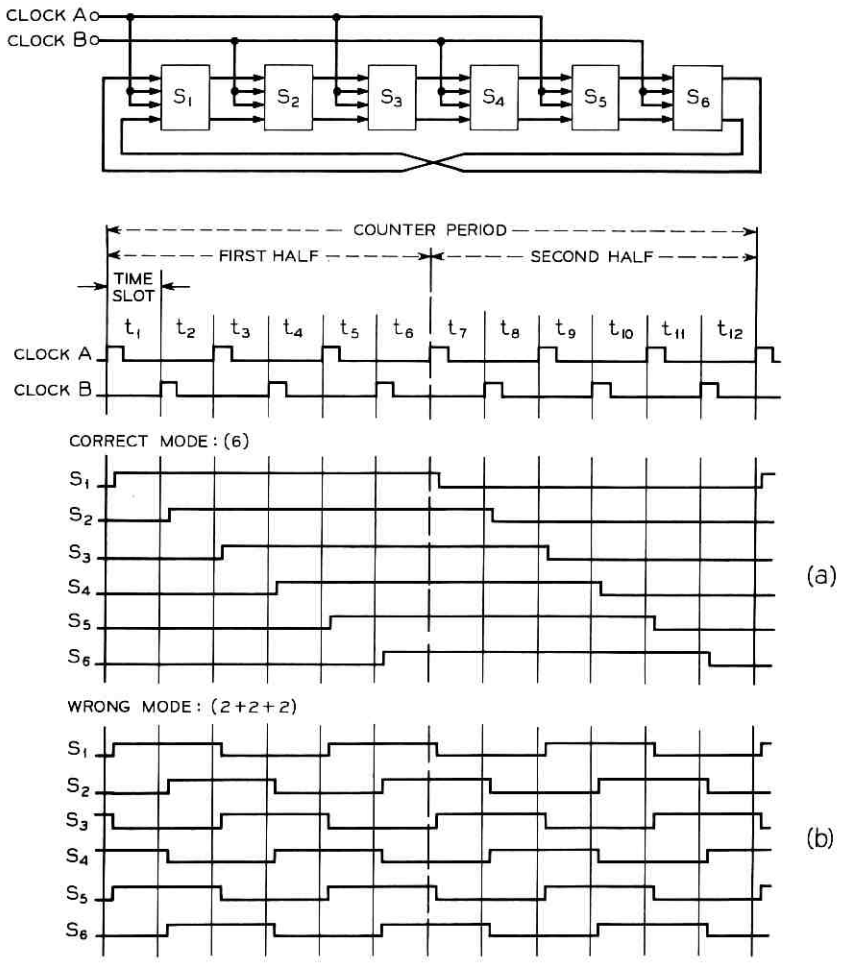


Fig. 2 — Double-phase twisted ring counter with even number of stages (six stages).

desired patterns from circulating for more than a very short time (typically less than one counter period).

III. GENERAL CHARACTERIZATION OF MODES

There is a unique way in which a pattern, that is, a sequence of states 0 or 1, is circulated around the counter ring. Any pattern is shifted by one stage per time slot, as can be seen from the pulse dia-

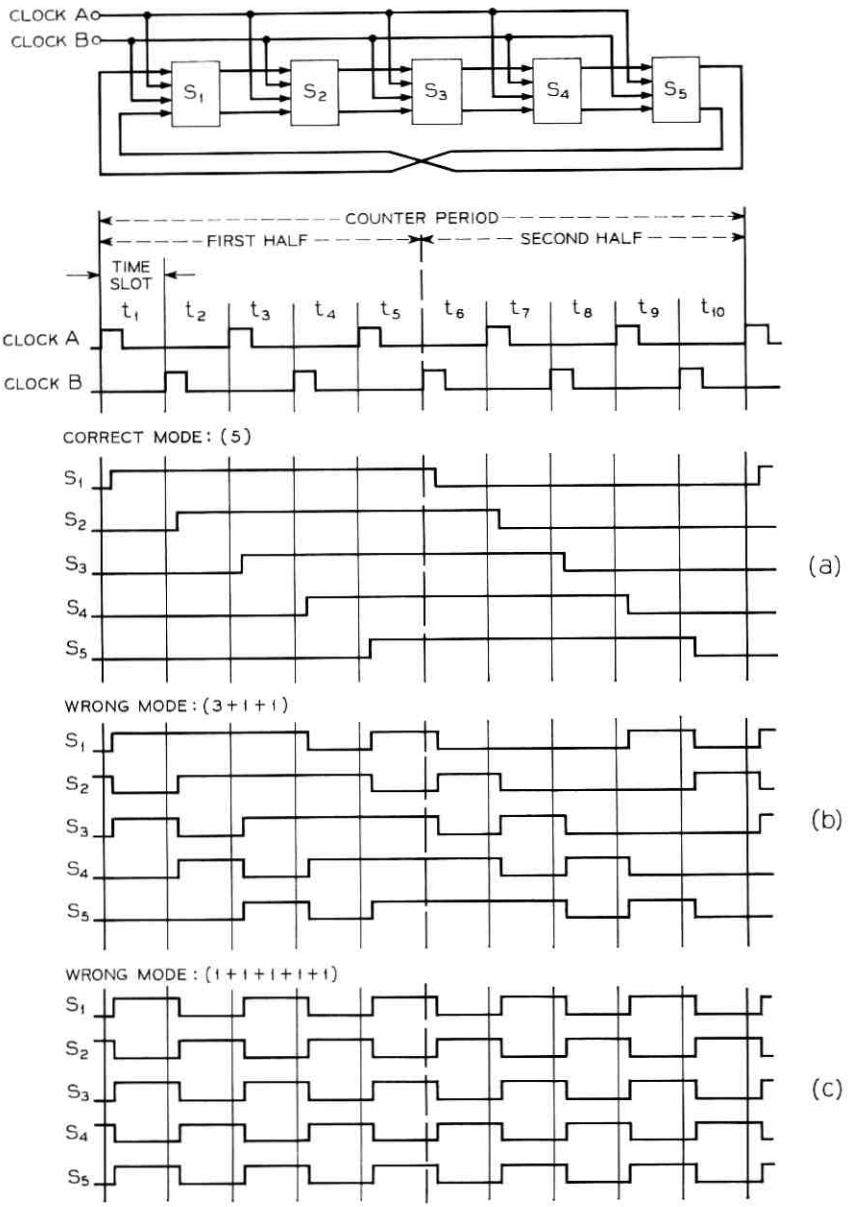


Fig. 3 — Double-phase twisted ring counter with odd number of stages (five stages).

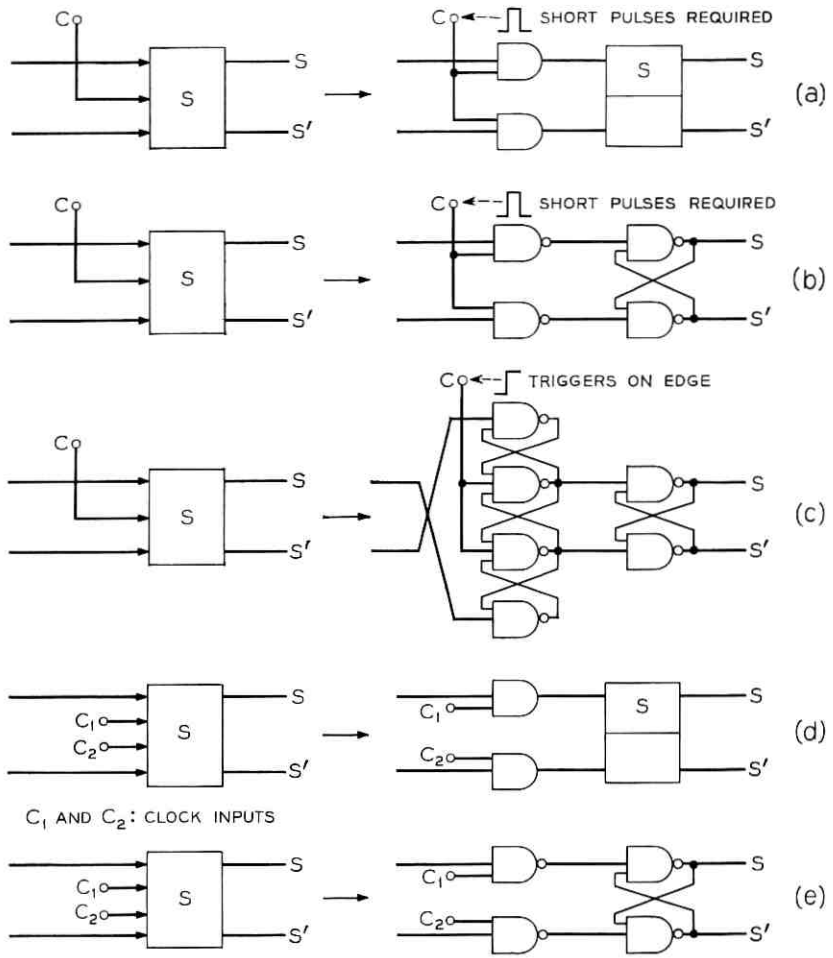


Fig. 4 — Implementation of counter stages with AND gates, NAND gates, and set-reset flip-flops; a, b, and c for single-phase counters, d and e for double-phase counters.

grams in Figs. 1, 2 and 3. The state of the last stage appears in inverted form at the first stage in the subsequent time slot. For a counter with n stages, the pattern, seen as a time sequence at each stage, repeats itself in inverted form after n time slots; the whole counter period is $2n$ time slots long.

This well-defined behavior allows us to reconstruct the entire pulse

diagram for a particular mode, if we only know the states of all n stages at any one time, or if we know a sequence of n states at any single stage. Therefore, a sequence of n binary digits uniquely describes a mode.

3.1 *Definitions*

(*i*) We will call the state of a particular stage in a particular time slot an *element*. An element can have a state 0 or 1.

(*ii*) Elements in successive time slots, or in successive stages that have the same state, form a *logic group*.

(*iii*) The *size* of a logic group (g_j) is the number of its elements.

(*iv*) The smallest size logic group of a particular mode has g_{min} elements.

(*v*) The positive direction of a sequence of elements corresponds to the sequence as observed on the positive time axis. This corresponds to a sequence backwards through the stages. (This can be illustrated with Fig. 1b. The sequence 1 1 1 0 1 appears at stage S_1 in the time slot sequence t_1, t_2, t_3, t_4, t_5, and it appears at time t_5 in the stage sequence S_5, S_4, S_3, S_2, S_1.)

3.2 *Description*

For describing one particular mode, it is sufficient to write the size and sequence of the logic groups g_j that are built by n elements. The following symbolic notation is used:

$$(g_1 + g_2 + g_3 + \cdots + g_x)$$

where

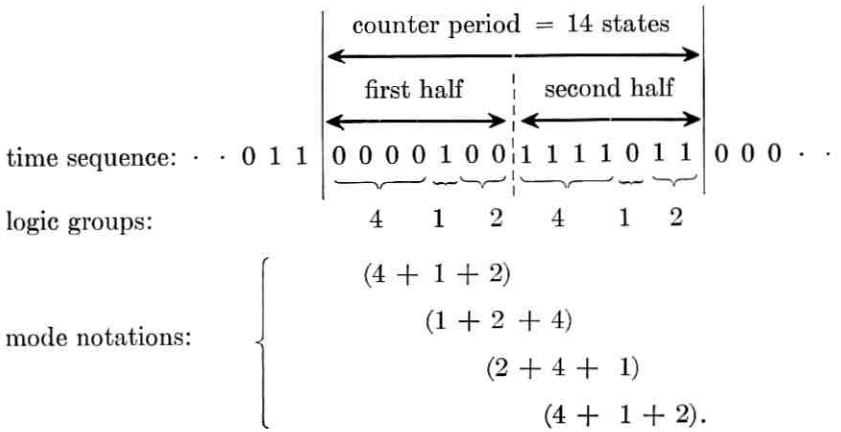
$$\sum_{j=1}^{x} g_j = n = \text{number of stages}$$

$$x = \text{odd number.}$$

For example, (3 + 1 + 1) denotes a mode of a 5-stage counter, with three logic groups, the first containing three elements, the second and third containing one element each (shown in Fig. 1b).

This symbolic notation describes one half of the periodic cycle. Since each half is always the complement of the other half, the elements of the first and the last logic group in the mode notation have the same state. Therefore, the number x of logic groups in this notation is always an odd number. This is illustrated with a 7-stage counter, for which a time sequence of states, as observed on the oscillo-

scope connected to one of the stages, may look like this:



The period consists of $2n = 14$ states. Describing this particular mode, the logic groups built by $n = 7$ elements can be written in three different ways: $(4 + 1 + 2)$, $(1 + 2 + 4)$, and $(2 + 4 + 1)$.

These mode notations are cyclic permutations. Hence they are equivalent and describe the same mode. The 7-stage counter could have another mode with the same set of logic groups. This different mode can be described by the following three equivalent mode notations: $(4 + 2 + 1)$, $(2 + 1 + 4)$, and $(1 + 4 + 2)$. If a certain wrong mode can exist, all possible permutations can exist also.

The correct mode always is the one with $x = 1$, that is, with one single logic group of size n. All other possible modes with $x \geqq 3$ are wrong modes.

IV. PREDICTION OF POSSIBLE MODES

4.1 *Possible Logic Groups*

Not all possible partitions of n into an odd number x of logic croups result in a possible mode, because there are some restrictions in possible logic group sizes g_j for the different counter types.

In single-phase counters, the logic groups can have any even or odd number of elements, up to n, since in any time slot, either a "1" or a "0" can be shifted from any stage to the following stage (Fig. 1). This is not so in double-phase counters.

In double-phase counters with an even number of stages (Fig. 2), a clock pulse A can shift either a "1" or a "0" to any odd-numbered

stage from the preceding stage, and a clock pulse B can shift either a "1" or a "0" to any even-numbered stage from the preceding stage. This results in the restriction that only logic groups with an even number of elements can appear in a possible mode.

In double-phase counters with an odd number of stages (Fig. 3), a clock pulse A can shift a "1" to any odd-numbered stage and a "0" to any even-numbered stage, and a clock pulse B can shift a "0" to any odd-numbered stage and a "1" to any even-numbered stage, always from the preceding stage. This results in the restriction that only logic groups with an odd number of elements can appear in a possible mode.

4.2 *Examples of Possible Modes*

We are now able to predict all possible modes of a twisted ring counter with n stages by breaking n into an odd number of logic groups in all possible ways, taking the restrictions of possible logic group sizes into account. This is shown in three examples.

Example 1: A single-phase counter with $n = 6$ stages can have six different possible modes:

(6)	correct mode
(4 + 1 + 1)	wrong modes.
(3 + 2 + 1)	
(3 + 1 + 2)	
(2 + 2 + 2)	
(2 + 1 + 1 + 1 + 1)	

In this counter type, the logic groups can have an even or odd number of elements.

Example 2: A double-phase counter with an even number of $n = 6$ stages (Fig. 2) has only two possible modes:

(6)	correct mode
(2 + 2 + 2)	wrong mode.

In this counter type, the logic groups can only have an even number of elements. Because of this restriction, there are always fewer wrong modes than in a single-phase counter with the same number of stages.

Example 3: A double-phase counter with an odd number of $n = 9$

stages has ten different possible modes:

(9)	correct mode
(7 + 1 + 1)	wrong modes.
(5 + 3 + 1)	
(5 + 1 + 3)	
(5 + 1 + 1 + 1 + 1)	
(3 + 3 + 3)	
(3 + 3 + 1 + 1 + 1)	
(3 + 1 + 3 + 1 + 1)	
(3 + 1 + 1 + 1 + 1 + 1 + 1)	
(1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1)	

In this counter type, the logic groups can only have an odd number of elements. In general, the higher the number n of stages, the higher is the number of wrong modes.

4.3 *Experimental Verification of Predicted Modes*

Many counters of the three types shown in Figs. 1, 2, and 3 have been built, with various numbers of stages, and with different types of stages, including all types shown in Fig. 4. All of the predicted modes for these counters have actually been observed. Any desired mode can be induced by presetting all stages before turning the clock pulses on, but only the possible modes will be able to circulate without being altered.

V. SUPPRESSION OF WRONG MODES

All wrong modes can be suppressed by adding a certain small number of circuit connections. A general method for finding the necessary and sufficient additional connections for any twisted ring counter is to find criteria that are common to all wrong modes but do not appear in the correct mode. By suppressing these criteria, all wrong modes will be prevented. To find these common criteria, it is useful to define the concept of common logic groups.

5.1 *Common Logic Groups*

For a particular counter, the common logic groups represent the set consisting of the smallest logic groups (g_{min}) from each wrong

mode. For example, the 9-stage double-phase counter, whose wrong modes are listed in example 3 of Section 4.2, has two common logic groups of sizes 1 and 3. Each wrong mode contains at least one of the common logic groups. Taking the g_{min}-values of all wrong modes as common logic groups results in the smallest possible set of logic groups with the property of each wrong mode containing at least one of these logic groups.

The size of the smallest common logic group (m_k) is equal to the smallest g_{min}-value of all wrong modes, $g_{min\ min}$. That is

$g_{min\ min} = 1$ for single-phase counters,
$g_{min\ min} = 1$ for double-phase counters with odd number of stages, and
$g_{min\ min} = 2$ for double-phase counters with even number of stages.

The size of the largest common logic group (m_0) is equal to the largest g_{min}-value of all wrong modes, that is, $g_{min\ max}$:

$$m_0 = g_{min\ max} \quad \text{of} \quad \sum_{j=1}^{x} g_j = n$$

with $x \geqq 3$ for wrong modes. Every possible partition of the above sum represents a possible wrong mode with a certain value g_{min}. The maximum of this value for all possible partitions is $g_{min\ max}$. It occurs with the minimum value of $x = 3$ and is

$$g_{min\ max} \leqq n/3.$$

The largest common logic group is therefore

$$m_0 \leqq n/3, \tag{1}$$

the next possible logic group size equal or less than $n/3$. This is

$$m_0 \geqq (n - 2)/3 \quad \text{for single-phase counters,} \tag{2}$$

$$m_0 \geqq (n - 4)/3 \quad \text{for double-phase counters.} \tag{3}$$

This results is only a single m_0-value in each case, when the restrictions of possible logic group sizes are taken into account. Combining the latter and expressions (1), (2), and (3) into a single expression, we get for the largest common logic group m_0:

$$m_0 = \frac{n}{3} - \frac{2p}{3} \cdot \Delta \qquad 0 \leqq \Delta \leqq 1 \tag{4}$$

with Δ chosen to make m_0 an integer, and

n = number of stages
p = 1 for single-phase counters

$p = 2$ for double-phase counters
$m_0 = 1, 2, 3, 4, \cdots$ for single-phase counters
$m_0 = 2, 4, 6, 8, \cdots$ if n = even } for double-phase counters.
$m_0 = 1, 3, 5, 7, \cdots$ if n = odd }

The set of common logic groups for a particular counter consists of the smallest and the largest common logic groups and all possible sizes of logic groups between. It is given in Table I for counters up to 20 stages. For single-phase counters with two stages and for double-phase counters with two or four stages there are no common logic groups, since these counters do not have any wrong mode.

5.2 *Suppressing the Common Logic Groups*

Suppressing all common logic groups in a counter leads, by definition, to the prevention of all possible wrong modes, and does not introduce any new modes. This section shows that there is a subset of common logic groups (Table II) whose suppression is sufficient for

TABLE I — COMMON LOGIC GROUPS

(Common logic groups are all different g_{min} values of all wrong modes)

Number of stages n	For single-phase counters	For double-phase counters	
		With even number of stages	With odd number of stages
2	—	—	
3	1		1
4	1	—	
5	1		1
6	2 1	2	
7	2 1		1
8	2 1	2	
9	3 2 1		3 1
10	3 2 1	2	
11	3 2 1		3 1
12	4 3 2 1	4 2	
13	4 3 2 1		3 1
14	4 3 2 1	4 2	
15	5 4 3 2 1		5 3 1
16	5 4 3 2 1	4 2	
17	5 4 3 2 1		5 3 1
18	6 5 4 3 2 1	6 4 2	
19	6 5 4 3 2 1		5 3 1
20	6 5 4 3 2 1	6 4 2	

suppressing all common logic groups and is thereby sufficient for preventing all wrong modes.

5.2.1 *Method of Suppressing a Group*

If we want to suppress a particular common logic group of size m_i, we must prevent one of the following two patterns consisting of an undesired sequence of ones and zeros

$$\cdots 0\,0\,1\,1\,1\,1\,1\,1\,1\,1\,0\,0\cdots$$

$$\text{or} \quad (\cdots 1\,1\,\underbrace{0\,0\,0\,0\,0\,0\,0\,0\,0}_{m_i \text{ elements}}\,1\,1\cdots)$$

from circulating around the counter ring. The inverse pattern, in parentheses, always appears with the first one. This suppression can be accomplished by preventing stage S_x from switching from "0" ("1") to "1" ("0") whenever stage S_{x-1-m_i} is in state "0" ("1"). The position of the patterns immediately before suppression is:

$$\begin{array}{ccc} & S_{x-1-m_i} \qquad\qquad S_{x-1}\ S_x \\ & \downarrow \qquad\qquad\quad \downarrow\downarrow \\ \cdots & 0\,0\,1\,1\,1\,1\,1\,1\,1\,1\,0\,0\cdots \end{array}$$

$$\text{or} \quad (\cdots 1\,1\,\underbrace{0\,0\,0\,0\,0\,0\,0\,0\,0}_{m_i \text{ elements}}\,1\,1\cdots).$$

If stage S_x does not switch to "1" ("0") with the next clock pulse, the logic group of size m_i is prevented from passing through stage S_x. It is sufficient to suppress only one of the two patterns, since the inverse of it is then suppressed automatically.

This suppression can be implemented by adding a circuit connection from the output of stage S_{x-1-m_i} to the input of stage S_x, preventing S_x from switching from "0" to "1" whenever S_{x-1-m_i} is in state "0." This circuit connection, shown in Fig. 5a, bridges m_i stages, and therefore is called a "bridging connection"; its associated parameter m_i is called a "bridging parameter."

The bridging connection could also be made on the inverse side of the stages S_{x-1-m_i} and S_x, thus preventing S_x from switching from "1" to "0" whenever S_{x-1-m_i} is in state "1." These two bridging connections are equivalent, and one of them is sufficient. However, if both connections are applied for each m_i-value, a wrong mode is cleared within half a counter period instead of a full period.

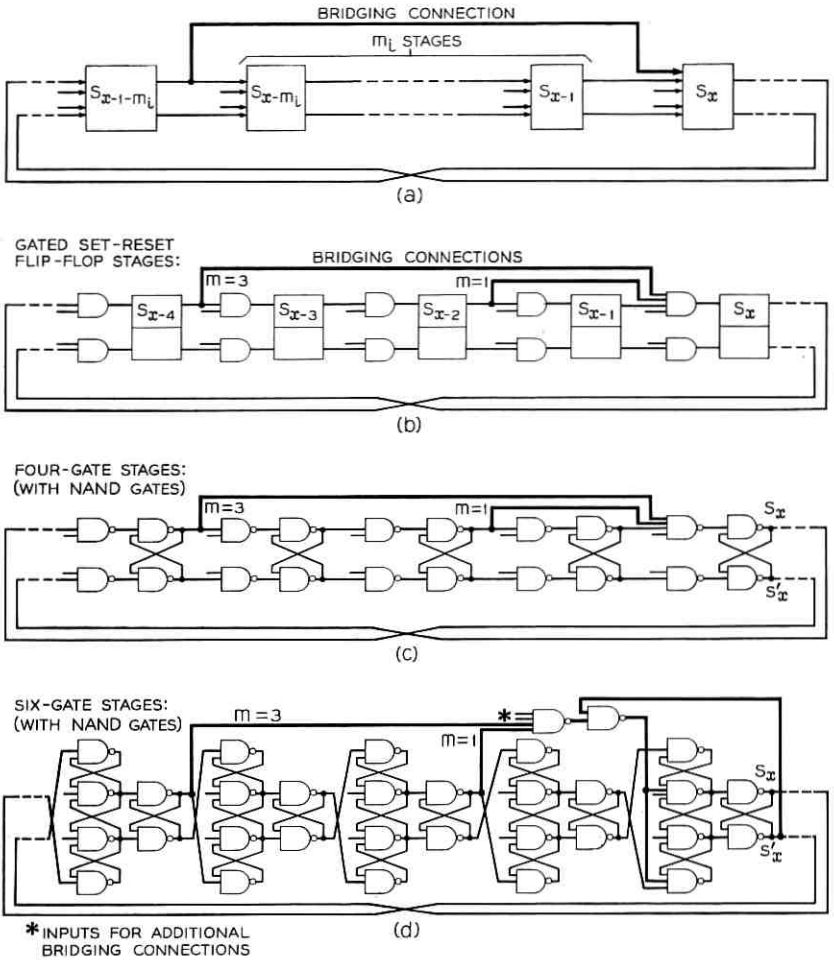


Fig. 5 — Suppression of wrong modes by adding bridging connections, bridging m_i stages. Part a shows the principle; b, c, and d show an example with the two bridging connections $m = 3$ and $m = 1$ for counters with different types of stages.

S_x may be any particular stage of the counter, but it should be the same stage for all bridging connections (although this is not essential with many counters). The correct mode is not affected by this inhibition, since in the correct mode S_{x-1-m_i} is always in state "1" ("0") when S_x is switched from "0" ("1") to "1" ("0") because m_i is always smaller than n.

5.2.2 *Automatically Eliminated Wrong Modes*

If we suppress one common logic group of size m_i by using the described method, we prevent the pattern

$$\cdots 0 \underbrace{\cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot}_{m_i} 1\ 0 \cdots$$

and hence eliminate not only the wrong modes containing the common logic group m_i, but also all other wrong modes that show this pattern at any one position.

The remaining wrong modes, which do not contain this pattern, require additional steps for their prevention. It can be shown that if a mode with $g_{min} < m_i$ satisfies both of the following two conditions, it does not contain the above pattern and therefore is not eliminated by suppressing m_i:

(*i*) All possible sums of the elements of an even number $2v$ of consecutive logic groups must be $\leqq m_i$ for at least one value of v ($v = 1$ or 2 or 3 $\cdots$).

(*ii*) All possible sums of the elements of an odd number $2v + 1$ of consecutive logic groups must be $\geqq m_i + 1$ for the same value of v that satisfies condition *i*.

Example: Suppose we have a single-phase counter with 19 stages, and we suppress the common logic group of size $m_i = 6$ by adding a bridging connection bridging 6 stages as shown in Fig. 5a. Would the mode (2 + 3 + 3 + 1 + 5 + 1 + 4) be suppressed?

We check whether this mode satisfies both conditions. Condition *i* is satisfied with $v = 1$, since all pairs of consecutive numbers in the mode notation (2 + 3, 3 + 3, 3 + 1, 1 + 5, 5 + 1, 1 + 4, 4 + 2) sum up to $\leqq 6$. That is, all sums of the elements of a pair ($2v$) of consecutive logic groups are $\leqq m_i$. Condition *i* could not be satisfied with $v > 1$ in this example. Condition *ii* is also satisfied with $v = 1$, since all triplets of consecutive numbers in the mode notation (2 + 3 + 3, 3 + 3 + 1, 3 + 1 + 5, 1 + 5 + 1, 5 + 1 + 4, 1 + 4 + 2, 4 + 2 + 3) sum up to $\geqq 7$. That is, all sums of the elements of a triplet ($2v + 1$) of consecutive logic groups are $\geqq m_i + 1$.

The above mode satisfies both conditions, and therefore would not be eliminated by suppression of the common logic group of size $m_i = 6$.

5.2.3 *Sufficient Subset*

Suppression of a particular common logic group of size m_i generally does not prevent wrong modes with $g_{min} > m_i$, but it does prevent some of the wrong modes with $g_{min} < m_i$. In the remaining unsuppressed modes with $g_{min} < m_i$, which all satisfy the two conditions stated in Section 5.2.2, the largest possible g_{min}-value, called $g_{min\ max}$, follows from condition i:

$$\sum_{j=1}^{2v} g_j \leqq m_i \qquad (v = 1 \text{ or } 2 \text{ or } 3 \cdots).$$

Every possible partition of this sum delivers a value g_{min}. The maximum of these g_{min}-values for all possible partitions is $g_{min\ max}$. It occurs with the minimum value of $v = 1$ and is

$$g_{min\ max} \leqq m_i/2.$$

This is the next lower common logic group size m_{i+1} that must be suppressed:

$$m_{i+1} \leqq m_i/2. \tag{5}$$

m_{i+1} is the next possible logic group size equal to or less than $m_i/2$, which is

$$m_{i+1} \geqq (m_i - 1)/2 \quad \text{for single-phase counters,} \tag{6}$$

$$m_{i+1} \geqq (m_i - 3)/2 \quad \text{for double-phase counters.} \tag{7}$$

This results in only a single m_{i+1}-value in each case, when the restrictions of possible logic group sizes are taken into account.

Combining the restrictions and the inequalities (5), (6), and (7) into a single expression, we get for the next lower common logic group m_{i+1} that must be suppressed:

$$m_{i+1} = \frac{m_i}{2} - (p - \tfrac{1}{2}) \cdot \Delta \qquad 0 \leqq \Delta \leqq 1 \tag{8}$$

with Δ chosen to make m_{i+1} an integer, and

$p = 1$	for single-phase counters
$p = 2$	for double-phase counters
$m_{i+1} = 1, 2, 3, 4, \cdots$	for single-phase counters
$m_{i+1} = 2, 4, 6, 8, \cdots$ if n = even $m_{i+1} = 1, 3, 5, 7, \cdots$ if n = odd	for double-phase counters.

If we suppress m_i, it is sufficient to supress m_{i+1} as the next lower common logic group, since suppression of m_i prevents all wrong modes with $m_i \geqq g_{min} > m_{i+1}$. Recursion formula (8) determines the maximum spacing of successive common logic group sizes m_i to be suppressed for sufficiently suppressing all common logic groups within the covered range. By extending this range from the largest common logic group m_0 to the smallest common logic group m_k, we get the sufficient subset of common logic groups

$$m_0 , m_1 , m_2 , \cdots , m_k$$

that must be suppressed for preventing all wrong modes. m_0 is determined by expression (4); m_1 through m_k are obtained by expression (8).

5.2.4 *Necessary Subset*

The m_i-values resulting from expressions (4) and (8)

$$m_0 , m_1 , m_2 , \cdots , m_i , \cdots , m_k$$

always represent a *sufficient* subset of common logic groups to be suppressed for preventing all wrong modes. But for some particular counters, the *necessary* subset $m_0 , m_1 , m_2 , \cdots , m_i$ may be smaller by a few m_i-values. That is, the smallest values $m_{i+1} \cdots m_k$ of the set are not necessary. There is not a simple expression like (4) and (8) for giving only the necessary m_i-values but, for a particular counter, they may be found by using the two conditions in Section 5.2.2, which have not yet been used to their full extent in Section 5.2.3. In a first step, the last value m_k is left off and a check is made whether any wrong mode exists that could satisfy both conditions for the remaining m_i-values. Such modes can be found by listing all possible combinations of logic groups that satisfy those two conditions (for $1 \leqq v \leqq m_i/2$). If there is no mode consisting entirely of these listed combinations, m_k is not necessary. In the next step, m_{k-1} is left off, repeating the procedure, until the last necessary value m_i is found.

For counters up to 20 stages, Table II gives the sufficient m_i-values (bridging parameters) according to expressions (4) and (8), with the unnecessary ones in parentheses.

5.3 *Implementation in Different Counter Circuits*

Each bridging parameter m_i denotes one bridging connection, bridging m_i stages, which has to be added to prevent wrong modes (as described in Section 5.2.1 and shown in Fig. 5a). Figures 5b, c, and d

TABLE II — BRIDGING PARAMETERS m_i

(Numbers without parentheses denote the necessary and sufficient bridging connections.)

Number of stages n	For single-phase counters m_i	For double-phase counters	
		With even number of stages m_i	With odd number of stages m_i
2	*	*	
3	1		1
4	1	*	
5	1		1
6	2 (1)	2	
7	2 1		1
8	2 (1)	2	
9	3 (1)		3 (1)
10	3 1	2	
11	3 (1)		3 (1)
12	4 2 (1)	4 (2)	
13	4 2 1		3 (1)
14	4 2 (1)	4 2	
15	5 2 (1)		5 (1)
16	5 2 (1)	4 (2)	
17	5 2 (1)		5 (1)
18	6 3 (1)	6 (2)	
19	6 3 1		5 (1)
20	6 3 (1)	6 2	

* No bridging parameters because these counters have no wrong modes.

show the bridging connections for the values $m = 3$ and $m = 1$ for counters with different types of stages. In counters with 6-gate stages, as shown in Fig. 5d, additional gates are required for proper suppression of common logic groups without impairment of the correct mode. For not impairing the correct mode, a feedback connection is required from the output of stage S_x. These counters need one additional gate if there is one bridging connection or two additional gates if there is more than one bridging connection.

As an example, we obtain for a 3-stage single-phase counter only one bridging parameter $m_0 = m_j = 1$. This means that only one bridging connection is needed, bridging one stage. Figure 6 shows three possible locations of the bridging connection. If bridging connections pass the twist, they must also be twisted, as illustrated in Figs. 6b and c.

For double-phase counters with an odd number of stages, one also has to make sure that the signal from stage S_{x-1-m_i} does not reach stage

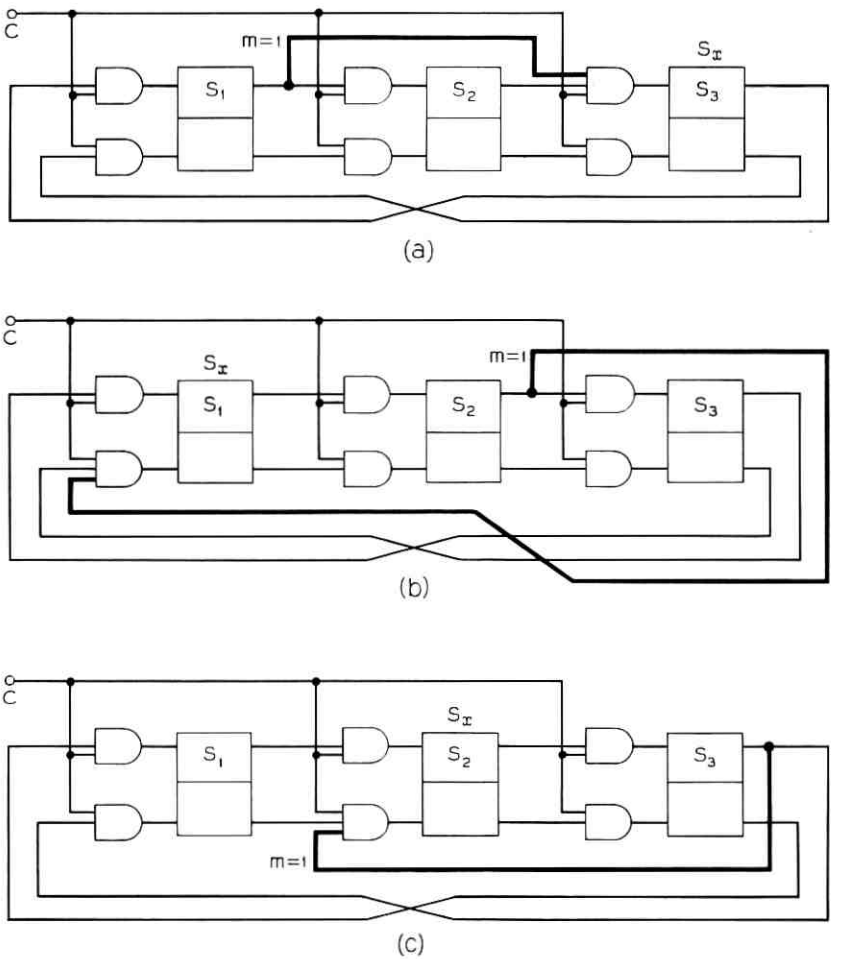


Fig. 6 — Single-phase counter with three stages requiring one bridging connection, $m = 1$. Parts a, b, and c are three equivalent solutions. If a bridging connection passes the twist, as in b and c, it must also be twisted.

S_x earlier than the signal from stage S_{x-1} caused by the same clock pulse. Otherwise a pattern $\cdots 1 + 1 + 1 + 1 + 1 \cdots$ might not be prevented under certain worst case propagation delays of the logic circuits involved. It is easy to assure this timing condition if logic gates are used that also provide a complementary output (as is the case in emitter-coupled gates). Figure 7 shows such an example with NOR/OR gates. In

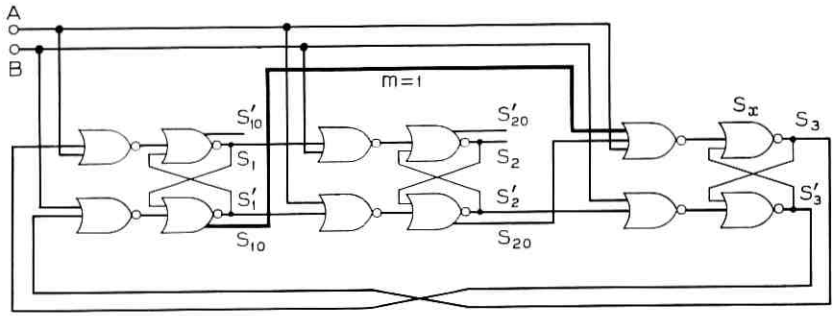


Fig. 7 — Double-phase counter with three stages (with NOR/OR gates). Use of complementary gate outputs (OR output S_{10} instead of NOR output S_1, and S_{20} instead of S_2) for increasing the permissible gate propagation delay tolerance range.

the case of a wrong mode (1 + 1 + 1) in this double-phase counter with three stages, output S_{10} appears one gate propagation delay later than S_1 upon an input pulse A, and output S_{20} one propagation delay earlier than S_2 upon the same input pulse A. This is sufficient to meet the above timing condition. If complementary gate-outputs are not available, a small delay may be introduced into the bridging connections. This additional timing condition does not exist in single-phase counters and in double-phase counters with an even number of stages.

5.4 *Experimental Verification*

Proper suppression of all wrong modes by bridging connections determined according to the described procedures has been verified experimentally with counters of all three types (Figs. 1, 2, and 3), with different stages (Fig. 4) and with many different values of n. Counters for which the necessary set of bridging connections is smaller than the sufficient set resulting from the formulas were given special attention.

5.5 *Summary: Suppression of Wrong Modes*

A small number of additional circuit connections (bridging connections) are sufficient for suppressing all wrong modes in a twisted ring counter. The bridging connections are determined by the bridging parameters m_i, which can be found by the formula:

$$m_0 = \frac{n}{3} - \frac{2p}{3} \cdot \Delta \qquad 0 \leqq \Delta \leqq 1$$

$$m_{i+1} = \frac{m_i}{2} - (p - \tfrac{1}{2}) \cdot \Delta$$

with Δ chosen to make m_0 and m_{i+1} integers, and

$$
\begin{aligned}
n &= \text{number of counter stages} \\
p &= 1 && \text{for single-phase counters} \\
p &= 2 && \text{for double-phase counters} \\
m_i &= 1, 2, 3, 4, \cdots && \text{for single-phase counters} \\
m_i &= 2, 4, 6, 8, \cdots \text{ if } n = \text{even} \\
m_i &= 1, 3, 5, 7, \cdots \text{ if } n = \text{odd} && \text{for double-phase counters} \\
i &= 0, 1, 2, 3, \cdots, j, \cdots, k.
\end{aligned}
$$

Each of the resulting bridging parameters m_i denotes one bridging connection in the circuit, which bridges m_i stages (Fig. 5). The bridging connection can be located anywhere in the counter ring; if it passes the twist, it must also be twisted. See Fig. 6.

The resulting $k + 1$ bridging parameters denote a sufficient set of $k + 1$ bridging connections in every case. For certain counters, however, the necessary set of $j + 1$ bridging connections is slightly smaller; it can be determined by the procedure described in Section 5.2.4.

Table II gives the $k + 1$ bridging parameters according to the above formula for different counter types up to 20 stages. The bridging parameters denoting unnecessary bridging connections according to the above procedure are in parentheses.

VI. CONCLUSION

Tools and methods for predicting and suppressing wrong modes in twisted ring counters have been developed. As a result we have gained a better insight into the multimoding mechanism and obtained a simple method for preventing multimoding. This method is summarized, and the required additional circuit connections are given in Table II for counters up to 20 stages.

VII. ACKNOWLEDGMENTS

I wish to thank Mr. D. Koehler for many helpful suggestions and his critical reading of the manuscript, and Mr. W. J. Kusner for experimentally verifying various results of this study.

REFERENCES

1. Ware, W. H., "The Logical Principles of a New Kind of Binary Counter," Proc. IRE, *41*, No. 10 (October 1953), pp. 1429–1437.

2. Tarczy-Hornoch, Z., "Five-Binary Counting Technique Makes Faster Decimal-Counting Units," Elec. Design, *9,* No. 2 (January 1961), pp. 34–37.
3. Millman, J., and Taub, H., *Pulse, Digital, and Switching Waveforms,* New York: McGraw-Hill Book Co., Inc., 1965, p. 682.
4. Boag, T. R., "Practical Applications of an IC Counter/Shift Register," Elec. Design News, *10,* No. 7 (June 1965), pp. 18–33.
5. "TTL Integrated Circuits: Counters and Shift Registers," Texas Instruments, Inc., Application Report SC 10660 (March 1968), pp. 28–29.

Synthesis of Rational Transfer Function Approximations Using a Tapped Distributed RC Line With Feedback

By DAVID A. SPAULDING

(Manuscript received June 26, 1968)

This paper describes a simple procedure for synthesizing an active distributed RC network which, by using dominant poles and zeros, realizes a very accurate approximation of an arbitrary stable rational transfer function. The network uses a single uniformly distributed RC line with taps spaced along its length. A linear combination of tap voltages is added to the input signal to form the driving voltage for the RC line; the output signal is also a linear combination of the tap voltages.

The network offers a number of significant advantages. Since it realizes a nearly rational transfer function, the approximation problem can be conveniently solved using readily available results on rational function approximation. Also, the network uses only one uniform RC line, the transfer function can be changed simply by changing resistor values, and the frequency can be scaled by minor connection changes. Thus one standard network with minor modification is useful for a wide variety of applications.

This paper develops the design procedure and derives the various sensitivity functions of importance. Two example designs are carried out: an approximation to a second-order low-pass transfer function and an approximation to a second-order band-pass transfer function with a Q of 100. *The sensitivities for the examples are very reasonable and the measurements made on laboratory models indicate excellent agreement with theoretical predictions.*

I. INTRODUCTION

The progress being made in miniaturizing electronic circuits has stimulated a continuing interest in the synthesis of networks using distributed RC components. Numerous techniques are available for synthesizing transfer functions using distributed RC components in conjunction with various active network elements.[1] Generally, these synthesis procedures are applicable only if the transfer function has

a very special form. This form is not a rational function of the complex frequency variable s but involves hyperbolic functions of s. If the problem posed to the network designer were merely to realize given transfer functions of this special form using distributed RC networks, there would be no difficulty.

However, the problem is generally not this but rather to realize a network which achieves certain system specifications such as band-limiting or pulse shaping. Thus, a realizable transfer function must be developed which approximates the specifications (that is, the approximation problem must be solved) before a network can be synthesized. Because the transfer functions realizable by distributed RC networks have a somewhat complicated form, the approximation part of the network designer's work is more difficult when using distributed RC networks. This fact has led to a continuing effort to develop distributed RC networks which realize rational transfer functions. Since rational functions are easier to manipulate, and many applicable results are readily available in the literature, the approximation problem is made much easier. This paper develops a simple procedure and network for realizing an accurate approximation to a rational transfer function using an active network incorporating a distributed RC line.

Available techniques for synthesizing rational transfer functions using distributed RC networks are documented by Heizer, Barker, Woo and Hove, and Fu and Fu.[2-6] Each of these techniques uses the fact, first demonstrated by Heizer, that some of the immittance parameters of a distributed RC line can be made rational functions of s by cutting the conducting layer of the RC line in a particular manner.

These synthesis techniques have some definite disadvantages. They require two RC lines with cuts in the conducting layer which depend upon the transfer function being realized; this is undesirable from a manufacturing point of view and makes tuning difficult. Also, the synthesis procedure involves a test to determine that the curve cut in the conductor satisfies certain restrictions, that is, it does not "attempt" to create a negative capacitance in the line. If it does, a new try at the design is required. Fu and Fu eliminate this problem at the expense of a significant increase in circuit complexity.[6]

Recently techniques have become available for approximating rational transfer functions by using the dominant poles and zeros of distributed networks. A few representative approaches are those of

Kerwin, Bello and Gausi, and Wyndrum. Kerwin's approach, in general, requires the use of lumped components.[7] Bello and Gausi consider only low-pass transfer functions and use different configurations to realize an arbitrary transfer function.[8] Wyndrum's technique also deals only with low-pass transfer functions.[9]

The synthesis technique described here offers advantages over the other available techniques since only one uniformly distributed RC line is used and it is capable of realizing an accurate approximation to an arbitrary transfer function. In addition, the design procedure is very simple.

II. TRANSFER FUNCTION OF UNIFORM RC LINE WITH FEEDBACK

Chen and Levine[10, 11] have suggested that filters could be built using a uniform RC line driven by an input voltage source and having the output formed as a linear combination of the voltages appearing along the line as in Fig. 1. This procedure is useful in some cases but is not general enough because it synthesizes transfer functions by using zeros of transmission. What is needed in addition to zeros are poles; poles can be realized by using feedback as in Fig. 2.

The network of Fig. 2 consists of a uniform RC line with taps spaced along its length. The tap voltages are appropriately scaled by the infinite input impedance coefficients a_i and added to the input signal to form the driving voltage for the line. The output voltage is the sum of the tap voltages appropriately scaled by the infinite input impedance coefficients b_i. The RC line is the three-layer structure shown in Fig. 3, where it will be assumed that there is no voltage variation in the y direction.

To determine the voltage transfer function of the network of Fig.

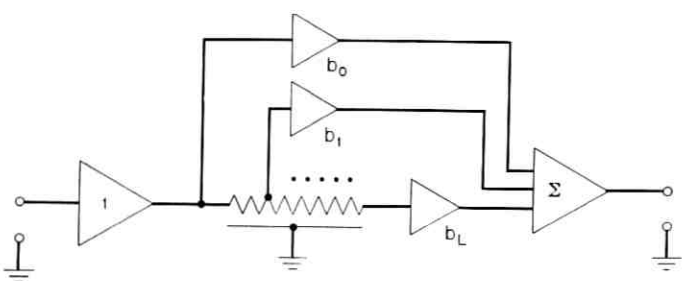


Fig. 1 — Tapped RC line.

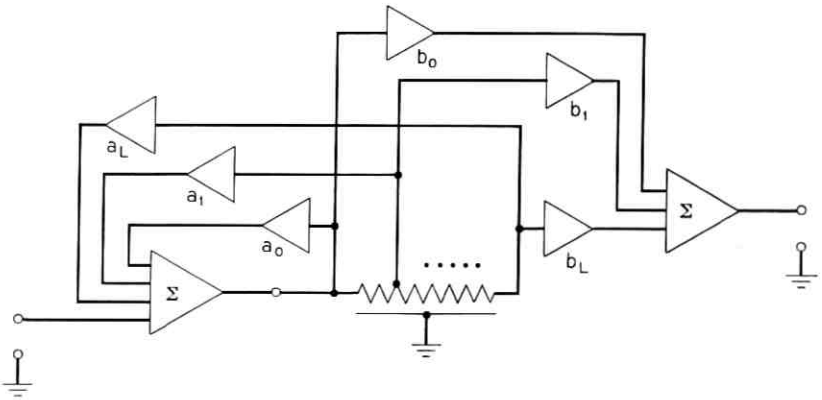


Fig. 2 — Tapped RC line with feedback.

2, we first determine the voltage gain $G_i(s)$ from the input of the line to a point x_i meters from the input. The result is[1]

$$G_i(s) = \frac{\cosh (l - x_i)(rcs)^{\frac{1}{2}}}{\cosh l(rcs)^{\frac{1}{2}}},$$

where l is the total length of the line in meters, and r and c are the resistance and capacitance per meter. If the distances l and x_i are constrained to be integral multiples* of some fixed length d_0, that is, $l = Ld_0$ and $x_i = id_0$, and we let $\tau = rcd_0^2$, $G_i(s)$ becomes

$$G_i(s) = \frac{\cosh (L - i)(\tau s)^{\frac{1}{2}}}{\cosh L(\tau s)^{\frac{1}{2}}}. \tag{1}$$

Using (1) the voltage transfer function of the network of Fig. 2 becomes

$$G(s) = K \frac{\sum_{i=0}^{L} b_i \cosh (L - i)(\tau s)^{\frac{1}{2}}}{\sum_{i=0}^{L} c_i \cosh (L - i)(\tau s)^{\frac{1}{2}}}, \tag{2}$$

where $c_0 = 1$ and $c_i = -a_i$ for $i \neq 0$†. The real constant K is such that $b_i = 1$ for the smallest i for which $b_i \neq 0$. By making the substitution $p = \exp (\tau s)^{\frac{1}{2}}$ and factoring the resulting polynomials in p, it can be

* For any set of x_i and l, a small enough d_0 can be found that error in this assumption is negligible.

† α_0 has been set to zero which can be done without any loss of generality.

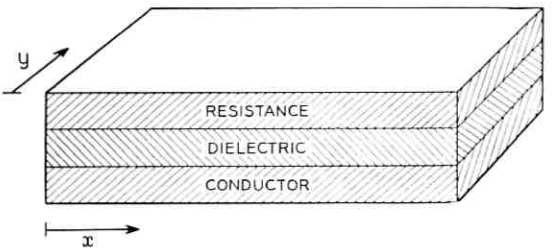


Fig. 3 — Uniform RC line.

shown that (2) can be factored into the form

$$G(s) = K \frac{2^{R-L} \prod_{i=1}^{R} [\cosh (\tau s)^{\frac{1}{2}} - Z_i]}{\prod_{i=1}^{L} [\cosh (\tau s)^{\frac{1}{2}} - P_i]} \tag{3}$$

where $1 \leqq R \leqq L$ [unless the numerator in (2) is unity in which case the numerator of (3) is 2^{1-L}] and the quantities P_i and Z_i are real or occur in complex conjugate pairs.

Before considering the question of stability, we will determine the locations of the poles and zeros of $G(s)$. Notice that, in spite of the fact that $(s)^{\frac{1}{2}}$ is involved, $G(s)$ is single valued. To determine the pole (zero) locations, we set the denominator (numerator) factors in (3) equal to zero and solve for s. For a typical denominator factor $(\cosh (\tau s)^{\frac{1}{2}} - P_i)$ we calculate the s-plane pole positions to be

$$\tau s_i = \ln^2 | p_i | - (\arg p_i + 2n\pi)^2 + 2j \ln | p_i | (\arg p_i + 2n\pi) \tag{4}$$

where $n = 0, \pm 1, \pm 2, \ldots$ and $p_i = P_i + (P_i^2 - 1)^{\frac{1}{2}}$. The term p_i comes from the solution of a quadratic equation which has two roots. However, these roots are always reciprocals of one another and, as can be seen from the form of (4), these two values of p_i give the same s_i. Hence, only one of them need be used. A simple check shows that each of the poles resulting from the single term $(\cosh (\tau s)^{\frac{1}{2}} - P_i)$ as given by (4) is simple.*

When P_i is real, the s_i given by (4) are on the negative real axis for $| P_i | \leqq 1$ and occur in complex conjugate pairs for $| P_i | > 1$. When P_i is complex, (3) involves a term $[\cosh (\tau s)^{\frac{1}{2}} - P_i^*]$ which gives poles that are the complex conjugates of those of (4).

It is easy to see from (4) that the infinite set of poles generated by one

* For $P = \pm 1$ double roots occur but not for $P = +1$ with $n = 0$.

denominator factor lie on a parabola given by

$$\sigma = \frac{\ln^2 | p_i |}{\tau} - \frac{\omega^2 \tau}{4 \ln^2 | p_i |}. \tag{5}$$

Figure 4 shows the location of these poles in the normalized, $\tau = 1$, s-plane. The poles due to the term $[\cosh (\tau s)^{\frac{1}{2}} - P_i]$ are indicated by single circles and those due to the term $[\cosh (\tau s)^{\frac{1}{2}} - P_i^*]$ by double circles. Similar comments hold in the case of numerator factors in (3).

Knowing the locations of the poles of $G(s)$ permits the question of stability to be answered easily. For simplicity we assume that in (2) $b_0 = 0$. If this is not the case, $G(s)$ can be separated into the sum of a constant plus a $\hat{G}(s)$ which is of the form of (2) where $\hat{G}(s)$ has the same denominator as $G(s)$ but different numerator and $b_0 = 0$. The constant gain is stable. With $b_0 = 0$, $G(s)$ is stable, that is, its impulse response remains bounded for large values of time, if all the poles lie in the left

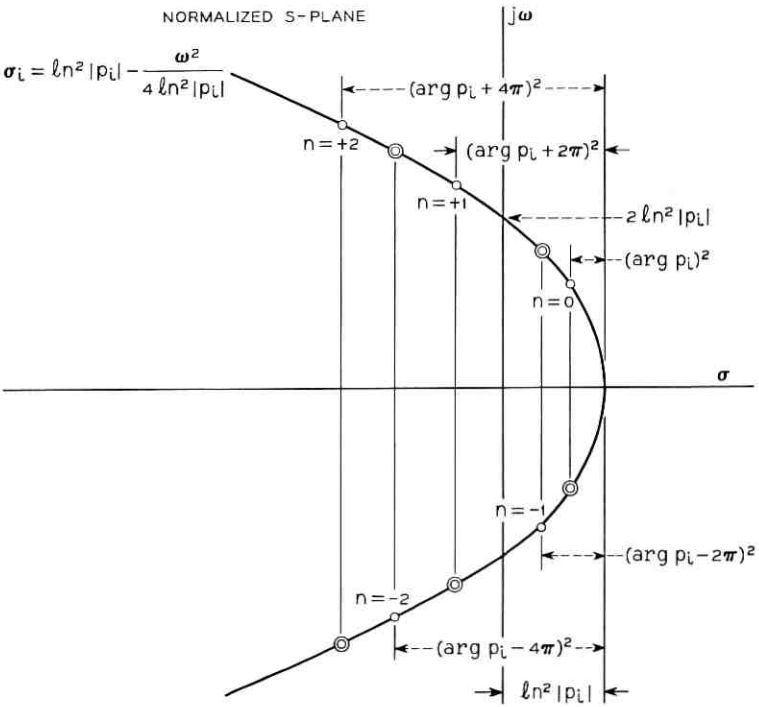


Fig. 4 — S-plane roots resulting from a pair of complex conjugate factors in $G(s)$.

half of the s-plane and those on the $j\omega$ axis are simple. This result can be proved by finding the inverse transform of $G(s)$ by using the Cauchy residue calculus.[12]

Notice that $G(s)$ is a meromorphic function and $G(s) \to 0$ as $s \to \infty$. The Laplace inversion relation is written as

$$g(t) = \lim_{n\to\infty} \frac{1}{2\pi j} \int_{\alpha - jy_n}^{\alpha + jy_n} G(s)e^{st}\, ds.$$

This integral is evaluated by closing the contour in the left half plane so that it does not pass through any poles of $G(s)$ and encloses a finite number of poles. The value of the closed contour integral is determined by the residues of the poles enclosed. As $n \to \infty$ $y_n \to \infty$ and the contours in the left half plane become larger without bound. Using Jordan's lemma[13] the integral over the left half plane contour approaches zero and $g(t)$ is determined. For large values of time the behavior of $g(t)$ is dominated by that pole with the most positive real part. The stability requirement follows directly from this.

III. TRANSFER FUNCTION SYNTHESIS

A glance at Fig. 4 shows that, if the $n = 0$ pole is close to the $j\omega$ axis, the response of the network will approximate that of this single pole alone for values of ω near the pole. An examination of (4) shows that this dominance can always be made to occur by an appropriate selection of τ. From (4) the pole positions in the s-plane are proportional to τ^{-1}. Therefore, by decreasing τ the poles become more widely spaced and hence those near the $j\omega$ axis become more dominant. Since p_i can be adjusted so as to cause the $n = 0$ pole to be arbitrarily close to the s-plane origin, a decrease in τ can be offset, for the $n = 0$ pole, by changing p_i. Therefore, the $n = 0$ pole can be made dominant. Hence, a rational transfer function can be approximated by the system considered here by making its dominant poles and zeros match those of the desired rational function. To calculate the feedback and feed forward coefficients of (2) we calculate the P_i and Z_i of (3) by using the desired pole or zero for s_i in

$$\left.\begin{matrix} P_i \\ Z_i \end{matrix}\right\} = \cosh\,(\tau s_i)^{\frac{1}{2}} \tag{6}$$

and multiply the factors in (3).

The scale factor τ controls the dominance of the $n = 0$ poles and zeros; the dominance improves as τ is reduced. A lower limit on practical

values of τ occurs because the network sensitivity generally deteriorates with reduced values of τ. An upper limit on τ occurs because the $n = 0$ poles are restricted to the shaded area of Fig. 5. If τ is large enough so that a desired dominant pole s_i lies outside the shaded region, the network will realize this pole for a nonzero value of n. It is clear from Fig. 4 that the network will then have an $n = 0$ pole with a more positive real part than that of s_i . This pole can destroy the desired dominance or cause instability if it lies in the right half s-plane. The region permitted for $n = 0$ poles in Fig. 5 is determined from (4) by setting $n = 0$, substituting a value for ω_i and solving for the most negative value of σ_i . The resulting restriction is

$$0 \geqq \sigma_i \geqq \frac{\omega_i^2 \tau}{4\pi^2} - \frac{\pi^2}{\tau}. \tag{7}$$

Except in rather unusual situations τ will be much smaller than the maximum implied by (7).

To synthesize an approximation of a given rational transfer function, the following simple steps are performed.

(*i*) τ is selected so that all the poles of the transfer function lie in the region shown in Fig. 5, and the resulting $n = 0$ poles and zeros realized by the RC line are dominant.

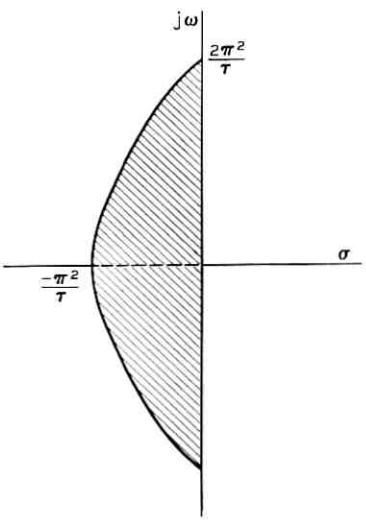


Fig. 5 — Permitted $n = 0$ pole positions in s-plane.

(*ii*) The desired transfer function poles and zeros are used in (6) to determine the P_i and Z_i which, when substituted in (3) and multiplied out, yield the feedback and feed forward coefficients for the network.

(*iii*) The exact response of the network is calculated using (2) or (3) to verify that a good approximation has been achieved.

As pointed out in the examples in Section VI, a wide range of values of τ gives a very accurate approximation. Thus, a little experience will then make step *iii* unnecessary. The selection of τ also affects the sensitivity of the network; hence sensitivity considerations may determine the best value of τ.

IV. SENSITIVITY

One of the most important aspects of any active network synthesis technique is its sensitivity to various parameter variations. In addition, sensitivity results are necessary to show how a physical network may be tuned to achieve an accurate realization of the requirements. Of the several different sensitivity functions that could be derived, we have chosen to consider the relative changes of the poles and zeros with a variety of parameters. These seem to give good physical insight into the behavior of the circuit and result in reasonably concise expressions. The sensitivity functions derived are the relative changes of the poles resulting from relative changes in feedback coefficients, τ, tap positions, and tap loading. Similar results hold for the zero sensitivity functions. The details of the derivations are contained in the Appendix.

If λ_j is the pole in question and the sensitivity of that pole to some parameter X is defined as

$$S_X^{\lambda_j} = \frac{\partial \lambda_j}{\partial X} \frac{X}{\lambda_j},$$

and $P_q = \cosh (\tau\lambda_q)^{1/2}$ where λ_q is the qth pole, then we have the following:*

(*i*) Pole sensitivity to feedback coefficients:

$$S_{a_i}^{\lambda_j} = \frac{a_i \cosh (L - i)(\tau\lambda_j)^{\frac{1}{2}}}{2^{L-2}(\tau\lambda_j)^{\frac{1}{2}} \sinh (\tau\lambda_j)^{\frac{1}{2}} \prod_{\substack{k=1 \\ \neq j}}^{L} (P_j - P_k)}. \tag{8}$$

* λ_j is assumed to be a simple pole.

Simple relations for determining the numerator of (8) are given in Appendix equations (20) and (21).

(*ii*) Pole sensitivity to RC product, τ:

$$S_{\tau}^{\lambda_j} = -1. \tag{9}$$

(*iii*) Pole sensitivity to improper tap spacing*

$$S_{l_i}^{\lambda_j} = -\frac{ia_i \sinh (L - i)(\tau\lambda_j)^{\frac{1}{2}}}{2^{L-2} \sinh (\tau\lambda_j)^{\frac{1}{2}} \prod_{\substack{k=1 \\ \neq j}}^{L} (P_j - P_k)}, \qquad i \neq L \tag{10a}$$

$$S_{l_L}^{\lambda_j} = -\frac{L \sum_{k=0}^{L-1} c_k \sinh (L - k)(\tau\lambda_j)^{\frac{1}{2}}}{2^{L-2} \sinh (\tau\lambda_j)^{\frac{1}{2}} \prod_{\substack{k=1 \\ \neq j}}^{L} (P_j - P_k)}. \tag{10b}$$

(*iv*) Pole sensitivity to tap loading:

$$S_{g_i}^{\lambda_j} = -\frac{g_i R \cosh (L - i)(\tau\lambda_j)^{\frac{1}{2}} \sum_{k=0}^{i-1} c_k \sinh (i - k)(\tau\lambda_j)^{\frac{1}{2}}}{2^{L-2}\tau\lambda_j \sinh (\tau\lambda_j)^{\frac{1}{2}} \prod_{\substack{k=1 \\ \neq j}}^{L} (P_j - P_k)} \tag{11}$$

where g_i is the conductance loading the ith tap and $R = d_0 r$.

V. SECOND ORDER DESIGN EQUATIONS

For the case where $L = 2$, that is, where the RC line is realizing an approximation to a second order transfer function $H(s)$, the design and sensitivity relations given above take on the very simple forms below (λ and ρ, which are complex, are the pole and zero positions in the upper left half s plane):

$$P = \cosh (\tau\lambda)^{\frac{1}{2}}, \qquad Z = \cosh (\tau\rho)^{\frac{1}{2}}$$

$$a_0 = 0,\ a_1 = 4\ \mathrm{Re}(P),\ a_2 = -(1 + 2 \mid P \mid^2)$$

$b_0 = b_1 = 0,\ b_2 = 1$ for $H(s)$ with no finite zeros

$b_0 = 1,\ b_1 = -4\ \mathrm{Re}(Z),\ b_2 = 1 + 2 \mid Z \mid^2$ for $H(s)$ with finite complex zeros

---

* $l_i d_0 = x_i$ where x_i is the distance from the input of the RC line to the ith tap. Nominally $l_i = i$.

$b_0 = 0, b_1 = 1, b_2 = -1$ for $H(s)$ with one zero at zero

$$S_\lambda^\tau = -1$$

$$S_{a_1}^\lambda = \frac{2P \text{ Re } (P)}{j(\tau\lambda)^{\frac{1}{2}} \sinh (\tau\lambda)^{\frac{1}{2}} \text{ Im } (P)}, \qquad S_{a_2}^\lambda = \frac{-(1 + 2 \mid P \mid^2)}{2j(\tau\lambda)^{\frac{1}{2}} \sinh (\tau\lambda)^{\frac{1}{2}} \text{ Im } (P)}$$

$$S_{l_1}^\lambda = \frac{-2 \text{ Re } (P)}{j \text{ Im } (P)}, \qquad S_{l_2}^\lambda = \frac{2P^*}{j \text{ Im } (P)}$$

$$S_{g_1}^\lambda = \frac{-g_1 RP}{2j\tau\lambda \text{ Im } (P)}, \qquad S_{g_2}^\lambda = \frac{g_2 RP^*}{j\tau\lambda \text{ Im } (P)}$$

$$S_{b_0}^\rho = -\frac{2Z^2 - 1}{2j(\tau\rho)^{\frac{1}{2}} \sinh (\tau\rho)^{\frac{1}{2}} \text{ Im } (Z)} \text{ for } H(s) \text{ with finite complex zeros.}$$

For the case where $H(s)$ has two complex zeros, the zero sensitivities are the same as for the poles with ρ and Z replacing λ and P, except for $S_{b_0}^\rho$ which is given. For the case where $H(s)$ has a zero at zero and at infinity, the sensitivity of the zero at zero is infinite (due to the normalization by $1/\rho$), but unnormalized,

$$\frac{\partial \rho}{\partial b_1} = \frac{\partial \rho}{\partial b_2} = -\frac{2}{\tau} \quad \text{and} \quad \frac{\partial \rho}{\partial g_2} = -\frac{2R}{\tau}.$$

Other sensitivities not given are zero.

VI. EXAMPLES

Two examples of approximations to second order rational transfer functions will be worked out and compared with experimental results achieved with a thin film line. The two functions to be approximated are, normalized in frequency,

$$G_1(s) = \frac{(s/4)^2 + 1}{s^2 + (2)^{\frac{1}{2}} s + 1} \tag{12}$$

$$G_2(s) = \frac{0.01\, s}{s^2 + 0.01\, s + 1}. \tag{13}$$

The first is a noncritical low-pass function with a pair of zeros on the $j\omega$ axis and the second is a band-pass function with a Q of 100.

For the low-pass function (12), sensitivity is not a problem because the poles are very low Q. Therefore, τ can be selected to satisfy (7) and to insure dominance of the poles. Letting $\tau = 1$ we have the following results:

$$\lambda = (-1 + j)/(2)^{\frac{1}{2}}, \qquad P = 0.646 + 0.313j,$$

$$\rho = +j4, \qquad Z = 0.342 + j1.9$$

$$a_0 = 0, \quad a_1 = 2.59, \quad a_2 = -2.032$$

$$b_0 = 1, \quad b_1 = -1.36, \quad b_2 = 8.5, \quad K = 0.0544$$

$$S^{\lambda}_{a_1} = 3.33 \angle \mathbf{154}^\circ, \qquad S^{\lambda}_{a_2} = 3.65 \angle -\mathbf{52}^\circ$$

$$S^{\lambda}_{l_1} = 4.12 \angle \mathbf{90}^\circ, \qquad S^{\lambda}_{l_2} = 4.6 \angle -\mathbf{116}^\circ$$

$$S^{\lambda}_{g_1} = 1.15 g_1 R \angle -\mathbf{19}^\circ, \qquad S^{\lambda}_{g_2} = 2.3 g_2 R \angle \mathbf{109}^\circ$$

$$S^{\rho}_{b_0} = 0.508 \angle \mathbf{125}^\circ, \qquad S^{\rho}_{b_1} = 0.16 \angle -\mathbf{137}^\circ, \qquad S^{\rho}_{b_2} = 0.512 \angle -\mathbf{37}^\circ$$

$$S^{\rho}_{l_1} = 0.359 \angle \mathbf{90}^\circ, \qquad S^{\rho}_{l_2} = 2.03 \angle -\mathbf{170}^\circ$$

$$S^{\rho}_{g_1} = 0.126 g_1 R \angle \mathbf{80}^\circ, \qquad S^{\rho}_{g_2} = 0.254 g_2 R \angle \mathbf{100}^\circ$$

Figure 6 shows a block diagram of the experimental circuit, the theoretical response, and the measured results. Notice that the theoretical response realized by the RC line and that of the rational function cannot be distinguished on the scale used for this figure, since they differ by 1 percent at most.

In the case of $G_2(s)$ which has a pole with a $Q = 100$, dominance is achieved for a wide range of values of τ for which (7) holds, and the selection of τ is influenced primarily by sensitivity considerations. The parameter τ affects the sensitivity in a rather complicated way as can be seen from the various sensitivity relations. An examination of the pole sensitivity to coefficient variations has shown that S^{λ}_{a2} has a rather broad minimum in the range $2 \leqq \tau \leqq 14$ and that S^{λ}_{a1} goes to zero in this range when $a_1 = 0$. Therefore, without an exhaustive study to determine an optimum value of τ, we select that value which gives $a_1 = 0$, that is, $\tau = 4.94$. With this value of τ the following result:

$$\lambda = -0.005 + j \qquad P = j2.34$$

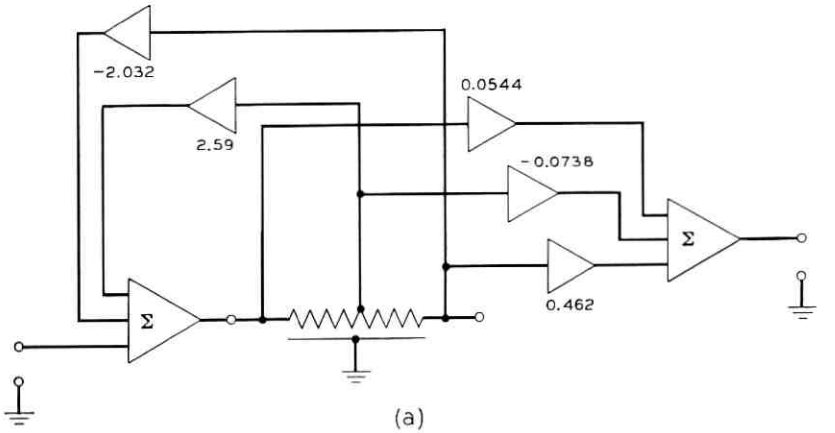
$$a_0 = a_1 = 0 \qquad a_2 = -11.95$$

$$b_0 = 0, \quad b_1 = 1, \quad b_2 = -1 \qquad K = 0.051$$

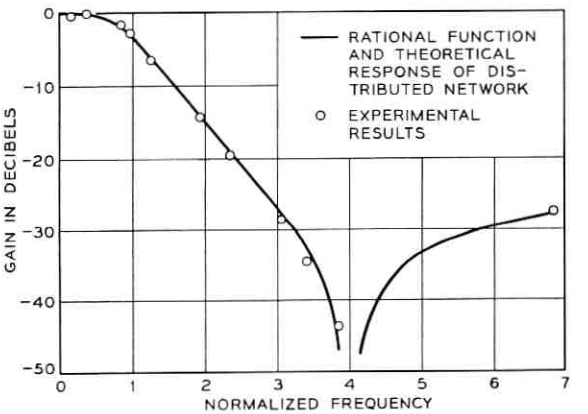
$$S^{\lambda}_{a_2} = 0.452 \angle -\mathbf{45}^\circ \qquad S^{\lambda}_{l_2} = 2 \angle \mathbf{180}^\circ$$

$$S^{\lambda}_{g_1} = 0.1015 g_1 R \angle \mathbf{90}^\circ \qquad S^{\lambda}_{g_2} = 0.203 g_2 R \angle \mathbf{90}^\circ$$

$$\frac{\partial \rho}{\partial b_1} = \frac{\partial \rho}{\partial b_2} = 0.405 \angle \mathbf{180}^\circ \qquad \frac{\partial \rho}{\partial g_2} = 0.405 R \angle \mathbf{180}^\circ \text{ for the zero at zero.}$$



(a)



(b)

Fig. 6 — (a) RC line with feedback approximating $G_1(s)$. (b) Gain vs frequency for $G_1(s)$.

The block diagram of the experimental circuit, the theoretical response, and measured results are shown in Fig. 7. The difference between the theoretical gain of the RC line and that of the rational function is not noticeable since it is approximately 0.1 percent over the frequency range shown in the figure.

The sensitivity of this network is quite acceptable. $S^{\lambda}_{a_2}$ can be controlled by the proper selection of impedance levels; l_2 , a_2 and τ can be stabilized so that the values of $S^{\lambda}_{l_2}$, $S^{\lambda}_{a_2}$ and S^{λ}_{τ} are satisfactory. l_2 should not change after manufacture; a_2 can be made to depend only on the

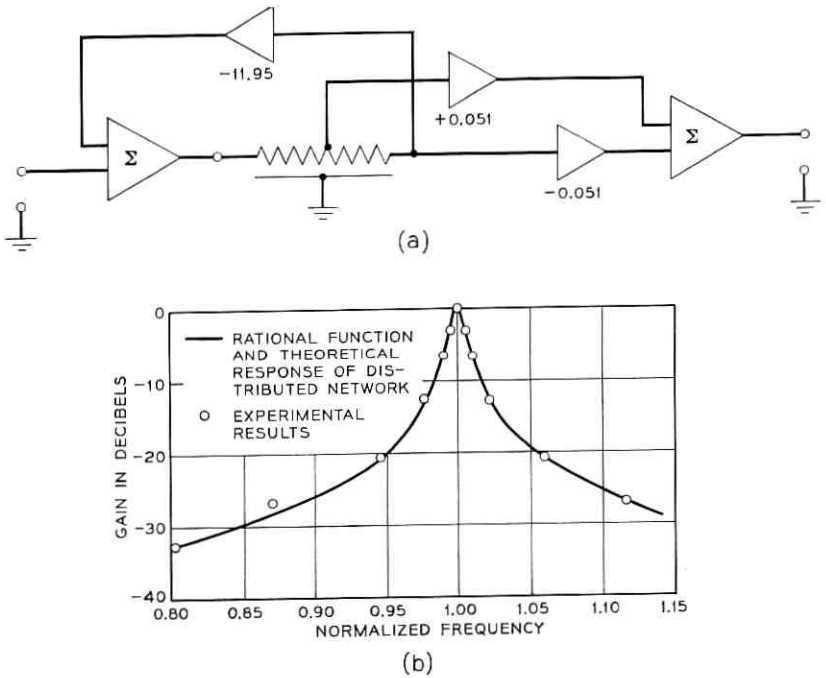


Fig. 7 — (a) RC line with feedback approximating $G_2(s)$. (b) Gain vs frequency for $G_2(s)$.

ratio of two resistors which track with temperature and τ can be stabilized by selecting the temperature coefficients of the resistive and capacitive materials of the line to be negatives of one another.*

Several final notes concerning the network are in order. By isolating the taps on the line with emitter followers when necessary, it is possible to reduce to two the number of operational amplifiers in the network used for combining and scaling, one for the feedback voltages and another for the feed-forward voltages. When several of these networks are cascaded, one of these two can be eliminated by using an operational amplifier from the succeeeding network. One RC line can be constructed with a large number of taps. Then by selecting the appropriate set of taps, the line can be used for a variety of purposes and at different frequencies.

* Tantalum resistors on a substrate can be made to track within ± 5 ppm/°C and RC products can be made to track within ± 30 ppm/°C.

Although only second-order examples were worked out and built, it is not unreasonable to expect that advances in building thin film RC lines and resistors using tantalum may eventually yield the stability of the various parameters required to make higher-order realizations possible.

VII. CONCLUSIONS

A network has beeen described which uses a single uniform RC line with feedback to approximate an arbitrary rational transfer function. The design procedure is simple as is the physical network. Theoretical calculations indicate that the transfer function realized by the RC line is an accurate approximation of the desired rational transfer function and measurements made on experimental circuits agree well with the theory.

VIII. ACKNOWLEDGMENT

The author thanks W. W. Armstrong for his helpful discussions.

APPENDIX

Derivation of Sensitivity Expressions

This appendix derives the sensitivity expressions given by (8) through (11). The sensitivity of quantity λ to parameter α is defined as

$$S_{\alpha}^{\lambda} = \frac{\partial \lambda}{\partial \alpha} \frac{\alpha}{\lambda}. \tag{14}$$

If s_j is a network pole and $D(s)$ is the denominator of (2), $D(s_j) = 0$. The equation $D(s_j) = 0$ defines s_j as an implicit function of the parameters in $D(s)$. By differentiating the equation $D(s_j) = 0$ with respect to a parameter α, we can determine the quantity $\partial s_j/\partial \alpha$. This result will hold for general values of the various parameters in $D(s)$. For the particular case when all the parameters in $D(s)$ have their nominal values, s_j will in fact be one of the desired network poles, that is, $s_j = \lambda_j$. Furthermore, the factorization used in going from (2) to (3) can then be used to simplify the expression for $\partial \lambda_j/\partial \alpha$. The sensitivity of λ_j to α is then determined by using (14). A similar procedure using the numerator of (2) gives the sensitivity functions of the zeros.

A. 1 Sensitivity of Poles to Feedback Coefficients

From (2) the denominator of $G(s)$ is

$$D(s) = \sum_{k=0}^{L} c_k \cosh (L - k)(\tau s)^{\frac{1}{2}}. \tag{15}$$

If s_j is a root of $D(s_j) = 0$, we have

$$\frac{\partial}{\partial c_i} D(s_i) = 0 = \left[\sum_{k=0}^{L} c_k (L - k) \frac{\sinh (L - k)(\tau s_i)^{\frac{1}{2}}}{2(\tau s_i)^{\frac{1}{2}}} \tau \right] \frac{\partial s_i}{\partial c_i} + \cosh (L - i)(\tau s_i)^{\frac{1}{2}}.$$

Setting all the parameters to their nominal values gives $s_i = \lambda_i$. As shown in (18), the term in brackets is nonzero if λ is a simple pole and $\tau\lambda_i \neq -n^2\pi^2$ where n is a nonzero integer. Therefore, solving the above equation gives

$$\frac{\partial \lambda_i}{\partial c_i} = -\frac{\cosh (L - i)(\tau\lambda_i)^{\frac{1}{2}}}{\sum_{k=0}^{L} c_k (L - k) \dfrac{\sinh (L - k)(\tau\lambda_i)^{\frac{1}{2}}}{(\tau\lambda_i)^{\frac{1}{2}}} \tau}. \tag{16}$$

As was done in (3), (15) can be factored, when all parameters have their nominal values and $P_k = \cosh (\tau\lambda_k)^{\frac{1}{2}}$, into

$$D(s) = \sum_{k=0}^{L} c_k \cosh (L - k)(\tau s)^{\frac{1}{2}} = 2^{L-1} \prod_{k=1}^{L} [\cosh (\tau s)^{\frac{1}{2}} - P_k]. \tag{17}$$

Differentiating this equation with respect to s gives

$$\sum_{k=0}^{L} c_k \frac{(L - k) \sinh (L - k)(\tau s)^{\frac{1}{2}}}{2(\tau s)^{\frac{1}{2}}} \tau = 2^{L-1} \sum_{m=1}^{L} \tau \frac{\sinh (\tau s)^{\frac{1}{2}}}{2(\tau s)^{\frac{1}{2}}} \prod_{\substack{k=1 \\ \neq m}}^{L} [\cosh (\tau s)^{\frac{1}{2}} - P_k],$$

and letting $s = \lambda_j$, we have

$$\sum_{k=0}^{L} c_k \frac{(L - k) \sinh (L - k)(\tau\lambda_j)^{\frac{1}{2}}}{2(\tau\lambda_j)^{\frac{1}{2}}} \tau = 2^{L-2} \tau \frac{\sinh (\tau\lambda_j)^{\frac{1}{2}}}{(\tau\lambda_j)^{\frac{1}{2}}} \prod_{\substack{k=1 \\ \neq i}}^{L} (P_i - P_k). \tag{18}$$

(18) is nonzero provided λ_i is a simple pole and $\tau\lambda_i \neq -n^2\pi^2$ where n is nonzero integer.

Using this and $\partial c_i / \partial a_i = -1$, $i \neq 0$, with (16) gives

$$S_{a_i}^{\lambda_i} = \frac{a_i \cosh (L - i)(\tau\lambda_i)^{\frac{1}{2}}}{2^{L-2}(\tau\lambda_i)^{\frac{1}{2}} \sinh (\tau\lambda_i)^{\frac{1}{2}} \prod_{\substack{k=1 \\ \neq i}}^{L} (P_i - P_k)}. \tag{19}$$

The numerator can be simplified by expressing a_i in terms of the P_k by using (17) and by factoring $\cosh(L-i)(\tau\lambda_j)^{\frac{1}{2}}$. The results are

$$\begin{aligned} a_0 &= 0 \\ a_1 &= 2\sum_{i=1}^{L} P_i \\ a_2 &= -L - 2\sum_{i=1}^{L} P_i\left[\sum_{\substack{k=1\\ \neq i}}^{L} P_k\right] \\ a_3 &= 2(L-1)\sum_{i=1}^{L} P_i + \frac{4}{3}\sum_{i=1}^{L} P_i\left\{\sum_{\substack{j=1\\ \neq i}}^{L} P_j\left[\sum_{\substack{k=1\\ \neq i,j}}^{L} P_k\right]\right\}^* \\ &\vdots \end{aligned} \tag{20}$$

and

$$\cosh(L-i)(\tau\lambda_j)^{\frac{1}{2}} = 2^{(L-i-1)}\prod_{k=1}^{(L-i)}\left\{P_j - \cos\left[\frac{2k-1}{2(L-i)}\pi\right]\right\}, \qquad i \neq L. \tag{21}$$

A. 2 Sensitivity of Poles to Tap Position

The tap positions are directly proportional to the integers k in (15). If k in (15) is replaced by l_k which is no longer constrained to be an integer and s_j is a root of the resulting $D(s)$, we have

$$D(s_j) = \sum_{k=0}^{L} c_k \cosh(l_L - l_k)(\tau s_j)^{\frac{1}{2}} = 0.$$

As in the previous section, differentiating with respect to l_k, solving for $\partial s_j/\partial l_k$, using the nominal values $l_k = k$ so that $s_j = \lambda_j$, and using (18), we have

$$S_{l_i}^{\lambda_i} = -\frac{ia_i \sinh(L-i)(\tau\lambda_i)^{\frac{1}{2}}}{2^{L-2}\sinh(\tau\lambda_i)^{\frac{1}{2}}\prod_{\substack{k=1\\ \neq j}}^{L}(P_j - P_k)} \qquad i \neq L \tag{22}$$

and

$$S_{l_L}^{\lambda_i} = -\frac{L\sum_{k=0}^{L} c_k \sinh(L-k)(\tau\lambda_i)^{\frac{1}{2}}}{2^{L-2}\sinh(\tau\lambda_i)^{\frac{1}{2}}\prod_{\substack{k=1\\ \neq j}}^{L}(P_j - P_k)}. \tag{22}$$

* Divide a_i by 2 if $i = L$.

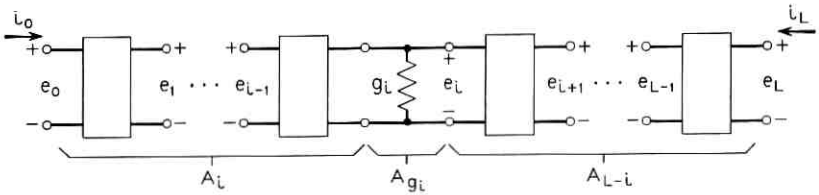


Fig. 8 — RC line model for loading analysis.

A.3 Pole Sensitivity to Tap Loading

To calculate the sensitivity of a pole λ_i to the loading at the ith tap by a conductance g_i, we assume that all other taps are not loaded, that is, $g_j = 0$ for $j \neq i$, and calculate the voltage transfer function from the input to the various taps. Having found these, we calculate the denominator of the system transfer function, $D(s)$, and proceed to calculate $\partial\lambda_i/\partial g_i$ in the same way as was done in Section *A.2*.

To calculate the voltage transfer function we will use the chain matrix description of the line which is

$$A_k = \begin{bmatrix} \cosh k(\tau s)^{\frac{1}{2}} & Z_0 \sinh k(\tau s)^{\frac{1}{2}} \\ Z_0^{-1} \sinh k(\tau s)^{\frac{1}{2}} & \cosh k(\tau s)^{\frac{1}{2}} \end{bmatrix}$$

where $Z_0 = (r/cs)^{\frac{1}{2}}$ and kd_0 is the length of the line. The line, loaded by g_i at the ith tap, can be considered as the cascade connection of an RC line of length id_0 connected to a two-port consisting solely of g_i which in turn is connected to an RC line of length $(L - i)d_0$ as seen in Fig. 8. The chain matrix of g_i is

$$A_{g_i} = \begin{bmatrix} 1 & 0 \\ g_i & 1 \end{bmatrix}.$$

From Fig. 8 and the properties of the chain matrix we have for $0 \leqq k < i$

$$\begin{bmatrix} e_k \\ i_k \end{bmatrix} = A_{i-k} A_{g_i} A_{L-i} \begin{bmatrix} e_L \\ -i_L \end{bmatrix} \triangleq B_k \begin{bmatrix} e_L \\ -i_L \end{bmatrix}$$

and for $i \leqq k \leqq L$

$$\begin{bmatrix} e_k \\ i_k \end{bmatrix} = A_{L-k} \begin{bmatrix} e_L \\ -i_L \end{bmatrix} \triangleq B_k \begin{bmatrix} e_L \\ -i_L \end{bmatrix}.$$

With $i_L = 0$ the above relations give the voltage transfer functions

from the line input to the kth tap as

$$G_k(s) = e_k(s)/e_0(s) = B_{k11}/B_{011}$$

where B_{k11} is the (1, 1) element of the matrix B_k .

The matrices B_k for $0 \leqq k < i$ and $i \leqq k \leqq L$, are

$$B_k = \begin{bmatrix} \left\{ \begin{array}{l} \cosh (L-k)(\tau s)^{\frac{1}{2}} \\ \quad + g_i Z_0 \sinh (i-k)(\tau s)^{\frac{1}{2}} \cosh (L-i)(\tau s)^{\frac{1}{2}} \end{array} \right\} & \left\{ \cdots \right\} \\ \{ \quad \cdots \quad \} & \{ \cdots \} \end{bmatrix}$$

and

$$B_k = \begin{bmatrix} \cosh (L-k)(\tau s)^{\frac{1}{2}} & \{ \cdots \} \\ \{ \cdots \} & \{ \cdots \} \end{bmatrix},$$

respectively, which give the gains $G_k(s)$ as

$$G_k(s) = \frac{\cosh (L-k)(\tau s)^{\frac{1}{2}} + g_i Z_0 \sinh (i-k)(\tau s)^{\frac{1}{2}} \cosh (L-i)(\tau s)^{\frac{1}{2}}}{\cosh L(\tau s)^{\frac{1}{2}} + g_i Z_0 \sinh i(\tau s)^{\frac{1}{2}} \cosh (L-i)(\tau s)^{\frac{1}{2}}},$$

for $0 \leqq k < i$, and for $i \leqq k \leqq L$

$$G_k(s) = \frac{\cosh (L-k)(\tau s)^{\frac{1}{2}}}{\cosh L(\tau s)^{\frac{1}{2}} + g_i Z_0 \sinh i(\tau s)^{\frac{1}{2}} \cosh (L-i)(\tau s)^{\frac{1}{2}}}.$$

By using these equations in the expression for the gain of the feedback structure and multiplying the numerator and denominator of this expression by B_{011}, we have the following expression for the denominator, $D(s)$.

$$\begin{aligned} D(s) &= \sum_{k=0}^{L} c_k B_{k11} \\ &= \sum_{k=0}^{L} c_k \cosh (L-k)(\tau s)^{\frac{1}{2}} \\ &\qquad + g_i Z_0 \cosh (L-i)(\tau s)^{\frac{1}{2}} \sum_{k=0}^{i-1} c_k \sinh (i-k)(\tau s)^{\frac{1}{2}}. \end{aligned}$$

Now proceeding as in the previous sections, let $D(s_j) = 0$ and differentiate with respect to g_i to get

$$\frac{\partial s_j}{\partial g_i} \sum_{k=0}^{L} \frac{c_k (L-k)\tau}{2(\tau s_j)^{\frac{1}{2}}} \sinh (L-k)(\tau s_j)^{\frac{1}{2}}$$

$$+ \frac{R}{(\tau s_j)^{\frac{1}{2}}} \cosh (L-i)(\tau s_j)^{\frac{1}{2}} \sum_{k=0}^{i-1} c_k \sinh (i-k)(\tau s_j)^{\frac{1}{2}} + g_i(\cdots) = 0.$$

With $g_i = 0$, $s_j = \lambda_j$ and the above gives with (18)

$$S_{g_i}^{\lambda_j} = -\frac{g_i R \cosh (L - i)(\tau\lambda_j)^{\frac{1}{2}} \sum_{k=0}^{i-1} c_k \sinh (i - k)(\tau\lambda_j)^{\frac{1}{2}}}{2^{L-2}\tau\lambda_j \sinh (\tau\lambda_j)^{\frac{1}{2}} \prod_{\substack{k=1 \\ \neq j}}^{L} (P_j - P_k)}.$$

A. 4 Sensitivity to RC Product Changes

It is assumed that the product RC $= \tau$ of the line is uniform but not correct. The sensitivity of the poles to changes in τ is easily seen to be

$$S_\tau^{\lambda_j} = -1$$

since τ always appears multiplying s in the transfer function and is a frequency scale factor.

REFERENCES

1. Chirlian, F. M., *Integrated and Active Network Analysis and Synthesis,* Englewood Cliffs, New Jersey: Prentice-Hall, 1967.
2. Heizer, K. W., "Distributed RC Networks with Rational Transfer Functions," IRE Trans. Circuit Theory, *CT-9,* No. 4 (December 1962), pp. 356–362.
3. Heizer, K. W., "Rational Parameters with Distributed Networks," IEEE Trans. Circuit Theory, *CT-10,* No. 4 (December 1963), pp. 531–532.
4. Barker, D. G., "Synthesis of Active Filters Employing Thin Film Distributed Parameter Networks," 1965 IEEE Int. Conv. Record, *13,* pt. 7, pp. 119–126.
5. Woo, B. B. and Hove, R. G., "Synthesis of Rational Transfer Functions with Thin Film Distributed Parameter RC Active Networks," Proc. National Electronics Conference, Chicago, 1965.
6. Fu Y. and Fu, J. S., "Synthesis of Active Distributed RC Networks," IEEE Trans. Circuit Theory, *CT-13,* No. 2 (September 1966), pp. 259–264.
7. Kerwin, W. J., "Analysis and Synthesis of Active RC Networks Containing Distributed and Lumped Elements," Ph.D. dissertation, Stanford University, Stanford, Calif., 1967.
8. Bello, V. and Gausi, M. S., "Design of Active Distributed RC Low-Pass Networks," Eleventh Midwest Symp. Circuit Theory, Lafayette, Ind., May 13–14, 1968.
9. Wyndrum, R. W., Jr., "Active Distributed RC Networks," IEEE J. Solid State Circuits, *SC-3,* No. 3 (September 1968), pp. 308–310.
10. Chen, W. I. H., "Synthesis of Transversal **RC** Filters in the Time Domain," Ph.D. dissertation, New York: Columbia University, 1967.
11. Chen, W. I. H. and Levine, R., "Computer Design of Transversal Filters Using Thin Film RC Dispersion," Tenth Midwest Symp. Circuit Theory, Lafayette, Ind., May 1967.
12. Doetsch, G., *Guide to the Applications of Laplace Transforms,* London: D. Van Nostrand, 1961, pp. 195–197, and 217.
13. Doetsch, G., *Handbuch der Laplace-Transformation,* Vol. 1, Birkhäuser, 1950, pp. 224, 276.

Contributors to This Issue

NOACH AMITAY, B.Sc., 1953, and Dipl. Ing., 1954, Technion, Israel Institute of Technology; M.Sc., 1957, Ph.D., 1960, Carnegie Institute of Technology; Bell Telephone Laboratories, 1962—. Mr. Amitay has been involved in electronic instrumentation, magnetic devices, tapered transmission lines, and electromagnetic field theory as applied to phased array antennas and wave scattering. Member, IEEE and Sigma Xi.

WERNER BLEICKARDT, Dipl. El. Ing., 1960, and Dr. Sc. Techn., 1963, both from Swiss Federal Institute of Technology, Zurich, Switzerland; Hasler Ltd., Bern, Switzerland, 1961–1965; Bell Telephone Laboratories, 1965—. With Hasler Ltd., Mr. Bleickardt worked on the design of pulse code modulation terminals for telephony. At Bell Telephone Laboratories he has been concerned with analytical studies of synchronization in PCM systems and is presently engaged in the design of high-speed digital circuits for PCM multiplex terminals. Member, IEEE.

T. M. BUCK, B.S., 1942, Muskingum College; M.S., 1948, and Ph.D., 1950, University of Pittsburgh; National Lead Co., 1950–52; Bell Telephone Laboratories, 1952—. Mr. Buck's work has been concerned mainly with chemical problems and surface properties of semiconductor materials and devices. He participated in developing silicon p-n junction radiation detectors for the *Telstar*® communications satellites and other space vehicles. He has been supervisor of a group developing materials and processes for silicon diode array targets for TV camera tubes. He is studying solid surfaces and ion implantation. Member IEEE, Sigma Xi, Phi Lambda Upsilon, Sigma Pi Sigma.

H. C. CASEY, JR., B.S. (E.E.), 1957, Oklahoma State University; Ph.D. (E.E.), 1964, Stanford University; Hewlett-Packard Co., 1957–62; Bell Telephone Laboratories, 1964—. Mr. Casey has been studying light absorption and emission in semiconductors and is now studying impurity solubility, diffusion, and injection luminescence in III–V compound semiconductors. Member, American Physical Society, Sigma Xi, Eta Kappa Nu, Sigma Tau.

M. H. Crowell, B.S. (E.E.), 1956, Pennsylvania State University; M.S. (E.E.), 1960, New York University; Bell Telephone Laboratories, 1956—. From 1956 to 1963 he was engaged in designing and developing an electron beam encoder for a high-speed pulse code modulation system. Since 1963 he has been engaged in a study of optical modulators which use the electric field-induced shift in the absorption spectrum of a semiconductor. During this period he has also studied the characteristics of intracavity modulated lasers. Member, Tau Beta Pi, Eta Kappa Nu.

John V. Dalton, B.S., 1964, Rutgers University; RCA Semiconductor Division, 1958–62; Bell Telephone Laboratories, 1962—. Mr. Dalton has worked on semiconductor device development, including studies of ion migration in thin insulator films on semiconductor surfaces and developing a silicon diode array camera tube target for *Picturephone*® visual telephone. Member, ECS.

Corrado 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.

Louis H. Enloe, B.S.E.E., 1955, M.S.E.E., 1956, Ph.D. (E.E.), 1959, University of Arizona, Tucson. Bell Telephone Laboratories 1959—. His early work was in modulation and noise theory in connection with space communications. Later work has been with lasers, coherent light, and holography with emphasis upon communication and display. He is head of the Opto-Electronics Research Department. Member, IEEE, Phi Kappa Phi, Sigma Xi, Tau Beta Pi, Pi Mu Epsilon, Sigma Pi Sigma.

Victor Galindo, B.S. (E.E.), 1954, New York University; M.S. (E.E.), 1962, and Ph.D. (E.E.), 1964, University of California, Berkeley; Hughes Aircraft Company, 1954–1957, 1958–1960; M.I.T. Lincoln Laboratory, 1957–1958; Bell Telephone Laboratories, 1964—. Mr. Galindo has been engaged in applying electromagnetic theory to studies of microwave transmission devices, antennas, and phased arrays. Member, Eta Kappa Nu, Tau Beta Pi, IEEE.

Eugene I. Gordon, B.S. in Physics, 1952, City College of New York; Ph.D. in Physics, 1957, Massachusetts Institute of Technology; Bell

Telephone Laboratories, 1957—. Mr. Gordon is Director of the Electro-optical Device Laboratory concerned with gas lasers and their application as well as devices for the *Picturephone*® visual telephone. He is an associate editor of the IEEE Journal of Quantum Electronics. Member, APS, Phi Beta Kappa, Sigma Xi.

WILLIAM C. JAKES, JR., B.S.E.E., 1944, M.S.E.E., 1947, Ph.D. (E.E.), 1949, Northwestern University; Bell Telephone Laboratories 1949—. Mr. Jakes was originally engaged in research on microwave antennas and propagation. In 1959 he was assigned to the Project Echo satellite communication experiment as project engineer and later participated in the Telstar® experiment. He received an honorary doctorate from Iowa Wesleyan College and the Northwestern University Alumni Award in 1962 for contributions to satellite communications. In 1963 he became Head of the Radio Transmission Research Department, where fundamental studies of the use of microwaves for high-capacity mobile telephone systems are currently in progress. Fellow, IEEE; Member, Commission 2, U.R.S.I.; Eta Kappa Nu, Pi Mu Epsilon, Sigma Xi.

JAMES MCKENNA, B.Sc. (Math), 1951, Massachusetts Institute of Technology; Ph.D., (Math), 1961, Princeton University; Bell Telephone Laboratories, 1960—. Mr. McKenna has done research in quantum mechanics and classical electromagnetic theory. He is now doing a study of optical waveguides.

J. A. MORRISON, B.Sc., 1952, King's College, University of London; Sc.M., 1954 and Ph.D., 1956, Brown University; Bell Telephone Laboratories, 1956—. Mr. Morrison has been doing mathematical research in a variety of problems in mathematical physics and applied mathematics. His recent interests have included perturbation techniques, especially averaging methods as applied to perturbed nonlinear oscillations. He has been a Visiting Professor of Mechanics at Lehigh University during the fall semester 1968–69. Member, American Mathematical Society, SIAM, Sigma Xi.

R. E. PARKIN, B.Sc. (Eng.), 1962, Ph.D., 1965, Imperial College (University of London); Bell Telephone Laboratories, 1965—. Mr. Parkin is concerned with computer-aided design techniques with particular emphasis on hybrid integrated technology. Associate Member, IEE.

STEPHEN 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, 1965–1968. Mr. Rappaport pursued theoretical studies of signal processing for radar and data communications. Now he is on leave of absence from the Laboratories to be with the State University of New York at Stony Brook, N. Y. Member, IEEE, Tau Beta Pi, Eta Kappa Nu, Sigma Xi.

STEPHEN O. RICE, B.S., 1929, D.Sc. (Hon.), 1961, Oregon State College; Graduate Studies, California Inst. of Tech., 1929–30 and 1934–35; Bell Telephone Laboratories, 1930—. In his first years at the Laboratories, Mr. Rice was concerned with nonlinear circuit theory, especially with methods of computing modulation products. Since 1935 he has served as a consultant on mathematical problems and in investigations of telephone transmission theory, including noise theory, and applications of electromagnetic theory. He was a Gordon McKay Visiting Lecturer in Applied Physics at Harvard University for the Spring, 1958, term. He is a fellow of the IEEE and received that society's Mervin J. Kelly Award in 1965.

CHARLES B. RUBINSTEIN, B.E.E., 1959, City College of New York; M.E.E., 1962, New York University; Bell Telephone Laboratories, 1959—. Mr. Rubinstein has been engaged in investigations of magneto-optical phenomena with particular reference to optical rotators for device use, and he has been concerned with the use of holography in visual communication systems. He is now engaged in research on color perception and color rendition in complex scenes. Member American Physical Society, Optical Society of America, Eta Kappa Nu, Tau Beta Pi.

DAVID A. SPAULDING, A.B., 1959, M.S., 1960, Dartmouth College; M.S., 1961, Ph.D., 1965, Stanford University; Bell Telephone Laboratories, 1967—. Mr. Spaulding has been concerned with network studies for data transmission systems. Member IEEE, Phi Beta Kappa, Sigma Xi.

MICHAEL YAMIN, B.S., 1949, Polytechnic Institute of Brooklyn; Ph.D., 1952, Yale University; Mellon Institute, 1953–56; Bell Telephone Laboratories, 1957—. Dr. Yamin has been working on surface problems related to semiconductor device development, especially those concerned with thin passivating films such as silicon dioxide.